NYC green & yellow taxi trip time data analysis

ABSTRACT

The project focuses mainly on performing data mining tasks on a combination of 2 data sets coming from 2 sources. NYC Green Taxi and NYC Yellow Taxi. The intention of this project is to create a model that is trained on the data records fetched for the month of December 2018 since it comprises of the new year which is prominently the major week for people opting for taxi trips. The projects outcomes indicate on correctly predicting whether or not the trip will result into a tipped ride or not. The report also depicts the various analysis that have been done to assist visual aid to the dataset.

1. INTRODUCTION

The following term project is aimed at working with the humongous collection of taxi trip time data provided from the NYC Green and Yellow cabs. This project is focused on generating meaningful and insightful information about passenger patterns and taxi hot spots in and around NYC. The data is collected from different parent organisations namely the NYC Green taxi[1] and the NYC Yellow taxi union[3]. This diversification in capturing data from two different parent organisations from UCI data repository along with NYC data science repository will help us clarify our concepts further strengthening our command over the concepts of data cleaning and preparation for fruitful data analysis.

Before, moving further let us understand our problem domain and the reason why we are participating on the following data analysis task. Many a times we have seen that we focus on larger domains looking for insights and have difficulties in generating fruitful information, however, we often forget that simple things in life share the highest amount of data. Moving on the same line of principle, we would like to shed some light on the motivation behind the selection of this topic in particular. "Taxi trip details" :- The name in itself speaks volumes about what we are referring to at first hindsight we can say that we are trying to identify and map people who rely merely on public transport / people who do not prefer to drive their own vehicle. Secondly, one can also say that this project is pinpointed towards a structural research to aid fuel a startup invitation based on taxi rides in and around the NYC. Yes, both of these are accurate insights one can pin to just by the name of our project, however, we are looking to throw more light on many other observations one could possibly think of being the person behind that wheel of study ie. The average cost of a particular trip originating from a particular location. This helps understand which area's of the city have a higher possibility of generating a decent trip time. We are also looking forward to find solutions to understand which areas of the NYC generate trips that tip the most, etc. Moving forward let's understand our data sets.

Let us first dive deep into the difference between two taxis and the motive behind using two taxis.There are mainly Yellow Medallion Taxis in New York City and it is used for the people of NYC to commute[2]. It provides the transportation specifically through street-hails[4]. There is no way of pre arranging the Yellow taxis. On the other hand, there are taxis called as Street Hail Livery commonly known as Green taxis. You can either street hail these taxis after 110Th street or in the out-bounds of NYC[2] or let out your vehicle get a license and then pick up some one else. These cabs were usually brought up in this area because there was a scarcity of yellow cabs while led to illegal street hails. They can also arrange a pre-defined trip. There are some interesting facts that we are going to explain and visualize about this data set. Yellow Medallion taxis has boosted the employment rate in NYC[5]. Lack of Yellow taxis were giving rides to pre booked cabs. The prices are much higher than the pre booked cabs so it is invested by a private investor and they helped the TLC to lease out drivers[2].

Plan of action for this term project:

- These data sets contains a lot of data sorted according to months, the first aim would be to merge these data sets and convert it into a single entity so that we could apply some pre processing techniques and move to the next step.
- Cleaning the data set and storing them in a RDBMS would be our next aim.
- The next part would be doing some statistical analysis on the data set and analysing the distribution of how time vs distance works. Here we will provide with some intuitive statistically inclined graphs which will help us to analyse the data efficiently.
- Obviously merging two data sets of such a huge size would lead to anomalies and discrepancy in data. Hence, we will perform feature engineering and various Data cleaning techniques to resolve and get a clean data set for Data mining task.
- Finally we will do a Data mining task (Mostly a regression tasks because it will be fun and intuitive to play with numbers)

2. PROGRESS

We were successfully able to download the 115M row record database for the Yellow Taxi database and subsequently 50M row record database for the Green taxi data set. The data set being in raw form we had to procure the base step for any data mining task aka " data cleaning ". We resahped our target outcome of the project to reduce overhead on query processing timeline by filtering the same objectives thought of earlier but now for a strict timeline of 7 days for the week encapsulating the new year, provided that people bound to travel outwards usually to places which attract more number of people often making parking a pain, people tend to prefer cabs more often than not. This timeline also sutis perfectly for our analysis as more and more people travelling outbound also factor in the amount of people who rely on public transport for their daily commute in a hustling city like of NYC. This also ridicules the thought of such strata of the society to use public transport on that timeline as people tend to spend a bit more money on themselves during such occasions also the target analysis for this experiment is based on "tip-amounts" during the ride.

We are trying to factor in a correlation is to when does a ride earn more tip, based on the distance travelled ? based on the total amount of the ride ? Also, if there was a tip provided what percent of the trip amount was that based on - was it the total trip cost consisting the local surcharges and taxes / just the base fare excluding the taxes. Also we want to identify trips originating from which fare point generates the highest amount of tips is it the Airport ? is it someplace people travel the most for their personal leisure like a club / bar / restro ?

The above factors can only be identified and worked upon when we are ready with a schema to work upon. For this we performed our data cleaning procedure. Steps Considered:-

- The date and trip originating time should fall between 12/24/2018 - 01/01/2019
- Only the attributes which help add value to the analysis excluding attributes which explore individual taxes
- The payment method
- The minimum fare should be greater than 2.5
- Trip distance should be greater than 0
- Target attribute tip_amount

our goal is to showcase which trip origin and destination aka trip route is the best to shot for a cab driver to generate a tip and then if it does which route generates the highest percentage of the total amount. We have used python dataframe to clean our data set coming from the NYC org website as the original raw source. We used our limiting conditions mentioned above for both the yellow and green data sets. Later we merged both the dataframes to create one single datasource for our data analysis. We then converted the csv sources and loaded them into our RDBMS using MySQL. Let's talk about the filteration based on the payment method: The dataset has various payment methods:

- 1: Credit Card
- 2: Cash
- 3: No charge

	trip_distance
count	1,664,993.00
mean	3.08
std	3.97
min	0.01
25%	0.99
50%	1.67
75%	3.20
max	140.58

	tip_amount
count	1,664,993.00
mean	1.50
std	2.36
min	0.00
25%	0.00
50%	1.00
75%	2.06
max	175.08

	total_amount
count	1,664,993.00
mean	15.77
std	13.63
min	2.50
25%	8.16
50%	11.30
75%	17.30
max	570.06

Figure 1: Statistical Analysis

- 4: Dispute
- 5: Unknown
- 6: Voided

To have a fair rationale, we only are considering the type of trips where a formal method of payment was guaranteed. i.e. Cash / Credit. Thus, to keep the dataset cleaned on these parameters we also procured a filtration method in the noted above step. There were many - amounts associated with the types 3-6 which usually would denote a refund / dispute as mentioned and this data usually becomes the outlier and pulls the analysis skewing away from the single point of thought that the analysis should be on, hence, the decision was made to remove these rows. Let us now see the reports generated from the statistical analysis section:

3. DATA CLEANING THE PROCESS

In this section of the report we would like to point out the various methods and tasks performed on the semi-cleaned data procured from the second phase. The data now clear of anomalies was subjected to screening over the attributes that remain available for feature selection. The feature selection process comprised of a single major step of comparison of correlation between the available attributes and the target class. The features with least interest were subsequently removed from the dataset. Various standard SQL queries were run on the dataset on MySQL to understand the nature and spread of the target class which are detailed and explained

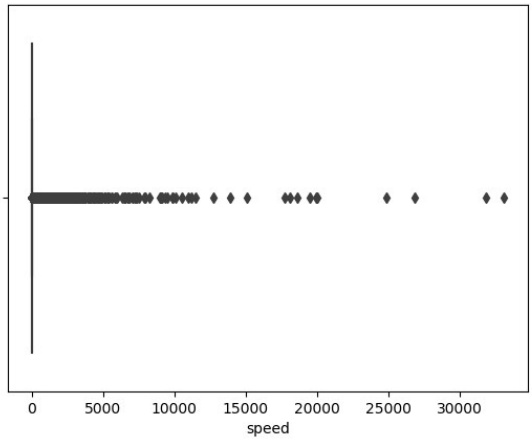

Figure 2: Box Plot Analysis

later in this report. The features removed were type of payment method, and type of cab ie. yellow or green. We also resolved the datetime objects like pickup_datetime and dropoff_datetime from datetime objects to numerical values and then resolved them further into sections like afternoon, morning and evening grouping them for better analysis. The same was done with location of pickup and dropoff to rationalise the dataset based on which part of new york has better tippers for their rides. The final target class of tip amount was normalised using the one hot encoding method of a yes/no type based on all tips which are 0 are a No and rest with a monetary value are Yes.

4. DATA VISUALIZATION

To better understand your data one should always visualize their data. Let's see some plots that we did from our side on this dataset.

From the figure we can clearly see there are many outliers here but prominently most of the rides the speed is between 10-13km/hr. Therefore, the avg speed of people in NY is 10-13km/hr. Next let's visualize a bit about the number of rides for a week nearing the new years eve in 2018.

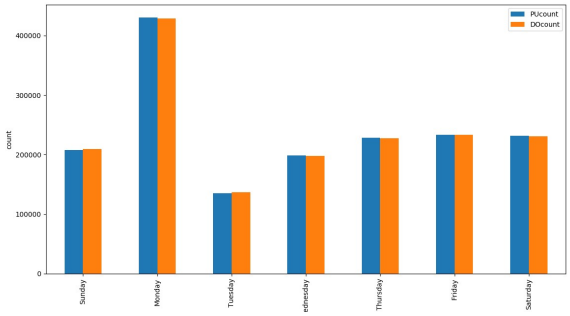

Figure 3: Trips in a week

We see that the highest number of rides are the most for the Monday. Just like any laxed up weekend the business is booming on the following day for city riders. To more detailed analysis we also dug deep to find out the average distribution of rides per hour on a given day. To enhance the same we used the monday's data which had the highest reported rides.

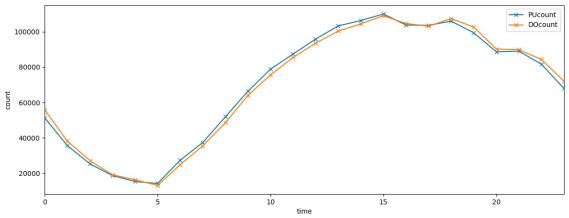

Figure 4: Trip by time

We clearly see that the graph peaks up as the day progresses, around the hours 15-20 is the highest peak meaning 3:00pm to 8:00pm speaks like the time people take the ride back home from offices.

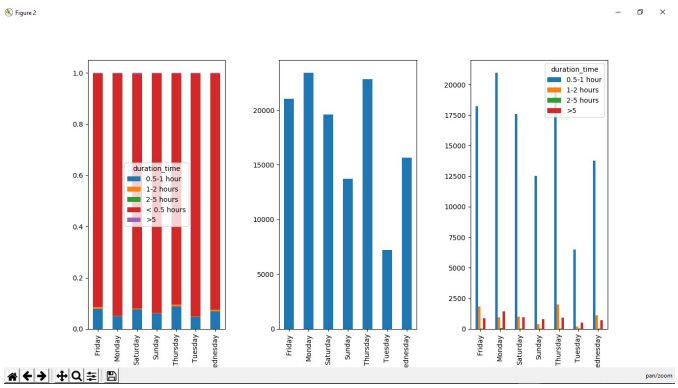

Figure 5: Trip duration analysis

The above 3 figures represent various different visuals on the same criteria that is trip time. However, with each plot the meaning and the data that has been identified is completely different. For example the first example is a stack bar graph for all rides based on their completion time. The red dominating the short rides and the rest is a known fact that uber and cabs are usually for a short distance. The next picture is a plot of the long rides relatively less visible in the earlier plot. The last plot is the last plot is to showcase the irregularities reported in the data as data shows more 5hour drives than 1-2 hour drives combined which points to the fact that the data is a bit buggy. Or maybe the NYC traffic is a real demon. But our analysis shows that the amount of big rides are usually pretty beast like 5 hour drives or else they usually are the smaller ones comprising a time lesser than 30 minutes.

Let us see a few more plots on trip distance since we saw enough plots on the duration, distance is the next big factor that matures towards a tip for the driver as the longer the ride the larger the base fare the more is the 10% component.

As we proved earlier most of the rides take 10-13km/hr and in this graph there are various points between 20-40 which is taking more than 15-20 hours. That might be an

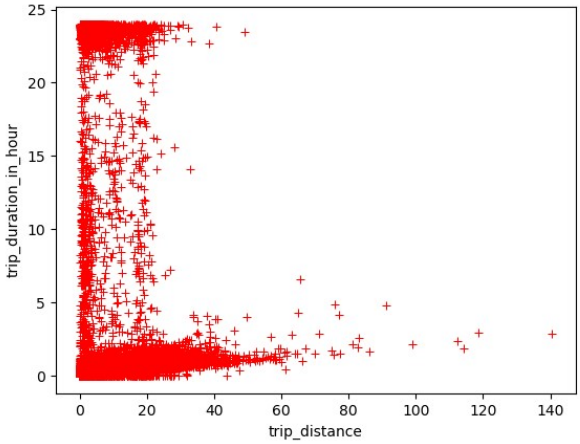

Figure 6: Trip by distance

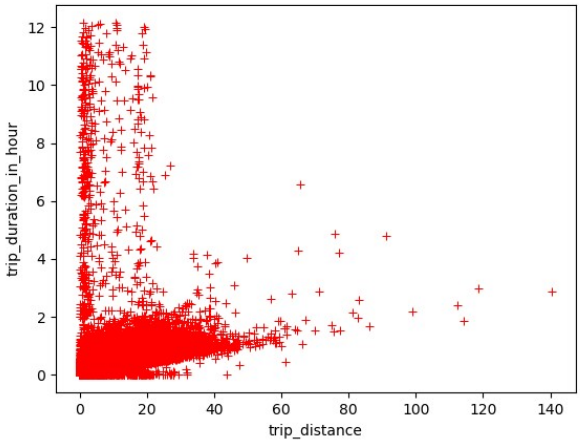

Figure 7: Trip by distance logarithmic duration

anomaly and could be due to extreme traffic congestion. After doing log transformations and providing linear relationship we came to know that the data could be cleaned more and brought down to only those data whose logarithmic duration was less than 2.5 hours.

Now, as we clearly visualized our data from multiple standpoints its time we dig deep in the actual mining step of our project.

5. DATA MINING

5.1 Methodology

The data mining activity performed on the dataset was of classification, we here based on 1.6M record worth data are set to predict based on 9 qualifying attributes whether a ride is set to record a tip or not. The process post cleaning was sent to the trainer, the trainer algorithm used are both SVM and Random forest Classification. The trainer program was fed with around 1.2M worth data records to make sure we do work on a big data platform, the data had an even mix of tipped and not tipped rides to make our learner robust and not have skewed results. The testing set was worth 0.4M records roughly 25% of the original dataset.

5.2 Results

The algorithm returned a 98% accuracy for the random forest algorithm and a 95% accuracy for the SVM algorithm. The accuracy was calculated using the accuracy package from sklearn. The results indicate that the dataset we cleaned was absolutely as per industry standards and duely fit for learning. The data model thus prepared is fit for distribution now in its cleaned form for industry giants like Uber to perform deep analysis on enhancing the driver incentives. The defacto analysis performed on the dataset alongside the other factors mentioned was about the location which generated the most tips was none other than the airport. The longest rides were also originating from the airport, however, the highest sum of tips provided combined did not come from the airport area meaning people travelling long distances do not tend to honor the standard 10% norm of tips on their rides. The Manhattan area received the highest pickup and dropoff's.

6. LEARNING

To breifly summarise our learning outcome we have itemised them for better understanding.

- Importance of Visual analysis on uncleaned data
- Importance of visual analysis on cleaned and ready for analysis data
- Importance of clearing Missing values
- Importance of Data Preparation
- Importance of having a clear vision for performing analysis task
- Team work

7. CONCLUSION AND FUTURE WORK

This is the part of the project where we weigh our effort throughout and provide constructive feedback on our overall work, we would like to report that the expectation moving into this project at Phase 1 was well established in our previous para based on Data mining, the results we obtained were not what we expected as a blind guess earlier. However, we can conclude saying that the data we prepared throughout for our mining exercise was unique and could not be replicated for any other task. The results are naive as we targeted only 1 attribute from the present 12 for our analysis. The data is huge and sparse there are many possibilities to perform multiple analysis projects picking up any of the 9 eligible attributes selected for this particular task. The accuracy reported for this project is a sample for this particular model, we can confirm that a drop in about 10% can be noticeable if random data points are fed to the system as real time data as this data corresponded to only a month ie. December of 2018. However, we can also mention that the extension to this project can be in various dimensions.

8. REFERENCES

[1] Green taxi dataset. https://data.cityofnewyork.us/Transportation/2017-Green-Taxi-Trip-Data/5gj9-2kzx. Accessed: 2020-02-17.

[2] Nyc gov taxi data. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2020-02-17.

[3] Yellow taxi dataset. https://data.cityofnewyork.us/Transportation/2017-Yellow-Taxi-Trip-Data/biws-g3hs. Accessed: 2020-02-17.

[4] U. Patel. Nyc taxi trip and fare data analytics using bigdata. 10 2015.

[5] Y. Tang. Big data analytics of taxi operations in new york city. *American Journal of Operations Research*, 09:192–199, 01 2019.