Stevens Institute of Technology
School of Business
Business Intelligence & Analytics Program

A Snapshot of Data Science
Student Poster Presentations
Corporate Networking Event – November 27, 2018
Forward

This document reproduces the posters presented by students of the Business Intelligence & Analytics (BI&A) program at a Corporate Networking event held at Stevens Institute on November 27, 2018. The event was attended by approximately 80 company representatives and approximately 200 students and faculty members.

The posters were presented by students at all stages in their academic programs from their first semester through their final semester. The research described in each poster was conducted under the guidance of a faculty member. The broad range of research topics and methodologies exhibited by the posters in this document reflects the diversity of faculty research interests and the practical nature of our program.

For background, the first poster describes the BI&A program. Founded in spring 2012, with just 4 students, the program now has over 180 full-time and part-time masters of science students and 50 graduate certificate students. For the second year in a row it has been ranked 7th in the nation by The Financial Engineer. As illustrated in the first poster, a distinctive feature of the program is its three-layer structure. In the professional skills layer, business and communication skills are developed through workshops, talks by industry leaders and an active student club. In the second layer, the 12-course curriculum covers the concepts and tools associated with database management, data warehousing, data and text mining, web mining, social network analytics, optimization and risk analytics. The curriculum culminates in a capstone course in which students work on a research project – often in conjunction with industry associates. Finally, in the technical skills layer, students attend a series of free weekend boot weekend camps that provide training in industry-standard software packages, such as SQL, R, SAS, Python and Hadoop.

The 69 student posters in this document represent a broad array of research projects. We are proud of the quality and innovativeness of our students’ research and of their hard work and enthusiasm without which this event would have been impossible.

Chris Asakiewicz, Ted Stohr and Alkis Vazacopoulos
Business Intelligence & Analytics Program
Stevens Institute of Technology
www.stevens.edu/business/bia
<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Student Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1*</td>
<td>The Business Intelligence &amp; Analytics Program</td>
<td>BI&amp;A Faculty, Weiyi Chen, Shiyue Ren</td>
</tr>
<tr>
<td>2</td>
<td>HSFL: Technology to Support Teaching &amp; Research</td>
<td>HSFL Faculty</td>
</tr>
<tr>
<td>3</td>
<td>Integrated Marketing Plan: Pennsylvania Market LLC</td>
<td>Shuting Zhang</td>
</tr>
<tr>
<td>4*</td>
<td>UBS WM Branch Prediction</td>
<td>Shuting Zhang, Harsh Kava</td>
</tr>
<tr>
<td>5</td>
<td>Employee Branding Research Glassdoor.com Company reviews analysis</td>
<td>Shuting Zhang, Siyan Zhang</td>
</tr>
<tr>
<td>6</td>
<td>Analysis of Olympic athletes' data in 120 years history</td>
<td>Yanzhao Liang, Yingjian Song, Hao Xu, Zhihao Yang</td>
</tr>
<tr>
<td>7</td>
<td>Predicting the outcome of a shot</td>
<td>Haitao Liu, Yang Liu, Jiawei Xue</td>
</tr>
<tr>
<td>8*</td>
<td>Predicted churn rate reduction for Telephone Services using Marketing Analytics</td>
<td>Jiawei Xue, Sucharitha Batchu, Suguna Bontha, Suprajah Suresh, Yang Liu</td>
</tr>
<tr>
<td>9</td>
<td>Budget allocation optimization for natural disaster preparation</td>
<td>Haohan Hu, Jiahao Shi, Lianhong Deng, Nifan Yuan</td>
</tr>
<tr>
<td>10</td>
<td>Revenue and Cost Optimization for a Clothing Supply Chain</td>
<td>Liran Zhang, Mingxin Zheng, Weifeng Li, Zeyu Shao</td>
</tr>
<tr>
<td>11</td>
<td>Trip Master: A configuration tool for designing travel experience in New York</td>
<td>Shunyu Zheng, Sisi Xiong, Xinghong Liu, Yang Wu, Yiyi Liang</td>
</tr>
<tr>
<td>12</td>
<td>Predicting Airbnb Prices in Washington D.C.</td>
<td>Ankur Morbale, Matthew Rudolph, Kyle Eifler, Victoria Piskarev, Arthur Krivoruk, Gaurav Venkataraman, Sarvesh Gohil</td>
</tr>
<tr>
<td>13</td>
<td>Function Approximation Using Evolutionary Polynomials</td>
<td>Aleksandr Grin</td>
</tr>
<tr>
<td>14</td>
<td>Correlating Long-Term Innovation with Success in Career Progression</td>
<td>Adam Coscia</td>
</tr>
<tr>
<td>15</td>
<td>Car Sales Analysis</td>
<td>Lulu Zhu, Xin Chen, Yifeng Liu, Yuyi Yan</td>
</tr>
<tr>
<td>16</td>
<td>NBA Data Visualization Analysis</td>
<td>Xin Chen, Xiaohao Su, Xiang Yang</td>
</tr>
<tr>
<td>17</td>
<td>Hot Wheels Analysis at NYC Yellow Taxi</td>
<td>Abhitej Kodali, Nikhil Lohiya</td>
</tr>
<tr>
<td>18</td>
<td>Customer Revenue Prediction for Google Store Products</td>
<td>Abhitej Kodali, Nikhil Lohiya</td>
</tr>
<tr>
<td>19</td>
<td>Clustering Large Cap Stocks During Different Phases of the Economic Cycle</td>
<td>Nikhil Lohiya, Raj Mehta</td>
</tr>
<tr>
<td>20*</td>
<td>FIN-FINICKY : Financial Analyst’s Toolkit</td>
<td>Nikhil Lohiya</td>
</tr>
<tr>
<td>21</td>
<td>Group Emailing using Robotic Process Automation</td>
<td>Pallavi Naidu, Abhitej Kodali</td>
</tr>
<tr>
<td>22</td>
<td>Cognitive Application to Determine Adverse Side Effects of Vaccines</td>
<td>Pallavi Naidu, Kathy Chowaniec, Krishanu Agrawal</td>
</tr>
<tr>
<td>23</td>
<td>Predict Potential Customers by Analyzing Bank's Telemarketing Data</td>
<td>Shreyas Menon, Pallavi Naidu</td>
</tr>
<tr>
<td>Index</td>
<td>Poster Title</td>
<td>Authors</td>
</tr>
<tr>
<td>-------</td>
<td>------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
</tr>
<tr>
<td>25</td>
<td>Lending Club – How to Forecast the Loan Status of Loan Applications.</td>
<td>Tsen-Hung Wu, Shreyas Menon</td>
</tr>
<tr>
<td>26</td>
<td>Visualization of Chicago Crime</td>
<td>Zihan Chen, Xuanyan Li</td>
</tr>
<tr>
<td>27</td>
<td>Predictive Model for House Pricing</td>
<td>Zihan Chen, Xuanyan Li</td>
</tr>
<tr>
<td>28</td>
<td>DOE for Amazon Recommendation Email</td>
<td>Siyan Zhang, Biyan Xie, Xuanyan Li</td>
</tr>
<tr>
<td>29</td>
<td>Predicting trends in bike sharing program</td>
<td>Zixuan Wang, Shuqiong Chen, Kevin Walsh</td>
</tr>
<tr>
<td>30</td>
<td>NBA player management Optimization</td>
<td>Jingchen Lan, Shan Jiang, Shuqiong Chen, Zixuan Wang</td>
</tr>
<tr>
<td>31</td>
<td>Improvement of medical wire manufacturing</td>
<td>Zixuan Wang, Jingchen Lan</td>
</tr>
<tr>
<td>32</td>
<td>Identify the safety level of precincts in New York City</td>
<td>Tianyu Liu, Chen Liao, Yu Hong, Xiangxiang He</td>
</tr>
<tr>
<td>33</td>
<td>Credit Analysis: Loan Default Prediction</td>
<td>Jimit Sanghvi, Ho Ben Wesley Cheung, Xinhang Wang</td>
</tr>
<tr>
<td>34</td>
<td>Better Photography using Design of Experiments</td>
<td>Kumar Bipulesh, Ping-Lun Yeh, Sibo Xu, Sanjay Kumar Pattanayak</td>
</tr>
<tr>
<td>35</td>
<td>Driver Safety using CNN &amp; Transfer Learning</td>
<td>Kumar Bipulesh</td>
</tr>
<tr>
<td>36</td>
<td>Machine Learning to Predict US GC</td>
<td>Sanjay Pattanayak, Smriti Vimal</td>
</tr>
<tr>
<td>37</td>
<td>UBS Pitch 2018 1st Prize Winners Machine Learning &amp; Automation</td>
<td>Monica Vijaywargi, Poojan Gajera, Rohan Gala, Sanjay Pattanayak, Xunyan Li</td>
</tr>
<tr>
<td>38</td>
<td>Classifying Restaurant Rating</td>
<td>Xiaojun Zhu, Jhao-Han Chen, Haiping Sun</td>
</tr>
<tr>
<td>39</td>
<td>Consumer Analytics for Restaurant Preferences using Yelp User Reviews</td>
<td>Xiaojun Zhu, Haodong Zhao, Yuhan Su</td>
</tr>
<tr>
<td>40</td>
<td>Portfolio Optimization using Python</td>
<td>Jhao-Han Chen, Jiamei Wang, Liang An, Ming Wei, Xuanzhu Luo</td>
</tr>
<tr>
<td>41</td>
<td>Radiology Assistant</td>
<td>Amit Kumar, Jayesh Mehta, Yash Wanve</td>
</tr>
<tr>
<td>42</td>
<td>Predicting Change in Bitcoin Prices Using Public Sentiment of the Cryptocurrency on Twitter</td>
<td>Thomas Treshock, Michael Ficke, Robert Frusina</td>
</tr>
<tr>
<td>43</td>
<td>Surface-Enhancement Raman Scattering of Urine for Risk Assessment of Prostate Cancer</td>
<td>Yiwei Ma, Yanbo Wang, Guohao Gao</td>
</tr>
<tr>
<td>44</td>
<td>Customer Churn Rate Analytics: Predictive Analysis Model for Retaining Customers</td>
<td>Shangjun Jiang, Shan Jiang, Hongyi Chen</td>
</tr>
<tr>
<td>45</td>
<td>Analysis of Absenteeism at Works</td>
<td>Ian Christie, Rossano Rossi, Neelansh Prasad, Steven Toal</td>
</tr>
<tr>
<td>46</td>
<td>Reddit user’s top ten attentions about world news</td>
<td>Yu Hong, Yuyang Tian, Mingjun Han, Ran Yi</td>
</tr>
<tr>
<td>47</td>
<td>Analysis of avocado based on other data in multiple US markets</td>
<td>Tianyu Liu, Yuyang Tian, Mingjun Han, Ran Yi</td>
</tr>
<tr>
<td>Index</td>
<td>Title</td>
<td>Authors</td>
</tr>
<tr>
<td>-------</td>
<td>----------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
</tr>
<tr>
<td>48</td>
<td><strong>Supply Chain Analytics Strategy Development</strong></td>
<td>Chad Vaske</td>
</tr>
<tr>
<td>49</td>
<td><strong>UBS Pitch 2018 1st Prize Winners: Branch Location Selection using Machine Learning</strong></td>
<td>Minyan Shao, Yuankun Nai, Fan Yang</td>
</tr>
<tr>
<td>50</td>
<td><strong>Predicting Overall Health from Behavioral Risk Factor Surveillance Survey Data</strong></td>
<td>Malik Mubeen, Erika Deckter</td>
</tr>
<tr>
<td>51</td>
<td><strong>Financial Distress Assessment by Text Mining Risk Factors of 10k Report</strong></td>
<td>Siwei Wang, Haochen Liu, Yuzhen He, Yiru Yang</td>
</tr>
<tr>
<td>52</td>
<td><strong>Ship Detection Along Maritime Shipping Routes with Convolutional Neural Networks (CNNs) Methodology</strong></td>
<td>Kevin Walsh, Erdong Xia, Ping-Lun Yeh</td>
</tr>
<tr>
<td>53</td>
<td><strong>Analysis of Topic Trends in Biomaterial Research</strong></td>
<td>Jingsi Gong, Yuhan Hu, Shivi Jain, Shuai Wang</td>
</tr>
<tr>
<td>54</td>
<td><strong>Prediction of Black Friday Sale Using Machine Learning</strong></td>
<td>Minzhe Huang, Shuo Jin, Jiaqiang Lu, Raj Mehta, Jingmiao Shen</td>
</tr>
<tr>
<td>55</td>
<td><strong>Who are the most important authors in the Biomaterial Research?</strong></td>
<td>Raphaël Presberg, Niraj Chaurasia, Medhavi Uniyal</td>
</tr>
<tr>
<td>56</td>
<td><strong>Fraud Detection for credit card transactions</strong></td>
<td>Pranav Prajapati, Sonali Jhohri, Rumeng Zuo, Qian Liu</td>
</tr>
<tr>
<td>57</td>
<td><strong>What makes a good TED talk?</strong></td>
<td>Suprajah Suresh, Sucharitha Batchu, Aneesh Shinde</td>
</tr>
<tr>
<td>58</td>
<td><strong>Optimizing London Fire Station Resources to Better Serve the Community</strong></td>
<td>Sonali Johari, Pranav Prajapati, David McFarland, Erika Deckter, Marielle Nwana</td>
</tr>
<tr>
<td>59*</td>
<td><strong>Stack Watson: The Friendly S.O. Bot</strong></td>
<td>Smit Mehta, Xue (Calvin) Cao</td>
</tr>
<tr>
<td>60</td>
<td><strong>Optimal Portfolio Rebalancing using Reinforcement Learning</strong></td>
<td>John-Craig Borman</td>
</tr>
<tr>
<td>61</td>
<td><strong>Global Burden of Tuberculosis, 1990-2013</strong></td>
<td>Erdong Xia, He Li, Wenlei Feng</td>
</tr>
<tr>
<td>62</td>
<td><strong>Can We Predict Wine Quality with Machine Learning?</strong></td>
<td>Patrick Curran, Smit Raval</td>
</tr>
<tr>
<td>63*</td>
<td><strong>Opinion Mining: Tracking public emotions on Twitter</strong></td>
<td>Dhaval Sawlani</td>
</tr>
<tr>
<td>64</td>
<td><strong>Predicting customer churn for a bank using logistic regression</strong></td>
<td>Rushabh Vakharia, Ameya Swar, Rashmi Khurana</td>
</tr>
<tr>
<td>65</td>
<td><strong>Quantum Computing Applied to Financial Portfolio Optimization</strong></td>
<td>John Robson</td>
</tr>
<tr>
<td>66*</td>
<td><strong>Object Detection in Autonomous Driving Car</strong></td>
<td>Amit Agarwal, Pravin Mukare, Taru Tak</td>
</tr>
<tr>
<td>67</td>
<td><strong>Energy Price Forecasting using Deep Recurrent Neural Networks</strong></td>
<td>John Robson</td>
</tr>
<tr>
<td>68</td>
<td><strong>Dark Pool Stock Trading Prediction Model</strong></td>
<td>Z. Yao, X. Chen</td>
</tr>
<tr>
<td>69</td>
<td><strong>Intelligent Targeting - Bank Marketing Data</strong></td>
<td>Suprajah Suresh, Sucharitha Batchu, Aneesh Shinde</td>
</tr>
</tbody>
</table>
Master of Science
Business Intelligence & Analytics

The Master of Science in Business Intelligence and Analytics (BI&A) is a 36-credit STEM program designed for individuals who are interested in applying analytical techniques to derive insights and predictive intelligence from vast quantities of data.

The first of its kind in the tri-state area, the program has grown rapidly. We have approximately 200 master of science students and another 50 students taking 4-course graduate certificates. The program has increased rapidly in quality as well as size. The average test scores of our student body is top 75 percentile. We have been ranked #7 among business analytics programs in the U.S. by The Financial Engineer for the last 2 years.

**PROGRAM ARCHITECTURE**

- **Social Skills**
  - Written & Oral Skills
  - Team Skills
  - Job Skills Workshops
  - Industry speakers
  - Industry-mentored projects

- **Disciplinary Knowledge**
  - SQL, SAS, R, Python
  - Software "Boot" Camps
  - Course Projects
  - Industry Projects

- **Technical Skills**
  - Hadoop, SAS, DB2, Cloudera
  - Trading Platforms: Bloomberg
  - Data Sets: Thomson-Reuters, Custom

- **Infrastructure**
  - Hanlon Lab – Hadoop for Professionals
  - MOOCs
  - Workshops
  - Team Skills
  - Job Skills Workshops
  - Industry speakers
  - Industry-mentored projects

**PROGRAM PHILOSOPHY/OBJECTIVES**

- Develop a nurturing culture
- Race with the MOOCs
- Develop innovative pedagogy
- Migrate learning upstream in the learning value chain
- Continuously improve the curriculum
- Use analytics competitions
- Improve placement
- Partner with industry

**CURRICULUM**

- **Organizational Background**
  - Financial Decision Making

- **Data Management**
  - Data Management
  - Data Warehousing & Business Intelligence
  - Data and Information Quality *

- **Optimization and Risk Analysis**
  - Optimization & Process Analytics
  - Risk Management Methods & Simulation.*

- **Machine Learning**
  - Data Analytics & Machine Learning
  - Advanced Data Analytics & Machine Learning*

- **Statistics**
  - Multivariate Data Analytics
  - Experimental Design

- **Social Network Analytics**
  - Network Analytics
  - Web Mining

- **Management Applications**
  - Marketing Analytics*
  - Supply Chain Analytics*

- **Big Data Technologies**
  - Data Stream Analytics*
  - Big Data Technologies
  - Cognitive Computing*

- **Practicum**
  - Practicum in Analytics

- **Electives** - Choose 2 out of 11

**STATISTICS**

- **Demographics**
  - Full-time: 180
  - Part-time: 19
  - Gender:
    - Female: 44%
    - Male: 56%

- **Admissions**
  - 2013F: 101
  - 2014F: 157
  - 2015F: 351
  - 2016F: 591
  - 2017F: 725
  - Accepted: 48, 84, 124, 287, 364
  - Rejected: 34, 34, 186, 257, 307
  - In system/other: 19, 39, 41, 46, 53

- **Placement**
  - Starting Salaries (without signing bonus):
    - $65 - 140K Range
    - $84K Average
    - $90K (finance and consulting)
  - Data Scientists: 23% Data Analysts: 30% Business Analysts: 47%

  - Our students have accepted jobs at for example:
Lab Projects

Hanlon Financial Systems Lab provides hardware and software techniques to support academic research, including:
- Academic research projects
- Joint projects with other divisions
- Master thesis projects

Research Projects:

- Rare Events:
  - We developed a multivariate framework for the detection and analysis of rare events in high-frequency financial data. The connection between the rare events and liquidity facilitates the further development of market liquidity indices and early warning systems for critical market events.
  - Pricing Volatility Derivatives
    - "We propose a detailed structural approximation general stochastic volatility models. The method is applied to price various volatility derivatives, for example variance swaps.

- Market Liquidity
  - We are trying to investigate how different liquidity measures behave with respect to each other as well as what is the dimensionality number of liquidity measures can be reduced without loss of information. In order to address the preceding question, we utilized correlation based clustering method.

- Robotics Application Platform: Integrated Development (RAPID)
  - The project is an effort to put together the up-to-date software and hardware techniques to S&E related robotics domain (platform for future applications. The robotics platform is designed to operate completely independent of human operator. Several targeted applications include consumer electronics devices and multiple areas of research.

Joint Projects

- JHF Projects
  - The goal of this project is to create a test-bed platform for simulating the behavior of modern high-frequency (HF) financial markets with much greater realism than the current models. The SHFT Platform operates with live, real-time, tick-level market data.
  - Surge Projects
    - The objective of this project is designing models which could simulate the reliability of each prediction based on observation during short time "time-slices" to select the best forecasts. This is a joint project between Hanlon Financial System Laboratory and the Davidson Laboratory.

- Predicting S&P500 Component
  - Financial Physics (PGY 427) Lab, (June 2016). The primary goal of this project is to develop a model to help predict the next run S&P500 Company to be added to the index. The objective is to predict the set of companies that could be added to or deleted from the S&P 500 index to gain profit from taking positions in these companies before the announcement of the constituents.

- Copula Methods in CDO Tranche Dependence Structure
  - Jingqi Zhao, Nan Zhao, and Zhenyu Ma. Master in Financial Engineering. Graduated May 2016. This study proposes a CDO tranche valuation based on elliptical copulas and Archimedean copulas. The intensity model by Duan and K (1999) for default probability is assumed rather than structural model by Merton (1974). Furthermore, the recovery rate here is set at 40%. It applies a bottom-up method, one factor Gaussian copula model and top-down method, Archimedian copula model, to calibrate dependence structure between single name CDOs in the pool.

- Xiuping Zheng and Wenting Zhao. Master in Financial Engineering. Graduated May 2016. The Heston lattice-based volatility model can explain volatility smile and skewness while the Black-Scholes model assumes constant volatility. With the explicit option pricing formula derived by Heston. This study uses the Local-Squares Fit to calibrate and to backtest the result. Using this method in the real market behavior, it can provide the recommendation of choosing initial parameter for stocks in different market behavior.

A new lab (Hanlon Lab) is under construction and will be opened for courses in Fall 2016. If you wish to discuss support for your project or possible collaboration with the Hanlon Financial Systems Laboratory, please contact Demis@stevens.edu or dhsu@stevens.edu.

---

Lab Courses

- FE506 Technical Writing in Finance
  - In this course the students learn to write research type articles for financial. It is an integral part of the FE500 special problems in Financial Engineering.

- FE512 Database Engineering
  - Teaches the foundations of database systems and SQL in applications in financial institutions.

- FE514 VBA In Finance
  - Teaches the students Excel usage at a high level using VBA for front office applications in financial institutions.

- FE515 R In Finance
  - Provides an extended coverage of the Bloomberg terminals with focus on financial data for derivatives.

- FE516 MATLAB/fore Finance
  - Fundamental MATLAB programming using financial data and applications.

- FE517 SAS for Finance
  - Fundamental SAS programming using financial data and applications.

- FE520 Python for Finance
  - Fundamental Python programming using financial data and applications.

- FE522 C++ Programming in Finance
  - Teaches the foundations of C++ programming as applicable to financial engineering.

- FE523 GPU Computing in Finance
  - Teaches the foundation of C++ programming as applicable to financial engineering.

- QF302 Financial Market Microstructure & Trading Strategies
  - Offers students an understanding of the main micro-structural features of financial markets, and the opportunity to test and practice different trading strategies.

- QF427 & QF428 Student Management Investment Fund (SMIF)
  - The course is intended as an Advanced course for Stevens-Howe QF-Fund BT and possibly other students considering the pursuit of an investment management career. Enrollment is by application only and only top students are in the course.

---

If you wish to discuss support for your project or possible collaboration with the Hanlon Financial Systems Laboratory, please contact Demis@stevens.edu or dhsu@stevens.edu.
Keywords:
• Marketing strategy, New Business, Competitor Analysis
• Data oriented Marketing

Business Background:
• Pennsylvania Market LLC is a Food Hall (18400 square ft). It locates at the Pennsylvania Building, strip district in Pittsburgh, PA which is 1 mile away from downtown area.
• PAMarket includes restaurants, shops, winery, bar, and social area for classes, workshops, and meetups. Grand opening time: mid-June 2018.
• PAMarket’s mission: make it a destination for everyone.

Business Question:
Business owners asked for a Marketing strategy

Project Approaches:

1. PA Market LLC
   - Demographic studies
   - Suggested Segmentation

2. Case Studies: Marketing Strategic Plan
   - Big Brands: Eataly
   - Digital Marketing
   - Marketing Aspects

2. Case Studies – Eataly

Business Model
1. Evolving consumer Preferences- customization, customer service, etc
2. High-end culinary market, food and experience
3. Commitment to traditional Italian cuisine, transfer of knowledge

Digital Marketing Channel plays important role.
Facebook, twitter, Instagram (total ~30k followers)
Good quality website

Marketing Strategic Plan: Analytics Aspects

1. Website
   Track customer journey Learn who are customers Customer engagement
   PAMarket need a website, suggest using Google Analytics, SEO, event calendar, FAQ, online support, subscription to newsletter, etc.

2. Data Storage
   • Daily transaction data: better promotions, forecasting, effective inventory and budgeting
   • Customer data: address, email, delivering data for customer engagement to decrease churn rate
   • Advertising data: A/B testing, coupons find effective advertisement
   • Survey data
   • Government data: annually local information

Marketing Strategic Plan: Marketing Aspects

1. Affordable food
2. Trendy food/events
3. Influence by social media

Greater downtown people

College & University students

Pittsburgh residents

3. Case Studies: Nearby Competitors

Smallman Galley
- 6000 square ft
- 0.1 miles away
- 4 restaurants
- 1 bar, 1 coffee shop
- Food hall
- Restaurant Incubator
- Food hall
- Restaurant Incubator
- Healthy food oriented
cashless system
- support local vendors/ producers
- Happy hours
- In store events
- Happy hours
- Online ordering
- Digital Marketing
- Event calendar
- Fine design website
- Facebook, Twitter, Instagram, Yelp

Oxford Market
- 12000 square ft
- 1.5 mile away
- 5 restaurants
- 1 bar
- Healthy food oriented
cashless system
- support local vendors/ producers
- Happy hours
- Online ordering
- Digital Marketing
- Event calendar
- Fine design website
- Facebook, Twitter, Instagram, Yelp

References

Other strategies:
• Online food ordering system/Group order with discount option
• Cross-store promotion (BYOF)
• Cooking classes/schools/Social nights

Tech Companies within 20min (5miles)

Solve parking problem, benefit for the city as well!
• Collaborate with nearbytech companies, PAMarket can advertise on the cars, trucks, paddle carts

Let companies/colleges organize Workshops & Meetups
• Many activate meetups (~2 events / week)
• Bring people in for social, dating, music, workshops, meetup events

http://www.stevens.edu/bia
Wealth Management Branch Prediction

Authors: Shuting Zhang, Harsh Kava
Instructors: Prof. David Belanger, Prof. Edward Stohr, Prof. Khasha Dehnad

Keywords:
• Python, Tableau
• Supervised Learning
• Hybrid Data Science Modeling

Business Questions:
Identify 3 new locations in US as UBS' wealth management branches

Objectives:
1. Who & where are target WM customers
2. How to use machine learning to predict new branches?

Data & Machine Learning Challenges:
1. No Ready-to-use dataset
2. Data under different level: zip/city/county/state level
3. Missing data
4. No pre-labelled data for machine learning models

Data Sources & Feature Engineering:

Demographics:
Population
Gender
Ethnicity
Median Age ...

City Amenities
Luxury Store Sales
Private schools
Hospitals
Airport ...

City Amenities:

U&S & Competitors:
Competitor locations
Competitor Sales
UBS Locations
UBS sales ...

Commercial & Housing:
IRS data
Housing Units
Median Home Value
Startups ...

Machine Learning Approaches:
K-means clustering
1. Median household income
2. Median house value
3. WM branch sales volume
4. Population
5. Number of competitors

Cosine Similarity:
Choose zip codes similar to UBS existing branches
1. Include all features under zip code level
2. Select zip codes with high similarity scores

Feature Selection:
Step 1: Eliminate highly correlated features $152 \to 106$
Step 2: Machine learning algorithm pick the important features. Random Forest, $106 \to 24$ columns

Modeling:
Top 20% of existing zip code with high sales volume: most qualified locations (label=1)

渊 6 Different Machine Learning Algorithms
渊 Cross Validations (5 folds) + Grid Search

Machine Learning Models:
<table>
<thead>
<tr>
<th>Model</th>
<th>Test Accuracy Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.96</td>
</tr>
<tr>
<td>SVM</td>
<td>0.96</td>
</tr>
<tr>
<td>KNN</td>
<td>0.90</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.98</td>
</tr>
<tr>
<td>XG Boost</td>
<td>0.98</td>
</tr>
<tr>
<td>Stacking</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Result Validation:
Averaging score vs ANN scores

Results & Conclusion:
1. ANN ranking have similar results as average scoring method, which indicates 6 models were optimized and worked well
2. Distance of recommended cities were calculated in Python
3. To determine final locations, we recommend UBS starts marketing research on these candidate cities then determine the final locations
4. This program is automated in Python and data exploration is done in Tableau as seen in the demo.

http://www.stevens.edu/bia
Employee Branding Research
Glassdoor.com Company reviews analysis
Authors: Shuting Zhang, Siyan Zhang
Instructor: Rong (Emily) Liu

Keywords & Programs:
- Python, MySql, Tableau, Excel
- Text Mining, Natural Language Processing
- K-means, Non-negative matrix factorization, Topic Modeling, Doc2Vec

Business Questions:

Companies:
1. Employee branding: What are your employees saying about the company?
2. Is anyone hurting the company's reputation in a bad way?
3. How can we solve potential problems and attract talented people?

Job Seekers:
1. How many previous employees left good/bad reviews?
2. Will I fit into the company culture? What problems might I face into?

Sample detected Topics:

Insights:

Keywords & Programs: • Python, MySql, Tableau, Excel • Text Mining, Natural Language Processing • K-means, Non-negative matrix factorization, Topic Modeling, Doc2Vec

Business Questions:

Companies:
1. Employee branding: What are your employees saying about the company?
2. Is anyone hurting the company's reputation in a bad way?
3. How can we solve potential problems and attract talented people?

Job Seekers:
1. How many previous employees left good/bad reviews?
2. Will I fit into the company culture? What problems might I face into?

Conclusion:
1. NMF is the most effective algorithm, Doc2vec is the worst due to the short length of the reviews.
2. NMF provided the most identified topics.
3. Lemmatization: removing stop words might change the review meaning.
4. Small companies are very different from the big companies.

Future improvement:
1. Getting more labeled reviews and improve the algorithms' performance.
2. Try different data processing methods, such as: no split on reviews, stemming, etc.
3. Optimize the algorithm.
4. Gather companies from different industries.
5. Integrated the company info data, predict what are the most factors to determine the best 20 employers of the year on Glassdoor.com.

Q: What about hidden information?

Project Approaches: Data process

Business Questions

Reviews Analysis

Pros

Cons

Data Exploration

Topic detection by NLP

Goal: Avoid multi-topic labeling

Example:

<table>
<thead>
<tr>
<th>No.</th>
<th>Original</th>
<th>Lemmatized-pro</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&quot;After working on and proving some foundational skills&quot;</td>
<td>work prove foundational skill</td>
</tr>
<tr>
<td>2</td>
<td>I was quickly brought on board major</td>
<td>quickly bring board major</td>
</tr>
<tr>
<td>3</td>
<td>Important projects</td>
<td>important project</td>
</tr>
<tr>
<td>4</td>
<td>Every bit of work I was assigned was meaningful and important</td>
<td>every bit work assign meaningful important</td>
</tr>
<tr>
<td>5</td>
<td>Good snacks</td>
<td>good snack</td>
</tr>
</tbody>
</table>

Project Approaches: NLP Algorithms selections

K-means clustering: cosine similarity
NMF: tf-idf
Latent Dirichlet Allocation (LDA) (sklearn): raw term count (probabilistic graphical model)
Doc2vec + K-means clustering: Paragraph vectors + clustering

http://www.stevens.edu/bia
Introduction
We have selected data for all athletes in the history of the Olympic Games. These data contain:
1. Data for all athletes.
2. Age, height, weight, championship status, etc., of all athletes.
3. The order of athletes is ranked by last name.

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Gender</th>
<th>Age</th>
<th>Height (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Peiyi Jiang</td>
<td>M</td>
<td>24</td>
<td>1.90</td>
</tr>
<tr>
<td>2</td>
<td>Katsusaburo</td>
<td>N</td>
<td>23</td>
<td>1.70</td>
</tr>
<tr>
<td>3</td>
<td>Kumar Jeevan</td>
<td>M</td>
<td>24</td>
<td>1.75</td>
</tr>
<tr>
<td>4</td>
<td>Breda Ross</td>
<td>M</td>
<td>24</td>
<td>1.84</td>
</tr>
<tr>
<td>5</td>
<td>Dzmitry Jauke</td>
<td>F</td>
<td>23</td>
<td>1.80</td>
</tr>
</tbody>
</table>

Age Analysis
The two figures contain the athlete's age and gender data. By analyzing these data we can find:
1. What is the age of the athletes? What age of athletes is more suitable for the Olympics?
2. What is the best age for male athletes? What is the best age for female athletes?

Gold Medals In Each Country
From the figure we can see which countries are stronger, which countries are better at the Summer Olympics, and which countries are better at the Winter Olympics.
The figure above show the gold medals for the Summer Olympics. The figure below shows the gold medal for the Winter Olympics.

Number Of Sports
The figure depicts the changes in sports as the year progresses. One of the sharp declines is due to war, and other years are steadily increasing.
Predicting the outcome of a shot
Team: Haitao Liu, Yang Liu, Jiawei Xue
Instructor: Amir H Gandomi

Objectives
• Use the data on shots taken during an NBA season to find the most important features that affect the shooting result
• Build machine learning models to get the relationships between different features and the shooting result
• Interpret the models to get insight into players’ shooting performance

Data Understanding
Using boxplot to find relationships between variables.

Further check the relationships between interested variables.

Data processing
Using box plot to detect outliers of the variables and replacing the outliers with the average value of that feature.

Using heatmap to build a correlation matrix between different features of the dataset.

Principal Component Analysis
In order to reduce the number of variables without losing much of the information, we conduct principal component analysis with nine categories to achieve this goal.

Imputing Missing Values
We found the 5567-missing value in column SHOT_CLOCK

The reason these missing values exist is that the blanks mean zero offense time. More specifically, when a player catches the ball, there’s no time left for shooting. So we replaced these blank value with zero.

Modeling
• Split the dataset into the training and testing subsets in a ratio of 4:1
• Create the prediction models on the training subset using Naïve Bayes, Linear Discriminant Analysis, Logistic Regression, XGBoost, and Ensemble.
• Validate the prediction results using the testing subset.

Accuracy of each models

Result & Future Work
• The XGBoost perform the highest accuracy.
• These models are not ideal enough and have a great potential to enhance their accuracy.
• In the future, we will add more features to increase the accuracy of the model so that we can make a more accurate prediction to give pieces of advice to teams for training, coaching, and making playing strategies.

http://www.stevens.edu/bia
Motivation
The churn rate analysis is crucial for telephone service companies because the cost of retaining an existing customer is far less than acquiring a new one. Companies from these sectors often have customer service branches which attempt to win back defecting clients, because recovered long-term customers can be worth much more to a company than newly recruited clients.

Technologies
• Using Python to identify the relationships among different features of the dataset
• Using Tableau to generalize plots for visualization
• Using Solver to build optimization model and find the optimum solutions

Data Understanding
• Correlation Matrix

![Correlation Matrix Diagram]

• Further understanding

Optimization Model
Marketing Optimization Model

<table>
<thead>
<tr>
<th>Tenure</th>
<th>Estimated Churn No.</th>
<th>Retain-Confidence</th>
<th>Gross Month Avg-Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.12)</td>
<td>1319</td>
<td>30%</td>
<td>$69.75</td>
</tr>
<tr>
<td>(12.24)</td>
<td>308</td>
<td>40%</td>
<td>$83.66</td>
</tr>
<tr>
<td>(24.36)</td>
<td>227</td>
<td>50%</td>
<td>$90.70</td>
</tr>
<tr>
<td>(36.48)</td>
<td>123</td>
<td>50%</td>
<td>$98.46</td>
</tr>
<tr>
<td>(48.60)</td>
<td>81</td>
<td>50%</td>
<td>$101.20</td>
</tr>
<tr>
<td>(60.72)</td>
<td>44</td>
<td>50%</td>
<td>$103.44</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tenure</th>
<th>Discount for 1st Year</th>
<th>Discount for 2nd Year</th>
<th>Discount for 3rd Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.12)</td>
<td>10%</td>
<td>10%</td>
<td>25%</td>
</tr>
<tr>
<td>(12.24)</td>
<td>15%</td>
<td>15%</td>
<td>35%</td>
</tr>
<tr>
<td>(24.36)</td>
<td>17%</td>
<td>17%</td>
<td>40%</td>
</tr>
<tr>
<td>(36.48)</td>
<td>19%</td>
<td>19%</td>
<td>45%</td>
</tr>
<tr>
<td>(48.60)</td>
<td>24%</td>
<td>24%</td>
<td>55%</td>
</tr>
<tr>
<td>(60.72)</td>
<td>28%</td>
<td>28%</td>
<td>65%</td>
</tr>
</tbody>
</table>

Discount Details:
For 1st Year: X1
Increased by:
For 2nd Year: X2=X1*(1+a)
For 3rd Year: X3=X2*(1+b)

Estimated Cost
Price of Internet Service

<table>
<thead>
<tr>
<th>Months For Discount Year</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>39,570</td>
<td>39,570</td>
<td>92,419</td>
</tr>
<tr>
<td>12</td>
<td>17,894</td>
<td>17,894</td>
<td>41,790</td>
</tr>
<tr>
<td>12</td>
<td>11,863</td>
<td>11,863</td>
<td>27,280</td>
</tr>
<tr>
<td>12</td>
<td>7,110</td>
<td>7,110</td>
<td>16,405</td>
</tr>
<tr>
<td>12</td>
<td>5,722</td>
<td>5,722</td>
<td>13,285</td>
</tr>
<tr>
<td>12</td>
<td>3,674</td>
<td>3,674</td>
<td>8,580</td>
</tr>
</tbody>
</table>

Estimated Financial Effect
Total Marketing Period/Months Expected Total Income: $53,643
Total Expected Total Cost: $29,165

Classification Model (Churn)

Split data
Training set: Testing set = 4:1
Fit model
Validation

Predicting Results

Receiver Operating Characteristic (ROC)
Motivation -- Why we choose this topic:

During 1997 and 2017, The economic damages of natural disasters are getting bigger and bigger; The United States ranked the first place of economic losses all over the world.

Top 10 states selected:

<table>
<thead>
<tr>
<th>State</th>
<th>GDP</th>
<th>Pop</th>
<th>Feet</th>
<th>Bud</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texas</td>
<td>2,056,072.4</td>
<td>3,056,072.4</td>
<td>1,700.0</td>
<td>0</td>
<td>14,812.24</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>2,759.01</td>
<td>223,759.01</td>
<td>1,300.0</td>
<td>0</td>
<td>500.00</td>
</tr>
<tr>
<td>New York</td>
<td>728,350.81</td>
<td>1,792,387.6</td>
<td>850.0</td>
<td>0</td>
<td>7,415.40</td>
</tr>
<tr>
<td>Missouri</td>
<td>342,407.18</td>
<td>1,792,387.6</td>
<td>1,000.0</td>
<td>0</td>
<td>16,389.61</td>
</tr>
<tr>
<td>Mississippi</td>
<td>126,029.20</td>
<td>126,029.20</td>
<td>800.0</td>
<td>0</td>
<td>2,670.47</td>
</tr>
<tr>
<td>Louisiana</td>
<td>284,111.46</td>
<td>284,111.46</td>
<td>300.0</td>
<td>0</td>
<td>500.00</td>
</tr>
<tr>
<td>Florida</td>
<td>1,135,263.3</td>
<td>1,135,263.3</td>
<td>100.0</td>
<td>0</td>
<td>29,748.68</td>
</tr>
<tr>
<td>California</td>
<td>3,268,754.1</td>
<td>3,268,754.1</td>
<td>2,900.0</td>
<td>0</td>
<td>17,380.22</td>
</tr>
<tr>
<td>Alabama</td>
<td>240,552.98</td>
<td>240,552.98</td>
<td>500.0</td>
<td>0</td>
<td>2,090.19</td>
</tr>
</tbody>
</table>

Min subject: 47.69

P = population
G = GDP of State
E = Elevation of State
b = Budget

Constraint:
\[ \sum_{i=1}^{n} b_i \leq 10000M \]
\[ b > 500 M \]

\[ F_i(D) = \frac{P * G}{E*(1+b_i)^2} \]

Technology:
- We use R for generating the result NLP files of the mathematical models and the visualization through GGPlOT.
- We use NLP Solver for solving the optimization models.

Current & Future Work:
- Development formulas to build the model to analyze the past data and choosing the best model to predict the future.
- Seeking more factors and trying to find frequency which would multiply the formula to make the model better.
- Development a visualization work of the budget allocated on the target states.

Estimation of the future:
- The expected trend of natural disasters is showing an upward trend.
- Economic damage will rise as a share of gross domestic product (GDP), which provides a measure of the nation’s ability to pay for that damage.

http://www.stevens.edu/bia
Revenue and Cost Optimization for a Clothing Supply Chain
Authors: Weifeng Li, Liran Zhang, Mingxin Zheng, Zeyu Shao
Instructor: Alkiviadis Vazacopoulus

Problem Statement
- We need to develop a supply chain strategy to maximize profit of the clothing manufacturing facilities.
- Our objective is to increase the revenue, reduce the transportation cost, labor and raw material costs.

Constraints
1. All the materials that needs transportation cannot be less than the quantities in our plan.
2. The suppliers have fixed locations. Selecting the right suppliers is part of the optimization methodology.
3. The price of transportation cost is constant and depends on the distance.
4. Different type of employees can only complete one job in one period.
5. We introduce minimum and maximum order quantities.
6. We introduce labor costs in our optimization model.
7. We introduce raw material costs that depend on the location of the supplier.

Methodology
- We modeled our problem using mixed integer programming.
- We had incorporated binary variables for selecting the right suppliers.
- We ran different scenarios to find a robust optimized solution.
- We used Excel Solver.

Observation

Conclusions
- We managed to solve a reasonable size problem using the Excel Solver.
- We can extend our model and solve larger instances using a commercial solver.
- We have the ability to use scenario generation using different demand patterns.
Introduction

New York City, the world's economic, commercial, financial, media, political, educational, and entertainment center, and the world's largest city. Therefore, NYC is naturally very attractive to tourists from all over the world. However, many tourists may not be able to plan the trips to NYC due to their limited time and budget, thus missing out on various spots. Our project optimizes the travel experience by judging the factors that influence travel, such as traffic, cost, popularity, and number of attractions. We help visitors planning their trips and helping them get the best travel experience.

Based on this existing situation, we introduced the Analytic Hierarchy Process Method by selecting the priority factors from the travel, trying to complete a good trip route which includes most spots with a limited budget. We also use Travelling Salesman Method to design a route.

Experiment

Model development

Analytic Hierarchy Process ------ A structured technique for organizing and analyzing complex decisions, based on mathematics and psychology. AHP can be used to make decisions in situations where multiple objectives are present.

To initialize, we choose 7 famous scenic spots and then use the AHP decision tool to make a priority ordering with the criteria of Spots Quality, Popularity, Cost and Transportation.

<table>
<thead>
<tr>
<th>Spots</th>
<th>Quality</th>
<th>Popularity</th>
<th>Cost</th>
<th>Transportation</th>
<th>Overall score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time square</td>
<td>0.349</td>
<td>0.271</td>
<td>0.205</td>
<td>0.337</td>
<td>0.3</td>
</tr>
<tr>
<td>Statue of Liberty</td>
<td>0.115</td>
<td>0.019</td>
<td>0.184</td>
<td>0.046</td>
<td>0.141</td>
</tr>
<tr>
<td>Wall street</td>
<td>0.067</td>
<td>0.288</td>
<td>0.142</td>
<td>0.089</td>
<td>0.141</td>
</tr>
<tr>
<td>Met museum</td>
<td>0.056</td>
<td>0.082</td>
<td>0.168</td>
<td>0.280</td>
<td>0.181</td>
</tr>
<tr>
<td>Empire building</td>
<td>0.028</td>
<td>0.044</td>
<td>0.305</td>
<td>0.081</td>
<td>0.071</td>
</tr>
<tr>
<td>WTC</td>
<td>0.052</td>
<td>0.017</td>
<td>0.144</td>
<td>0.13</td>
<td>0.071</td>
</tr>
<tr>
<td>Central park</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sum</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

From the AHP analysis, there is a priority among 7 spots. If you only want to visit top 5 spots in New York, we recommend you to choose Time square, Statue of Liberty, Wall street, Met museum and Empire building.

Results

<Travel route>

met museum of art ------ empire state building ------ time square ------ wall street ------ statue of liberty

Future work

Concerning the next step, we aim to make a new model by adding extra measures to customize the design, regarding the price of every level and the satisfaction weight, and to make the model satisfy the utility when given a few constraints in cost.
Introduction

• **Real World Problem:** Airbnb has come up with an affordable alternative to hoteling accommodations and has become a successful competitor. However, the prices of Airbnb homes are not typically consistent due to surges either through tourism or events. This makes it difficult for its consumers to accurately know the price for their stay, which could affect the outcome of their trip.

• **Client:** Airbnb and its consumers

• **Research Question:** Identify factors that affect pricing of Airbnbs in Washington D.C. and create an optimization equation to best predict Airbnb pricing

Data and Scope

• **Technology:** R programming and MS Excel


• **Data Size:** 790 rows of Airbnb homes

• **Data Cleaning:**
  - Omitted NAN rows
  - Omitted non-numerical columns that we could not model, such as host verifications
  - Omitted outliers in prices

Model Approach

\[ y \sim x1 + x2 + x3 + x4 \ldots + x23 \]

• Applied with linear regression, logistical regression, stepwise regression, and ANOVA to find factors most affecting price

Results

|                  | Intercept | Estimate | Standard Error | T value | Pr(>|t|) |
|------------------|-----------|----------|----------------|---------|----------|
| Bathrooms        | 1.192e-03 | 4.238e+00| 2.813          | 0.00504 |
| Accommodations   | -5.057e+00| 1.637e+00| -3.089         | 0.00208 |
| Security         | -3.814e-01| 1.838e-01| -2.074         | 0.03837 |
| Deposit          | -9.242e+00| 3.614e+00| -2.557         | 0.01074 |
| Bedrooms         | 8.566e-01 | 4.153e+00| 8.610          | 0.00132 |

After running the regressions, these 5 factors stood out due to their low p-values. These are the most important factors affecting price. Below feature residual charts that show correlation.

Conclusion and Real World Impact

• The most important factors that affect pricing are accommodations, location, security deposit, 30 day availability, and number of bathrooms.

• This research is for Airbnb’s consumers to become aware of what factors affect pricing the most. One way to minimize cost on Airbnb is by booking more than 30 days in advance.
Function Approximation Using Evolutionary Polynomials

Author: Aleksandr Grin
Advisor: Professor Amir H. Gandomi

INTRODUCTION

Fitting curves to data is an important statistical analysis tool. A cottage industry of various methodologies has evolved which aims to fit trends into data. This is an important tool because it allows us to generalize processes and predict outcomes for data which has not previously been tested or obtained.

There are various methods which work to optimize this task, ranging from polynomial regression to neural network function approximation. Genetic algorithms also hold a place amongst the slew of toolsets available for curve fitting. They are often used to find polynomial constants in regressions, thereby optimizing the process.

In this project we examined a new approach to polynomial curve fitting by attempting to reduce the number of constants in the problem to two. We also combined this novel function representation with genetic programming and regression. The resulting structure has been termed a “polynomial network”. With this novel structure we hope to reduce the computational complexity of curve fitting and thereby optimize the process compared to other methods.

ALGORITHM

RESULTS

CONCLUSION

Having constructed the algorithm as described we have shown that the algorithm can in fact perform curve fitting. We used target functions to simulate data to fit, but the result is promising especially with the chaotic weiserstrass function. The algorithm still has certain parts that require optimization and further fixes, but on the whole we have achieved our goal of demonstrating viability for this approach. Thus, further work can be concentrated on optimizing this approach and testing against other algorithms.

Acknowledgments:

I would like to extend my sincere gratitude to professor Gandomi for allowing me to pursue a research area in a very interested area providing all the knowledge and experience he had in the field. I would also like to thank Steven’s for providing the opportunity to experience the research field firsthand.

Author: Aleksandr Grin
Advisor: Professor Amir H. Gandomi

INTRODUCTION

Fitting curves to data is an important statistical analysis tool. A cottage industry of various methodologies has evolved which aims to fit trends into data. This is an important tool because it allows us to generalize processes and predict outcomes for data which has not previously been tested or obtained.

There are various methods which work to optimize this task, ranging from polynomial regression to neural network function approximation. Genetic algorithms also hold a place amongst the slew of toolsets available for curve fitting. They are often used to find polynomial constants in regressions, thereby optimizing the process.

In this project we examined a new approach to polynomial curve fitting by attempting to reduce the number of constants in the problem to two. We also combined this novel function representation with genetic programming and regression. The resulting structure has been termed a “polynomial network”. With this novel structure we hope to reduce the computational complexity of curve fitting and thereby optimize the process compared to other methods.

ALGORITHM

RESULTS

CONCLUSION

Having constructed the algorithm as described we have shown that the algorithm can in fact perform curve fitting. We used target functions to simulate data to fit, but the result is promising especially with the chaotic weiserstrass function. The algorithm still has certain parts that require optimization and further fixes, but on the whole we have achieved our goal of demonstrating viability for this approach. Thus, further work can be concentrated on optimizing this approach and testing against other algorithms.

Acknowledgments:

I would like to extend my sincere gratitude to professor Gandomi for allowing me to pursue a research area I was very interested in while providing all the knowledge and experience he had in the field. I would also like to thank Steven’s for providing the opportunity to experience the research field firsthand.

http://www.stevens.edu/bia
Correlating Long-Term Innovation with Success in Career Progression

Adam Coscia
Instructors: Aron Lindberg, Ph.D., Amir Gandomi, Ph.D.

Motivation

- **Successful** individuals and businesses in all fields explore new innovations and/or exploit successful ones.
- **Intervals** of strategy exploration and exploitation may affect **long-term success**, independent of career type and field.
- **Developing** long-term innovation models to maximize career success is the goal!

Intermediate Results

- Sample size: 2373 **EVE Online players** tracked over a 37-month period from May 2015 through May 2018.
- **Investment trajectories** overlaid with moving average of success to visually determine relationship
- Pearson Correlation employed to measure **association** within player timeline.

**Comparing Trajectory and Success of Sample Player**

![Graph showing trajectory and success comparison](image)

**Key Finding:** Time evolution of strategy is often unpredictable on an individual basis!

- Preliminary distributions of correlations between performance and investment across players shows little evidence to claim dependency.

Data Source

- **EVE Online**

Analysis

- **Scrape** 37 months of player data from [zkillboard.com](http://www.zkillboard.com) using **Scrapy** for Python in 4 days.
- **Unpack** and **Clean** API data using **pandas** for Python.
- **Conduct** data analysis using **pandas**:
  i. Develop **Investment** and **Performance Series** from observations of player strategy upon death (killmails, see above)
  ii. **Measure association** between performance and investment to predict groupings
  iii. **Cluster** series by **rolling average** investment
  iv. **Compare** success of each cluster using **weighted average** of kill/death

Future Research

- **Compare** moving averages of success across groups of similar trajectories
- **Visualize** clustering trajectories
- **Assess** strategy bias, implications for careers outside of gaming.
- **Further** considerations:
  i. More ship types would increase player sample size and reduce strategy bias.
  ii. Performance observations could be weighted based on contribution to kill.

http://www.stevens.edu/bia
Introduction:
In order for car manufacturers to increase sales and boost profits, analysis is performed to find which factors are important for car sales and predict sales price. In addition, we cluster the cars using a machine learning clustering model recommendation system and maximizing propensity to buy, to help salesmen develop strategies.

Experiment:

Data processing:
The dataset has 14 columns and represent different characteristics. In addition we did some pre-processing:
1. replaced Nan and change datatype of the variables that belong to many categories
2. encoded category variables
3. split datetime into year, month, and day
4. Performed EDA to find the factors’ correlation and found that Ford had the highest sales and price. Manufacturer Jaguar got the highest value loss after 4 years.

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Sales in thousands</th>
<th>Engine Size</th>
<th>Horsepower</th>
<th>Weight</th>
<th>Fuel efficiency</th>
<th>Year</th>
<th>Month</th>
<th>Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10,500</td>
<td>1.8</td>
<td>4.0</td>
<td>4,500</td>
<td>22.0</td>
<td>2015</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>12,000</td>
<td>2.0</td>
<td>4.5</td>
<td>5,000</td>
<td>25.0</td>
<td>2016</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>11,500</td>
<td>1.9</td>
<td>4.2</td>
<td>4,800</td>
<td>22.5</td>
<td>2017</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>10,750</td>
<td>1.8</td>
<td>4.0</td>
<td>4,500</td>
<td>23.0</td>
<td>2018</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

Conclusion:
• R2 of the linear regression model is 0.89. This is because there are several variables that can explain the dependent variable “Price in thousands”. For example, the “4-year resale value” has 0.95 correlation with the dependent variable, “Engine size” has 0.62, and “Horsepower” has 0.83. In other words, those three factors are the top three important variables for determining the price.
• Clustering may differ when we use different distance functions. The best number of clusters for k-mean method is 3, while complete or maximum linkage methods require more clusters, because maximum linkage method is not efficient at dealing with the “outliers”.

Practical implications:
Part 1: We use the linear regression model to predict two cars’ price (in thousands), which one is 24.92 and another is 16.05. We combine the analysis of local market price preferences.

Part 2: We use the clustering results to build a car recommendation system. We can show customers similar cars and optimize our sales price and propensity to buy a car. Based on the price and the features we can maximize our revenues and profits.
NBA Data Visualization Analysis
Xin Chen, Xiaohai Su, Xiang Yang
Instructor: Alkiviadis Vazacopoulos

Introduction
We managed to scrape data related to NBA teams and players from Internet. The dataset contains:
- 41 columns and 1057 rows
- field performance information of over 300 NBA players of 30 NBA teams in the last 3 years
- Salary and Geographic information of players and teams

Geographic Overlook
Find out which division the NBA teams are from:
Altogether 6 divisions, which are Division Atlantic, Central, Northwest, Pacific, Southeast and Southwest.

Find out which cities the teams are from:
There are in total 27 cities of America and Canada, from which the teams come.

Team Analysis
By analyzing, we want to find out
- What is the success rate of scoring of every team?
- What is the composition of every team based on the age and experience?
- Is this team good or bad? What is the weakness of this team?

Salary Analysis
We create a heat map and word cloud for this analysis on team level and player level, respectively.

Geographic Overlook
Find out which division the NBA teams are from:
Altogether 6 divisions, which are Division Atlantic, Central, Northwest, Pacific, Southeast and Southwest.

Find out which cities the teams are from:
There are in total 27 cities of America and Canada, from which the teams come.

Team Analysis
By analyzing, we want to find out
- What is the success rate of scoring of every team?
- What is the composition of every team based on the age and experience?
- Is this team good or bad? What is the weakness of this team?

Salary Analysis
We create a heat map and word cloud for this analysis on team level and player level, respectively.

Player Analysis
As long as we select a single player in the “Team Composition”, the other visualizations will show this player’s performance in the last 3 years.
INTRODUCTION

OBJECTIVE
It is a large dataset about the New York City Taxi Fare (Approx. 55 Million Flows) containing information on pick up / drop off points, time, date, fare, etc. The idea is to explore the visualizations on a large-scale data with Tableau and further write a prediction algorithm for the testing dataset.

BACKGROUND

- New York City is US largest Metropolitan area with population of 20.1 million.
- In the last year, there were 302,599 licenses issued to the drivers, where as the budget amounted to $46,890,009.
- Such a large market tells a lot about the taxi requirements, patterns, and fairs which are not only useful for the customers but also potential market entrants.

PREDICTION MODELLING

- Calculated Fare build up using NY Taxi fare, to account for additional fare due to surcharge and discount rates available.
- Divided by datetime to categorical timeframe to account the change in values to date time observation and convert time-series data to a normal dataset to apply Machine Learning Models.
- LightGBM (Introduction) - LightGBM is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with the following advantages:
  i. Faster training speed and higher efficiency.
  ii. Lower memory usage.
  iii. Better accuracy.
  v. Capable of handling large-scale data.
- Leaf-wise may cause over-fitting when data is small, so LightGBM includes the max_depth parameter to limit tree depth. However, trees still grow leaf-wise even when max_depth is specified.

CONCLUSION

- Base Fare increased in the mid 2012, where the average fare per ride increase exponentially, because of Uber and Lyft being the new market entrants.
- Yellow Taxis on an average charged approximately 5$ per mile distance covered in New York.
- The lowest ride fares are observed during early morning time frame approximately between 3 a.m. to 5 a.m.
- Future scope would be to bring in data from Uber and Lyft to analyze the rise of taxi ride sharing in New York City and the affect on the prices.

http://www.stevens.edu/bia
INTRODUCTION

OBJECTIVE
Analyzing Google Merchandize Store customer dataset to predict revenue per customer. As the dataset is very large, 31GB (Training & Testing), which includes traffic source, session, device, geoNetwork, page views, transaction revenue. Being a huge dataset, only 2 million rows were used in the project.

BACKGROUND
• Since, only a small percentage of customers produce most of the revenue.
• As such, marketing teams are challenged to make appropriate investments in promotional strategies.
• Google Products are loved by the most people, but the buying is only done by few people who visit the site.

PREDICTIONS
• Combined score of XGBoost, Catboost and LGBM was used to predict the score.
  • LGBM Model:
    LightGBM is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed.
    Negative gradient of the loss function:
    \[ \frac{\partial L(y, f(x))}{\partial f(x)}f(x) = f^{(m-1)}(x) \]
  • XGBoost Model:
    XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.
    Different from GBM, XGBoost tries to determine the step directly by solving.
    \[ \frac{\partial L(y, f^{(m-1)}(x) + f_m(x))}{\partial f_m(x)} = 0 \]
  • CatBoost Model:
    CatBoost is a machine learning algorithm that uses gradient boosting on decision trees.

REFERENCES
• https://www.kaggle.com/hakkisimsek/plotly-tutorial-4, Plotly Tutorial, Kaggle
• https://www.kaggle.com/karkun/sergey-ivanov-msu-mmp, Kaggle Kernel

DATA VISUALIZATIONS

CONCLUSION
• Transaction revenue is predicted with an RMSE of 1.607865
• Being very large dataset, only a portion was used in training the model due to limited computational power at disposal.
• This project is still under progress for testing on a better system to further reduce RMSE.
• Future scope would be to employ decision statistics to pick the elements from dataset that would help in the prediction model.
Clustering Large Cap Stocks During Different Phases of the Economic Cycle

Students: Nikhil Lohiya, Raj Mehta
Instructor: Amir H. Gandomi

Introduction

**OBJECTIVE**

We tried to provide a set of securities that behave similarly during a particular phase of the economic cycle. For this project, the creation of sub-asset classes is done only for large-cap stocks.

**BACKGROUND**

Over time, developed economies such as the US are becoming more volatile and hence the underlying risk of securities has risen. This project aims to identify the risks & potential returns associated with different securities and to cluster similar stock similarities according to their Sharpe ratio, volatility, and an average return of stocks for a better analysis of the portfolio.

Flow - Project

- **Data Acquire**
  - Data of Large Cap Stocks & US Treasury Bonds is gathered directly using APIs.
  - The data potentially consists of 2 time frames i.e. Recessionary & Expansionary Economy.

- **Data Preprocessing**
  - This segment included the application of formulae to calculate the pre-required parameters. (Eq 1,2,3,4).

- **Analysis**
  - This segment consists of K-means clustering Analysis done on the Large Cap Stocks. (K=22) (500 Stocks)
  - The clustered securities are then further tested for their correlation among the sub asset classes.

- **Results**
  - The results of K-means clustering varies in the range (9 to 45)
  - There were some outliers in our analysis as well.

Mathematical Modelling

- Here we take daily returns for all the 500 securities.
  \[ R = \frac{p_t - p_{t-1}}{p_{t-1}} \times 100 \]  
  Eq1
- Average Return and Volatility.
  \[ \mu_I = \frac{\sum_{t=1}^{n} R_t}{n} \]  
  Eq2
  \[ \sigma_I = \sqrt{\frac{\sum_{t=1}^{n} (R_t - \mu_I)^2}{n-1}} \]  
  Eq3
- Sharpe Ratio calculation for the Securities.
  \[ SR_I = \frac{(\bar{R} - \bar{R})}{\sigma} \]  
  Eq4
- Correlation Matrix between the clustered securities following the cluster formation.

Results

- **Clustering of Stocks during Recovery phase**
  - K-means plot shows that the stocks are clustered with similarities by their Sharpe ratio, volatility, and average return. There are 9 graphs in total, and 2 of them are displayed above for the expansion and recession phase. The x-axis shows the ticker/symbol of sp500 and the Y-axis shows the Cluster number. If we hover on the dot in the graph, it shows the ticker along with its cluster number, and the variable used for clustering. We used Silhouette and visually inspected the data points to find the optimal value of k, which turns out to be 22.

Conclusion & Future Scope

- With the above methodology, we have been able to develop a set of classes which behave in a similar fashion during each phase of the economic cycle.
- The same methodology can be extended to different asset classes available online.
- Application of Neural Networks can significantly reduce the error in cluster formation.
- Also, application of different parameters such as Valuation, Solvency or Growth potential factors can be included for clustering purposes.
- Next, we plan to add leading economic indicator data to identify the economic trend and to perform the relevant analysis.

http://www.stevens.edu/cia
FIN-FINICKY: Financial Analyst’s Toolkit

Author: Nikhil Lohiya

INTRODUCTION

OBJECTIVE

An open source application which is a one stop shop for Stock Market Data analysis, Portfolio Management, Real Estate Investments and Equity Analytics and can be accessed by the users from any device. A C2C application designed for the individuals still using Excel based analysis for basic calculations on the REITs, ARIMA and GARCH Models.

BACKGROUND

In the world of finance and mathematics there are endless set of instruments and models. In an attempt to create the different models online as a wevtool, I came up with this prototype tool taking into account for the basic calculations in the Econometrics Sector, Real Estate Sector, and Equity Industry. Since, the models in the above-mentioned sectors are a world in themselves, I tried to integrate as many as possible for the purpose of this application.

Link: https://nlohiya.shinyapps.io/Fin-Finicky/

FORMULA BOOK

- ARIMA(1,0,0): \( \hat{Y}_t = \mu + \phi_1 Y_{t-1} \)
- GARCH(p,q): \( y_t = a_0 + a_1 y_{t-1} + \cdots + a_p y_{t-p} + \varepsilon_t = a_0 + \sum a_i y_{t-i} + \varepsilon_t \)
- VAR(1,2): \( y_{2,t} = \phi_2,1Y_{1,t-1} + \phi_2,2Y_{2,t-1} + \varepsilon_{2,t} \)
- Sharpe Ratio: \( SR_j = \frac{(R_j - R_f)}{\sigma_j} \)
- Portfolio Return: \( E(R) = p_1 R_1 + p_2 R_2 + \cdots + p_n R_n \)
- Portfolio Variance \( = \sigma_A^2 \sigma_A^2 + \sigma_B^2 \sigma_B^2 + \cdots + \sigma_n^2 \sigma_n^2 \)
- Net Operating Income(NOI) = (1 - Vacancy Loss) * (Gross Rental Income) * (Prop. Size) + Other Income - Operating Expense
- Value = NOI / Capitalization Rate
- Loan to Value Ratio = Loan Amount / Value
- REITs:
  - Net Asset Value = (NOI/CAP) - Debits - Liabilities + Cash Amount / Shares
  - Price to Funds from Operations (FFO) = (Funds from Operations * (Industry Avg. Multiplier of FFO)/Shares)
  - Price to Adjusted Funds from Operations (AFFO) = (FFO - Non Cash Rent - Recurring Maintenance) * (Industry Avg. Multiplier of FFO)/Shares
- Risk Returns:
  - CAPM = Rf + (B x Equity Risk Premium)
  - Fama-French Three Factor Model:
    \( R_i - R_f = \beta (R_m - R_f) + (\alpha_i \times SMB) + (\alpha_i \times HML) \)
- And More…

APPLICATION LAYOUT

Here are a few screenshots from the application GUI:

1. Main Page
2. Portfolio Analytics & Real Estate Investments
3. REITs Calculations

END RESULT & FUTURE SCOPE

- This prototype application is a base for developing a large-scale application useful for financial analysts.
- In future, I plan to include a comprehensive blog explaining the use of the given formulae/models and addition of remaining models in the given sectors.
- The domain of fixed income securities is vast and constantly evolving. I plan to include the models from this sector and a comprehensive toolkit for risk management tools in a sister application.

http://www.stevens.edu/blg
Group Emailing using Robotic Process Automation
Authors: Pallavi Naidu, Abhitej Kodali
Instructor: Prof. Edward Stohr

OBJECTIVE
To automate the group mailing service for the Business Intelligence and Analytics club using Blue Prism - a Robotic Process Automation.

BACKGROUND
Robotic Process Automation is a form of business process automation technology based on software robots or artificial intelligent workers. Blue Prism software enables business operations to be agile and cost effective by automating manual, rule-based, and repetitive back-office processes. The Blue prism tool offers a flowchart-like designer with drag and drop features to automate each step of a business process.
Currently, the BIA club has a tedious process for sending a group mail to the list of students in the club. The member data is stored in an excel sheet. The email IDs from the excel sheet have to be copied each time and pasted to the address tab whenever a mail has to be sent to a group. The process is tiresome and mundane as the records are sorted manually and there is a chance of manual error.

METHODOLOGY
- Ran process flow analysis on Signavio software to verify the effectiveness of bringing a new solution to the current existing system in place
- Created a database of students who are currently enrolled in the BI&A program.
- Created a bot using BluePrism RPA software and VB.net to automate the process of sending group emails.

PROJECT FLOW

CONCLUSION & FUTURE SCOPE
- The efficiency of the process improved tremendously from 15 minutes to 2 minutes.
- Future scope involves processes to be integrated between departments for easier integration of student details.

http://www.stevens.edu/bia
Cognitive Application to Determine Adverse Side Effects of Vaccines

Authors: Pallavi Naidu, Kathy Chowaniec, Krishanu Agrawal
Instructor: Dr. Chris Asakiewicz

OBJECTIVE
To develop a cognitive chatbot application that would enable the public to discover the potential symptoms of a particular vaccine based on their demographics using past reported events from the VAERS Dataset. The vaccine bot would be featured on a medical website to attract potential users, but could be expanded to doctors and more experienced medical professionals.

BACKGROUND
The Centers for Disease Control and Prevention (CDC) and the U.S. Food and Drug Administration (FDA) maintain a database of adverse reactions to vaccines, called the Vaccine Adverse Event Reporting System (VAERS). According to the CDC, over 30,000 VAERS reports are filed each year. By using this data, our chatbot would help users know the symptoms of a reaction to a vaccine and the number of days after which it would manifest. This would help users to be aware of and be prepared for any adverse events in the future.

DISCOVERY ARCHITECTURE
Below is the standard architecture for Watson Discovery:

To extract the symptom data, we used the Discovery Language Query concept by building queries and integrating them with our application.

PROJECT FLOW

VAERS dataset of 2000 records

Cleaned and converted the files to JSON format to upload to Watson Discovery

Queries using Watson Discovery Language and Natural Language Query features

API to give symptoms by connecting Watson Assistant & Discovery applications

Vaccine Side Effects Chatbot which asks user demographics and gives results from API accordingly

RESOURCES
• IBM Watson Assistant & Discovery services
• API using IBM Cloud Functions
• Languages: Python & JavaScript
• Bluemix/IBM Cloud deployment

FUTURE SCOPE
• Can be expanded to include pharmaceutical drug symptoms from FDASide effects database
• Allow doctors and pharmacists to use chat bot in advising and helping diagnose symptoms in patients
• Expand vaccine dataset for more accurate results

http://www.stevens.edu/bia
Predict Potential Customers by Analyzing Bank’s Telemarketing Data
Authors: Shreyas Menon, Pallavi Naidu
Instructor: Prof. David Belanger

OBJECTIVE
To develop a predictive model and analyze customer attributes to help banks enhance their success rates for Telemarketing campaigns.

BACKGROUND
Banks most often use telemarketing campaigns to target potential customers and sell products like term deposits, credit cards, etc. The strategic goal of such campaigns is to enhance business. The process involves direct calls over a fixed line or a cellular network. Agents interact with the customers and persuade them to subscribe.

However, most banks fail to identify the important attributes of the customers who subscribe to their products. Also, if a customer has been called several times, there is risk of losing a prospective subscriber. Such careful selection of attributes to target best set of clients needs extensive analysis of the already available data. An extensive analysis is reported here with final objective to help banks decide on the best possible set of parameters that would lead to a subscription.

RESULTS

Random Forest
Accuracy: 68.92%

![Random Forest Accuracy Graph]

Logistic Regression
Accuracy: 64.43%

![Logistic Regression Accuracy Graph]

PROJECT FLOW
- Converted biased data set to unbiased data set
- Box and Whiskers Histogram
- Categorical variables to dummy variables
- Random Forest Accuracy:
- Logistic Regression Accuracy:

RESOURCES
Languages: Python and R

FUTURE SCOPE
- Cross-selling other banking and financial products to the targeted customers
- The model can be used for passive marketing wherein the customer base is contacted via an email or social media

http://www.stevens.edu/bia

23
Quora - Answer Recommendation Using Deep Learning Models
Authors: Tsen-Hung Wu, Cheng Yu, Shreyas Menon
Instructor: Rong Liu

Background
Quora is a question-and-answer website where askers can post questions or find answers. Around 38 millions of questions have been asked on Quora. To be more specific, we focused on the topic of Bitcoin discussed on Quora because it becoming a popular topic recently.

Problem Statement
To help Quora construct a platform where it attracts more users to find solutions, our goal is to provide the answer recommendation using language understanding models.
• For Quora
  More engagement and active users on the Quora website.
• For askers
  Spend less time on finding the answers given a new question.
The answer recommendation suits the needs of askers.

Data Understanding
• In total, the database of Quora contains more than 48.4k questions under the topic of Bitcoin.
• 303.6k users have subscribed to follow the topic of Bitcoin on Quora.
• 6788 questions and 23389 answers have been scraped using the self-designed crawler in Python, a dynamic way to collect raw string data on websites.

Modeling & Methodology
• Traditional Machine Learning Algorithms: SVM, Random Forest
  SVM F1-Score
  Random Forest F1-Score

• Deep Learning Model: Convolutional Neural Network - network design
  CNN train history

• CNN train history
  Parameter tuning: max number of words in the corpus, document length of questions, and document length of answers, overall CNN network.

Results and Evaluations
• Selected the best similarity score threshold to achieve the optimal model performance on test data (CNN)
  • Model Performance Report on test data: 1992 pairs of questions and answers
    precision recall f1-score support
    0.0  0.81  0.97  0.88   989
    1.0  0.96  0.78  0.86  1003
    avg / total 0.89  0.87  0.87  1992

Conclusions
• We recommend the most reliable answer given a new question to askers. Thus, it is time-efficient since askers no longer need to wait the solutions day by day.
• CNN is a powerful deep learning model. The major advantage of it is to extract the “critical” words or patterns among questions and answers.
• The applications of CNN is broad not only limited to the document searching (recommendation) tasks.

Deployment (Demo)
New question is: How can I invest in Bitcoin
New answer is: Thanks for A2A. Disclaimer: Crypto market is highly volatile. Invest the amount that you can afford to lose in a single moment. There are multiple platforms across the ... The predicted similarity is: 0.82

http://www.stevens.edu/bia
LendingClub – How to Forecast the Loan Status of Loan Applications? 

Use Machine Learning Algorithms to Predict the Probability of Defaulting

Authors: Tsen-Hung Wu, Shreyas Menon
Instructor: Rong Liu

Background

- **LendingClub** is a peer-to-peer fintech company lending money to loan applicants by finding resources from individuals. Until now, 42 billion dollars have been borrowed, and 2.5 million customers are active on the platform.
- The role of LendingClub is to provide a platform that screens borrowers, facilitates the transaction, and services the loans.

Problem Statement

- To keep the business thriving, the most important problem is to forecast whether an approved loan application will default in the future or not. Otherwise, defaulting applications might jeopardize LendingClub’s reputation and bring loses.
- Precisely classifying a loan application into levels of loan status can spot the problematic loan applications beforehand.

Data Understanding

- 0.1 million of loan applications with 145 features between Jan 2018 to Mar 2018 (Q1 2018).
- Broad categorization of 145 features into 8 sections.
  - a. User (general) b. User (finance) c. Loan (general)
  - d. Current loan repayment e. Secondary applicant info
  - f. Hardship g. Settlement h. Response
- After performing feature engineering tasks, 18 additional features have been generated.

Project Flow

- **Preprocess Data**
  - Outlier & Missing data imputation, one-hot encoding, dummy features generation
- **EDA**
  - Density distribution plots, box plots, scatter plots, heat map, feature selection with statistical testing
- **Build Model**
  - Regularized logistic regression + Tree-based models + Light GBM
- **Model Validation & Fine Tuning**
  - Attain the best hyper-parameters of models and validate models to prevent overfitting issues.

Modeling & Methodology

- **Multi-classification prediction**
  - **Regularized Logistic Regression**
  - **Random Forest**
  - **Gradient Boosting Decision Tree (GBDT)**
  - **Light Gradient Boosting Machine**

  - **Optimization and Tuning**: Applied Bayesian Optimization with five folds cross-validation to determine the hyper-parameters of models.

  - **Why Bayesian Optimization?**
    1. Hyper-parameter optimal searching
    2. Fewer evaluations

Results and Evaluations

- **Response Variable**: Loan status (six levels)
- **Model Comparison**

  - **Top 10 feature importance**

Conclusions

- **Business Insights**
  1. After attained the optimal model, we can form a formal decision rule by deploying the best machine learning model to inspect a loan application which has a higher probability of defaulting in the future.
  2. Furthermore, once a loan application is issued, we can keep tracking the future loan status of it by days, weeks, or months to see the state transition.
  3. However, due to the privacy issue, LendingClub didn’t provide “id” for each application. If this variable is given, we can observe the state transition by each loan application. The corresponding risk-adjusted actions can be taken into consideration on applications turned into higher probability of defaulting.
- **Feature Importance Summary**
  - Several features are important predictors and ranked in the top ten feature importance list, indicating that these features contribute more on the response variable. Therefore, these powerful features need to be maintained precisely by engineers.
- **Model Comparison Summary**
  - Favorable results rely on the appropriate input data, properly dealing with data, and the right choice of algorithms. Finally, Light Gradient Machine outperforms than other algorithms. A combination of models can also be considered to use.

http://www.stevens.edu/bia
Visualization of Chicago Crime
Authors: Xuanyan Li, Zihan Chen
Instructor: Prof. Alkiviadis Vazacopoulos

Introduction
Using Tableau and Python, we show the distribution of each kind of crimes in Chicago city on the map. This poster also tries to explore the relationship between prime crime type and time period (Hour, Month & Date), geographic information, criminal citizenship and FBI code.

Data Source:
Crimes - 2001 to Present, Chicago Data Portal Website https://data.cityofchicago.org/

Hours – Location Relationship
These graphs show the different Chicago crime distribution in each hour. We select 5:00 am and 5:00 pm to display the result.

Prime Type – Location & Month Relationship
These four pictures utilizing maps and bar graphs to show the trend and distribution of each crime in different year, month and location. We select Narcotics to demonstrate the results.

Arrest – Type & Location Relationship
Here are the results of whether a criminal was been arrested in different time period, crime type and location information.

Domestic – Prime Type Relationship
Following graphs show how criminal citizenship relate with crime prime type, as well as the relationship with each location.
Predictive Model for House Pricing
Authors: Xuanyan Li, Zihan Chen
Instructor: Prof. Alkis Vazacopoulos

Introduction
- Our project utilizes advanced machine learning algorithms to build the predictive model for house pricing, using 80 different features and 3,000 instances.
- Strict feature selection based on Statistical tests and intuitive data visualization, as well as advanced regression techniques including Random Forest, Neural Network, Gradient Boosting and Linear Regression.
- Features including: House location, land slope, neighborhood, construction date, etc.
- Data Source: https://www.kaggle.com

Results
- **Stacking model**
  - Random Forest, Gradient Boosting, Ridge, MLP Regressor & Linear Regression
  - Evaluation: Root-Mean-Squared-Error (RMSE)

- **Using all features:**
  - Evaluation:
    - Root-Mean-Squared-Error (RMSE)
  - Applied feature selection:
  - Feature selection based on Statistical tests and intuitive data visualization, as well as advanced regression techniques including Random Forest, Neural Network, Gradient Boosting and Linear Regression.
  - Features including: House location, land slope, neighborhood, construction date, etc.
  - Data Source: https://www.kaggle.com

- **Using random forest to select important features**
  - Final features: 38

- **Results based on feature selection is not good as the former one --- overfitting**

Conclusion
- Applying the Random Forest algorithm, the top five factors which affect the house price are:
  - Overall material and finish quality;
  - Size of garage in car capacity;
  - Kitchen quality;
  - Exterior material quality;
  - Height of the basement.
- Ensemble methods – stacking of regression models win when compared with single regression methods. The mean square error is less than 0.64 in our final result, which gave us a position of top at 25% in the competition.

http://www.stevens.edu/bia
DOE for Amazon Recommendation Email

Authors: Siyan Zhang, Xuanyan Li, Biyan Xie
Instructor: Prof. Chihoon Lee

**Keywords & Programs:**
- JMP
- Fractional Factorial Design

**Business Questions:**
How can a company effectively attract people to visit the website and convert the browsers into buyers?

**Objectives:**
1. What are the important factors in Amazon recommendation emails?
2. How to apply fractional factorial design to test the best combination of factor levels

**Project Approaches:**
- Business Questions
- Design of Experiment
- Data Collection
- Analysis of Results
- Conclusion & Recommendation

**Experimental Design:**
1. Fractional factorial design with 6 factors and 2 blocks:

<table>
<thead>
<tr>
<th>Factors</th>
<th>Lower Level (-)</th>
<th>Higher Level (+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Sender’s name</td>
<td>Amazon.com</td>
<td>Amazon.com</td>
</tr>
<tr>
<td>2 Subject</td>
<td>Laptops...</td>
<td>HP Flyer Red 15.6...</td>
</tr>
<tr>
<td>3 Ad’s content</td>
<td>Popular product on Amazon (e.g. Amazon Echo)</td>
<td>Similar product with recommended ones (e.g. Samsung Odyssey 15.6)</td>
</tr>
<tr>
<td>4 Ad’s placement</td>
<td>On the top of the email</td>
<td>On the bottom of the email</td>
</tr>
<tr>
<td>5 Provide rating and review numbers</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>6 Product list sorting by price</td>
<td>Randomly</td>
<td>From lowest to highest</td>
</tr>
</tbody>
</table>

| Block Membership status | 1: Non-prime | 2: Prime |

2. 16 recommendation emails were created with 3 responses for each.
3. Target Audience: Students at Stevens Institute of Technology

**Results:**

**Significant Factors:**
- Subject
- Rating and Review
- Sender’s name * Subject
- Sender’s name * Sort by price

(*) denotes interaction effect

**1. Pareto Plot:**

**2. Normal Plot:**

**Regression Equation: Expected Response Under**

Desire of click = through (Y) = 5.6458 + 0.5625 * (1) + 0.5208 * (-1) - 0.6042 * (-1) * (1) + 0.6458 * (-1) * (-1) - 0.7048

**Best Experiment Setting: Recommendation:**

Desire of click = through (Y) = 5.6458 + 0.5625 * (1) + 0.5208 * (-1) - 0.6042 * (-1) * (1) + 0.6458 * (-1) * (-1) - 0.7048

- Sender’s name: Amazon.com
- Subject: Specific products
- Provide rating and review numbers: Yes
- Product list sorting by price: Random

**Future improvement:**
1. To improve responses: a) increase replications; b) use click-through rate as a response;
2. To acquire better data: randomly select students from a school’s enrollment list;
3. To improve survey method: send emails to respondents simultaneously.

http://www.stevens.edu/bia
Predicting trends in bike sharing program

Authors: Zixuan Wang, Shuqiong Chen, Kevin Walsh
Instructor: Amir H. Gandomi

Introduction

Problem:
• Predicted the amount of riders under different conditions by multiple linear regression and multiple polynomial regression
• Predicted whether the bikes would be heavily used

Business value:
• This model will help companies to distribute their bikes in a reasonable way.
• Companies will reduce their cost on bike dispatch and help to increase their retention rate of customers to increase business awareness.

Data understanding and processing

• Dataset: 17,379 bike sharing records with 11 variables, including continuous, binary, and categorical variables.

| Weekday | Holiday | Plasma | Month | Year | Weather | Hdry | Temp | Humidity | Wind | Season
|---------|---------|--------|--------|------|---------|------|------|----------|------|--------
| 1       | 0       | 0      | 0      | 0    | 0       | 0    | 0    | 0        | 0    | 0      
| 2       | 0       | 0      | 0      | 0    | 0       | 0    | 0    | 0        | 0    | 0      
| 3       | 0       | 0      | 0      | 0    | 0       | 0    | 0    | 0        | 0    | 0      
| 4       | 0       | 0      | 0      | 0    | 0       | 0    | 0    | 0        | 0    | 0      
| 5       | 0       | 0      | 0      | 0    | 0       | 0    | 0    | 0        | 0    | 0      
| 6       | 0       | 0      | 0      | 0    | 0       | 0    | 0    | 0        | 0    | 0      
| 7       | 0       | 0      | 0      | 0    | 0       | 0    | 0    | 0        | 0    | 0      

• Data insights: The whole dataset contains continuous 724 days’ record. From the exact date information from dataset, we can also conclude that this dataset is record from 2011-2012.

• Feature engineering: Understanding dataset helps us extract more information about from the raw data. Now we can impute date and year label for each row. For periodic features, such as hour and day, we can use polar coordinates to transform them, so each point can be calculated through trigonometric function. In this way, we can replace “Month”, “Season”.

• Outliers:

From the plot between different variables, we find some outliers. After observing the detail data record, we use different method to deal with the outliers.

| Correlation Coefficient | Season | Temperature | Humidity | Wind | Hdry | Date | Hdry | Temp | Humidity | Wind | Season
|-------------------------|--------|-------------|----------|------|------|------|------|------|----------|------|--------
| 0.5                    | 0.4    | 0.3         | 0.2      | 0.1  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0      | 0.0  | 0.0    
| 0.4                    | 0.3    | 0.2         | 0.1      | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0      | 0.0  | 0.0    
| 0.3                    | 0.2    | 0.1         | 0.0      | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0      | 0.0  | 0.0    
| 0.2                    | 0.1    | 0.0         | 0.0      | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0      | 0.0  | 0.0    
| 0.1                    | 0.0    | 0.0         | 0.0      | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0      | 0.0  | 0.0    

Conclusion & Future Work

• Regression model is highly descriptive. People would like to use bikes in a warmer daytime.
• Random Forest brings the best result for classification problem.
• The number of riders in 2012 was visibly increasing compared with that in 2011. We suggested the company to analyze the operational strategy since there is little difference of external factors between 2011 and 2012.

http://www.stevens.edu/bia
NBA player management Optimization

Authors: Zixuan Wang, Jingchen Lan, Shuqiong Chen, Shan Jiang
Instructor: Alkiviadis Vazacopoulos

Introduction

➢ The Problem:
More and more NBA teams tend to organize a star team, such as Warriors (16-18), Heat (11-13). Our work is to help the team to choose players to form a competitive team within a reasonable salary constraint and discriminate players who have potential to be super stars in future. In this paper, by adopting an optimization solution we organized the most competitive team, by predicting each player’s potential.

➢ Modeling:
This model can help team managers to use their budget to construct a highly competitive team. It also provides a method to judge players and to make them perform well.

Analysis of players and positions

➢ Data processing:
- Dropped some variables with high multicollinearity like ‘3P’
- Used minimum to fill null value in the column ‘3P%’
- Deleted some interactive positions like ‘PF-C’

| Year | Player Pos Age Tm PB RS TS% USG% BPM VORB 3P% 2P% TRB STL AST BLK TOV |
|------|-----------------|---------------|-----------|-----------|----------------|----------------|----------------|----------------|-----------|----------------|---------------|----------------|----------------|----------------|----------------|----------------|----------------|-----------|----------------|---------------|
| 2012 | Jeff Adrien PF  25  HOU 11.2 0.650 16.0 -7.7 -1.1 0.000 0.458 22 1 0 2 2  |
| 2012 | Amare Stoudemire SG 28  PHI 14.7 0.594 19.1 1.8 1.5 0.398 0.004 197 169 36 13 85 |
| 2013 | Bismack Biyombo PG 27  UT 11.3 0.027 28.4 -16.2 -0.3 0.122 0.405 2 1 0 0 0  |
| 2013 | Solomon Alabi C  23  TOR 14.2 0.418 17.2 -4.1 -8.1 0.000 0.361 47 3 2 8 5  |
| 2012 | Cole Aldrich C  23  OKC 17.7 0.502 14.0 0.3 0.1 0.000 0.524 40 3 8 8 9  |

Order

➢ Position classification:

- Three classes: C, SG&PG, PF&SF
- Used BPM value as criterion
- C: BPM is almost decided by 2P%
- PF&SF:BPM is positive correlated with TS%
- SG&PG:BPM is negative correlated with TS%


➢ Player classification:

Players in different positions show a different distribution. Used GVF (0.8) to determine the number of leaves on each position.

The distribution of ‘SG’, ‘PG’ and ‘C’ shows a polarization. Both the number of players with good and bad performance is large. Players in ‘PF’ and ‘SF’ are relatively equal.

Conclusion & Future Work

- According to our model, we recommend teams which want to choose new players to study on ‘TS%’ and ‘Score’ by positions. Players with outstanding score ability are easy to be favored by team and audience. However, if the position doesn’t match his ability in the association, his performance may have a negative effect on team.
- The whole NBA eliminate-select process can be extended to the personnel departments of various companies. Based on this model, the company can recruit the most talented candidates and align them to employees who contribute the least to the company.

http://www.stevens.edu/
Improvement of medical wire manufacturing

Authors: Zxuan Wang, Jingchen Lan
Instructor: Chihoon Lee

Medical wire is a very inconspicuous device but plays a big role during the surgery. The strength of the Medical wire is critical.

**Problem:**
- Design an experiment with four main factors to test the effects of factors or interactions to the Yield Strength of the medical wire.

**Variables**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spool ID</td>
<td>Identify spools used in the experiment. 1-48</td>
</tr>
<tr>
<td>Block</td>
<td>Doctors who handle the machines: 1=Doctor 1, 2=Doctor 2, 3=Doctor 3.</td>
</tr>
<tr>
<td>Machine</td>
<td>Drawing Machine: X1=Type 1, X1=Type 2.</td>
</tr>
<tr>
<td>Angle</td>
<td>Die Reduction Angle: X2=8.5-10, X2=11.1-12.1.</td>
</tr>
<tr>
<td>Diameter</td>
<td>Die Bearing Length: X3=Short 15% - 20%, X3=28% - 36%.</td>
</tr>
<tr>
<td>Yield Strength (YS/UTS)</td>
<td>Estimates the Yield Strength in Medical Wire.</td>
</tr>
</tbody>
</table>

**Implications:**
- This experiment will help manufacturers to select a more suitable method to improve efficiency and quality on their operation.

**Fractional factorial design**

- **Design of experiment:** 24-1 fractional factorial with 4 factors, 3 blocks and 5 replications.
- We used 24 different spools at this level of experiment.

**Outliers:**
- Conducted a regression analysis using the design table directly.

**R-Square adj=28%**

So it is necessary to explore the data and try to exclude outliers.

**Regression equation:**

\[ Y = 0.41 \times B + 0.35 \times AB + 0.16 \times BC + 93.4 \]

**Effect graph:**
- From the main effect and interaction plots we can make a conclusion that optimization of the process may be a consequence of the combination of multifactor.

**Conclusion:**
- We find the manufacturing of medical wires is a multi-factor interactive process. Experiments of changing one factor at a time may not maximize the YieldStrength.
- For a medical wire manufacturer, we recommend using a short bearing length and wide reduction angle with type 2 machine (even machine is not a significant factor).
Identify the Safety Level of Precincts in NYC
Authors: Chen Liao, Yu Hong, Tianyu Liu, Xiangxiang He  Instructor: Feng Mai

Introduction
• Identify the safety level of each precinct and borough according to past complaints records
• Explore the relationship between offense-level, time and precincts in New York City.

Data Understanding
• New York Police Department Public Data
• The origin dataset contains 23 columns, we only use 7 of them
• 2,714,699 complaints in total, from 2012 to 2015

<table>
<thead>
<tr>
<th>Date</th>
<th>Type</th>
<th>Level</th>
<th>Borough</th>
<th>precinct</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>12/31/2015</td>
<td>COMPLETED</td>
<td>FELONY</td>
<td>BROOKLYN</td>
<td>103</td>
<td>-73.946241</td>
<td>-40.6820863</td>
</tr>
<tr>
<td>12/31/2015</td>
<td>COMPLETED</td>
<td>FELONY</td>
<td>QUEENS</td>
<td>104</td>
<td>-73.945624</td>
<td>-40.687034</td>
</tr>
<tr>
<td>12/31/2015</td>
<td>COMPLETED</td>
<td>FELONY</td>
<td>MANHATTAN</td>
<td>28</td>
<td>-73.943925</td>
<td>-40.685963</td>
</tr>
<tr>
<td>12/31/2015</td>
<td>COMPLETED</td>
<td>MISDEMEANOR</td>
<td>QUEENS</td>
<td>305</td>
<td>-73.945644</td>
<td>-40.687030</td>
</tr>
<tr>
<td>12/31/2015</td>
<td>COMPLETED</td>
<td>MISDEMEANOR</td>
<td>MANHATTAN</td>
<td>15</td>
<td>-73.943925</td>
<td>-40.685963</td>
</tr>
</tbody>
</table>

• New York City Population (CITYPOPULATION)

<table>
<thead>
<tr>
<th>Borough</th>
<th>CITYPOPULATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bronx</td>
<td>1,969,796</td>
</tr>
<tr>
<td>Brooklyn (King County)</td>
<td>2,484,713</td>
</tr>
<tr>
<td>Manhattan (New York City)</td>
<td>2,934,717</td>
</tr>
<tr>
<td>Queens</td>
<td>2,093,960</td>
</tr>
<tr>
<td>Staten Island/Brooklyn (County)</td>
<td>2,093,960</td>
</tr>
<tr>
<td>Staten Island</td>
<td>1,457,963</td>
</tr>
</tbody>
</table>

Networks and Analysis
Overall view
Overall Network Communities
Networks Based on Different Offense Level
Number of Crimes vs Special Days

Data Preparation
Network Construction
• For each precinct, calculate the number of complaints base on different type (attempted / completed) and offense level (misdemeanor / violation / felony).
• Assign the scores base on type, level and number of complaints.
• Calculate the average longitude and latitude of each precinct.
• Compute the pair-wise Euclidean distance of precincts base on standardized scores, longitude and latitude.

Analysis
Crime Rate vs Precinct
QAP Tests(Quadratic Assignment Procedure)
ERGM(Exponential random graph model)
The ERGM plots above show the vital variables which can affect the network connection most.

Conclusion
• Precinct 14 in Manhattan and precinct 75 in Brooklyn has the highest crime rate and number of complaints.
• During Christmas and Thanksgiving, more misdemeanor and felony type of crimes, but less violation type of crimes than normal days in most precincts.
• According to QAP tests, precincts have almost the same probability of crime occurrence.

• The ERGM plots illustrates that the boroughs, the communities and the transitivity contributes the most to the network connection.
• Any question about this poster please contact cliao4@stevens.edu, yhong5@stevens.edu, ttlu31@stevens.edu, xhe19@stevens.edu

http://www.stevens.edu/bia
Background
Credit analysis is used to determine the risk associated with repayment of the loan. It helps to understand the creditworthiness of a business or a person. The financial crisis in the year 2008 has brought public awareness of the importance of risk management and management.
The project aims to address the mis-specified & outdated stress testing model, and non-informative data problem, which was two of the main reasons of the 2008 financial crisis, with the help with Data Visualization and Machine Learning.

Analysis
The distribution plot shows the transition of Borrower’s Credit Score from 2000 to 2017. It clearly explains the reason for the 2008 Financial Crisis. The frequency of loans given to lower credit score (Subprime Mortgages) is more before the crisis than compared to the frequency during and after the crisis. Subprime Mortgages was one of the main reason for the 2008 Financial Crisis.

Technology /Methodology

Machine Learning and Prediction
According to the ROC curve above, the AUC value for SVM (85%) and RF (94%) models are both higher than logistic regressions (81%). RF models have the highest accuracy and Naive Bayes model has the lowest one (74%).

Random Forest is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification).

Consider, \( X = x_1, \ldots, x_n \) - training set, \( Y = y_1, \ldots, y_n \) - response, the number of samples/trees: B:
Bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:
For \( b = 1, \ldots, B \):
1. Sample, with replacement, \( n \) training examples from \( X, Y \); call these \( x_b, y_b \).
2. Train a classification tree \( f_b \) on \( x_b, y_b \).

After training, predictions for unseen samples from testing set can be made by taking the mode of the \( f_b \) values (classification).

<table>
<thead>
<tr>
<th>Best Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.92</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.95</td>
<td>0.92</td>
<td>0.94</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
</tr>
</tbody>
</table>

The above table give precision, recall, and F1-score of the Random Forest model. It has very high precision recall and f-score, thus making the best prediction of loan default.

Conclusion
The project compared result of different machine learning models. Random Forest (RF) scored the best result (AUC= 0.94), and Naive Bayes had the worst result (AUC=0.74). Logistic Regression is the common model that Office of Federal Housing Enterprise Oversight (OFHEO) and other studies to model the mortgage loan default and prepayments, which the model accuracy can be improved if supervised machine learning technique has been implemented.
Better Photography using Design of Experiments
Authors: Sibo Xu, Ping-Lun Yeh, Kumar Bipulesh, Sanjay Pattanayak
Instructor: Dr. Chihoon Lee

The Problem
- Photography Enthusiasts buy expensive DSLRs but shoot in Auto Mode
- Common thinking that a high-end camera automatically makes the photographs amazing

Why Design an experiment?
- To understand what factors influence the quality of a photograph most in a setting
- To help photographer improve their skills in shooting pictures

Do camera settings affect the image quality?

<table>
<thead>
<tr>
<th>Aperture</th>
<th>White Balance</th>
<th>Shutter Speed</th>
<th>Metering Mode</th>
<th>Angle</th>
<th>Stabilizer</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Cloudy</td>
<td>1/100</td>
<td>Center-weighted</td>
<td>Left-hand</td>
<td>On</td>
</tr>
</tbody>
</table>

Results
Fractional Factorial Design

Regression Equation for Predicting Ratings

Max Rating obtained: 9.27
Optimal Setting: Aperture[L1], White Balance[L2], Shutter Speed[L1], Metering Mode[L1]

One Factor at a Time (OFAT):

<table>
<thead>
<tr>
<th>Aperture</th>
<th>White Balance</th>
<th>Shutter Speed</th>
<th>Metering Mode</th>
<th>Distance</th>
<th>Stabilizer</th>
<th>Judge 1</th>
<th>Judge 2</th>
<th>Judge 3</th>
<th>Judge 4</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>3.29</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3.13</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>3.5</td>
<td>3.37</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>8</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>9.08</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9.00</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9.00</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>8</td>
<td>10</td>
<td>9</td>
<td>9.00</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>7</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>9.00</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>7</td>
<td>8</td>
<td>10</td>
<td>9</td>
<td>9.00</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>7</td>
<td>8</td>
<td>10</td>
<td>9</td>
<td>9.00</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>7</td>
<td>5</td>
<td>8</td>
<td>10</td>
<td>9.00</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>9.5</td>
<td>9.25</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>9.5</td>
<td>9.25</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>9.5</td>
<td>9.25</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9.5</td>
<td>9.25</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>9.5</td>
<td>9.25</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9.5</td>
<td>9.25</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>9.5</td>
<td>9.25</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9.5</td>
<td>9.25</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>9.5</td>
<td>9.25</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9.5</td>
<td>9.25</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>9.5</td>
<td>9.25</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>9.5</td>
<td>9.25</td>
</tr>
</tbody>
</table>

Response: Ratings (1-10 scale) provided by a Judge for each photograph.

Fractional Factorial Design 2^6-2:
- Choose two levels (+) and (-) of each factor
- Conduct with Replication and Blocking effects

Replication Effect is the repetition of an experimental condition so that the variability associated with the phenomenon can be estimated.

Blocking Effect reduces unexplained variability. For this experiment 4 of the judges acted as the blocks.

One Factor at a Time (OFAT):
- Rate one setting (all +) at the beginning and change the first factor of the setting to negative level, then compare their rating scores
- Pick best rated setting and change the second factor for the next round comparison
- Until there is no other factor to change then stop
- 84 run of OFAT was performed.

Conclusion
- Aperture, Shutter speed, White Balance, Metering Mode, Block, and Interaction of Aperture and Shutter Speed are significant factors.
- Performing replication and blocking helped identify the significant factors with blocks serving as significant factor.
- OFAT design helps us quickly identify the optimal setting for photography, although it doesn’t guarantee the same measurable precision for the quality.
Driver Safety using CNN & Transfer Learning
Author: Kumar Bipulesh
Instructor: Dr. Christopher Asakiewicz

The Problem
**Distracted Driving**
- With constant incoming messages and alerts on our mobile devices, the fear of missing out (FOMO) and desire to multi-task distracted driving is a big issue.
- In 2015, **3000** were killed and **425,000** were injured as a result of distracted driving.

**Driver Drowsiness**
- Falling asleep after long-tiring work schedules or during long boring journeys are equally life threatening.

The Proposed Solution
- A Neural Network driven framework that has the ability to classify with great precision what a driver is doing based on real time image analysis
- Ability to alert (sound signals) in case of detection of drowsiness or distraction
- Using in-expensive mobile phones with front-facing facing camera as the input (driver’s image), analysis (prediction of behavior) and output (sound alerts)

Solution Approach

A. Image Analysis for Distraction Detection

**Data Cleaning**

![Image](http://www.stevens.edu/bia)

Transfer Learning with Deep Convolutional Network VGG16

Transfer Learning can be applied to CNNs when there is insufficient data or computational power. Using a model with pre-trained weights in a much larger database, the model can be enhanced to solve one problem. In this case it is the 10 class prediction based on image being analyzed.

B. Image Analysis for Drowsiness

```python
# initialize dlib’s face detector (HOG-based) and then create
# the facial landmark predictor
print("[INFO] loading facial landmark predictor...")
detector = dlib.get_frontal_face_detector()
predictor = dlib.shape_predictor(args["shape_predictor"])

# grab the indexes of the facial landmarks for the left and
# right eye, respectively
(lStart, lEnd) = face_utils.FACIAL_LANDMARKS_IDXS["left_eye"]
(rStart, rEnd) = face_utils.FACIAL_LANDMARKS_IDXS["right_eye"]
```

Findings
- Using pre-trained VGG16 model and retraining it for distracted driver data significantly reduces training time
- Transfer Learning helps in training a deep CNN even on regular CPUs with great model performance and accuracy, and this is apt to fine tune pre-trained models
Machine Learning to Predict US Green Card Approvals
Authors: Smriti Vimal, Sanjay Pattanayak
Instructors: Prof. Chris Asakiewicz, Prof. Khasha Dehnad

Motivation
- Thousands of application for Green Card are filed every year. GC approval is critical and is a first step towards attaining US citizenship. Through this project the companies get an insight of the application features and probability of approval.
- Companies and individuals are keen to get the GC application approved as they have a lot in stake.
- Through this project we try to apply ML algorithms to predict the approval/denial of an application.
- We also try to visualize the data and find important features for predicting the application status.

Data Analysis
- Total Data Observations: 374,000
- Total Features: 154

Feature Importance

Experiment
- Converting different units of time into single unit.
- Converting salary into bins of salary range.
- Selecting features with more than 330,000 data.
- Converting case labels into binary.
- Converting state names into labels.
- Data imputation with mean, mode.
- Converting features into numerical values using Label Encoder.
- Importing Logistic Regression Classifier, GridSearchCV, train_test_split and accuracy metrics from sklearn.
- Importing k-Nearest Neighbors Classifier and performing classification.
- Importing RandomForestClassifier from sklearn for classification
- Importing GradientBoostingClassifier from sklearn.
- Analyzing all the results of all classification method.

Performance

Method and Results
- The data selection and analysis gives important insights, that are visualized above.
- Data imputation, data binning, and conversion of units of measurement into single units is vital for ML algorithm’s function.
- Feature importance helps to analyze and give insight into the various features and their contribution to ML algorithm.
- Classification results are measured by the accuracy, precision and recall of the algorithm.
- Python sklearn library is primarily used for ML.
- Python matplotlib is used for visualization
- Further analysis and prediction capabilities are possible with this ML model.
UBSPitch 2018 1st Prize Winners
Machine Learning & Automation
Authors: Monica Vijaywargi, Poojan Gajera, Rohan Gala, Sanjay Pattanayak, Xunyan Li
Mentors: Prof.Stohr, Prof.Daneshmund, Prof.Dehnad, Prof.Belanger, Wonmoh Li, Vasuki Neelgiri

Pitch
- UBS is always looking for more innovative ways to connect with and provide value to its clients and prospects. In pursuit of this goal, we are applying Machine Learning to identify new locations for branch offices.
- For this competition, we are tasked with using Machine Learning to identify the next 3 locations where UBS should open a Wealth Management branch.
- As part of this, we are required to come up with a solution that is completely automated and can be repeated upon request.

Data Source Selection
- **Public Sources**: (Income Tax Returns Data (Along with the Tax Slabs, Average Zillow Home Value, Unemployment Data)
- **Private Sources**: Charitable Donors, Finance Start-ups Data, Competitor Data, Buying Style)
- **Scraping Websites**: (Barron’s List, Political Donations)

Feature Selection
- **ML model Errors**: Bias Error, Variance Error, Intraducible Error, Feature Selection
- **Correcting Model Error**: Feature Importance, Mean decrease in impurity (Gini/Entropy), Mean decrease accuracy
- **Benefits**: Important features selection, Increase accuracy, Reduce error, Reduce Overfitting, Cost and effort Saving

Automation
- **Weights keeping attributes of HNWIs into consideration**
- **The Top 3 Cities!!!**
  - [Map Image of Top Cities]

Method and Results
- Selection of data and features is vital. Collecting HNWI attributes’ data and competition data are vital.
- Using clustering enables to convert unstructured data into structured data so that ML algorithms can be applied to find cities.
- Multiple algorithms are used to enhance the prediction of the cities. Ensemble and individual classification algorithms enable the best city selection.
- Feature selection helps to reduce the errors and effort required for data collection and cleaning.
- Automation involved website design that can be used both on a handheld mobile device or a computer desktop or laptop.
- Automation enables the leadership team to select/prioritize the features based on which they want to select a location.

http://www.stevens.edu/bia
Classifying Restaurant Ratings
Authors: Xiaojun Zhu, Jhao-Han Chen, Haiping Sun
Instructor: Amir H. Gandomi

Introduction
In our daily life, people often use mobile applications to view a restaurant's rating and decide which restaurant to eat. Therefore, figuring out which variables have a greater impact on restaurant rating is important for entrepreneurs while starting a restaurant. The goal of our project is to find the best-classified method to discriminate the good and the bad restaurant according to the variables in the dataset we used.

Experiment

Data Collection
Our data comes from website Kaggle and the original dataset includes 11 features such as restaurant category, station, Review Number, Dinner Rating, Dinner Price and so on.

Here is the explanation of some features

<table>
<thead>
<tr>
<th>station_class0</th>
<th>Low density</th>
<th>C_Cafe</th>
<th>FirstCategory for Cafe</th>
</tr>
</thead>
<tbody>
<tr>
<td>station_class1</td>
<td>Middle Density</td>
<td>C_European</td>
<td>FirstCategory for European</td>
</tr>
<tr>
<td>station_class2</td>
<td>High Density</td>
<td>C_Japanese</td>
<td>FirstCategory for Japanese</td>
</tr>
<tr>
<td>C_Asian</td>
<td>FirstCategory for Asian</td>
<td>C_Noodle</td>
<td>FirstCategory for Noodle</td>
</tr>
<tr>
<td>C_Bar</td>
<td>FirstCategory for Bar</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data Cleaning
Firstly, as we process the raw data, we found that categories such as BBQ, Mexico and Seafood are rare, so we delete those data.

Then, to make it easier to analysis, we separate variables ‘station’ and ‘First Category’ into dummy variables.

Finally, as our target cell is dinner rating, we take the rating more than 3.07, which is the median of the rating, as good restaurant and rating less than 3.07 as bad restaurant.

Feature Selection
After cleaning the data, we generate a correlation coefficient matrix to see which factor should we use in our analysis. The result shows that variables “station_class1” and “station_class2”, variables “C_Bar” and “C_Japanese” have high correlation which are greater than 0.5, so we decide to drop variables “station_class1” and “C_Japanese”.

Model
Our group test four classification machine learning algorithms to build model including Logistic Regression, Naïve Bayes, K-Nearest Neighbors (KNN), and Linear discriminant analysis.

Result
By testing and comparing the accuracy of several models, finally KNN is selected as the best model to fit our data set. Also while K = 5, the Model has the highest accuracy rate.

10 randomly chosen restaurants to test our best model:

The result of the classification have high accuracy rate based on our model.

Conclusion
• We achieve AUC of up to 0.8601 by using KNN model with K = 5.
• KNN algorithm is sensitive to distance between variables, therefore it is better to normalize/standardize the features
Consumer Analytics for Restaurant Preferences using Yelp User Reviews
Authors: Xiaojun Zhu, Haodong Zhao, Yuhan Su
Instructor: Feng Mai

Motivation
- **For Customers:**
  With the advent of the Internet age, customers can gather information of many restaurants online. Furthermore, they can also read reviews or post their own reviews.
- **For Companies:**
  High-quality reviews are valuable because they are helpful for restaurants to improve themselves. And high-quality reviews shows the active users' behavior patterns, this will be helpful for restaurants to attract more users.

Introduction
- **Objective:**
  Find the relationship between different kind of users and their preference for restaurants.
- **Active user:**
  They are willing to post their own reviews and always spend more money than others.
- **Dataset:**
  Yelp business and users' reviews.
  Business raw data: 188,593 companies.
  Reviews raw data: 5,996,995 reviews.

Approach
- **Clean yelp raw data and filter restaurant data out of the total data.**
- **Based on users and the quantity of reviews they posted on Yelp, category users into active users or non-active users.**
- **Match users with restaurants, including features of them.**
- **Analyze which restaurants the active users prefer to go.**
- **Visualize the result.**

Data Analysis Result

<table>
<thead>
<tr>
<th>Dataset Type</th>
<th>Active User</th>
<th>NonActive User</th>
<th>1Star</th>
<th>2Star</th>
<th>3Star</th>
<th>4Star</th>
<th>5Star</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>Reviews</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

http://www.stevens.edu/bia
Portfolio Optimization using Python

Authors: Liang An, Jhao-Han Chen, Xuanzhu Luo, Jiamei Wang, Ming Wei
Instructor: Professor Alkis Vazacopoulos

Project Overview

Our project objective is to select the best asset allocation. We compare the expected return, expected volatility, and the Sharpe ratio while minimizing the variance (risk) and maximizing the Sharpe ratio. All the analysis and results are generated using Python.

Modeling

1. Select 10 stocks from different sectors
2. Calculate annual mean returns of each stock
3. Calculate covariance
4. Initialize weights randomly
5. Calculate expected annual mean return, variance, and standard deviation of the portfolio
6. Used Monte Carlo Simulation to produce random weight vectors and records the expected returns and variances of the random combination.
7. Optimization 1: Maximize Sharpe Ratio
8. Optimization 2: Minimize Variance (Risk)

Methods

I. Monte Carlo Simulation
We use Monte Carlo Simulation to run 10,000 different randomly generated weights for the individual stocks and then calculate the expected return, expected volatility and Sharpe Ratio for each of the randomly generated portfolios.

II. Efficient Frontier
The efficient frontier is the set of optimal portfolios that offers the highest expected return for a defined level of risk or the lowest risk for a given level of expected return. Portfolios that lie below the efficient frontier are sub-optimal because they do not provide enough return for the level of risk.

III. Sharpe Ratio
The ratio describes how much excess return you are receiving for the extra volatility that you endure for holding a riskier asset

Conclusion

When we use Monte Carlo simulation in Python, the expected return and Sharpe Ratio are higher when the objective is to maximize Sharpe Ratio than the expected return and Sharpe Ratio when the objective is to minimize risk.

However, the expected volatility have little difference, for example, one of the expected volatility is 0.147, and the other is 0.143. We can ignore the difference after considering the higher expected return and Sharpe Ratio we obtained by maximizing the Sharpe Ratio, and take the weight of the portfolio when maximizing the Sharpe Ratio as the optimal weight.

From the optimal weight we acquired, we can build a portfolio with 9.9% annual return and 0.674 Sharpe Ratio by investing 15% in GOOGL, 35.1% in JNJ, 32% in KO, 2.1% in VZ and 15.8% in WMT. Investors can find their targeted stocks to customize stocks portfolio from the results achieved in this model.
OBJECTIVE: To create a model that recognizes different x-ray types by exposing them through an input device.

INTRODUCTION: Machine learning as a technology that has several implementations and the scope of growth is beyond current human comprehension. During our research to choose a topic we came across an idea in which we wanted to expose an X-ray copy to our model through an input device and it would give an output which would tell us whether the lungs are infected with a disease or not.

DATA: We derived our dataset online over which we implemented our neural network to develop the model. The dataset consists of multiple folders with x-rays of different diseases segregated accordingly.

MODELING: We have used Convoluted Neural Networks (CNN) for analyzing our input i.e. x-ray images.

The CNN architectures make the assumption that the inputs are images, which allows us to encode certain properties into the architecture. These then make the forward function more efficient to implement and vastly reduce the amount of parameters in the network.

We used 3 convolutional layers and 3 maxpooling layers of 2x2 size in our CNN.

Our model experiences a log loss of 0.8.

OPENCV: We have utilized OpenCV library for computer vision. We have used a camera as input device through which the x-ray images are fed.

RESULTS: After training our model on the dataset we then fed a sample x-ray from the input device and the program generates an output which tells us whether the set of lungs have contracted a disease or not and if yes, then which disease would it be.

The heatmap on the output image tells us about the probability of whether that area is infected or not. The green patches on the lower side of both the lungs show positive prediction of an infection.

We tested some random x-ray to test the model. Example: This image shows the symptoms of effusion and when we tested it using our model we also got the same result. [0,1,0,0]

CONCLUSION CNNs compare images piece by piece. The pieces that it looks for are called features. By finding rough feature matches in roughly the same positions in two images, CNNs get a lot better at seeing
Predicting Change in Bitcoin Prices Using Public Sentiment for Cryptocurrency on Twitter

Introduction

OBJECTIVE
We tried to provide a measure of sentiment that could accurately predict future Bitcoin prices based on the sentiment from well-known accounts on Twitter regarding the cryptocurrency.

BACKGROUND
When trying to predict the pricing surges of Bitcoin, they seem to appear due more to consumer sentiment rather than the actual performance of Bitcoin. The use scenario would be using the non-traditional consumer sentiment, through the general consensus of Twitter environment regarding Bitcoin. The forecasting problem is the price of Bitcoin is highly unpredictable and highly volatile due to it trading more on public opinion rather than professional analysis reports.

Flow - Project

Flow Diagram:
- Creation and implementation of code to acquire tweets and sentiment.
- Downloading historical Bitcoin prices dating back to start of set of tweets.
- This segment included the application of formulae to calculate the Like and Retweet sentiment parameters. (Eq 1.2).
- Tweets based on a Twitter-API binary sentiment of either positive or negative.
- Tweets with likes and retweets above the average number help in supporting the correlation.
- So far, we are able to predict Bitcoin’s future change in price with 62.5% accuracy.
- We are confident that we can significantly increase our predictive accuracy in the future.

Equations & Modelling

- Average Likes and Retweets.

\[ L_t = \frac{2018.11.12 L_t}{n(t)} \quad \text{Eq1} \]

\[ R_t = \frac{2018.11.12 R_t}{n(r_t)} \quad \text{Eq2} \]

- Models
  - Logistic Regression
  - LDA and QDA
  - K-Nearest Neighbor
  - Decision Trees
  - Support Vector Classifiers
  - Bagging, Boosting, Random Forest

Results

- Bitcoin Price over the last 238 Days

After running a Logistic Regression, we were able to identify a strong correlation between Bitcoin’s price movement and tweets made on Twitter about Bitcoin. Using a variety of classification algorithms in Python, we were able to predict Bitcoin’s future change in price, especially using Quadratic Discriminant Analysis.

The x-axis shows different days out of a range of 238 days, ending with 11/12/2018, and the y-axis shows the price of Bitcoin and the number of like on tweets regarding Bitcoin on certain days. Looking at the relation between the two graphs, there is a sense of correlation. On days with a large number of like such as approximately Day 25-40 and right around Day 100, there is an increase in the price of Bitcoin in the following days.

Conclusion & Future Scope

- With the above methodology, we have been able to identify a correlation between the Twitter sentiment and Bitcoin prices.
- The same methodology can be extended to different type of equities and their public sentiment.
- Also, application of different parameters on Twitter as well as an extension to other social media platforms could help us more accurately predict a public sentiment to price relation.
- Next year, we plan to add more in-depth variables and models to continue to analyze a correlation between the sentiment and price.

http://www.stevens.edu/bia
Overview

- New and novel approaches that will enable clinicians to differentiate indolent and lethal prostate cancer so as to aggressively treat only the latter type.
- Distinguishing the indolent cases from the minority of lethal ones in order to minimize overtreatment intervention.

Introduction

- Prostate cancer is the most commonly diagnosed non-cutaneous cancer in men in the US, with approximately 1 in 6 men will be diagnosed with prostate cancer at some point in their lives.
- Not all cancers are equal since prostate cancer is heterogeneous and can follow multiple paths. An indolent case is induced that can cause no harm during patients’ lifetime.
- PSA (prostate specific antigen) level cannot predict prostate cancer with high degree of accuracy.

Research Objectives

- We hypothesize that SERS can be utilized to detect biomarker(s) in the urine of prostate cancer patients that could serve as an indicator of severity of cancer.
- Statistical analysis method, PCA-LDA, can be applied into differentiate SERS spectra of indolent and lethal cases.

Our Methods

- Carrying out SERS investigation of banked urine specimens from two groups of prostate cancer patients. Group A—low risk, indolent cases. Group B—high risk, lethal cases.
- Conducting principle component analysis (PCA) and linear discriminant analysis (LDA) of SERS spectra for classification.
- Establishing the correlation between outcome of the prostate cancer.
- The SERS findings and clinical

Current Results

- SERS spectra of indolent and lethal cases.
- The scatter plot of LDA of binary classifications after PCA-LDA treatment.
- 20 urine samples are analyzed, 16 in the indolent group and 4 in the lethal group. For each sample, the SERS measurement is repeated 6 times. In total, 120 SERS spectra are acquired and used.
- PCA is used to visualize the variances between groups with corresponding to score plots. PCs with higher associations with their class identities are preferentially used in LDA to achieve a better classification performance.

Impact Statement

- This discriminant of biomarker(s) will further increase the accuracy of the current prediction of tumor aggressiveness, and guide physicians to choose a safe approach, increase the survival rate.
- Decrease the possibility of leading overtreatment and avoids the risk in both economical cost and long-lasting side effects.
- The method identifies the novel biomarker(s) also can be applied in other disease treatment which is closely related with the urethra system.

http://www.stevens.edu/bia
Customer Churn Rate Analytics: Predictive Analysis Model for Retaining Customers

Authors: Shangjun Jiang, Shan Jiang, Hongyi Chen
Instructor: Amir H Gandomi

INTRODUCTION

- **Key Words:** churn rate, classification analysis, optimization
- **Background:**
  The U.S. telecom market continues to witness intense competition, every company launches competitive pricing and service plan to attract customers and increase sales. Though it’s easier to acquire new customers, keeping customers sticking around is more important to build a thriving business. Customer churn rate is a critical metric that determines your overall business success. And this creed does not only apply to the telecom industry, but to the entire commercial market.
- **Project Goal:**
  By conducting analysis to the data of customer’s contract info and different services usages, to find out the top 5 key variables to influence customers' decision on whether stay or not, and to suggest the telecom company to work deeper on these five variables to generate more competitive service plan, so to decrease churn rate and increase revenue.

PROCESS FLOW

**MODELLING**

Linear Discriminant Analysis

Coefficients of linear discriminant:

<table>
<thead>
<tr>
<th>Account Length</th>
<th>New Voicemail Messages</th>
<th>State Call Charges</th>
<th>Total Calls</th>
<th>Total Calls in State</th>
<th>Total Revenue</th>
<th>Total rains</th>
<th>Total Churn Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.0089</td>
<td>0.6783</td>
<td>0.4306</td>
<td>0.2325</td>
<td>0.1384</td>
<td>0.1712</td>
<td>0.1346</td>
<td>0.0157</td>
</tr>
</tbody>
</table>

For KNN, we choose 5 as K base on the ROC of the model.

The confusion Matrix of testing dataset gain as above with accuracy rate as 76%.

Random Forest

When mtry = 4, we get the minimal error rate, so we implement this K value into the model. Then we get the top 5 significant variables in this model. The accuracy rate for Random Forest is 94%.

COMPARISON OF 5 MODELS

- **Logistics Regression**

  | Coefficient | Standard Error | z value | Pr(>|z|) |
  |-------------|----------------|---------|----------|
  | Intercept   | 0.7423         | 0.2087  | 3.557    |
  | account     | 0.6732         | 0.4306  | 1.569    |
  | voicemail   | 0.2325         | 0.1384  | 1.569    |
  | churn rate  | 0.1712         | 0.1346  | 1.569    |
  | rain        | 0.1346         | 0.0157  | 8.712    |

  After performing a stepwise regression, we get a model with higher accuracy, and the confusion matrix shown as above with accuracy rate as 76%.

- **Decision Tree**

  The confusion Matrix shown as above with accuracy rate as 77%.

- **Forest**

  The confusion Matrix shown as above with accuracy rate as 87%.

FUTURE WORK & RECOMMENDATION

1. With top 5 significant variables that affect customers’ decisions the most, company can use other algorithms to figure out how to adjust their service plan to better compensate customers needs and decrease churn rate.
2. When the model was put into practical use in real world, company can add more variables based on the actual situation for more accurate & suitable optimization results.

http://www.stevens.edu/bia
Introduction: Problem:
We would like to identify employees that abuse sick leave while also using the available demographic information to identify factors that could explain and predict problematic employees. We will also recommend company policies that could address some of these factors that influence absenteeism.

The database was created with records of absenteeism at a courier company in Brazil from July 2007 to July 2010.

Data Understanding:

Key independent variables used in our model includes:
Distance from work, Transportation Expense, Age, Education, # of Children, Social drinker, Social smoker, # of Pets, Body Mass Index.

Data Preparation:

Being the goal of the study the identification of problematic employees, we created a categorical variable for classification. Using the average and a half standard deviation of missing hours per disease as reference points, we identified three levels: Above Average, Average, and Below Average. The ensuing categorical analysis was aimed at single out only the critical ones. Our allocation can be seen in Figure 1.

Modeling:

While we used the K-means classification technique to confirm our initial classification, we did not use it to remove outliers because we needed it for our principal component analysis (PCA).

We are using predictive analysis by applying techniques like PCA & factor analysis to predict problematic employees.

Results & Discussion:

Our PCAs seemed to indicate there was some clustering of employees that took a below average number of hours from work. As employees distance themselves from this centroid, they become more likely to be either average or above average. This is visualized in Figure 3.

One use of this information would be to aim to hire employees that are closer to the centroid as they would have a higher chance of being better employees. The specific allocation of the PCs can be seen in the chart below. The first four PCs explain 80% of the overall variable.

Conclusion:

We identified four overall factors that seem to influence absenteeism. These are: education level, the difficulty of getting to work, social engagement/responsibility, and physical constitution. These factors can be seen in Figure 2 below.

To refine this analysis, we would recommend collecting additional reference points for the average hours absent by type of disease.

Finally, we would find that difficulty of getting to work and physical constitution are the 2 factors that have the greatest influence on absenteeism. Policies/practices to positively influence these should reduce absenteeism.
Reddit user's top ten attentions about world news
Authors: YuHong, Yuyang Tian, Mingjun Han, Ran Yi
Instructor: Rong Liu

Introduction
- An analysis about the top 10 topics in reddit world-news community.
- Reddit users' attitude towards different events happened in the world.
- Different topics and post titles will trigger different comments. By this way we can predict what kind of news will receive negative or positive responses!

Data processing
- Scraped 10,000 posts from reddit created during October 2018
- Clustering each post and classify them into different topics.
- Sentiment analysis, calculate the positive/negative score about the comments. Get user’s attitude towards different themes.
- Data classification, compute the heat score of each post, predict what kind of report will trigger people's interests.

The proportion of topics
- This is the top ten topic in reddit world-news community

User's attitude towards the topics
- The positive/negative score represents people's preference

Most frequent words about the topics
- This figure represents the word used the most to describe the topic Trump&Putin and weed legalisation

Conclusion
- These day the most intriguing news is the Saudi's murder, over ¼ users were discussing about its last month. And reddit users are also like German/Mekel topic, their attitudes towards this topic is really active. Maybe because the reddit users share some mimic or slangs about this topic.

http://www.stevens.edu/bia
Analysis of avocado based on other data in multiple US markets
Authors: Tianyu Liu, Tianyu Yang, and Ran Yi
Advisor: Amir H Gandomi

Motivation
- Missing value is annoying for data scientists, we want to find a way to construct a model to impute missing values from the existing values.
- For this project, we intend to use all variables of avocado (such as price, total sale amount) to compute the missing values to classify the type of avocado which is important for a salesmen who wants to decide which type he/she should invest more money into.

Data pre-processing

Correlation

Change in correlation matrix

The correlations between variables in the original data are quite high. We used normalize methods to deal with the data.

Scatter plot matrix

Outliers

Results & Evaluation

Confusion Matrix

ROC curve

Conclusions
- From the models we built and the predicted results (confusion matrices and the ROC curves) we can say that linear discriminant analysis is the best model.
- At first the data of each variables are highly correlated to each other, in this condition, all methods for classification will perform bad because the multi-collinearity problem. After normalization, it gets better.

http://www.stevens.edu/bia
INTRODUCTION

Many supply chain executives recognize the importance of investing in supply chain analytics, however, it is rare for firms to have a holistic strategy in place that will allow them to achieve significant value from analytics and enable a more productive supply chain. This paper outlines a strategy for supply chain analytics.

OBJECTIVES

The contributions of this paper are three-fold:
1. It presents a spectrum for firms to self-evaluate their level of maturity in supply chain analytics and identify opportunities for improvement.
2. It outlines an approach for developing a holistic supply chain analytics strategy, and
3. It presents several recommendations to bear in mind when executing the supply chain analytics strategy as well as several considerations to evaluate the success of the strategy.

EXECUTION RECOMMENDATIONS

- Align analytics investments with prioritized capabilities in order to realize the greatest impact
- Leverage the right data at the right time in the right way by giving careful focus to how the data is used, stored and shared
- Test new technologies and focus on the user experience by following an agile methodology and training the end users
- Build a flexible operating model that balances centralization with keeping the analysis close to the decision maker
- Focus on adoption by developing a governance structure, communication approach, and interactive experience

EVALUATING SUCCESS

Several questions to evaluate level of implementation success:

- Are we achieving the expected value from our analytics investments?
- Do stakeholders have a clear understanding of where our data is and how it is being used to make better decisions?
- Are we using our analytical tools for their intended purpose?
- Are our analytical capabilities effectively organized and are we appropriately managing our analytics talent?
- Are we building a culture of data driven decision making across the end to end supply chain?
Introduction

This project was entered in the Machine Learning Challenge for the 2018 UBS pitch and shared first prize. The purpose was to find three new branch locations for UBS wealth management.

UBS provided the cities and zip codes of their existing wealth management branches in the US. We collected other data and built two supervised machine learning models in the project:

The first is a common model with 7 classifiers and a stacking classifier where the zip-codes with existing branches are labeled as 1 while others 0.

The second model generated features based on the distance matrix instead of a single zip-code and combined several business objectives to the model label.

Model Approach

**Oversampling:** Stratifiedkfold / SMOTE

**Basic Classifiers:**
- Logistic Regression
- Gaussian Naive Bayes
- Decision Tree: Cross-Entropy Criterion
- Bagging: Random Forest/ Extra Tree
- Boosting: Gradient Boosting/ Adaboost

**Parameter Selection:** Least Training Error

**Evaluation:** AUC, confusion matrix

Data Collection

**Census:**
Employment/Median Income/Education/Housing/Population

**Simply Analytics:**
Expenses/ Very Rich People/ Health care and medical/ Travel agency/ Weather/ Financial Banking services

**Web scraper:**
Zip-code location/ Competitors' information

Data Processing

**Growth Rate:** Compute 5yrs Growth Rate to indicate the future development of that area

**Normalization:** Using population and number of households and Min-Max

Exploratory Data Analysis

**Feature Selection**

**Correlation Matrix:** removing high correlated features

**PCA:** reducing the dimensions

Major Findings

Generated the most important features in different models

- For the model using original data, the most important features for UBS to locate their wealth management branches is the number of competitors and advisors.
- According to relabeled data, population and wealth condition of that region are more important.

http://www.stevens.edu/bia
Predicting Overall Health from Behavioral Risk Factor Surveillance Survey Data

Authors: Malik Mubeen and Erika Deckter
Instructor: Amir H. Gandomi

Problem Statement

• The Centers for Disease Control and Prevention performs an annual health survey via telephone.
• Data from the 2015 survey conducted in New York State have been used to build a prediction model to determine overall health of the survey respondents.
• Data Reference: https://health.data.ny.gov/Health/Behavioral-Risk-Factor-Surveillance-Survey-2015/rcr8-b3jj
• The data contain 12,357 survey responses.
• Overall health is classified as either “Poor or Fair Health” or “Good or Better Health.”

Data Preparation

• Survey responses without an overall health class value were removed from the data (67 rows).
• In general, unknown values for each variable were replaced with the most common value for each response.
• Continuous variables were capped at a reasonable maximum value, and missing values were replaced by the mean.
• The data set was divided into training data (75%) and test data (25%).

Classification Model Comparison

<table>
<thead>
<tr>
<th>Classification</th>
<th>Logistic Regression</th>
<th>Linear Discriminant Analysis</th>
<th>Naive Bayes</th>
<th>K-Nearest Neighbors (K=6)</th>
<th>K-Nearest Neighbors (K=15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fair or Poor Health Accuracy Rate</td>
<td>72.6%</td>
<td>71.7%</td>
<td>71.7%</td>
<td>69.6%</td>
<td>65.4%</td>
</tr>
<tr>
<td>Good or Better Health Accuracy Rate</td>
<td>77.7%</td>
<td>78.2%</td>
<td>78.2%</td>
<td>66.7%</td>
<td>74.3%</td>
</tr>
</tbody>
</table>

Overall Accuracy Rate

Using Kaiser’s Rule, dimensions can be reduced to 38 principal components (versus 77 original variables).
Approximately a 50% reduction.

Results & Evaluation

Sensitivity of Model Accuracy to K for K-Nearest Neighbor Model

Sensitivity vs. Specificity for Various Classification Models

Principal Component Analysis

Using Kaiser’s Rule, dimensions can be reduced to 38 principal components (versus 77 original variables).
Approximately a 50% reduction.

The first 38 principal components explain 71% of the total variance.

http://www.stevens.edu/bia
Text Mining 10-K Filling Report: Predicting Financial Distress Using Risk Factors
Authors: Siwei Wang, Haochen Liu, Yuzhen He, Yiru Yang
Instructor: Professor Rong Liu

Introduction
Motivation: Financial distress indicates a probability a company goes to bankruptcy. Financial distress can be unveiled from several warning signs of company’s financial performance, such as poor profits, negative cash flow, declining relationship with the bank, etc. The "Risk Factor Section" in the 10-K annual report includes company’s explanation of the risks it faces, which contains information regarding future firm fundamentals that is not captured by the quantitative information.
Objective: Improve financial distress (edf) prediction based on risk factors in 10-K filling report from 2012 to 2016.
Key Words: Text Mining, Natural Language Process, Topic Modelling, Predictive Modeling, Python

Project Pipeline
Dataset: 12,103 observations
Yearly Range: 2012 - 2016
Variables:
- gvkey: Unique key for each company
- year: Year for each 10-K report
- edf: Financial distress score
- actual_10k_path_raw: Path to documents
- risk_factor_text: Scraped ‘Risk Factor’

Data Preparation
- Text Mining
- Data Cleaning
- Structuring Cleaning
- Text: Numpy, OS, Regular Expression

Natural Language Processing
- Basic statistics analysis
- Text analysis
- Sentiment analysis
- Topic Modeling
- Model: Random, Naive, Decision, xgboost

Modeling
- Training: Using previous year’s data to train model
- Model: Random forest

Prediction
- Regression
- Classification
- Random forest

Conclusion & Future Work
Conclusion
- Topic model (15 topics):
  1. Downstream Risks
  2. Insurance Risks
  3. Company Operation Risks
  4. Insuficient Resource Risk
  5. International Operation Risks
  6. Company Operation Risks
  7. Product Design Risks
  8. Commodity Price Risks
  9. Healthcare Spending Risks
  10. Project Loan Risks
  11. Shareholder’s Interest Risks
  12. Input Prices Risks
  13. Infrastructure Disruption Risks
  14. Regualtions Risks
  15. Investment Risks

- After studied the correlation between negative words and edf, we found that generic sentiment analysis using word lists does not add predictive power according to the nature of “Risk Factor” section.
- The meaningful results from sentiment analysis is that we observed new words showed up over the years.

Limitations and future work
- Medium correlation coefficient among texts and edf.
- The correlation coefficient between sentiment results and edf reveals that only text from ‘Risk Factor’ is not sufficient to support edf. In the future, the research will add both textual and numerical dimensions to explain edf.
- For textual dimension, industry of companies will be considered. For numerical dimensions, the research will consider adding the text feature we got to the original Moody’s model to investigate whether we can improve the prediction accuracy of financial distress.

http://www.stevens.edu/bia
The Business Problem
There are tons of ships navigating the ocean every day, but few can be detected by satellites or other means quickly and accurately. How can we automate detection and quantification of ships found in satellite imagery, and then make these results easy to access?

Data and Scope
We used segments of satellite imagery submitted to the public by Airbus for use in a Kaggle classification competition. This dataset contains a database of more than 200,000 small images of tankers, commercial ships, or fishing ships.

- Sample Images:

Ship Existence Rate
- 0: Ships do not exist in the images
- 1: Ships exist in the images
- Existing Rate is around 35%

Methodology Approach (Concept)
CNNs consist of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of convolutional layers, pooling layers, fully connected layers, and normalization layers.

Methodology Approach (Model)
Convolutional layers apply a convolution operation to the input, passing the result to the next layer. The convolution emulates the response of an individual neuron to visual stimuli. Each neuron processes data only for its receptive field.

Conclusion
The test accuracy is more than 80% which is excellent. We interface this with a chatbot made in IBM Watson Analytics Studio so that certain routes or areas can be investigated through asking a chatbot.
**Background**

A great number of biomaterials have been strongly developed in many fields of academic research. For academic publishing companies like Weily, the prediction of topics trends can make contribution to the marketing strategy. Our objective is to study the topic trends in biomaterial research and investigate effective features which possibly indicate the emerging and shrinking of topics. Especially, the influence of social medias is discovered.

**Topics Extraction Process**

**Data Source:** Web of Science Database  
**Search Terms:** "Biomaterials" or "Biomedical Materials"  
**Record Count:** 43480  
**Timeline:** 1972-2018

<table>
<thead>
<tr>
<th>Article Keywords</th>
<th>Unification</th>
<th>TF-IDF Filtering</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>LDA Model</td>
<td>Topics</td>
<td></td>
</tr>
</tbody>
</table>

**General Overview**

- Density visualization reveals the hottest topics are regeneration, microstructure, alloy and glycol.
- Topics like bone, hydroxypatite, polymer dominated the field before 2010
- Bone, hydrogel, polymer, chitosan and tissue engineering have stayed on the top 10 topics list for more than 10 years
- Hydrogel, scaffold and nanoparticles have emerged as rising hot topics since around 2009

**Emerging and Shrinking Analysis**

- The emerging trend is discerned with the positive coefficient of the linear fit, while the shrinking trend is correlated with the negative one.
- The emerging topics includes tissue engineering, scaffold, hydrogel, chitosan, regeneration, microstructure, nanoparticles, while the topics of polymer, bone and adhesion are shrinking.

**Feature Analysis - Journal Impact Factor**

- Topics like Regeneration, Tissue, Biomaterials and Scaffolds, Tissue, Bone etc have high Journal impact factor, which are also emerging topics as analysed before.
- Topics like Microscopy, Laser, dna, beta and strength etc have very low Journal impact factor which are also shrinking topics.

On those journals with very high impact factors, the emerging group has 17 documents published while the shrinking group has only 7 during the early years.

**Feature Analysis - Platform**

Firstly, we focus on articles that published on Journal and Open Access. Using LDA model, the 24 topics were generated. Most trends of topics are consistent with each other. However, there are some exceptions differ at certain periods.

We also compared the topic trend of top 10 topic words among Journal, Open Access on Web of Science and articles on Google Scholar. For all top 10 topic words, the topic trend extremely differs between journals and open access, and the topic trend of Google Scholar stays flat before 2015, and then increases abruptly after 2015 for all 10 topics. Thus the social media like Google Scholar may not be a good indicator to discern the topic trends.

**Conclusion and Future Work**

- We successively verified the previous teams’ work on topic trend analysis in terms of two different method on generating topics.
- Both TF-IDF and LDA models affect the topic classification. Also the update of database have a great influence on topic determination.
- The weight of topics in review portions has a positive correlation with the emerging trend of topics.
- The Journal Impact factor is a strong indicator of emerging & shrinking trend.
- Type of platforms for which articles were published also plays an important role in topic trends.
- In future we will ascertain whether the social media like blogs impact the topic trends and structured a predictive model based on all effective features.
Prediction of Black Friday Sale using Machine Learning

Authors: Erdong Xia, He Li, Wenlei Feng
Instructor: Dragos Bozdog

Business Problem

- A retailer wants to predict Black Friday sale including customers’ consumption level and product categories based on transactions record which contains 500k observations and 12 variables related to customers’ profile.

Exploratory of Data Analysis

- The proportion of male customers versus female customers is nearly 1:4.
- Customer aging from 18-45, unmarried who live in City Chave the highest willingness to purchase.

Methodology

- Used machine learning measures including Decision Trees, Random Forest and Support Vector Machine (SVM) to predict customers’ purchasing level (Gold, Silver, Premium & Low).
- Used Random Forest algorithm to examine the importance of customers’ factor related with product category, and to predict customer’s willingness to buy for each product category.
- Used Apriority algorithm to find the correlations between category variables and serve as the basis for the Recommendation System.

Results

- After different machine learning measures were applied in predicting customers' purchasing power, Random Forest provided us the best model result with 60% accuracy.
- 3 categories (C7, C10 and C12) of products with a relative low purchase rate 28%, 55% and 63% were applied in predicting customers' purchase intent to a specific type of product.

<table>
<thead>
<tr>
<th>Category</th>
<th>Purchase Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>C7</td>
<td>28%</td>
</tr>
<tr>
<td>C10</td>
<td>55%</td>
</tr>
<tr>
<td>C12</td>
<td>63%</td>
</tr>
</tbody>
</table>

- City is the most important variable to predict purchase intent. Marital status, Age and Occupation are also helpful in specific prediction model.
- The accuracy of predicting customer's purchase willingness to different categories of product is around 70%.
- 41 groups of product collections that customers are most likely to buy together have been found based on purchasing frequency record.

<table>
<thead>
<tr>
<th>Category</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 1</td>
<td>1,5,8,14</td>
</tr>
<tr>
<td>Top 2</td>
<td>5,8,14,16</td>
</tr>
<tr>
<td>Top 3</td>
<td>1,8,14</td>
</tr>
<tr>
<td>Top 4</td>
<td>1,2,5,8</td>
</tr>
<tr>
<td>Top 5</td>
<td>8,14,16</td>
</tr>
<tr>
<td>Top 6</td>
<td>2,5,8</td>
</tr>
<tr>
<td>Top 7</td>
<td>1,8</td>
</tr>
<tr>
<td>Top 8</td>
<td>1,8,16</td>
</tr>
<tr>
<td>Top 9</td>
<td>5,8,16</td>
</tr>
</tbody>
</table>

Conclusion

- To boost Black Friday sale, customers with beneficial features can be assigned more marketing budgets and provided with product collection offers in accordance with our result.

http://www.stevens.edu/bia
Who are the most important authors in the Biomaterial Research?

Authors: Minzhe Huang, Shuo Jin, Jiaqiang Lu, Raj Mehta, Jingmiao Shen
Instructor: Christopher Asakiewicz; Sponsored: John Wiley & Sons

Motivation
• Academic prosperity brings us tons of authors and papers. For a company like John Wiley & Sons, it is beneficial to predict the most valuable authors in the next few years so that the industry can switch the attention to those potential leaders.
• Besides this objective, we also focus on analyzing the relationship within the "Author Citation Network", "Number of Publications" and other factors to find out the most related features to determine the most important author.

Methodology
Key Word: BioMaterial
Year Range: 1982 ~ 2018
Tool: Python + VOSviewer + CitNetExplorer:
Model: Xgboost

Approach
1. Write script to automatically download citation info from Web of Science and process Data Cleaning
2. Process Feature Correlation Test
3. Use VOSviewer and CitNetExplorer to process both the Overall Citation Network and Yearly Citation Network
4. Fetch data from SemanticScholar and do Data Cleaning, including “Influential citation Count” and “Citation Velocity”
5. Conduct Feature Engineering, including the below:
   • Number of Publications
   • Publication Frequency
   • Career Length
   • Influential Citation Count
   • Citation Velocity
5. Label the training data if “InfluentialCitationCount” > 100
6. Label the test data if “CitationVelocity” > 100
7. Use XGBoost to fit the data
8. Evaluate the model by Score Matrix
9. Get our 5 Most Important Authors
10. Get our 5 Least Important Authors

Insight
Feature Correlation Test

Conclusion
5 Most Important Authors (By Score)

<table>
<thead>
<tr>
<th>Devendrapand Santhana Panneer</th>
<th>Sylvia G Simpson</th>
<th>Lu Wang</th>
<th>Chien Hung Li</th>
<th>Yan Li</th>
</tr>
</thead>
<tbody>
<tr>
<td>789.598</td>
<td>788.951</td>
<td>788.670</td>
<td>788.413</td>
<td>788.402</td>
</tr>
</tbody>
</table>

5 Least Important Authors (By Score)

<table>
<thead>
<tr>
<th>T.W. Forest</th>
<th>Yufang Zhu</th>
<th>Jeonghun Kim</th>
<th>Nestor Schor</th>
<th>Rod H. Smallwood</th>
</tr>
</thead>
<tbody>
<tr>
<td>238.00</td>
<td>239.32</td>
<td>239.44</td>
<td>239.44</td>
<td>239.56</td>
</tr>
</tbody>
</table>
Fraud Detection for Credit Card Transactions
Authors: Raphael Presberg, Niraj Chaursasia, Medhavi Uniyal
Instructor: Dr. Christopher Asakiewicz

Introduction
Fraud detection has become one of the most critical challenges for Companies. For this project, we help a credit cards company by detecting suspicious credit Card transactions.

Business Question
How to detect, alert and prevent fraudulent credit card transactions?

Data Set Discovery
- 284 807 transactions
- Target variable 0 or 1 (if fraud)
- Event Rate: 0.17%

Event rate is representing the ratio of positive instance in a dataset

Re-Sampling Techniques
- Under Sampling: create a sample of the non fraudulent transactions
  Re-sampled training DataSet:
  • 329 non fraudulent transactions
  • 329 fraudulent transactions
- Cluster Based: K-means clustering independently applied to minority and majority class instances
  Re-sampled training DataSet:
  • 502 non fraudulent transactions
  • 502 fraudulent transactions

Results
We have performed several classification algorithms to detect the fraudulent transactions. The following results have been obtained on an untouched testing Dataset containing 163 positive instances.

Under Sampling
- RandomForest
- KNN
- XGBoost

K-Means Sampling
- RandomForest
- KNN
- XGBoost

Business Decision
Based on the previous result, we have an excellent outcome in detecting fraud with my XGBoost algorithm trained on re sampled data
I would then implement this model into my IT ecosystem to detect fraudulent credit cards transaction in real time.

Technical Challenges
- Handle a highly unbalanced dataset
- Find & Develop the fitting Machine Learning Algorithm to avoid the metric trap

Conclusion
Working on fraud detection was an exciting challenge and integrating our model into and IT ecosystem would be a fantastic opportunity

http://www.stevens.edu/bia
What Makes A Good TED Talk?
Authors: Pranav Prajapati, Sonali Johari, Rumeng Zuo, Qian Lu
Instructor: Feng Mai

Introduction
Motivation:
- Deduce the elemental reason why TED talks are considered the benchmark for influential speeches. What makes them special?
- Sentiment analysis and emotional introspection of over 2500 TED talks conveys the uniqueness of the expert speakers and the trends in their speeches, giving an insight into the psyche of successful minds.

Key methods: Sentiment analysis, network analysis, text mining, content similarity, topic modelling, LDA, K-means clustering.

Speakers and Themes
Interestingly, it was observed that while technology was the most popular theme of the TED talks, Writers were the most dominant when it came to popular professions.

Power Words
Based on the number of views and comments of each TED talk, the Top and Bottom 500 talks were selected. The histogram describes the power words of the Top 500 talks, while the word cloud compares the power words of the Top and Bottom 500 Talks.

Methodology
- Sentiment analysis using bag of words technique and a lexicon based approach has been used.
- Network analysis by content similarity analysis.
- Topic modelling between three different talks on women empowerment (TED, UN & Political speech) was carried out to gain insight on the differences of power words based on the intent of deliverance.

NRClexicon labels words across multiple emotional states. The NRC lexicon tags words.

Plutchik’s emotions
Since TED talks connect with the audience, they have high trust count.

The BING lexicon categorizes words in a binary fashion into positive and negative categories. Comparing the timeline of sentiments of top and the bottom talks, the top talks end more optimistically.

The AFINN lexicon is a list of English terms manually rated for valence with an integer between -5 (negative) and +5 (positive). It exhibits how TED speakers use a storytelling approach, sharing with the viewers their journey of overcoming failures and achieving success.

Recommendation of similar speakers
By calculating cosine similarity between artists, a network of related speakers was created. The example here shows a subgraph of the result of recommended speakers when the user is interested in talks similar to Elon Musk.

Topic Modelling
Unsupervised learning methods like LDA & K-means clustering depict the top terms for 3 topics for different speeches on women empowerment.

We can see that Topic 1 depicts a TED talk. The K-means does not seem to perform as well as LDA. It couldn’t distinguish between the UN and TED talks well enough. Both methods can be tuned for better performance.

Conclusion
While TED talks are special because of the stories they convey, performing sentiment analysis on them generated some interesting results on the similarities of these talks.
- We can analyze that although the TED talks are of progressiveness, they mostly revolve around wisdom, women, family and passion.
- The greatest quality a TED speaker possesses is the ability to gain the trust of the audience. Also, creating an atmosphere of anticipation while storytelling is characteristic trait of great TED speakers.

http://www.stevens.edu/bia
Optimizing London Fire Station Resources to Better Serve the Community

Authors: Sonali Johari, Pranav Prajapati, David McFarland, Erika Deckter and Marielle Nwana
Instructor: Ted Stohr

Motivation
By simulating real-world emergency scenarios, fire station resources can be efficiently deployed to each incident while minimizing overall travel distance for the fire engines. This analysis also showed the impact of adding additional resources to existing fire stations in order to better serve the community.

Data
Using data provided by the London Fire Brigade as well as information from Kaggle, we were able to obtain a historical database for over 85,000 fire incidents for 2017 (from January to October).

Key Methods
Integer Programming Optimization, Simulation, Great Circle distance and Sensitivity Analysis

Optimization Model

Inputs
- Distance Matrix, D
- Delay Factor Matrix, F
- \( f_i \) = randomly generated factor (between 0 and 1) to simulate arrival delays
- Effective Distance Matrix, E
- Availability Vector, A
- \( d_{ij} \) = number of fire engines available at \( j \)th station

Decision Variable
- \( s_{ij} \) = if fire engine is dispatched to incident \( i \) from station \( j \)
- \( Q_i \) = if fire engine is not sent to incident \( i \) from station \( j \)

Constraints
- \( Q_i = 1 \), one fire engine is dispatched to each incident
- \( \sum_j s_{ij} \leq \alpha \), the total number of fire engines dispatched from a station cannot exceed the available number

Output
- Minimize the Total Effective Distance
- \( \sum_i \sum_j s_{ij} d_{ij} \)

Results
Sample Result Output for Select Simulation Periods

<table>
<thead>
<tr>
<th>Period</th>
<th>Incidents</th>
<th>Engine</th>
<th>Euston</th>
<th>Holloway</th>
<th>Islington</th>
<th>Kentish Town</th>
<th>Paddington</th>
<th>Soho</th>
<th>West Hampstead</th>
<th>Total Fire Engines</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>9</td>
</tr>
</tbody>
</table>

Sensitivity Analysis

Fire Engine Availability

<table>
<thead>
<tr>
<th>Engine</th>
<th>Dowgate</th>
<th>Euston</th>
<th>Holloway</th>
<th>Islington</th>
<th>Kentish Town</th>
<th>Paddington</th>
<th>Soho</th>
<th>West Hampstead</th>
<th>Total Fire Engines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Model</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Sensitivity +1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>15</td>
</tr>
</tbody>
</table>

Sensitivity analysis was performed using actual fire engine counts from London Fire Brigade’s fleet list (as of September 2017).

Simulation

- 9,600 simulated time periods (15-minute intervals over 100 days)
- Azero-truncated Poisson distribution was used to determine the number of incidents in each simulation period
- Incidents for each time period were selected using a random draw of a subset of the historical data
- The Integer Programming (IP) optimization model was applied to each simulation period
- The model assumed fire engines are deployed at the end of each 15-minute period and do not return for 30 minutes (i.e., a fire engine deployed in the previous two simulation periods cannot be used in the current period

http://www.stevens.edu/bia
Stack Watson: The Friendly S.O. Bot
Authors: Smit Mehta, Xue (Calvin) Cao
Advisor: Prof. Chris Asakiewicz

Stack Overflow Bot

- Provide real-time help to programmers on trivial programming issues using the wealth of information already existing on the Stack Overflow website
- This will also help in reducing the workload on administrators that have to monitor incoming questions for duplicates

<table>
<thead>
<tr>
<th>Stack Watson</th>
<th>Stack Watson</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi there! How can I help you today?</td>
<td>Hi there! How can I help you today?</td>
</tr>
<tr>
<td>How do I select the optimum Learning Rate?</td>
<td>Where can I learn the basics of Deep Learning?</td>
</tr>
<tr>
<td>Here is what I found...</td>
<td>I did not find anything substantial. Please go ahead and post your question on the website</td>
</tr>
<tr>
<td>If you use a convex loss function you always have one optimum point, and you will always be able to find it. I have done some calculations...</td>
<td></td>
</tr>
<tr>
<td><a href="https://stackoverflow.com/questions/ask">Link</a></td>
<td></td>
</tr>
</tbody>
</table>

- “Stack Exchange Data Dumps” by Stack Exchange, Inc. via archive.org; specifically the Data Science Stack Exchange
- Due to the high volume of data and limited resources, we will limit the scope of this project to a particular topic area
- The data consists of all the questions with the following tags: <machine-learning> and <neural-network>

Architecture

1. The interface between a user and Stack Watson, facilitated by Watson Assistant.
   - User will ask “natural language queries” which WA will pass to the AI agent
   - Watson Assistant will present the answer with the highest confidence level
   - If an answer is not found, it will prompt the user to post it on SO website
2. Topic Modeling is used to categorize the question into different tags to ensure relevance
3. Watson Discovery Service interacts with the Knowledge Base (“KB”) to return relevant answers. The threshold for relevancy can be set by us depending on the level of training provided to the application
4. KB is regularly updated with new questions being added to the SO website through automated extraction from the SEDE data dumps

Impact & Future Scope

- When someone posts a question on Stack Overflow, they have to wait for some time before they can get an answer (sometimes it’s even days!)
- A chatbot for trivial questions would eliminate the time lag and make the programmer more productive
- This application would also save the subject matter expert time and as a result focus on more pressing and important matters

**Future Scope:**
- Scale it up with more training data and also include other stack exchange websites content
- Adding additional features such as checking the quality of the questions being asked (another time saving option for mods)
- Return links to additional info and sample code (if available) by connecting to a central repository

http://www.stevens.edu/bia
The Portfolio Rebalancing Problem

- Portfolios, like their underlying assets, have risk and return characteristics that naturally evolve over time with the market.
- Rebalancing helps investors successfully navigate a portfolio across market regimes given a particular risk/return based objective.
- The objective of the portfolio rebalancing problem is to make a decision at each point in time to rebalance or not while minimizing costs sustained by the portfolio.
- Reinforcement learning provides the ideal modelling and optimal solution framework to a problem commonly solved by heuristics in the investment management industry.

The Model

- Given a portfolio of \( N \) assets with portfolio weights \( w^t = [w_{1t}, \ldots, w_{Nt}]^T \), our goal is to maintain a portfolio that tracks the target portfolio as closely as possible while minimizing transaction costs.
- The portfolio can be rebalanced every month.
- Normal returns are assumed:
  \[
  w_{t+1} = (1 + \eta_t)(w_t + u_t), \quad \text{where } \eta_t \sim N(\mu, \sigma)
  \]
- The objective to be minimized is the sum of (i) tracking error, (ii) transaction costs, and (iii) expected future costs.

Methodology: Reinforcement Learning

Algorithm 1: Calculate \( J_t(w^t) \), \( w^t \in W, t \in \{0, 1, \ldots, T\} \)

Let \( w_{init} \in W \) be the initial allocation

\[
T = 240, \quad \gamma = 0.9
\]

\[
J_T(w_T) = 0, \forall w_T \in W
\]

for \( t = T - 1 \) to 0 do

\[
J_t(w_t) = \infty, \forall w_t \in W
\]

for \( i = 1 \) to \( |W| \) do

\[
J_t(w_i) = \sum_{w' \in W} \mathbb{P}(w'|w_i) \times [G(w_i, u_t, \eta_t) + \gamma J_{t+1}(w')]
\]

end for

end for

\[
J_0^*(w^*) = \min_{w \in W} J_0(w)
\]

\[
w_0^* = w^* - w_{init}
\]

return \( w_0^* \)

Computational Results

- Q-Learning can deal with the curse of dimensionality as the number of assets \( N \) grows.
- Actions: Rebalancing Decisions \( (u_t) \), State Variable: Portfolio Allocation \( (w_t) \).
- Stage Cost:
  \[
  \langle w_t, u_t, \eta_t \rangle = \langle w_{t+1}, w_{t+1} \rangle + \epsilon(w_t, w_{t+1})
  \]

  \[
  \epsilon(w_t, w_{t+1}) = \text{Transaction Costs} + \text{Tracking Error}
  \]

  State Variable Costs

Conclusion

- Reinforcement learning provides an objective \textit{optimal} decision as a solution while heuristic methods provide ad-hoc sub-optimal decisions.
- The solution model can be flexibly adapted to meet an investor’s unique constraints.
- Future research should look to implement the reinforcement learning model with alternative assumptions to the normal multiplicative dynamic model as well as higher dimensional portfolios.

References


Global Burden of Tuberculosis, 1990-2013
Authors: Mingrui Wang, Wei Yang, Hefen He, Sicheng Zhang, Huiting Fang, Haiping Sun, Siqi Zhang
Instructor: Alkiviadis Vazacopoulos

Introduction
We managed to scrape data related to Tuberculosis and some related information (1990 - 2013) from Internet. The dataset contains:
- 47 columns and 5121 rows;
- Estimated prevalence of Tuberculosis and estimated mortality of Tuberculosis (HIV positive and negative all included); All estimated data have ratio about per 100,000 population;
- Total population and country geo-location (continent included) data;

Geographic Information
To find out which area has the most serious incident rate of Tuberculosis. We could see the South Africa and Southeast Asia is under the burden of Tuberculosis.

Total Incident Number Analysis
To find out which area has the biggest number of total incident. Using the dashboard below is easy to compare all the regions.

Relation Analysis
To find out the relation between HIV and mortality of TB.

Time Analysis
We create a time series map to find out which year is the worst year of Tuberculosis control.

Analysis of Africa incidents
We concluded that Africa is the worst area of controlling TB, so we did more detailed (by countries) analysis about this region.

http://www.stevens.edu/bia
Can We Predict Wine Quality with Machine Learning?
Authors: Patrick Curran & Smit Raval
Instructor: Alkiviadis Vazacopoulos

Introduction
Machine learning is an component of artificial intelligence where a computer is programmed with the ability to self-teach and improve its performance of a specific task. It’s fundamentally changing the way we live our lives, with applications ranging from healthcare to transportation.

Another possible application for machine learning is predicting human taste. Companies like Netflix and Google use machine learning to generate suggestions for what should be your next click. However, machine learning is not as commonly used in predicting how someone will grade the quality of food and drink. This project explores the possibility of using machine learning to accurately predict the quality of red wine.

Experiment
This experiment follows the “Machine Learning Process”, shown below:

![The Machine Learning Process Diagram]

The first steps were to gather and clean the data. The group acquired the following characteristics for 1,600 different types red wines: fixed acidity, volatile acidity, citric acid content, residual sugars, chlorides, sulfur dioxide content, density, alcohol content, and quality rating. Because we are attempting to create an algorithm to predict quality rating, this dataset becomes what is known as supervised data. Data becomes supervised when the information for the desired outcome is known. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used to make future predictions.

Results
RapidMiner gave a specific prediction for all 1,600 red wines in the dataset, and the results were checked for accuracy. Quality was measured on a 3 to 8 scale in the dataset.

| Average Points Off By | 0.1595 |
| Correct Quality Guess | 60.35% |
| Correct Guess +/- 1 Point | 97.37% |
| Correct Guess +/- 2 Points | 99.94% |

This test concluded that the most accurate machine learning method for predicting red wine quality was deep learning. A code was then setup to test the machine’s algorithm against the group’s set of data.

The next step was deciding which machine learning process should be used in creating a prediction algorithm. Using a powerful program called RapidMiner, the data was analyzed for which process best predicts wine quality.

![Accuracy Chart]

http://www.stevens.edu/bia
Opinion mining: Tracking Public Emotions on Twitter

Author: Dhaval Sawlani
Instructor: Ted Lappas

Introduction

The outbreak of the internet and social network presents a new set of challenges and opportunities in the way information is searched, retrieved and interpreted.

Opinions expressed on blogs and social networks are playing an important role in influencing everything from the products people buy to which US president should you support.

Thus there is a need for an application which will not only retrieves facts, but also enable the retrieval of opinions. Such an application can be used to understand user-product relationship in a more profound manner and can also help to aggregate opinions on political candidate or issues with more consistency.

Application architecture

Process

1. Login to the web app on https://35.231.96.132:5006
2. User inputs hashtags or search terms on the Search bar of the application
3. With the help of Parallel processing and multi-threading techniques, we scrape 4x times more Twitter data in half the time as compared to the native TwitterAPI
4. Preform data pre-processing obtaining cleaner version of the Tweets by stemming and removing stop words and punctuations
5. Extract the Emoji from the Tweets; Emoji helps us understand the context of the Tweet as people use it to convey emotions on a very large scale
6. Perform Sentiment Analysis, using sci-kit learn and extract the Percentage Positive, Negative and Neutral sentiments of Tweets
7. Perform Topic-modeling, an NLP technique used to understand the breath of the textual conversations

Results for #MAGA

Conclusion

1. Emoji Analysis gives us an in-depth idea about how people are conveying their emotions
2. Word cloud summarizes the most frequent words used; helping to understand the most popular words related to #MAGA
3. Sentiment Polarity concludes the % outreach of Positive, Negative and Neutral Sentiments on Twitter for #MAGA
4. Emotion Radar breaks down 8 major human emotions into % distribution; 29.05% tweets have Joyous emotions associated

http://www.stevens.edu/bia
Predicting customer churn rate for a bank using logistic regression
Authors: Ameya Swar, Rashmi Khurana, Rushabh Vakharia
Instructor: Prof. Khasha Dehnad

Introduction:
Customer churn has become one of the top issues for most banks. It costs significantly more to acquire new customers than it costs to retain existing ones, and it costs far more to re-acquire defected customers. In fact, several empirical studies and models have proven that churn remains one of the biggest destructors of enterprise value for banks and other customer intensive companies. For our project, we have the data for a bank which has its branches in 3 different European countries. There are a lot of factors that customers use to consider another alternative. In our project, some of the factors that we have considered to predict churn are age, gender, salary, balance, geography, is the customer an active member of the bank or not, does the customer have a credit card with the bank or not, customer’s tenure, etc. Using the provided information, we have used logistic regression to predict which customers are most likely to exit the bank in the near future.

Experiment:
Dataset → Cleaning → EDA → Conclusion → Logistic Regression

Results:
After performing logistic regression, we see that out of the initial 12 factors, the following 7 factors tend to have a greater impact on customers deciding to leave or exit the bank.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-3.9760</td>
<td>0.2312</td>
<td>296.8376</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>CreditScore</td>
<td>1</td>
<td>0.00666</td>
<td>0.00280</td>
<td>5.6501</td>
<td>0.0175</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>-0.0727</td>
<td>0.0257</td>
<td>797.3454</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Balance</td>
<td>1</td>
<td>-2.65E-6</td>
<td>5.139E-7</td>
<td>26.8299</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>NumOfProducts</td>
<td>1</td>
<td>0.1010</td>
<td>0.0471</td>
<td>4.5985</td>
<td>0.0320</td>
</tr>
<tr>
<td>IsActiveMember</td>
<td>1</td>
<td>1.0718</td>
<td>0.0576</td>
<td>346.0764</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
<td>-0.5306</td>
<td>0.0545</td>
<td>94.8968</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Germany</td>
<td>1</td>
<td>-0.7608</td>
<td>0.0633</td>
<td>144.3322</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

When you compare the CAP (Cumulative Accuracy Profile) for the training and test data, you see that the CAP curve for test data is a little rough. This is due to the fact that the data used for testing had only 1000 records whereas the one for training had 10000 records. In spite of that, the accuracy rate as calculated using the confusion matrix is 81% for training data and 76% for the test data.

Conclusion and future work:
We can conclude that the geography, balance, age, number of products a customer has with the bank, credit score, gender and whether the customer is an active member of the bank or not play a very important role in predicting customer churn. The bank can devise strategies accordingly and reduce churning.

In the future, we plan to use a classification algorithm like Random Forest or decision trees to identify which customers have a high risk of churning. Although random forests have certain advantages against decision trees, such as resistance to overfitting and more robust results, it is important to make sure that we have a quite large number of different variables for the trees to be trained differently.

http://www.stevens.edu/bia
Quantum computers can harness quantum physical effects not available to conventional computers: **Superposition, Entanglement and Tunneling.**

**Superposition** is the ability of a quantum system to be in multiple states at the same time until it is measured. Quantum states can be added together ("superposed") and the result will be another valid quantum state; and conversely, that any quantum state can be represented as a sum of two or more other distinct states. A quantum logical qubit state, as used in quantum information processing, which is a quantum superposition of the "basis states" \( |0 \rangle \) and \( |1 \rangle \). The principle of quantum superposition states that if a physical system may be in one of many configurations - arrangements of particles or fields - then the most general state is a combination of all of these possibilities.

**Entanglement** is a quantum mechanical phenomenon in which the quantum states of two or more objects have to be described with reference to each other, even though the individual objects may be spatially separated. As a result, measurements performed on one system seem to be instantaneously influencing other systems entangled with it.

**Tunneling** is the transitioning through a classically-forbidden energy state. Consider rolling a ball up a hill. If the ball is not given enough velocity, then it will not roll over the hill. For a quantum particle moving against a potential hill, the wave function describing the particle can extend to the other side of the hill. This wave represents the probability of finding the particle in a certain location, meaning that the particle has the possibility of being detected on the other side of the hill, as if the particle has ‘dug’ through the potential hill.

**Quadratic Unconstrained Binary Optimization (QUBO)**

- Problem Formulation and QUBO
- Entangled QUBO
- 1-Persistent Entity-Variant Portfolio Optimization
- Number Encodings
- New Model For Quantum Portfolio Optimization

Quantum annealing (QA) enhances optimization heuristic exploiting superposition, tunneling, and entanglement. The D-Wave 2X (1000 qubits) quantum annealer achieves significant run-time advantages to simulated annealing (SA) and quantum Monte Carlo (QMC) running 108 times faster than running a single processor core.

**D-Wave Quantum Annealing Computer**

Quantum annealing is a generic approximate method to search for the minimum of a cost function (multivariable function to be minimized) through a control of quantum fluctuations. Quantum annealing is used for combinatorial optimization problems with discrete variables. Many practically important problems can be formulated as combinatorial optimization, including machine learning for clustering, distribution of components in factories, and route optimization in traffic.

Finding efficient methods to solve such optimization problems is of enormous social significance, which is the key reason why quantum annealing attracts much attention. Also, of current research interest are the sampling problem for machine learning.

**Quantisim Cloud Solution**

For the continuous wave QA, we can exploit the QA solution method to the "original" Hamiltonian \( H_{2Q} \) in the given system by

\[
H_{2Q} = \sum_{i=1}^{N} E_i |i\rangle \langle i| + \sum_{i < j}^{N} J_{ij} |i\rangle \langle j| + H.C.
\]

where \( |i\rangle \) is the qubit wavefunction of each qubit. Then, utilizing the QA method to the "original" Hamiltonian \( H_{2Q} \) in the given system, the state of the QA system will approach the ground state of the system. To observe the state of the QA system, the state of the system will have to be measured, which will destroy the superposition of the QA system.

**New Model For Quantum Portfolio Optimization**

**Back to Number Encodings II**

Define the following encoding metric as

\[
\|\mathbf{x}\|_2 = \sum_{i=1}^{N} x_i^2
\]

**Back to Number Encodings III**

Define the following encoding metric as

\[
\|\mathbf{x}\|_2 = \sum_{i=1}^{N} x_i^2
\]

**Q-Matrix Formulation**

Define the metric:

\[
\frac{1}{|\mathbf{x}|} \quad \frac{1}{|\mathbf{x}|} \quad \frac{1}{|\mathbf{x}|} \quad \frac{1}{|\mathbf{x}|} \quad \frac{1}{|\mathbf{x}|}
\]

and \( Q = [q_{ij}] \) for any suitable \( n \).

**Matrix Formulation of Max Risk-Averse Portfolio Optimization Problem**

**Number Encodings**

Define the following encoding metric as

\[
\|\mathbf{x}\|_2 = \sum_{i=1}^{N} x_i^2
\]

**Back to Number Encodings II**

Define the following encoding metric as

\[
\|\mathbf{x}\|_2 = \sum_{i=1}^{N} x_i^2
\]

**Q-Matrix Formulation**

Define the metric:

\[
\frac{1}{|\mathbf{x}|} \quad \frac{1}{|\mathbf{x}|} \quad \frac{1}{|\mathbf{x}|} \quad \frac{1}{|\mathbf{x}|} \quad \frac{1}{|\mathbf{x}|}
\]

and \( Q = [q_{ij}] \) for any suitable \( n \).

**Matrix Formulation of Max Risk-Averse Portfolio Optimization Problem**

**Number Encodings**

Define the following encoding metric as

\[
\|\mathbf{x}\|_2 = \sum_{i=1}^{N} x_i^2
\]

**Back to Number Encodings II**

Define the following encoding metric as

\[
\|\mathbf{x}\|_2 = \sum_{i=1}^{N} x_i^2
\]

**Q-Matrix Formulation**

Define the metric:

\[
\frac{1}{|\mathbf{x}|} \quad \frac{1}{|\mathbf{x}|} \quad \frac{1}{|\mathbf{x}|} \quad \frac{1}{|\mathbf{x}|} \quad \frac{1}{|\mathbf{x}|}
\]

and \( Q = [q_{ij}] \) for any suitable \( n \).
**INTRODUCTION**

Object detection is an essential component for autonomous driving cars. Accurate detection of vehicles, street buildings, pedestrians and road signs could assist self-driving cars the drive as safely as humans. However, conventional classification-after-localization methods are in slow in real-time situations. We need an object detection model which can detect object with high accuracy while also running in real-time.

**OBJECTIVE**

Given an image taken by a camera mounted on top of the car, our objective would be to successfully detect a car in the image and put a bounding box around the car.

**MODEL**

We used a specialized Convolutional Neural Network algorithm called **YOLO** (You only look once). It assists in real time object prediction. Solution is implemented with Python, TensorFlow and Keras.

YOLO requires a large dataset and is computationally very expensive to train. Hence, weights that have been pre-trained on Microsoft's COCO dataset have been used. MS-COCO contains 91 labelled object types in 328K images.

**Implementation Details:**

- Our model runs a pre-processed input image through a Deep CNN.
- We filter through all the boxes using non-max suppression.
- Filter out detected object classes with low probability.
- Use Intersection over Union (IoU) to get the final bounding box.
- We get output image of detected objects and corresponding bounding boxes.

**CONCLUSION**

This implementation provides a practical object recognition example that will enable autonomous applications such as self-driving cars. It allows the car to build an accurate mapping of its surrounding which will allow it to steer safely in complex surroundings. For now, it can be used to augment human driving capabilities. The object recognition task can be paired with segmentation and GPS to mark lanes, pedestrian cross-walks, etc. to develop more robust solutions.
In the U.S. the process of deregulation and the introduction of competitive electricity markets have been reshaping the power sector. Among the several sources for electricity generation (like nuclear, hydro, solar, wind, fossil fuels, etc.), 90% of all U.S. electricity generation comes from non-renewable sources (Natural Gas, Coal & Oil).

Natural Gas is the most used fuel for electricity generation and its cost is the determinant factor in the wholesale electricity price which is formed by the highest generation cost for the demanded electricity.

Determine future demand and prices are fundamental in the energy generation, transmission, and distribution for reducing costs, supplying the demand, and improve the decision-making for future investments that affect directly the residential customers and economic activity.

Natural Gas achieve an excellent forecasting result for its real value with a a very low MSE (Mean Squared Error), less than 4%. Peak forecasting, this can be improved by using the data for natural gas inventories and storage, also supply and demand.

Diesel also has an excellent forecast with a MSE less than 6%, is very important to predict the peak prices and most of the algorithms fail in forecasting a correct peak price what increase the risk for generators, transmitters, and consumers.
**Introduction**

- Dark Pools is an important area in financial markets with high-frequency trading. Unlike open stock exchanges, transactions at dark pools are operated under asymmetric information and secretive protocols.
- Since there is little transparency of trade executions, trying to find liquidity is challenging.
- We develop machine learning methods to analyze and predict patterns in liquidity of dark pools.
- Basic data structure:
  a. 3 Months data : 2 months training, 1 month validation (June-August 2017)
  b. Venues: UBSA, CAES, DBSX, KNMX, LEVL
  c. Generated features: L1, L3, L5
  d. Main Additional Features:
     - StartTime, Symbol, Venue, VenueType, SecurityCategory, Sector, MktCap, Adv20d
- Main issue: extremely high class imbalance (Trade class less than 1% total data size).

**Feature Selection & Results**

- Feature selection is the next key step in our machine learning methodology
- In this step we compare all features against each other and select an “information optimal” subset of features
- We followed a two-pronged approach to feature-selection:
  1. Automatic feature selection based on statistical methods such as: Pearson correlation, Maximal Information, Regularized Methods, Mean Decrease Impurity, Mean Decrease Accuracy, Stability Selection, and Recursive Feature Elimination.
  2. The results of 1. is pruned based on area knowledge to select a subset of features on which to focus.

**Two main methods: Naïve and NNet**

**Naïve Methods:**
- Based on generated features
- Simple to implement

**Neural Networks:**
- Ubiquitous with solid theoretical foundation
- Flexible and scalable to big-data problems
- Binary prediction: Will order will be a trade or not? Can output probabilities too
- Finds structural patterns in the data, exploits the given features, and generates new features based on these

**Conclusion**

- Good performance of Naïve methods:
  - Follows a “simplicity” trend observed in financial optimization models
  - High cost of increased precision while maintaining reasonable recall values
  - Importance of oversampling methods:
  - Essential to help NNET to focus on trades instead of orders
  - Weighted NNET via resample exhibits best performance
  - Current implementations of SMOTE methods are computationally expensive
  - No clear winners on scaling, standardization, PCA
  - Reliance on domain knowledge: Deep Learning methods promise to extract patterns directly from data
  - Bootstrap inspired and stacked methods to deal with data imbalance
  - Data volume and complexity: design a more robust data management system
  - Hardware: Limitations on computation can be overcome with cloud computing (AWS)
  - Software: Limitations on Python Scikit learn, argues for more robust tools like Google’s Tensor Flow
**INTRODUCTION**

- Data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.
- Marketing campaigns were based on phone calls.

**OBJECTIVE**

- The classification goal is to predict if the client will subscribe a term deposit.
- To perform exploratory data analysis and find the best machine learning classification algorithm to better fit the data.

**DATA ANALYSIS**

The data has 6 continuous and 11 categorical variables.

- Data is left skewed and therefore creating imbalanced dataset.
- Imbalanced dataset is treated with random under and over sampling techniques.
- The ages factor has a medium dispersion and do not seem to relate with other variables.
- Jobs, Marital status and Education are the best factors to analyze if a customer will subscribe.

**RESULTS**

Algorithm is picked using cross validation results, least False Positive rates and high precision on negative predictions. Area under curve (AUC) metric is used to measure accuracy.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>90.0</td>
</tr>
<tr>
<td>SVC</td>
<td>86.0</td>
</tr>
<tr>
<td>Random Forest</td>
<td>90.0</td>
</tr>
<tr>
<td>XGBoost</td>
<td>91.0</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td>91.0</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>90.0</td>
</tr>
</tbody>
</table>

**CONCLUSION**

- Focus should be on reducing the false positive rate rather than accuracy to advertise to all prospective customers.
- K nearest neighbors (KNN) has the least false positive number of just 111 customers.
- The months of March and December have the highest probability of getting customers to enroll.
- Duration of the call plays an important factor. Generally, calls more than a minute can convince people to subscribe to a term deposit.

[http://www.stevens.edu/bia](http://www.stevens.edu/bia)