STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Business Intelligence & Analytics

A Snapshot of Data Science

Robert Bach

November 21, 2019

POSTER BOOK
November 21, 2019
Poster Show and Recruiting Event

Page Number	Poster Title
1	Fundamentals are fundamental
2	Chicago Crime Data
3	Analysis of Transportation Speed in New York City
4	Retail Analytics and Sales Forecasting
5	WINE ME UP in the USA
6	Design of Experiments on Machine Learning Hyperparameter Tuning
7	Redefining Cancer Diagnosis
8	NYSE Stock market prediction with technical analysis using Long short-term memory
9	YouTube Data NLP Analysis
10	The Analysis and prediction based on NBA Data of players and teams
11	Predicting Airbnb Prices in Jersey City Using Machine Learning
12	Customer Lifetime Value Optimization
13	Cookbook Consumer Behavior Analysis from Amazon.com
14	Data Science Approach to Automate Granting/Revoking Employee Computer Access Needs
15	Statistical Analysis and Optimization of Service Time at a Local Cafe
16	Towards an accurate prediction of customer behavior for IBM Watson marketing data, using Deep Learning
17	Loan Approval Prediction based on Customer History
18	Movie Recommendations based on Viewer History
19	Academic Chatbot for the BI&A Program
20	Flight Scheduling Optimization for American Airlines — based on network flow model
21	Predicting Researcher Performance in the Medical Marijuana Field
22	Marketing Analysis on The PA Market
23	Da Vinci - The Human Library
24	Text Mining 10-K Filling Report Predict Financial Distress Using Risk Factors
25	Prediction of Credit Card Default
26	Disaster news classification
27	Lack of Data? See How Amazon Comments Help Music Analysis and Classification
28	Forecasting Stock Prices, Predicting Earnings Per Share and Classifying Bankruptcy
29	Australia weather analysis for popular tourist city
30	Spotify: Song Genre Prediction
31	Rotten Tomatoes: Movie Genre Prediction
32	Using Bayesian Networks to Model Kickstarter Project Success Rates

33	Churn Warning Prediction on Commercial Bank's Customer
34	Optimizing Online Fraud Transaction Detection
35	Exploring relationship between social media and Bitcoin value with machine learning models
36	Quantify the impact of AI on jobs
37	Predicting Researcher Performance in Depression Treatment
38	Optimization of NYPD Response to Crime in Manhattan
39	Exploring several popular methods on analyzing Candlestick Chart
40	The Opioid Crisis in Numbers
41	Crop Optimization in India
42	Visualization and prediction of awards in a crowed-source environment
43	Plagiarism Detector – Batch Prediction and Cloud Deployment
44	End to End Automated Data Pipeline of Yelp
45	Using the K-means Clustering Analysis for the Unleveraged Portfolio
46	APTOS Blindness Detection
47	Player Recommendation System
48	Malarial Cell Detection Using Transfer Learning
49	Tip Prediction
50	Customer Churn Rate Analytics: Predictive Analysis Model for Retaining Customers
51	News Article Genre Prediction
52	Relationship Between World Market Indices and U.S Market Movement Prediction
53	RFM model: Customer Segment & Quantify Customer Value
54	Academic Advisor (Gil)
55	Watson Schedule Assistant
56	Cognitive Student Advisor
57	Stevens Jr Watson Assistant
58	An Analysis of Topic Prediction Indicators in the Field of Big Data
59	Olympics Analysis (Athens 1896 - Rio 2016)
60	Game Analysis for Evaluating the Strategy of Players
61	Financial Distress Prediction
62	Predicting Probability of Loan Repayment
63	Analyzing Real-Time Streaming Twitter Data
64	Credit Card Fraud Detection
65	Real-Time Fraud Detection System
66	Analysis of single-cell RNA sequences for high, medium, and low CD8 expression cells

Fundamentals are fundamental

Stefano Vanotti, Piyush Dwivedi, Rohan Kadu

Instructor: Alkiviadis Vazacopoulos

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Business Intelligence & Analytics
April 30th, 2015

Introduction & Goals

When choosing stocks, amateurs tend to look solely at prices and expectations around the company's business strategy. Could they beat the market's performance by focusing exclusively on fundamentals?

Is there a single indicator that can give good confidence about a company's performance?

If a company is solid now, can it provide returns in the long term?

Should an investor focus on earnings or on the company's overall profitability?

Experiment

- Take a big set of data with thousands of companies' balance sheets at 12/31/16
- Clean up the dataset (1000+ companies)
- Use numerical analysis to score and rank the companies according to 3 different financial aspects:

1. Margin (Operational and net margin)
2. Debt-level (Current Ratio and Cash Coverage Ratio)
3. Cash Flow (Operating Cash Flow Ratio, Operating Cash flow/share)
4. Overall score which is a mix of the past 3 scores

- Find the one year return and return to today (11/15/19) and for the top and worst 10 of each category
- Compare the different portfolios and their performances
- Run predicting algorithms with machine learning tools to check if a single indicator can be used to make prediction
- Use data visualization tools to gain a deeper understanding of the data

Results

Portfolios Comparison	1 Year Return (12/31/16 -12/31/17)	12/31/16 to today return (11/15/19)
Best Overall	47%	71%
Best Cash	15%	34%
Best Margin	23%	18%
Best Debt	52%	38%
S&P500	18%	37%
Russel 2000	13%	18%
Worst Overall	-26%	-40%
Worst Cash	-5%	-15%
Worst Margin	-8%	-37%
Worst Debt	-8%	-19%

- Overall score can provide excellent result in both short and long term, providing also the worst performance for the lowest scoring companies
- The cash score-based portfolio is the worst performer but also the safest for the worst companies
- Statistical analysis on the numerous companies proved that the overall score has a closer correlation to the return to today than the one-year return
- This indicates that a financially sound company can have "bad" years but in the longer run it will show good returns
- It is interesting to see how the overall score is very highly correlated with operating margins but has a negative correlation with the percentage of operating income that can be converted to net income

Conclusion

- Analysis of fundamentals can lead to great financial performance, especially if all factors are considered
- Companies that are financially sound tend to do well in the long term
- EPS is not a good predictor of a company's performance and operating margin is a more accurate predictor.

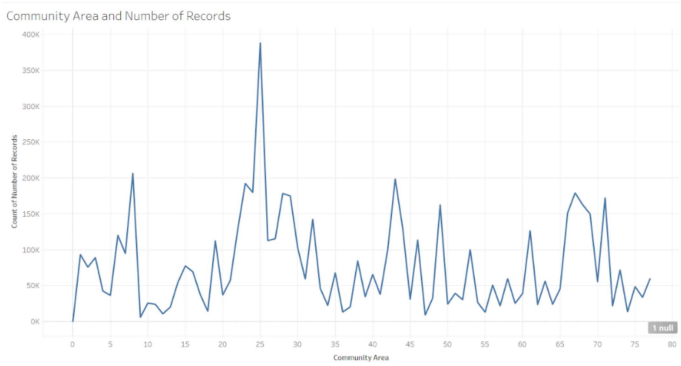
Chicago Crime Data

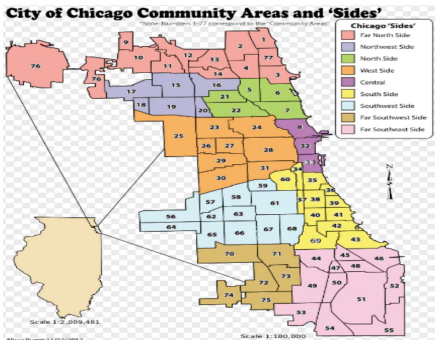
Spencer Cook, Emily Nunez
Professor Vazacopoulos

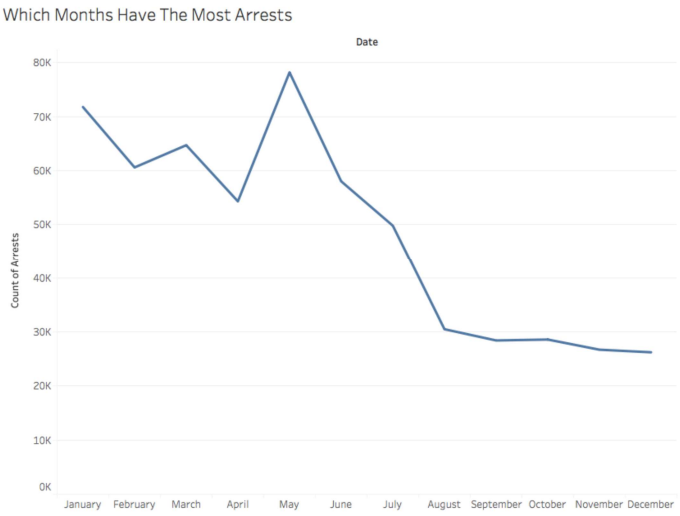
STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics
November 21, 2019

Motivation

- Chicago has consistently been ranked amongst the top cities in the nation in terms of crime
- We decided to use data gathered from 2012-2016 to determine the most dangerous areas in Oakland and the most dangerous times of day and times of the year in order to reduce future crime

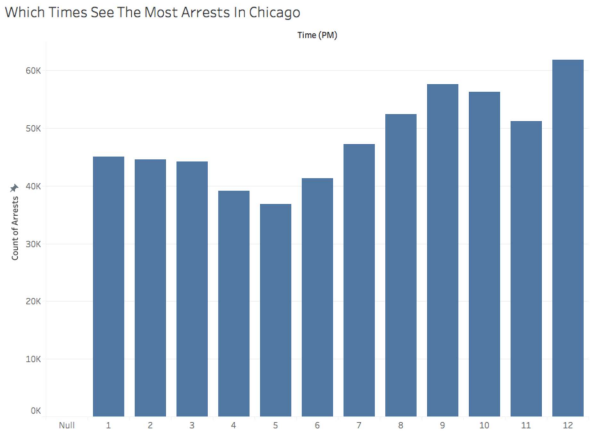






Technology

- **Excel –** Used to gather and clean the data
- **Tableau –** Used to develop visualization models to display findings



Conclusion

- Based on the data, most crimes are between 9pm-12am, in the first half of the year and in Westside

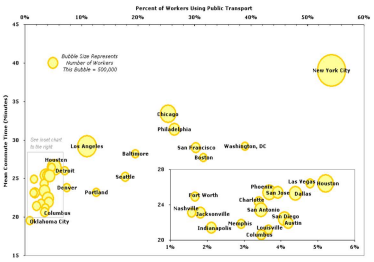
Analysis of Transportation Speed in New York City

Team: • Aditya Paldiwal • Fu Yang • Tingyi Lu
Instructor: Dr. Chihoon Lee

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

1. Motivation

- New York City is plagued by the longest mean travel time for commuters at 39 minutes among major U.S. cities.
- It is imperative to identify the factors affecting travel time to plan for an efficient transport infrastructure to facilitate traffic flow and restrain urban sprawl.

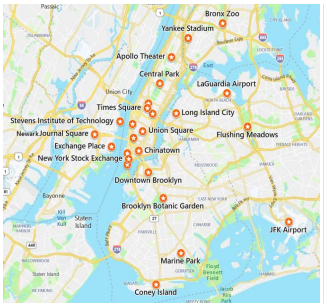


2. Objective

- This experimental design project aims to explore and analyze the influence of time, day and seasonality on the mean travel time across two major modes of transportations; driving and public transit using a fractional factorial design.

3. Data Collection

- Google Map Distance Matrix API was used in conjunction with MS Excel to simulate distance and travel time amongst a set of 25 randomly chosen locations across the New York Metropolitan Area.
- 40,000 observations were collected with an error of 0.6%.



Distance and Travel Time Calculator

Key	
Request URL	https://maps.googleapis.com/maps/api/distancematrix/xml?origins=$1&destinations=$2&mode=$3&key=$k
Origin	Stevens Institute of Technology
Destination	World Trade Centre
Travel mode	driving
URL	https://maps.googleapis.com/maps/api/distancematrix/xml?origins=Stevens+Institute+of+Technology&destinations=World+Trade+Centre&mode=driving&key=
Response	=WEBSERVICE(C13)
Status	=FILTERXML(C15,"//status[1]")
Distance	=FILTERXML(C15,"//distance[1]/text")
Travel Time	=FILTERXML(C15,"//duration[1]/text")

4. Design of Experiment

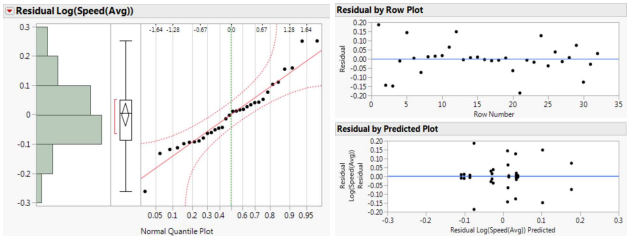
Fractional Factorial Design

- The experiment has 5 factors, thus $2^{5-1} = 16$ runs.
- One replication, hence, the total number of runs = 32.

Factors	Level (-)	Level (+)
Mode of Transport	Driving	Public Transit
Day of the Week	Weekday	Weekend
Time of the Day	Rush Hour	Non-Rush Hour
Public Holiday	Yes	No
Distance	< 10 kms	>= 10 kms
Area (Blocking Factor)		
Manhattan	Rest of New York City	
Response Variable		
Speed = Distance Travelled / Time		

5. Design Evaluation

- The standard regression assumptions we need to check are that the residuals/errors are 1) independent, 2) normally distributed, and 3) have Constant Variance.



6. Regression Analysis

- The R Square is 0.83 meaning 83% of the variance in the response variable 'speed' can be predicted from the factors.
- The P-value of the model is < 0.05, hence the independent variables can reliably predict the response variable.
- Regression Equation:

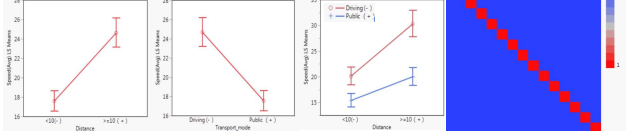
Log(Speed) = 3.04 + (0.17)Transport Mode - (0.17)Distance

Summary of Fit

RSquare	0.831786
RSquare Adj	0.799437
Root Mean Square Error	0.121294
Mean of Response	3.035268
Observations (or Sum Wgts)	32

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	1.8914729	0.378295	25.7130
Error	26	0.3825171	0.014712	Prob > F
C. Total	31	2.2739900		<.0001*

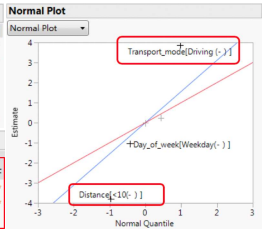


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.0352677	0.021442	141.56	<.0001*
Transport_mode[Driving (-)]	0.1708486	0.021442	7.97	<.0001*
Day_of_week[Weekday(-)]	-0.039726	0.021442	-1.85	0.0753
Time_of_day[Non-rush hour (+)]	0.0063093	0.021442	0.29	0.7709
Holiday[No (+)]	0.0067605	0.021442	0.32	0.7551
Distance[<10(-)]	-0.168094	0.021442	-7.84	<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Transport_mode	1	1	0.93405597	66.5120	<.0001*
Distance	1	1	0.90417980	64.3846	<.0001*
Transport_mode*Distance	1	1	0.04253828	3.0291	0.0928



7. Results

- Mode of Transport and Distance have P-value < 0.05 so they are statistically significant to predict log(Speed).
- If P-value was taken as 0.1 as per business norms, Day and Transport_Mode*Distance would have been significant.
- The impact of time, day and holidays or the interaction between these factors was found to be insignificant.

8. Practical Implications

- Factors such as condition of transport infrastructure, population density and signal cycle length can be used to make model efficacious.
- Outlier detection techniques can be used to identify the routes with abnormally long travel time. Regression analysis can be performed on a set of such routes to find causality.

Retail Analytics and Sales Forecasting

Team: • Aditya Paldiwal • Fu Yang • Tingyi Lu • Xiaomin Yang
Instructor: Dr. Alkis Vazacopoulos

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

1. Motivation

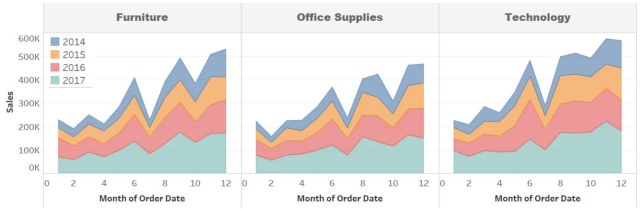
- To maintain a competitive edge in a fast-growing marketplace, it is becoming increasingly necessary for retail companies to look for proactive methods of harnessing new and extensive data sources in unique ways.
- Analytics can help retailers achieve deeper understanding of their customer data and offer actionable insights into the consumer's minds or market trends.

2. Objective

- This project aims to perform time series analysis on retail data and use forecasting models to estimate the future value of the sales.

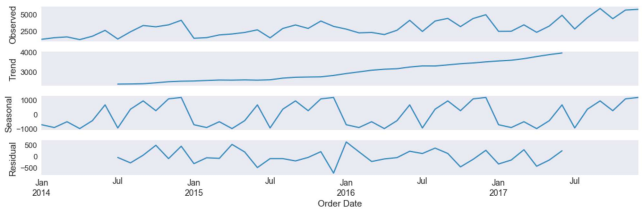
3. Data

- Global Superstore is a retail dataset of a fictious online vendor who deals in three categories of products – Furniture, Office Supplies and Technology.
- The stacked area chart shows the sales for each of the three categories of products from 2014 to 2018.

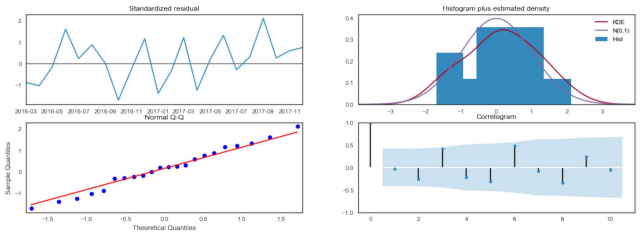


4. Time Series Analysis

- Time-series decomposition: It decomposes time series into three distinct components: trend, seasonality, and noise .
- The sales are always low at the beginning of the year and high at the end of the year.



- The diagnostic plot are used to evaluate the basic regression assumptions that the residuals are 1) Independent, 2) Normally distributed and 3) have Constant Variance.



5. Forecasting

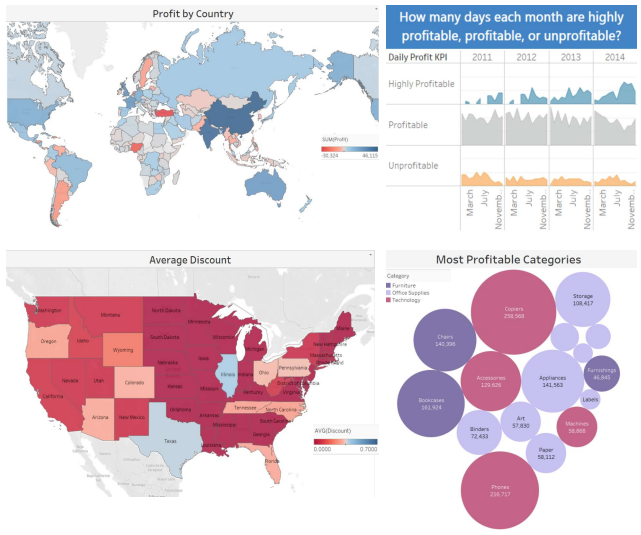
The algorithms used for forecasting are

- Holt-Winters (Triple Exponential Smoothing)
- Box-Jenkins (SARIMA)
- Facebook Prophet

6. Conclusion

- All the three models are capturing the sales seasonality.
- The forecast becomes less reliable the further we move in the future as the confidence intervals keeps getting larger.

7. Visualizations



WINE ME UP in the USA

Giselle Cruz
Samantha Lederle
Carinne Werthaim
Instructor: Alkis Vazacopoulos
*Scan the QR codes for our LinkedIn Profiles!

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Analytics M.B.A.

MOTIVATION

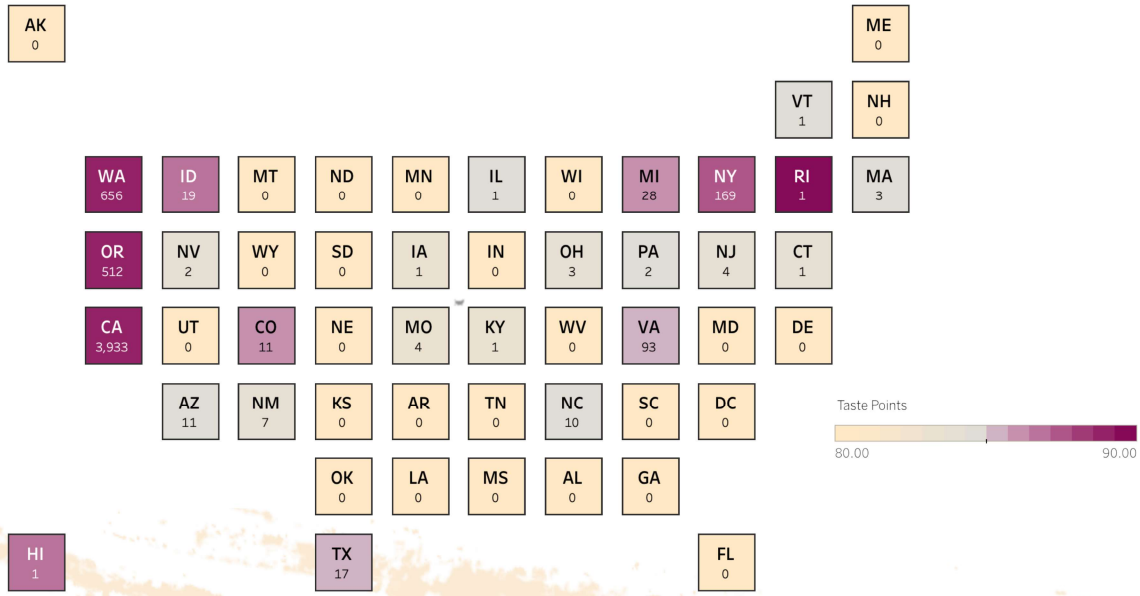
There's a vast and unimaginable large world of wine out there. Here, we take a look at a slice of that, particularly in our very own country, the **United State of America**.

BACKGROUND

Wine Enthusiast gives us a deep look into various wines from all over the world. Their expert reviewers who have become connoisseurs in all things vino, rate wines based on a **Vintage Point System**.

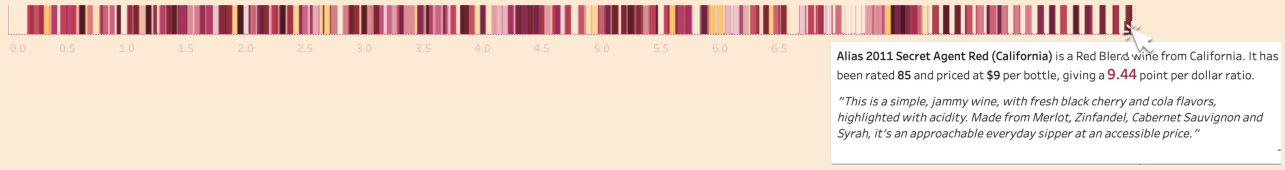
Wine Enthusiast has reviewed over **5,491 wineries within the United States** and their wide collection of reds, whites and in-betweens.

80-82	83-86	87-89	90-93	94-97	98-100
Acceptable	Good	Very Good	Excellent	Superb	Classic

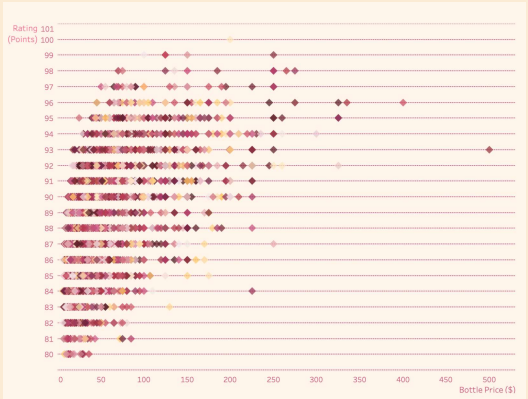


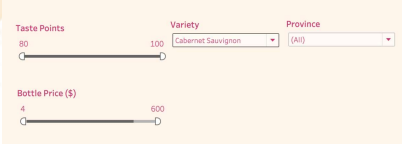
POINT PER DOLLAR

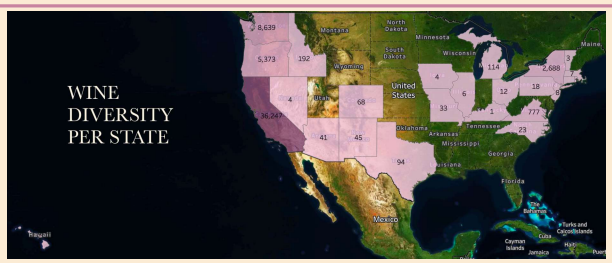
Which wine bottle gets you the most bang for your buck? According to our viz, it is **Alias's Secret Agent Red from California**.



WINE SELECTOR







USING TOOL TIPS

For a deeper insight into what some of the marks mean, hovering over it can reveal additional information such as the exact review of each bottle that the Wine Enthusiast connoisseur wrote. Editing and using Tool Tips is one of our favorite features in Tableau.

Want to see our viz live on your own desktop? Scan this QR Code!

SCAN ME

Design of Experiments on Machine Learning Hyperparameter Tuning

Chen Liao, Yu Hong, Haitao Liu, Tianyu Liu
Instructor: Chihoon Lee

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

Introduction

- We designed the experiment table for hyperparameters-tuning in LSTM (Long Short-Term Memory)
- The hyperparameters including learning rate, decay rate, number of neurons, epochs, dropout probability.
- Analyzed share prices of UPS and FedEx with 14 predictors, including trading prices and commodity prices.
- Take MSE as response
- Training size = 1000
- Testing size = 475
- The Factors:

FACTOR	NEURON	LEARNING RATE	EPOCH	DECAY RATE	DROPOUT PROBABILITY
+	28	0.002	20	0.95	0.5
-	14	0.001	10	0.9	0.3

- Block: UPS, FedEx

Plackett-Burman Design

1. PB Design Table with fold-over

keep_prob	epochs	neurons	LR	decay_rate
+	-	-	+	-
-	+	-	-	+
+	-	+	-	-
-	+	-	+	-
-	-	+	-	+
+	+	+	+	+
-	+	+	-	+
+	-	+	+	-
-	+	-	+	+
+	-	+	-	+
+	+	-	+	-
-	-	-	-	-

2. The Result Of FedEx

The number of neurons is the significant factor

Regression Function:

*Y=55.7254+12.915*neurons*

Parameter Estimate Population

Term	Estimate	t Ratio	Orthog Coded	Orthog t-Ratio	Prob>|t|
Intercept	85.4554	17.8128	85.4554	17.8128	<.0001*
keep_prob[-]	12.5896	1.8556	-3.3654	-0.7015	0.4920
epochs[-]	10.6946	1.5763	3.1522	0.6571	0.5195
neurons[-]	-1.4192	-0.2092	14.2346	2.9671	0.0083*
LR[-]	-22.1917	-3.2709	-18.6821	-3.8942	0.0011*
decay_rate[-]	13.5596	1.9986	9.5881	1.9986	0.0610

Each Orthog Estimate is conditioned on the effects before it

3. The Result of UPS

The number of neurons and the learning rate the most significant factors

Regression Function:

*Y = 85.4554 – 22.1917*learning _rate – 1.4192*neurons*

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	55.725417	4.293206	12.98	<.0001*
keep_prob[-]	-4.97125	6.07151	-0.82	0.4236
epochs[-]	3.735	6.07151	0.62	0.5461
neurons[-]	12.915	6.07151	2.13	0.0475*
LR[-]	-4.1225	6.07151	-0.68	0.5058
decay_rate[-]	5.2375	6.07151	0.86	0.3997

Full Factorial Design

1. The Result Of FedEx

Regression Function:

Y = 84.8991-9.2074*keep_prob-20.0888*epochs-10.8681*neurons-27.0628*LR+6.2385*keep_prob*neurons+7.8832*keep_prob*LR+14.3469*neurons*learning_rate

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
✓	✓	Intercept	84.8991092	1	0	0.000	1
☐	✓	keep_prob[--]	-9.2074365	3	10661.39	7.557	0.00025
☐	✓	epochs[--]	-20.088797	1	25510.27	54.248	8.6e-10
☐	✓	neurons[--]	-10.868061	3	21083.11	14.945	2.91e-7
☐	✓	LR[--]	-27.062749	3	58279.88	41.311	3.2e-14
☐	☐	decay_rate[--]	0	1	1513.993	3.355	0.07242
☐	☐	keep_prob[--]*epochs[--]	0	1	289.4903	0.611	0.43764
☐	✓	keep_prob[--]*neurons[--]	6.23848418	1	2460.171	5.232	0.02598
☐	✓	keep_prob[--]*LR[--]	7.88317168	1	3928.341	8.354	0.00547
☐	☐	keep_prob[--]*decay_rate[--]	0	1	677.8342	1.453	0.23319
☐	☐	epochs[--]*neurons[--]	0	1	170.1199	0.358	0.55229
☐	☐	epochs[--]*LR[--]	0	1	1222.841	2.678	0.10743
☐	☐	epochs[--]*decay_rate[--]	0	1	1462.59	3.234	0.0776
☐	✓	neurons[--]*LR[--]	14.3469217	1	13011.42	27.669	2.34e-6
☐	☐	neurons[--]*decay_rate[--]	0	1	1480.174	3.276	0.07579
☐	☐	LR[--]*decay_rate[--]	0	1	703.7908	1.510	0.23433

2. The Result of UPS

Regression Function:

*Y=56.4484-13.3547*learning_rate*

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
✓	☐	Intercept	56.7982647	1	0	0.000	1
☐	☐	keep_prob[--]	0	1	879.7353	0.743	0.39219
☐	☐	epochs[--]	0	1	3031.514	2.638	0.10951
☐	☐	neurons[--]	0	1	436.8367	0.367	0.54715
☐	✓	LR[--]	-13.354735	1	11369.75	9.638	0.00287
☐	☐	decay_rate[--]	0	1	265.0657	0.222	0.6393
☐	☐	keep_prob[--]*epochs[--]	0	1	893.8288	0.755	0.3884
☐	☐	keep_prob[--]*neurons[--]	0	3	4846.917	1.396	0.25303
☐	☐	keep_prob[--]*LR[--]	0	2	1850.952	0.779	0.46349
☐	☐	keep_prob[--]*decay_rate[--]	0	1	24.3379	0.020	0.88716
☐	☐	epochs[--]*neurons[--]	0	1	1393.321	1.185	0.2807
☐	☐	epochs[--]*LR[--]	0	1	2365.606	2.039	0.15842
☐	☐	epochs[--]*decay_rate[--]	0	1	381.7233	0.320	0.57366
☐	☐	neurons[--]*LR[--]	0	2	1273.172	0.531	0.59048
☐	☐	neurons[--]*decay_rate[--]	0	3	2966.066	0.831	0.48202
☐	☐	LR[--]*decay_rate[--]	0	2	3013.18	1.289	0.28306

3. Full Factorial Design with Block

Regression Function:

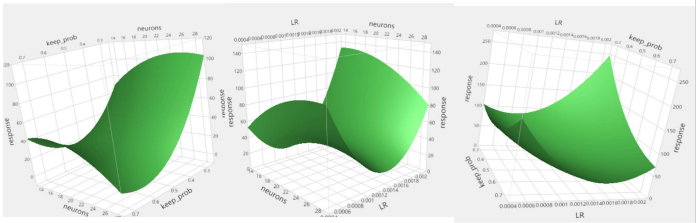
*Y=72.8236-15.2222*Block +14.1114*epochs-23.5138*LR +9.2846*decay_rate*

Parameter Estimates

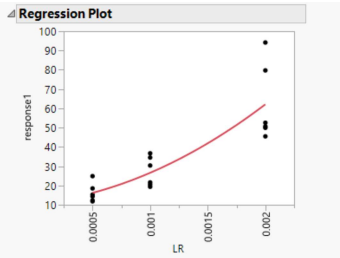
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	72.823559	3.852642	18.90	<.0001*
Block[1]	-15.22219	3.829456	-3.98	0.0002*
keep_prob[-]	2.9097103	3.852642	0.76	0.4532
epochs[-]	14.111371	3.852642	3.66	0.0005*
neurons[-]	3.3734603	3.852642	0.88	0.3849
LR[-]	-23.51377	3.852642	-6.10	<.0001*
decay_rate[-]	9.284566	3.852642	2.41	0.0192*

Response Curve

- **Fedex:**
  - LR: (0.0005, 0.001, 0.002) ->0.00125
  - Epochs:(10, 20, 30) ->20
  - Keep Probability:(0.3, 0.5, 0.7) ->0.5
  - Neurons(14, 21, 28) -> 21
  - Decay Rate -> 0.95



- **UPS**
  - LR -> 0.0005
  - Epochs ->20
  - Decay Rate -> 0.95
  - Keep Probability->0.5
  - Neurons ->14



Redefining Cancer Diagnosis

Authors: Aditya Manchanda, Kritika Karamchandani, Siddharth Singh, Qiudi Zhu, Ashima Munjal
Instructor: Prof. Edward Stohr

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

MOTIVATION

A lot has been said during the past several years about how precision medicine and, more concretely, how genetic testing is going to disrupt the way diseases like cancer are treated. But this is only partially happening due to the huge amount of manual work still required. A major problem faced by MSK hospital is to manually verify the presence of cancer in genetic variations from the clinical literature. With our machine learning algorithm we are classifying if a given genetic mutation/variation based on evidence from text-based clinical literature.

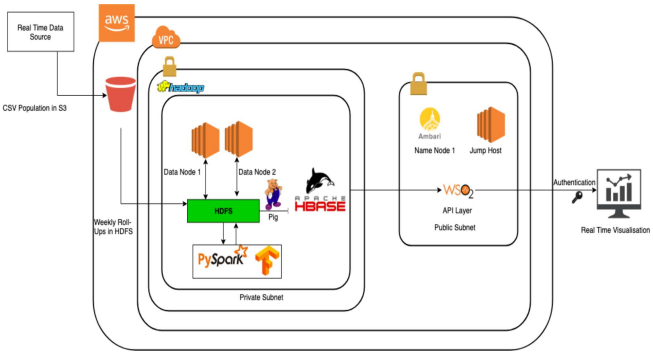
BUSINESS OBJECTIVES

- No low-latency requirement.
- Interpretability is important.
- Errors can be very costly.
- Probability of a data-point belonging to each class is needed.

DATA OVERVIEW

- Source: https://www.kaggle.com/c/msk-redefining-cancer-treatment/data
- We have two data files: one contains the information about the genetic mutations and the other contains the clinical evidence (text) that human experts/pathologists use to classify the genetic mutations.
- Both these data files have a common column called ID
- Data file's information:
  - Training variants (ID , Gene, Variations, Class)
  - Training text (ID, Text)

ARCHITECHTURE



TECHNOLOGY/METHODOLOGY

- Python, Node.js
- AWS
- Hadoop: Spark, Pig, HDFS, Hbase
- WSO2 & PowerBI
- Performance Metric: Multi-class log loss, Confusion Matrix

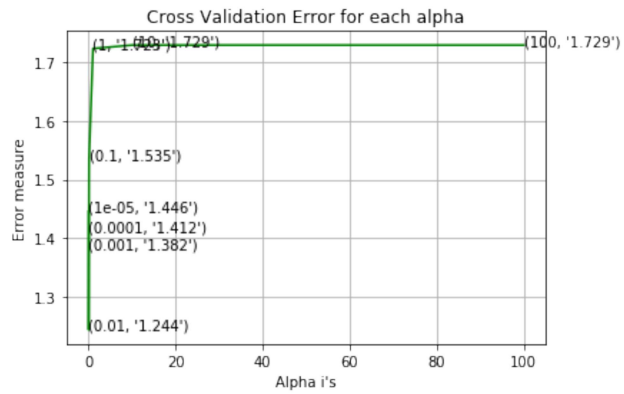
INTERPRETABILITY

As the problem involves predicting such sensitive results regarding cancer, we have taken interpretability as the topmost priority into account. Most of the results of predicted class are backed by information of most important features which led to the prediction of that class,

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0153 0.1199 0.0029 0.0151 0.0121
Actual Class : 7
--------------------------------------------------
28 Text feature [constitutively] present in test data point [True]
29 Text feature [cysteine] present in test data point [True]
49 Text feature [cdnas] present in test data point [True]
76 Text feature [flt1] present in test data point [True]
79 Text feature [concentrations] present in test data point [True]
82 Text feature [gaiix] present in test data point [True]
96 Text feature [technology] present in test data point [True]
101 Text feature [inhibited] present in test data point [True]
```

IMPLEMENTATION

The features Genes, Variation and Text were taken as the feature for the classifier as the log loss calculated from logistic regression was less that the random model. We applied the following models Naïve Bayes, KNN, Logistic Regression, Linear SVM, Random Forest, Stacking Classifiers and Majority Voting Classifier on the model by tuning hyper parameter for the alpha and then calculated the log loss and the number of misclassified points with the confusion matrix. Now for the purpose of implementing this model in production we used logistic regression for which given a query point we calculated the probabilities of the nine classes and the reasoning and stored in our data store Hbase. Once the data is in Hbase we displayed them in dashboards in Power BI.



RESULT

The following table contains the log los and number of misclassified points after applying the data and also if they were interpretable in terms of output.

Model	Encoding	Train	CV	Test	Misclassified Points (%)
Naïve Bayes	OHE	0.9	1.27	1.21	39.84
KNN	Response	0.71	1.13	1	39.47
LR + Balancing	OHE	0.61	1.14	1.05	34.77
LR + Imbalance	OHE	0.63	1.19	1.05	36.28
SVM + Balancing	OHE	0.74	1.13	1.06	36.47
Random Forest	OHE	0.7	1.19	1.1	36.84
Random Forest	Response	0.05	1.33	1.21	46.8
Stacking Classifier	OHE	0.66	1.18	1.09	36.24
Majority Voting	OHE	0.91	1.23	1.15	33.9

CONCLUSION

Using the above methodologies what we observed is that logistic regression with class balancing was the most effective and interpretable model in classifying cancer. Also it was good in giving out the reasoning with the most important features that contributed towards the prediction of a particular class. Having an efficiency of 66% in classifying can really save up pathologist's time and contribute to speedy treatment of patients which in turn can lead to a huge impact on the current cancer treatment.

NYSE Stock market prediction with technical analysis using LSTM Recurrent Neural Network.

Souraj Krishnan
Instructor: Edward A. Stohr

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Whom does it concern?

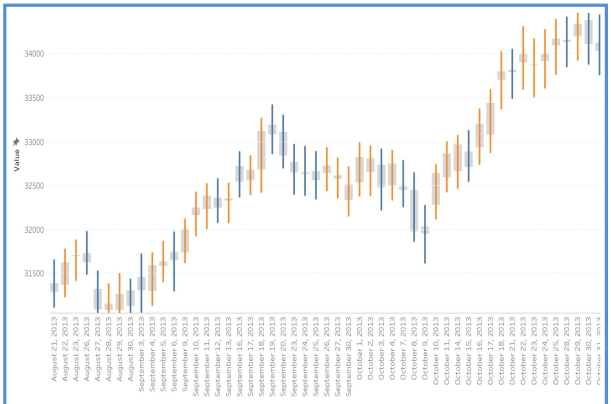
- Trading approximately 1.46 billion **shares** each day, the **New York Stock Exchange** (**NYSE**) is the leading **stock exchange** in the world.
- The main goal of any **investor** is to **minimize risk** and **maximize return**. Hence stock market prediction is important to both Investors and the organization they invest in.

Problem statement

- Number of organizations today, have started turning to machine learning, for the task of forecasting stock prices.
- The most common approach that can be used to achieve this task is **Linear Regression** however, this model isn't reliable while dealing with **time series.**
- **Long short-term memory** is an artificial **recurrent neural network (RNN)** that is ideal for analyzing time series
- In this experiment **RSI, Bollinger bands, MACD, Moving Average and EMA** indicators have been used with **LSTM** to predict stock prices and show that the accuracy of this model is higher than that of a **Regressor.**

Technology used:

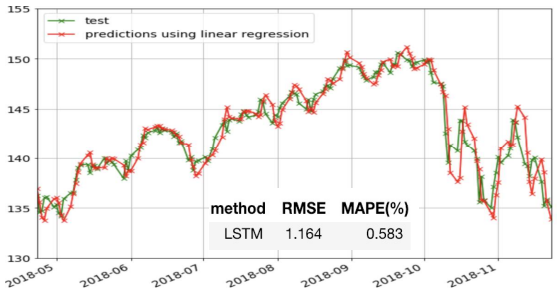
- **Python** for preprocessing, exploratory data analysis, computing technical indicators, and training the model –modules used are **Pandas, Numpy, and Scikit-learn.**
- **TensorFlow** library used for modeling the Recurrent Neural Network.
- For visualization **Python** modules – **Matplotlib** and **Seaborn** were used, along with **Tableau.**



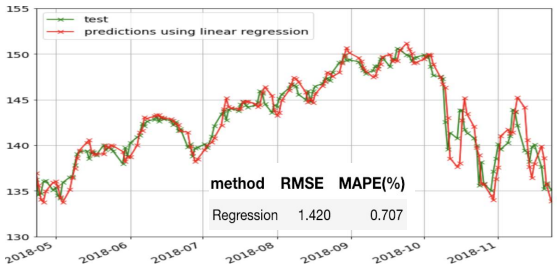
(Open High Low Close graph of WLTW in tableau)

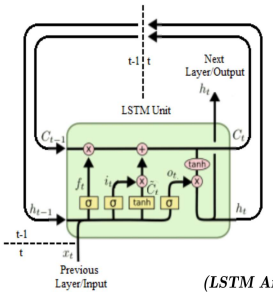
Result

Using **LSTM (Recurrent Neural Network)** with **technical indicators** we achieve a **mean absolute percentage error** of **0.583**



Using a **Linear Regression model** with the same **technical indicators** we achieve a **mean absolute percentage error** of **0.707**

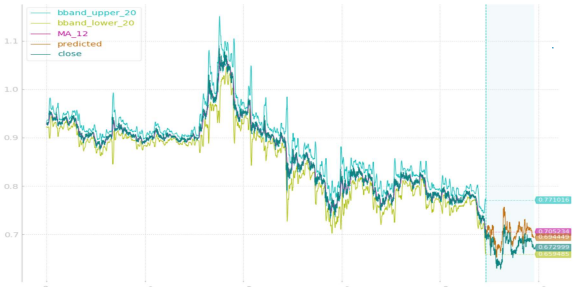




$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$
$$o_t = \sigma\left(W_o\,[h_{t-1}, x_t] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

(LSTM Architecture)

The blue region shows the technical indicators and prices predicted using the Recurrent Neural Network.



Future work

It would be interesting to work on **longer forecasting horizons**, with indicators such as **ARIMA** and **triple exponential smoothing**.

YouTube Data NLP Analysis

Rishabh Kasat, Yuning Yang, Shunchao Chen, Ziyu Ma, Zi Wang
Instructor: Dr. Edward Stohr

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

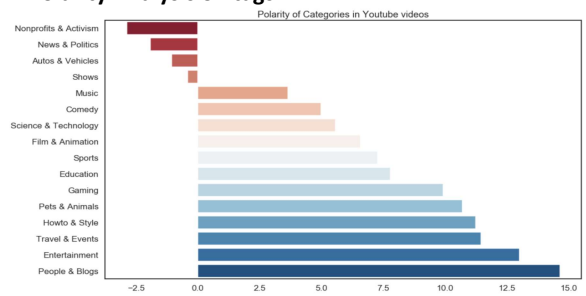
Introduction

- YouTube maintains a list of the top trending videos on the platform.
- YouTube uses a combination of factors including measuring user's interactions (number of views, shares, comments and likes) to determine if a video is trending.

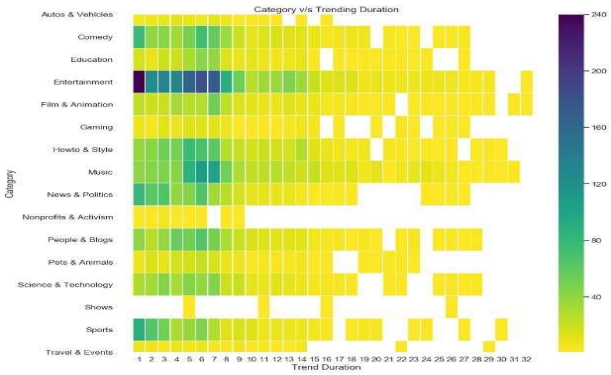
Objectives

- To discover the viewer's perception of videos in different categories based on analyzing the tags, description, likes and dislikes, views.
- To further develop a classification and verify if videos have been tagged in the right category.
- To explore the correlation between video publish time and probability to trend.

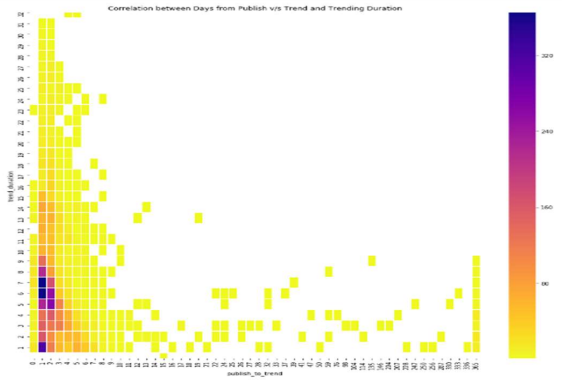
1. Polarity Analysis on tags



2. Heatmap of the trending duration of videos



3. Heatmap of the publish time of trending videos



Modeling

1. Multinomial Naïve Bayes (Baseline Model)

Likelihood

Class Prior Probability

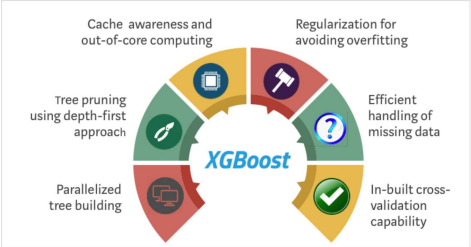
$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability

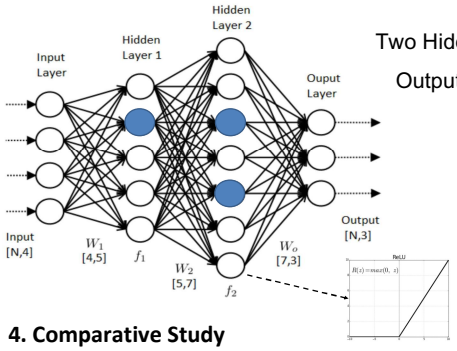
Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

2. XG Boost



3. Artificial Neural Network (3 layers)



Two Hidden Layer: *Relu*

Output Layer: *Softmax*

Optimization:

Adam

Early Stopping

Drop out

Grid Search

4. Comparative Study

	Naïve Bayes	XGBoost	ANN
Accuracy	81%	94.3%	97.8%
Time (mins)	1.2	6.3	21

Conclusion

- Our team built and testified models from statistics, machine learning and deep learning aspects and improve accuracy from 81% to 97.8%.
- Our team developed a model to recognize the best publish dates and trending probability of a video. The model showed that videos published during the first 8 days of the year are more likely to trend comparing to any other time in the year.
- According to the polarity analysis result, Entertainment and Peoples & Blogs acquired the most positive sentiment score.

The Analysis and prediction based on NBA Data of players and teams

Enchen Liu, Jifeng Ye
Instructor: Alsi Vazacopoulos

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics
Nov 21st, 2019

Introduction

For team managers, coaches and other relevant practitioners and even fans, it is essential to know more about the players of their team. Every player's performance affects the outcome of the game, but how much it can be? In this project, we explored the relationship between the data based on NBA players and teams, including data from twitter and Wikipedia. In addition, due to the limited salary cap, we have also estimated the players' annual salary in order to find the reference salary payable to each specific player.

Method

1. Tableau for visual data analysis
2. Rapid miner for data prediction
3. Microsoft Excel for data correlation analysis

Result:

Chart 1: Does individual player performance impact a team's win?

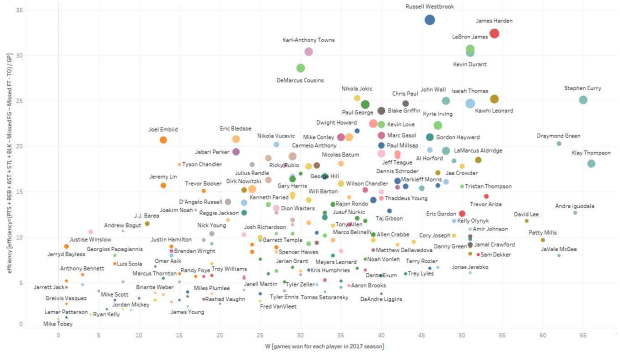


Chart 2: Which player contributes more to the team?

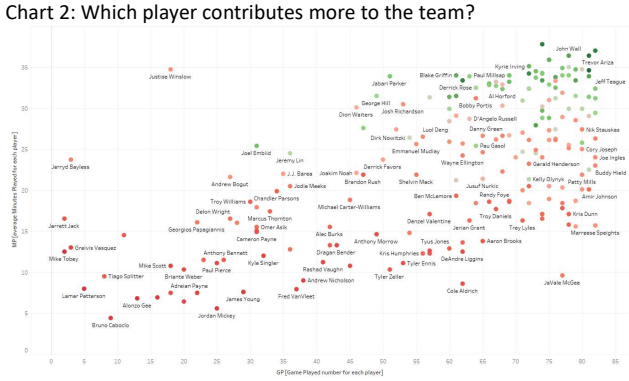
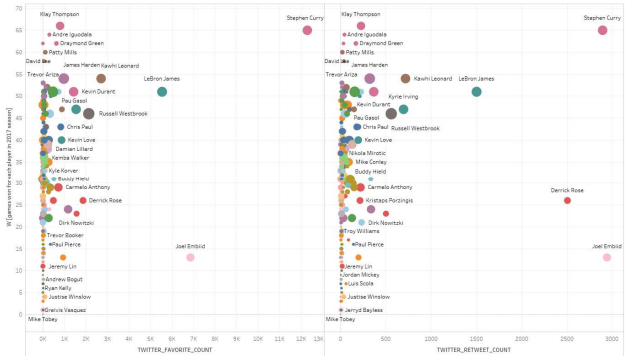


Chart 3: Will the players get popular if they are doing good on court?



Results

Chart 4: If the player are paid more, will they contribute more to the team?

Chart 5: Twitter Follow Count Millions, PTS, Wins RPM vs Endorsement

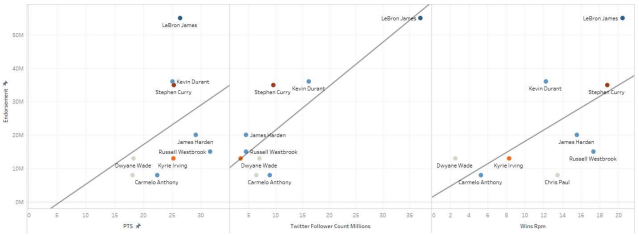


Chart 6: PTS, Wins Rpm vs Salaries

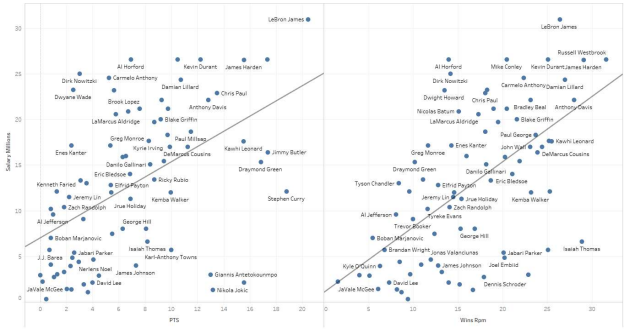


Chart 7: Correlation between different data

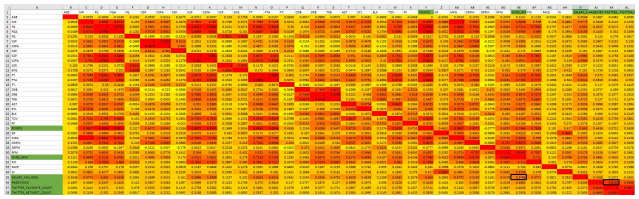


Chart 8: The prediction of annual salary from Rapid Miner(Sample)

Row No.	SALARY_MI...	prediction(S...	POSITION	TEAM	AGE	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%
1	26.500	24.916	PG	OKC	28	34.000	10.200	24	0.425	2.500	7.200	0.343	7.700	16.800	0.459
2	22.100	21.293	C	IND	23	36.100	10.200	20.300	0.505	0.500	1.800	0.289	9.700	18.600	0.524
3	24.330	27.422	PG	POR	26	35.900	8.800	19.800	0.444	2.900	7.700	0.370	6	12.100	0.492
4	17.640	19.303	PG	CLE	24	35.100	8.300	19.700	0.473	2.500	6.100	0.431	5.900	13.600	0.505
5	5.983	14.730	C	MIN	21	37	9.800	18	0.542	1.200	3.400	0.357	8.500	14.700	0.582
6	18.310	13.746	SF	IND	20	35.900	8.300	18	0.461	2.600	6.600	0.393	5.700	11.400	0.501
7	6.016	14.819	SF	MIN	21	37.200	8.800	19.100	0.452	1.300	3.500	0.356	7.400	15.600	0.473
8	25.140	21.057	PF	LAC	27	34	7.900	15.900	0.493	0.900	1.900	0.336	7.200	14.100	0.514
9	14	14.106	PG	PHX	27	33	6.800	15.700	0.434	1.900	4.700	0.395	5.200	11	0.477
10	15.890	12.166	PG	MIA	30	33.700	7.300	15.400	0.475	1.600	4	0.405	5.700	11.400	0.499
11	2.240	11.950	SG	MIN	21	37.200	6.900	15.100	0.459	2.600	6.600	0.387	4.400	8.500	0.515
12	15.850	16.736	SF	DEN	20	33.900	5.300	11.900	0.447	2	5.100	0.389	3.300	6.800	0.491
13	20.870	15.905	PF	ATL	31	34	6.200	14.100	0.442	1.100	3.500	0.311	5.100	10.600	0.486
14	4.320	10.853	PF	NY	21	32.800	6.700	14.900	0.450	1.700	4.800	0.357	5	10.200	0.493
15	21.320	14.141	PG	NY	26	32.500	7.200	15.300	0.471	0.200	0.900	0.217	7	14.300	0.487
16	17	9.773	SG	ORL	24	32.900	6	13.700	0.439	1.900	5.300	0.356	4.100	8.400	0.491
17	6	13.946	PG	UTAH	30	31.500	5.900	12.400	0.477	1.900	4.800	0.403	4	7.600	0.523
18	12.390	10.495	SG	HOU	29	31	5.500	13.500	0.406	3.300	8.800	0.372	2.200	4.700	0.466
19	6.568	8.223	SG	OKC	24	33.200	6.100	13.900	0.442	1.900	5.300	0.361	4.300	8.700	0.491
20	2.960	9.450	SG	MIA	20	36.100	6.100	14.400	0.424	1.800	4.700	0.395	4.300	9.700	0.438

Predicting Airbnb Prices In Jersey City Using Machine Learning

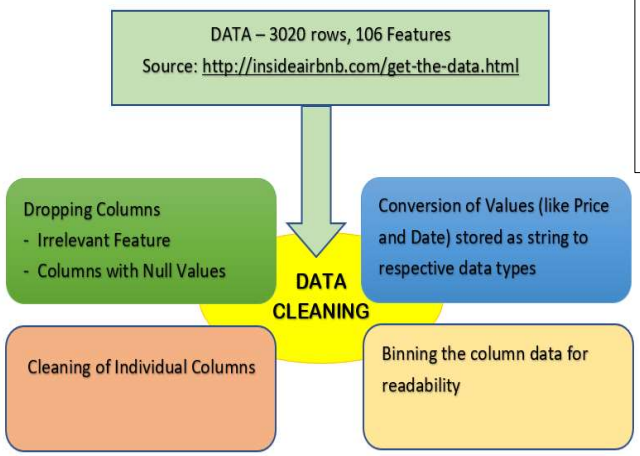
Author: Pranil Bhavsar, Bhakti Dalvi, Ayushi Singh, Biying Gao, Xinyi Li, Chao He
Instructor: Professor Edward A Stohr

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction

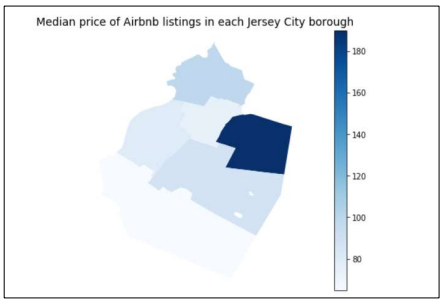
- **Real world problem:** Airbnb has to set the right price for the property which satisfies not only the host but also is competitive enough for the market and demands of the customers.
- **Purpose:** This projects aims at optimizing the pricing functionality. The model sets the right price according to the important features which seem to have affected the rate of the property over the years.

Data Preparation



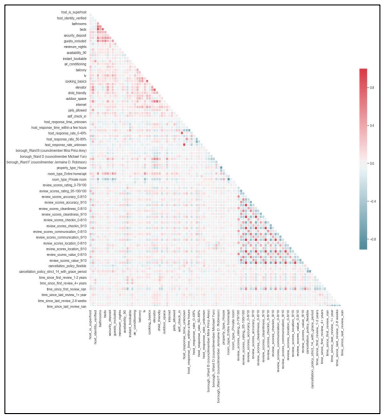
Exploratory data analysis

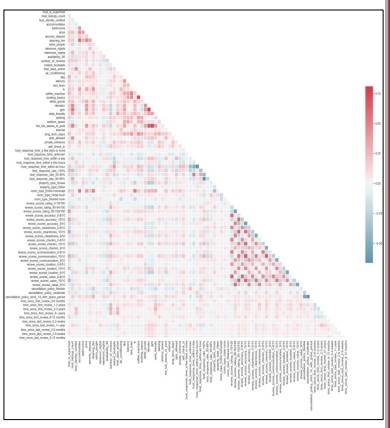
- Categorical features were explored and plotted, to gain insights and to determine whether or not they should be included in the final model.
- According to the chart, we could know that the **largest number and median price** of Airbnb listing in Jersey city are both from **Newport**.



Preparing the data for modeling

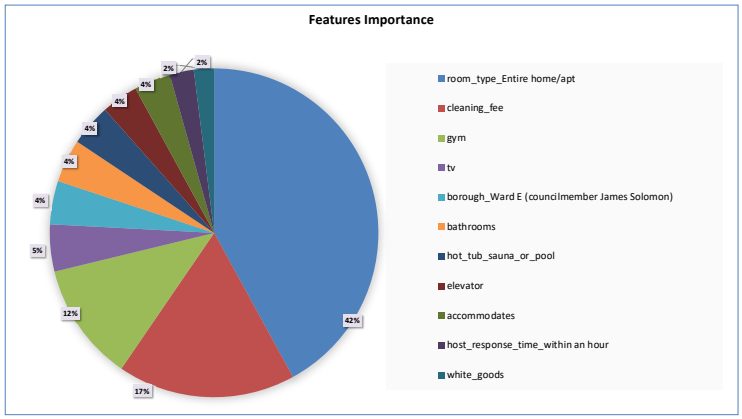
- The features are assessed for multi-collinearity. The following heatmap is observed and accordingly redundant columns are eliminated





XGBoost Model

- Training/ Test Split : 80%/20%
- Training MSE: 0.0882, Validation MSE: 0.119
- Training R^2: 0.8546, Validation R^2: 0.7907



Results

- This model guides Airbnb/Hosts to optimally set their rent and maximize their profits. Besides, the result helps them to understand the features that affect customer behavior and help them improve upon the customers overall experience.
- As noted above in the pie chart, room type, cleaning fee and gym are the factors which influence the price the most.

Directions for Future Work

- We can use Sentiment Analysis upon receiving the reviews, which could lead the model to be more accurate and reliable.
- One of the imperative methods, incorporating image quality, could be used.
- Natural Language Processing can be to used to operate the listing descriptions and reviews. Hyperparameter optimization can be used as well in order to generalize different data patterns during the learning process.

Customer Lifetime Value Optimization

Yubo Feng, Yanran Min, Fu Yang, Xuan Zhao, Ruizhi Hao
Instructor: Somayeh Moazeni

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction

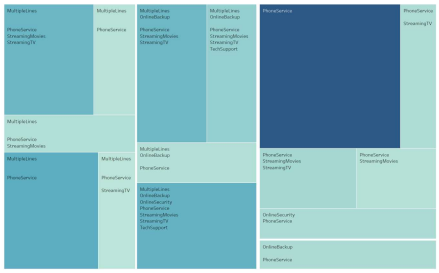
Decision making based on customer lifetime value(CLV) is playing an increasingly critical role in various industries. CLV is the expected net profit a company can get from a customer. With the concept of CLV, the effect of a business decision can be translated into the amount of CLV increased[1]. Furthermore, it encourages a company to shift focus from short-term profit to long-term profit[2].

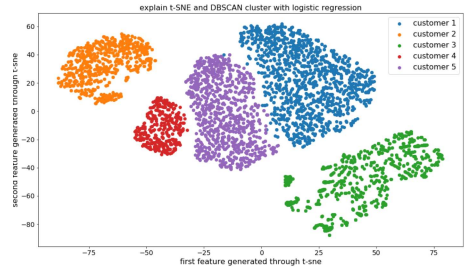
Technology

- Explored data through manifold learning and visualized it using python and Tableau.
- Segmented customer into five types using DBSCAN and Logistic Regression.
- Performed market basket analysis with Apriori; Designed ten service packages.
- Computed and validated CLV using churn rate estimated by Random Forest.
- Used solver to build CLV optimization model.

Data Exploration & Customer Segmentation
↓
CLV Estimation & Model Validation
↓
CLV Optimization

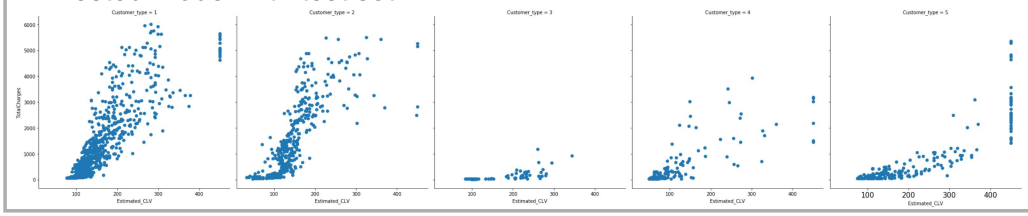
Data Exploration & Customer Segmentation

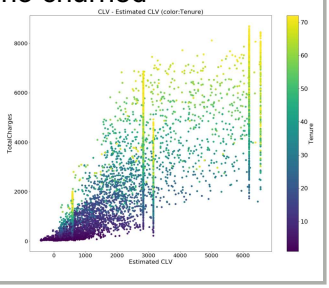




CLV Estimation & Model Validation

- Estimated probability of churn with Random Forest
- Computed expected CLV; compared it with total charges of the customers who churned
- Built linear transformation between total charges and estimated CLV
- Tested model with test set





CLV Optimization

$$\max \sum_{i,j} CLV_{ij} \cdot \frac{N_{ij}}{N}$$

$$st\ \Delta \boldsymbol{p} \le \boldsymbol{p_0} \cdot 10\%$$

$$CLV_i = k_i \frac{MC * CR}{1 + CR - \bar{i}} + b_i$$

$$CR(\boldsymbol{p}) = CR(\boldsymbol{p_0}) - \boldsymbol{a}^T(1 - \boldsymbol{m}^T \Delta p)\Delta \boldsymbol{p}$$

CLV_{ij}: Customer Lifetime Value of Customer i who purchase package j

$CR(\boldsymbol{p})$: Churn rate at price level $\boldsymbol{p}$, $\boldsymbol{p}$: Price of ten services packages

MC: Monthly charge, $\bar{i}$: Interest rate

N_{ij}: Number of customer i who buy package j

$\boldsymbol{a}^T$: Price elasticity $\boldsymbol{m}^T$: Marginal effect diminishing ratio

k_i: Over Exaggerate ratio of churn rate for customer i

Customer life value after discount

	customer type 1	customer type 2	customer t	customer type 4	customer type 5
Package 1	$ 2,308	$ 1,351			
Package 2	$ 1,716	$ 2,572			
Package 3	$ 3,950	$ 3,207			
Package 4		$ 6,702			$ 2,747
Package 5		$ 6,287			$ 2,747
Package 6		$ 7,185			$ 2,747
Package 7		$ 997		$ 1,481	
Package 8			$ 388		
Package 9		$ 7,185			$ 2,747
Package 10	$ 6,436	$ 7,185			
Rest					

			Discount limit
Sum discount of package 1	$ 1.14	<=	$ 8.0
Sum discount of package 2	$ 0.75	<=	$ 8.0
Sum discount of package 3	$ 1.57	<=	$ 9.0
Sum discount of package 4	$ 5.51	<=	$ 5.5
Sum discount of package 5	$ 5.51	<=	$ 5.5
Sum discount of package 6	$ -	<=	$ 5.5
Sum discount of package 7	$ 2.50	<=	$ 2.5
Sum discount of package 8	$ -	<=	$ -
Sum discount of package 9	$ -	<=	$ 9.0
Sum discount of package 10	$ -	<=	$ 11.5

Conclusion & Reference

- Real churn rate is proportional to estimated churn rate.
- CLV for low-tenure customers are overestimated, while CLV for high-tenure customers are underestimated
- Mean square error of CLV on test set is 1034.
- **Price of DSL & Fiber Optic** has high impact on CLV.
- The customers who have subscribed for more than 46 months have high CLV.
- The customers who buy all the services and the customers who have subscribe for more than 46 months **are loyal enough that no discount is needed**.

[1] Berger, P. D., Nasr, N. I. (1998). Customer lifetime value: Marketing models and applications. Journal of Interactive Marketing. 12 (1): 17–30
[2] Farris, Paul W., Neil T. Bendle, Phillip E. Pfeifer, David J. Reibstein (2010). Marketing Metrics: The Definitive Guide to Measuring Marketing Performance. Upper Saddle River, New Jersey: Pearson Education

Cookbook Consumer Behavior Analysis From Amazon.com

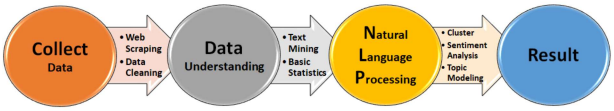
Xing Fang, Yating Liu, Zi-Qi Liu, Tianrui Wang
Instructor: Professor Rong Liu

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

Introduction

This project aims to analyze the content and the reviews of the best-selling cookbooks on Amazon and derive valuable consumer behavior patterns for the publishing industry.

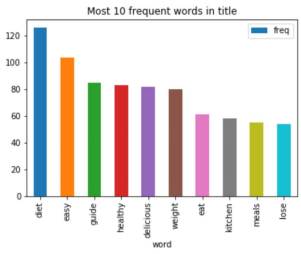
Project Flow

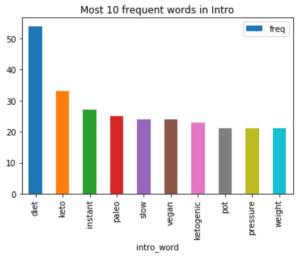


Collect Data and Understanding

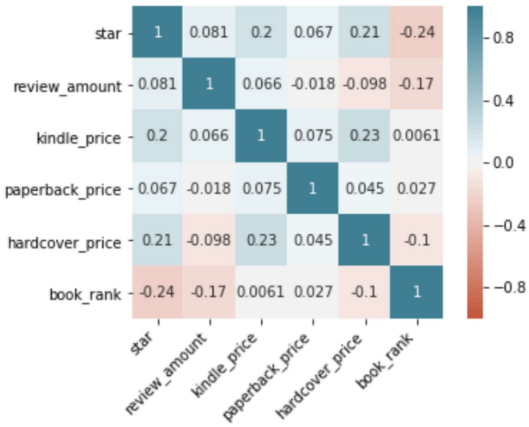
- **Web Scraping**
  - We used Selenium Webdriver to scrape the data of 1200 books from 'Cookbooks, Food & Wine' category on Amazon.com.
  - There are 9 main categories we are interested
- **Exploratory Data Analysis**
  - **WordCloud of Book Titles**

  - **Top words: Title and Intro**



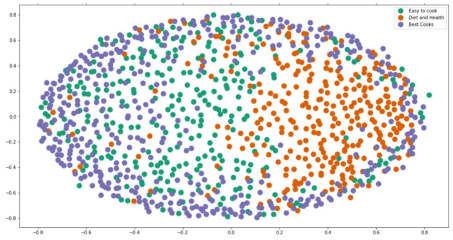


  - **Variables Correlation**



Natural Language Processing

- **K-Means Clustering (K = 3) on Book Information**



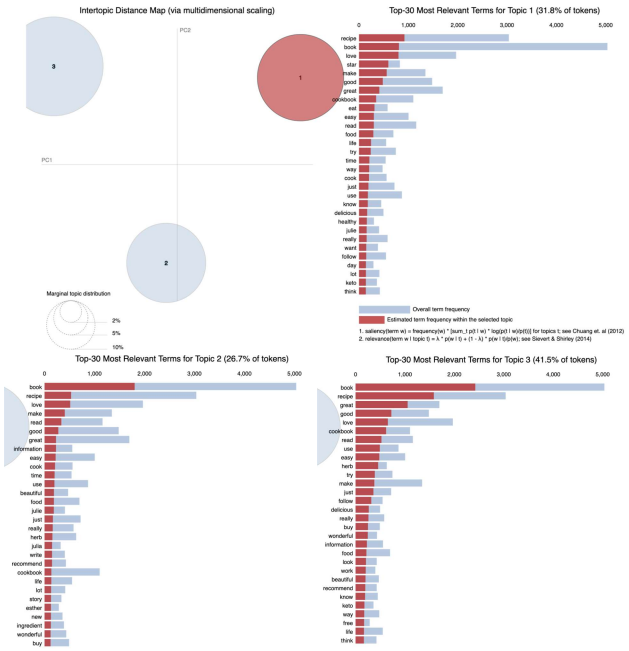
Cluster 1: Easy to cook | Cluster 2: Diet and Health | Cluster 3: Popular Chefs

- **Sentiment Analysis Based on Clustering**

(Accuracy calculated based on Amazon Sentiment Tag)

	0: Easy-to-Cook	1: Diet & Health	2: Famous Chefs
Sentiment Accuracy	0.70	0.81	0.65

- **Topel Modeling on Book Reviews**



Conclusion

- Current Trends of Cookbooks: Easy-To-Cook, Diet & Health (Keto, Vegan, Gluten), Famous Chefs
- **Authors, Price, Waste-of-time**
  - constantly appears in negative reviews
- **Erythritol, Carb, Sweet, Calorie**
  - Body Fitness is a main concern
- **Herbs, Detox, Ayurveda**
  - Seeking Cure From Food?
- Infused Water seemed to be a tough topic to write on.

Data Science Approach to Automate Granting/Revoking Employee Computer Access Needs

Aditya Gade, Ninad Kulkarni
Instructor: Edward Stohr

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Motivation

•When an employee at any company starts work, they first need the computer access necessary to fulfill their role.

•Employees are manually allowed or denied access to resources over time. You must create an algorithm capable of learning from this historical data to predict approval/denial for an unseen set of employees.

•In this data science project, we built an auto-access model that minimizes the human involvement required to grant or revoke employee access.

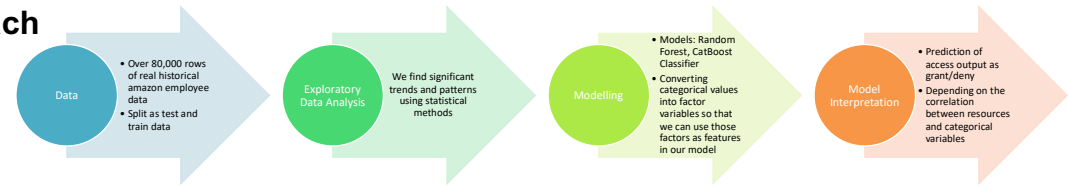
Technology

•Python for the implementation of model approach.

•Numpy library for linear algebra, Pandas library for data processing, CatBoost library for modelling and predicting.

• Shap for interactive visualizations, you can analyze your model by switching parameters for both abscissa and ordinate

Model Approach



EDA & Feature Engineering

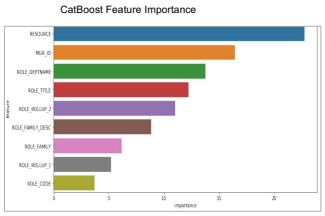
Test Set

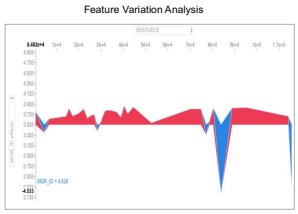
	id	RESOURCE	MGR_ID	ROLE_ROLLUP_1	ROLE_ROLLUP_2	ROLE_DEPTNAME	ROLE_TITLE	ROLE_FAMILY
count	58921	58921	58921	58921	58921	58921	58921	58921
unique	58921	4971	4689	126	177	466	351	2749
top	58921	4675	770	117961	118300	117878	118321	117906
freq	1	1989	276	37658	7731	2261	7433	11100

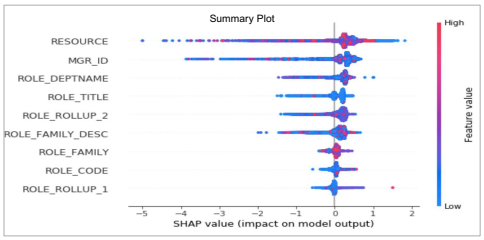
Train Set

	ACTION	RESOURCE	MGR_ID	ROLE_ROLLUP_1	ROLE_ROLLUP_2	ROLE_DEPTNAME	ROLE_TITLE	ROLE_FAMI
count	32769	32769	32769	32769	32769	32769	32769	32769
unique	2	7518	4243	128	177	449	343	2358
top	1	4675	770	117961	118300	117878	118321	117906
freq	30872	839	152	21407	4424	1135	4649	6896

These features are actually codes for different properties of an employee: **manager id, company role code, resource** and other. Thus, the data contains categorical features.







In the above diagram, every employee (instance/row in our dataset) is represented by one dot in each row. The x position of the dot is the impact of that feature on the model's prediction, and the color of the dot represents the value of that feature for that exact employee. Here we can see that **'ROLE_ROLLUP_1'** and **'ROLE_CODE'** features have low impact on the model prediction, and for most of employees their impact is almost zero.

Modelling: CatBoost Classifier

CatBoost introduces the following algorithmic advances:

1. Categorical features support:

You do not need to preprocess categorical features (like one-hot encoding), just have to specify some hyperparameters.

2. Better overfitting handling:

CatBoost uses the implementation of ordered boosting, an alternative to the classic boosting algorithm.

3. Other useful features: Fast and easy-to-use GPU-training, Missing value support (***nan_mode*** HP), great visualization,

AUC (Area Under Curve) Illustration

- **True Positive Rate = TP / (TP + FN)** Percentage of approved access that are correctly predicted
- **False Positive Rate = FP / (FP + TN)** Percentage of denied access that are incorrectly predicted (Denied Access which are actually Approved Access)
- **AUC= TP+FP/(TP+FP+TN+FN)**

Model Performance	AUC	No. of Iterations
Before Feature Engineering	0.8394	412
After Feature Engineering	0.9002	452

Conclusion

- Our model strikes right balance and chooses optimal features and number of trees you want to build. It shows solid results training on unprocessed categorical features
- This Data Science approach can automatically determine access privileges as employees enter and leave roles within a company, thus saves significant amount of time and money.

Statistical Analysis and Optimization of Service Time at a Local Cafe

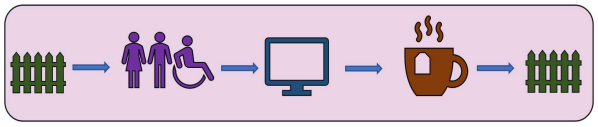
Saloni Kapoor; Malvika Reddy Kandi; Padmini Sapre
Instructor: Dr. Edward Stohr

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

Motivation

•Analyze statistical nature of total service time for customers arriving at America's Cup on the campus of Stevens Institute of Technology

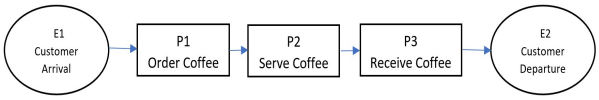
•Reduce the amount of time spent by a customer to receive order



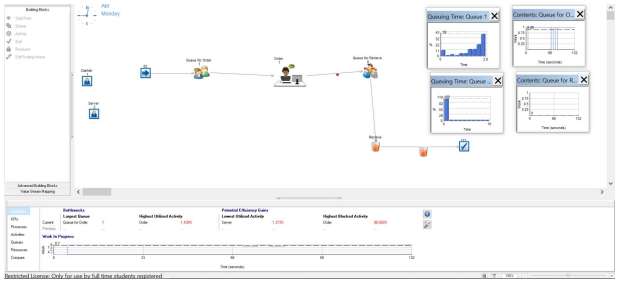
Model Approach

Total Service Time = Coffee Serving Time + Counter Serving Time
Average Counter Service Time = 127.5442 Seconds
Average Coffee Service Time = 143.1349 Seconds

EPC Model



This is a model which shows the imitation of the current process flow at the cafe. It consist of a single cashier and a single server which is causing bottlenecks.



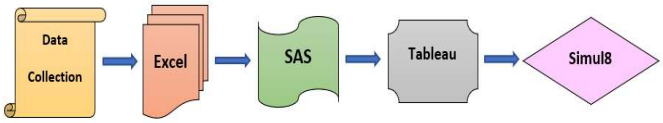
Result

The optimized simulation model consists of 2 servers which ensures that the cashier and servers work at the same speed. The efficiency has increased by 43.7% which indirectly causes a reduction in total service time.

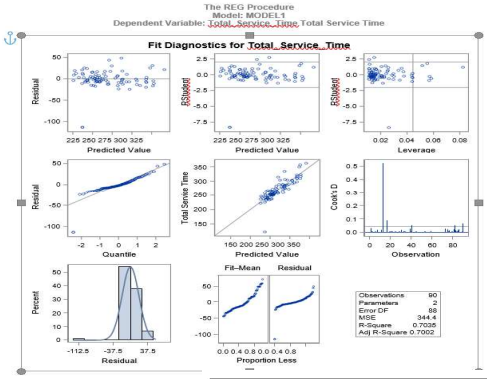
Future Scope

The scope of the project can be extended by adding data pertaining to various stages of making a sandwich and then optimizing each stage.

Technology

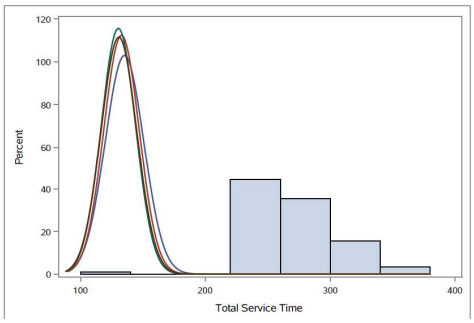


Statistical Analysis

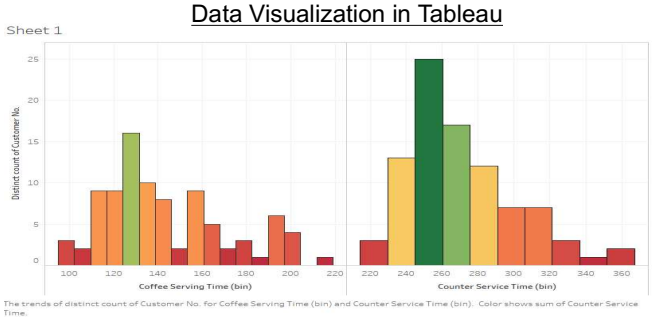


Normality fit in SAS
Result:
Not Normal, Appears to follow Gamma Distribution

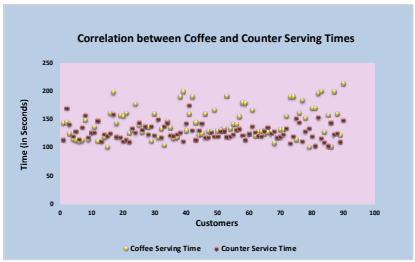
Gamma Diagnostics in SAS
Result:
Inferred as Gamma Distribution



Data Visualization in Tableau



Result: Confirmed as Gamma Distribution



From the bubble graph, we can infer that counter service time and coffee service time are highly positively correlated.

Towards an accurate prediction of customer behavior for IBM Watson marketing data, using Deep Learning

Souraj Krishnan | Minal Chandra | Kanish Shah

Instructor: Edward A Stohr

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Business Intelligence & Analytics

Motivation

- In present times, predictive analysis has become a vital aspect in business decision making of major companies in various industries such as the insurance industry.
- Insurance companies are looking for insights into the reason for low customer response rate, which is affected by various factors.

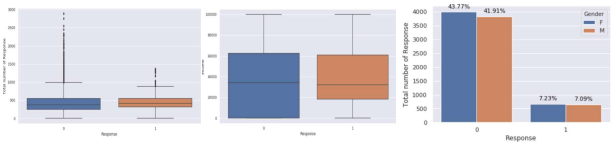
Problem statement

The experiment involves classifying customer response for IBM Watson marketing customer value dataset for car insurance policies as **"yes"** or **"no"** using three machine learning models namely:

1. **Multinomial Naïve Bayesian algorithm**
2. **Random Forest**
3. **Artificial Neural Network**

Exploratory Data Analysis

- **Missing values** were replaced with average measures and **Outliers** were identified with Z-score calculation and removed.
- **Feature engineering:** Understanding dataset helps us extract more information about from the raw data. **Label-encoder** was used to turn categorical variables like **"State", "Policy type",** etc. to numerical values.
- **Understanding data using visualizations:**



Feature Selection

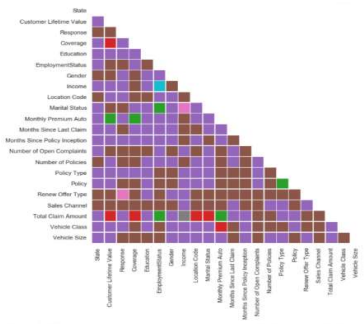
- Normalizing the data fields using **min-max scaler.**

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

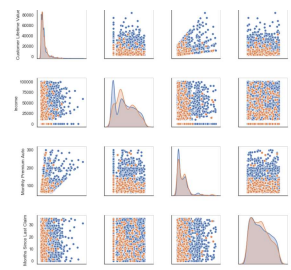
- Creating a **correlation matrix** and heatmap to understand the correlation between independent variables.

	State	Customer Lifetime Value	Response	Coverage	Education	Gender	Income	Location Code	Marital Status	Months Since Last Claim	Months Since Policy Inception
State	1.000000	0.008778	0.002473	0.002047	0.015755	0.004933	0.004620	0.005240	0.020371	0.002126	0.013965
Customer Lifetime Value	0.008778	1.000000	0.008930	0.168040	0.026038	0.013610	0.024366	0.004854	0.026003	0.011517	0.009418
Response	0.002473	0.008930	1.000000	0.000642	0.000285	0.004396	0.011932	0.001729	0.094843	0.016597	0.002952
Coverage	0.002047	0.168040	0.000642	1.000000	0.021573	0.014448	0.010646	0.020176	0.002956	0.016925	0.010658
Education	0.015755	0.026038	0.000285	0.021573	1.000000	0.002206	0.018709	0.011533	0.005675	0.003021	0.004140
Gender	0.004933	0.013610	0.004396	0.014448	0.002206	1.000000	0.015138	0.003108	0.034027	0.025522	0.004198
Income	0.004620	0.024366	0.011932	0.010646	0.018709	0.015138	1.000000	0.027517	0.231673	0.026715	0.000875
Location Code	0.005240	0.004854	0.001729	0.020176	0.011533	0.003108	0.027517	1.000000	0.016452	0.016733	0.007030
Marital Status	0.020371	0.026003	0.094843	0.002956	0.005675	0.034027	0.231673	0.016452	1.000000	0.007757	0.029345
Months Since Last Claim	0.002126	0.011517	0.016597	0.016925	0.003021	0.025522	0.026715	0.016733	0.007757	1.000000	0.042959
Months Since Policy Inception	0.013965	0.009418	0.002952	0.010658	0.004140	0.004198	0.000875	0.007030	0.029345	0.042959	1.000000

Correlation matrix



heatmap

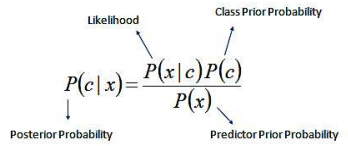


Pair plot

Modeling

We implemented the following Machine Learning models and compared the accuracy scores:

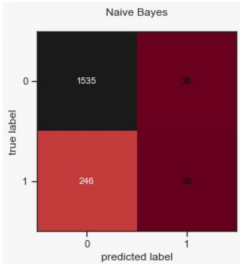
1) Naïve Bayes Classifier



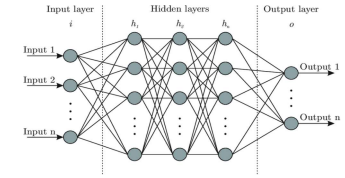
$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

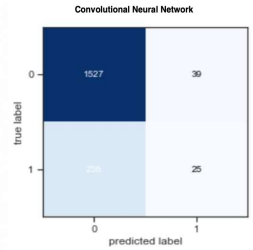
	precision	recall	f1-score	support
0.0	0.86	0.98	0.92	1561
1.0	0.43	0.08	0.13	266



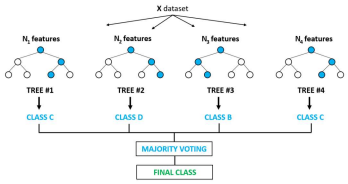
2) Artificial Neural Network



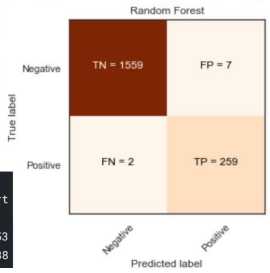

```
accuracy_score: 0.8659003831417624
              precision    recall  f1-score   support

       False       0.96      0.89      0.93      1704
        True       0.25      0.51      0.34       123
```



3) Random forest classifier




```
Accuracy score: 0.9879605983217804
              precision    recall  f1-score   support

           0       1.00      0.99      0.99      2353
           1       0.93      0.98      0.96       388
```



Conclusion and future work:

- We observe that Random forest classifier has the highest precision of the three classifiers 98.79% for the task of classification.
- In the future it would be interesting to use clustering for customer segmentation to market personalized policies to specific customers.

Loan Approval Prediction based on Customer History

Authors: Sujay Kumar Tatineni, Aishwarya Kandalam
Instructor: Dr. Mahmoud Daneshmand

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

INTRODUCTION

Aim: To predict loan eligibility of customers using Classification algorithms based on customer history

Models:

K- Nearest Neighbor

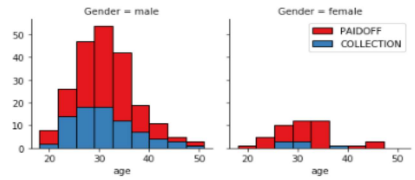
Logistic Regression

Decision Tree

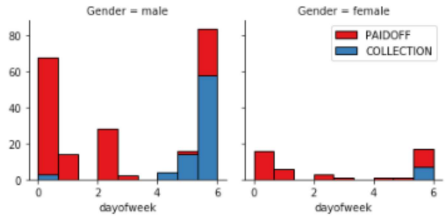
Support Vector Machine

Project Flow

- The data set consists of info of about 346 customers. The dataset is first split into training (80%) and testing(20%)
- Data visualization and Preprocessing



- Feature selection/ extraction



- Using Normalization the redundancy and dependency of data is reduced and prepared for testing
- Applying the classification algorithms: KNN, Decision Tree, Support Vector Machine, Logistic Regression

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.67	0.63	NA
Decision Tree	0.72	0.74	NA
SVM	0.80	0.76	NA
LogisticRegression	0.74	0.66	0.57

MODELING

K Nearest Neighbor(KNN) – Using Minkowski distance equation the nearest neighbors for the classification are obtained

$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

Decision Tree - Decision tree builds classification or regression models in the form of a tree structure. The entropy criterion is used for this

Support Vector Machine - SVM is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

Logistic Regression - Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical)

CONCLUSION

- Using the classifiers, we get the highest accuracy for the SVM Classifier
- This in turn calls for a proper prediction model which can predict the required probability by taking in consideration the independent parameters.
- There is a lot of scope to improve the prediction model. With many parameters and advanced prediction and classification algorithms available, the prediction model can be improved and updated constantly to give more accurate predictions. And with prescriptive analytics and advanced machine learning techniques coming into play, the model's prediction ability can be further trained efficiently and its ability to predict more accurate values can be taken to a whole new level.

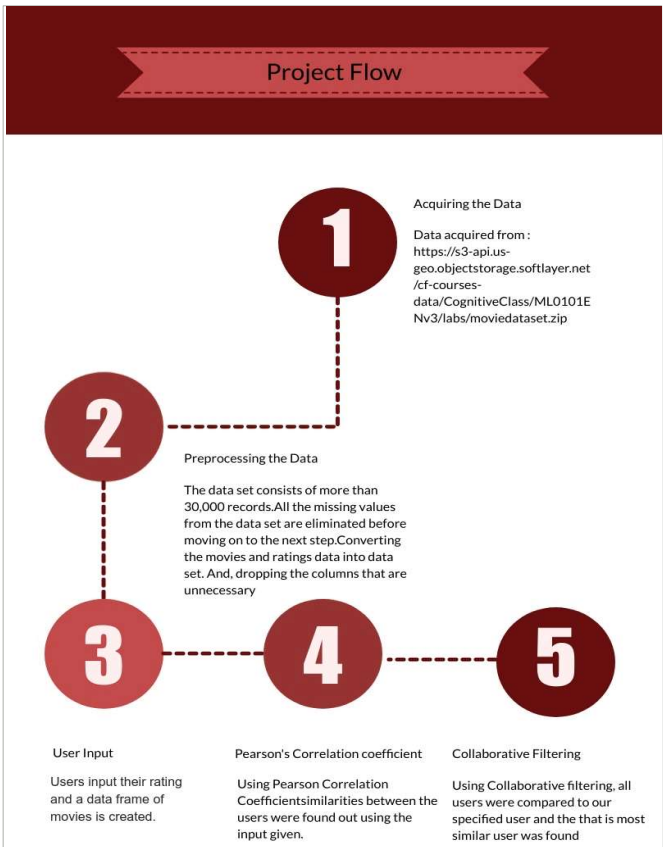
Movie Recommendations based on Viewer History

Author: Sujay Kumar Tatineni
Instructor: Dr. Mahmoud Daneshmand

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction

Aim: To create a User Based movie recommendation system using Collaborative Filtering



Pearson Correlation Coefficient: if you have two vectors X and Y, then, pearson(X, Y) == pearson(X, 2 * Y + 3). This is a pretty important property in recommendation systems because for example two users might rate two series of items totally different in terms of absolute rates, but they would be similar users (i.e. with similar ideas) with similar rates in various scales

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Modelling:

Collaborative Filtering:

- In the context of **recommendation systems**, collaborative filtering is a method of making predictions about the interests of user by analyzing the taste of users which are like the said user. The idea of filtering patterns by **collaborating multiple viewpoints** is why it is called collaborative filtering.

WHY COLLABORATIVE FILTERING?

- The main difference between collaborative filtering and content-based filtering is conceptual. Where content-based filtering is built around the attributes of a given object, collaborative filtering relies on the behavior of users. This approach has some distinct advantages over content-based filtering:
- **It benefits from large user bases.** Simply put, the more people are using the service, the better your recommendations will become, without doing additional development work or relying on subject area expertise.
- **It's flexible across different domains.** Collaborative filtering approaches are well suited to highly diverse sets of items. Where content-based filters rely on metadata, collaborative filtering is based on real-life activity, allowing it to make connections between seemingly disparate items (like say, an outboard motor and a fishing rod) that nonetheless might be relevant to some set of users (in this case, people who like to fish).
- **It produces more serendipitous recommendations.** When it comes to recommendations, accuracy isn't always the highest priority. Content-based filtering approaches tend to show users items that are very similar to items they've already liked, which can lead to filter bubble problems. By contrast, most users have interests that span different subsets, which in theory can result in more diverse (and interesting) recommendations.
- **It can capture more nuance around items.** Even a highly detailed content-based filtering system will only capture some of the features of a given item. By relying on actual human experience, collaborative filtering can sometimes recommend items that have a greater affinity with one another than a strict comparison of their attributes would suggest.

Conclusion

- Using collaborative filtering, similar movies based on initial user input ratings were recommended to the user.

Advantages

- Takes other user's ratings into consideration
- Adapts to the user's interests which might change over time

Disadvantages

- Approximation function can be slow
- Insufficient amount of users to approximate
- Privacy issues when trying to learn the user's preferences

Academic Chatbot For BIA Program

Haihan Hu, Lun Li, Xiaochi Ma

Instructor: Christopher Asakiewicz

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Background

In the start of a new academic year, all the candidates are buzzing with questions relating to admission, courses, tuitions and the like. It's undoubtedly a busy time for admission departments to this influx of questions quickly requires a lot of time and resources.

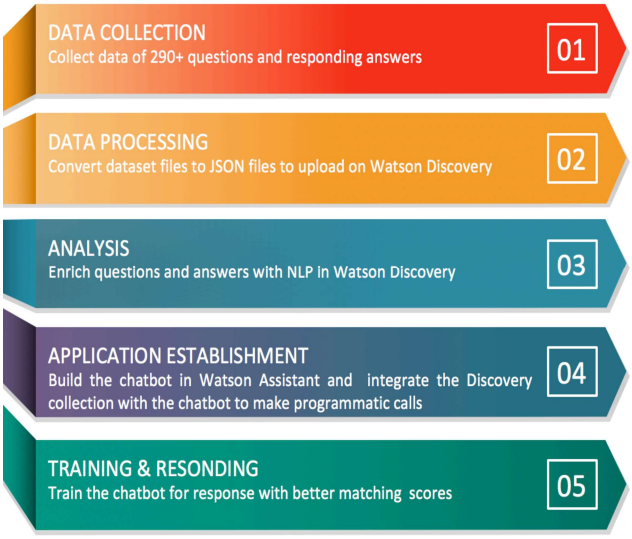
Objective

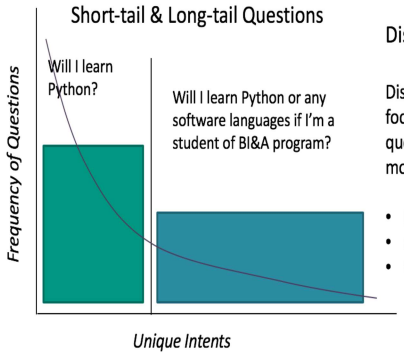
The main goal of this project is to develop an academic chatbot application which can chat with the applicants in a realistic manner and give relevant response. It has a benefit to school and the potential applicants by saving time, reducing frustration and helping the interest eventually turns into applications.

Data Source

- 290+ questions and corresponding answers
- Corpus form websites
  (BI&A program, School of business, Stevens)

Project Process





Discovery Services

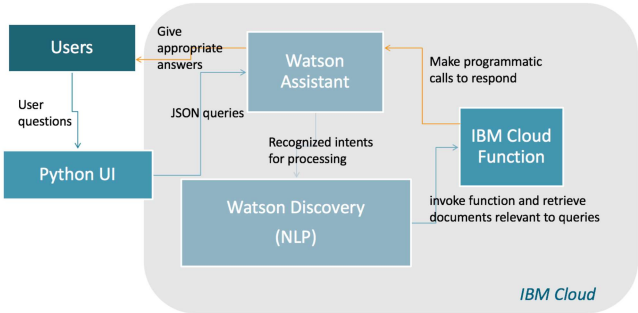
Discovery uses strategies that focus on processing long-tail questions and identifying the most appropriate answers:

- Keyword Extraction
- Concept Tagging
- Category Classification

Implementation And Results

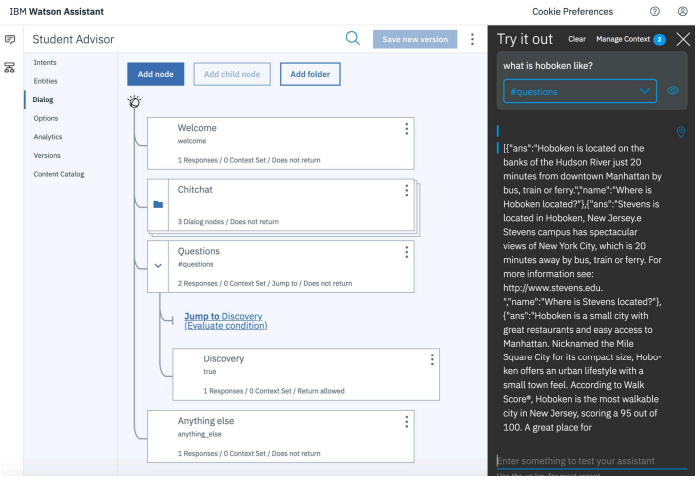
Implementation Details

- Use Watson Assistant create an academic chatbot. We combine the intent and the entities in the condition for the dialog node in Watson Assistant.
- Launch Watson Discovery collection and import structured Json format source content documents then add enrichment to our documents. Discovery Service takes a few steps as shown in the diagram below:



Combine Discovery collection with the academic chatbot by using IBM Cloud Functions to make programmatic calls from dialog node. With this connection we can dynamically revolving results rather than hardcoding response.

The final combined chatbot showed below:



Discussion

Limitations

Academic chatbot does not support multiple languages and dialects.

Academic chatbot cannot comprehensive the conversational tones.

Future Scope

Deploy the academic chatbot to different channel, E.g.. Slack, school official Facebook webpage.

Add advanced functions which can process audio information.

Flight Scheduling Optimization for American Airlines — based on Deficit Function and Network Flow Model

Author: Jinghao Wang, Han Zhu, Kun Liu, Nithin Das, Swithin Ruskin Roy
Instructor: Edward A Stohr

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics
Nov 14th, 2019

Introduction

We consider a periodic aviation schedule which is a finite collection of flights with given arrival/departure times. For this schedule, we solve two problems: 1) find the minimal number of aircrafts needed to carry out the flights; 2) maximize profits by removing several uneconomic flights.

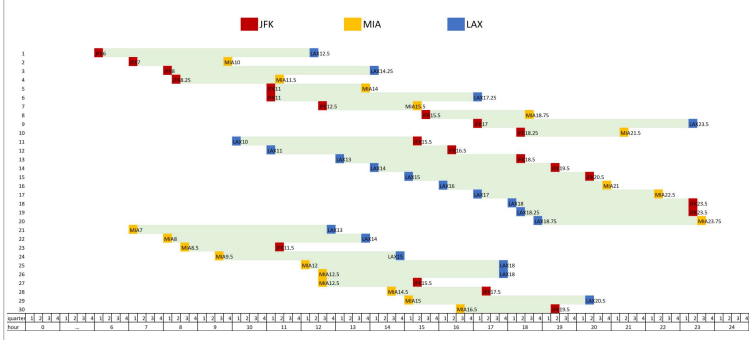
The crucial role in solving the problems is played by deficit function (DF) defined as the difference between departures and arrivals on interval $[0, t]$ for each terminal, and network flow model which performs very well in a flow selection linear optimization problem.

Solution to Minimal Number of Aircrafts

For a flight schedule like this:

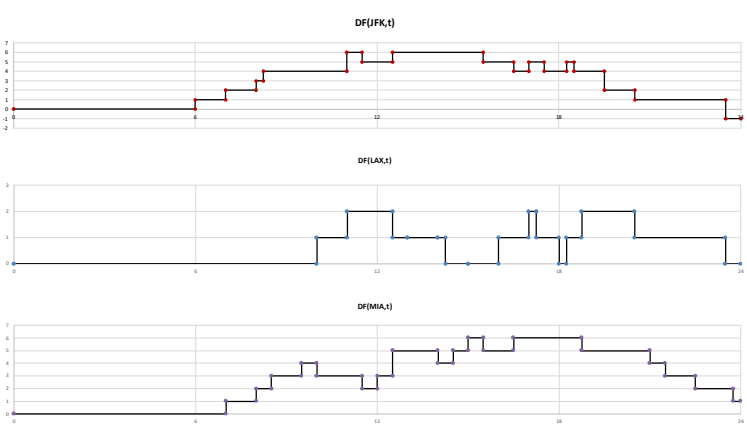
Flight No.	Departure + Time	Destination + Time
1	JFK6	LAX12.5
2	JFK7	MIA10
3	JFK8	LAX14.25
…	…	…
28	MIA14.5	JFK17.5
29	MIA15	LAX20.5
30	MIA16.5	JFK19.5

Step 1: Draw flight schedule graph



Step 2: Draw Deficit Function for each airport

Let us construct for each airport k its Deficit Function $DF(k,t)$, $0 \le t \le T$, which is defined for each t as the difference between the number of departures from k and the number of arrivals to k on the interval $[0, t]$.



Data Source

American Airlines has over 1,000 domestic flight routes across the country. Due to a high difficulty in incorporating all routes into the model, our team finally selects 3 major airports, which is JFK, LAX and MIA, and 30 flights that start and end at these airports.

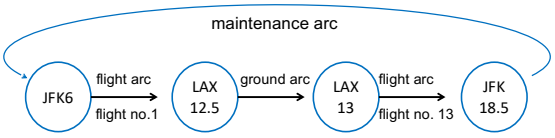
Data Preprocessing

Adjust data format and denote each flight as the following -- departure airport & departure time, destination airport & destination time, i.e., JFK8 to MIA11, meaning a flight that takes off from JFK at 8 a.m. and lands at MIA at 11 a.m.

Adjust flight time to Eastern Standard Time. Hence, flights from/to LAX are 3 hours later.

Solution to Maximal Profits

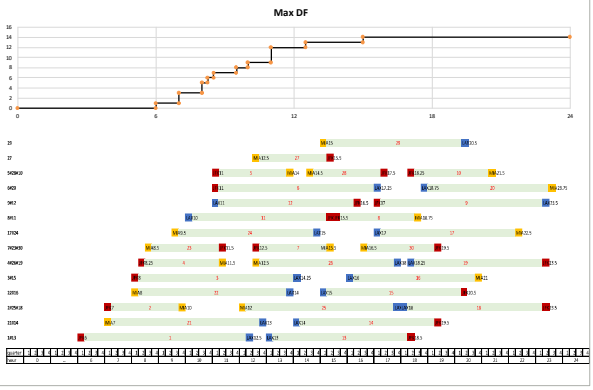
The network flow model consists of flight arc, ground arc and maintenance arc. Every two arcs are connected by one node, which determines the place and time of departure and destination. For each arc, find its revenue or cost. Then set flow balance of each node to 0; this is to make sure each aircraft passes through its arcs and nodes in a period of 24 hours.



Step 3: Draw the total DF and find the minimal number of aircrafts

The minimum number of aircraft required to service a set of flights equals the sum of the max values of n Deficit Functions. Here m represents the minimum fleet size.

$$m = \sum Max\{DF(k,t) | t \in [0,T]\}$$



Predicting Researcher Performance in the Medical Marijuana Field

Xiangxiang He, Hsuan-Ju Lin, Haitao Liu
Instructor: Dr. Chris Asakiewicz

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction

Background:
Public perception is shifting towards a wider acceptance of marijuana, and studies suggest that marijuana use among American adults is on the rise. Also, recent years many US states enact measures to decriminalize cannabis use and approve the medical use of cannabis. These phenomena draw our attention to marijuana.

Purpose:
The main goal of this project is to predict the potential researchers that might become the leading individuals in medical marijuana research using 3 year period data. We want to predict the performance of the 6th year for each author.

Feature Engineering

Citation score: The times of publication are cited by other research

Institution

- Institution score= The citation score of this institution in medical marijuana field * 0.7 + The citation score of this institution in all fields* 0.3
- If one author has multiple institutions, we assign the highest institution score to this author.

The weighted number of publications

- Journal score = Journal Impact Factor
- The weighted number of publications= The number of the publication in journal A * journal A score + The number of the publication in journal B * journal B score+....

The citation score of the author

- Time range: The 3 years since their first published the research related to medical marijuana.
- Citation score in medical marijuana field * 0.5 + Citation score in all science fields * 0.5

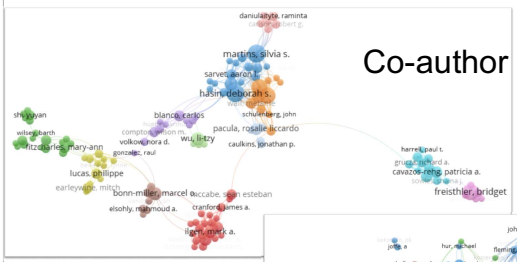
H-index of the author in medical marijuana industry.

Betweenness centrality from co-author network

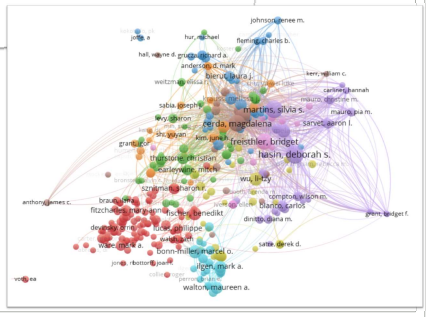
Closeness centrality from co-author network

PageRank from co-author network

Social Network



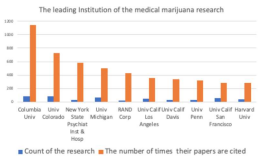
Co-author network

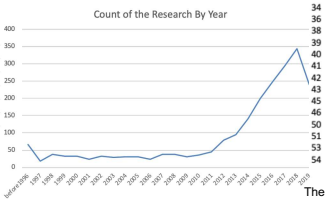


Citation Network by Author

Data Understanding

- Data Resource: Web of Science
- Research Topic: Medical Marijuana
- The total number records: 2221
- The total number authors: 6722





Rank	Word	Count
11	STATES	109
12	PAIN	103
15	ALCOHOL	95
16	RISK	95
17	LEGALIZATION	94
18	ADULTS	93
19	LAWS	85
20	CARE	83
22	ABUSE	75
23	ADOLESCENTS	73
28	STUDENTS	70
29	PRESCRIPTION	69
31	DRUGS	68
32	CHRONIC	66
34	DISORDERS	64
36	OPIOID	62
38	YOUNG	60
39	ADOLESCENT	58
40	HIV	56
41	RECREATIONAL	55
42	IMPACT	54
43	POLICY	54
45	PATTERNS	50
46	EPILEPSY	49
50	COLORADO	45
51	DISORDER	45
53	YOUTH	43
54	CANCER	42

The most frequent words in the data

Model

Input Data Sample

According to the distribution of the population, we separated the data into 3 groups by the citation score. We selected the sub-sample from these 3 groups to form the sample.(Around 10% of the whole data)

Method

1. Multiple linear regression model

To predict the performance of the author by the citation score of the 6th year

2. Logistic Regression model

If the citation score is greater than -0.35, class 1
If the citation score is greater than -0.35, class 0

To predict the performance of the author by the probability of class 1

Result

We predicted the 10 authors performance in their 6th year, the scores we gave them are shown as follow:

Name	Score (By Linear)	Score (By Logistic)
26	0.62	0.33
61	0.38	-0.33
2	0.69	0.012
62	0.42	-0.09
85	0.38	-0.25
48	0.37	-0.34
16	0.76	0.62
99	0.37	-0.38
56	0.45	-0.19

Dependent Variable	Coefficient of regression function
H-index	0.167
Citation3	0.25
Journal	0.128
Institution	0.157
betweenes s_centrality	-0.098
closness_ce ntrality	0.036
pageranks	-0.06
Interception	-0.018

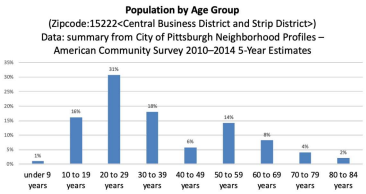
Marketing Analysis on The PA Market

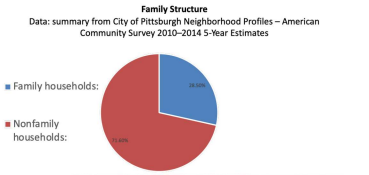
Haitao Liu, Hsuan-Ju Lin, Suprajah Suresh, Ying Lu
Instructor: Khasha Dehnad

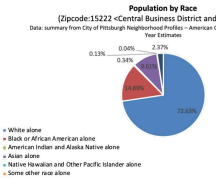
STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction

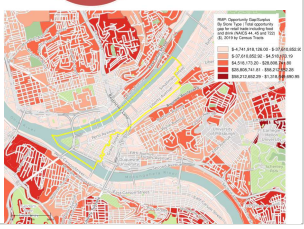
PA market restaurant is located in the heart of the Strip District in Pennsylvania, which is next to Central Business District, and it has 5 kitchen stations and drinking places. Our goal is to provide the manager with marketing strategies to increase its lunch revenue and attract it's customers to stay longer at night.





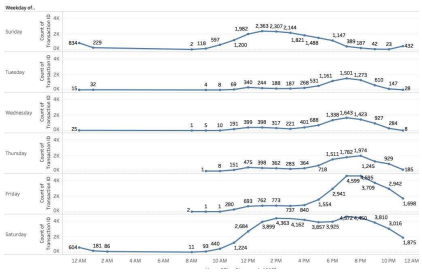


PA market is inside the yellow line region, where opportunity gap is negative. In this case, local retailers are attracting consumers of other trade area into their stores.

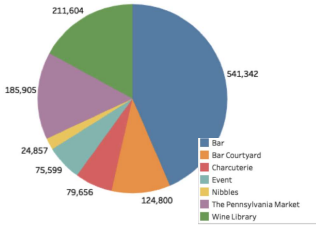


Sales Analysis

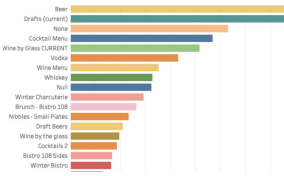
The number of items sold by hour



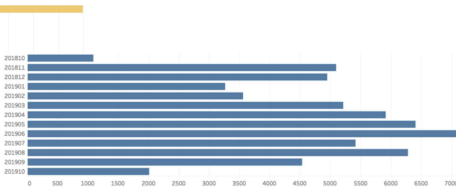
The net sales by location



Top selling categories

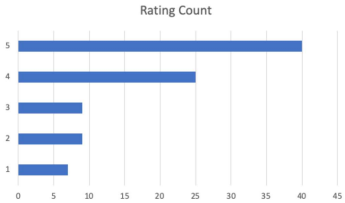


The number of items sold by hour



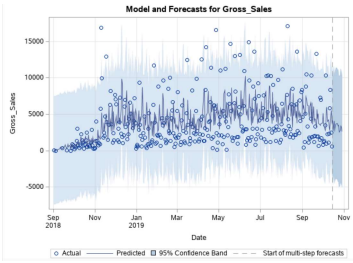
Review Analysis

1. Data source: yelp reviews + google map reviews
2. Total records: 91
3. model: Topic Extraction
   Sentiment Analysis
4. Disadvantage:
   - Long waiting time
   - Bad service of an employee
   - Mediocre food, overpriced food
   - Dirty dishes
5. Advantage:
   - Good environment
   - Special decoration
   - Fun Vibe

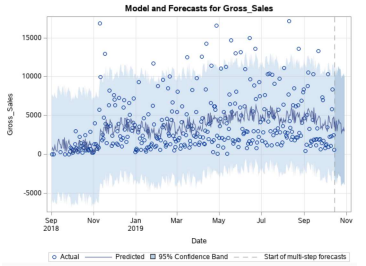


Sales Forecasting

- Time Range: (By Week)
  09/02/2018 - 10/14/2019
- Forecasting Time:
  The following 30 days
- Methods:
  1. Winter multiplicative model
  2. Additive Seasonal Exponential Smoothing
- Result:



Date	PREDICT
15OCT2019	4209.975225
16OCT2019	2680.7596329
17OCT2019	4119.971733
18OCT2019	2891.779518
19OCT2019	3379.1854916
20OCT2019	3628.5389194
21OCT2019	3733.5212386
22OCT2019	3669.5420204
23OCT2019	3362.089286
24OCT2019	3452.1506487
25OCT2019	2585.9367841
26OCT2019	2762.7339157
27OCT2019	3411.0730039
28OCT2019	3238.100495
29OCT2019	2615.8604473



Date	PREDICT
15OCT2019	4810.1038554
16OCT2019	3006.0195813
17OCT2019	4357.8706204
18OCT2019	3018.5979773
19OCT2019	3721.037803
20OCT2019	3935.506385
21OCT2019	4437.9052351
22OCT2019	4038.87309
23OCT2019	4082.5615166
24OCT2019	3939.2683054
25OCT2019	2789.549639
26OCT2019	3236.3884897
27OCT2019	3027.0563394
28OCT2019	3433.8103628
29OCT2019	2999.3112072

Marketing Solution

- It could target on majority which is between 20 to 40 years old, white and non-family households. Event: Bachelor Night, Pink Wednesday Night for ladies
- We can attract lunch crowd by offering the lunch special combo. The lunch special items can be selected by the top-selling items combined with the items that we want to promote
- The number of the customers on Tuesday is relatively lower than the other days. So we can arrange events on Tuesday night.
- Provide service training on all the stuffs to improve the service
- Give the promotion to loyal customers (refer code)

Future Work

- Conjoint analysis
  (To select items for food combo menu)
- Competitors research
  (It could fill the supply gap by analyzing surrounding restaurants)

Da Vinci – The Human Library

Amey Thombre, Athiban P
Instructor: Rong Emily Liu

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction

- Da Vinci is designed to provide its users with very personalized and unique recommendations based on a brief interaction with the user. It understands the user over the course of time.
- The concept of Human Libraries is on the rise across the globe. Da Vinci listens & collects stories, journeys and experiences of its users, stores it and narrates it to millions of users across the globe.

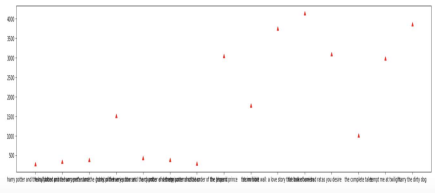
The Recommendation Engine

Da Vinci's book recommendation engine has been designed by harnessing the power of Web Mining coupled with 3 powerful recommendation mechanisms.

1. Content based Recommendation

Da Vinci Computes Pairwise Similarity Scores for all the books using the cosine similarity and sorts them based on the scores.

```
Please enter the name of the Book in Lower Case: harry potter and the deathly hallows
272                 harry potter and the half-blood prince
349                harry potter and the sorcerer's stone
411                 harry potter and the goblet of fire
2512                      harry potter series box set
455               harry potter and the chamber of secrets
403              harry potter and the prisoner of azkaban
301             harry potter and the order of the phoenix
5193                           the leopard prince
3086                                     storm front
6802    the invisible wall: a love story that broke ba...
7423                          the tale of one bad rat
5250                                   as you desire
1303                             the complete tales
5126                             tempt me at twilight
7090                             harry the dirty dog
Name: Book Name, dtype: object
```



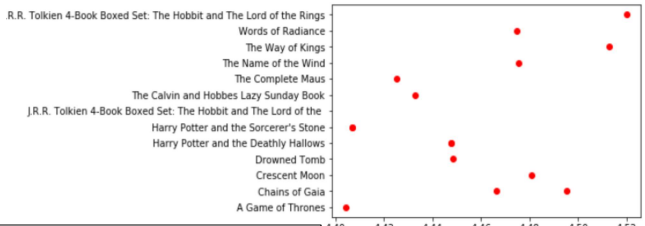
2. Collaborative Filtering

Da Vinci calculates similarity between items calculated using peoples ratings of those items

Book Name	1	2	3	4	5
A Court of Mist and Fury	Wild Man	Come Away with Me	Going Too Far	Glass Houses	Knight
A Walk to Remember	Wild Man	Eclipse	Lover Eternal	Darkfever	Fifty Shades Trilogy
Acheron	Lover Unbound	The Dex Files	Rock Chick Revenge	Rock Chick	The Last Olympian
All In	Lover Unbound	City of Glass	Release Me	My Life Next Doc	Seduction and Snack
Along for the Ride	A Court of Mist and Fury	Sweet Evil	Darkhouse	Fangirl	Reaper's Property
Angelfall	A Court of Mist and Fury	Sins & Needles	Going Too Far	Harry Potter and	The Marriage Barga
Anna and the French Kiss	Bloodlines	The Sea of Tranquilit	Reflected in You	Clockwork Angel	The Hunger Games
Archer's Voice	Lover Unbound	Knight	The DUFF: Designated Ugly	Rock Chick	Release Me
Backstage Pass	Lover Unbound	Chain Reaction	Fangirl	City of Glass	Into the Hollow
Bared to You	The Boy Who Sneaks in M	If I Stay	Bloodlines	The Hunger Game	Catching Fire
Beautiful Bastard	Lover Unbound	Chain Reaction	Fangirl	City of Glass	Into the Hollow
Beautiful Creatures	Wild Man	Perfection	A Walk to Remember	City of Ashes	City of Glass
Beautiful Disaster	If I Stay	Reflected in You	Love Unscripted	The Hunger Game	The Boy Who Sneak
Beautiful Stranger	Rock Hard	Mockingjay	Harry Potter and the Death	Fallen Crest High	Eleanor & Park
Because of Low	The Truth About Forever	On the Island	Naked	My Favorite Mist	Because of Low
Beneath This Man	A Court of Mist and Fury	The DUFF: Designate	Fallen Crest High	Fifty Shades Trilo	From Ashes
Bloodlines	The Coincidence of Calle	Anna and the French	Slammed	The Mighty Storm	The Mortal Instrume
Breaking Dawn	Rules of Attraction	The Iron King	Dark Lover	That Boy	Just for Now
Breathe	Graceling	Pride and Prejudice	The Sea of Tranquility	The Edge of Neve	Rock Me
Catching Fire	If I Stay	The Sea of Tranquilit	The Hunger Games	The Fault in Our S	The Edge of Never
Chain Reaction	Lover Unbound	City of Glass	Release Me	My Life Next Doc	Seduction and Snack
Charade	Clockwork Princess	The Notebook	Lover Avenged	Lover Eternal	The Gamble
Checkmate	Wild Man	Going Too Far	Sweet Evil	Seduction and Sn	Reaper's Property
City of Ashes	Lover Unbound	Lay Man	Ten Tiny Breaths	Sweet Evil	Sins & Needles

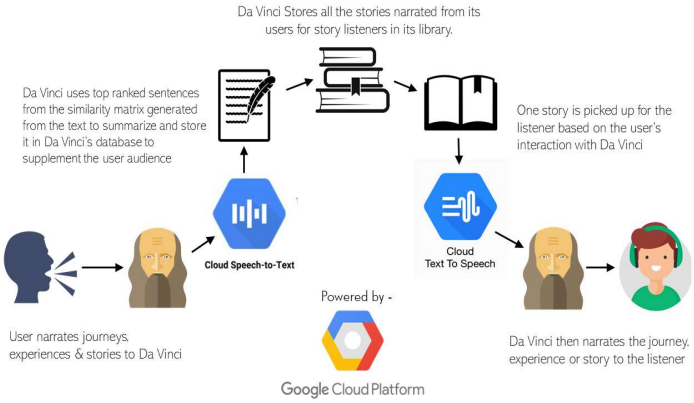
Fig: Flowchart of how our collaborative filtering system works

In addition to the other mechanisms, Da Vinci uses weighted average metric that is derived by normalizing the ratings.



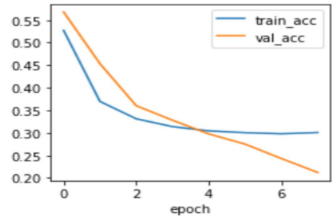
$$\text{weighted rating } (WR) = (v/(v+m))R + (m/(v+m))C$$

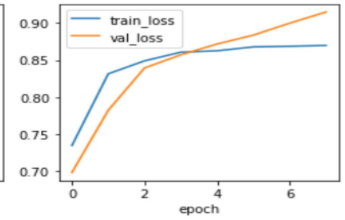
The Human Library Engine



Da Vinci's Human library engine categorizes user stories using emotion analysis as well as genre-based multiclass classification model built using Convolutional Neural Networks & Machine learning algorithms.

	precision	recall	f1-score	support
biography	0.86	0.89	0.88	28
children	0.75	0.56	0.64	16
fantasy	0.86	0.69	0.76	35
gay	0.86	0.78	0.82	23
graphic-novels	0.90	1.00	0.95	27
historical-fiction	0.77	0.90	0.83	30
historical-romance	0.93	0.96	0.94	26
history	0.81	0.68	0.74	25
horror	0.60	0.35	0.44	17
love	1.00	0.56	0.72	16
mystery	0.89	0.86	0.87	28
nonfiction	0.52	0.70	0.60	33
paranormal	0.72	0.82	0.77	22
romance	0.60	0.71	0.65	21
science-fiction	0.76	0.84	0.80	19
women	0.50	0.46	0.48	13
young-adult	0.54	0.61	0.57	23
avg / total	0.77	0.75	0.75	402





Current accuracy attained* = **86.98 %**

**Working towards further optimizing the model*

```
# predict
pred=model.predict(X_test)
# evaluate the model
scores = model.evaluate(X_test, Y_test, verbose=0)
print("%s: %.2f%%" % (model.metrics_names[1], scores[1]*100))
```

```
acc: 86.98%
```

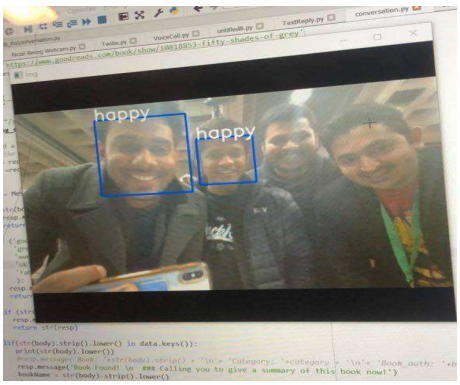
Feature Derivation & Emotion Analysis

Da Vinci derives emotions from the items based on its features.

	Name	Polarity	Subjectivity	Compound	Negative	Neutral	Positive	Emotion
0	Beautiful Disaster (Beautiful, #1)	0.01224042	0.5694315	-1	0.157	0.71	0.133	Love,Joy
1	Fifty Shades of Grey (Fifty Shades, #1)	0.03458903	0.5233159	-0.9954	0.135	0.728	0.137	Love,Joy
2	Hopeless (Hopeless, #1)	0.18532179	0.6000859	1	0.108	0.681	0.211	Love,Joy
3	The Fault in Our Stars	0.08053422	0.5543052	-0.9991	0.158	0.69	0.152	Love,Joy
4	Divergent (Divergent, #1)	0.09827393	0.5306314	1	0.104	0.729	0.168	Love,Joy
5	Slammed (Slammed, #1)	0.15986865	0.5706714	1	0.101	0.698	0.201	Love,Joy
6	Effortless (Thoughtless, #2)	0.20623599	0.5756692	1	0.089	0.69	0.22	Love,Joy
7	Easy (Contours of the Heart, #1)	0.18577932	0.5672696	1	0.116	0.691	0.193	Love,Joy
8	Bared to You (Crossfire, #1)	0.12701953	0.5566326	1	0.1	0.739	0.16	Love,Joy
9	Thoughtless (Thoughtless, #1)	0.06450958	0.5742567	-0.9996	0.173	0.668	0.159	Love,Joy
10	The Hunger Games (The Hunger Games, #1)	0.12850926	0.5197938	1	0.117	0.735	0.148	Love,Joy
11	Fallen Too Far (Rosemary Beach, #1; Too Far, #1)	0.12863951	0.5586169	1	0.111	0.699	0.19	Love,Joy
12	The Edge of Never (The Edge of Never, #1)	0.14973561	0.5443326	1	0.109	0.704	0.187	Love,Joy
13	Obsidian (Lux, #1)	0.15951647	0.5966561	1	0.099	0.677	0.224	Love,Joy
14	Walking Disaster (Beautiful, #2)	0.12120702	0.5485445	1	0.129	0.71	0.161	Love,Joy
15	City of Bones (The Mortal Instruments, #1)	0.08288836	0.5441439	1	0.099	0.739	0.162	Love,Joy
16	On Dublin Street (On Dublin Street, #1)	0.15839425	0.5547096	1	0.106	0.698	0.196	Love,Joy
17	Clockwork Angel (The Infernal Devices, #1)	0.13285989	0.5526003	1	0.104	0.708	0.187	Love,Joy
18	Pride and Prejudice	0.15063587	0.5304424	1	0.096	0.727	0.176	Love,Joy
19	Anna and the French Kiss (Anna and the French Kiss, #1)	0.12795272	0.5600886	1	0.115	0.688	0.197	Love,Joy
20	Hush, Hush (Hush, Hush, #1)	0.05024653	0.5397661	-0.9998	0.145	0.718	0.137	Love,Joy
21	Vampire Academy (Vampire Academy, #1)	0.16680075	0.5756972	1	0.123	0.696	0.181	Love,Joy
22	Wallbanger (Cocktail, #1)	0.16595423	0.5997667	1	0.098	0.685	0.216	Love,Joy
23	Fifty Shades Freed (Fifty Shades, #3)	0.10452274	0.5164469	1	0.109	0.71	0.181	Love,Joy
24	Gabriel's Inferno (Gabriel's Inferno, #1)	0.13750945	0.5715671	1	0.086	0.713	0.201	Love,Joy
25	Reflected in You (Crossfire, #2)	0.13541421	0.5568131	1	0.112	0.721	0.167	Love,Joy
26	Fifty Shades Darker (Fifty Shades, #2)	0.07430943	0.4862657	1	0.111	0.727	0.162	Love,Joy

```
Book Name : Fifty Shades of Grey (Fifty Shades, #1)
Book Review text
0.034589027180215
0.523315809091973
compound: -0.9954,
neg: 0.135,
neu: 0.728,
pos: 0.137,
Love,Joy,Anger,Fear
Book Name : Hopeless (Hopeless, #1)
Book Review text
0.185321786843620 7
0.600859301544365
compound: 1.0,
neg: 0.108,
neu: 0.681,
pos: 0.211,
Love,Joy,Sadness,Fear
Book Name : The Fault in Our Stars
Book Review text
0.080534222536432 9
0.554305159916277
compound: -0.9991,
neg: 0.158,
neu: 0.69,
pos: 0.152,
Love,Joy,Anger,Sadness
Book Name : Divergent (Divergent, #1)
Book Review text
0.098273926923508 72
0.530631369460903 9
compound: 1.0,
neg: 0.104,
neu: 0.729,
pos: 0.168,
Love,Joy,Surprise,Anger
```

Da Vinci is being developed further towards analyzing real time emotions of the user. For example: Da Vinci scales the level of happiness & recommends a joyous book on the other hand, if a person is extremely sad it will work towards facilitating an uplifting experience for the user



Text Mining 10-K Filling Report
- Predict Financial Distress Using Risk Factors

Yuzhen He, Yiru Yang, Siwei Wang, Haochen Liu
Instructor: Professor Rong Liu

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

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

pandas, Numpy, OS, Regular Expression

Natural Language Processing
- Basic statistics analysis
- Text analysis
- Sentiment analysis
- Hypothesis testing
- Topic Modeling

gensim, nltk, matplotlib, seaborn, wordcloud

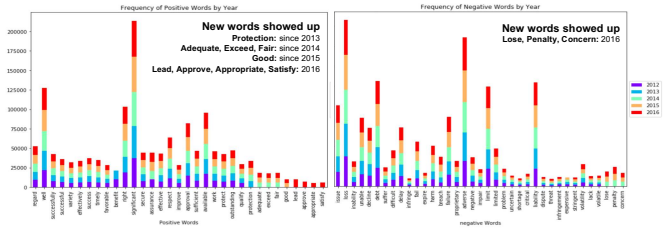
Modeling
- Training: Using previous year's data to train model
- Testing: Using data of year 2012 to 2015 to predict response variable (edf) of 2016

Prediction
- Regression
- Classification
- Random forest
- XGBoost

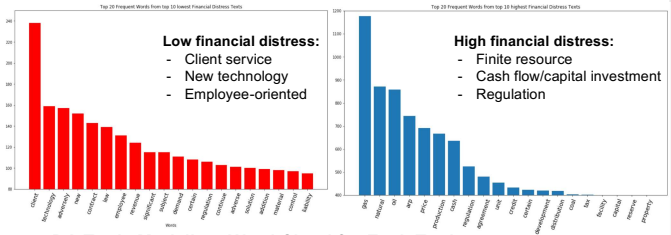
Scikit-Learn, StatsModels, RandomForestClassifier

Natural Language Process

- **Sentiment Analysis: Positive/Negative Words Across Years**



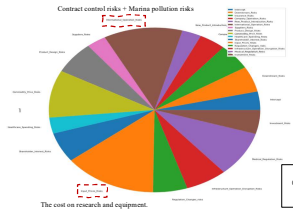
- **Hypothesis: Frequent Words VS. Financial Distress**



- **LDA Topic Modeling: Word Cloud for Each Topic**

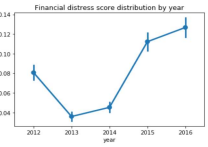
Modeling

- ***Prediction model***



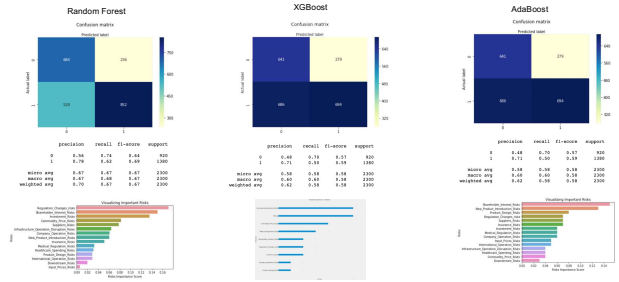
- ***Financial distress distribution***

Point estimates & confidence intervals



$$edf = -0.15381587*T1 + 0.21388551*T2 + 0.21651892*T3 + 0.17162309*T4 + 0.12585415*T5 + 0.51262567*T6 + 0.12708051*T7 + 0.27069798*T8 + 0.39627308*T9 + 0.13350712*T10 + 0.24905881*T11 + 0.60673596*T12 + 0.22353962*T13 + 0.27016040*T14 + 0.37805899*T15 + 0.20005294*Residual$$

- ***Model Performance Evaluation***
- ***Linear Regression***
- ***Classification (Random Forest, Logistic Regression, XGBoosting ,AdaBoosting)***



Conclusion & Future Work

- **Conclusion**
  - Topic model (15 topics):

0:Downstream Risks	1:Insurance Risks	2:Company Operation Risks
3:New Product Introduction Risks	4:International Operation Risks	5:Suppliers Risks
6:Product Design Risks	7:Commodity Price Risks	8:Healthcare Spending Risks
9. Shareholder's Interest Risks	10:Input Prices Risks	11:Regulation Changes risks
12:Infrastructure Disruption Risks	13:Regulation Risks	14:Investment Risks

  - After studied the correlation between negative words and edf, we found that generic sentiment analysis using word lists do 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 along 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 under consideration. For numerical dimensions, the research will consider to add the text feature we got to the original Moody's model to investigate whether we can improve the prediction accuracy of financial distress.

Prediction of Credit Card Default

Anirudh Sundararaghavan, Ramesh Balasubramani, Simran Jariwala
Instructor: Dragos Bozdog

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction

Default risk is the probability that a person will be unable to make the required payments on their debt obligation. Credit card lending is one of the major consumer lending products in the United States, representing roughly 30% of total consumer lending. Every instance of default results in a financial loss for the bank impacting the profitability, solvency and share price of the bank.

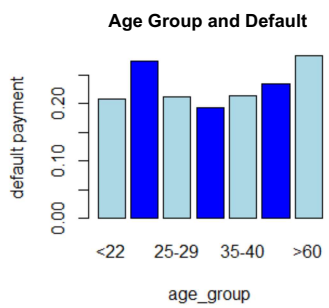
The ability to predict which customers are more likely to default on their credit cards or identifying factors that are strongest predictors of credit card default can significantly help banks protect against default risk.

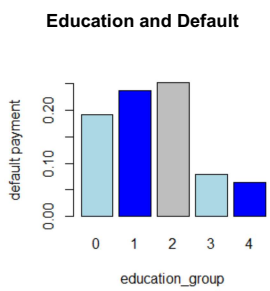
Dataset

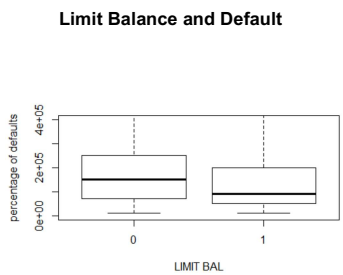
This dataset contains information on credit card statements of credit card clients in Taiwan from April 2005 to September 2005. We have a total of 30,000 rows (i.e. customer details and we have 24 features about each customer.

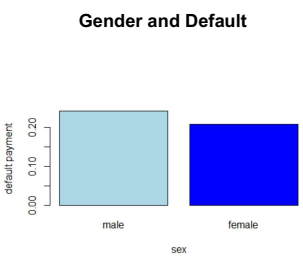
Limit Balance	*Sex*	*Marital Status*	*Education*	*Age*	*Repayment Status*	*Bill Statement*	*Prev. Payment*
Credit Limit of Customers	*Male or Female*	*Single, married or others*	*Graduate, University, High School or Others*	*Age of customers*	*Defaults by customer in prev. months*	*Amount due in prev. months*	*Amount paid in prev. months*

Exploratory Data Analysis









Data Transformation

While the dataset provides details with respect to the bill statement and amount payment of the clients. This data directly does not provide any information on the default risk of the clients.

Hence, we created a few transformed variables which categorize customers based on the trend in repayment status and proportion of bill repaid. Using R, we split the customers into the following categories based on trend in repayment and proportion paid.

Trends in repayment status and proportion

- *Majorly positive*
- *Majorly negative*
- *No Trend*
- *Mixed*
- *Positive*
- *Negative*

Conclusions

We used the dataset to fit multiple logistic regression models and used step wise regression models to identify strong predictor variables.

Based on our analyses, the following are the strongest predictors

- Latest repayment status
- Trend in repayment
- Limit Balance
- Education & Age

Results

Method Used	*Accuracy rate*	*Recall Rate*
Logistic Regression	*81.5%*	*22.82%*
Classification Tree	*83.06%*	*32.54%*

It is interesting to note that default is inversely proportional to limit balance. This highlights how banks may have higher scrutiny while issuing larger limit balances as compared to smaller limit balances. This could pose a risk of several lower limit balances accumulating to a large loss on the bank

Way Forward

This analysis gave us an indication of sources of default risk and how to identify them. We can continue to build robust predictive models using random forests and boosting algorithms. Further, importance should also be given to more detailed data collection with respect to nature of occupation of customers and credit scores.

Disaster News Classification

Ayushi Singh, Varsha R, Sparsha Hanumanthe Gowda, Shubham Shrivastava

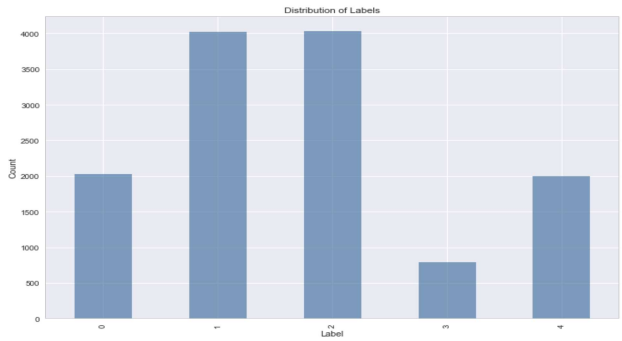
Instructor: Prof. Edward A. Stohr

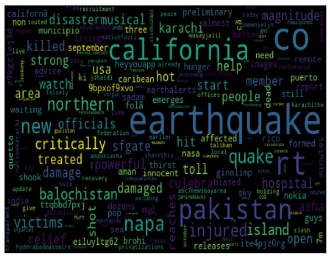
STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction:

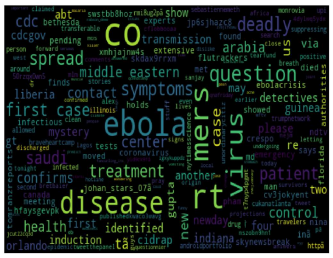
Twitter: One of the quickest sources of news and information. Tweets, twitter information feeds, are dispatched by registered users and reaching even to non-registered users. These are sometimes even ahead of traditional sources of mass news. We have developed machine learning model which detect informative disaster-related tweets. A dataset of tweets, collected from disaster tweets, was used in building the classifier models. Two machine learning algorithms, Naive Bayes and Support Vector Machine (SVM), are used to build models for the automatic classification of tweets, assessed across the metrics of accuracy, precision, recall, area under curve and measure. Experimental outcomes show that the model generated from SVM has better results than Naive Bayes. The interactions on social media, deeply distributed, decentralized and occurring in real time, provide vital breadth and immediacy of information required during emergencies. Here, tweets provide either first-person observations or bring relevant knowledge from external sources. Relevant tweets is beneficial in understanding and visualizing the scenario of affected parties. Tweets exceedingly fluctuate in terms of subject and content. The influx of tweets specifically in the event of a disaster may be overwhelming, consisting of socio-behaviors that include intensified facts search and data contagion.

Experiment:

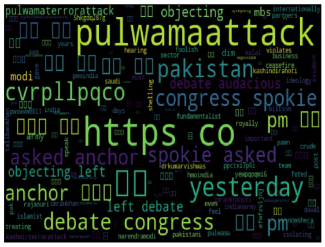




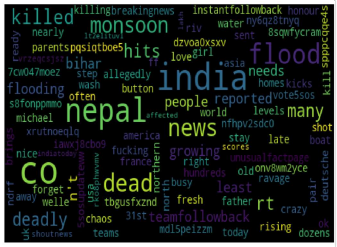
Label 1



Label 2



Label 3



Label 4

Label 0

Results:

1. Multinomial Naive Bayes:
With count vectorization
0.9408424041646948

Confusion Matrix :

```
[[375 12 20 2 5]
 [ 1 753 0 0 1]
 [ 8 0 831 0 1]
 [ 0 0 0 174 0]
 [ 40 21 17 0 316]]
```

2. Logistic regression:
With tfidf vectorization
0.9642995731470703

Confusion Matrix :

```
[[411 0 0 0 3]
 [ 5 750 0 0 0]
 [ 11 0 829 0 0]
 [ 1 0 0 173 0]
 [ 62 7 3 0 322]]
```

3. Random Forest classifier
With word2vec vectorization
0.9654519640321817

Confusion Matrix :

```
[[407 0 1 0 6]
 [ 1 753 0 0 1]
 [ 4 0 834 0 2]
 [ 0 0 0 174 0]
 [ 40 2 0 0 352]]
```

4. Artificial Neural Networks
With Text Tokenizera
0.9657610718273167

Confusion Matrix :

```
[[399  0  1  1 16]
 [  3 759  1  0 10]
 [  4  0 835  0  5]
 [  4  0  0 267  2]
 [ 38  2  3  2 335]]
```

Conclusion:

Twitter is preferred to broadcast disaster-related tweets. During disasters, subscribers use this medium to broadcast both informative and uninformative tweets where uninformative tweets outnumbered informative tweets with a 65% to 35% ratio. Although more uninformative tweets were posted, the informative tweets were rapidly sent because these were retweeted. This evidently implies that informative tweets contain vital & urgent information, providing significantly needed information for situational awareness of public and disaster response units. Disaster-related tweets can be automatically classified using Naive Bayes, Logistic Regression, Random Forest Classifier and neural networks. We see that neural networks performed better than the rest.

Future Scope:

We intend to further take this project further by converting it into a full-fledged news-reporting site which analyses data from Twitter as well as from Facebook status updates, Inshorts application.We would try to have more categories than we have now. This project would have a UI with a lot of more features where the user can also tweet or comment on the tweet along with repost and like features. Social media is consistently evolving. Therefore, a specific website only for urgent news display, may become quickly popular among the masses.

Secondly, we could use feature stacking, tuning model parameters and using ensemble models for improving accuracy of text classification. This work shows a clear behavioral difference in text classification when using sentiment-based features in addition to the usual word vector features.

Thirdly, we could use word topic models on top of tweet sentiments as features for text classification.

Fourthly, extraction of spatio-temporal information to identify different patterns in the data using story-graph can be coupled with Geographic Information Systems (GIS) in real time and location, helping disaster management and planning considerably.

Fifthly, we could expand the capabilities of crisis analytics tools like AIDR in real-time scenario and not post disaster. The ongoing work is only limited to classifying tweets based on information content (text or image), spatial and temporal characteristics of events in real space and time are completely ignored, which are vital for any disaster response planning, and thus should be looked into.

Lack of Data? See How Amazon Comments Help Music Analysis and Classification

Ziyu Ma, Fudi Liu, Xuhao Li, Zhongxiuzi Tian
Instructor: Professor Feng Mai

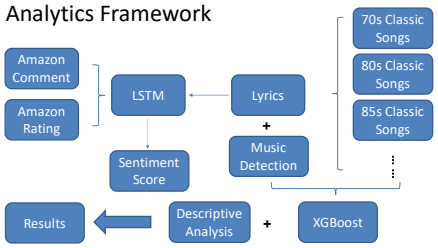
STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

Overview

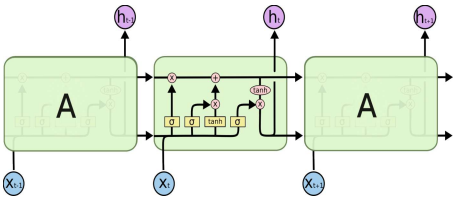
- **Purpose**: Extract insights from music website on classic songs in different decades through EDA, social network and predictive analysis
- **Key question**: Which songs are in the important position of each decades? What are differences between each decade? How to classify which song is much familiar with which decade style?
- **Findings**: Differences between decades on songs style, time, rhythm, information quantity; Good performed songs classification algorithm
- **Originality**: Instead of other papers that did music analysis and classification, our team first time compare differences in classic songs through decades and among this, we built a search and sorting model to search songs through decades, besides, music website is lack of users' data which provide lots of difficulties for sentiment analysis on songs, thus we use amazon comments data to train our sentiment classification model first and transfers to songs sentiment analysis. At all, crawled comments data from Amazon(10256 comments and ratings), songs data from NetEase Music(3500 songs with singer info, lyrics, duration, decades) and download songs from the website.
- **Keywords**: social network analysis, web analytics, text mining, transfer learning, machine learning, descriptive analysis, big data analytics

Framework & Modelling

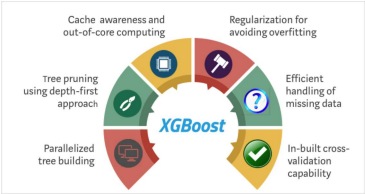
Analytics Framework



LSTM for Sentiment Analysis

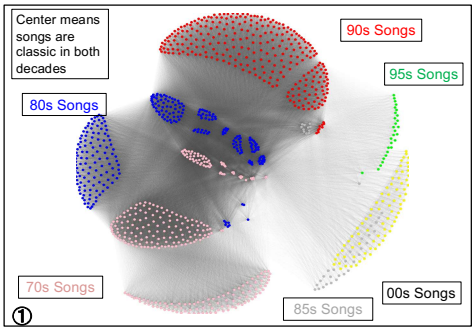


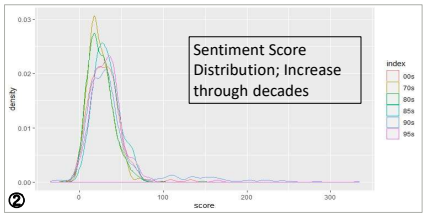
XGBoost for Songs Classification

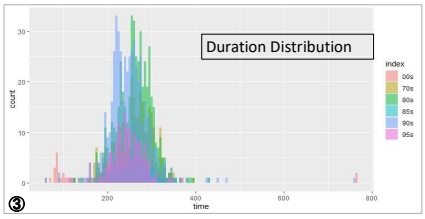


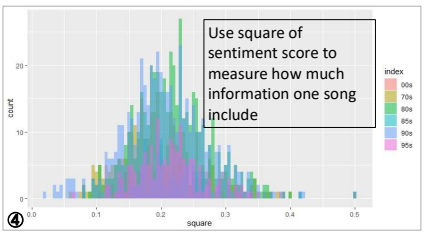
Insights

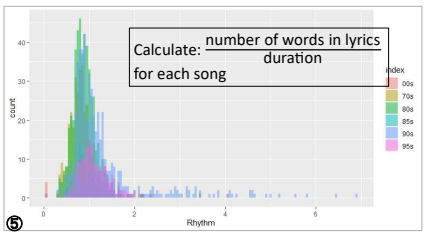
1. Social Network & Descriptive Analysis



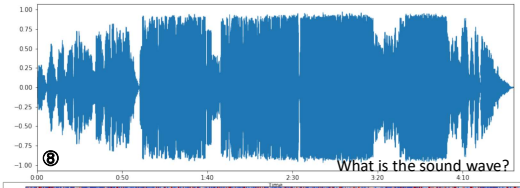


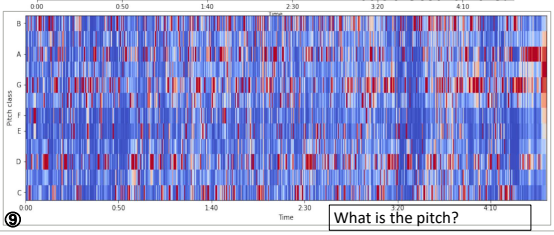


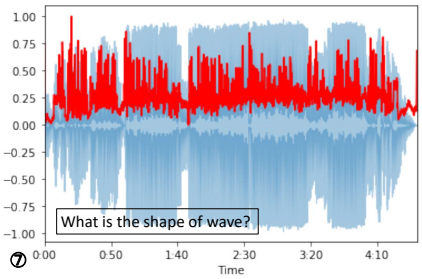


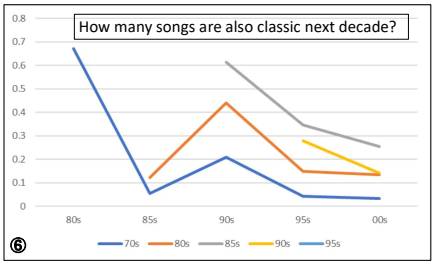


2. Music Classification Analysis









Model	Training Accuracy	Testing Accuracy
LSTM	96.27%	94.13%
XGBoost	87%	72.05%

⑩

Key Takeaways

- From ①, some songs are more important, they are bridges to other cliques, and the separation of 80s and 70s songs are most serious. From 70s to 00s, songs are increasingly positive(②), all distributions have long tails, with length increase, indicating great potential on positive songs, reflecting people situation through decades
- Songs durations are nearly no change through decades (③⑤), but lyrics per second increase, which means rhythm of songs through decades increase
- We used square of sentiment to measure how much information one song include(④), information quantity keeps increasing and 90s is a booming decade (⑥) and there is no relationship between how many songs a singer have and how many songs of this singer became classic ones
- To make classification, we need to capture music data using python. We selected sound wave, pitch, MFCC and Centroid as measures(⑦⑧⑨),combining with lyrics to do the NLP classification

Forecasting Stock Prices, Predicting Earnings Per Share and Classifying Bankruptcy

Pratik Sahoo and Swetha Kallam
Instructor: David Belanger

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

INTRODUCTION

Stock market is very dynamic in nature and it is a constant struggle to determine the stock value in near future. With the advancement of data science techniques, the challenges of predicting stock prices has been better addressed. This is our approach in forecasting stock prices, predicting earnings per share and classifying bankruptcy

TOOLS

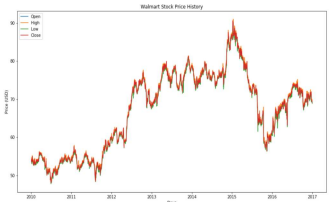
* Python (Sckit-learn, Pandas, Numpy, Matplotlib, Random Forests, Linear Regression, K-means)
* Tableau

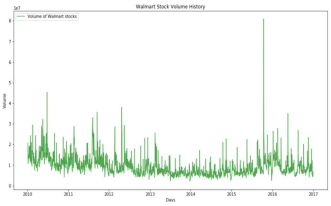
FORECASTING CLOSING STOCK PRICES

Data Understanding and Exploring

The NYSE prices dataset is a time series data set of 851264 records depicting the daily prices of stocks for 501 companies, spanning from years 2010 to 2016.

We choose to forecast closing stock price of Walmart. It had 1762 records with 7 columns.





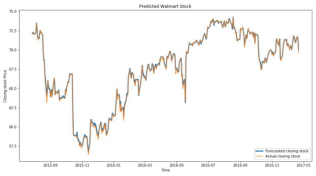
We couldn't do the regular predictions on this kind of data, as the stock market rapidly changes every day, it makes no sense if we are predicting todays stock value based on stocks of last 5 years .

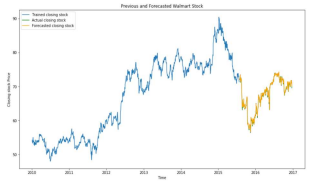
So we are predicting a particular day's closing stock value based on past five days.

date	open	close	low	high	volume
2010-01-04	53.740002	54.230000	53.669998	54.669998	20753100.0
2010-01-05	54.090000	53.689999	53.570000	54.189999	15648400.0
2010-01-06	53.500000	53.570000	53.419998	53.830002	12517200.0
2010-01-07	53.720001	53.599998	53.259998	53.750000	10662700.0
2010-01-08	53.430000	53.330002	53.020000	53.529999	11363200.0
2010-01-11	53.330002	54.209999	53.099998	54.439999	13987700.0
2010-01-12	54.000000	54.730000	53.860001	54.750000	15117000.0
2010-01-13	54.790001	55.009998	54.410000	55.200001	13290700.0
2010-01-14	54.730000	54.209999	54.160000	54.840000	13772000.0
2010-01-15	54.340000	53.680000	53.599998	54.549999	19087500.0
2010-01-19	53.910000	54.029999	53.500000	54.259998	14618500.0
2010-01-20	53.919998	53.860001	53.230000	53.939999	13308600.0
2010-01-21	54.070000	52.919998	52.840000	54.070000	16028700.0

Forecast Modelling

We used Linear Regression to train the model and it was trained with an accuracy of **83.97 %** .





Evaluating the Forecasting model

Performance Metrics of the Forecasting Model	
Metrics	Results
Mean Absolute Error	1.2513082
Mean Squared Error	3.4253527
Root Mean Squared Error	1.8507708
Mean Absolute Percentage Error	0.20973468
Accuracy	79.27 %

PREDICTING EARNINGS PER SHARE

Data Understanding and Exploring

The NYSE fundamentals data set has 1781 records with 77 feature metrics from annual SEC 10K fillings (2012-2016) .

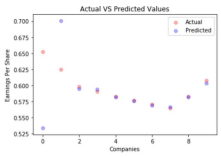
We started of with Min-Max scaling and since there were huge number of features, we had to scale them down.

Feature Selection

We used Radom Forests classifier for feature selection, since random forests considers correlated features as well, to over come this, we used co-relational matrix to trim out highly correlated features. We successfully got the 77 features down to 31 features.

Predictive Modelling

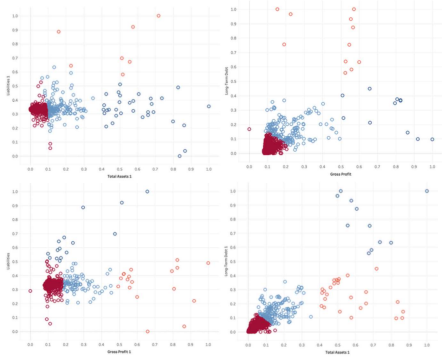
We used Random Forest Regressor for predicting earnings per share and training model was **80.94 %.**



Evaluating the Predictive model

Performance Metrics of the Forecasting Model	
Metrics	Results
Mean Absolute Error	0.02112
Mean Squared Error	0.00119
Root Mean Squared Error	0.04438
Mean Absolute Percentage Error	0.03317
Accuracy	96.66%

CLASSIFYING BANKRUPTCY



Using K Means Clustering, we classified 2 companies F and VZ that might go bankrupt in near future, using features like Total Assets, Gross Profit, Net Income, Total Equity, Long Term Debt.

CONCLUSIONS AND FUTURE WORK

* Linear Regression in our dataset has performed well for forecasting the shares. Nonetheless, we would like to delve into LSTM.
* Though, our predictive model using random forests has performed extremely well in predicting earnings per share in our scenario, it would be interesting to see how it would perform if we have enough data for each company.
* Using K-Means Clustering we classified companies that might go bankrupt in future as per the last 4 years trend, nonetheless the classification can be improved if we have more data for the companies, which would give us better results.

Australia weather analysis for popular tourist city

Yilin Cao, Liqi Cheng, Xuhao Li, Xiangyu Tang
Instructor: Somayeh Moazeni

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction

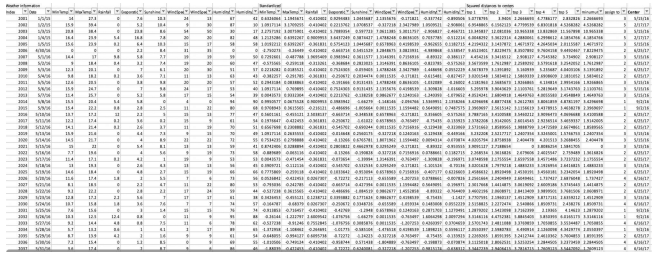
We use the dataset "rain in Australia". This dataset contains daily weather observations from numerous Australian weather stations including minimum temperature, maximum temperature, rainfall, evaporation, sunshine, windspeed at 9 am and 3 pm, and the humidity at 9 am and 3 pm. We set an ideal weather when the temperature is 24 ° C and the humidity is 40% to 70% and choose the group which data is closest to the ideal weather. Then reach the month which is most comfortable for the human and for travel.

Experiment

$$x_n = \frac{Temp - \frac{\sum_{n=1}^{n} Temp(n)}{n}}{stdTemp}, n\ stands\ for\ day$$

$$y_n = (X_{minTemp} + X_{maxTemp} + X_{rainFall} + X_{Evaporation} + X_{Sunshine} + X_{windSpeed} + X_{Humidity})$$

$$\sqrt{\sum_{0}^{n} (x_n - y_n)}$$



Results

Cluster centers and standardized values

	Column offs	12	13	14	15	16	17	18	19	20
Index	Date	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindSpeed9	WindSpeed3	Humidity9ar	Humidity3pm
1238	8/26/15	-0.680269	-0.748674	0.105513	-0.840138	-0.214738	-0.304294	-0.84634	0.836476	0.319992
1392	1/27/16	1.127941	0.249613	-0.14928	1.467471	-1.729827	0.267419	-0.579431	-0.08163	0.569977
1637	9/28/16	-1.127003	-0.515033	-0.336128	0.610359	1.117496	0.267419	-0.312522	-1.393211	-0.742444
1528	6/11/16	-1.148276	-0.833635	-0.336128	-0.708275	0.412196	1.267916	-0.045613	-0.934158	-1.242414
1298	10/25/15	-0.06335	0.249613	-0.336128	0.676291	0.804029	-1.161862	-0.045613	-0.343947	0.319992
Sum of dista	2095.554									

MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindSpeed9	WindSpeed3	Humidity9ar	Humidity3pm
20.240769	25.484615	4.3092308	8.2461538	2.6815385	16.423077	19.853846	71.738462	64.723077
10.621212	19.894697	1.4878788	3.7469697	8.4734848	22.106061	21.507576	55.856061	39.901515
13.236264	23.920879	0.4153846	6.8747253	10.430769	17.417582	19.824176	47.274725	34.307692
13.093571	20.803214	8.8707143	2.7364286	4.9464286	15.5	15.307143	80.435714	62.675
18.804745	27.655474	1.1364964	6.9076642	9.7021898	10.014599	22.018248	63.751825	52.981752

cluster centers and standardized values

	column offse	12	13	14	15	16	17	18	19	20
index	cluster cente	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindSpeed9	WindSpeed3	Humidity9an	Humidity3pm
2131	9/2/16	-0.4305843	-0.8144584	-0.3132611	-0.72272	-1.5839352	-0.3272179	-0.1718214	0.70377855	1.06400046
2427	6/25/17	-0.6722185	-0.9287932	-0.4104016	-0.486696	-0.6639588	0.32952492	0.41985395	-0.3377422	-0.0078257
2021	5/15/16	0.87424062	0.32888942	-0.4104016	0.2803822	0.46629782	0.32952492	-0.1718214	-0.8932199	-0.9535547
2418	6/16/17	-1.155487	-0.4714541	-0.4104016	-1.0177501	-0.4011085	-0.3272179	-1.2072533	0.9815174	0.55961168
2299	2/17/17	0.5601161	0.90056334	-0.4104016	0.87044233	0.93942852	-0.6555893	-0.4676591	-0.2683074	-0.7013603
Sum of distar	935.178817									

MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindSpeed9	WindSpeed3	Humidity9an	Humidity3pm
10.9652778	16.0791667	3.46944444	2.66666667	1.30138889	8.34722222	11.3472222	84.5416667	75.5277778
9.36036036	16.3189189	2.67927928	3.31351351	5.54234234	14.4324324	17.9189189	68.5585586	55.4684685
14.4142857	24.177381	0.61666667	6.76428571	7.93928571	15.8452381	17.4880952	57.0952381	44.8095238
7.22461538	16.0538462	0.85230769	1.50461538	6.21692308	6.98461538	7.76923077	85.0307692	61
14.3569892	26.4677419	0.72258065	7.20215054	9.80107527	7.2688172	13.2365591	70.1290323	48.6236559

Cluster centers and standardized values

	Column offse	10	11	12	13	14	15	16
Index	Date	MinTemp	MaxTemp	Rainfall	WindSpeed9	WindSpeed3	Humidity9ar	Humidity3pm
3297	11/7/15	0.56446461	0.56970942	-0.0185669	-0.3404069	-0.7920813	0.41510014	0.4077332
3559	8/15/16	-1.0885467	-1.0327149	-0.2773742	1.3063908	1.36417798	-0.1704024	-0.2346999
3859	6/25/17	-1.187234	-1.0618499	-0.2773742	-0.957956	-0.5960577	-0.1704024	-0.0594908
3307	11/18/15	-0.2003615	0.45316947	-0.2773742	-0.3404069	0.48207193	-0.8144551	-0.4099089
3440	4/13/16	0.44110555	0.16181959	-0.0670933	1.20346595	0.58008371	-0.0533019	0.05731517
Sum of dista	1575.55687							

MinTemp	MaxTemp	Rainfall	WindSpeed9	WindSpeed3	Humidity9ar	Humidity3pm
12.8665	22.54467	1.070051	12.03553	18.39594	65.49239	56.73096
20.16126	27.59948	4.889005	28.70681	34.8534	70.71204	68.89005
14.052	21.75867	2.224	29.70667	34.24	64.30667	60.97333
20.80584	28.1786	6.02179	13.76654	18.99222	79.37743	75.4786
17.46906	27.23525	0.625899	15.15108	27.44604	54.41007	54.28777

Conclusion

In Sydney, the best month is from January to March, The temperature from 20° C to 25 ° C is concentrated in these period. The humidity in these period is about 50%. As well as there is no more rainfall and strong wind.

In Melbourne, the best month is from January , The temperature from 14° C to 26° C is concentrated in these period. The humidity in these period is about 50% - 70%. As well as there is no more rainfall and strong wind.

Spotify: Song Genre Prediction

Author: Shiv Pratap Singh Tomar
Instructor: Dr. Mahmoud Daneshmand

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

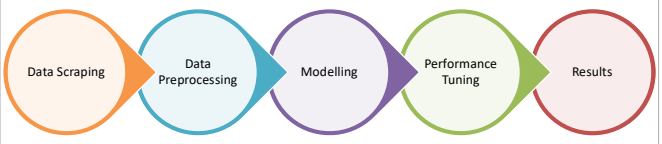
Introduction

Project Objective:

Song genre prediction through audio features of different measures.

Extracting real-life data from Spotify and implementing various machine learning algorithms to predict the 8 types of genre classes.

Project Flow:



Data Collection and Understanding

Web Scraping:

Scraped data using the Spotify API. Connected Spotify through python library known as "Spotipy". Added playlists of 8 types of genres and scraped around 8000 songs. Extracted audio features of all the songs from the different playlists.

Extracted the songs from the playlist in a list and further extracted the features in a .csv format through the use of pandas library in python.

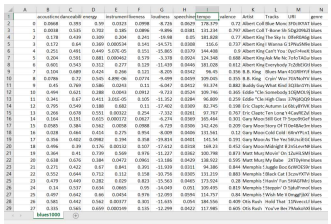
8 Genres: Jazz, Blues, Pop, Country, EDM, HipHop, Rock, R&B

9 Song Features extracted: Acousticness, Danceability, Valence, Energy, Speechiness, Liveness, Loudness, Instrumentalness, Tempo

Data Understanding:

Rows: 8,000, Columns: 9

Used pandas library to clean the 0 and 'NA' values in the data. Conversion of string classes in the genre column to single numeric value class.

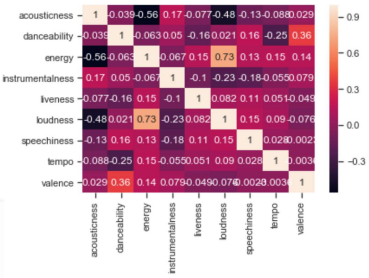


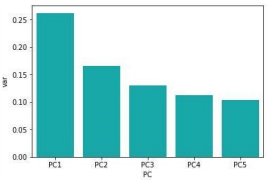
Removed unwanted columns like: Artist, URI, Tracks

Data Pre-Processing:

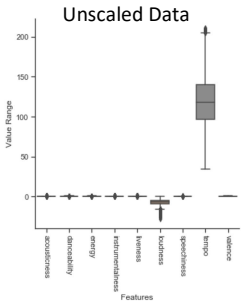
Feature Selection by PCA:

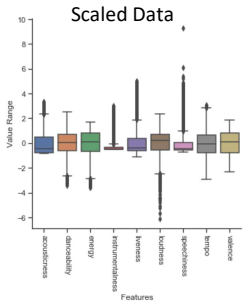
PCA was applied to reduce the number of parameters required to predict the target variable genre.





Data Scaling: Outliers were removed and the range of data was standardized for efficient training of the model.





Modelling

Supervised Learning:

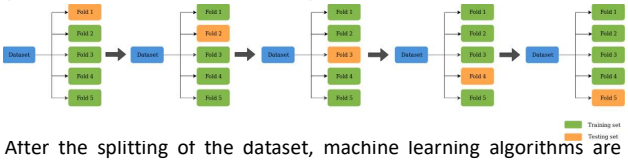
Supervised learning algorithms like Random Forest, Decision Tree, Naïve Bayes, Multi-Level Perceptron, K-Nearest Neighbors and Gradient Boosting are implemented. Sci-Kit learn library is used for applying all the deep learning algorithms.

Data Splitting:

Before feeding the data to the model, the dataset is divided into partitions of training and testing dataset. Two methods were used to split the data and the latter one was found to be more efficient.

Two partition splitting: The data was divided in 70:30 proportions, 70% for the training and 30% for the testing. Only drawback with this method was that, what if the values in the partitions are not spread evenly and the results might come biased.

K-Fold Cross Validation: It is a resampling procedure used for better training of the model for unseen data. The data is divided into K folds and the K-1 folds are used for training. In this way each of the folds get a chance to become the testing set at least once.



After the splitting of the dataset, machine learning algorithms are applied and selected based on the best accuracy. But there are some hyperparameters related to the model themselves which cannot be estimated from the data. Performance tuning was done using Gridsearch framework to tune those parameters.

Performance Tuning:

Tuning the hyperparameters of machine learning model provides us with the most accurate predictions. In this case, tuning the hyperparameters of Random Forest resulted in the highest accuracy.

```
from sklearn.ensemble import RandomForestClassifier
clf=RandomForestClassifier()
grid_param = {
    'n_estimators': [10,100,300,500,800,1000],
    'max_depth': [1,3,5,8,10,15,20,25],
    'criterion': ['gini','entropy']
}
gd_sr = GridSearchCV(estimator=clf,
                     param_grid=grid_param,
                     scoring='accuracy',
                     cv=5,
                     n_jobs=-1)
gd_sr.fit(X_train, y_train)
```

```
In [3]: runfile('D:/Stevens/Sem 2/MIS 637/Project/SpotifyClass/cvgsfinal.py', wdir='D:/
Stevens/Sem 2/MIS 637/Project/SpotifyClass')
Best Hyperparameters: {'criterion': 'gini', 'max_depth': 25, 'n_estimators': 1000}
```

Conclusion & Future Work

- What I learned from this project: how data when fed to a machine learning model can give us such fruitful insights and predictions.
- Genre of different songs can be predicted by analyzing the audio features of songs, without even listening to it.
- After comparing all the machine learning algorithms, Random Forest Classification tops the list by achieving the highest accuracy of prediction.
- Furthermore, I would like to create a model which will be integrated with the lyrics of all the songs, providing with textual analysis and extra parameters for the model to train.
- Using textual analysis with audio features will increase the accuracy of the prediction model even more.

Rotten Tomatoes: Movie Genre Prediction

Authors: Shiv Pratap Singh Tomar, Aayush Kalla, Nidhi Gamale
Instructor: Dr. Theodoros Lappas

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction

Project Objective:

Movie genre prediction based on attributes related to the movie's cast, director, studio, description, writer, etc. Extracted real-life data from Rotten Tomatoes and implemented 15 types of supervised learning and unsupervised learning algorithms for prediction.

Data Collection and Understanding

Web Scraping:

Data was scraped from the Rotten Tomatoes website. Used beautiful soup library in python to scrape the data. Since the website required to press a load more button after loading only 32 movie links, we used to reach at the bottom of the webpage of every genre. The links of different genre of movies was extracted and stored in a notepad file through python.

Each of the links were then opened and useful attributes were extracted and stored in .csv file through python.

Data Understanding:

Rows: 30,000 movie records, Columns: 20 attributes

12 Genres: Action, Horror, Classics, Comedy, Documentary, Drama, Sci-Fi, Art&Foreign, Mystery, Romance, Kids&Family, Animation

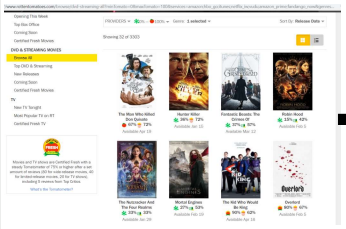
20 Movie attributes: Cast, Director, Writer, Studio, Runtime, etc.

There were many Null values, 'NA' values, missing values in the data. Pre-processing was needed before feeding this data for modelling.

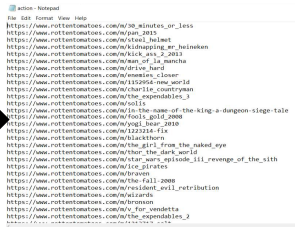
Data Pre-Processing:

Using Pandas library in python, we cleaned the raw data and removed the redundancies and the unwanted data.

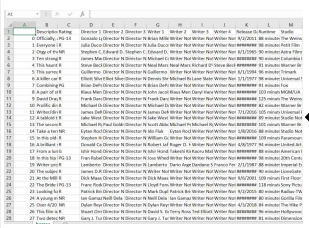
Steps Involved:



1. Scraping links of each movie by searching the title in HTML code on Rotten Tomatoes website using Selenium and Beautiful Soup



2. Movie links collected in a notepad using python for further gathering of parameters



4. Processed / Cleaned data after removing all the redundancies and missing values

3. Raw data with missing values and 'NA' values collected in a .csv format using pandas

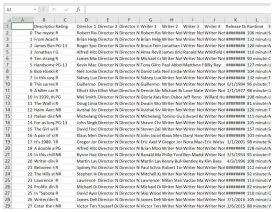
Modelling

For selection of the model for this type of prediction, we thought of using both supervised and unsupervised learning algorithms. We made changes in the data accordingly.

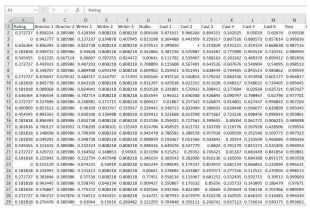
1. Supervised Learning:

Text-to-Numeric data conversion:

For better training and accuracy of the deep learning model, we converted all the text in the dataset into numeric data by mapping the unique values to their numeric indexes. After conversion to numeric data, the data was normalized for efficient and unbiased results.



Preprocessed Data



Normalized Numeric Data

Machine Learning Algorithms:

Implemented 15 types of classification and regression models and compared their accuracies for better understanding of the results.

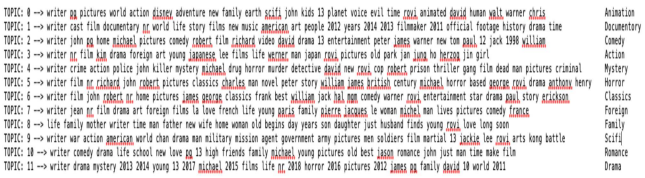
Applied Grid Search Cross Validation to implement multiple models at once and created an accuracy comparison graph.

2. Unsupervised Learning:

Textual Data Analysis:

In unsupervised learning we wanted our model to create different clusters of genres depending on the count of each term that appears in the data. Count of terms is then transformed into a Count Vector Matrix.

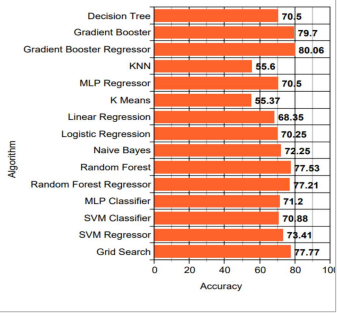
Training the model using Countvectorizer, provided by sklearn library.



Top 25 words from 12 topics output from model, clustering done manually into different genres

Conclusion & Future Work

- Both the supervised and unsupervised models were successful in predicting the genre of different movies.
- Comparing the different algorithms showed us that the Gradient Boosting Regressor achieved the highest accuracy amongst all the models.
- Further, we would like to include critic reviews for better prediction models.
- I learned, how count of words in a data can help a model train and do Predictions from it.



Using Bayesian Networks to Model Kickstarter Project Success Rates

Erika Deckter, Siddharth Kalra, Chenyang Ni, Vishnu Vijay
Instructor: German Creamer

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

Background

Kickstarter is a platform that allows artists to raise money to fund their future projects; however, project creators only get the money raised if the project successfully reaches its funding goal. But how can a project creator know whether their project is likely to get successfully funded before taking the time and effort to create a project page on the platform?

Data Description

Historical Kickstarter data was obtained from Kaggle for approximately 378,000 projects dating from 2009 to 2018.

Variables included in the Bayesian network graphs were:

- Funded (Was the project successfully funded?)
- Category
- Funding Goal (in USD)
- Campaign Length (derived from launch and deadline dates)

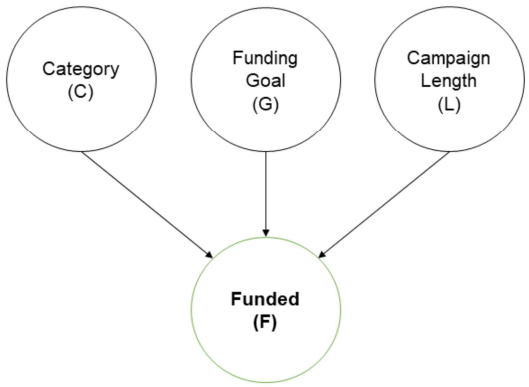
The dataset showed an overall funding success rate of 36%.

The dataset was split into a training dataset (80%) used to build the Bayesian network model and a test dataset (20%) used to validate the model.

Bayesian Network

The acyclical directed network graph represents the relationships between the various variables.

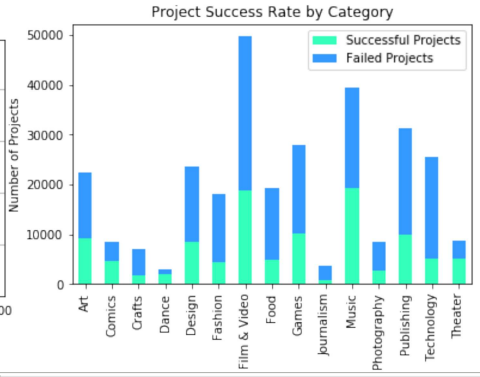
- Variables that are not directly connected in the graph are assumed to be conditionally independent.
- Numerical variables were discretized into a finite number of bins.

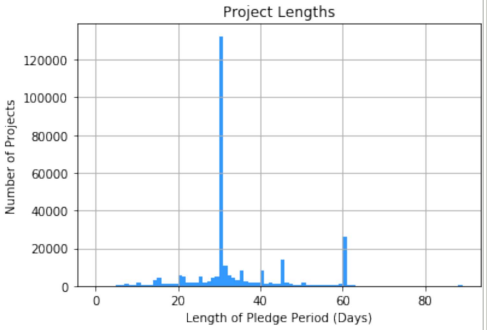


The Bayesian network can be used to calculate the probability of receiving funding for a project given a set of input conditions:

$$P(F|C,G,L) = \frac{P(F,C,G,L)}{P(C) * P(G) * P(L)}$$

Data Profiling





Results

The table below shows the probability of a project successfully receiving funding (F) given select combinations of category (C), project length (L) and funding goal (G) for both the model based on the Bayesian network and a test set of hold-out data. For these three project combinations, there is fairly good agreement between the test set and the model.

Model	P(F="Y"|C="Film & Video", L=(0,14], G=(0,10000]	P(F="Y"|C="Design", L=(30,60], G=(0,10000]	P(F="Y"|C="Comics", L=(30,60], G=(10000,20000]
Network Model	53.9%	46.7%	42.1%
Test Data	56.6%	45.8%	36.6%

Churn Warning Prediction on Commercial Bank's Customer

Wenqingyun Ma, Zhongxiuzi Tian, Yanan Liu, Siqi Jiang, Yiwei Zhang
Instructor: Professor Somayeh Moazeni

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction

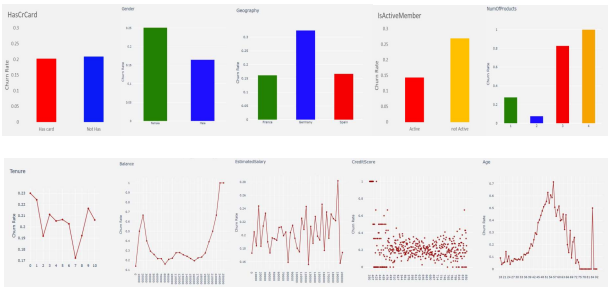
- The cost of attracting new customers can be five to six times more than holding on to an existing customers.
- Long term customers become less costly to serve, they generate higher profits, and they may also provide new referrals.
- Losing a customer usually leads to loss in profit for the bank.

Data description

- Data came from a bank. The dataset has 10000 observations.
- Attributes: RowNumber(int) CustomerId (ID) Surname(String) CreditScore(int) Geography(Country) Gender(String) Age(int) Balance(decimal) Tenure (int) NumOfProducts (int) HasCrCard (int) IsActiveMember (int) EstimatedSalary (decimal) Exited (int)

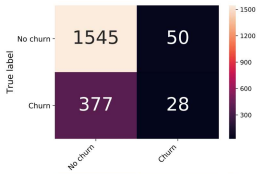
Model

- Random Forest
- Decision Tree
- Logistic Regression



single variable analysis

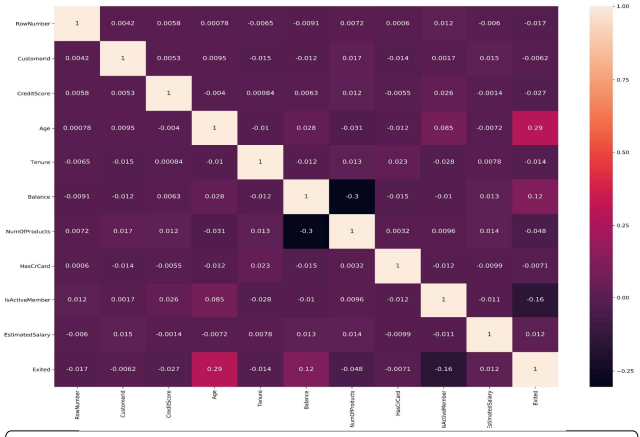
Accuracy of logistic regression classifier on test set: 0.79				
	precision	recall	f1-score	Support
0	0.80	0.97	0.88	1595
1	0.36	0.07	0.12	405
accuracy			0.79	2000
macro avg	0.58	0.52	0.50	2000
weighted avg	0.71	0.79	0.72	2000



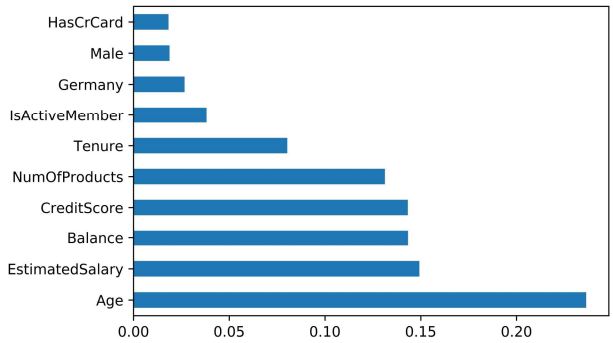
Accuracy of logistic regression classifier on test set: 0.67				
	precision	recall	f1-score	Support
0	0.67	0.65	0.66	1576
1	0.67	0.69	0.68	1610
accuracy			0.67	3186
macro avg	0.67	0.67	0.67	3186
weighted avg	0.67	0.67	0.67	3186

Unbalanced model AUROC: 0.6774286930608769
Balanced model AUROC: 0.7110078979726961

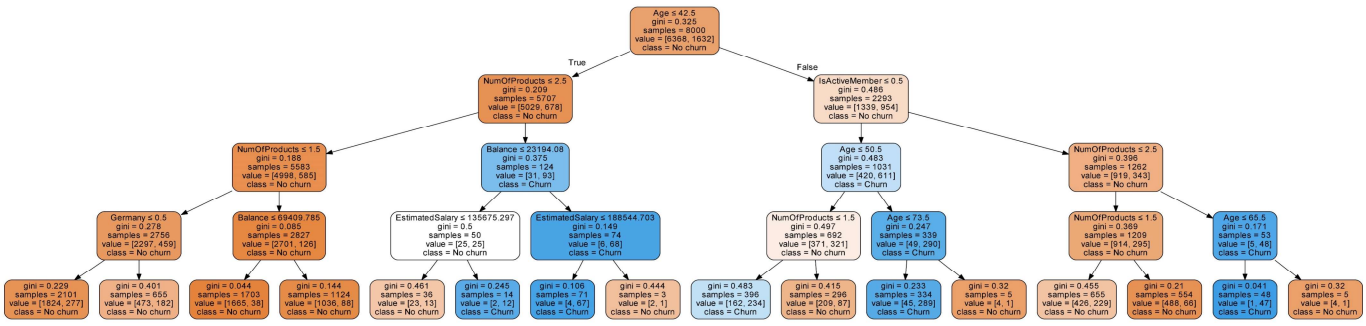
logistic regression



correlation between churn and each customer feature



relationship between features and customer churn base on random forest



relationship of features to customer churn base on decision tree

Conclusion

- Balanced logistic regression *can be used to* predict the loss of customers, and the accuracy is 0.689.
- The accuracy of random forest was 0.863. Analysis shows that age has the greatest impact on customer turnover.
- The accuracy of the decision tree model is 0.856, and the classification results from the decision tree show the characteristics of customer loss.
- Random forest has the highest accuracy among the three models.
- Future work: optimize bank's loan strategy to reduce the churn by offering different mortgage rates to different customers.

Reference

- Shirazi, Farid and Mahbobeh *Mohammadi. (2019). "A big data* analytics model for customer churn prediction in the retiree segment." Int J. Information Management 48, 238-253.
- Keramati, A., Ghaneei, H. & Mirmohammadi, S.M. (2016)."Developing a prediction model for customer churn from electronic banking services using data mining". Financial Innovation
- Chiang, D.-A., Wang, Y.-F., Lee, S.-L & Lin, C.J. (2003). "Goal-oriented sequential pattern for network banking churn analysis". Expert Systmens with Applications, 25(3), 293-302.
- Dudyala, A. K., & Ravi, V. (2008). "Predicting credit card customer churn in banks using data mining". International Journal of Data Analysis Techniques and Strategies, 1(1), 4-28.
- Bose, I., & Chen, X. (2009). "Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn". Journal of Organizational Computing & Electronic Commerce, 19(2), 133-151
- Sunkaraneni, Tarun. (2019). "Bank Customer Churn Modeling".www.kaggle.com/barelydedicated/bank-customer-churn-modeling. Public Domain

Optimizing Online Fraud Transaction Detection

Bhuvana Vellanki, Rajkumar Vijayakumar, Yug Patel, Yash Joshi, Manikandan Vinodh Kumar
Instructor: Edward A Stohr

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Motivation

- According to TD Bank's recent survey, the ***risk of payment fraud*** is the ***number one concern*** for 44% of financial industry
- According to Juniper Research, ***losses due to online and mobile payment frauds*** is estimated to be ***$22 billion this year***—and could go as high as $48 billion by 2023
- ***"Prevention is better than Cure"*** — The costs associated with being wrong are very expensive !

Dataset & Approach

- Dataset sourced from Vesta's ***real-world*** e-commerce transactions
- Re-classifying all categorical variables using ***One hot encoding***
- Replacing all ***NA's*** with the ***column means*** of the dataset
- ***Class Balancing*** after splitting into ***Training*** and ***Test*** data-sets
- Model trained on ***300000+*** observations and tested with ***170000+*** observations across ***242*** variables

Model Creation & Prediction

Confusion Matrix Weights		Predicted Fraud 0	Predicted Fraud 1
Actual	0	$ -	$ 20.00
Fraud	1	$ 50.00	$ -

Confusion Matrix Counts		Predicted Fraud 0	Predicted Fraud 1
Actual	0	167755	1240
Fraud	1	35432	46062

- The Average loss due to misclassification amounted to ~ $ 70 from the dataset. This is used as the weightage for false-positives and false-negatives with weightage for false-positives being more
- The model was compared against Linear Regression and was found to give a better optimized classification with better AUC
- Lastly, running solver on the model suggested a threshold value of 0.83 with a minimized 0.3 % loss of ~ $ 200,000 for an overall $ 79.83 M transaction amount

Metric for Loss

		Actual 0	Actual 1
Predicted	0	TP	FP
	1	FN	TN

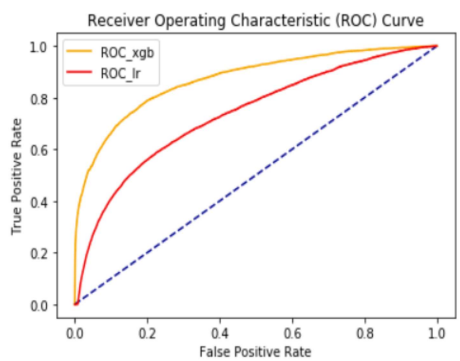
Expected Loss = $\sum p(o_i).v(o_i)$

Class priors => $p(\mathbf{p})$ = (TP+FN)/Total
$p(\mathbf{n})$ = (FP+TN)/Total

tp rate = TP/(TP+FN) *fp rate* = FP/(FP+TN)
fn rate = FN/(TP+FN) *tn rate* = TN/(FP+TN)

Metric for Loss = $p(\mathbf{p})$. [*tp rate*.(TP cost) + *fn rate*.(FN cost)] +
$p(\mathbf{n})$. [*tn rate*.(TN cost) + *fp rate*.(FP cost)]

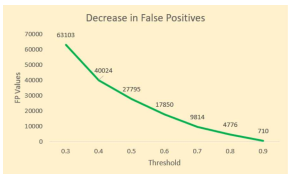
XGB & LR ROC

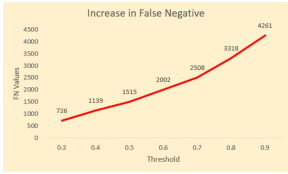


Sensitivity Analysis

Threshold	True Positive	True Negative	Fasle Positive	False Negative
0.3	107934	5399	63103	726
0.4	131013	4986	40024	1139
0.5	143242	4610	27795	1515
0.6	153187	4123	17850	2002
0.7	161223	3617	9814	2508
0.8	166261	2807	4776	3318
0.9	170327	1864	710	4261

Confusion Matrix





- Varying the threshold of prediction probability for XGBoost model to create a sensitivity analysis on false positives and false negatives
- Large decrease is observed in the false positive values which are predicted frauds but not actually fraudulent transactions
- There is a slight increase in the number of false negatives which are the transactions which cannot be certainly be classified as fraudulent

Future Scope

- Combining various model outputs and evaluate the confusion matrix to get better classification and improved accuracy.
- Perform time-series analysis and make use of variables such as transaction timing and deltas to classify fraudulent transactions
- Use techniques such as k-fold cross validation to get the aggregate of classifications and analuze the attributes of the resultant confusion matrix

Exploring the Relationship between Social Media and Bitcoin Value

Yixuan Wang, Wei Yang, Xinyi Li, Zexu Yang
Instructor: Feng Mai

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Keywords: Sentiment Analysis; Time Series Regression; Natural Language Processing (NLP)

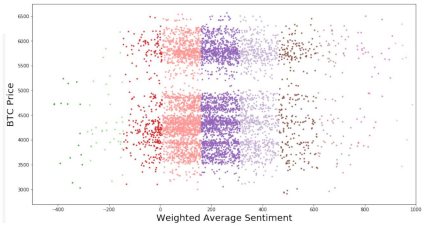
Motivation

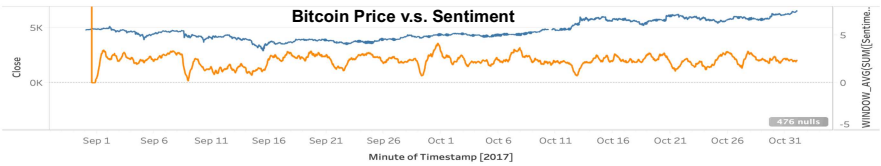
- With the popularity of Blockchain technologies and the increasing computing power invested in BTC mining, the decentralization and reliability of Bitcoin are promoting technology revolution and have been fully recognized by the market
- Based on previous researches, we hope to use textual data from Twitter, one of the largest social media platforms, to gain insights into the factors affecting the fluctuation of Bitcoin value

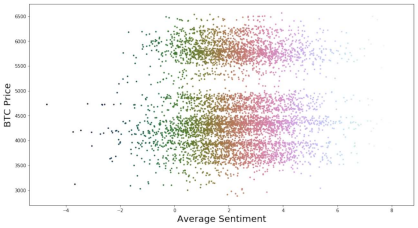
Data Manipulation

- Filtered stop-words, punctuation and whitespace
- Delivered Sentiment Analysis to distinguish positive and negative words
- Extracted and visualized the high-frequency words
- Quantified tweets features and integrated with Bitcoin value
- Analyzed the statistical relationships
- Visualized the results

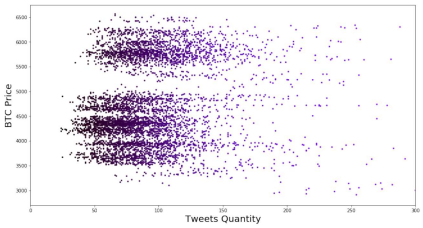
Data Analysis







- After applying the Moving Average method to these time series, we discovered that when people lose money by trading Bitcoin, they tend to publish negative words in Tweets. However, when they gain money from Bitcoin, they are less likely to showcase their success



- When Bitcoin value changes, the number of related comments from Tweet increases. Especially when Bitcoin value decreases, the number of negative words in Tweet increases dramatically

Future Works

- With the collection of more labeled datasets, we will train the Convolution Neural Network (CNN) models to better capture the underlying characteristics of tweets
- Related to Bitcoin textual data, we will apply the pre-trained models of the NLU module in IBM Watson, the fastText in Facebook, and make comparison and evaluation
- For the analysis part, we will explore regression models with trading volume, other cryptocurrency price indices, and then implement them over a longer period

Word Cloud generated by Tweets contents related to Bitcoin

Quantify the impact of AI on jobs

Ying Lu, Tianyi Wang, Yilun Yan Wei Yang
Instructor: Rong Liu

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Background

•As Artificial Intelligence (AI) becomes more intelligent, many activities in people's daily lives can be handed over to it. During the work, to improve efficiency, some activities that include high repetitiveness have been replaced by AI. In this analytics project, I propose some methods to evaluate AI influence on the jobs through the trends in job's description (JD) from 'indeed.com'. We may get insights in different JDs and provide indicators about a specific JD for the job seekers. Our model can help them accurately locate the required activities, and compare to what they really want to do. Additionally, they will know the proportion of this job can be replaced by AI. For recruitment website, they may also build activities label system for JDs, which may provide more supports for matching job seekers' need. Job activities: Each JD can be separated and categorized into several required activities or skills.

Current & Future Work

Text mining: Collecting Job Descriptions (JD) of at least 3 different positions from Indeed (4000 records of each), including the posting from NY, LA, Chicago, and Seattle.
Text filtering: Identifying the common structure of the JD, and removing unnecessary sections, such as company's descriptions, salary, offer, benefits.
Data Pre-processing: Cleaning, extracting and transform the unstructured text data with Natural Language Processing techniques. Preparing the structure data for the clustering models.
Exploratory data analysis: Exploring the potential features of the data; Checking the models required assumptions'; Identifying other possible analytics directions from the datasets.
Modeling: Clustering Model 1. K-Means or Latent Dirichlet Allocation (LDA) Clustering Model 2. Attention-based Aspect Extraction using DNN.

Evaluation

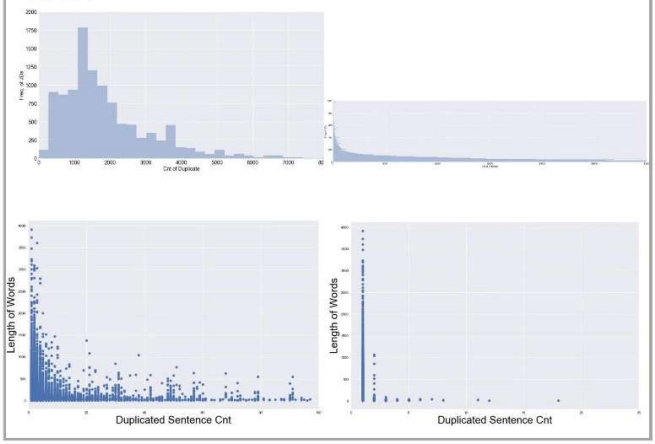
- Evaluate weight of AI factor on each activity based on two clustering results.
- Comparison with the weight of AI factor on each activity-based data from the Bureau of labor statistics (BLS)

Data Source

Using web scraping to get job description data from 'indeed.com'

	Job Description (suppose 3 jobs)
Data type	characteristic
Website	www.indeed.com
Sample	File Edit Format View Help Front Desk/Receptionist needed for a busy, highly organized chiropractic/physical therapy office. The person we are seeking must be very out-going, mature, dynamic professional, and able to maintain composure when the office is at its busiest times. The person needed for this position must also work well with co-workers and the doctor, and have excellent interpersonal skills. This position is very important because it is the first person all patients see, and is a representation of who and what we are. Essential Duties: Register patients in scheduling software, confirming accuracy. Schedule appointments; gather pertinent information needed to complete appointment. Prepare charts for the next day appointments. Confirm insurance status, eligibility and referrals when needed. Confirms insurance authorizations Answer multi-line phones. Check in and out patients, collect co-pays at time of service.

EDA



Model 1: K-Means

- Try different k from 3 to 60, k= 42 gets the best clustering result
- Continue to working on other models

Predicting Researcher Performance in Depression Treatment

Yilun Yan, Haihan Hu, Siyi Xu
Instructor: Prof. Christopher Asakiewicz

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction

The World Health Organization state that across the world, more than 300 million people of all ages live with depression. The number of Researchers from all round the world who focus on depression treatment and that of related research areas are increasing each year. The main purpose of this project is to find the most prominent authors in the field of depression treatment using bibliographic data analysis.

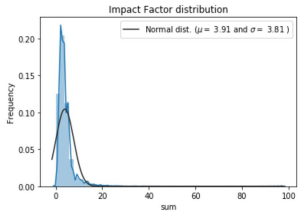
Data Preparation

- Scraping data from Web of Science
- Pre-processing 3800+ publication information involving article title, published year, journals, authors, impact factor of journals, h-index, funding status
- Using social network analysis methodology to build co-author network and calculate centrality degree of every author
- Predicting the authors future performance based on impact factors

Impact Factor Analysis

- Impact Factor(IF) is used to reflect the average number of citations divided by the total number of articles post on the journal recently

$$\mathrm{IF}_y = \frac{\mathrm{Citations}_{y-1} + \mathrm{Citations}_{y-2}}{\mathrm{Publications}_{y-1} + \mathrm{Publications}_{y-2}}$$

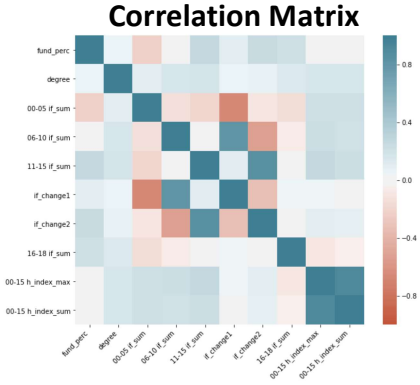


Variable Selection

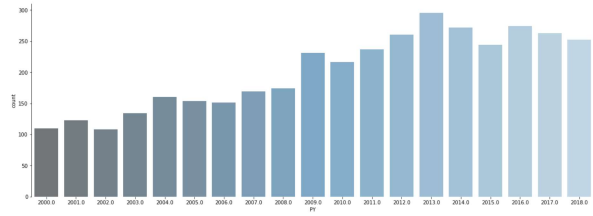
Y	x1(fund_perc)	x2(degree)	x3(00-05if_sum)	x4(06-10if_sum)
$y = \sum_T IF_T$ T: 2016~2018	$\frac{\sum_T article(fund)_T}{\sum_T article_T}$ T: 2000~2015	degree	$\sum_T IF_T$ T: 2000~2005	$\sum_T IF_T$ T: 2006~2010
x5(11-15if_sum)	**x6**(if_change0)	**x7**(if_change1)	**x8**(if_change2)	**x9**(00-15h_index_sum)
$\sum_T IF_T$ T: 2011~2015	$\Delta IF_{2000\sim2005}$ $=IF_{2005}-IF_{2000}$	$\Delta IF_{2006\sim2010}$	$\Delta F_{2011\sim2015}$	$\Delta h\text{-}index$ $= h_{2015}-$ $h_{PublicationYear}$

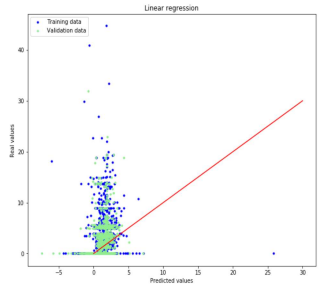
- h-index is defined as the maximum value of *h* such that the given author/journal has published *h* papers that have each been cited at least *h* times. It can assess both the number and the level of academic output for a researcher.

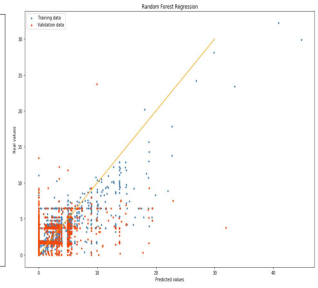
Correlation Matrix

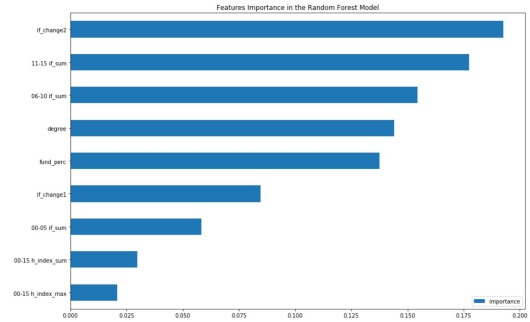


Modeling & Results









- Three graphs from top to bottom are: changes in the number of published articles (F1), the fitting results of LR and RF models (F2), and the importance of each variable in the RF model (F3)
- F1 showed the trend of topic discussion, and this topic (depression treatment) has caused a lot of discussion since 2009. Combine it with F3, we could find that time and discussion situation have a significant impact on IF.

Conclusion

Author	Ranking (Change of Impact Factor in last 5 years)	Ranking (Predicted Impact Factor)
Klein, Jan Philipp	*17*	*22*
Meiser-Stedman, Richard	*66*	*40*
Meltzer-Brody, S.	*67*	*57*
Furukawa, Toshi A.	*83*	*89*

- Through the model, we found that the first '5 years' after the topic was popular had the greatest impact on the author IF.
- The table shows the top 4 most promising researchers in the 2011-2015 period. They were not very influential at that time(as can be seen by the huge IF differences), but they have some influence many years later.

Optimization of NYPD Response to Crime in Manhattan

Ayusha Anand, Cindy Hsu, Yen-Cheng Liu, Kun Wu, Xinran Bai
Instructor: Professor Edward Stohr

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

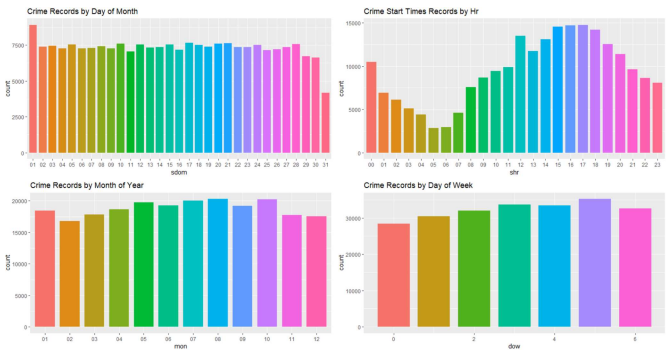
Motivation

- To prevent various types of crimes and increase police duty executive efficiency in Manhattan.
- This is done by predicting locations and frequency of crimes in the city and optimizing police response to crime by locating the nearest police precinct to mentioned crime scenes.
- By improving police response to crime we aim to minimize the overall crime rate in Manhattan.

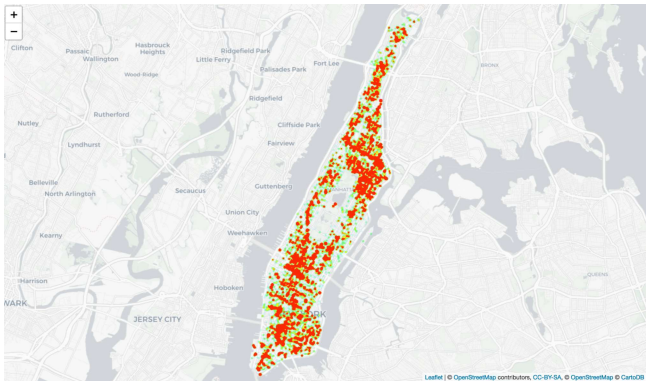
Design methodology

- R Studio to implement the k-nearest neighbor algorithm which returned the nearest police precinct to each crime scene.
- Excel was used to manage and clean the NYPD crime complaint data in 2017-2019.
- Tableau was used to create visualizations to better understand the distribution and frequency of crime occurrence in Manhattan.

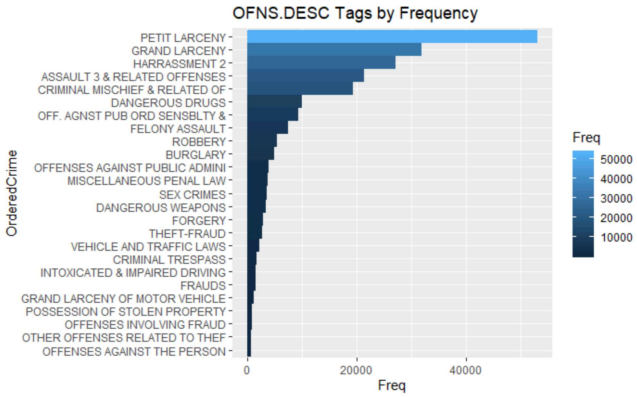
Frequency visualization



The graph above depicts the frequency of crime occurrence by month, day and hour. We can see the times at which crimes are most likely to take place.



The density map depicted above shows us the distribution of crimes that take place in Manhattan, by location.



This graph depicts frequency of crime classified by type of crime. We can see that the most frequently occurring crime is Petit Larceny which has the half of the total crime. And the least frequent is offenses against a person.

Results

Make location optimization to police force deployment in terms of finding the optimal locations of police force with the spatial distribution of crimes. It decreases the proportion of the police officers to cross jurisdictions.

Exploring several popular methods on analyzing Candlestick Chart

Author: Minghao Kang, Yuwen Jin
Instructor: Dragos Bozdog

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction :

Speculation on equity market has been popular for years, and Candlestick Chart is one of the most popular method for individuals to predict future trend.

However, not all patterns of the chart come from statistic calculation. Therefore, we want to do some test with randomly picked equities. Our goal is to try dozens of popular patterns, get accuracy and see if there is some improvement we can do. For example, we combine the pattern models with RSI index in order to increase the accuracy. (Using R software as our tool)

Empirical Analysis:

For each chosen pattern, we find detailed conditions, analyze them in R together with the RSI index to predict directional price changes. In all cases, we estimate the accuracy of the patterns with our data.

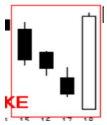
After finding all thed target patterns, we record temporary results (accuracy) into a table, add other conditions (RSI thresholds), estimate accuracy of each methods and record. Then, compare 2 tables for further explorations.

Sample code:

```
# Method 5
result.table.5 <- NULL
k <- 1
for (i in names) {
  df <- get(i)
  for (j in 11:(nrow(df)-13)) {
    if (df[j, 7] == 1
        && all(df[(j+1):(j+3), 7] == -1)
        && df[j-10, 2] < df[j, 2]
        && all(order(df[(j+1):(j+3), 4]) == 3:1)
        && all(order(df[(j+1):(j+3), 3]) == 3:1)){
      result.table.5 <- rbind(result.table.5,
                              c(df[(j+4):(j+6), 7],
                                df[j+6, 2]-df[j+4, 1],
                                df[j+13, 2] - df[j+3, 1],
                                df[j+13, 3] - df[j+3, 1],
                                df[j+3, 10]))
      rownames(result.table.5)[k] <- paste(i, "pattern", j, "to", j+2)
      k <- k + 1
    }
  }
}
colnames(result.table.5) <- c("1 day later trend", "2 days later", "3 days later",
                              "change in 3 days", "HIGH change in 10 days",
                              "LOW change in 10 days", "RSI")
head(result.table.5)
# In prediction it will go down
s.accuracy.5 <- sum(result.table.5[, 4] < 0)/nrow(result.table.5)
l.accuracy.5 <- sum(result.table.5[, 6] < 0)/nrow(result.table.5)
```

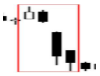
Sample patterns and observations.

1. Three line strike.



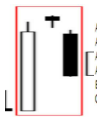
	change in coming 3 days	change in 10 days	RSI
Amazon.01 pattern 1014 to 1017	5.500000	-11.789993	60.78324
Amazon.01 pattern 1434 to 1437	-0.899994	-4.000000	72.11920
Amazon.01 pattern 1515 to 1518	3.239990	16.480011	64.77297
Amazon.01 pattern 2728 to 2731	26.390015	29.650025	74.72005
AmericanAirlines.01 pattern 119 to 122	-0.200001	5.610000	41.49236
Ashford.01 pattern 1953 to 1956	0.065317	0.475876	35.88098

2. Two Black Gapping



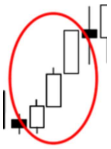
	change in 3 days	HIGH change in 10 days	LOW change in 10 days	RSI
Amazon.01 pattern 742 to 745	-2.140000	7.470001	4.350006	50.38472
Amazon.01 pattern 1041 to 1044	-0.970002	-7.110001	-12.960007	42.27018
Amazon.01 pattern 1442 to 1445	4.600006	6.950012	-1.669998	52.23582
Amazon.01 pattern 1558 to 1561	0.460022	8.100006	1.710022	43.08106
Amazon.01 pattern 2198 to 2201	27.159974	47.760010	37.659974	39.28171
Amazon.01 pattern 2718 to 2721	-2.289978	125.950074	101.790040	50.14921

3. Evening star



	change in coming 3 days	HIGH change in coming 10 days	LOW change in 10 days	RSI
Amazon.01 pattern 1558 to 1560	-5.480011	2.399994	-0.920013	46.97134
Amazon.01 pattern 2038 to 2040	7.320008	9.290009	4.970002	76.35085
Amazon.01 pattern 2789 to 2791	-33.400025	-8.450074	-40.550049	67.42439
AT&T.01 pattern 1314 to 1316	0.290001	-0.500000	-0.910000	68.95600
BankOA.01 pattern 1437 to 1439	0.090000	-0.030000	-0.310000	70.15064
CanadaG.01 pattern 316 to 318	-0.931504	45.678787	34.087234	51.87796

4. Three White solider



	change in 3 days	HIGH change in 10 days	LOW change in 10 days	RSI
Ashford.01 pattern 895 to 897	0.018787	0.212919	0.087672	68.77111
Ashford.01 pattern 1216 to 1218	0.450889	0.075148	-0.225445	63.69532
Ashford.01 pattern 1652 to 1654	0.144034	0.131509	0.025050	41.57469
CanadaG.01 pattern 3058 to 3060	44.709961	51.289917	19.549927	56.53065
CocaCola.01 pattern 1881 to 1883	0.580002	0.290001	0.020000	68.15405
CocaCola.01 pattern 2306 to 2308	0.690002	1.639999	1.260002	59.23636

5. Three Black Crows

	change in 3 days	HIGH change in 10 days	LOW change in 10 days	RSI
Amazon.01 pattern 193 to 195	0.630005	-5.830002	-9.010002	48.56879
Amazon.01 pattern 235 to 237	-1.599998	0.840004	-2.599998	55.70862
Amazon.01 pattern 281 to 283	0.309997	-6.940003	-12.310001	40.47632
Amazon.01 pattern 307 to 309	6.089996	1.119995	-1.640000	49.90076
Amazon.01 pattern 396 to 398	2.639999	13.169998	9.900001	52.40737
Amazon.01 pattern 408 to 410	3.459999	-1.089996	-3.069999	53.01861

Sample results :

without and with RSI

```
> result.table
         Predict direction Observation 3 days accuracy 10 days accuracy
Method 1                Up          35       0.6285714        0.6285714
Method 2              Down         135       0.3185185        0.5185185
Method 3              Down          14       0.2857143        0.5000000
Method 4                Up          13       0.7692308        0.7692308
Method 5              Down         983       0.2726348        0.6307223
         Predict direction Observation 3 days accuracy 10 days accuracy
Method 1                Up           8       0.5000000        0.8750000
Method 2              Down          37       0.2972973        0.4324324
Method 3              Down          11       0.2727273        0.5454545
Method 4                Up           2       1.0000000        1.0000000
Method 5              Down         328       0.2439024        0.5945122
```

Conclusions :

After testing on 10 years of data of 20 companies, we find that most of the condlestick charts are not very reliable (mostly with accuracy below 65%) while predicting directional price change. After adding RSI, accuracies become more extreme, some of the petterns seem to have better performance (above 80%), and some are very low (below 30%). Better choices depends on target investing periods.

The Opioid Crisis in Numbers

Leland Montgomery, Vinay Balani
Instructor: Alkiviadis Vazacopoulos

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics
November 21th, 2019

The Washington Post and the Charleston, West Virginia Gazette-Mail newspapers waged a year-long legal battle for access to the DEA's automated reporting system known as ARCOS. The government and the drug industry had sought to keep the database secret, but a court order made the data for 2006-2012 available. The Post opened the database to allow researchers like us to count how much hydrocodone and oxycodone went to states and counties, and which companies and distributors were responsible. Since the beginning of the opioid crisis in 1999, nearly 400,000 Americans have lost their lives to this tragedy.

Who Made the Pills?

Manufacturer	Total Doses	Market Share
SpecGx LLC (Mallinckrodt)	28.8 billion pills	38%
Actavis Pharma, Inc.	26.5 billion pills	35%
Par Pharmaceutical (Endo Pharmaceuticals	12 billion pills	16%
Nationwide Total	76 billion pills*	(2012-2016)

- 242 pills for every man, woman and child in the U.S.A.

Deaths from Opioid Overdosing 2006-2012: 100,000

Who Distributed the Pills? (2006-2012)

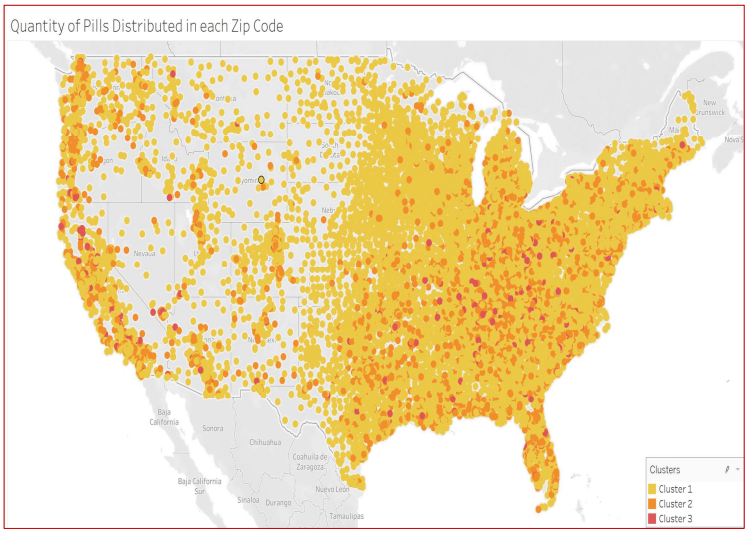
Distributor	Total Doses	Market Share
McKesson Corp.	14 billion pills	18.4%
Walgreens	13 billion pills	16.5%
Cardinal Health	11 billion pills	14.0%
AmerisourceBergen	9.0 billion pills	11.7%
CVS	5.9 billion pills	7.7%
Walmart	5.3 billion pills	6.9%
Smith Drug Co.	1.3 billion pills	1.8%
Rite Aid	1.3 billion pills	1.7%
Kroger	1.2 billion pills	1.6%

Walgreens distributed 41 pills for every man, woman and child in the U.S.A.

Who Prescribed these Pills?

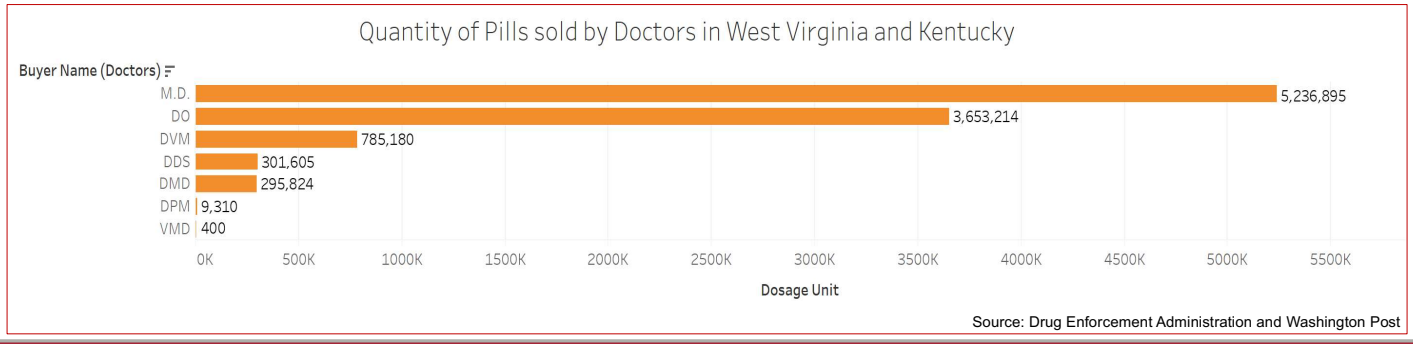
A Few of the more prominent examples:

	Haven Wall, Jr. M.D.	Prescribed 26,842 pills in Lewisburg, a town of 3,900 people, from 2006-12.
	James M. Minger D.V.M	This Veterinarian from Morgantown, WV prescribed 19,600 opioid pills in 7 years.
	Paul D. Miller, D.O.	Prescribed 9,500 pills in Terra Alta, WV, a town of 1,477 residents, in 7 years.
	Roland Chalifoux, D.O.	Prescribed 9,200 pills in a town of 1,926; federally indicted on 15 counts of health care fraud, 7 counts of mail fraud and 4 counts of wire fraud.



In Kentucky, Chain Stores Dominated the Market

	Largest Retailers	# of Pills	Market Share (KY)
Walgreens	Walgreen Pharmacy	60,039,060	11%
Kroger Pharmacy	Kroger Pharmacy	47,567,370	10%
Rite Aid Pharmacy	Rite Aid Pharmacy	35,270,500	7%
CVS pharmacy	CVS Pharmacy	29,805,400	6%
Walmart Pharmacy	Wal-Mart Pharmacy	29,574,580	6%
PCA Pharmacy	PCA Corrections Pharmacy	2,882,770	1%
ValuMed Pharmacy	Value-Med Pharmacy	2,780,060	1%
K kmart pharmacy	KMart Pharmacy	2,587,628	1%



Source: Drug Enforcement Administration and Washington Post

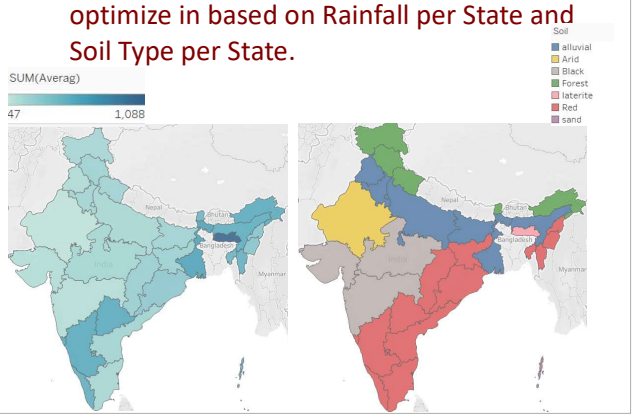
Crop Optimization in India

Kanav Chibber, Rishabh Shah
Instructor: Alkiviadis Vazacopoulos

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics
November 21, 2019

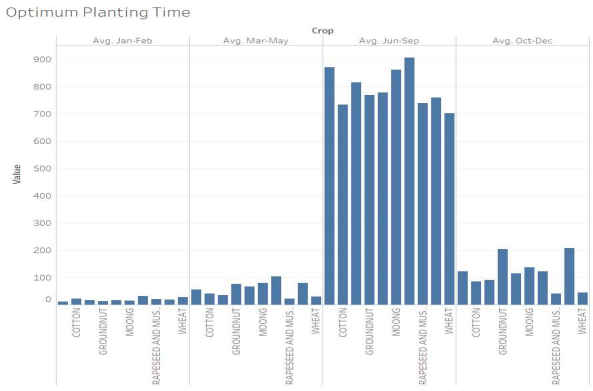
Introduction

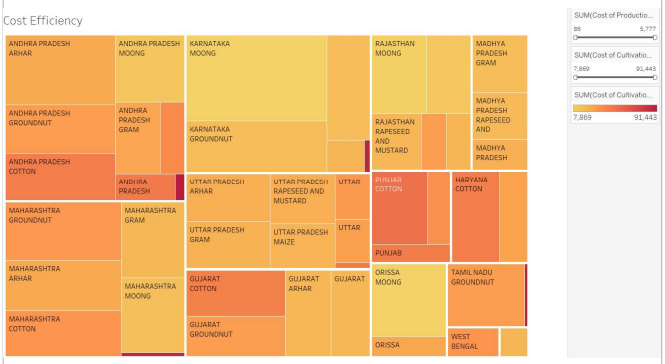
- We wanted to see which crop India should optimize in based on Rainfall per State and Soil Type per State.



Experiment (?)

- What we looked at was when the amount of rainfall was at its highest and the cost of cultivating and producing the crop





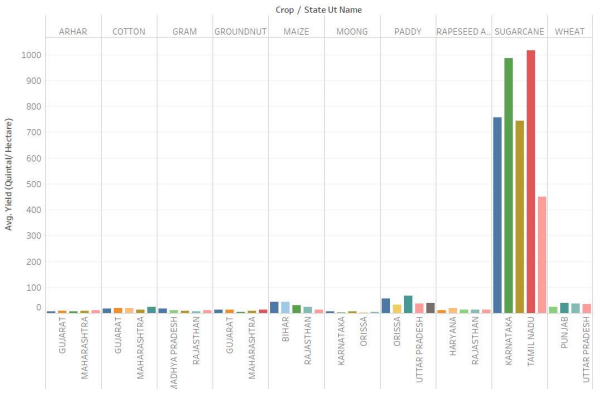
Results

- Based on the rainfall, soil types, and cost of production and cultivation, we saw that sugarcane has the highest yields

Average Yield per Annum

Crop	State												
	Andhr..	Bihar	Gujar..	Harya..	Karna..	Madh..	Maha..	Orissa	Punjab	Rajas..	Tamil ..	Uttar ..	West ..
ARHAR	6		10		7		9					10	
COTTON	18		19	20			13		24				
GRAM	17					10	8			7		11	
GROUNDNUT	12		13		5		9				12		
MAIZE	43	43			31					24		14	
MOONG	6				1		7	3		4			
PADDY	56							32	67			37	39
RAPESEED AND MU..			12	20		13				14		14	
SUGARCANE	758				986		744				1,015	449	
WHEAT						24			40	37		35	

Market Leader for Crop Share



Conclusion

- The conclusion we came to is that India should plant more Sugarcane plants as they have the highest yields and the lowest cost of production. This would help to boost India's economy.

Visualization and prediction of awards in a crowed-source environment

Mostaan L Saremi, Giusi Sotera
Advisor: Prof. Alkiviadis Vazacopoulos

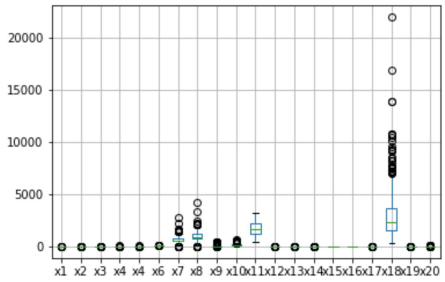
STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Statistical Learning & Analytics
Fall, 2019

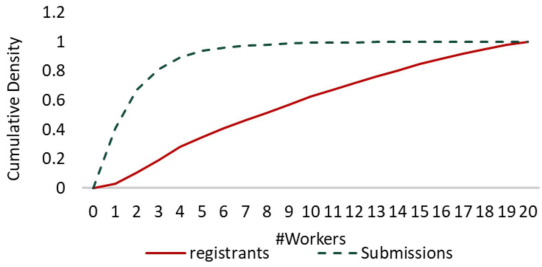
Data Understanding

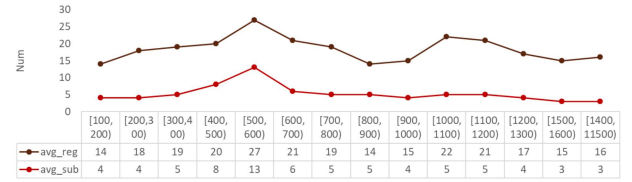
- The dataset used contains 514 component development task from Sep 2003 to Sep 2012, extracted from Top coder Website.
- All tasks are successfully completed, and the total award is divided in the top two winners with a 2:1 ratio.
- The awarded prize is dependent variable and other variables are independent

Data Visualization



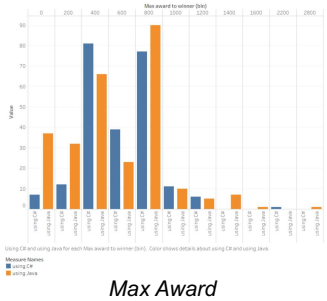
A simple plot of data to see the outliers



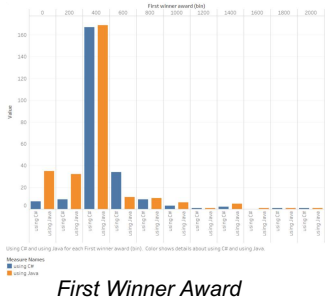


	[100, 200)	[200,300)	[300,400)	[400, 500)	[500, 600)	[600, 700)	[700, 800)	[800, 900)	[900, 1000)	[1000, 1100)	[1100, 1200)	[1200, 1300)	[1500, 1600)	[1400, 11500)
avg_reg	14	18	19	20	27	21	19	14	15	22	21	17	15	16
avg_sub	4	4	5	8	13	6	5	5	4	5	5	4	3	3

Scatter-plot of average Performance vs. Award



Max Award
vs
programing language



First Winner Award
vs
programing language

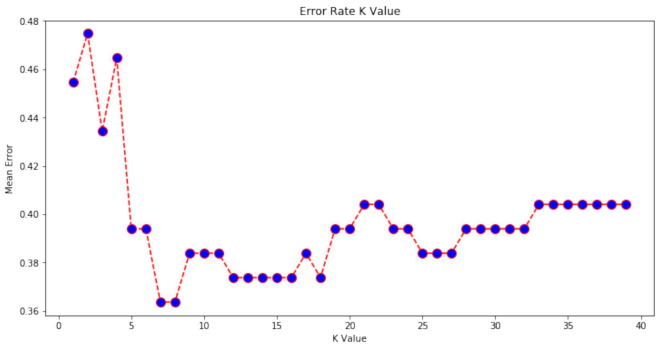
Results & Evaluation

- All different models of KNN, linear regression, Logistics regression, and Naïve Bisayan were applied to the data set in order to predict the optimum monetary prize.
- The best prediction was belonged to KNN (K=7) with 64% accuracy.
- By applying PCA the KNN accuracy decreased by 8% and reached to 56%.

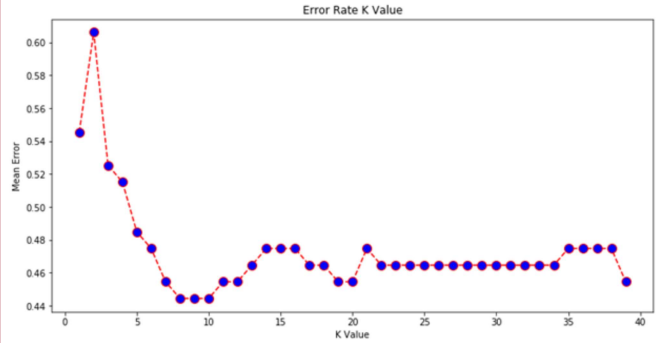
	y1_devAward
2	400
3	700
6	500
7	500
9	550
16	100
19	500
20	1800

	fit	lwr	upr
2	364.9198	314.06475	415.7749
3	509.8689	447.18459	572.5533
6	524.8586	406.19364	643.5236
7	459.7405	397.06985	522.4112
9	579.3995	524.48106	634.3179
16	312.7529	250.42057	375.0852
19	480.6162	421.35035	539.8820
20	664.2047	489.90968	838.4996

Linear regression results of 10 samples



KNN errors before applying PCA



KNN errors after applying PCA

Reference

- K. Mao, Y. Yang, M. Li, and M. Harman, “Pricing Crowdsourcing-based Software Development Tasks,” Piscataway, NJ, USA, 2013, pp. 1205–1208.
- Y. Yang and R. Saremi, "Award vs. Worker Behaviors in Competitive Crowdsourcing Tasks," ESEM 2015, pp. 1-10

Plagiarism Detector – Batch Prediction and Cloud Deployment

Pranav Prajapati, Sonali Johari
Instructor: Prof. Edward Stohr

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

Introduction

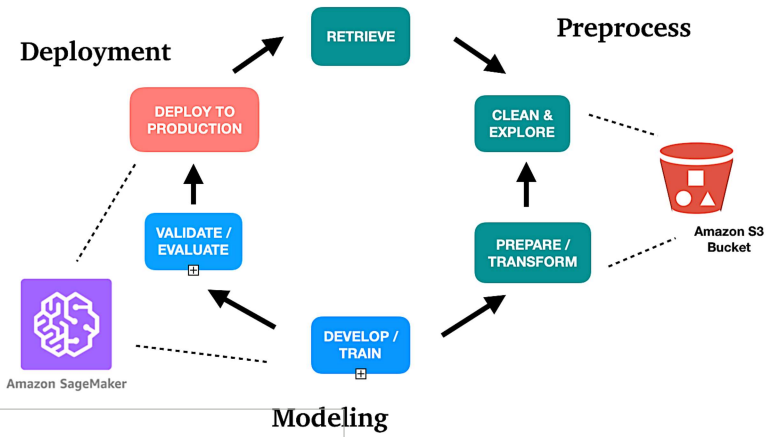
- Plagiarism detection is commonly used by universities and journals to maintain integrity of original content
- Detecting plagiarism is an area of active research – the task is non-trivial and the differences between content paraphrased and original content is not that obvious

Categories of Plagiarism

- Plagiarized : *cut, light, heavy*
- Non-Plagiarized: *non*
- Special, Source text : *org*

Data Preprocessing

- Implemented stratified random sampling to randomly split data by task and plagiarism amount
- Ensures an approximate 74% of training and 26% of testing data
- Text processing cleaned the data of punctuations and converted the entire text to lowercase



Numerical Columns

- According to set of rules, the categories were converted to numerical values

Class Labels

- The Goal is to create a binary classifier. Hence a binary class label 1 – plagiarized and 0 for not was made

Determining Plagiarism

- One of the ways to detect plagiarism is by computing similarity features - a measure of how similar a given text is to the original text.

Containment

- Containment is the ratio of the intersection of of the n-gram word count of the original text with the n-gram word count of the paraphrased text to the n-gram word count of the paraphrased text, given by:

$$\frac{\Sigma\, count(ngram_P) \cap count(ngramO)}{\Sigma\, count(ngramP)}$$

Longest Common Subsequence

		A	B	C	D
	0	0	0	0	0
B	0	0	1	1	1
D	0	0	1	1	2

no match: max of top/left values

		A	B	C	D
	0	0	0	0	0
B	0	0	1	1	1
D	0	0	1	1	2 +1

match: diagonal addition

- LCS is the longest string of words (or letters) that are the same between the original text and the paraphrased text
- Using dynamic programming, an optimal method was implemented

Feature Selection

- Multiple containment features for different n-grams
- Remove highly correlated features and selected the best ones for our model

Loading Data

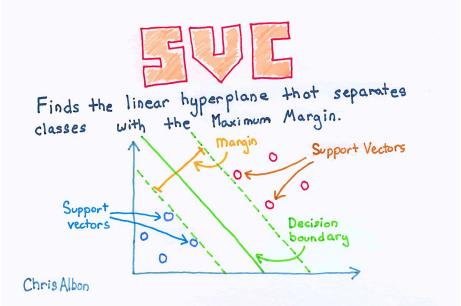
- Only Training data is loaded to Amazon S3 Buckets

Custom Scikit- learn Classifier in SageMaker

- With Amazon SageMaker, we can quickly and easily train our model and directly deploy them into production ready hosted environment
- It provides an integrated Jupyter authoring notebook instance for easy access without requiring to manage servers
- To run custom training script in SageMaker, an estimator was constructed with specific entry point and instance type

Support Vector Machines

- Once model is trained and deployed, we can create versions of different models and test in batches



Conclusion

- The model was successfully created with its endpoint being the actual function call that acts as a REST API
- The model was able to classify all the documents successfully with full accuracy
- Further scope – Use API Gateway and lambda function to deploy model to a web application

End to End Automated Data Pipeline of Yelp Reviews

Pranav Prajapati
Instructor: Joseph Morabito

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

Introduction

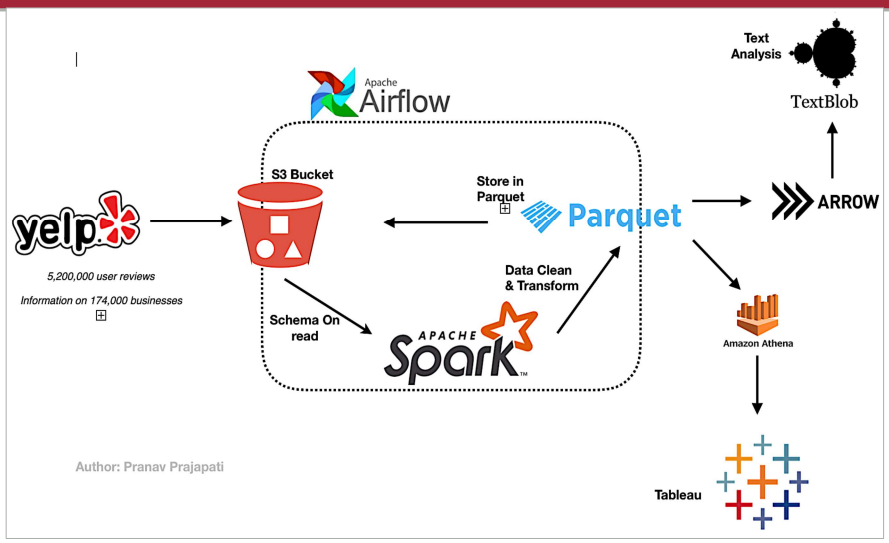
- The Star treatment – Chefs go all out to impress reviewers
- 137 million active Yelp Reviews

Evolution of Data Warehouse and why Data Lakes ?

- Abundance of unstructured data
- Rise of Big Data technologies
- Advanced Analytic Capabilities
- Schema on Read

The data

- Deep JSON and CSV file formats
- Business, Review, Tip, Check-in, User data
- Restaurants, Compliment tables created

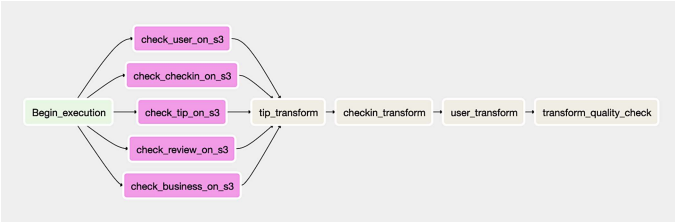


Overplanned Analytics Initiatives Are Doomed to Fail

Airflow and DAGs

- Use airflow to author workflows as directed acyclic graphs (DAGs) of tasks
- The rich user interface makes it easy to visualize pipelines running in production, monitor progress, and troubleshoot issues when needed
- We can automate tasks and schedule them whenever we want – Endless Possibilities !
- Essential Data Quality checks are added

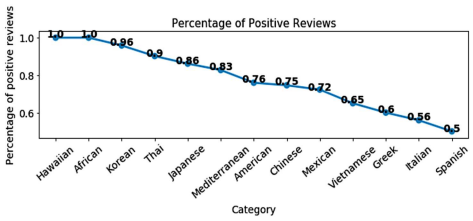
First DAG

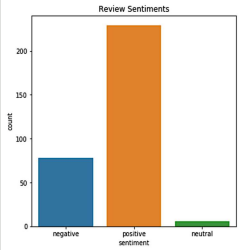


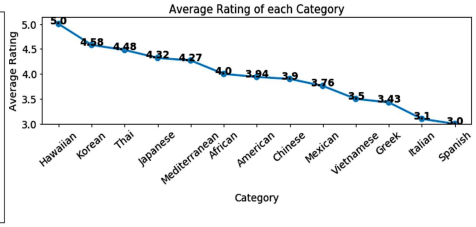
Second DAG

Analysis of the Restaurant Reviews

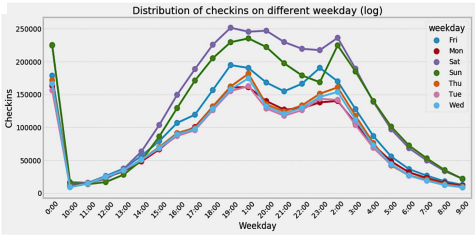
- Do people always Rate positively ?
- Which Cuisines are liked the most?
- Sentiment towards specific cuisines







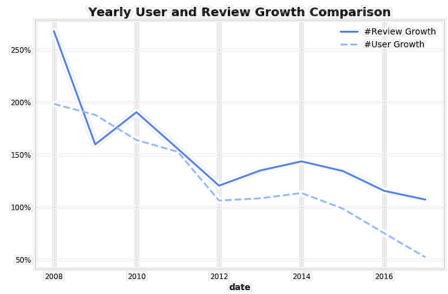
- Daily check-ins of users
- Partitioning the Data on S3 gives performance boost



59%** is the amount by which revenue of restaurants have increased when the average ratings increase by **1 star

Connect to your Dashboard

- Configure Amazon Athena Connector with Tableau to read the Data
- Create Business Intelligence Reports
- Present information in a more interactive manner



Conclusion

This Pipeline solves the logistics between data sources and those who need access to data to undertake further processing , visualizations, transformations, routing, reporting or statistical models

Scan here for detailed code and report

Using the K-means Clustering Analysis for A Unleveraged Portfolio

Fangchi Wu, Lantian Jia, Wenbo Yu
Instructor: Dragos Bozdog

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction

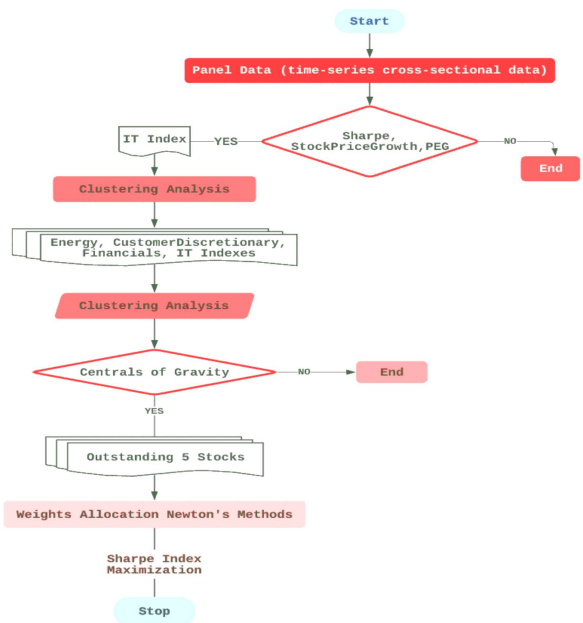
People of different ages maybe tend to have different investment choices in the stock market. In general, older people are more likely to choose conservative stocks to invest in, while younger prefer to invest in high-yield stocks. The risk and benefit are two important factors in stock investment. Young people in the 20-30 age range are becoming more conservative because of the limited amount of funds, and they focus on risks as well as benefits. For this practical problem, we will research the portfolio of stock in the project, which can provide some advice for those young people who want to maximize profits and minimize risks.

Data Process

Data Integration. After integrating the original data with the calculated data, we obtain a data set that totals 11 dimensional variables.

Handling outliers and missing values.

Standardizing data.




```
> kmeans(SumTable[2:6], 5)
K-means clustering with 5 clusters of sizes 2818, 5, 9, 17, 3

Cluster means:
           SPs          Vs           Bs         EPSs         PEs
1  0.006339092  0.13168503   -0.07011610   0.5687069  0.010862197
2  0.047286703  0.21907551   48.01199946   1.3891534 -0.015585296
3 -0.008745050 -0.04892740   -0.03630683 -46.8631756  0.005050115
4 -0.020504330  0.16485404    0.87183625  32.5814276 -0.018726724
5  0.055954692 -0.04004368 -188.02004282   0.9454212  0.023996164

 [737] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [769] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [801] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [833] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [865] 1 1 1 1 1 5 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 4 1
 [897] 1 1 1 1 1 1 1 1 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [929] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [961] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [993] 1 1 1 1 1 1 1 1
 [ reached getOption("max.print") -- omitted 1852 entries ]

Within cluster sum of squares by cluster:
[1] 26782.256  1860.762  6691.117  6991.168 13319.527
 (between_SS / total_SS =  73.6 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

Method and Technology

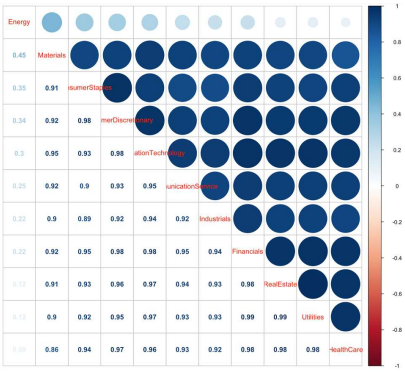
Pearson correlation coefficient analysis

K-means cluster analysis

Newton method to find the optimal solution of weight combination

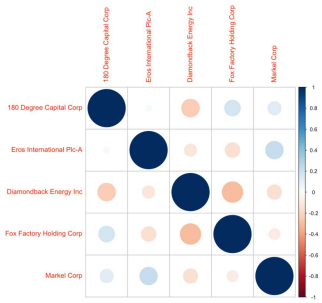
Results and Conclusion

The correlation coefficient between the rise and fall of the total stock price of each industry is obtained.



The k-means clustering of all stocks in 4 specific industries (The overall stock price rise negative correlation or positive correlation) is conducted to divide them into 5 clusters. And we get the center of gravity for each cluster.

According to the defined distance formula, the most representative stock is selected from each cluster, which is mathematically the point closest to the center of gravity of each cluster. They are 180 Degree Capital Corp, Eros International Plc-A, Diamondback Energy Inc, Fox Factory Holding Corp, Markel Corp.



Using machine learning, I made a portfolio of these five stocks, established the weights and reached the best sharpe ratio. The vector of weights is (0.0065%, 31.19%, 9.0%, 0.023%, 59.772%)

After deducting the risk-free return, the expected return of the portfolio is 5.7818%, the standard deviation is 2.45494% and the sharpe ratio is 2.315.

APTOS Blindness Detection

Jhao-Han Chen, Fang Zhang
Instructor: Chris Asakiewicz

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

Introduction

Problem Statement:

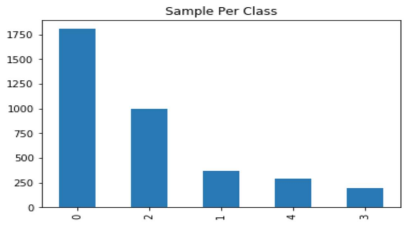
Millions of people suffer from diabetic retinopathy, the leading cause of blindness among working aged adults. Aravind Eye Hospital in India hopes to detect and prevent this disease among people living in rural areas where medical screening is difficult to conduct. So they set up a competition on kaggle.com for competitors to build model which can help ophthalmologists to identify potential patients.

- **Goal:** Maximize the quadratic weighted kappa, which measures the agreement between true labels and predicted results.
- **Purpose:** Being able to detect blindness before it happened.

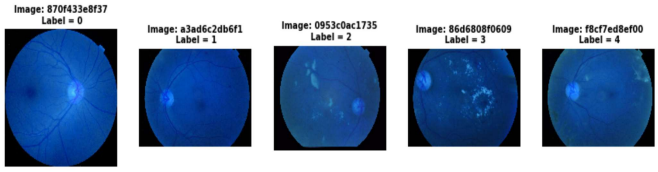
Data Understanding

A clinician has rated each image for the severity of diabetic retinopathy on a scale of 0 to 4:

0 - No DR
1 - Mild
2 - Moderate
3 - Severe
4 - Proliferative DR



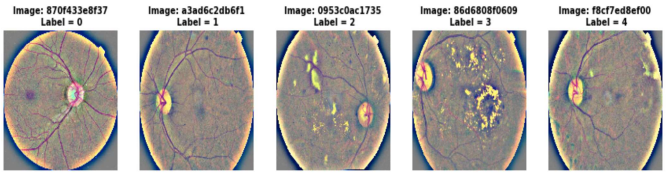
Training data contain 3662 samples, each sample contains one image and it's severity scale.



Preprocessing training image pipeline:

Load image → crop image → Resize image → Add noise

Final training images:



Models

CNN

- 4 conv layers, 2 fc layers, 175K parameters
- Deal with overfitting: Dropout layer, Batch normalization, data augmentation
- Accuracy: 0.694
- Kaggle submission:

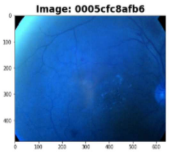
kernel4eb76f9017 (version 1/1) 6 days ago by Hank Chen From Kernel [kernel4eb76f9017]	Succeeded	0.552739	0.220898

Resnet50

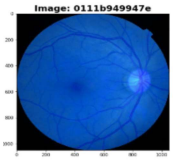
- 49 conv layers, 1 fc layers, 36M parameters
- Deal with overfitting: Dropout layer, Batch normalization, data augmentation
- Pretrain big dataset and fine tune the complete model
- Improve optimization: go deeper to the model, skip connections for curing vanishing gradient.
- Accuracy: 0.8428
- Kaggle submission:

kernel49d655b707 (version 3/6) 4 days ago by Hank Chen From Kernel [kernel49d655b707]	Succeeded	0.835671	0.622312

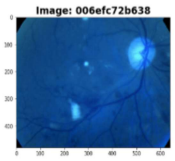
Prediction (test data)



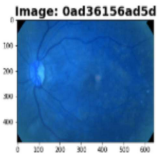
→ 2



→ 0



→ 3



→ 4

Future Work

- Try different models with more layers but less parameters but improve the accuracy.
- Try different cropping methods in order to efficiently extract useful features.

PLAYER RECOMMENDATION SYSTEM

FIFA

AUTHOR: SAHAJ SHUKLA, PREM PATEL

INSTRUCTOR: PROF. EDWARD STOHR

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

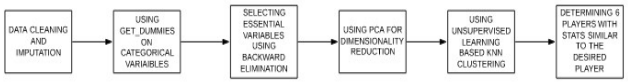
OBJECTIVE

To help the user create the best possible team with limited money. The team will be optimized in a way that maximum overall stats can be achieved using minimum money .

BACKGROUND

FIFA 19 by EA sports is one of the top trending games in 2019. Due to Its highly competitive nature, creating a good team with famous players is difficult since there are a lot of people bidding for the same players. However, if the stats of relatively less famous players are similar to the famous ones, it is a good idea to buy the cheaper players and achieve similar overall strength. Thus we can optimize the team spending minimum amount.

PROJECT FLOW



WORKING OF PCA

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i (t_1)^2_{(i)} \right\} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i (\mathbf{x}_{(i)} \cdot \mathbf{w})^2 \right\}$$

Equivalently, writing this in matrix form gives

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \{\|\mathbf{X}\mathbf{w}\|^2\} = \arg\max_{\|\mathbf{w}\|=1} \{\mathbf{w}^T\mathbf{X}^\mathbf{T}\mathbf{X}\mathbf{w}\}$$

Since $\mathbf{w}_{(1)}$ has been defined to be a unit vector, it equivalently also satisfies

$$\mathbf{w}_{(1)} = \arg\max \left\{ \frac{\mathbf{w}^T\mathbf{X}^\mathbf{T}\mathbf{X}\mathbf{w}}{\mathbf{w}^T\mathbf{w}} \right\}$$

The kth component can be found by subtracting the first $k-1$ principal components from $\mathbf{X}$:

$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X}\mathbf{w}_{(s)}\mathbf{w}_{(s)}^\mathrm{T}$$

and then finding the weight vector which extracts the maximum variance from this new data matrix

$$\mathbf{w}_{(k)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \|\hat{\mathbf{X}}_k\mathbf{w}\|^2 \right\} = \arg\max \left\{ \frac{\mathbf{w}^T\hat{\mathbf{X}}_k^T\hat{\mathbf{X}}_k\mathbf{w}}{\mathbf{w}^T\mathbf{w}} \right\}$$

THE IDEA OF K-NEARTEST NEIGHBOURS

Since we are using dimensionality reduction techniques, we can perform K nearest neighbors for clustering similar players based on the newly obtained reduced parameters. We will use the elbow curve method to determine the optimal number of nearest neighbors we need to minimize the error function whilst ensuring that the number of neighbors is not too high. We use unsupervised learning for this.

One method to validate the number of clusters is the *elbow method*. The idea of the elbow method is to run k-means clustering on the dataset for a range of values of *k* (say, *k* from 1 to 10 in the examples above), and for each value of *k* calculate the sum of squared errors (SSE).

REFERENCES FOR KNN:

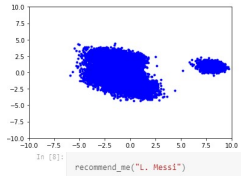
https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

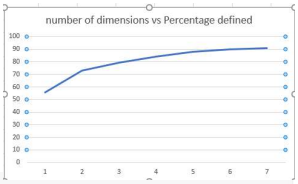
REFERENCES FOR PCA:

https://en.wikipedia.org/wiki/Principal_component_analysis

RESULTS

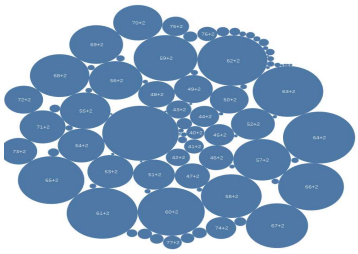
```
In [5]: explained_var
Out[5]: array([0.55520083, 0.17342652])
```



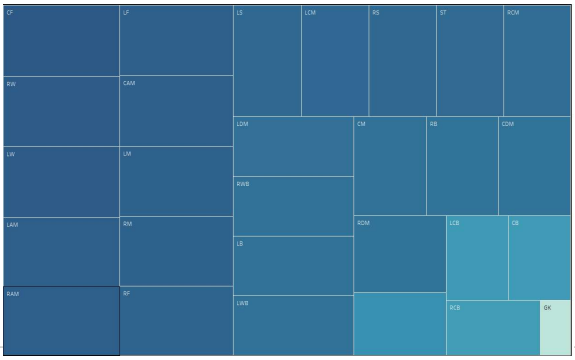



```
In [8]: recommend_me("L. Messi")
```

```
Here are 5 players similar to L. Messi :

Neymar Jr

K. Coman

D. Mertens

Cristiano Ronaldo

P. Dybala
```



Position	
Null	2.200
CAM	3.192
CB	2.095
CDM	2.571
CF	3.273
CM	2.767
GK	1.000
LAM	3.333
LB	2.619
LCB	2.057
LCM	2.883
LDM	2.676
LF	3.333
LM	3.099
LS	2.879
LW	3.343
LWB	2.639
RAM	3.111
RB	2.570
RCB	2.036
RCM	2.825
RDM	2.564
RF	3.100
RM	3.087
RS	2.897
RW	3.246
RWB	2.679
ST	2.825



CONCLUSION AND FUTURE SCOPE

Successfully generated a recommendation system for FIFA 19 dataset Open-sourced on Kaggle by EA sports. This should help the user to pick perfect players for their teams without investing a lot of money. The players with their stats and positions closest to the star players will be displayed and a very strong team can be built at much cheaper costs.

Future scope: We can have a much balanced side with better stats and chances of winning if we are provided with more parameters such as the Clubs the player plays for, Chemistry with other players in the team etc. , One can make a much more balanced team with high odds of winning at a relatively low cost.

KAGGLE PROJET REFERENCE:

https://www.kaggle.com/sahajshukla/a-recommendation-system-for-fifa-19

MALARIAL CELL DETECTION USING TRANSFER LEARNING

AUTHOR: SAHAJ SHUKLA, HARSH SHETTY

Instructor: PROF. EDWARD STOHR

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

OBJECTIVE

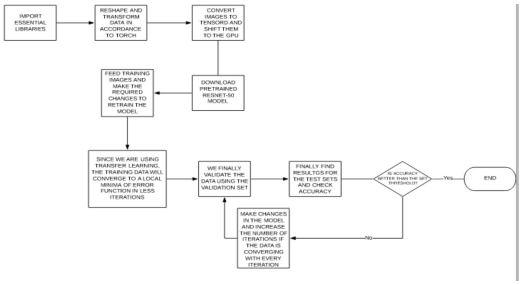
To detect the cells infected by Malaria parasite in order to detect malaria without human intervention.

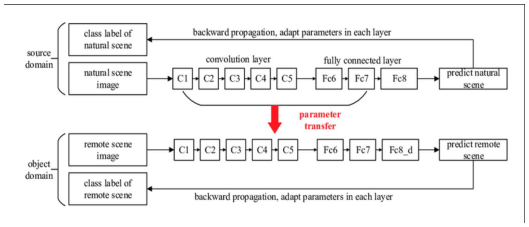
BACKGROUND

In 2017, there were an estimated 219 million cases of malaria in 87 countries. In countries with poor medical facilities, time is a key constraint.

A lot of human resources are wasted in the pathology labs to analyze the slides. This program based on neural networks will tackle this problem and automize the process of analyzing the sample slides in order to help the doctors and pathologists invest time in more crucial situations.

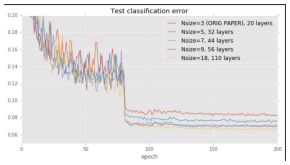
PROJECT FLOW

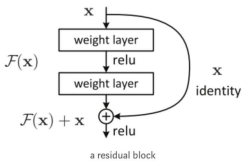




RESNET-50 MODEL

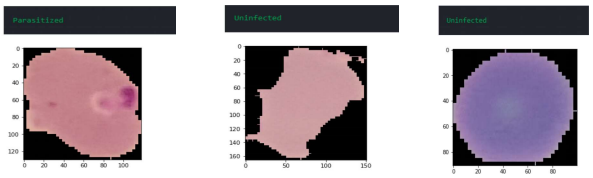
Since we are using the RESNET-50 Model as our transfer learning pre trained model, it is crucial to mention about the architecture of RESNET-50. Adding extra layers in a dataset does not always ensure a better performance. This problem is tackled in the RESNET model by ensuring skip connections after 2 layers to ensure that the performance does not degrade and remains the same even in the worst case scenario.





We use Torch (an open source machine learning library by Facebook) to carry out the operations with the neural network. The main idea of using torch is its speed and flexibility. Also, torch helps us draw dynamic maps before executing all iterations. Also, it is relatively light in comparison to its TensorFlow and would perfectly fit the needs of this program

RESULTS

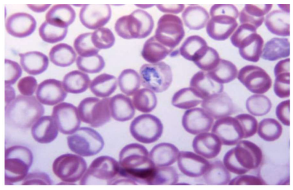


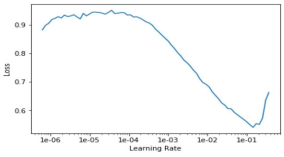

```
Test Loss: 0.331059

Test Accuracy: 85% (2359/2756)
```

IMPORTING RESNET-50 MODEL

```
Downloading: "https://download.pytorch.org/models/resnet50-19c8e357.pth" to /tmp/.torch/models/resnet50-
19c8e357.pth
100%|▉▉▉▉▉▉▉▉▉▉▉▉▉▉▉▉▉▉▉▉| 102502400/102502400 [00:01<00:00, 89692748.86it/s]

ResNet(
  (conv1): Conv2d(3, 64, kernel_size=(7, 7), stride=(2, 2), padding=(3, 3), bias=False)
  (bn1): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (relu): ReLU(inplace)
  (maxpool): MaxPool2d(kernel_size=3, stride=2, padding=1, dilation=1, ceil_mode=False)
  (layer1): Sequential(
    (0): Bottleneck(
      (conv1): Conv2d(64, 64, kernel_size=(1, 1), stride=(1, 1), bias=False)
      (bn1): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
      (conv2): Conv2d(64, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bias=False)
      (bn2): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
      (conv3): Conv2d(64, 256, kernel_size=(1, 1), stride=(1, 1), bias=False)
      (bn3): BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
      (relu): ReLU(inplace)
```

TRAINING SLIDE





CONCLUSION

We have automized the procedure of malarial detection using convolutional neural networks and transfer learning with pretrained RESNET-50 model.

FUTURE SCOPE

This model can be greatly improved by various methods and the accuracy rate can be improved. For this separate models may be trained and saved. Also, the improved model can be adopted by various organizations in several countries where health facilities are scarce and malaria like diseases are a great problem.

Tip Prediction

Authors: Sahaj Shukla, Harsh Shetty, Zhaonian Zhu, Chenxu Yi and Sonia

Instructor: Prof. Edward Stohr

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction

1. The objective here is to create a model so as to optimize and get the maximum tips from the customers
2. Introduction of various kinds of offers on specific days of the week is the prime way we intend to attain our goal
3. A critical challenge would be to bifurcate the customers based on gender, smokers/non-smokers, number of people on a table etc.
4. Creating an optimized model will help the restaurant to target a specific nature of customers with attracting offers

Technology

Using Python to analyze the dataset:

1. Backward elimination to select significant features.
2. Multiple linear regression to predict the tip amount.
3. R-Square to test the fitness of our model.

Using Tableau to visualize the data. Backward elimination: Which Involves starting with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable (if P value is greater than α (0.05)) whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically significant loss of fit

The Formula for Multiple Linear Regression Is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = expanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

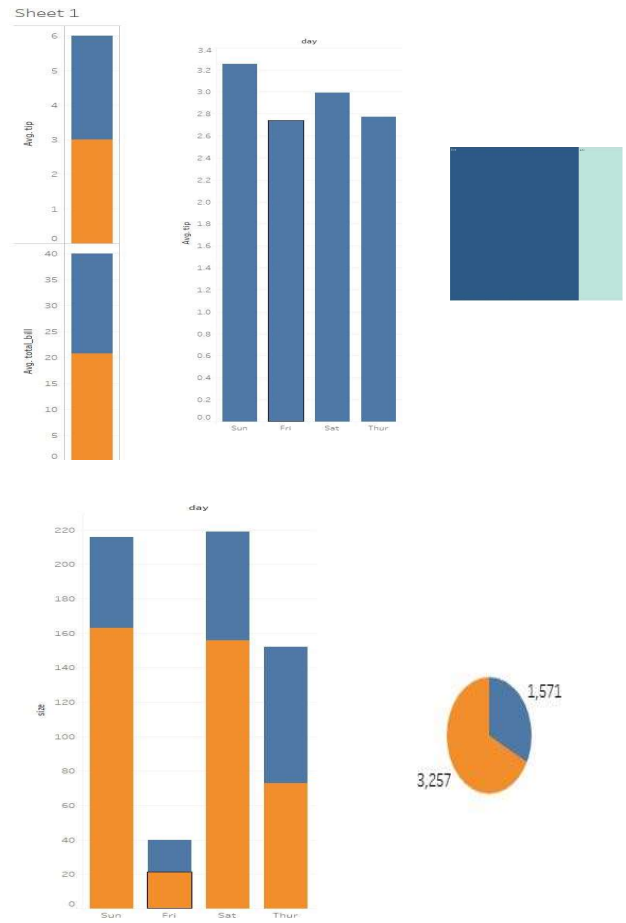
ϵ = the model's error term (also known as the residuals)

R2 is a percentage of response variable variation that is explained by its relationship with one or more predictor variables. In simple words R2 indicates the accuracy of the prediction.

$$R^2 \equiv 1 - \frac{SS_{error}}{SS_{total}}.$$

Results

Below are the graphs obtained from Tableu. We can see average tips are the highest on Sundays, from which can assume that when customers visit with their families or friends (basically a group), they generally tip more than on the weekdays. We also found that male smokers had a higher bill, but tipped significantly lesser than male non-smokers. A statistic that the restaurant will not mind, but for our agenda it poses a concern. 75% of the revenue is from dinners. But the customers who have lunch tend to tip slightly more than the latter. On an average there are more males than females on each day except for Thursday. One key observation made was that, females (15.69%) tip more than males (14.87%).



Conclusion

1. In order to optimize the tips, we conclude that several things can be done
2. The restaurant can keep special offers for women on Thursdays, since we observed that women frequent on this day of the week. Also, they can arrange for a few family-centric activities that can be done on the weekends since families tend to come more on Sundays.

Customer Churn Rate Analytics: Predictive Analysis Model for Retaining Customers

Shangjun Jiang
Instructor: Amir H Gandomi

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

INTRODUCTION

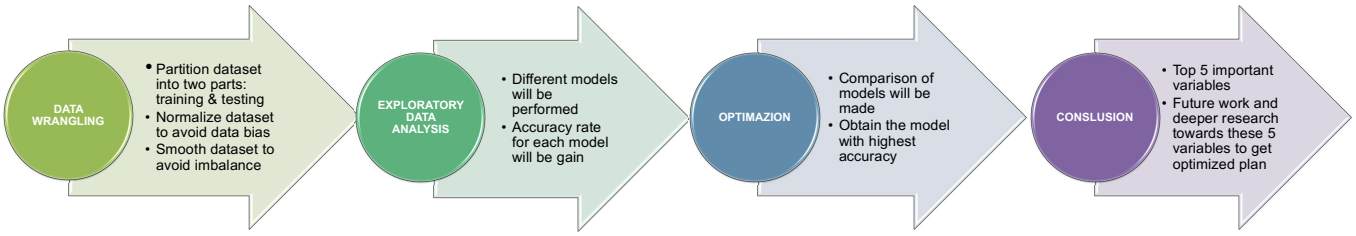
- **Key Words:** churn rate, classification analysis, optimization
- **Background:**

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 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.

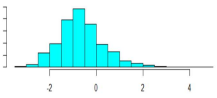
PROCESS FLOW

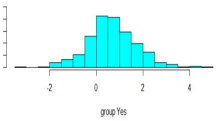


MODELLING

Linear Discriminant Analysis

```
Coefficients of linear discriminants:
                          LD1
accountlength        6.706381e-01
numbervmailmessages  1.945667e+00
totaldayminutes      1.273053e+03
totaldaycalls        5.480808e-01
totaldaycharge      -1.269688e+03
totaleveminutes     -7.447574e+01
totalevecalls        7.209265e-02
totalevecharge       7.623507e+01
totalnightminutes    2.215381e+02
totalnightcalls     -1.843684e-01
totalnightcharge    -2.200525e+02
totalintlminutes     8.449952e+01
totalintlcalls      -1.097834e+00
totalintlcharge     -8.349534e+01
numbercustomerservicecalls  4.241855e+00
internationalplanyes  1.741883e+00
voicemailplanyes    -1.761354e+00
```

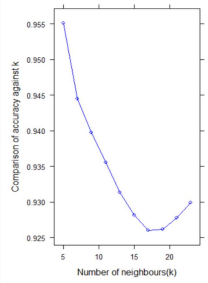





```
> table2
         Actual
Predicted  No Yes
     No   928  45
     Yes  293 167
> sum(diag(table2))
[1] 0.7641312
```

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

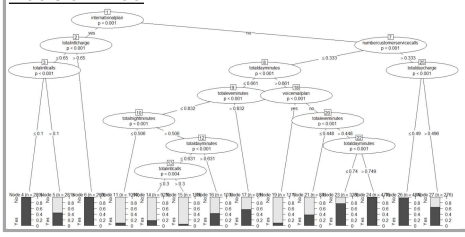
K-Nearest-Neighbor Analysis




```
> confusionMatrix(p_knn,testing$churn)
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  944  52
       Yes 277 160

               Accuracy : 0.7704
                 95% CI : (0.7477, 0.792)
    No Information Rate : 0.8521
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3669
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7731
            Specificity : 0.7547
         Pos Pred Value : 0.9478
         Neg Pred Value : 0.3661
             Prevalence : 0.8521
         Detection Rate : 0.6588
   Detection Prevalence : 0.6950
      Balanced Accuracy : 0.7639

       'Positive' Class : No
```

For KNN, we choose 5 as 'K' base on the ROC of the model. The confusion Matrix of testing dataset as shown on the left with accuracy rate as 77%.

Decision Tree



```
> confusionMatrix(p_decision,testin
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1081   53
       Yes  140  159

               Accuracy : 0.8653
                 95% CI : (0.8465,
    No Information Rate : 0.8521
    P-Value [Acc > NIR] : 0.08306
```

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

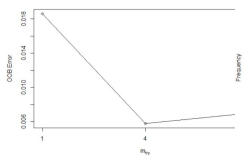
Logistics Regression

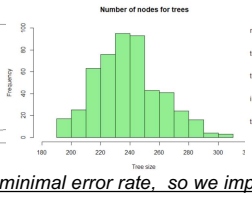
```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                -7.4021     0.3705 -19.981  < 2e-16 ***
accountlength               0.8230     0.2117   3.888 0.000101 ***
numbervmailmessages         2.6274     0.5181   5.072 3.95e-07 ***
totaldayminutes          1885.7622   722.1815   2.611 0.009022 **
totaldaycalls               0.6772     0.2810   2.410 0.015936 *
totaldaycharge          -1881.2581   722.2323  -2.605 0.009193 **
totalevecharge              2.4855     0.2555   9.727  < 2e-16 ***
totalnightminutes           1.8836     0.2835   6.645 3.03e-11 ***
totalintlminutes          103.2883    64.9816   1.589 0.111948
totalintlcalls             -1.4553     0.2802  -5.193 2.07e-07 ***
totalintlcharge         -101.9909    64.9846  -1.569 0.116540
numbercustomerservicecalls   5.6721     0.2317  24.479  < 2e-16 ***
internationalplanyes        2.4278     0.1144  21.229  < 2e-16 ***
voicemailplanyes           -2.3634     0.3138  -7.532 5.01e-14 ***
```

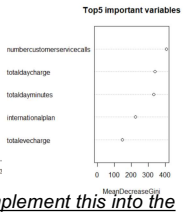
```
[1] 0.7620377
> logit_table_test
         Actual
Predicted  No Yes
     no   921  41
     yes  300 171
```

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

Random Forest

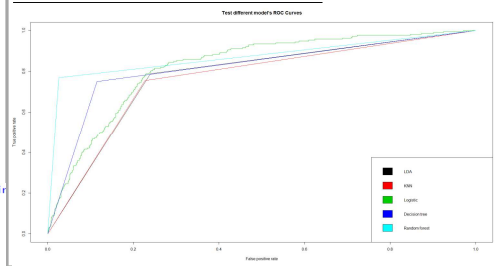






When mty = 4, we get the minimal error rate, so we implement this into the model. Then we get the top 5 significant variables in this model. The accuracy rate for Random Forest is 94%.

COMPARISION OF 5 MODELS



Base on the ROC curve as shown on the left, we can se that the Random Forest is the most accurate model.

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.

News Article Genre Prediction

Jainam Sheth
Instructor: Lappas T

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Motivation

•Initially Publication and author of the article played a crucial role in implementing prediction algorithm

•Analysis of words using lexicon library from the content had a significant contribution to develop the solution model

•The inherent combinatorial complexity, the presence of non-linearities and the requirements for 0-1 variables head the list of challenges.

Technology

•Excel for cleaning the unwanted values and normalizing the dataset for avoiding glitch

•Python for splitting the dataset into small groups for achieving high accuracy and building regression models.

•Grid Search for analyzing best classification model for predicting the dataset with high accuracy.

Current & Future Work

•Incorporate merging of publication and discarded unwanted articles with minimal contribution to dataset for data uncertainties.

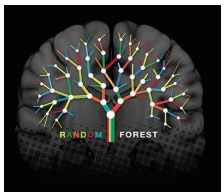
•Development of a unified optimization and framework for various authors of publication depending on the content of the article.

•Used Python for building prediction model and Excel to normalize the data to achieve better accuracy.

•Development of a prediction algorithm to account for multiple authors/publications which can predict number of articles published by an author in any publication with certain period.

Random forest

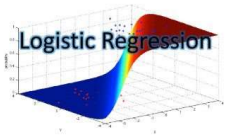
Accuracy: 65%



- **Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble.**
- **A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.**

Logistic regression

Accuracy: 75%



- **Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.**
- **In regression analysis, logistic regression(or logit regression) is estimating the parameters of a logistic model (a form of binary regression).**

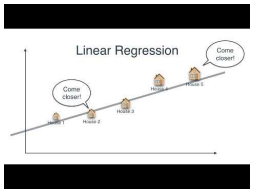
Decision tree

Accuracy: 59%

- **Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.**
- **We used statistical methods for ordering attributes as root or internal node.**
- **For our dataset, building a data regression model was more significant.**

Linear Regression

Accuracy: 78%



- **SGD Classifier implements regularized linear models with Stochastic Gradient Descent.**
- **This can perform as logistic regression and as faster as hyper tuning**

Conclusion:`

As the dataset that we were working on contained over 100k rows, we needed an algorithm that ensured a high accuracy. Logistic Regression gave the highest accuracy, but it was not much efficient to handle a large dataset. We tuned SGD Linear regression that gave us accuracy as high as Logistic regression and time complexity also reduced to a significant level.

Relationship Between World Market Indices and U.S Market Movement Prediction

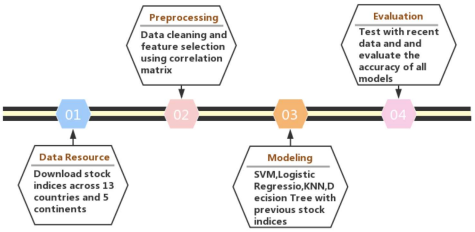
Jiali (Andrea) Chen,Rui Wang,Oluwadamilola Seweje
Instructor: Dragos Bozdog

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introducion

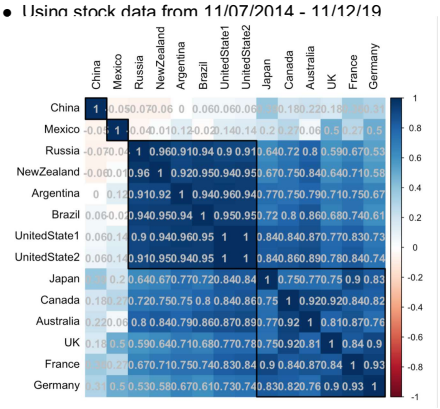
- Machine learning has gained more traction in recent years for tracing and predicting trends in the stock market.
- In this project, we use global stock indices data as parameters to predict the U.S stock market in machine learning algorithms to see if we get a better prediction.

Approach

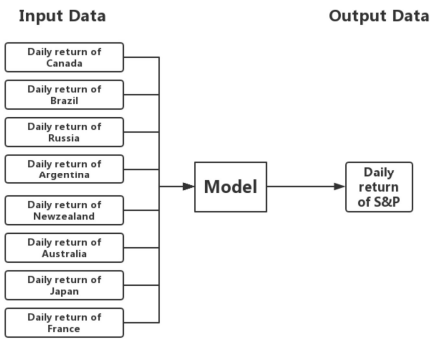


Data Imputation

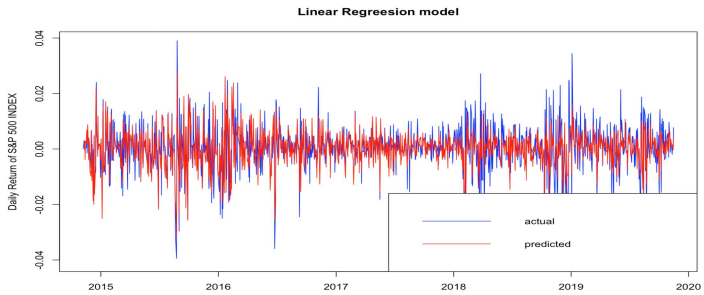
- We have selected stock indices from stock markets of 13 countries and 5 continents as parameters for machine learning algorithm to predict the U.S stock market.
- Using stock data from 11/07/2014 - 11/12/19

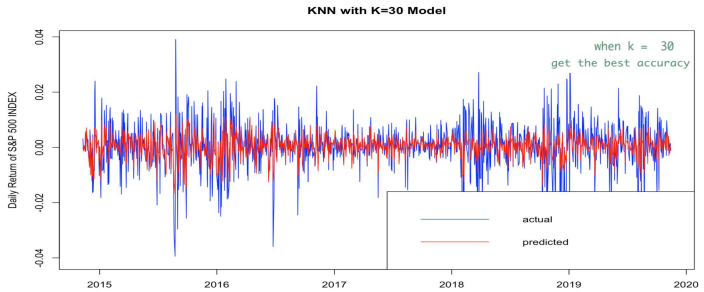


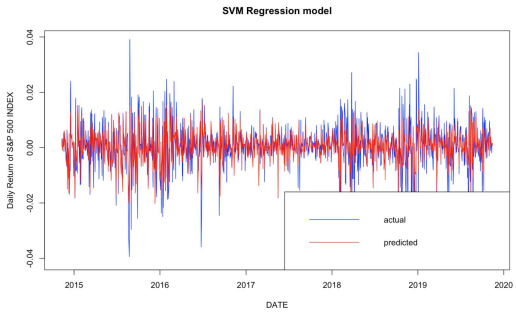
Feature Selection



Result







Model	Accuracy
Linear Regression	57.30%
KNN	61.65%
SVM	60.70%

Conclusion

- The correlation plot shows us North American and European markets are more strongly connected among themselves compared to other regions. The integration of Asian markets among other countries is found to be rather weak.
- Various machine learning based models are proposed for predicting daily trend of US stocks. Numerical results suggests high accuracy

Future work

- Explore other creative and effective methods that might yield even better performance on stock market forecasting.
- Models can be modified to take care of the currency,commodity(like silver,oil,gold), tax and fees in the trading process.
- Investigate the mechanism of short sale and maximize our profit even when the market is bullish.

RFM model: Customer Segment & Quantify Customer Value

Xintong Hou, Jiaqing Wang, Rui Mei, Rui Li, Jiahao Chen, Yixuan Wang
Instructor: Elizabeth A Gomez

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

Motivation

To identify:

1) which customers are the premium customers?
2) which customers are of high quality and risk of churn?
3) which customers are most likely to repurchase?
4) which customers are unprofitable?

Customer Type	R	F	M	Description
Best Customers	1	1	1	Premium Customers; Priority Processing
High-Spending Customers	0	1	1	No recent transactions; Emergency recall strategy should be adopted
Potential Customer	1	0	1	Key identification customers; Implement promotion strategy and make them re-purchase
Customer Retention	0	0	1	Potential valuable customers; Adopt retention marketing strategy
Active Loyal Customers	1	1	0	Customers with small contribution; Adopt promotion strategy
New Customers	1	0	0	Adopt the strategy of brand introduction and consumption stimulation
General Customers	0	1	0	Adopt advertising and promotion strategies
Churn Customers	0	0	0	Non-marketing focus; It's secondary work to recall them

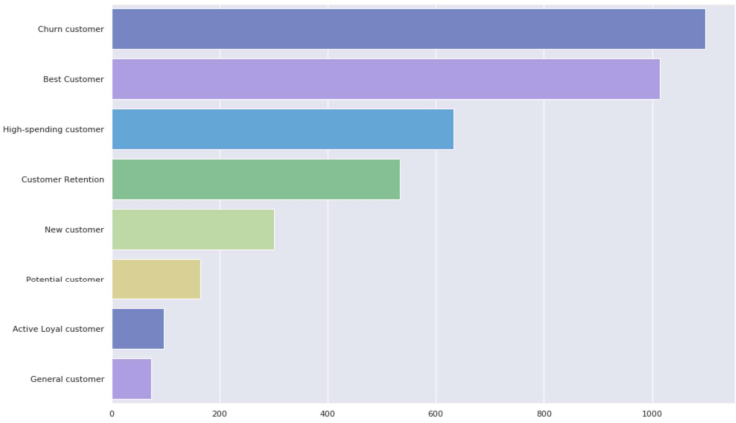
Current & Future Work

•Accurately identify customers, measure customer value and serve customers in a hierarchical way.

•Improve the activity, retention and payment rates of different types of users through various operational strategies.

•When different types of strategies are applied, the core indicators of the model should be changed.

•Select important customers to implement marketing strategy.



Technology

•Introduction to RFM model:

1) R(Recency): the interval between the last transaction of a customer. The larger the value, the longer the date the customer transaction took place.

2) F(Frequency): the number of transactions a customer has made in a year. The higher the value, the more frequently the customer trades.

3) M(Monetary): the total amount of a customer's transaction in one year. The larger the value, the higher the customer value.

•Python to make the data cleansing and visualization using Pandas, Heatmap, Matplotlib and Seaborn.

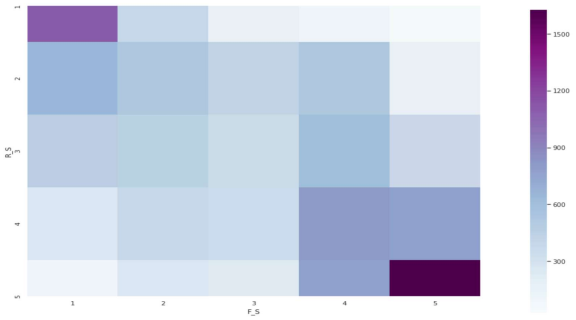
•There are many analysis tools could be used for RFM. SPSS Modular, SAS even Excel could be used to conduct modeling analysis and then dig deeper.

Conclusion---Frequency histogram

•Customer loss accounts for the largest proportion, which is also a challenge faced by many businesses.

•The proportion of important value customers is also very large, which indicates that the company's business is developing well.

•The proportion of new customers and potential customers is not high, indicating that the company may need to increase advertising and publicity efforts to attract more new customers.



Conclusion---Heatmap

•The three indicators of the company's current customers' RFM have obvious commonalities.

•The higher the R_S score and F_S score of the users, the larger the corresponding transaction value will be.

•The more right the ladder is, the higher the user loyalty is, and the lower the ladder is, the higher the user stickiness is.

Reference

1. Dataset: 'A transnational dataset of Online Retailer', Daqing Chen, London South Bank University. https://archive.ics.uci.edu/ml/datasets/Online+Retail#
2. Jim Sellers, Arthur Middleton Hughe. RFM Migration Analysis: A New Approach to a Proven Technique (Nov. 2019). http://www.dbmarketing.com/articles/Art123.htm
3. Mobius Solutions, 'RFM Segmentation' https://www.optimove.com/resources/learning-center/rfm-segmentation

Academic Advisor(Gil)

Gan Song

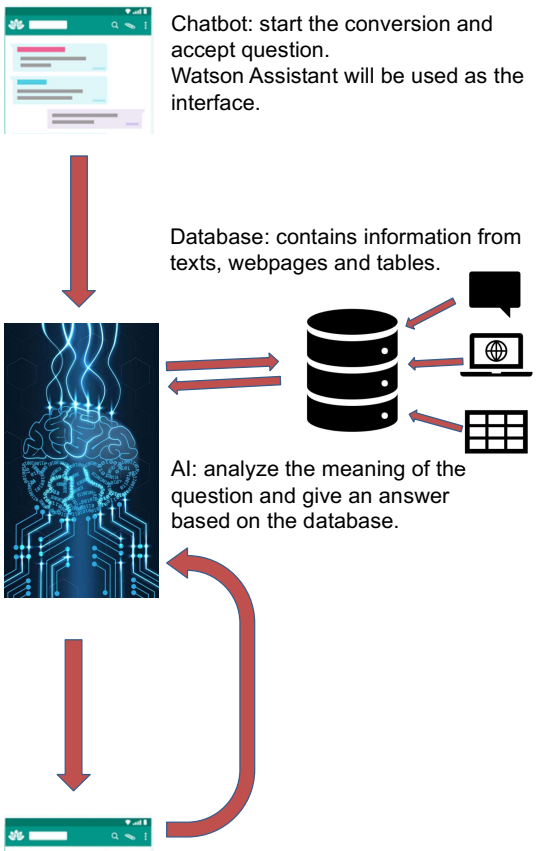
Advisor: Prof. Christopher Asakiewicz

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

Introduction

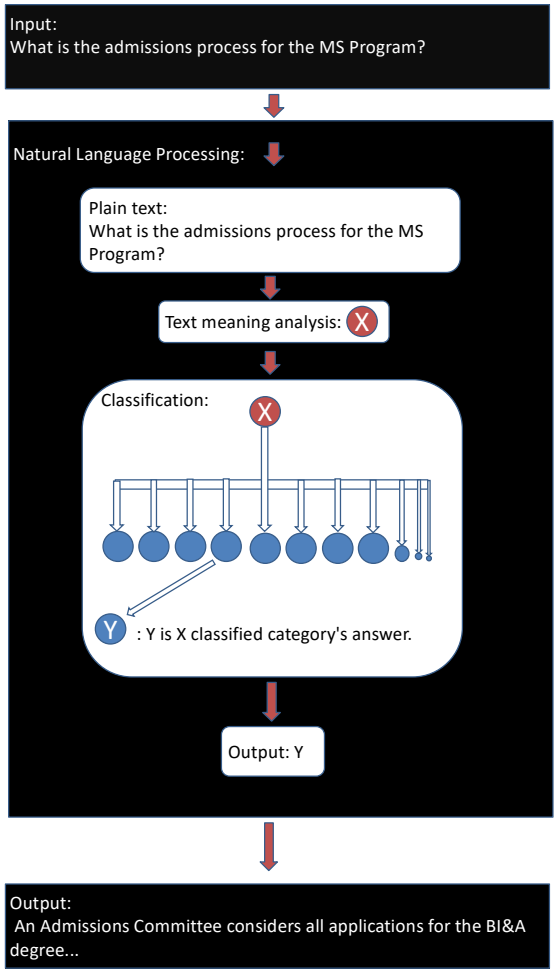
Students seeking information about their program in college often send emails to their advisors. This process is time consuming and very inefficient for the advisor, especially for most frequently asked questions. To improve the efficiency of the whole process and get a timely reply for students, the academic advisor for the BIA program, Gil, a conversational agent leveraging Watson Deep QA was created.

Structure



AI

Natural language processing is used as the AI for Gil. Gil's brain is trained using a database of BIA information, and Gil can only answer BIA related questions. Here is the anatomical map of Gil's brain.



Future work

Gil can only answer BIA related questions, but Gil should be able to expand its knowledge base with information from other academic programs or areas. So a universal web crawler for getting web information will be created to build the database for the other programs. In the future, Gil will be able to save time for students seeking information.

Watson Schedule Assistant

Chang Su, Yarong Liu, Shu Li
Advisor: Prof. Chris Asakiewicz

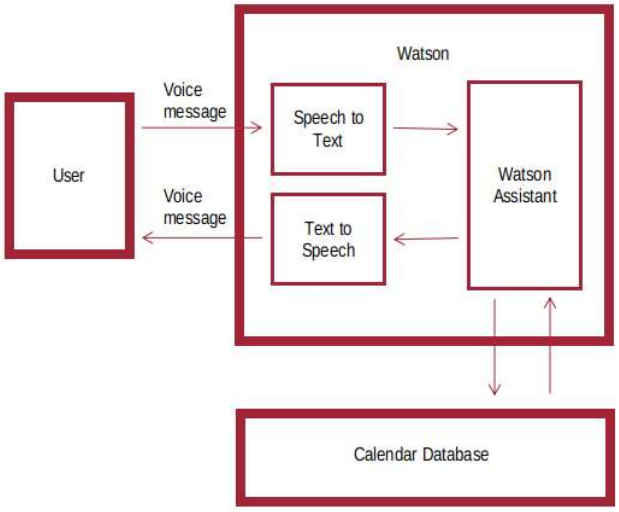
STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

Introduction

There used to be a time when arranging a schedule meant pulling out a notebook and flicking through the pages. It meant looking past the scribbles and the unreadable handwriting to find a spare space. It meant seeing the appointments you missed because… Well, forget your notebook and your handwriting this time.

Talk to Watson Schdule Assistant check what you gonna do today or any of your spare time.

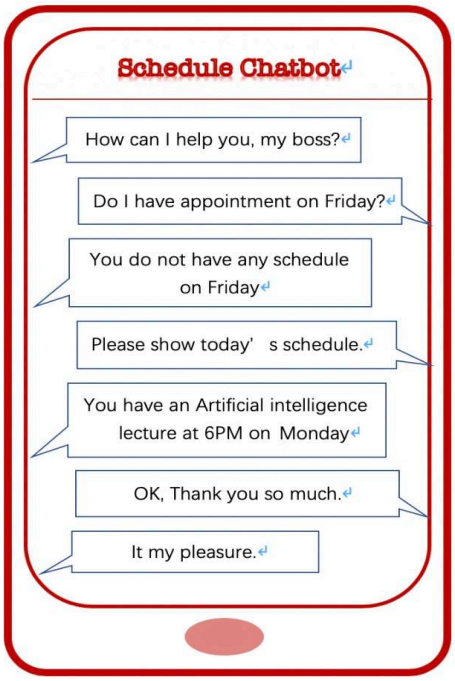
Application Architecture



To use the application:

Users would talk to the assistant, ask about their schedule. A user's voice would be converted to a text message, and Watson Assistant would analyze the text and send a request to the database to get the calendar data. Watson Assistant would then return data from database and convert it to voice message.

Result Display



This picture shows how we chat with Watson Assistant to check a personal calendar.

Business Impact

- Use Watson Schdule Assistant to improve your business and revenue.
- Use Watson Schdule Assistant to manage workflows.
- Use Watson Schdule Assistant so people can focus on more important things, boosting personal productivity.

Cognitive Student Advisor

Authors: Angad Singh, Hitesh Shetty
Instructor: Prof. Chris Asakiewicz

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Background

- Our goal was to develop a chatbot that would provide answers to prospective and current students.
- To do this we developed a system, that crawls the www.stevens.edu and extracts FAQs across several pages on the site.
- This system will keep the knowledge base up-to-date with the information on the Stevens website.

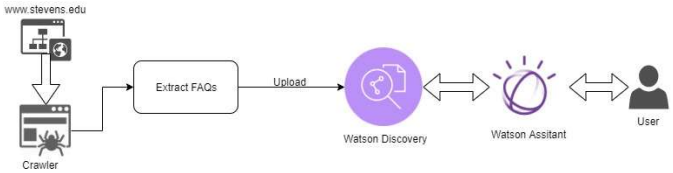
Data and Scope

- We were able to extract approximately 4K question and answers.
- As a proof of concept we have limited our project to a limited set of FAQs from the Stevens website.

Methodology

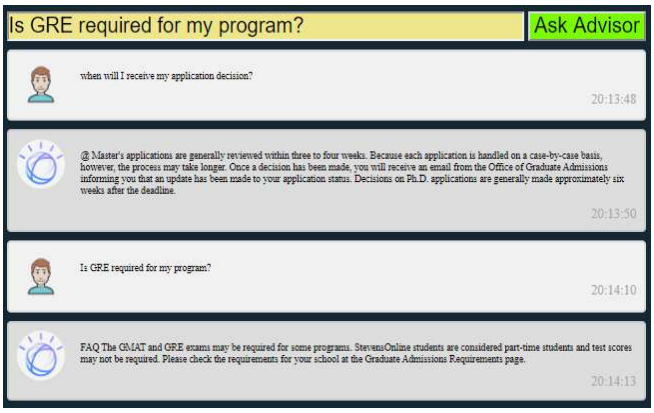
- Crawl www.stevens.edu to extract FAQ leveraging the pattern in the structure of the html.
- We limited our project to test with 1000 FAQs, since academic account is limited to a 1000 documents in Watson Discovery.
- Load the answer html documents to Watson discovery via the API.
- Train Watson Discovery in querying the knowledge corpus for improved response.

Architecture



- The extracted FAQs are uploaded to Discovery collection.
- Watson Assistant is integrated with Discovery to query each of the user's questions.
- Discovery will use NLP technique to understand the question and answer it with an appropriate level of confidence.

Cognitive Student Advisor



Future Scope

- Ingest a larger set of FAQs to Discovery to answer more questions.
- Train Watson Discovery further to produce answers with higher level of confidence.
- Create a sophisticated dialog flow for handling short tail questions.

STEVENS Jr WATSON ASISTANT

Albert Alaluf, Zhentao Du
Advisor: Prof. Christopher Asakiewicz

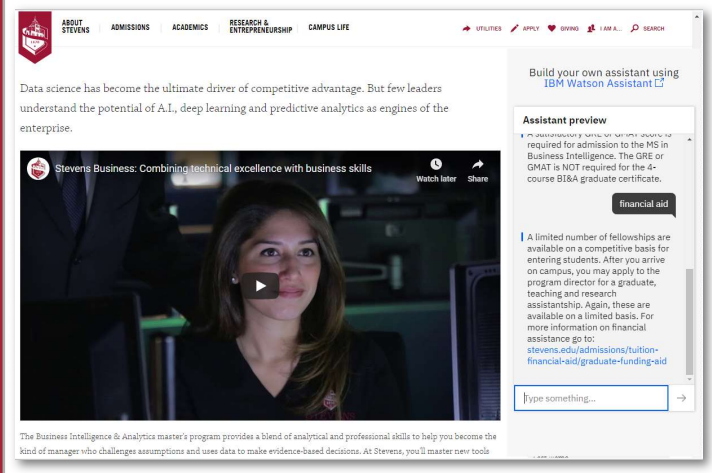
STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

THE BUSINESS PROBLEM

Every day admissions officers are answering the same questions from applicants visiting Stevens' admission site.

BUSINESS PROBLEM

- Spending hours to answer the questions
- Inefficiencies to complete their daily tasks
- Losing time and money

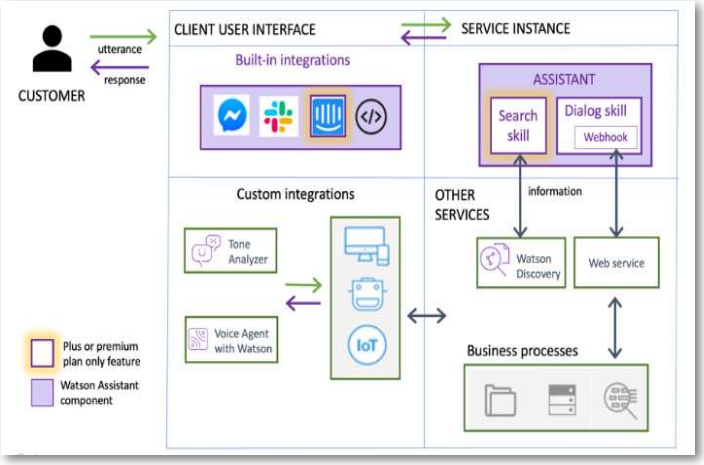


BIA main page with STEVENS Jr WATSON Assistant

STEVENS Jr Watson is an AI supported cognitive agent which can answer questions from multiple applicants, on a 24x7basis.

STEVENS Jr WATSON SERVICE

Uses IBM Watson Assistant, and Discovery Service.



1. Applicant asks a question of the agent
2. Agent replies if the answer is basic; but if it is complex, the agent uses the Discovery Service to find highly relevant answer.
3. In case the result is of low confidence, the agent connects the applicant with a live person.
4. The Knowledge Base is updated with the new questions and trained to improve the service accuracy.

An Analysis of Topic Prediction Indicators in the Field of Big Data

Tingyi Lu, Qiaoli He, Xinjie Tang
Instructor: Dr. Chris Asakiewicz

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Motivation

Topic analysis and trend prediction are of importance for governments and academic fundings.

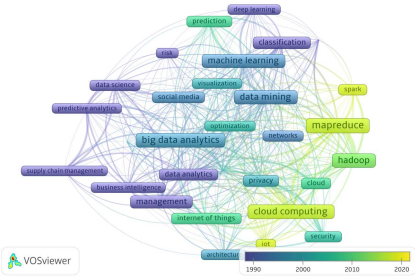
Introduction

The poster lends result of topic modeling to present an overview of most popular topics in time series
Then we select the hottest topics and try to find important indicators that are proven valid by machine learning algorithms.

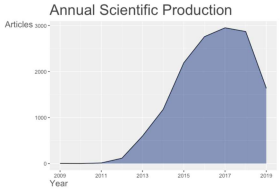
Technology & Tools

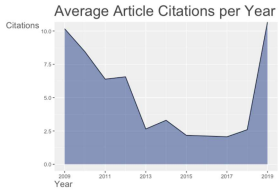
- Data mining from 14390 papers related with big data from Web of Science
- Python nltk package for data processing and topic modeling
- Vosviewer for visualization to find potential indicators
- Python sklearn package for indicators validation and model fitting

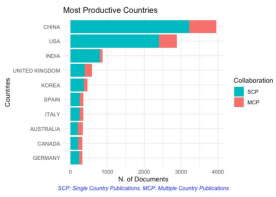
Experiment

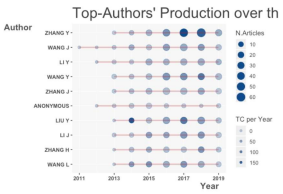


Word cloud from text mining on abstracts(left) Network graph of co-occurrence of key words(right)









Annual scientific production graph

Document production by country

Average article citations graph

Author's production by year

Conclusion

The most popular topics in the field of big data are big data analysis, data mining, cloud computing, Hadoop, machine learning, MapReduce, Internet of things, and analytics.

We believe topic modeling indicators might include region(country) of paper, year factors, authors abilities measure by h-index, author impact and citation factors under topics.

Olympics Analysis (Athens 1896 - Rio 2016)

Devanshu Chauhan

Instructor: Alkiviadis Vazacopoulos

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

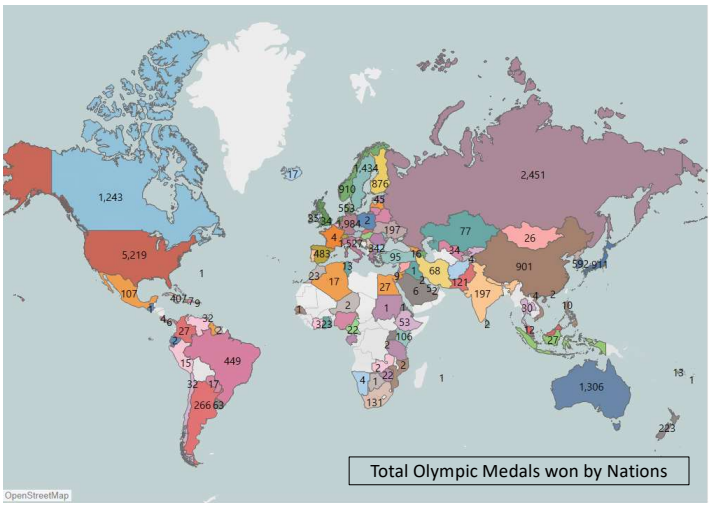
Motivation

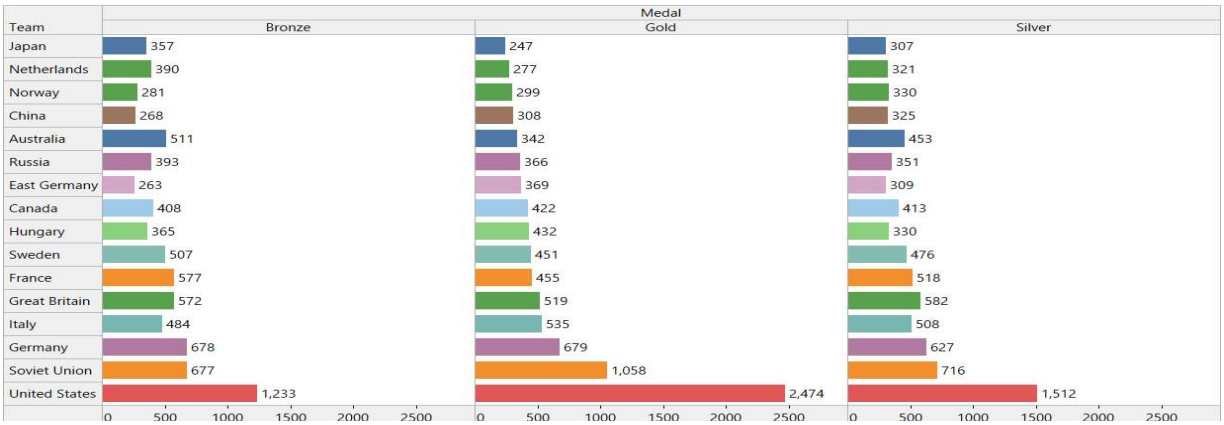
Tokyo 2020 Olympics games is just a couple of months away, and so the goal of this project is to analyze:

- Medals each country have won
- Know the Top 3 decorated Olympians of all time
- Sports in which USA is winning its most medals

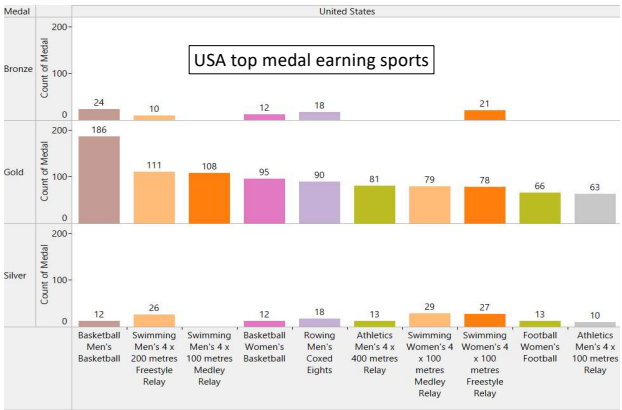
Tools

- Microsoft Excel to handle the data
- Tableau to make the necessary visualizations to understand the dataset easily





Medal Bifurcation of the Top 16 Nations



Inference

USA is not only a leader in Economic and Military Power, but its dominance in Olympics is also inevitable. USA has been doing exceedingly well in the Olympics as of 2019. It shows how committed USA is towards sports, as compared to the other nations.

With the Tokyo Olympics starting on **July 24, 2020**, let us embrace ourselves for another sports fiesta we are about to witness. Cheers!!

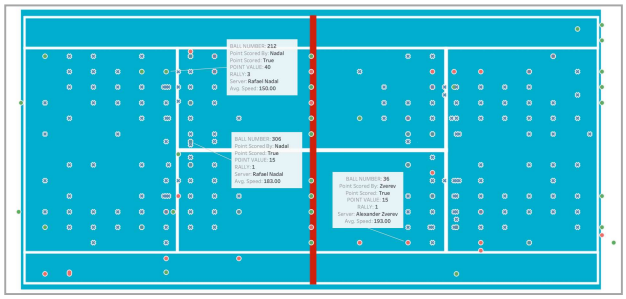
Game Analysis for Evaluating the Strategy of Players

Abhishek Singh, Aditya Shenoy
Instructor: Alkiviadis Vazacopoulos

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Motivation

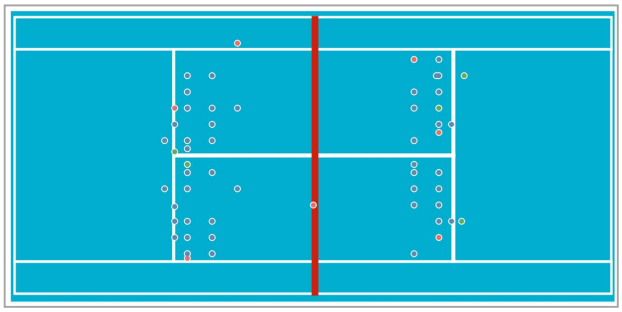
- Analyzing the advantage a serve has in a game of Tennis.
- Understanding the change of strategies employed by players during the game.
- Presenting the reasons that are responsible for the decision making of players during a match.

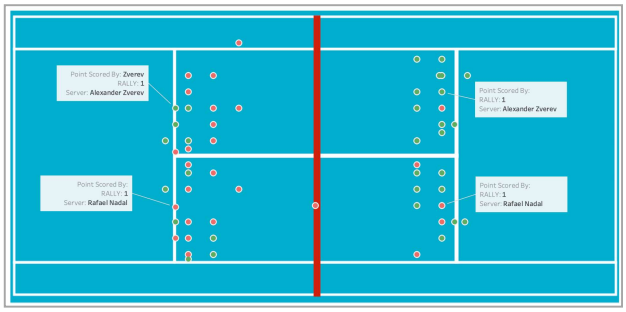


Current & Future Work

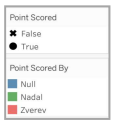
•Analyzed Ball to Ball Tennis match of Australian Open 2017 between Rafael Nadal and Alexander Zverev.
•Visualized the data to showcase the preferences of the players across multiple rounds.
•Incorporate stochastic elements to account for player movements, angle and speed to predict probability of each ball's success.
•Strategic analysis of players to find the winning chance of a player in a match between two players that hasn't occurred.

Game Analysis



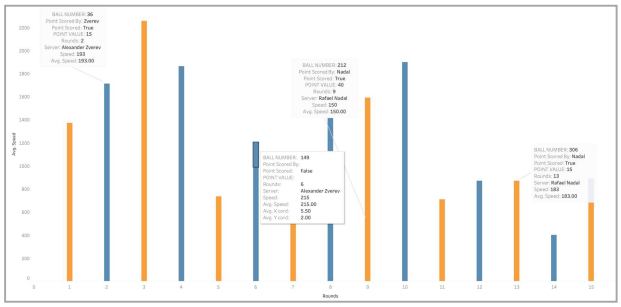


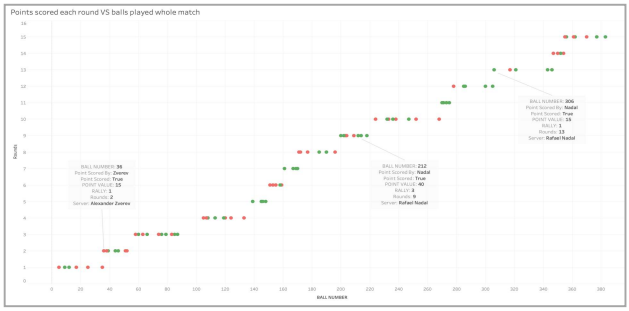
- Analyzing the serve of each player we can observe the focus of players when delivering the serve to maximize the chance of scoring a point during the serve.
- Closer the serve is to the center and ends of the serve court the higher the chances of scoring the point in that serve



Nadal serve plot on left shows he aims to deliver his serve that can maximize his chance of scoring the point.

- Analyzing Speed per round by server, We can observe that as the match progresses the utilization of speed to score a point drops focusing more on out playing opponent than brute strength.





- Analyzing Rally and Total Balls played across multiple rounds, we can observe the rallies decreases as player focus is to preserve as much energy as possible and score the point in less than 3 Rally.
- Preserving the stamina can help player utilize core strengths of the player and capitalize on the weaknesses exposed of the opponent.

Financial Distress Prediction

Heramb Dharmadhikari
Instructor: David Belanger

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

MOTIVATION

Credit scoring algorithms that assess the probability of a person to be a defaulter. Currently employed by banks to determine risk involved in granting a loan to an applicant. The algorithm implemented aims to improve the credit scoring by predicting the probability of financial distress experienced in the next two years by an applicant

TECHNOLOGY

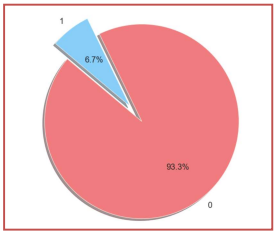
- Utilized Python panda library to read the data and obtain basic insights .
- Python matplotlib libraries for performing EDA and visualizations and for better understanding of data .
- Python libraries for applying Machine Learning Models, and evaluating them using multiple factors

CURRENT & FUTURE WORK

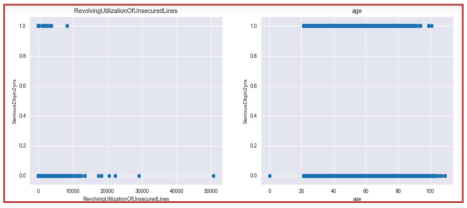
- Analyzed the dependency between multiple attributes using scatter plot
- Implemented logistic regression and random forest classifiers on an unbalanced dataset
- Balanced the dataset using SMOTE
- Implemented Random Forest classifier on balanced dataset
- Analyzed the confusion matrix and identified factors to reduce the true negative and false positives values
- Used Ensemble Methods, such as XG Boost, to improve the performance

Understanding the distribution of data

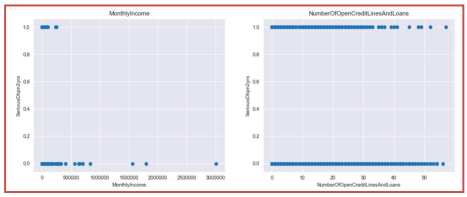
```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150000 entries, 0 to 149999
Data columns (total 12 columns):
Unnamed: 0                            150000 non-null int64
SeriousDlqin2yrs                      150000 non-null int64
RevolvingUtilizationOfUnsecuredLines  150000 non-null float64
age                                   150000 non-null int64
NumberOfTime30-59DaysPastDueNotWorse  150000 non-null int64
DebtRatio                             150000 non-null float64
MonthlyIncome                         120269 non-null float64
NumberOfOpenCreditLinesAndLoans       150000 non-null int64
NumberOfTimes90DaysLate               150000 non-null int64
NumberRealEstateLoansOrLines          150000 non-null int64
NumberOfTime60-89DaysPastDueNotWorse  150000 non-null int64
NumberOfDependents                    146076 non-null float64
dtypes: float64(4), int64(8)
memory usage: 13.7 MB
```

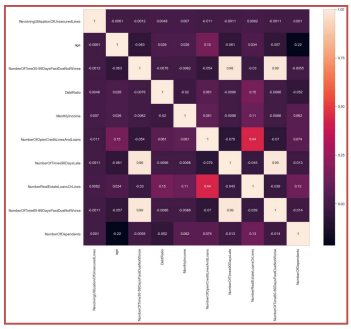


Observing the distribution the data is highly imbalanced



Using scatterplots to detect any outliers





CORRELATION HEATMAP

Reducing our features to the minimum number of most significant features. The basic rule of feature selection is that selecting features that are highly correlated to the dependent variable and also not highly correlated with each other as they show the same trend.

SCALING THE DATA

The learning model is important to scale the features to a range that is centered around zero so the variance of the features are in the same range. If the feature's variance is orders of magnitude more than the variance of other features, that particular feature might dominate other features in the dataset and our model will not train well which gives us bad model.

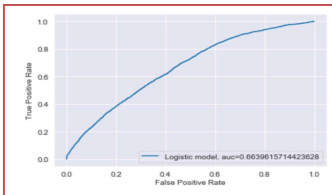
HANDLING THE UNBALANCED DATASET

SMOTE is an over-sampling method that creates synthetic **(not duplicate)** samples of the minority class, hence making the minority class equal to the majority class, Applying 'SMOTE' balances the data.

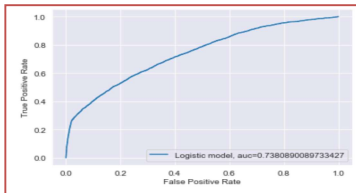
The dataset is not balanced and shows that **93%** of customers **will not** face financial distress.
In this situation, the predictive model developed using conventional machine learning algorithms could be biased and inaccurate.

After applying 'SMOTE', data is balanced. Balanced data is used for training our model and to improve the performance of our model by comparing different evaluation parameters.

Random Forrest is a bagging technique that uses Decision Tree as the base model. As we were dealing with classification problem using the Random Forest Classifier we can have significant improvement in our prediction

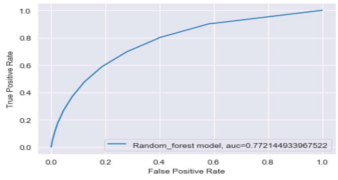



```
Confusion_matrix
 [[42031    39]
 [ 2881    49]]
Classification_report
              precision    recall  f1-score   support

           0       0.94      1.00      0.97     42070
           1       0.56      0.02      0.03      2930

    accuracy                           0.94     45000
   macro avg       0.75      0.51      0.50     45000
weighted avg       0.91      0.94      0.91     45000
```




```
Confusion matrix
 [[28641 13429]
 [ 1044  1886]]
Classification_report
              precision    recall  f1-score   support

           0       0.96      0.68      0.80     42070
           1       0.12      0.64      0.21      2930

    accuracy                           0.68     45000
   macro avg       0.54      0.66      0.50     45000
weighted avg       0.91      0.68      0.76     45000
```

```
Confusion_matrix
 [[38799  3271]
 [ 1847  1083]]
Classification_report
              precision    recall  f1-score   support

           0       0.95      0.92      0.94     42070
           1       0.25      0.37      0.30      2930

    accuracy                           0.89     45000
   macro avg       0.60      0.65      0.62     45000
weighted avg       0.91      0.89      0.90     45000
```



Predicting Probability of Loan Repayment

Team: Aditya Paldiwal
Instructor: Dr. Dragos Bozdog

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

EDA
- Data Cleaning
- Analyze Dataset Features
- Feature Visualization

Statistical Overview
- Pearson Correlation
- Hypothesis Testing
  - R-S Test
  - Z-Test
  - Chi Squared Test

Feature Engineering
- Principal Component Analysis
- One Hot Encoding
- Feature Scaling
- Mean Imputation

Optimization ⟷
- Machine Learning Pipeline
- Grid Search
- Stratified K Fold Validation
- Linear Discriminant Analysis
- Recursive Feature Elimination

Modelling
- Logistic Regression
- Random Forest
- K-Nearest Neighbor
- Multilayer Perceptron

Model Evaluation
- AUC-ROC
- Confusion Matrix

Workflow Diagram

1. Motivation

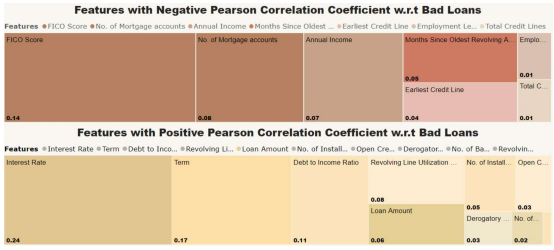
- Peer-to-peer lending is the practice of lending money to individuals or businesses through online services that match lenders with borrowers.
- The industry is valued at around $18 billion and is poised to grow at CARG of 19% to surpass $44 billion by 2024.
- However, the investors are often exposed to high credit risks without no insurance or government protection.

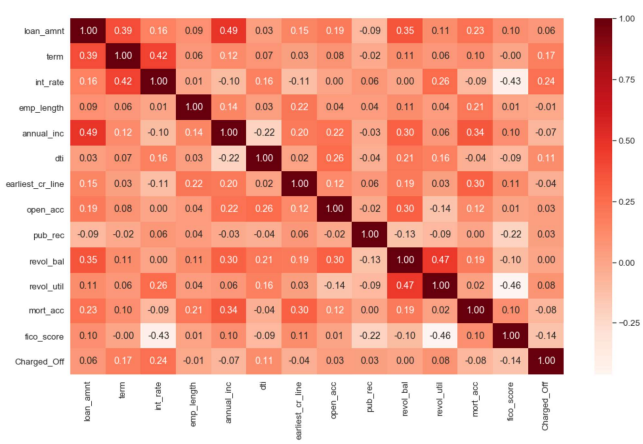
2. Objective

- To use the historical loan data to build machine learning models that can predict the probability of loan repayment (whether a loan will be fully paid back or charged off).
- Such a predictive model could help investors make better-informed investment decisions.

3. Feature Analysis

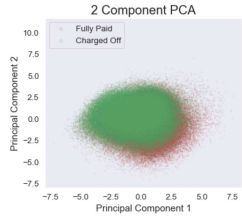
.• The variables most linearly correlated with the target variable "Loan Status" are interest rate, loan term, Fico Score and debt-to-income ratio.



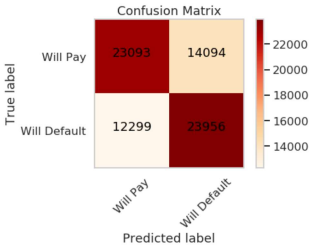


4. Machine Learning Models

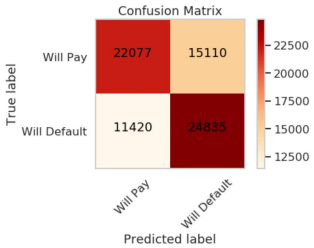
PCA Plot



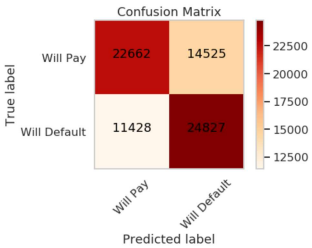
K-Nearest Neighbors



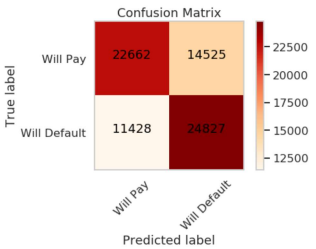
Random Forests



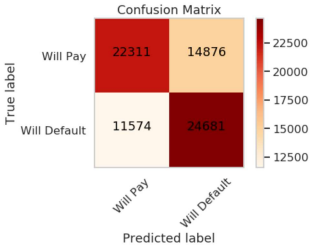
Multi-Layer Perceptron



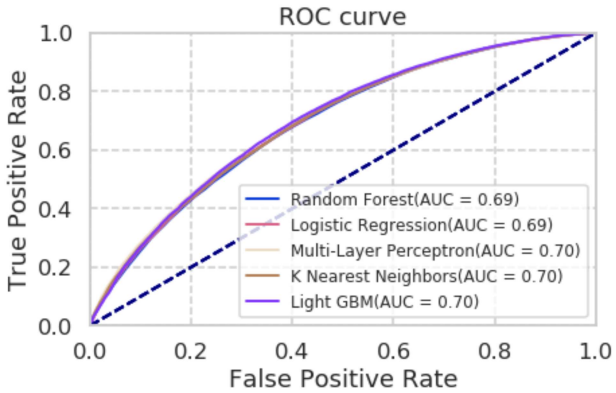
Light GMB



Logistic Regression



- The classifier Multi-Layer Perceptron has the highest accuracy score of 70.22%.



Analyzing Real-Time Streaming Twitter Data

Syed Zain Raza, Hadia Hameed, Sushmith Ramesh
Instructor: David Belanger

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY
Business Intelligence & Analytics

Motivation

- Businesses can use tweets to gain insights into customer needs. These insights can be used for understanding "Brand Perception" to improve products and launch new services.
- Organizations can connect with their target audience on a more personal level by leveraging their understanding of human thoughts and experiences through the analysis of social media interactions.

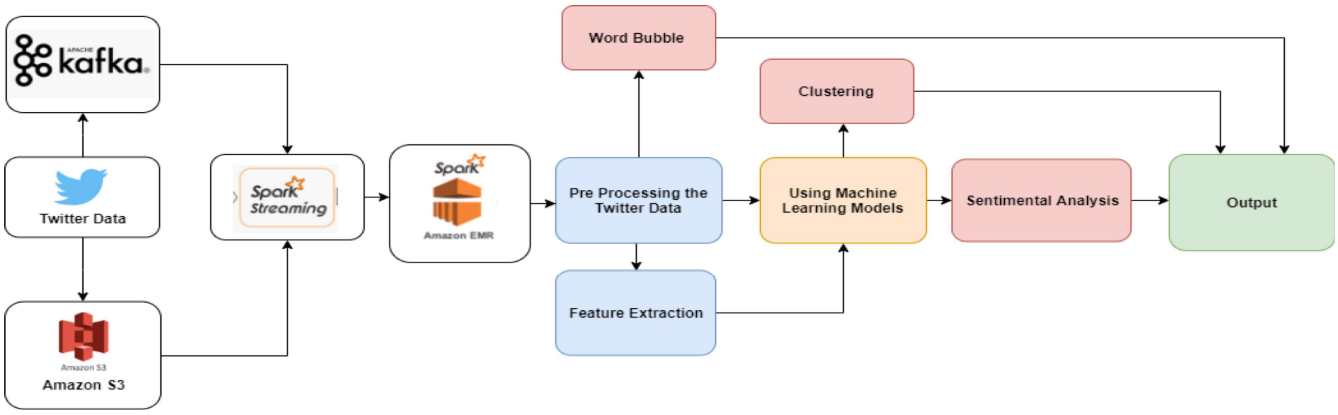
Technology

- **Twitter API:** Get streaming and static data.
- **Spark / Kafka:** Stream & perform data parallelism.
- **Amazon EMR – EC2**: Scale and increase the number of clusters achieving performance gains.
- **Python - NLTK:** Data preprocessing and visualization using Matplotlib.
- **Spark MLlib**: Use machine learning algorithms for sentiment analysis and clustering.

Current & Future Work

- Real time sentiment analysis to understand given trends.
- Visualizing word bubbles showing most frequent words for a given hashtag.
- Clustering tweets to compare user behavior and responses.

Architecture



Methods / Algorithms

- TF-IDF algorithm was applied to convert textual data to feature vectors
- K-means clustering using Euclidean Distance
- Removing stop words and punctuation from the streaming tweets and selecting top N words to make a word frequency bubble.
- Naive Bayes, Decision Trees, Logistic Regression and Deep Learning models for Sentimental Analysis
- Dividing the data into chunks to compare training time for different algorithms.
- Comparing time taken to train and perform analysis on single and multiple instances

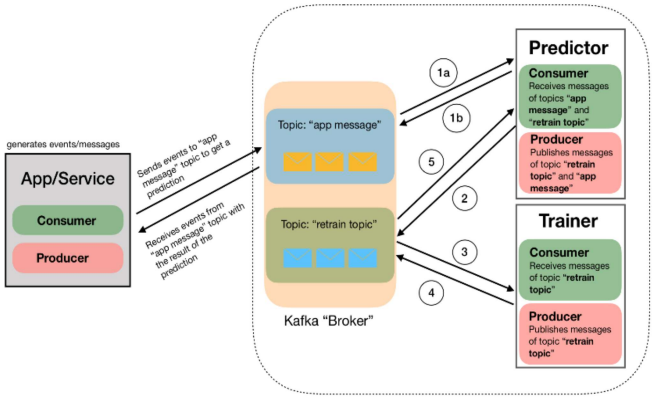
Credit Card Fraud Detection

Ayushi Chaturvedi, Abrar Alam
Instructor: David Belanger

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Motivation

- Increasing number of consumers are shifting to a non-cash society
- Advent of advanced computing power has led to the constant growth of non-cash transactions
- This implies that fraudulent transactions are anticipated to follow the same path



Introduction

- Credit card fraud occurs due to an unknown third party's attempt to leverage another person's data to make unsanctioned purchases
- About 47% of all global credit card frauds prevail in the United States
- Potential targets more often include consumers over 50 years of age

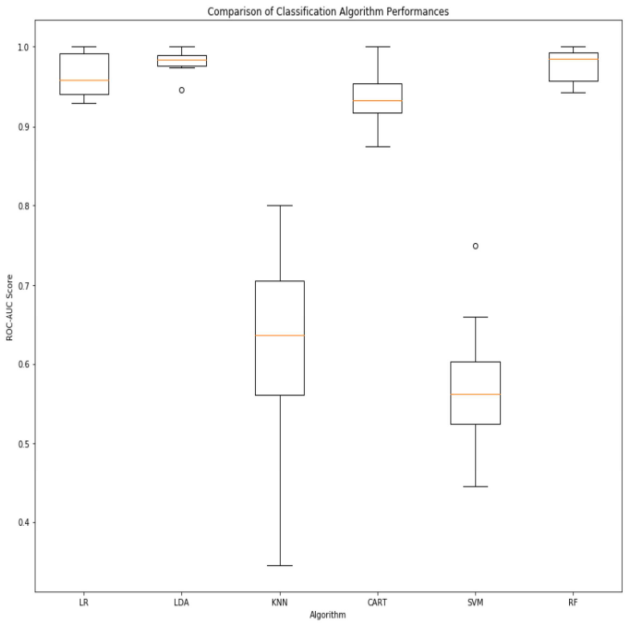
Technologies

- Amazon S3 bucket for data storage
- Amazon EC2 for running Apache Kafka and Zookeeper
- Amazon SageMaker for running notebook instances in the cloud
- Apache Zookeeper for running Kafka cluster and topics on EC2 instance
- Apache Kafka for real-time data streaming from S3

```
LR: 0.963655 (0.026242)
LDA: 0.982030 (0.014793)
KNN: 0.630392 (0.128784)
CART: 0.937942 (0.033581)
SVM: 0.570134 (0.083848)
RF: 0.975200 (0.021082)
```

Current and Future Work

- Kaggle Credit Card dataset uploaded to AWS S3 and extracted on SageMaker Jupyter Notebook instance using "boto3"
- Exploratory data analysis using Python3, including finding correlations between predictors
- Data preparation for a highly imbalanced dataset, including outlier detection and removal as well as dimensionality reduction
- Applying classification algorithms to test and compare ROC values
- Future work will include Kafka data streaming through EC2 instance



Real-Time Fraud Detection System

Mukunth Rajendran, Rashmi Khurana, Suprajah Suresh
Instructor: Dr. David Belanger

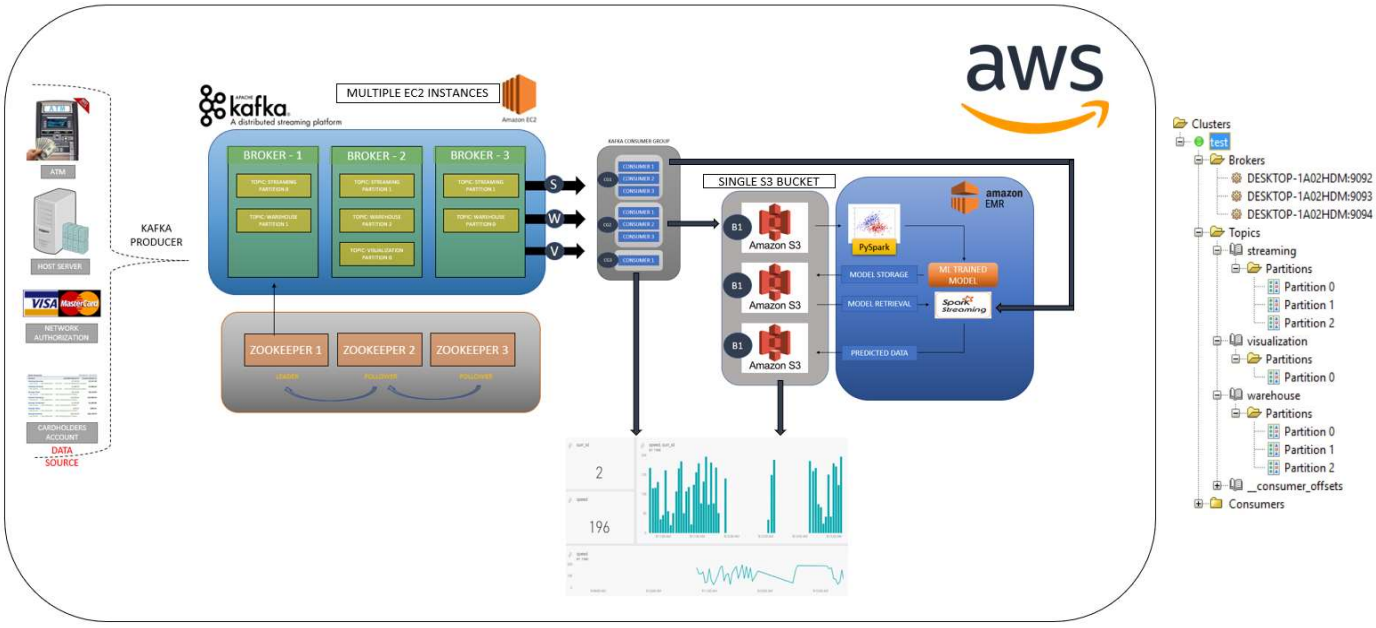
STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Objective:

Recognize fraudulent credit card transactions in real-time so that customers are not charged for unauthorized transactions. Instantaneous fraud detection is required to stop transaction completion. To build a scalable, multi-source stream processing system with parallel processing capabilities to predict fraudulent transactions, to provide system information, and to deliver visualization of data.

Data Pipeline:

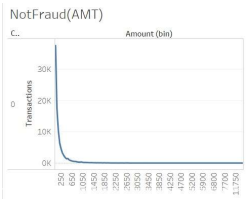
The system is built using Amazon web services to host the components on the cloud. EC2 instances are used to run a multi-node, multi-cluster Kafka structure. AWS S3 is used for storage of the streamed data, trained model and to run Spark SQL queries. Spark programs are run on Amazon EMR to predict the incoming stream. The outcome is to provide notifications for fraudulent behavior.

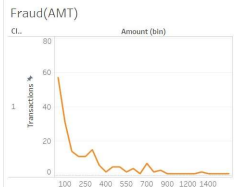


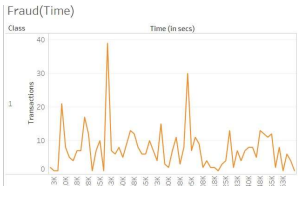
Modelling & Analysis:

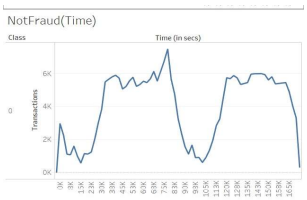
- Warehoused data is used for model building. Many Machine learning algorithms are used and the model with best performance is selected.
- The dataset is highly unbalanced but does not require preprocessing.
- Extracted the best performing model using cross-validation and saved the model in S3.
- Model is read by the Spark Streaming program and predictions are made in real-time.
- The Real-Time data is used for visualization and for triggering alarms.

Visualization:









Future Scope:

- Use other processing frameworks like Apache Flink and Apache Storm to process data in real-time and Apache Beam to provide a unified model that performs preprocessing.
- Improve performance of the system using multithreading and reactive programming for further optimization.

Analysis of single-cell RNA sequencing for various CD8 expression cells

Shiraz Bheda
Instructor: Dragos Bozdog

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Financial Analytics

Motivation

•Until recently, trends and data on gene patterns were analyzed using "Bulk-RNA sequencing" – that is, researchers would look for trends among pools of cells that typically numbered in the millions.

•Recent technical improvements have led to growing interest in an approach called "Single-cell RNA sequencing" – the analysis of gene expressions in singular cells, which occurs at a far more granular level than bulk-RNA sequencing.

•We attempt to profile genetic characteristics associated with various expressions of the cell-surface glycoprotein CD8.

Technology

•Single-cell RNA sequencing raw data is gathering using Illumina dye sequencing technology, which is used to determine base pairs in DNA and RNA.

•A formatted version of the dataset used for this analysis was published alongside a research publication from the New York Genomic Center.

•R was primarily used for data manipulation and analysis.

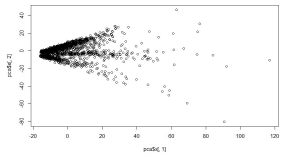
Current & Future Work

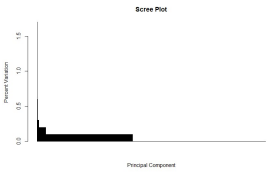
•The purpose of analyzing single-cell RNA sequencing for high, intermediate, and low cd8 expression cells, so that we may gain a greater understanding of the genetic profile and characteristics associated with key components of our immune system.

•At a high level, we also used our study to better understanding some of the unique challenges associated with analyzing high-dimensional data, alongside some of the diagnostics that ensure proper analysis.

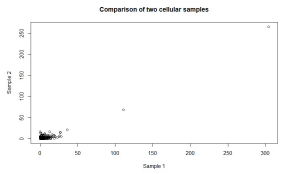
•We also incorporate traditional means of trend analysis such as PCA, multiple linear regression analysis, and clustering.

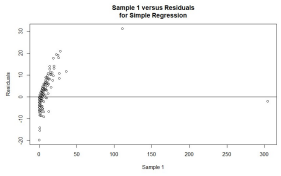
Analysis of Data Quality and Trends





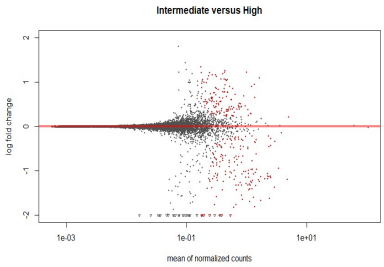
- There does appear to be a pattern in variance between the first two PCA factors; however, before proceeding, we want to analyze whether the variance captured by the first two PCA factor is representative of the entire dataset.
- While the first two PCA factors capture a relatively higher percentage of variation compared to other PCA factors, the reality is that the variation appears to be more or less equitably distributed amongst the PCA factors. Since our PCA factors fall below the threshold of statistical significance, we must come up with another method to analyze the dataset.



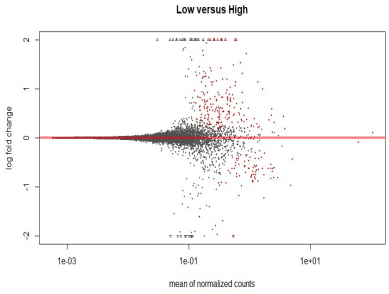
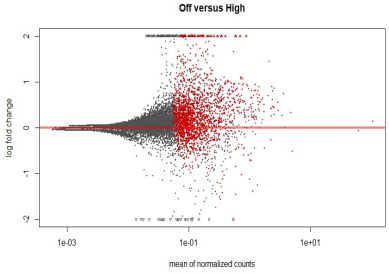


- Plotting datapoints of the first two cellular samples indicates that a relationship may exist. The shape of the data in the residual plot is curved, and the distribution of variance across the variables is not random at all. The plot would suggest that a nonlinear relationship exists between the two samples.

Analysis of expression patterns for various levels of expression of CD8



- Clearly, the majority of datapoints are close to zero along the y-axis in this plot. We observe similar pattern in the distribution of the "Low versus High" genetic expressions in the plot below. Therefore, the two treatment groups in both plots are highly similar in expression patterns.



- A significant number of data points (in particular, data points with p-values of less than 0.1) have a wide distribution across the y-axis. Therefore, the two treatment groups are highly differential in expression patterns.