Jushira Thelakkat - Projects

Predictive Modelling: Predict Price of Used Mercedes Cars — August 2017

The goal of this project was to develop the best possible model that predicts the price of used cars based on 15 predictor variables, including mileage, year, trim, color, and more.

The first step was to explore the dataset to ensure I understood all the predictor variables. After initial exploratory data analysis (creation of scatterplots for quantitative variables and box plots for categorical variables), I modeled the data using ***linear regression***. The p values of regression of price on each predictor gave insights about what variables could be the most and least important. But this model had a Root Mean Square Error of 8000+.

The predictor variables had to be narrowed down, hence used shrinkage methods such as ***Lasso and Ridge***. Cross-validation gave a best value of 35 for tuning parameter lambda. I narrowed down my predictors to only 7 from 15 after using these modeling methods.

Since multiple linear regression models do not work too well with categorical variables and non-linear relationships had to be handled in a better way, tree based models seemed like a wise choice. I first used ***Bagging*** where several trees are created and predictions are made by averaging the predictions of individual trees. I played around with different values of ntree, minimum node size and maximum number of nodes (mtry=15) to get the lowest RMSE on my training data. I came down to an RMSE of 6500 with bagging. Using Variable Importance plots, I also got the most important predictor variables. I wanted to check effects of ***Random Forest*** and ***Boosting*** too but the RMSE value did not significantly decrease and I saw no point in increasing the complexity of my model.

Finally, I built a model with Bagging for predicting the price of used Mercedes cars. With the test data too, the model gave a similar RMSE of 6250, which proved that there was no overfitting in the model.

Tools Used: The project was entirely done on R Studio.

Data Analytics Programming: 2014-15 NBA Season Analysis — August 2017

For our final group project in Data Analytics Programming (Python), we chose the NBA data set from Kaggle for our analysis.

Performed extensive exploratory analysis on this dataset after initial data cleaning. Looked at the effect of rest on teams' performance, different types of shots attempted (pull up, catch and shoot etc.), effect of shot distance and shot time on accuracy.

Also analyzed if the players in the 2014-15 season showcased the Hot Hand phenomenon. This was the most interesting part of our project. Hot Hands phenomenon states that once a player makes a shot, the accuracy of his successive shots increases. Once we wrote the code in R and ran this analysis on the top four MVP (Most Valuable Players) of that season, we saw that "hot hands" was just a fallacy. We then

reasoned out that this could be because of other factors such as – added pressure of making the next shot once a shot is made or overconfidence, that might have led to failure of hot hands phenomenon.

We then used our data to make a prediction about the outcome of a shot. After splitting data into training and test sets, we ran a random forest with n=100 and mtry=2 to get an accuracy of 65%. With the Variable Importance plot, we also found the most important predictors to be shot distance, type of shot (2pts or 3pts) and defender distance. The analysis we did in this project could help teams, betters and players in making better decisions.

Tools Used: The project was partially done in Python and partially done in R Studio.

Text Analytics: **Brand level text analysis on car reviews on Edmunds.com** **September 2017**

I developed a crawler/scraper to fetch messages posted in Edmunds.com discussion forums and pushed the output of the crawler to a .csv file.

I fetched around 10000 posts about cars from a general topics forum to have a wide variety of car brands and models to discuss about. Using frequency counts, I found the top ten brands being talked about, used lift ratios to find associations between these brands and used a Multi-Dimensional Scale (MDS) map to show these brands.

I then moved on to do attribute analysis wherein I found the top five attributes of cars mentioned in the discussions. These were general attributes like performance (that included pick-up and acceleration). I then analyzed what attributes are most strongly associated with the top brands.

Based on my analysis, I came up with suggestions for both the product as well as marketing/advertising managers of these brands.

Tools Used: Python, Excel

Text Analytics: **Amazon Product Reviews – Men's Athletic Shoes** **September 2017**

Built a web crawler to scrape 2000 product reviews for the top men's athletic shoe of four most competitive brands- Nike, Asics, Adidas and New Balance from Amazon.com

After massive data preprocessing including tokenization, stemming and lemmatization, I calculated lift ratios to build MDS map for these brands and looked at different attributes being associated with each brand. These attributes included style, comfort, price, durability etc.

Analyzed lift ratios further to understand whether the brand image was in line with what consumers thought about the brands

Built recommendation systems for product and brand managers of each of these brands to improve customer acquisition, satisfaction and retention

Tools Used: Python, Excel

***CAPSTONE* – VISA Fraud Analytics** **Fall 2017, Spring 2018**

Problem Statement:

To predict the likelihood of VISA's e-commerce merchant being a "BUST OUT" merchant.

A bust-out merchant in VISA's dictionary is defined as a merchant with malicious intent; one who sets up fake businesses to make the customer pay in return for no goods or services.

Data:

Transactional data given for 10000 merchants over two years.

Methodology:

To forecast merchant behavior, I have performed time series analysis using ARIMA and other forecasting techniques for short term as well as long term forecasting.

Using the given data, I first created new features such as Stability, Size, Age etc. to cluster the merchants into homogenous groups. With the clusters now ready, I built predictive models (random forests, boosting, neural networks) to identify fraudulent merchants for each of these clusters.

Business Value:

VISA plans to use this analysis to improve its payment ecosystem. In long term, VISA plans to build a merchant tracking tool and mitigate merchant frauds.

Tools I will be using: Excel (PIVOT tables), Python, SQL, Hadoop, Google Colaboratory