Chicago Taxi Tip

Problem Statement:

Taxi driving can be a pretty lucrative occupation, however an important portion of revenue for drivers are tips. Tipped professions are usually paid less for their time or hourly rate to account for customer tips. The goal of any tipped worker is to optimize their time or model to obtain the greatest tip amount to compensate for this reality. So our objective for this project is to come up with a way to predict the tip amount for a taxi driver in Chicago using many variables with the open taxi data provided.

One of our methods is to train a Linear Regression model to predict the tip amount from a trip based on many factors such as time, distance, and area. Our second method is to create a Classification model to predict/classify tips as low, medium, or high amounts. Using our models, a driver could predict what trips would be the best for them. A taxi company could also optimize their drivers for greater tip amounts, as in Illinois, companies are required to pay a minimum wage if tips don't account for enough revenue for the driver.

Data Collection:

In order to gather the relevant data for creating our models, we chose to take advantage of the Chicago Data Portal. The data portal provides a dataset called "Taxi Trips" (https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew) that keeps track of all individual trips going back to 2013, and has several variables for each observation. However, the dataset has over 188 million observations for a file size of 10+ gigabytes. So we chose a smaller sample of 100,000 taxi trips in order to run the code. From here we moved on with this sample to prepare the data and address inconsistencies.

Data Preparation:

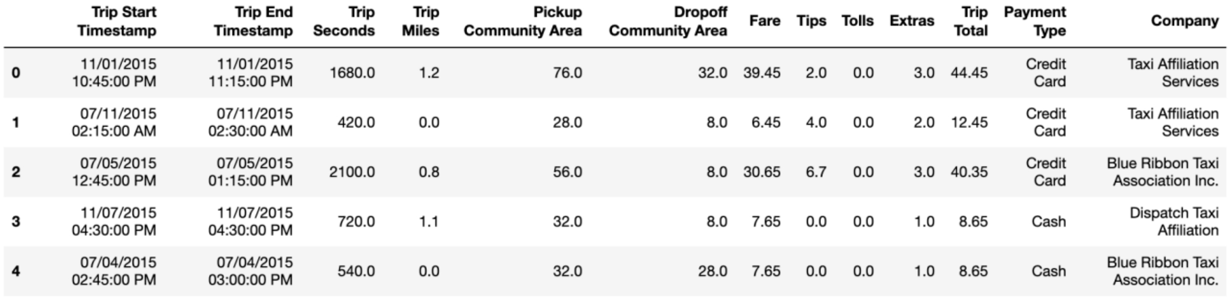
Data preparation was done in multiple steps and using different techniques.

1- First of variables that obviously are not effective towards the amount of tip were removed. Removed variables are:

- Trip ID, Taxi ID , Pickup Census Tract , Dropoff Census Tract , Pickup Centroid Latitude , Pickup Centroid Longitude , Pickup Centroid Location , Dropoff Centroid Latitude, Dropoff Centroid Longitude , and Dropoff Centroid Location.

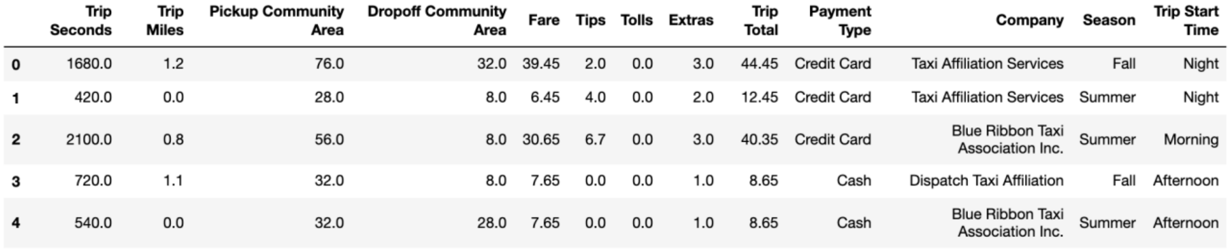
Remaining variables are:

- Trip Start Timestamp, Trip End Timestamp, Trip Seconds, Trip Miles, Pickup Community Area, Dropoff Community Area, Fare, Tips, Tolls, Extras, Trip Total, Payment Type, Company



	Trip Start Timestamp	Trip End Timestamp	Trip Seconds	Trip Miles	Pickup Community Area	Dropoff Community Area	Fare	Tips	Tolls	Extras	Trip Total	Payment Type	Company
0	11/01/2015 10:45:00 PM	11/01/2015 11:15:00 PM	1680.0	1.2	76.0	32.0	39.45	2.0	0.0	3.0	44.45	Credit Card	Taxi Affiliation Services
1	07/11/2015 02:15:00 AM	07/11/2015 02:30:00 AM	420.0	0.0	28.0	8.0	6.45	4.0	0.0	2.0	12.45	Credit Card	Taxi Affiliation Services
2	07/05/2015 12:45:00 PM	07/05/2015 01:15:00 PM	2100.0	0.8	56.0	8.0	30.65	6.7	0.0	3.0	40.35	Credit Card	Blue Ribbon Taxi Association Inc.
3	11/07/2015 04:30:00 PM	11/07/2015 04:30:00 PM	720.0	1.1	32.0	8.0	7.65	0.0	0.0	1.0	8.65	Cash	Dispatch Taxi Affiliation
4	07/04/2015 02:45:00 PM	07/04/2015 03:00:00 PM	540.0	0.0	32.0	28.0	7.65	0.0	0.0	1.0	8.65	Cash	Blue Ribbon Taxi Association Inc.

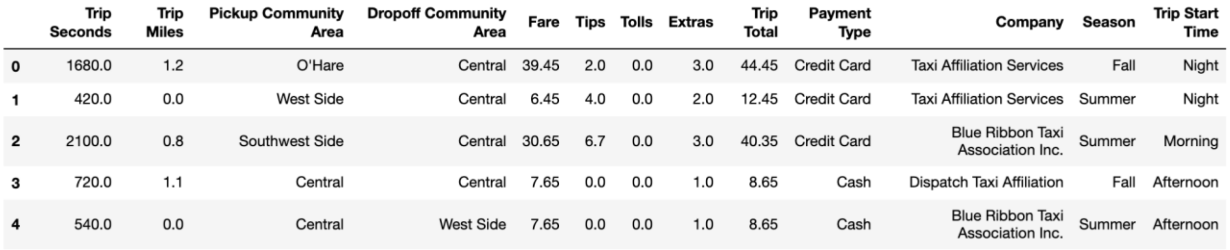
2- Trip start/end timestamps are provided in date-time format (e.g., 11/01/2015 10:45:00 PM). In order to make this information useful we needed to convert these data into useful and computer interpretable data. Trip start/end timestamps were converted to seasons and time of the day (e.g., morning, afternoon, evening, and night). Since start/end time of these trips would happen within the same season and most probably within the same time of the day (this is not very sensitive since there are different definitions for morning, afternoon, etc), only one column for season and time of the day was created.



	Trip Seconds	Trip Miles	Pickup Community Area	Dropoff Community Area	Fare	Tips	Tolls	Extras	Trip Total	Payment Type	Company	Season	Trip Start Time
0	1680.0	1.2	76.0	32.0	39.45	2.0	0.0	3.0	44.45	Credit Card	Taxi Affiliation Services	Fall	Night
1	420.0	0.0	28.0	8.0	6.45	4.0	0.0	2.0	12.45	Credit Card	Taxi Affiliation Services	Summer	Night
2	2100.0	0.8	56.0	8.0	30.65	6.7	0.0	3.0	40.35	Credit Card	Blue Ribbon Taxi Association Inc.	Summer	Morning
3	720.0	1.1	32.0	8.0	7.65	0.0	0.0	1.0	8.65	Cash	Dispatch Taxi Affiliation	Fall	Afternoon
4	540.0	0.0	32.0	28.0	7.65	0.0	0.0	1.0	8.65	Cash	Blue Ribbon Taxi Association Inc.	Summer	Afternoon

3- Number of missing values for each variable was calculated. Among all variables except for dropoff community and payment method missing values are less than 50 among 100,000 observations. Dropoff Community and payment methods for 3062 and 469 observations were missing, respectively. Missing values were imputed using the most frequent class for categorical data and mean value for numerical values.

4- Community areas in Chicago are defined using numbers from 1 to 76 (as shown in Fig.1). These numbers are not meaningful for computer. In other words, machine learning algorithms would treat these numbers as distances and for example would compare 1 to 77. However both of these numbers as shown in Fig.1 belong to the same side: Far North Side. Using the following map, pickup/dropoff community areas were converted to sides which would be more useful toward interpretation data and prediction of desired variable.



	Trip Seconds	Trip Miles	Pickup Community Area	Dropoff Community Area	Fare	Tips	Tolls	Extras	Trip Total	Payment Type	Company	Season	Trip Start Time
0	1680.0	1.2	O'Hare	Central	39.45	2.0	0.0	3.0	44.45	Credit Card	Taxi Affiliation Services	Fall	Night
1	420.0	0.0	West Side	Central	6.45	4.0	0.0	2.0	12.45	Credit Card	Taxi Affiliation Services	Summer	Night
2	2100.0	0.8	Southwest Side	Central	30.65	6.7	0.0	3.0	40.35	Credit Card	Blue Ribbon Taxi Association Inc.	Summer	Morning
3	720.0	1.1	Central	Central	7.65	0.0	0.0	1.0	8.65	Cash	Dispatch Taxi Affiliation	Fall	Afternoon
4	540.0	0.0	Central	West Side	7.65	0.0	0.0	1.0	8.65	Cash	Blue Ribbon Taxi Association Inc.	Summer	Afternoon

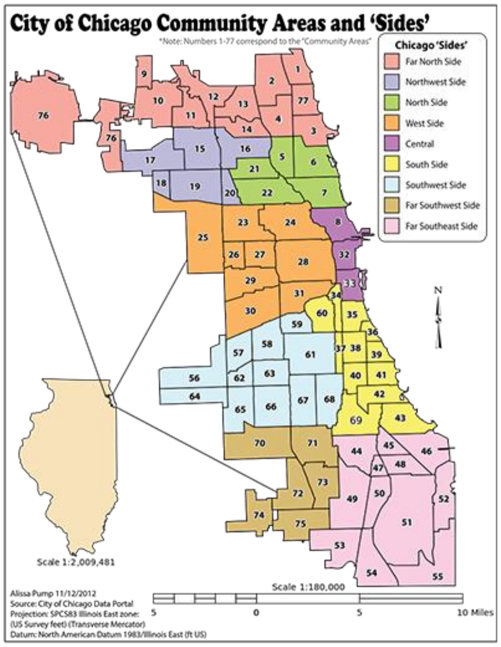


Fig.1 Chicago community areas and sides

5- Before converting categorical values into numbers which will be explained in the next step, we need to reduce the number of unique categories for each categorical variable. For example, there are 49 taxi companies among 100,000 observations. If we sort these companies by their frequency, number of observations for the last 42 companies combined are 2667 out of 100,000 observations which seems insignificant. In order to prevent some issues that will rise in the next step, the number of companies was reduced to 7 and kept all other companies under the name of "other". Same process was done for payment method and pickup/dropoff community areas.

6- In order to make categorical data interpretable for our machine learning algorithms, we need to convert this data into numbers. For this purpose, one hot encoding was implemented. One hot encoding adds column based on the number of unique values for each categorical data. If we didn't reduce the number of companies to 8, we would end up with 49 columns only for taxi companies. This is not desirable since it would increase the dimensionality (complexity) of our data without having any useful information.

7- Finally, for classification purposes, “Tips” was categorized as “0” when the amount of tip was zero, and as “1” when the amount of tip was less than $10 and as “2” when the amount of tip was more than $2.

Data Exploration:

Using summary statistics for three classes of tip (no tip, less than $10, and more than $10) suitable variables and plot types were identified.

First count of trips for each class using bar chart was plotted which is shown in Fig. 2. Based on this chart in 63201 trips customers did not pay any tip. In 36086 trips, paid tip was less than $10 and in only 713 trips, amount of tip was more than $10.

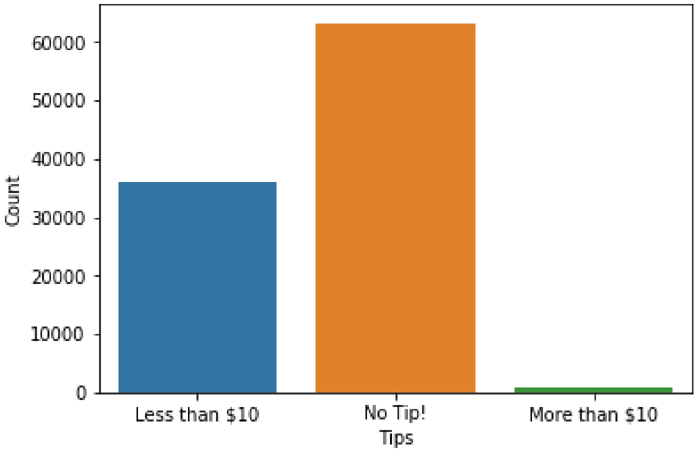


Fig. 2. Count of trips for each class of tips

Next, average trip seconds, average trip miles, average fare, average extras, average tolls, and average trip totals were plotted for each class of tip using bar chart and show in Figs 3-8. As shown in fig. 3 and fig.4, as the duration and mileage of trip increase, customers are more likely to pay tips. Further increase of duration and mileage results in higher tips. Increase of duration and mileage of trip, naturally cause the increase of fare and trip total. As shown in fig.5 and fig.8 as the fare and trip total increase, customers are more likely to pay tips. Further increase of these variables result in higher amount of paid tips. As shown in fig.6 the higher the extras the higher probability of being paid more than $10 as a tip. This is reasonable, since customers asking for extra service are more likely to pay higher amounts of tips. Finally, as shown in fig.7 average amount of tolls in trips which tips with amounts of more than $10 were paid is significantly higher than other trips. However, we need to consider average amount is still very low (0.04) and is very unlikely to be an important factor in amount of paid tip.

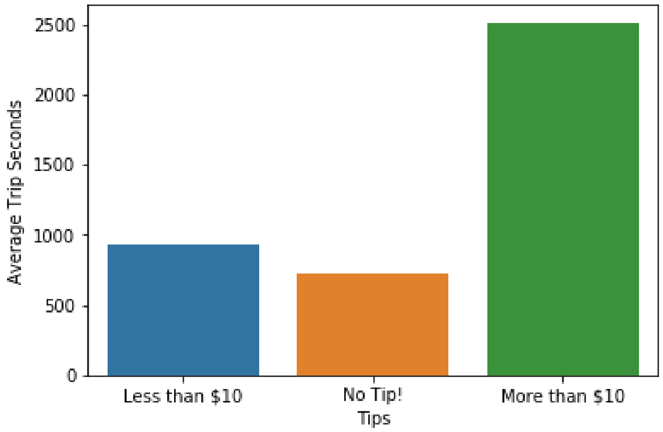


Fig. 3. Average trip seconds for each tip class

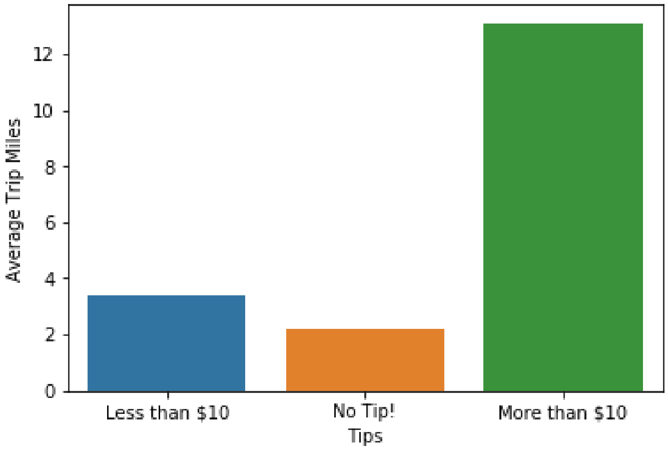


Fig. 4. Average trip miles for each tip class

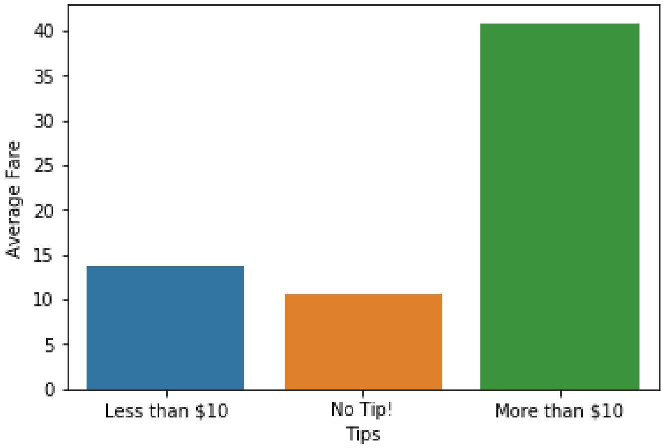


Fig. 5. Average Fare for each tip class

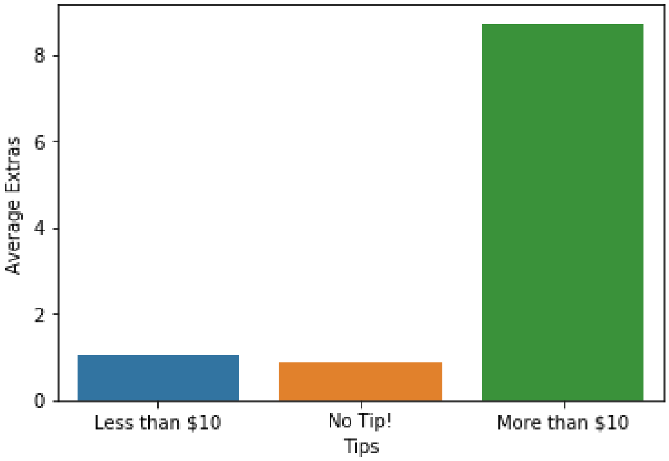


Fig. 6. Average extras for each tip class

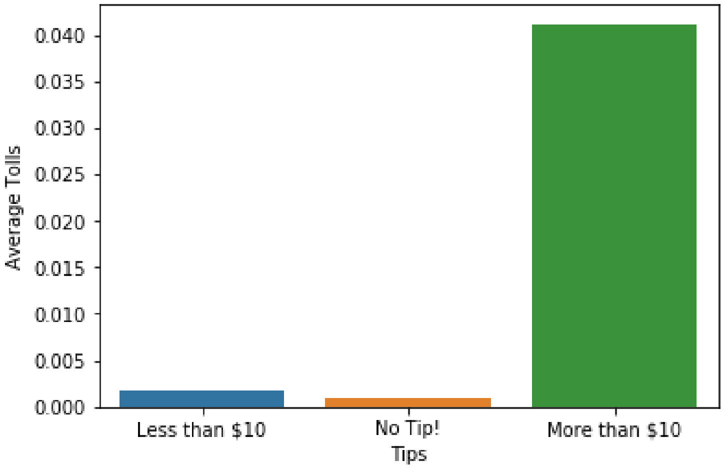


Fig. 7. Average trip seconds for each tip class

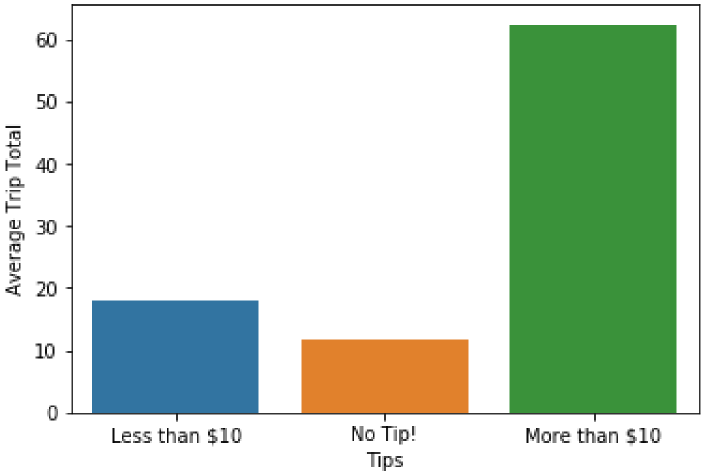


Fig. 8. Average trip seconds for each tip class

In fig. 9 average pickup sides for each class of tips is shown in three separate pie charts. Based on fig. 9 customers who are picked up from central side are more likely to not pay any tip or pay less than $10. On the other hand, customers who are picked up from O’Hare airport are more likely to pay more than $10 as a tip. Based on fig. 10, customers who are dropped off at central are more likely to not to pay any tip or pay less than 10$. On the other hand, customers with destination of O’Hare and Northwest side are more likely to pay more than $10 tips. We can also conclude that customers picked up/dropped off at O’Hare airport are very unlikely not to pay any tip.

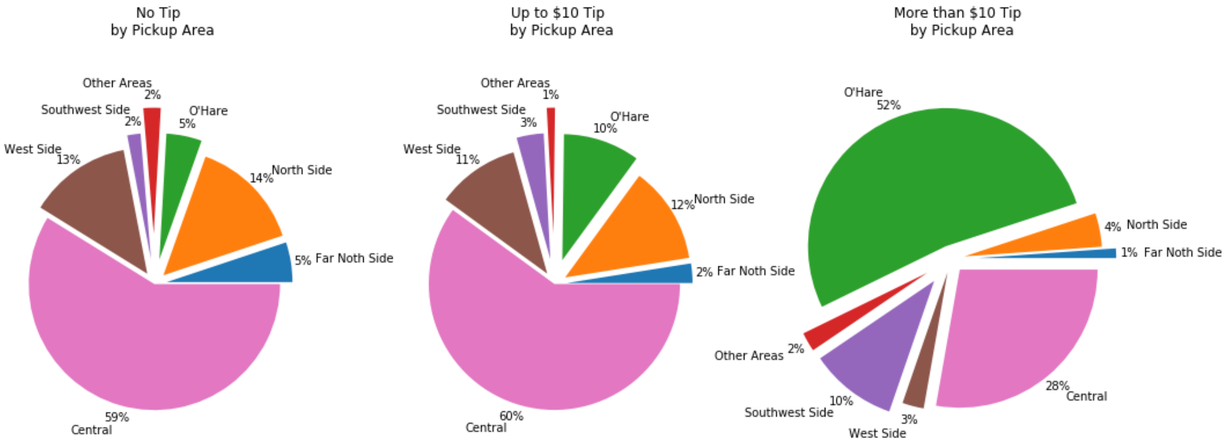


Fig. 9. Average pickup sides for each tip class

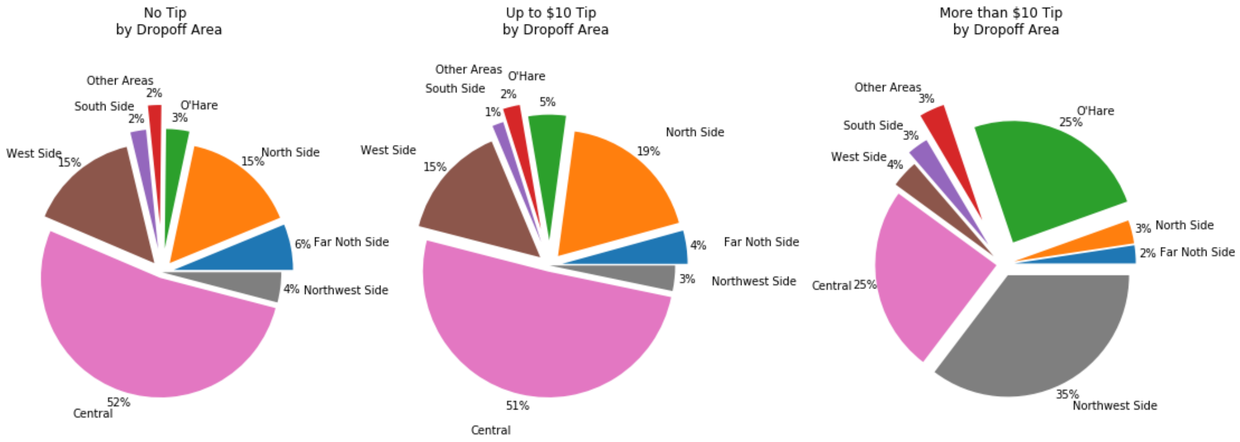


Fig. 10. Average dropoff sides for each tip class

Average payment type for each class of tips is shown in fig. 11. Based on this figure, almost all the customers who didn’t pay any tip used cash as their payment method, and almost all the customers who paid any tip (less or more than $10) used credit card as their payment method.

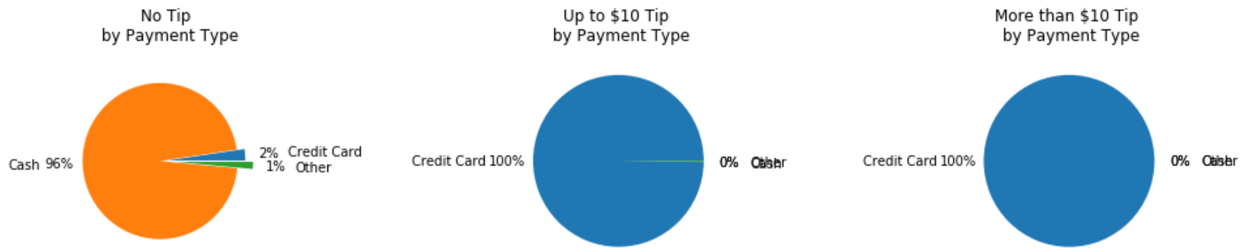


Fig. 11. Average payment method for each tip class

Data Modeling:

As mentioned earlier, we chose to create a Linear Regression model, and a Classification model using our data.

Regression

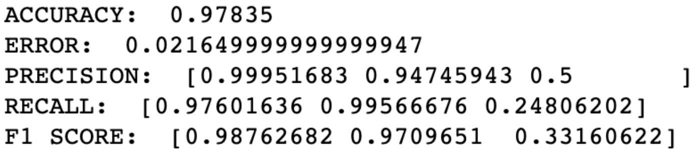
For the linear regression model we split the data as 80 percent training, and 20 percent testing data. The result of this training was a model with a correlation coefficient of almost 64%. This could be better, but likely increased by using more data. A lot of the observations show that many people do not tip at all as well, which makes the tip harder to predict. At the second attempt instead of all variables, number of variables was reduced to half (17 variables were removed). Even though the number of variables was reduced, correlation coefficient remained at 64%. However, we know that adjusted R^2 is the important factor not R^2 itself. By reducing the number of predictors and keeping the R^2 at the same range, adjusted R^2 increases. The relevant variables that were used are Trip Seconds, Trip Miles, Fare, Tolls, Extras, Payment Types, and some important pickup/dropoff locations like O’Hare, central, etc.

Classification

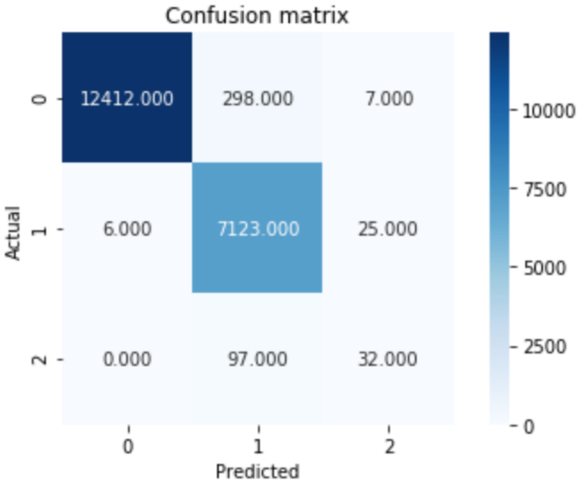
For classification 3 different algorithm was used.

- Decision tree with the variables and also subset of variables was tested.
- Support vector machine with ‘rbf’ kernel using all and subset of variables.
- And Finally, Random Forest which also is our best classifier. To create this classifier, we used a grid search algorithm to find the best combination of max_depth and n_estimators among 28 possible combinations using 10-fold cross-validation. After we found the best parameters to have highest F1_score, those parameters were used to train the actual random forest. For this model, accuracy was 97.84% and F1_score was [0.987 0.971 0.331]. All the classifier scored low for F1_score for the class label 2 (more than $10 tip) since the total number of observations for this class

was very low (713 out of 100,000 observations) compared to two other classes and classifiers did not have enough observations for training.



```
ACCURACY:  0.97835
ERROR:  0.021649999999999947
PRECISION:  [0.99951683 0.94745943 0.5       ]
RECALL:  [0.97601636 0.99566676 0.24806202]
F1 SCORE:  [0.98762682 0.9709651  0.33160622]
```


Confusion matrix
Actual
Predicted