Maple Tan

DATASCI420

May 20th, 2018

**Capstone Project: Taxi Tip Classification**

**Section 1: Problem Description**

In the service industry, tips often make a large portion of a worker’s wage and it is not uncommon for these workers to rely heavily on tips just to make a living. This is also the case with taxi drivers and it would be beneficial for them to know which factors could help them obtain higher tips from their customers.

The goal of this project is to create a multiclass classifier that predicts the range in which a customer tips their taxi driver. The tips in the data has a very long tailed distribution, with around half the tips being less than $1. The tip categories will be separated into low, medium and high tiers and the model would predict a trip’s tip classification depending on certain factors such as weather, trip duration or payment method.

**Section 2: Data Description**

The provided taxi data from class compromises of two different datasets. One describes the taxi trip such as time duration, amount of people in taxi, pick up and drop location, etc. and the other describes the taxi fare such as payment method, total amount of the trip and the tip paid. To clean up the taxi data, the two separate CSV files were merged based on both the hack license number and the date and time of the taxi trip. Certain columns such medallion, vendor id, the breakdown of charges of the payment, rate code, etc. have also been removed from the data set due to redundancy.

On top of the taxi data provided in class, there will also be integration of weather data for New York in 2013 from the [National Centers for Environmental Information](https://www.ncei.noaa.gov/) website. The weather data set is comprised of data taken from several different weather stations in the Central Park area of New York. There are three features to take note: average temperature throughout the day, precipitation amount and snow amount. Within the data, there were several chunks of missing data from each the columns and a good amount of NaN data to deal with. In particular, the first three months of average temperature data were missing from the dataset.

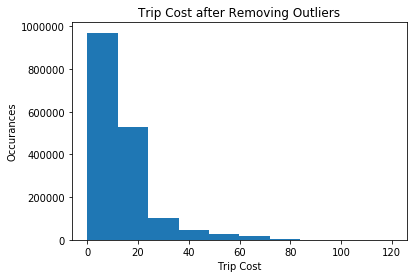
The cleaning up of the data was done in parts. The missing precipitation and snow data were simply replaced with 0 but the missing average temperature required a little more work. To help deal with the missing data, the different features from the weather stations will be averaged based on day. So rather than have data from several different weather stations, we will have an average between all the weather stations for that specific day. To fill the missing average temperature data from the first three months, we found average New York temperature from the [National Weather Service Forecast Office](http://w2.weather.gov/climate/xmacis.php?wfo=okx) website, which actually does draw data from National Centers for Environmental Information, where we got out first dataset. From there, we replaced all the NaN temperature data from the first three months of the year with the data we found from the National Weather Service Forecast Office site.

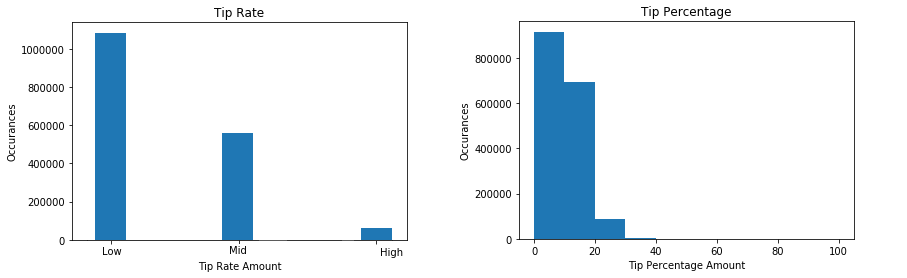
Once the weather dataset was cleaned up then it was integrated with the main taxi dataset based on its date. This way, each trip will have its average weather data for the day. Though it is an average and the weather specifically during the trip, it can provide a good idea of the driving conditions during that day.

Later in the paper, we describe a pattern seen in the data where longer trips tended to lead to both higher raw tip amounts and higher percentages tips being given. To see if we can explore that more, we also added a column labeled ‘rush hour’ that indicates whether or not the taxi trip occurs during typical rush hour times of 6:00 AM – 10:00 AM and 4:00 PM-8:00 PM. The value would be True if the trip occurs during these rush hour times and False if otherwise. Typically, rush hours would have an increased amount of traffic, and thus theoretically, a longer trip time as well.

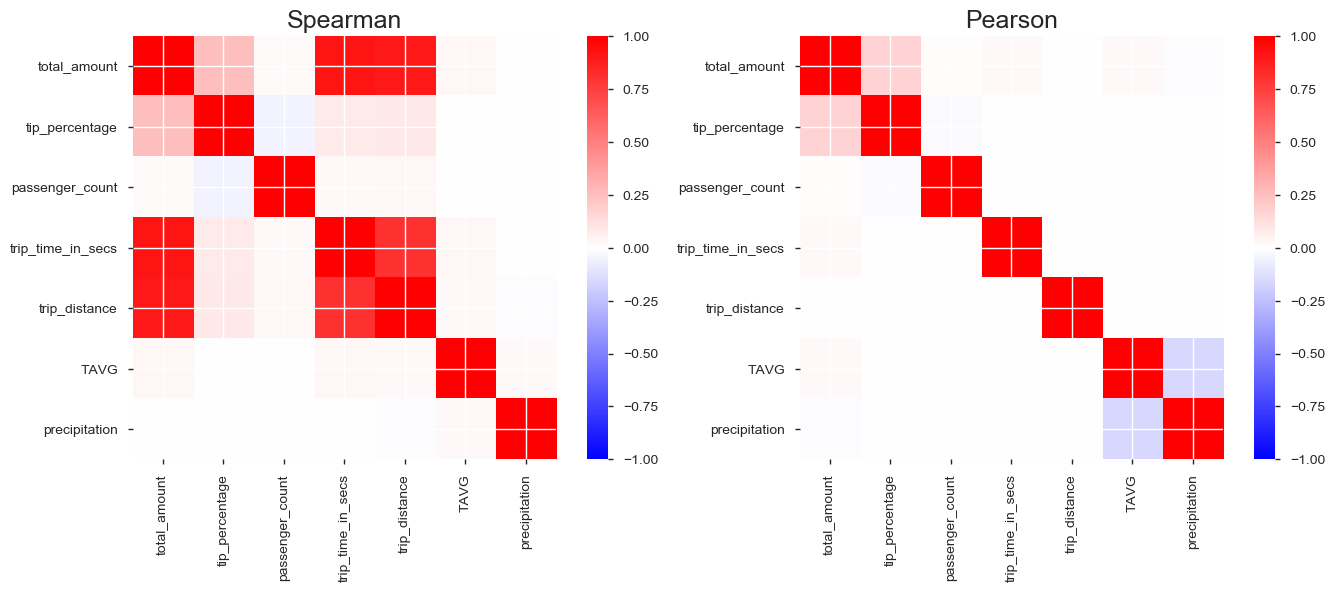
**Section 3: Data Exploration and Analyzing Report**

The taxi trip cost has a very long tailed distribution and we see that the majority of the trip’s cost falls between 0-20 dollars. As such, it could be profitable to analyze the data based on tip percentage rather than the raw tip amount. So theoretically, it would be better to maximize the tips a driver gets from a drive since trips will tend to fall around the same price. When you look at the tip data, we see that around half the tips are less than $1. So, let’s say that the typical tip is 15%. As a result, we will define the tiers as the following: a low tip consists of tips less than 15%, medium tips consist of tips between and including 15%-20% and a high tip consists of anything more than 20%. This means that around 64% of all tips fall into the ‘low’ category which follows the trend of a long-tailed distribution as seen in the figures below:





Furthermore, when we look only at the taxi data, we see that there is a positive correlation between the original price of the taxi trip and tip percentage. Needless to say, the majority of people tip depending on the original amount and while it is for sure that we would see an increase of the raw tip amount with an increased trip duration (due to the trip costing more), it follows that longer trips also seem to be positively correlated with tip percentage to a slight amount.



We also see that customers who pay with card are much more likely to tip higher, or even tip in general compared to customers who pay with cash or with any other method. This could be due to how card payments are processed, since there is almost always a prompt asking the customer for an amount they wish to tip when paying with card. In comparison, paying and tipping with cash is a little more inconvenient because it requires having loose change to tip.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tip Rate vs Payment Type** | | | | |
|  | **Tip Rate** |  |  |  |
| **Payment Type** | Low | Medium | High | **Totals** |
| Card | 299791 | 559127 | 59781 | **918699** |
| Cash | 778186 | 42 | 11 | **778239** |
| Disputed Fare | 1162 | 7 | 1 | **1170** |
| No Charge | 4021 | 4 | 1 | **4026** |
| Unknown | 851 | 100 | 63 | **1014** |
| **Totals** | **1084011** | **559280** | **59857** | **1703148** |

There were also some interesting things to note for tipping percentages in relation to if the trip was during rush hour or not. If you look at the numbers, then we see that the majority of all taxi rides happen outside of rush hour. About 64% of all trips occurring outside of rush hour. However, in terms of the distribution of the tip totals, we also see that the percentage of high, medium and low tips given are similar regardless of if the tip was given during rush hour or not. This could indicate that rush hour is not especially correlated to the percentage in which a person tip.

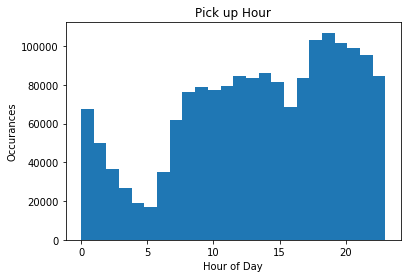
|  |  |  |  |
| --- | --- | --- | --- |
| **Tip Range** | **Yes Rush Hour** | **Not Rush Hour** | Total |
| High | 20952 | 38905 | 59857 |
| Mid | 201131 | 359052 | 560183 |
| Low | 391872 | 692153 | 1084025 |
| Total | 613955 | 1090110 | 1704065 |

|  |  |  |
| --- | --- | --- |
| **Percentages out of Total** | | |
| **Tip Range** | **Yes Rush Hour** | **Not Rush Hour** |
| High | 0.03412628 | 0.035689059 |
| Mid | 0.327598928 | 0.329372265 |
| Low | 0.638274792 | 0.634938676 |
| Total | 1 | 1 |

In terms of other interesting observations, after looking at various other columns such as trip time and trip distance, it follows that a lot of the distributions similarly follow a very long tailed distribution. In the case of trip time and distance, this would imply that the majority of people and a very large amount of them ride the taxi for a relatively short amount of time and distance. Another common theme among all the features are that many have very extreme outliers that entirely skew the distribution if not filtered out.



There also appears to be a large volume of taxi trips throughout the day with the only time it dwindles relatively is around 5:00 AM. Even at the lowest occurrence, there are still around 20,000 trips taking place around 5:00 AM. It appears that people taking taxi trips remain a constant high between 10:00 AM to around 5:00 PM and then there is an even larger spike throughout the night until the taxi trips start dwindling around midnight.



**Section 4: Feature Engineering**

The first feature engineering we need to do is how to categorize the tips given to a taxi driver. As discussed before, due to the nature of tips, many people tend to tip depending on the original amount paid. So rather than categorizing the tips based on the raw amount, it would be better to categorize the tips based on the original amount paid because most trips fall around the same amount.

Following this, there was also the question as to how to quantify the weather on a during a given taxi trip. Given that the tip percentage is positively correlated with trip duration, we can try to find a way to connect trip duration and the weather data in our dataset. Within the weather dataset, there are columns which quantity the average amount of precipitation such as rain and snow on a given day. Typically, we see that the more rain or snow that occurs on a given day, the worse the traffic usually is. As such, we can create a new variable called ‘precipitation’ that will give us a value that describes both how much rain and snow that occurs on a given day and how much it may affect the traffic in New York. We will define this column mathematically as the sum of the amount of rain and two times the amount of snow. Since generally, snow is much to drive in than rain, it would make sense to give accumulating snowfall a higher weight than rain.

**Section 5: Feature and Model Selection/Performance**

In our current data frame, we have our target value: tip percentage, which describes how much tip a customer gave compared to the original price of the trip and our features:

|  |  |
| --- | --- |
| **Feature** | **Description** |
| payment\_type | Method customer used to pay for trip |
| total\_amount | Total Amount paid on trip |
| passenger\_count | Number of passengers on trip |
| trip\_time\_in\_secs | Duration of trip in seconds |
| trip\_distance | Distance of trip in miles |
| TAVG | Average temperature for the day |
| precipitation | Measure of how much snow and rain for the day |
| rush\_hour | True if trip starts during rush hour, false otherwise |

Due to the categorical nature of both our target value and in the payment type value in our features, we also used the LabelEncoder function to transform those categorical variables into numerical ones before feeding all the values into our different models.

The models we attempted to use are: Naïve Bayes Classifier, DecisionTree Classifier, RandomForest Classifier, K Nearest Neighbors and SVM. From the beginning, it was clear that the RandomForest Classifier and SVM took a relatively longer time to train compared to the rest of the models. In fact, SVM took over half hour to train and so it unfortunately had to be eliminated from our list of models to consider. Another thing to note was that from simple accuracy calculations – out of the all trials that where we trained and tested our models, the accuracy of the models was all relatively stable and did not especially fluctuate wildly between the different trials. At most, we often saw maybe a two or three percent difference in each of the metric values.

Below is a table that shows the model metrics of the various different models that were attempted:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metrics of a Single Trial** | | | | |
| **Model** | **Accuracy** | **Recall** | **Precision** | **Runtime** |
| Naïve Bayes | 0.389 | 0.335 | 0.296 | A few seconds |
| DecisionTree | 0.927 | 0.848 | 0.846 | A minute or so |
| RandomForest | 0.8 | 0.579 | 0.539 | Several Minutes >5 mins |
| K Nearest Neighbor | 0.774 | 0.563 | 0.559 | A minute or so |
| SVM | n/a | n/a | n/a | Too Long to run >30 Mins |

In terms in the models themselves, it seems that that the Naïve Bayes performed the poorest out of the bunch and DecisionTree performed the best. It seems in general that the methods that utilizes trees perform best out of the models tested. Though, K Nearest Neighbors also did a rather decent job in terms of accuracy.

**Section 6: Future Improvements**

There are many things we can experiment and use to try to make the model better. For example, we could have incorporated more types of data. If there was more time, we have the option to incorporate some actual New York traffic data such as congestion or accidents that may delay the taxi trip. This would theoretically be more detailed than only including a field for typical rush hour time, which ultimately might not even be applicable to New York. We could also home in more on the specific regions and make use of latitude and longitude columns provided in the taxi data. Certain pick up locations could experience more traffic and delays than others and that could result in longer trip times and tip percentages given. Furthermore, even to a certain extent, every large city is divided by income bracket in some way and it could be possible that one income bracket would be more willing to tip than the other.

Another option would be to try more different types of models and feature selection. We see that a lot of the data is heavily skewed, so a model such as logistic regression could give better results than the ones we tired out. In a similar vein, we could have also scaled/normalized some of our features especially since a lot of models work better with certain distributions such as linear or normal distributions. In terms of the categorical variables used, rather than using a LabelEncoder, different methods could also have been used instead. For example, we could have one hot encoded the categorical variables or used Risk Values to replace them.

One thing to note in the model performance as well, is that the accuracy is rather high for the DecisionTree Model, so one has to question as to if the model is overfitting. One way we can address this is to use split the data into three different sets: a training set, a testing set, and a validation set and test the validation set. Then we will train and test all the models as usual but then use the validation set against the DecisionTree Model and see if the accuracy still holds up.

**Section 7: Responsibilities**

All data cleaning, model building, figure creation and write up was done by Maple Tan.