Portfolio Milestone

Masters of Science - Applied Data Science

SUID: 455606416
March 19th, 2021
https://github.com/pete_math/portfolio_milestone

Table of Contents

Overview

This paper is the Graduation Portfolio Milestone for my master's degree in Applied Data Science at Syracuse University. It is also a summary of my growth and reflection regarding my experiences. It will showcase my knowledge about data science and machine learning through the significant academic projects I completed in the program's coursework. The program has provided various courses to prepare me to take on multiple data science related tasks such as data mining, text mining, time-series data analysis, information visualization, deep learning, etc. The highlighted academic projects cover different areas of expertise and skill sets, including machine learning, data or text mining using tools such as Python, Apache Spark, SQL, and R. This portfolio will demonstrate the achievements of the overarching learning objectives laid out by the program.

Describe a broad overview of the major practices' areas in data science.

While still a relatively young field, data science is fundamentally changing the ways organizations in all industries conduct business. For example, companies can devise marketing strategies that are targeted and effective. Banks and credit card companies can identify and even prevent potential fraud. Medical researchers are learning to spot brand-new patterns within our DNA that will continue to bring about breakthroughs in how we live and treat diseases.

Data science, namely, is a discipline about data. The courses in Syracuse's MSADS degree helped me to understand the topic involves a wide range of focuses, including data analysis, text mining, business intelligence, deep learning, machine learning, and many more. Each study area can further break down into various tasks, such as data collection, exploration analysis, data transformation and preprocessing, model development, and derived business insights. Data science is an important and complex topic that cannot be capsuled in this paper. Still, I picked academic projects covering different areas of expertise - (text mining, time-series big data analysis, information visualization, and data mining) that highlight the program's skills.

Collecting and Organizing Data

Data comes in various structured and unstructured formats. One of the program's foundational topics was exploring multiple ways to obtain data and structure it. Before any level of analysis can be performed, data must first be collected and organized into a usable format.

In classes like IST652, IST707, and IST718, I learned to obtain data from various sources and formats like CSV, JSON, and Twitter using Twitter's APIs. Also, I learned to parse web pages using libraries like python's BeautifulSoup and pandas' readHTML. In IST659 and SCM 651, we worked with office tools such as SQL Server, Excel, and Microsoft Access.

Project 1: Collecting and Organize Data within a relational database

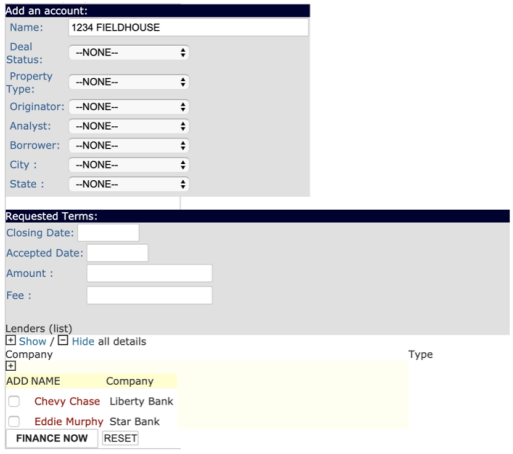
This project aimed to design a database for a web-based commercial real estate deal (sales and loans) tracking system for commercial real estate brokers (both sales and financing). The database tracked records consisting of one or more properties, the quotes an investor or lender would provide, and the fees or revenue that a broker received for each deal. The workflow consisted of moving the deal through a pipeline from prospect to close. Each milestone is called a deal status (ex: active, accepted, and closed). The deal will be placed in the lost deal status if the broker cannot find an investor or lender.

Building out the Database

The first task was to create a data dictionary of all of the different variables that needed to be collected. This data dictionary requires developing an understanding of the data and its use within the business context. This first task culminated with a conceptual model that outlined best practices for storing and organizing the different variables.

The second step was to design the logical model and the entity diagram. These diagrams are useful in visualizing how the database tables and fields relate to each other. Understanding these relationships is crucial when building a database. Upon completion, the logical model should flush out design concerns by providing an exact mapping of each relational table and the tables' primary and foreign keys used to join the tables together. At this point, each table's field data type, such as int, float, text, varchar, will also have been defined.

Finally, I generated the database and inserted test data using SQL within a stored procedure. A stored procedure is code that is saved within the database to make code reuse easier.



Learning Objectives

This project contributed to the successful execution of the learning objectives. I developed a first-hand understanding of the database development lifecycle, while creating a database and database objects using popular database management software like Microsoft SQl server and Microsoft access. I successfully proved I can collect data via a web form and organized into structured data, and stored in a relational database using SQL. This project also contributed to the ability to deliver actionable insights to subject matter experts.

Identify patterns in data via visualization, statistical analysis, and data mining.

Project 2: Identifying Patterns when evaluating an NCAA coaches salary

Introduction

The goal of this project was to perform a mock analysis to assist the Syracuse University's athletic director in understanding the key metrics that can be used to quantify the total pay for Dino Barbers as head coach of the football club. Syracuse's football club finished the 2018 season ranked 15th in the nation. 2019 was the polar

opposite. The school ended with a losing record and dropped down to 71st despite being in one of the weakest conferences, the ACC.

Syracuse is negotiating a new contract with Mr. Barber. The athletic director would like to determine if it is possible to build a linear model that can help to predict the total pay the coach should receive? The remainder of this analysis will attempt to answer the question.

Data Questions:

- What is the recommended salary for the Syracuse football coach?
- What would his salary be if we were still in the Big East?
- How good is our model?

About the Data

The primary dataset included coaches' salaries from the NCAA FBS Division 1 school. In addition, the dataset was supplemented with the following additional datasets:

- FBS Coaching Record (2019): Coach's previous season's record and all-time win-loss percentage. The number of years the coach has coached the team. Source: Supplied by the Athletic Department.
- Football Power Index (2019): The college football team ranking in 5 groups one to 5. With 1 being considered one of the top 20 schools last season and group 5 being the worse. Source: http://www.espn.com/college-football/statistics/teamratings
- Graduation School Rate (2017): While NCAA football is a big-money making venture, the majority of the students participating in the games are in school for a great education. Coaches ideally should help to drive that education. Source: https://web3.ncaa.org/aprsearch/gsrsearch
- School Stadium (2019): The school's home stadium size. https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_FBS_football_stadiums
- NFL Players (2019): The number of current NFL players the school has produced. Source:https://www.ncaa.com/news/football/article/2019-09-03/colleges-most-represented-2019-nfl-rosters

Once the data was scrubbed and merged, the final features were selected with 114 teams set for evaluation.

Total Pay

Before delving into the relationships of each feature with total pay. I looked at what the average total pay per conference looks like, as illustrated below. With the exception of a

few notable programs, the majority of coaches make close to 2 million per year. It should be expected that going into negotiations, the baseline for an experienced coach within the Power 5 (first 5 schools listed above) conferences will start there.

Conference	Sum of TotalPay	Count of Coach	Min of TotalPay	Max of TotalPay	nfl_roster	teams	avg
SEC	$65,008,379.00	14	$2,350,000.00	$8,307,000.00	339	14	$4,643,455.64
Big Ten	$60,256,192.00	14	$1,830,000.00	$7,600,000.00	253	14	$4,304,013.71
ACC	$48,073,154.00	14	$1,831,580.00	$6,543,350.00	215	11	$4,370,286.73
Big 12	$36,163,301.00	10	$1,701,109.00	$5,500,000.00	131	10	$3,616,330.10
Pac-12	$34,681,433.00	12	$1,500,000.00	$4,377,500.00	189	12	$2,890,119.42
AAC	$16,562,677.00	11	$1,000,000.00	$2,600,000.00	100	14	$1,183,048.36
Mt. West	$12,071,254.00	12	$486,504.00	$1,800,000.00	65	12	$1,005,937.83
C-USA	$10,921,091.00	14	$500,000.00	$1,425,000.00	59	14	$780,077.93
MAC	$6,959,534.00	12	$412,500.00	$1,125,000.00	46	12	$579,961.17
Sun Belt	6506500	10	$390,000.00	$850,000.00	21	10	$650,650.00
Ind.	$4,929,080.00	6	$419,640.00	$2,129,638.00	35	6	$821,513.33

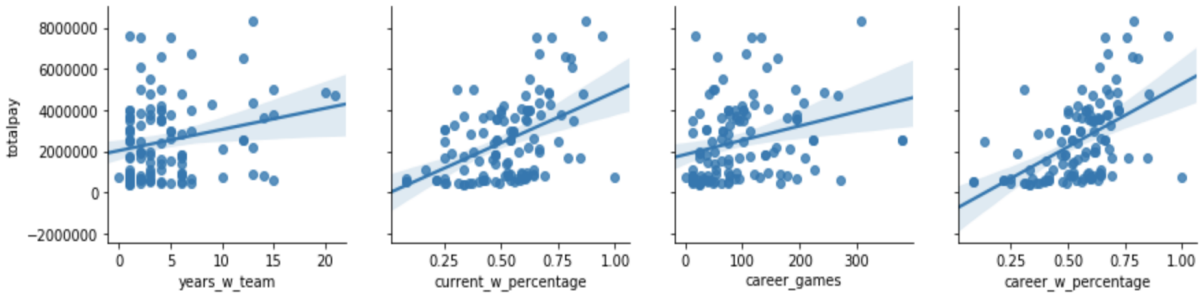
Also another indicator, players on an NFL roster, appears to be a strong indicator of total pay. More on that later.

Relationships

In order to identify the final features to be used for linear modeling. The dataset was analyzed to determine which features had a relationship with the dependent variable total pay.

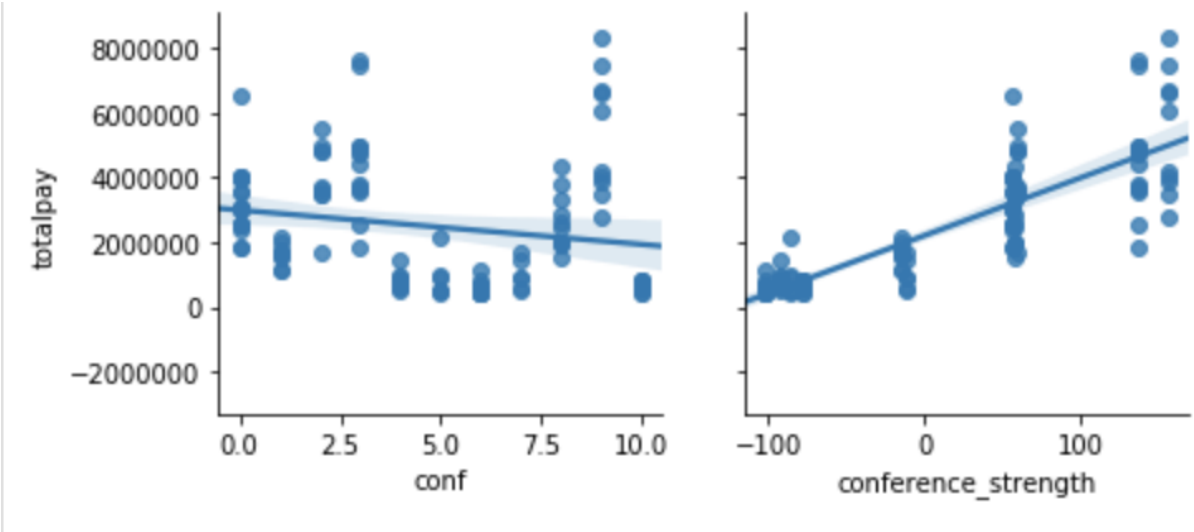
Coaches Record

It appears there is a relationship between two of the features within a coach's record and total pay. The two that jump out are current_w_percentage and career_w_percentage. Both account for the win percentages both from last season and over the course of a coach's career.



Teams Performance

How the team performs is a strong indicator of total pay for the coach.



Stadium Size and Number of Current NFL Players

Both Stadium and the number of current NFL players point towards both the popularity of the school and the number of players that will go on to become even more popular within the NFL. It would stand to reason, if you are training and producing NFL caliber players, you will win games and attract large-number of paying fans to watch live and on a device.

Analysis

3 different models were used for the regression analysis, Ordinary Least Squares Regression, and Mixed Linear Regression.

Model 1 - Ordinary Least Squares Regression (all features): To determine the statistically significant features all features were included within the model. Of the features selected, 5 turned out to be statistically significant. Where the significance is defined as having a p-score of five-percent or below. Many of the features associated

with the coach's record, career wins and a career winning percentage did not turn out to be as significant as the features associated with the team's performance such as FPI, efficiency, and rank. All three team performance features penalized coaches with low to negative scores. Another key driver was the number of players currently in the NFL. Lastly stadium capacity proved to be significant.

The model returned an R-Squared Adjusted value of 80% in training and 80% with the test dataset. This translates to 80% of the model's variance can be attributed to the features included in the model. The next model will carry over Rank, Efficiency, FPI, Capacity, and in_nfl_cnt. We will also include conference encoded features.

Model 2 - Ordinary Least Squares Regression (all statistically significant features): While only using statistically significant features, the model's R-squared value dropped. This is to be expected, as having more features is a known way to trick the R-squared value in improving. The underlying RMSE slightly dropped for model 2, with the MAE remaining the same. In terms of prediction, model 1 was closer to the actual total pay. This can be explained but the level of "noise" included in the model by the less significant features.

Model 3 - Mixed Model Linear Model: (all statistically significant features grouped with the conference): Of the first 3 models, this model performed the worst. Again the drop in R-squared can be explained by removing a feature. However, the MAE has gone up 774k. This means that our model is off by 774k when making predictions.

Of the 3 models, model 2 was chosen to complete the remaining predictions.

Results

2019 was not a particularly strong year for Syracuse's football club. As such, key coefficients such as rank, efficiency, and FPI penalized Dino Barbers when predicting total pay. Put another way, for the 2020 season, if Barbers is able to improve the team's efficiency, FPI score and therefore move up in rank, his predicted total pay will improve. Based upon the recent staff changes, most notably firing his defensive coach, clearly, Barber understands this.

What's really nice about using the team predictors to evaluate the coach's total pay, is these metrics can be tracked and predicted weekly on a game by game basis. Armed with this data, Syracuse's athletic director can set-up KPIs that can be tied into Coach Barber's next contract.

Suggested Salary for Syracuse Head Coach is $2.22 million with the actual salary $2.4 million. 200k is a significant drop in value. However, this can be explained by each of the team performance coefficients. For example as the team's FPI improves, it will have a positive impact on Coach Barber's Salary.

Learning Objectives

In this project i was tasked with obtaining data, scrubbing the data, and exploring it using techniques including descriptive statistics,summarization, and visualizations. Based upon the exploratory data phase, I successfully modeled relationships between the data sources to create a final dataset based upon my mocked client's needs. Finally, I was able to provide insights and recommendations to my mock client using various data mining and quantitative techniques.

Throughout the course of the entire program we were tasked to complete projects and assignments that displayed our knowledge of visualization, statistical analysis, and text/data mining. The core assignments and final projects required that I solve real world problems using regression, clustering, and classification. I feel confident that whatever phase of the project i may be working in, descriptive, diagnostic, predictive or prescription, I will be able to draw upon my experiences in the program to complete the work.

Develop alternative strategies based on the data.

Data scientists need to be aware of the various challenges presented when implementing a machine learning project into production. Often what may work in the sandbox environment may not be practical for production. Trade-offs are required. In the project described below, the goal was to build out a dashboard to track the opioid epidemic. While running the model, we found training time was extensive. Addressing the training and prediction time required changing strategies and exploring an entirely new set of tools.

Project 4: Analyzing the Opioid Crisis

Project Overview

My team's goal was to understand the opioid epidemic in New York and provide a dashboard-like tool to help health and law enforcement officials manage the crisis. This crisis is an ongoing, urgent problem throughout our society. The team analyzed the number of opioids (measured in Morphine Milligram Equivalents or MMEs) distributed to pharmacies over time, making forecasts for future distributions, and trying to predict potential hot spots of opioid misuse.

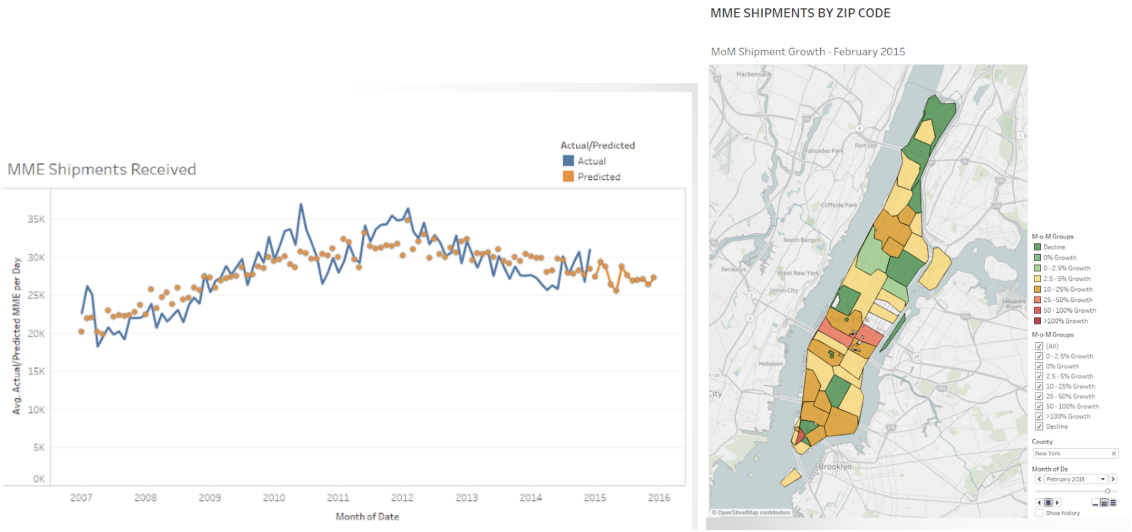
About the Data

We obtained the dataset from the Drug Enforcement Administration's Automated Reports and Consolidated Ordering System (DEA ARCOS). The initial dataset contained over 550M records for all states from 2006 - 2014. We aimed to analyze New York State only.

At close to 25 gigabytes, the dataset would run out of memory while loading the records into Google Colab. Our group opted to narrow our analysis down to Hydrocodone and Oxycodone records in New York state.

Forecasting

I created the forecasting model using Python, Apache Spark, and the Facebook Prophet Arima model for time series. As highlighted below by the dashboard view, the results were both accurate and actionable. However, training and testing the model took significant time. If we wanted to provide a real-world scalable prototype tool, we'd need to find a way to decrease the time required to process and forecast data.



Sample dashboard: Forecast Actual vs. Predicted. Zipcode showing the month-over-month change

Improving training times

To improve the scalability of our prototype, I explored the Rapids.ai machine learning GPU framework. The Rapids.ai framework overcomes large datasets' longer training times due to running on CPU hardware. Rapids.ai speeds up the training and processing time by running parallel processes on the GPU's (graphic's card) hardware. Rapids provides a Python API interface that closely mirrors the popular NumPy and pandas packages. Using the Rapids.ai Holt-Winters forecasting model, I retained the same accuracy level from the Spark/Prophet model; however, it cut the forecasting time down from 4 mins to 3 secs.

Learning Objectives

This project forced us to develop alternative strategies based on the data's size and our computing environment. Initially, the plan was to analyze the entire dataset. However, we had to scale back our analysis strictly to New York state and two drugs. Also, to improve our dashboard application's scalability, we identified a framework in Rapids.ai that improved our model's performance while not sacrificing its accuracy.

We also had to find a way to pass data between Rapids.ai and Apache Spark. Copying data into the common CSV format was a waste of precious memory that we could not spare. We settled on the compressed data format, Apache Arrow. The Arrow memory format supports zero-copy reads for lightning-fast data access without serialization overhead. Using Arrow provided a language-agnostic framework to communicate between Rapids.ai and Apache Spark.

I learned most of the techniques I used for the Opioid analysis while completing the homework labs for the program's Big Data course. One assignment required that I forecast home prices for all zip codes within the country to identify the top three zip codes for investment. I used Facebook's Prophet model and Apache Spark. Another assignment required that I use alternative techniques to solve a well-known computer vision problem. In addition to working with the neural network library Keras, I used Rapids.ai random forests and k-nearest neighbor models.

Develop a plan of action to implement the business decisions derived from the analyses

Almost all of the projects and assignments completed in the iSchool's Masters in Data Science curriculum and those highlighted in the earlier objectives have required mastering the ability to communicate technical analysis to a non-technical audience effectively. As a student progresses through the data science program at Syracuse, they will find that the ability to share their findings through presentations, reports, and group discussion is a crucial requirement and will serve as a long-term benefit in the student's professional life.

Project 5: Developing a plan of action to gain an advantage in NFL Fantasy Football

Project Overview

Fantasy sports betting has become a multi-billion-dollar industry. Our project aimed to predict fantasy football outcomes and identify the associated drivers of these outcomes.

About the Data

The lack of publicly available National Football League (NFL) data sources has been an obstacle in creating modern, reproducible football analytics research. While clean play-by-play data is available via open-source software packages in other sports (e.g., nhlscrapr for hockey; PitchF/x data in baseball; the Basketball Reference for basketball), the equivalent datasets are not freely available for researchers interested in the analysis of the NFL. This analysis used the ArmchairAnalytics.com dataset, which is considered one of the premier resources for American Football statistical information.

The Armchair Analytics dataset consists of 30 tables and over 7000 data points, spanning from 2000-2018. Of the thirty tables, data from five tables were combined to create the final dataset for further analysis. The five selected tables centered around the Play-by-Play data and the player's availability. Also, tiers for coaching and defense were imputed.

Data Preprocessing

I took several steps to transform the initial dataset and condense it down to the critical variables required to predict a player's points.

Key Steps:

- Combine the key fields from Armchair Analysis CSV dataset tables.
- Perform EDA to determine if there are any issues with the data
- Calculate the player's total fantasy points and compare with Armchair Analysis's fantasy points from NFL.com
- Create user-defined variable metrics
- Run Association Rules Mining and R's Boruta package and compare results for final feature selection.

Feature Selection: Association Rules Mining and Boruta:

Given that there were over 300 data points. One of the critical aspects of this analysis was identifying the features that best contributed to a player's scoring potential. Using Associate Rules and R's Boruta package (a wrapper built around the random forest classification algorithm)

Association Rules Mining:

Association Rule Mining (also called Association Rule Learning or Arules) is a common technique used to find associations between many variables. Arules is often used by grocery stores, e-commerce websites, and anyone with large transactional databases. The team utilized the association rules mining package to explore the data and inform which variables are potentially significant. The team ran the item frequency plots and the Apriori algorithm. The

team sorted by high support and lift and filtered on RHS => “Fantasy Tier 1”, to determine which variables were associated with favorable fantasy outcomes.

We found seven features that have a high impact in predicting success in fantasy football:

- Market Share
- Opposing Defensive Teams Strength
- Player Availability
- Practice Status
- Depth Chart Position
- Coaching
- Average TDs Scored

Boruta Library:

Boruta is an all relevant feature selection wrapper algorithm, capable of working with any classification method that output variable importance measure (VIM); by default, Boruta uses Random Forest. The method performs a top-down search for relevant features by comparing original attributes' importance with importance achievable at random, estimated using their permuted copies, and progressively eliminating irrelevant features to stabilize that test. In our case, the results confirmed many of the variables identified within the association rules evaluation. With one notable exception, Boruta rejected “player_injury_status” as not being significant.

Modeling and Analysis

Armed with the features listed above, I ran the following models: SVM, KNN, K-means, Random Forest, and Decision Trees, to determine the best predictor for a player’s scoring output. The Decision tree, with a tuned length of 10 and repeated cross-validation, produced a training/test accuracy of 89% and 85%, respectively. The model also yielded a high Kappa number and low P-value, lending more support for its accuracy. The tree split seven times, with Market Share being the most crucial of these splits. Other features with high importance were Average Snaps, Average Pass Attempts, Average Yards per game, Average Running Touchdowns, and Average touches/targets per game. Having identified the key features and a relatively accurate model, the team produced a final excel deliverable that anyone can use to predict a player's scoring output.

Reflections and Learning Goals

Two factors that contributed to this project's success were feature engineering and focusing on providing an actionable plan with a deliverable for the targeted user-group.

I started with 300 data points, the features that finally went into the models needed to be engineered or put another way, created from the original dataset. Useful features represent the data's underlying structure more accurately and, therefore, make the best model. I learned that features could be engineered by decomposing or splitting features from external data sources or aggregating or combining features to create new features.

Secondly, we wanted to provide both an actionable plan and a tool so that anyone could implement our analysis using Microsoft excel. The excel file we constructed clearly outlined the required steps (which was to enter a small set of variables simply), and the model would provide a score.

Lastly, this project allowed my team and I to work with several tools and techniques such as R, Dataframes, SqlDF, feature-engineering, Random Forests, Association Rules, Decision Tree, SVM, K-means, KNN, and GGplot.

Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization

Data science and the outcomes it delivers can be complicated and hard to explain. I feel the iSchool prepared me to present approaches and findings to a non-technical audience in every class. The project below highlights how I presented critical findings to a mock telecom company's executives looking to stem customers from leaving its service.

Project 6: Predicting Customer Churn for a Telecom company

Overview

This analysis focused on the behavior of telecom customers who were more likely to leave the platform. My team and I intended to find out the demographics of customers through EDA and use predictive analytics techniques to determine the most likely churn customers.

The project's goal was to present our analysis to the mock telecom heads, along with our recommendations on moving forward. The final deliverable would be presented to business subject matter experts and not data scientists. With that in mind, we

condensed our powerpoint presentation down to the 5 most pertinent slides. The goal of the presentation and deck was to present our findings and recommend a customer retention strategy.

About the data

The dataset from Kaggle contained 7,000 customers and 21 attributes, including demographic data, services, and contract details. Out of the entries, 5,000 were active customers, and 1,800 were churned, demonstrating that the dataset is highly unbalanced.

Key Steps:

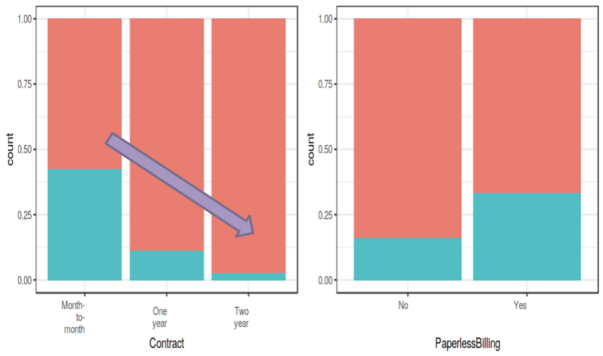
- Cleaning the Categorical features
- Standardizing Continuous features
- Creating derived features
- Creating dummy variables for factor variables
- Creating the final dataset
- Splitting the data into train and validation sets.
- Logistic Regression to Predict Churn
- K-Means Clustering to confirm the best customer segments

Reported Results

Contracts Are Key

- If a customer has a **one-year contract** and not using paperless billing, they are unlikely to churn
- If a customer has a **month-to-month contract**, and is in the **tenure group of 0-12 months**, and uses paperless billing, they are more likely to churn (**78% chance of churn)**
- As contract length increases, the likelihood of churn decreases.
- The strongest correlation is whether someone has the internet or not. Contract, multiple lines, and paperless billing also are strong indicators.

DATA FINDINGS: CUSTOMER CHURN MODEL



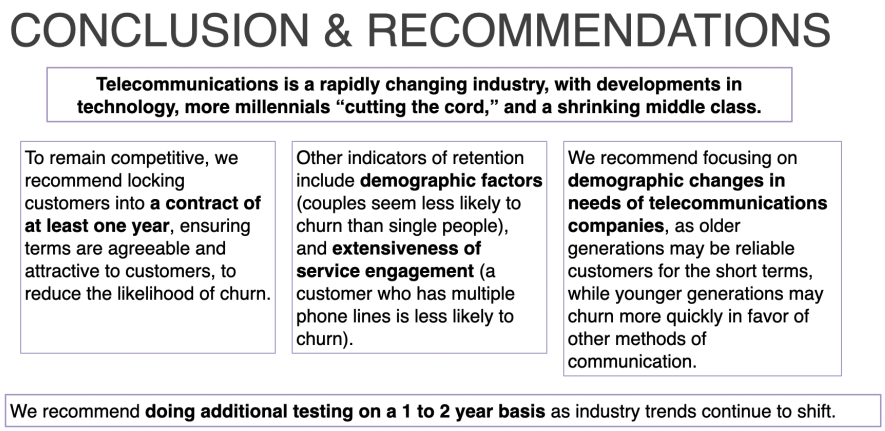
As contract length increases, likelihood of churn decreases

CONTRACTS ARE KEY

- If a customer has a **one-year contract** and not using paperless billing, they are unlikely to churn
- If a customer has a **month-to-month contract**, and is in the **tenure group of 0-12 months**, and uses paperless billing, they are more likely to churn (**78% chance of churn)**

Telecommunications is a rapidly changing industry, with developments in technology, more millennials "cutting the cord," and a shrinking middle class.

- To remain competitive, we recommend locking customers into a contract of at least one year, ensuring that terms are agreeable and attractive to customers and reduce the likelihood of churn.
- Other retention indicators include **demographic factors** (couples seem less likely to churn than single people) and **extensiveness of service engagement** (a customer who has multiple phone lines is less likely to churn).
- We recommend focusing on **demographic changes in telecommunications companies' needs**, as older generations may be reliable customers for the short term. In comparison, younger generations may churn more quickly in favor of other methods of communication.
- We recommend **doing additional testing on a 1 to 2-year basis** as industry trends continue to shift.



Learning Objectives

This project provided the opportunity to organize and analyze transaction information using data mining techniques and visualization to identify patterns for customer targeting. It was also necessary to develop a plan of action to quantify the insights developed in this analysis, translating to measurable and actionable recommendations. Ethical considerations were also needed to ensure that customer segmentation and profiling were free of bias, using demographic information to profile the previous behavior rather than using said information to explain their behavior. This project allowed the data to guide the analysis, requiring alternative strategies to be developed as observations were made within the data.

Lastly, It's not enough just to have the technical know-how to analyze data, create predictive models, and so on – communication skills are equally important. Using the 5-page PowerPoint deck, I feel we effectively explained our approach and provided clear, concise, actionable takeaways to improve the business.

Synthesize the ethical dimensions of data science in practice.

The MSADS program professors drilled home the point that good data science requires both privacy and transparency. Data scientists need detailed data to provide sophisticated products and services. At the same time, data scientists should be aware of the privacy trade-offs required to collect massive amounts of data.

In terms of transparency, data scientists have a lot of freedom to determine answers to data questions. For example, in IST707, one homework project required that we use various machine learning techniques to identify the Federalist Papers author. I found that the potential that it was one author over another came down to the usage of a single word. Or, put another way, I could choose who I wanted the author to be simply by leaving in or out that word.

Learning Objectives

As I completed the asynchronous content, homework, and projects, I feel the professors emphasized that there are consequences for our actions as data scientists. Therefore, it is vital to ask critical questions about data security, user privacy, and potential biases in finding insights.

Lastly, we can build trust both inside and outside the organizations where we may be working by being transparent.

Conclusion and Reflection

My goal in taking the Syracuse ADS program was to develop tools and techniques that can be applied to my existing and future clients. As a career consultant, I'm often tasked with identifying ways and methods to produce value for a business unit or an organization. After completing the master's program, I feel I can take data from disparate sources, analyze the data, and create actionable insights based upon my findings. With each course's completion, both the incredible capabilities of data analysis and the inherent complexities that come with gathering and interpreting data became more and more apparent. The Applied Data Science program at Syracuse gave me a good foundation of how to model, visualize, and interpret data that will be useful in any number of fields. I believe this portfolio demonstrated the achievements of the overarching learning objectives laid out by the program. The learning objectives are listed below, along with how this program has helped me achieve these goals:

1. Describe a broad overview of the major practices' areas in data science.
2. Collect and Organize data.
3. Identify patterns in data via visualization, statistical analysis, and data mining.
4. Develop alternative strategies based on the data.

5. Develop a plan of action to implement the business decisions derived from the analyses
6. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization
7. Synthesize the ethical dimensions of data science in practice.

This program not only cultivated my technical abilities, such as programming languages, statistics, and machine learning algorithms, but it also nurtured my soft skills, including teamwork, analytical thinking, and perceptive and open-minded. I came into the applied data science master program with little knowledge about machine learning, yet I graduated from the program with abundant hands-on experience and expertise in many areas. From those experiences, I discover that knowledge is rigid, however, how you apply and interpret is powerful and agile. Though the professors taught me plentiful concepts, I found putting them into real-world applications was far more beneficial. I learned to always remain curious and suspicious of what you know and what you see, and never jump to conclusions too quickly. It is vital to look at things with an open mind and embrace what the data may lead you. Last but not least, when conducting a project, plan ahead but also leave space for imagination as things do not always go as planned.