Project Report

Name: Preetesh Verma
EntryNo.:2018EEB1171

Name: Tanya Gupta
EntryNo.:2018EEB1149

Foundations Of Data Science
Course Code:CS521

Abstract

On-demand, app-based ride services like Uber have become an important part of today's transportation system with its flexibility and quick responsiveness. But often we see that on booking a cab from a location, the time required by the cab driver to reach the spot may vary from 5 minutes to hours. In this project, methods and models were used to explore and forecast Uber Cab Services Customer Data, that includes number of pickups at random locations and random time in Austin,Texas, to give useful insights to the owner about demand or load at a specific location and time that will help to reduce this delay. For this purpose **Sarima model** was used.The State of the Art technique **LSTM(Long Short Term Memory)** is also used for the same purpose.Methods, such as **XGBoost and Linear Regression** have also been applied to estimate the number of trips been made and the distance travelled.A recommendation based on **demographic filtering** has also been prepared to figure out the most highly rated travel locations in the city along with other features.

1 Introduction

Often we see that on booking a cab from a location, the time required by the cab driver to reach the spot may vary from 5 minutes to hours. This causes unsatisfactory response at the end of customer for the services provided. Here are some common associated problems associated with the cab drives:

- The drivers are exhausted.
  Sometimes drivers are too tired to service the request.Most taxi operators give drivers the option of logging in and out of the system as per their choice.Though the company has recommended that they stay logged in to the system for a minimum of 12 hours but not hard and fast.

- Drivers do not make enough.
  Even after investing long hours of commitment for the service the monthly take-home salary may not be enough.

- Traffic problem.
  Most of the times, in a city where traffic moves at 9 km per hour during peak traffic time ultimately result in cab driver not landing up on time.

- Receiving the booking information late
  It might also be a problem or error from the section that handles interface between customer and the cab company, or a result of technical issues with the booking software.
- Affect of external factors such as rain,humidity wind on the number of trips being demanded by the people.Also if a particular day is a public holiday then will the number of trips been made in the city soars up or maintains the same level.
- Availability of the drivers in the most demanded area all the times at the appropriate locations so as to reach the customer in as little time as possible.
- Problems with the vehicle.
  Assuming that a vehicle is on the go for 16 hours a day, there is not enough time to put the vehicle for maintenance or service. This sometimes can result in vehicle break down.

Problem Statement:

To find number of pickups, given location coordinates(latitude and longitude) and time, in the query region and surrounding regions as well as predicting the optimal location for placing the drivers to optimize the output along with the complete analysis of the dataset to derive insightful inferences.

Justification:

In our project we have used the previously available data with their pick up and drop points with hourly update. Then using a time series model we have tried to forecast the relative load or demands in that particular area for a certain period of the day.Using this data will help the owner as well as cab drivers to deliver satisfactory services.

This information will enable the drivers to set their logging time accordingly, as at time of low rush only a few will be required to log in while others may rest and at the peak hour others may be placed beforehand at the location.

This reduces the total time drivers have to remain logged in and thus may do some other work that will resolve their issue of low income. As the drivers are preplaced, they may avoid the traffic easily and reach the spot without much delay. Also the issue of information delay or problem in vehicle can be dealt with as drivers will have their mindset so will act accordingly on their discretion.Also a demographic filtering would help both the drivers and the users to identify the highly recommended cars for travel as per other users ratings.

Model used:

AUTO REGRESSION is a time series model that uses observations from previous time steps as input to a regression equation to predict the value at the next time step.To find the order of the Auto Regression we use the Partial Auto Correlations Curve.

MOVING AVERAGE in statistics, a moving average is a calculation to analyze data points by creating a series of averages of different subsets of the full data set.To predict the order of the moving average we use the Auto correlation Coefficients Curve curve to find the order.

For the working of this project we have used inbuilt models provided by statsmodel python library for time series analysis and forecast.The models used are ARIMA and SARIMAX.SARIMA is just like ARIMA with the addition of seasonal component. This package return the model with lowest AIC(Akaike information criterion).. Data sets for this problem were taken from data.world website .

Neural Networks have also been used to implement LSTM(Long Short Term Memory). LSTM are known as gated neural networks as well because these gates can learn which data in a sequence is important to keep or throw away. By doing that, it can pass relevant information down the long chain of sequences to make predictions. Demographic Filtering has also been employed to the cars and road trips to identify the best of the lot in the city to travel.XGBoost and Linear Regression techniques have been employed to predict the number of trips that would be occurring. Tkinter library has also been employed to generate a GUI based output which requires the user to enter the start location and end location co-ordinates and would suffice the user with the path connecting the two locations.

2 Literature Survey:

As early as 1970s, the autoregressive integrated moving average (ARIMA) model was used to predict short-term freeway traffic flow [M. S. Ahmed, A. R. Cook, "Analysis of freeway traffic time-series data by using Box–Jenkins techniques"].

Since then, an extensive variety of models for traffic flow prediction have been proposed by researchers from different areas, such as transportation engineering, statistics, machine learning, control engineering, and economics.[Lv, Y., Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: a deep learning approach."]

ARIMA model integrates the above two statistics models(AUTO REGRESSION and MOVING AVERAGE) for forecasting. An inefficient taxi service system leads to more empty trips for drivers and longer waiting time for passengers and introduces unnecessary congestion on the road network.Taxis in urban areas such as NYC are equipped with GPS devices, providing second by second location information. The unprecedented amount taxi trip data generated from GPS equipped taxis allow researchers to directly observe and analyze the system performance and recommend various taxi related services.Another key factor in the demand of the taxi is the season.

SARIMA MODEL or Seasonal ARIMA model was first suggested in 2004.A seasonal ARIMA model uses differencing at a lag equal to the number of seasons (s) to remove additive seasonal effects. As with lag 1 differencing to remove a trend, the lag s differencing introduces a moving average term. The seasonal ARIMA model includes autoregressive and moving average terms at lags.

STATE OF THE ART TECHNIQUE–

Using recurrent neural networks **LSTM Long Short Term Memory** to predict the time series data.Sima Siami Namini has published paper in 2018 comparing the performance of ARIMA and LSTM for time series forecasting.Unlike modeling using regressions, in time series datasets there is a sequence of dependence among the input variables.Recurrent Neural Networks are very powerful in handling the dependency among the input variables. LSTM is a type of Recurrent Neural Network (RNN) that can hold and learn from long sequence of observations. The algorithm developed is a multi-step univariate forecast algorithm

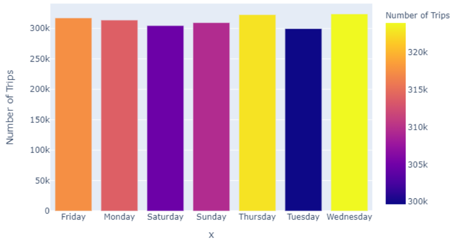
Another technique is using the **MAPA i.e. Multiple Aggregation Prediction Algorithm** algorithm proposed in 2013 by Kolassa and Kourentzes.The algorithm takes advantage of the time series transformations that can be achieved by non-overlapping temporal aggregation. Temporally aggregating a time series can cause various of its components to become more or less prominent with direct effects on model identification and estimation. MAPA uses multiple temporal aggregation levels, allowing multiple views of the data to be considered during model building and subsequently combined in a final forecast. Demographic filtering in combination with collaborative filtering was proposed in 2005 by F.O.IsinkayeaY,FolajimibB and Ojokohc which would give a hybrid model by combining the ratings, features and demographic information about items in a cascade hybrid recommendation technique.A similar technique has been employed by us to generate the recommendations but without much use of the collaborative filtering with the model primarily focusing on distribution of the items and ratings given by past users.

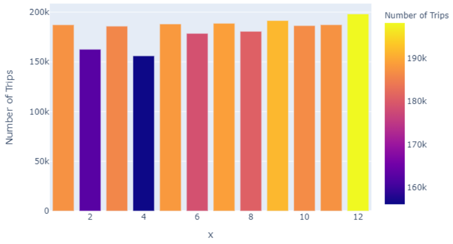
3 Results:

3.1 Data Exploration:

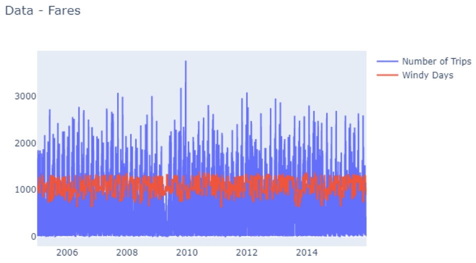
EDA is an important step in Data Analysis as it helps in understanding the data and also to draw insights from the same.The initial Exploratory Data Analysis of the data shows that the number of cab drives demanded in the city is varying a lot with every day.But some very common pattern is also observed.

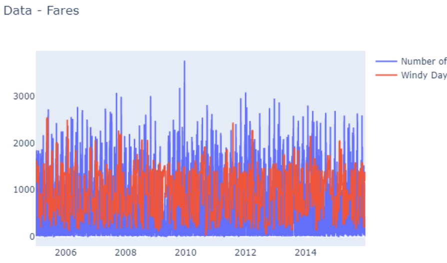
- The data shows that the number of cab demands was almost independent of the day of the week as over the course of time the number of trips undertaken on each week day almost amounted equally.While month wise distribution of the number of trips clearly shows that the number of cab drives is maximum in the months of December and January showing a direct affect of holidays while the number drops to the lowest in the month of April (a possible reason could be school exams).



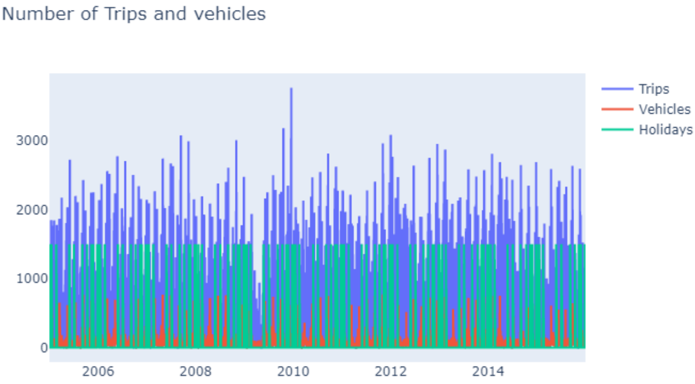


- We can clearly see that both wind and humidity features have little to no effect on the number of trips being made on those days.The graphs below show that the number of trips being made on a windy or rainy day and a normal is almost same.The graphs show precipitation levels and wind speed respectively.

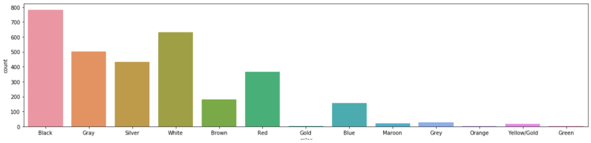


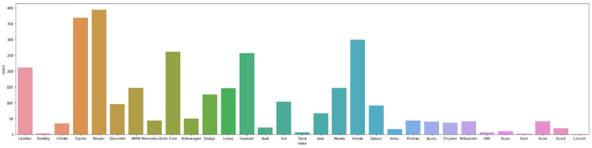


- **On every holiday the number of cab drives demanded is increased significantly.**The fact can be seen in the below graph.Thus showing SEASONALITY also supported by the monthly drives graph above.The number of cab demands soars on days when it is a holiday.

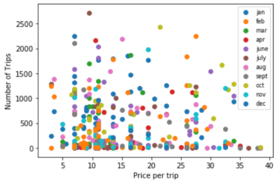


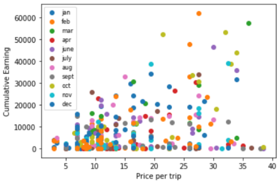
- Another important takeaway from the data is the location of the pick and drop points is that the most of the pickups and drops and centered around the city with a few around the airport area region as well but no where else.(negligible amount of outliers present in the data).This helps in up-voting the performance of the SARIMA predictor.
- a common and quite understandable fact is the gradual increase in the number of cab drives demanded annually with each passing year thus showing an **upward trend**.
- Here are two graphs showing the distribution of the features of the car such as the company and color which were used in demographic filtering based on users ratings.



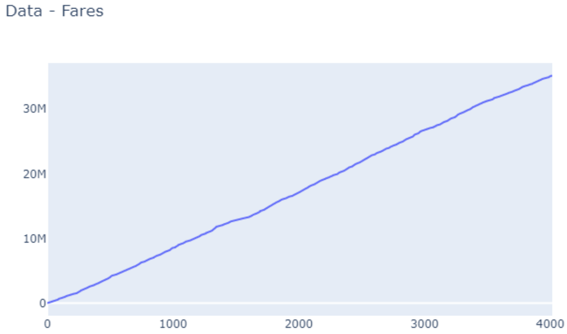


- Here are two graphs showing the cumulative earning and daily number of trips as scatter plot with month being the label.

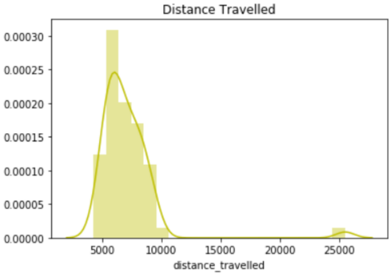




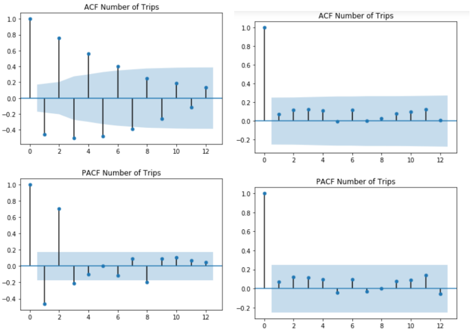
- Here is another observation regarding the cumulative total gross earning of Uber from the city of Austin in a period of almost 10 years with the scenario mostly following a linear pattern which is quite understandable given the continued popularity of the Uber services as a mean of communication.



- On viewing the data in monthly and weekly basis the distribution remains more or less same but on viewing it in bimonthly basis there is a significant difference with the peak being clearly shifting towards the right side or the latter time side.
- The distance distribution of the Uber cab drives is also pretty interesting with the bulk of the rides being for distances between 5 to 10 kilometers.



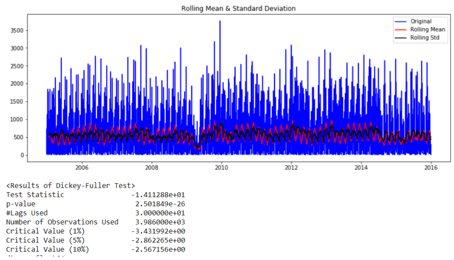
- The PCAF and ACF curves for the two time series data also give the p and q values for the SARIMA model which proves the series to be of the first order.



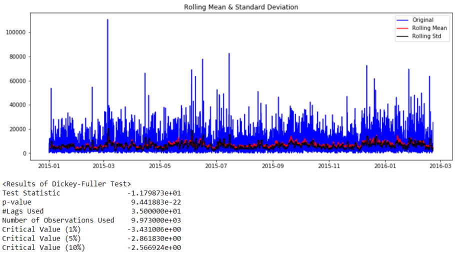
3.2 SARIMAX:

ACF curve gives us the auto correlation of the series with its lagged values(describing how well our present values relate to past values) while the PCAF curve tells the correlation of the residuals(what remains after removing the effects of the already explained) with the next lag value. The given time series on visualising through graphs proofs to be stationary via rolling mean test with the mean remaining more or less constant. The above results are also verified by the Dickey fuller test which is a theoretical way of predicting whether or not a time series has seasonality and trend.The **Augmented Dickey-Fuller test** is a type of statistical test called a unit root test.The intuition behind a unit root test is that it determines how strongly a time series is defined by a trend.There are a number of unit root tests and the Augmented Dickey-Fuller may be one of the more widely used. It uses an autoregressive model and optimizes an information criterion across multiple different lag values.

The null hypothesis of the test is that the time series can be represented by a unit root, that it is not stationary (has some time-dependent structure). The alternate hypothesis (rejecting the null hypothesis) is that the time series is stationary. On running the Dickey Fuller test we find that the unit root does not exist and so the null hypothesis is rejected and the series is STATIONARY.Here are the p values.



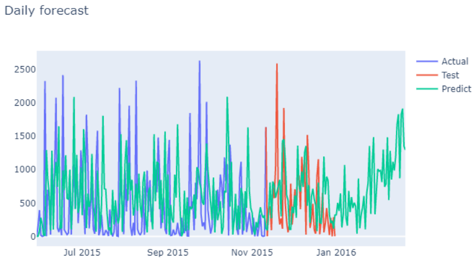
```
<Results of Dickey-Fuller Test>
Test Statistic                -1.411288e+01
p-value                        2.501849e-26
#Lags Used                     3.000000e+01
Number of Observations Used    3.986000e+03
Critical Value (1%)           -3.431992e+00
Critical Value (5%)           -2.862265e+00
Critical Value (10%)          -2.567156e+00
```

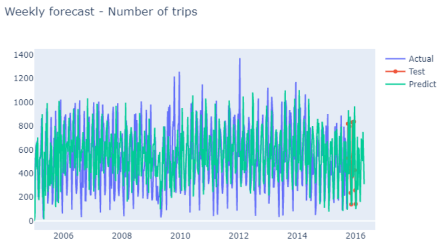


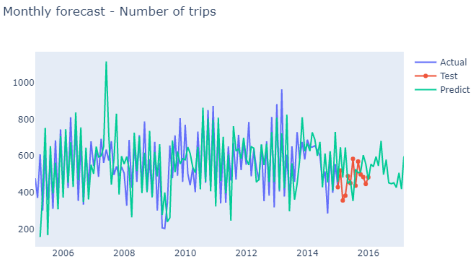
```
<Results of Dickey-Fuller Test>
Test Statistic                -1.179873e+01
p-value                        9.441883e-22
#Lags Used                     3.500000e+01
Number of Observations Used    9.973000e+03
Critical Value (1%)           -3.431006e+00
Critical Value (5%)           -2.861830e+00
Critical Value (10%)          -2.566924e+00
```

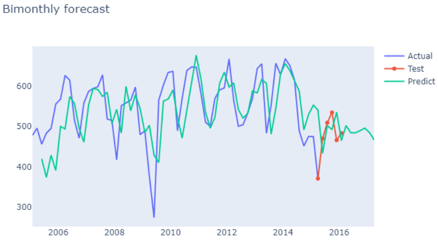
From the above information it is clear that the data did not necessarily needed to be first decomposed to make it stationary (but for the better functioning of SARIMA model we did).

For decomposing the model we used the statsmodel library .To account for the Seasonality as well we used the SARIMAX model and used the time lag of 12 months to fit the data and predict the number of total pickups.The data was divided into weekly,monthly,bi-monthly basis and individually the model was being fitted. Here are the results of the SARIMA model on different aggregated scales.







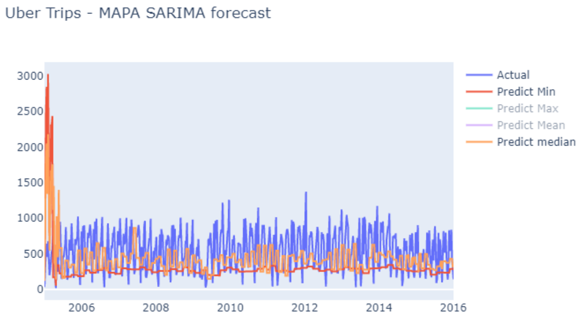


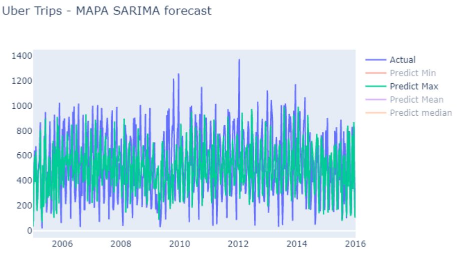
The results in different time aggregated scales also form the basis for the next used algorithm which is MAPA.The results of SARIMA have been pretty good for all the time-scales.

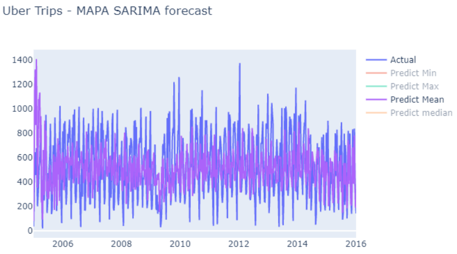
3.3 Multiple Aggregate Probability Algorithm:

Multiple time series are constructed from the original time series,using **temporal aggregation** .These derivative series helps in strengthening or attenuating the signals of different time series components.A time series forecast method is employed in each of the series and thereafter the the series is reconstructed using different mathematical methods**forecast combination**.

The MAPA algorithm runs the data in the form of weekly ,monthly and bimonthly basis with SARIMA as forecasting method,which accordingly values the significance of a particular attribute and returns the model thereafter . The metric of max and mean are doing pretty reasonable job on the data while the rest are initially overshooting the original series as seen in the following graph.



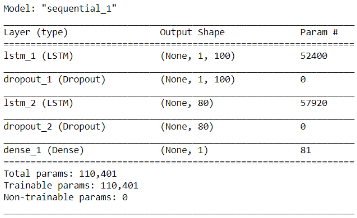




The MAPA models performance is in line with the results of the SARIMAX model. The RMSE error kept on decreasing on as the model was trained on data grouped in larger time frames.With the model doing pretty well on the test data in all the cases.

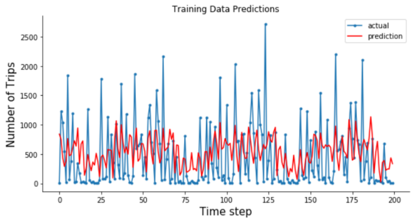
3.4 LSTM Results and Comparison

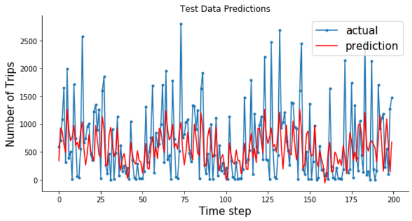
The state of the Art technique LSTM is also used by us to perform the forecasting task. The model description is as follows:



```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_1 (LSTM)                (None, 1, 100)            52400
_________________________________________________________________
dropout_1 (Dropout)          (None, 1, 100)            0
_________________________________________________________________
lstm_2 (LSTM)                (None, 80)                57920
_________________________________________________________________
dropout_2 (Dropout)          (None, 80)                0
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 81
=================================================================
Total params: 110,401
Trainable params: 110,401
Non-trainable params: 0
_________________________________________________________________
```

The model comprises of two LSTM layers of 128 and 64 units along with dropout layers to tackle the issue of vanishing gradients.LSTM is capable of handling sequential data so it is preferable to be employed in time-series data.Here are the results on training and test data.



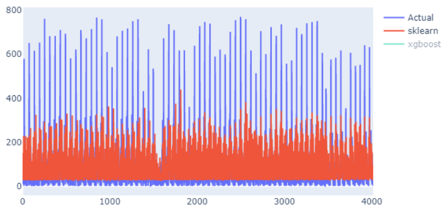


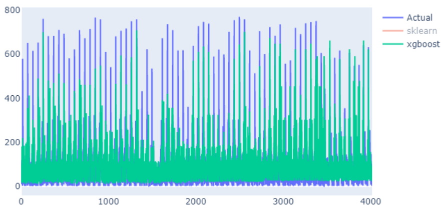
The performance is LSTM is not bad but falls short in comparison with the SARIMA and MAPA algorithms with reason being the simplicity of the data which does not necessarily needs a complex neural architecture for execution as LSTM provide non linear solution in contrast to the statistical methods which are linear.Sarima model produces the best results of the three algorithms used.LSTM works better if we are dealing with huge amount of data and enough training data is available, while SARIMA is better for smaller dataset

3.5 Linear Regression and XGBoost

Linear Regression and XGBoost techniques have been employed to compute the distance travelled by the Uber cars.Linear Regression finds a linear relation between the features to predict the output value by reducing the loss function using Gradient Descent optimization technique.Whereas XGBoost is Gradient Boosting Algorithm.

Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made.Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. Here are the results for the same.

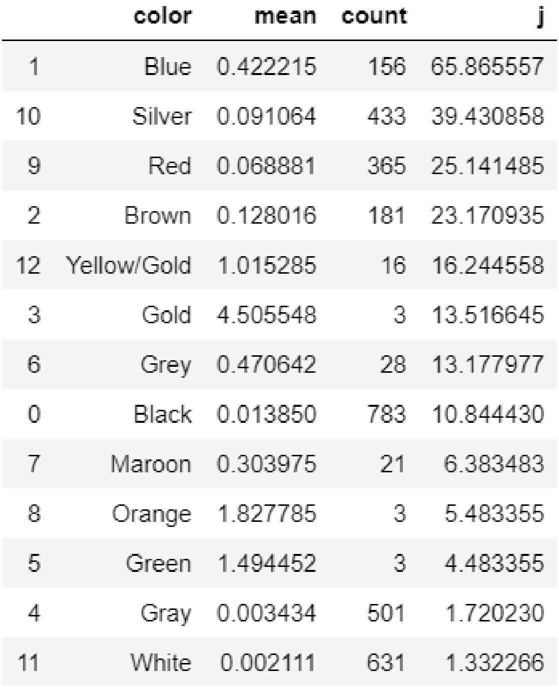




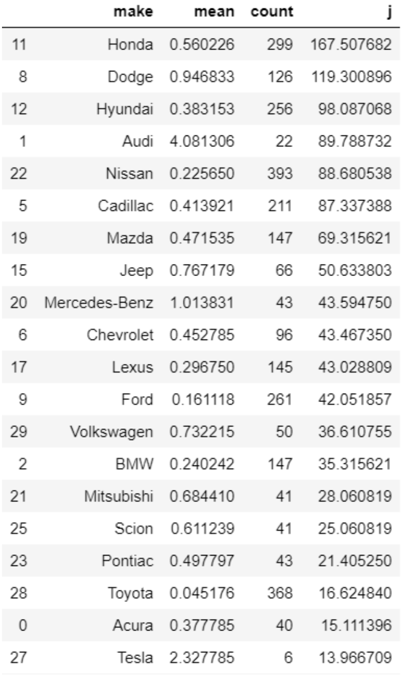
Clearly the XGBoost algorithm outperforms Linear Regression quite easily. For location case we have only used the SARIMA algorithm since most of the trips are concentrated in the centre of the city and in the coordinates system the difference is almost negligible since a decimal value change in the latitude and longitude value is equal to kilometres of change on the ground.

3.6 Recommendation System and GUI

We have used demographic filtering of the dataset along with the user provided ratings to develop a ranking system for the roads in the city.The EDA of the car features has been mentioned above and the ranking of the car features based upon the distribution and the ratings provided by the users are as follows with column named 'j' showing the final ratings.



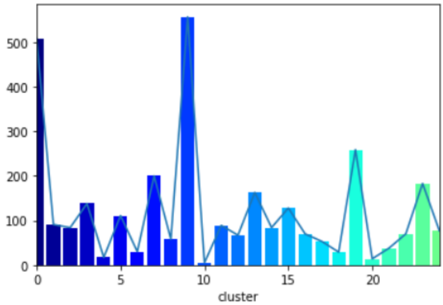
	color	mean	count	j
1	Blue	0.422215	156	65.865557
10	Silver	0.091064	433	39.430858
9	Red	0.068881	365	25.141485
2	Brown	0.128016	181	23.170935
12	Yellow/Gold	1.015285	16	16.244558
3	Gold	4.505548	3	13.516645
6	Grey	0.470642	28	13.177977
0	Black	0.013850	783	10.844430
7	Maroon	0.303975	21	6.383483
8	Orange	1.827785	3	5.483355
5	Green	1.494452	3	4.483355
4	Gray	0.003434	501	1.720230
11	White	0.002111	631	1.332266

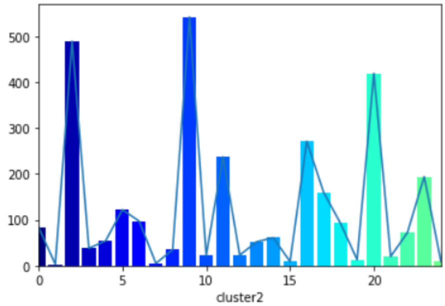


	make	mean	count	j
11	Honda	0.560226	299	167.507682
8	Dodge	0.946833	126	119.300896
12	Hyundai	0.383153	256	98.087068
1	Audi	4.081306	22	89.788732
22	Nissan	0.225650	393	88.680538
5	Cadillac	0.413921	211	87.337388
19	Mazda	0.471535	147	69.315621
15	Jeep	0.767179	66	50.633803
20	Mercedes-Benz	1.013831	43	43.594750
6	Chevrolet	0.452785	96	43.467350
17	Lexus	0.296750	145	43.028809
9	Ford	0.161118	261	42.051857
29	Volkswagen	0.732215	50	36.610755
2	BMW	0.240242	147	35.315621
21	Mitsubishi	0.684410	41	28.060819
25	Scion	0.611239	41	25.060819
23	Pontiac	0.497797	43	21.405250
28	Toyota	0.045176	368	16.624840
0	Acura	0.377785	40	15.111396
27	Tesla	2.327785	6	13.966709

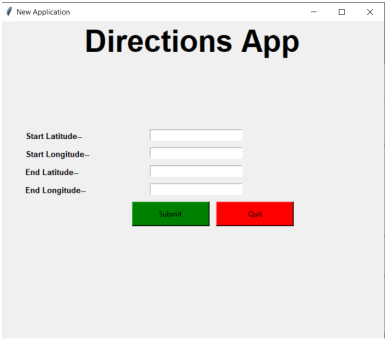
Blue Honda and Silver Hyundai stands out in the demanded car brands and color.

Thereafter we used the K-Means Algorithm to divide the city into 25 clusters based upon the most frequent pickups and drops.This would help us in breaking the big city into small segments.Here is the distribution of the same followed by a map in map visualisation.





We have also used the Tkinter library provided by Python to prepare a GUI which would take input from the user and thereafter project the path between the entered coordinates by the user on a map. Here is the image of the dialog box of the GUI.The GUI created by us is an attempt to bridge the gap between the user and the technology as it would be suffice for the user to view the location he want to travel in the form of co-ordinate map.(though this has already been done by Google Maps but an attempt has been made to perform the similar task in a different manner)



New Application

Directions App

Start Latitude--

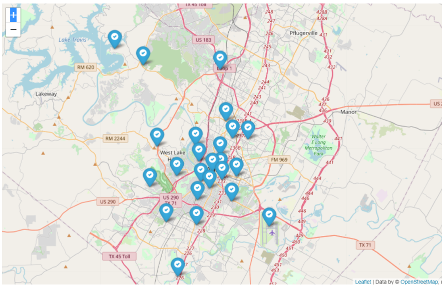
Start Longitude--

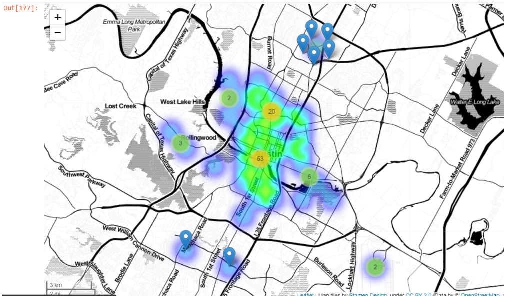
End Latitude--

End Longitude--

3.7 Map Visualisation

Here are the maps which we have generated using the Folium library.The first map shows the cluster centres created by k-Means algorithm while the second one shows the preferred location to place the drivers in the form a heat map to maximize the output of the company as they would be closer to the customers.The map visualisations are made so as to provide a better pictorial representation of the otherwise tabular data and also so that the understanding is enhanced.





Out[177]:
+
−
Emma Long Metropolitan Park
Bee Cave Road
Lost Creek
West Lake Hills
Rollingwood
Southwest Parkway
West William Cannon Drive
Walter E Long Lake
Decker Lane
2
20
3
53
6
2
3 km
2 mi
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap,

4 Conclusion and Learnings

Uber data used in the project after being analysed provides several meaningful insights regarding the travelling habits of the people of the city of Austin,Texas in USA.The weekday and monthly analysis is also supported by the common understanding of the fact that on holidays people travel more to visit parks,malls etc. The gradual increase in the revenue generated by the Uber is also in support with the growth of the company in the global market of Tours and Travels.The predicted locations of the optimum placement of the drivers would certainly help the company in improving it's ratings and user-experience.

There were several learning aspects in the project given to us. Time-Series Analysis is one of the main tasks in Machine Learning in the present day world and it includes stock market prediction as well as sequential data analysis. Usage of both traditional methods along with the modern day neural networks has improved the understanding of the time series data and ability to do forecasting. EDA of the dataset has also been an important step in the project which has certainly improved graphical visualisation skills. The addition of the Tkinter library along with the map visualisation has been a new learning experience as these things were used for the first time. Employment of Gradient Boosting and the MAPA algorithms has also been new with the both algorithms being relatively new in their respective fields of use.

****