**12/10/2019**

Edwin Kamgang

Data mining

Chicago taxi trips

This dataset is a record of Taxi trips reported to the City of Chicago in its role as a regulatory agency from the year 2013 to present. It is in the government public domain and can be downloaded from the website <https://catalog.data.gov/dataset/taxi-trips>. There are 188 million rows and 23 columns in the dataset and each row of is a taxi trip with different features that describe it such as distance traveled during the trip, time taken for the trip, pickup and drop-off locations, price and tip just to name a few. In this project we are trying to determine if these features have a correlation with the price and if so we will train a models which can make some predictions such predicting the tip and price of a taxi trip given the distance or pickup/drop off location for example, Which drop-off areas have the highest average tip, how does trip duration affect fare rates for trips lasting less than 90 minutes. This is useful information that can help a private owner of a taxi company firm in Chicago manage availability of taxi rides and increase his profit by increasing taxi availability in parts of the city where customers request taxi more often or where customers take longest and most expensive trips. It might also be useful to a taxi driver who is looking for features that will make a trip with the high tip amount.

A screenshot of a cell phone

Description automatically generated

This dataset is about 36GB and contains over 188 million taxi trips. Reading this dataset requires large memory and most laptops don’t have enough memory to even process half of this dataset. Luckily pandas provide a way to get around this by using a parameter called “chunksize”. chunksize allows us to divide this dataset into smaller, manageable chunks. For this dataset, we will store each 100k trips into a file, then we will pick a random file for this assignment.

A picture containing screenshot

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Now we reduce the dataset to 44.2MB, but we can still save computation time by removing all the unnecessary and duplicate columns. First, we will remove the “Pickup Centroid Location” and “Dropoff Centroid Location” because both of these columns contain duplicate information. Secondly, we will remove “Pickup Census Tract” and “Dropoff Census Tract” since we have the exact pick-up and drop-off locations.

A screenshot of a cell phone

Description automatically generated

According to city of Chicago’s official website, the base fare of taxi trips in Chicago is $3.25, so any trip where the price is less than or equal to $3.25 is either miscalculated or inaccurate information. This dataset contains number of trip where the price is less than or equal to 3.25. One way to correct the price of these trips is to use the “Trip seconds” and “Trip Miles” to predict the price of the trips, but “Trip seconds” and “Trip Miles” for these trips equal to 0. Not only that but also “Trip start Timestamp” and “End start Timestamp” are exactly the same. In this case we can’t calculate the price of the trips so will just drop these rows.

A screenshot of a social media post

Description automatically generated

A picture containing screenshot

Description automatically generated

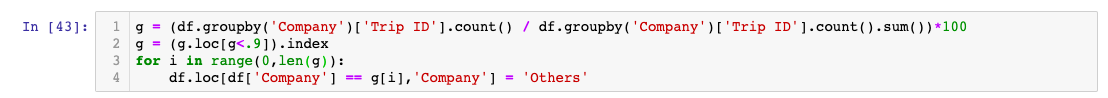
There are number of missing values in “Trip Miles” column, but unlike the fare column, we have other columns to fill this column, one way to fill this column is to calculate average distance travel in a second and multiply that to “Trip seconds” and that will give you very good approximation of number of miles travel.



A screenshot of a cell phone

Description automatically generated

There are 47 different companies in this dataset, but 95% of the trips are done by only seven companies. To easily visualize their market share we will label “Others” to any company that has less than 0.9% of market share.



A screenshot of a social media post

Description automatically generated

We can’t not predict how much someone will tip using only this dataset since we don’t have any information about the passenger but what we can do is predict weather someone will likely tip depending on their pickup location and duration of their trip. To do this we will create new categorical columns called “Tip”, and this Tip will be true if the passenger tip over $ 0.1 and will be False otherwise.



A screenshot of a cell phone

Description automatically generated“Trip start Timestamp” column contain the time stamp of each trip and to visualize and analyze the demand of each hour we will create new column called “Time” and this column will include the hour of the trip and weather it was Am or PM. To do that we can uses “slice” to only get the hour and am/pm. Also, we will be inserting this column next to the “Trip start Timestamp” column.

A screenshot of a social media post

Description automatically generated

According to this dataset, the average trip is about two and a half miles, and that makes sense since almost 70% of the pick-ups occur in and around downtown, and passenger tent to take taxis to a walkable distance. The average taxi trip in Chicago lasts 10 minutes and that is very high considering the average trip is only two and a half miles, and that is either due to the traffic in downtown or drivers start and kept running their meters while they load and unload their passengers belongs. The average trip costs only $10, but some drivers made more depending on the tips they got, even though the average tip is low, but we found a number of passengers who added a tip that is more than the cost of the trip, some passengers tip as much as $50.

A close up of a map

Description automatically generatedA close up of a logo

Description automatically generated

A close up of a logo

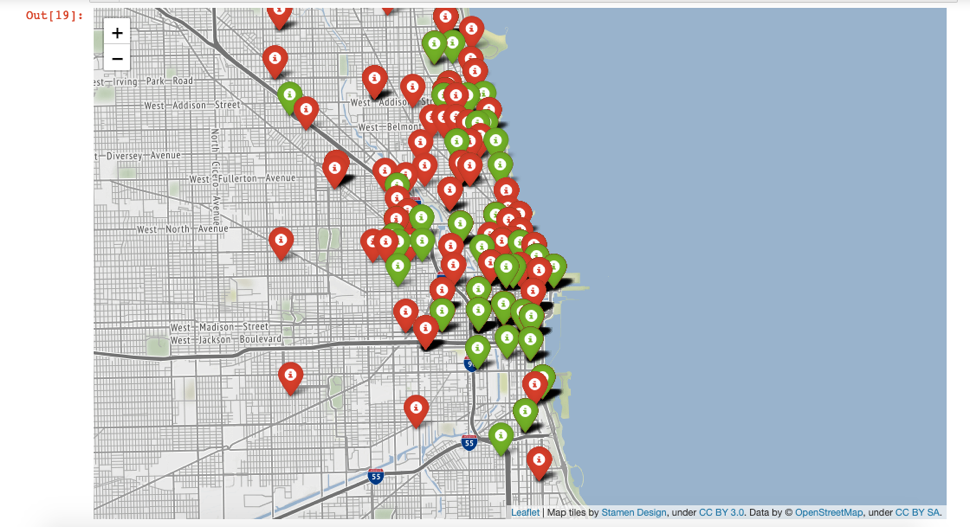
Description automatically generated

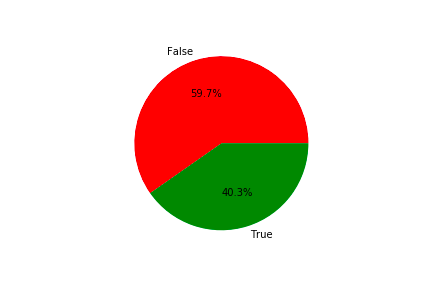
A close up of a logo

Description automatically generated

If we assume each taxi has one driver, then there are 2773 drivers in this dataset, and the average driver made $500 during this time frame. Only 6% of the drivers in this dataset made over $1000.

We tried to predict whether someone will tip or not, depending on their pick-up or drop-off location, in a  logical way people who live in high-end neighborhoods or going to certain part of the city will likely to tip and people who live in a low-income area will not tip, but we found that is not the case. Even though there are places like “Chicago midway international airport” where almost all passenger tip, but that is not the case in other places and people tend not to tip base on solely on their location but other factors contribute their decision, like the service they received or how clean their taxi was. Almost 60% of passengers in this dataset don’t tip and people who pay using credit cards are more likely to tip.





There are over 40 different companies in this dataset, but it turns out only 7 companies did over 95% of the trips. Taxi Affiliation services did over half of the trips, which is surprising that one company is responsible for more than half of all taxi trips in Chicago. These numbers may be different if we evaluate all the 188 million taxi trips in this dataset but considering we pick random 100k trips, this is still surprising that one company is this big while all other companies share less than 50% of the market.

A close up of a card

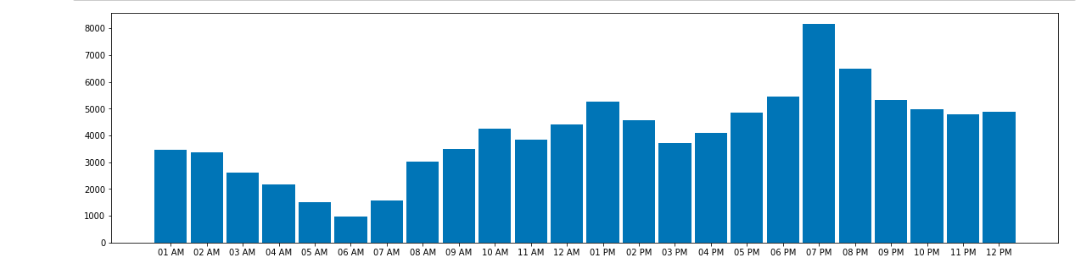
Description automatically generated

There are 100 thousand trips in this dataset and as can see the heatmap below, the majority of the trips are in and around Chicago downtown. Some of most popular pickup location includes, “O'Hare international airport”, “Chicago midway international airport” and “Lincoln park”. The longest trip in this dataset is about 550 miles and the shortest trip is less than a mile.

A close up of a map

Description automatically generated

According to this dataset, the prime time for taxi drivers is 7-8 pm, while the slowest is 5-6 am. As you can see from the graph below, the demand for taxis varies depending on the time of the day. The need for taxis in the early morning is very low while the evening hours are the busiest. That is not surprising since the majority of the trips in this dataset are downtown, and people usually go out at night. One thing that explains why the demand for taxis is so high from 7 pm to midnight is that people tend to drink and choose not to drive while they intoxicated, and they choose the safe way and take taxis.



The correlation graph below shows “Fare” and “Trip seconds” have 0.82 correlation, that makes sense since the longer trip takes, the more money a passenger pays, Also “Trip Seconds” and “Trip Miles” have 0.6 correlation, which lower than what we expected, but it also makes sense, because some short trips may take longer due to traffic, construction or weather and some longer trip take shorter time due to less traffic or the time of the trip. “Fare” and “Pickup community Area” have 0.47 correlation, that is not that high but it shows how people in certain places tend to go the same direction, for example, people in the airport who take taxis are usually visitors and they tent to go hotels in the downtown area.

**A screenshot of a cell phone

Description automatically generated**

***Logistic regression:***

In the column named "Tips" which represents the tip amount for the trip, there is an even amount of instances where there is a tip to where there is no tip so one of our prediction goal was to see if we could accurately predict the likely hood of a customer paying a tip after a trip. For this we used logistic regression because it is a suitable statistical method for analyzing data where the target class is dichotomous (there are only two possible outcomes), in this case, weather a customer is going to tip or not. It does this by fitting the plotted data into an s shaped fit line where the datapoints at the top represent one class and those at the bottom represent another class. So, it calculates the probability of data belonging to one of those two classes.



The first step was to create a new column with Boolean(true or false) content representing trips where the tip was greater than zero as true and trips where the tip was zero as false and named that column "Tip".(Note: "Tips" is the column for tip amount and "Tip" is the true or false representation of the tip).



Next we have to determine which features affect the tip he most, in other words the features which are highly correlated to the tip. For this we used the heatmap and observed the column "tip" for features which have a correlation higher than 0.1.



"Trip seconds", "Trip Miles" and Pickup Community Area" are the only columns that meet these criteria. The following diagrams show that even though the correlation is low there is a moderate separation of the two classes at the top and the bottom and this pattern can be used to fit a logistic curve for logistic regression.

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A requirement for using logistic regression is that the independent variables have a weak correlation between each other, however the column "Trip seconds" and "Trip Miles" have a correlation of 0.57 which is pretty high so I ended up using only the features "Trip seconds" and "Pickup Community Area" as the independent variables. These two variables have different units, so we standardize them so that they are scaled on a mean of zero and a standard deviation of +1 and -1. This is important because we don't want one variable with higher overall values to overwhelmingly affect the prediction more than the others(for this project we actually showed results with and without standardizing to show how standardizing affects the model) .We then divided the data into training and test set in a ratio 80:20 then create an instance of logistic regression for the model as in the following piece of code:

A screen shot of a computer

Description automatically generated

Sklearn has a library called metrics that can score the result of your prediction model using the test set. We got an accuracy score of 0.620403009417583 with and without standardizing which is pretty low. To improve the score we used more independent variables be decreasing the required correlation to 0.4 and got the following: 'Fare', 'Trip Miles', 'Trip Seconds', 'Pickup Community Area ', 'Trip Total' which gave us way better scores; 0.8442679012995212 without standardizing and 0.8404798232230231 with standardized data.

Logistic regression performed better with more independent variable because with more variables you have more information about the data which helps in the accuracy of probability calculation and therefore better classification. Overall, all those features have a relatively low correlation with the "tip" class(the highest is 0.12) so the better course of action will be to aim for using more features to get more information on the data for the model than to use a minimum correlation value and choose those features that meet it. In a way it's like compensating for the lack of information due to low correlation by using many independent variables with low correlation.

***Clustering:***

For the clustering algorithm we used k means clustering. k-means is an algorithm the finds the k number of clusters for the data which means the number of clusters that best differentiate the variables. To do this it initializes centroids with random data points, and it keeps repeating until the is no change to the centroids. It then computes the sum of the squared distance between data points and all centroids and assigns each data point to the closest cluster (centroid). The final step is to compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

Since k means algorithm is unsupervised, there is no target class using during the clustering, the clustered are create based on an underlying pattern that the has to discover or can use to prove or disprove an initial hypothesis. For this project our hypothesis is that taxi trips of the from taxi provider company hence similar features so will cluster together. That is for example people will choose a certain taxi company to travel short distance and another to travel long trips. Another interpretation of these clusters is that people in a certain community area tend to choose a particular taxi company.

Before performing the clustering, we said PCA on the data which is basically an algorithm used to reduce the number of dimensionalities in a dataset while conserving the correlation with the dependent variable. It finds the features that affect the target class the most and also if there are 2 independent features that are highly correlated with one another, PCA will remove one of them since they portray the same information about the class, in doing so it avoid overfitting of the data when performing a fitting model. PCA also help with visualization on a dataset with more than 3 columns which cannot be plotted on a single graph such as this one.

The graph shows that most of the information on the data are contained in principal components 1, 2 and 3 so those are ones we choose to go forward. The following scattered graph show the relationship between these three principal components.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

To find the right number of clusters for the data, we draw an inertia graph using the principal component. The graph shows a curve which represents an information gain at each point, the point where the change in curvature decrease and becomes constant is called the elbow point and it tells the right number of clusters for the dataset.

A screenshot of a cell phone

Description automatically generated

The elbow point for this data is 3 so we cluster the data using 3 clusters and get the following results.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Looking at the following pie chart we can see that our clusters match our hypothesis because the blue cluster taxi affiliation services has most of the taxi trips followed by blue ribbon taxi association then dispatch Taxi affiliation. Our interpretation from this clustering algorithm is that taxi trips from the same company share similar features.



**Multiple linear regression**

For this dataset, we decided to predict the fare of trips. There are many companies in this dataset, and they calculate their price using different factors, some calculate the price base on the duration of the trip, and some calculate using the duration and the distance of the trip. Some companies charge extra for Luxury taxis, bigger taxis, and Special-Needs taxis. Even though we don’t know the types of taxis, we can still predict the fare of trips using the duration of the trip and the pickup location. We used multiple linear regression. Multiple linear regression is an algorithm that uses multiple independent variables to predict the dependent variable and, in our case, we have multiple factors that contribute to the price of the trip, for example, the location of the pickup, duration of the trip and the drop-off location. After performing this algorithm, using only the pick-up location and the duration of the trip, the result was good, but we added a drop-off location and we got a mean squared error of 5, which is good considering is the fare range of 0 to $500.

A screenshot of a cell phone

Description automatically generated A picture containing object

Description automatically generated