Springboard Data Analytics Course

Ultimate Challenge

October 2019

John L. Parsons

**Data Analysis Ultimate Challenge Summary**

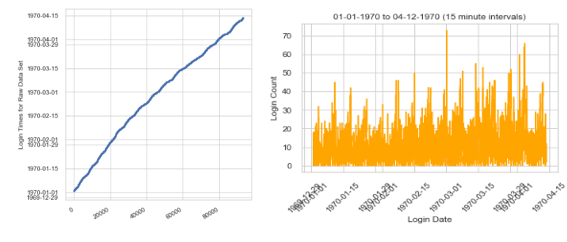
The Ultimate Challenge is a take home assignment that will contain a total of three sections to analyze two json data files from Springboard. The first data file contains simulated login counts and the file is called logins.json file. The second login file is called ultimate\_data\_challenge. json and will contain a total of 12 attributes and the goal is to build a predictive model.

The first phase of this report will contain the Exploratory Data Analysis section from the login.json file and the goal is to aggregate the logins in 15-minute intervals and explain the patterns seen in the data and provide any additional insight from the EDA section. The second phase will design an experiment dealing with a toll bridge and its impact on the activity level between two cities. The managers have decided to reimburse the residents for the toll costs to increase the flow of activity between both cities. The experimental design for this project will be discussed in this section. The last section will contain the best predictive model that can accurately predict rider retention from the population in six months from data collected in January of 2014. There is a total of 11 input attributes and the target variable is trips\_in\_first\_30\_days. The managers have agreed to assign an active rider value on a user if they have been active in the past 30 days from the dataset.

**Phase I: Exploratory Data Analysis.**

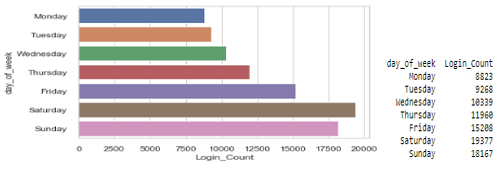
The json file was converted to the dataframe in Panda and then transformed to the timestamp. The login\_time was then resampled every 15 minutes and the number of logins were counted for each fifteen-minute increment. Figure 1 shows the results from the login counts from the raw data set compared to the data set that aggregated the login times every 15 minutes for this challenge. The login frequency from the raw data set shows a nice upward trend from the initial login date of 1970-01-01 to the last login date of 1970-04-13. The aggregated data set also shows an increase in login frequency’s, but it is much more cyclic or a weekly/biweekly period.

**Figure 1: Login Frequency with the raw and aggregated dataset.**



The minute, hour, day, day\_of\_week, month and year was created from the aggregated login times to determine where the time cycles were occuring during this four month period. The day\_of\_week variable and week attributes both had very interesting results for the cyclic pattern of the login counts. Figure 2 shows the Login counts per day of the week and there is an increase in login counts from Monday (8,823) up to Saturday (19,377) and then Sunday it drops

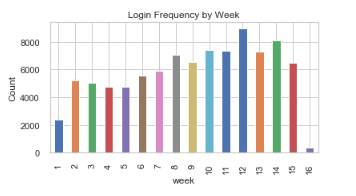
**Figure 2: Login counts by day\_of\_week attribute.**



down to 18,167 logins. Saturday and Sunday are the most active times for this population and Friday is the third most active time for the day of the week. The login times do not say where this activity is occuring, but this population is very active on Friday to Sunday and less active from Monday to Thursday times of the week.

The login counts were also calculated on the week and Figure 3 shows the results. The overall trend is an increase in activity from week 1 to week 15 and then a dramatic drop in activity for week 16. Weeks 2, 8 and 12 were the peak of activity and then activity decreased and began the cycle again.

**Figure 3: Login counts by week attribute.**



There are two areas for possible data quality problems in the dataset. PD Profiler shows the login counts had a total of 407 zero counts for this time period or was 4.2% of the dataset. This should be investigated to determine if there was no activity during this time period or the sensors are not recording any activity. The second area to focus on is week 16 which ends on April 13, 1970. Figure 3 shows almost no activity for this week and the data should be removed for future analysis or should of continued until the week ended to aquire all the data for a complete four month analysis in this study.

**Part 2: Experiment and Metric Design.**

The goal of this section is to initiate an experiment that will determine if a two-way toll between two cities is reimbursed, then will this encourage driver partners to become more available in both locations. The driver partners tend to spend more time in Gotham City at night and to be more active in Metropolis City during the day. Traffic is more evenly mixed during the weekend.

The metric to be used for this experiment will be based on two outcomes to determine the success of the new toll-free program. The first metric will measure the level of activity going across the bridge and compare this to the control from the previous measurements (data collected from the most recent data). The second metric will measure the sales tax for both cities and compare this to the previous values. Several assumptions will be put in place to interpret the results and provide a valid recommendation to the city managers. First, the toll bridge is assumed to be the only access between both cities and citizens cannot drive to another bridge to gain access. Second, when the tolls are reimbursed to the drivers, we are assuming the toll booths are no longer in service and this will be a regular state or federal highway connecting both cities. Finally, the population is closed and the number of residents from both cities stay the same and residents do not leave or enter the study. The data collected from the previous months is assumed to be the control measure.

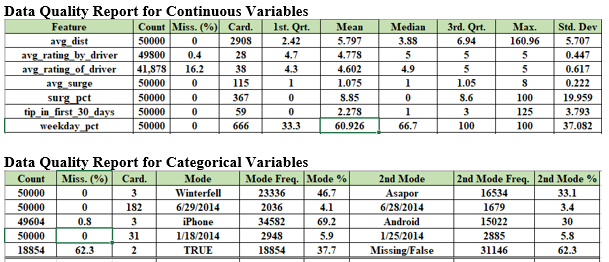
The new toll-free program needs to increase activity levels which will generate more revenue than was collected from the toll booths to be successful for this project. The project will use AB Testing or ANOVA Models to determine if the activity across the bridge has significantly increased in activity at a 0.05 p level and driver partners are no longer exclusively staying in one city verses the other city. The second treatment will measure sales tax from gas stations which are appropriated to highway purposes and sales tax from local establishments to determine if the toll-free program significantly increased revenue for both cities. The revenue collected needs to generate more money than the toll roads generated for this time period to be successful. The bottom line for the success of the program is to increase the revenue from both cities which offset the revenue lost from the reimbursement of the tolls and can be measured in increased activity across the bridge and local sales tax.

The study needs to be short to minimize the assumption of the closed population and the previous data collected which is the control metric is assumed to be an accurate measurement for the experiment in this study. Another approach that could be used is to cut the tolls in half to see how this affects traffic across the bridge. Does the reduction in tolls actually increase the revenue from both cities and offset the revenue lost by the fifty percent cut in fees?

**Phase 3: Predictive Modeling**

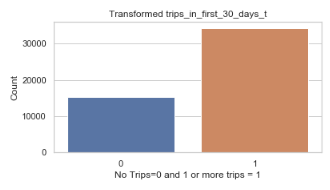
The ultimate data challenge json file contains a total of 12 attributes for this section. There were seven numeric, four categorical and one Boolean set of variables in the data set and the Data Quality Report (Figure 4) can be seen for all 12 attributes. The target variable is trips\_in\_first\_30\_days and if a rider was active anytime in this time period, they were considered retained and if no activity was recorded then they were not retained. The target variable was

**Figure 4: Data Quality Report.**



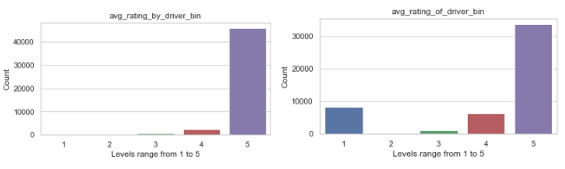
transformed to a binary variable and any activity greater than 0 was binned to a level of 1 and no activity was binned to a level of 0. The count plot can be seen in Figure 5 and this is an unbalanced data set with 34,292 riders as being classified as retained and only 15,312 riders as not being retained.

**Figure 5: Transformed Target Variable Count Plot**



The remaining data cleansing steps will be briefly explained in this section. The phone variable dropped the missing values and the remaining two levels were binned into a binary categorical variable. The Android phone was given a level of 1 and the iPhone was given the level of 2. The city variable had a total of three categories and was binned into three levels. Astapor was binned to level 1, King’s Landing was binned to level 2 and Winterfall was given a level 3. The ulitmate\_black\_user was a Boolean attribute and all missing values were given a value of 0 and the True riders were given a level of 1. The average rating by and of the driver had values between 0 and 5 for this dataset. The levels were binned into a total of five categories and any value between 0 to 1 was given a level of 1 this continued until any level between 4 and 5 was given a level of 5. There were missing values for both datasets and these were binned into the 1 level. The avg\_rating\_by\_driver only had 200 missing values but the avg\_rating\_for\_driver had 8,122 missing values and should be noted during the final analysis. The count plots can be seen in figure 6 below. The correlation matrix was created in Python and

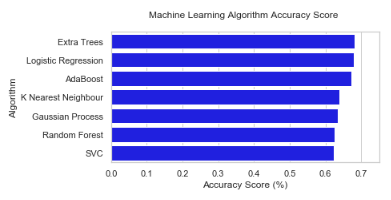
**Figure 6: Comparison of the binned average rating by and for the driver.**



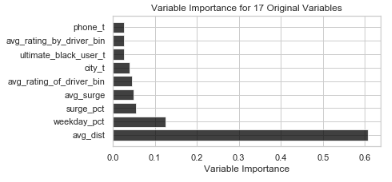
two variables were highly correlated with each other and these are avg\_surge and surge\_pct and only one of these variables will be used in the final model to minimize cardinality between these two variables.

A total of seven models will be used to determine the champion model for this report. These models are; Random Forest, KNN Classifier, Logistic Regression, Extra Trees Classifier, SVC, AdaBoost and Gaussian Process Classifier. The data was randomly sampled and only 25% of all rows will be used for the seven models. The data was split 65:35 for the training and validation set and the scoring was based on Accuracy of the model and all variables were used for the initial run except the last\_trip\_date and signup\_date. The variable importance was calculted from the Random Forest model and the results can be seen in Figure 8 for this run.

**Figure 7 Accuracy Scores for all 7 Models.**



**Figure 8 Variable Importance for the Random Forest Model.**

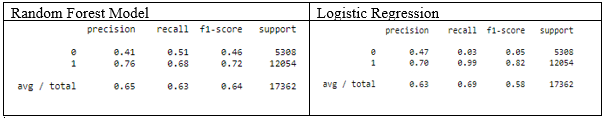


The accuracy scores ranged from 62.2% for the SVC Model up to 68% for the Extra Trees Model (Figure 7). The most important variable for the Random Forest is avg\_dist and this is the average distance in miles taken per trip in the first 30 days. The weekday\_pct and the surge\_pct were the second and third most important variables and phone\_t was the least important variable.

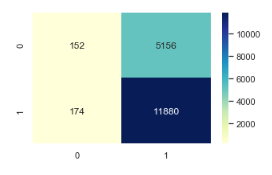
The Random Forest model with a balanced approach and the Logistic Regression model with no balanced approach were re-run and only five variables were included in the run. These are; avg\_dist, weekday\_pct, avg\_surge, avg\_rating\_of\_driver\_bin, city\_t and trips\_in\_first\_30\_days\_t was included in this run. The entire dataset was used in the analysis and the balanced approach was used because the target variable is unbalanced.

The results from the final run for both models show interesting results. The accuracy scores from the previous runs are close to 70% and the model does a really good job in predicting the True Positive Values and this can be seen the Logistic Regression Confusion Matrix (Figure 10). The model classifies 11,880 out of 17,362 as TP and then only 152 as being True Negative (TN). The model has a lot of False Positive values and has a hard time classifying the TN in the dataset. The balanced approach was used in the Random Forest and the Recall and f1 values increased from 0.03 and 0.05 for Logistic Model up to 0.51 and 0.46, but it only accurately predicted 8,251 TP values but correctly identified 2,685 of the TN values in the data set.

**Figure 9: Classification Report for the Random Forest and Logistic Models.**



**Figure 10 Logistic Model Confusion Matrix.**



The model that can be used depends on the business goals and if the business goals are only concerned with predicting the active rider then the logistic model works well. However, if the business wants to know both the active and inactive riders for the six-month time period then the balanced approach needs to be implemented for the data set.