Supervised ML- Regression

TRANSPORT DEMAND PREDICTION



* **By Kaiwalya Zankar**

**DataScience Trainee At AlmaBetter**

* **INTRODUCTION-**

Transport demand forecasting is to predict future transport demand when establishing transport plans within a given budget

Transport demand is a quantitative input to evaluate supply strategy of transport facilities and land use planning.

Presented as travel volume based on transport system usage, including transport facilities and transport services.

The derived demand was created by continuous interaction of transport systems and activity systems

* **OBJECTIVE-**
* Data Prepping/Wrangling.
* Checking the Null values for cleaning the Dataset for further analysis.
* Checking the unique values for Analyzing the Dataset for further analysis
* Exploration of Neighborhood group, Neighborhood, Room type, Price, Reviews.
* Correlation between variables.
* The main objective of project is to create a predictive model using traffic data provided to us and historic bus ticket sales data from Mobiticket to predict the number of tickets that will be sold for buses into Nairobi from cities
* **PROBLEM STATEMENT-**

this challenge asks you to build a model that predicts the number of seats that mobiticket can expect to sell for each ride, i.e. for a specific route on a specific date and time. There are 14 routes in this dataset. All of the routes end in a Nairobi and originate in towns to the North-West of Nairobi towards Lake Victoria.

The towns from which these routes originate are:

* Awendo
* Homa bay
* Kehancha
* Kendu bay
* Keroka
* Keumbu
* Kijauri
* Kissi
* Mbita
* Migori
* Ndhiwa
* Nyachenge
* Oyugis
* Rodi
* Rongo
* Sirare
* Sori

These routes from these 14 origins to the first stop in the outskirts of Nairobi takes approximately 8 to 9 hours from the time of departure. From the first stop in the outskirts of nairobi into the main bus terminal, where most passengers get off, in central business district, takes another 2 to 3 hours depending on traffic.

The three stops that all these routes make in Nairobi (in order) are:

1. Kawangware: the first srop in the outskirts of Nairobi
2. Westlands
3. Afya Centre: the main bus terminal where the most passengers disembark

All of these points are mapped here.

Passengers of these bus (or shuttle) rides are affected by Nairobi traffic not only during their ride into the city, but from there they must continue their journey to their final destination in Nairobi wherever that may be. Traffic can act as a deterrent for those who have the option to avoid buses that arrive in Nairobi during peak traffic hours. On the other hand, traffic may be an indication for people’s movement patterns, reflecting business hours, cultural events, political events, and holidays

dataset prepping-

the dataset has 51645 observations having 10 columns, datatset contain 0 null values.

the target variable is “number\_of\_tickets”

Data Description

Nairobi Transport Data.csv (zipped) is the dataset of tickets purchased from Mobiticket for the 14 routes from “up country” into Nairobi between 17 October 2017 and 20 April 2018. This dataset includes the variables: ride\_id, seat\_number, payment\_method, payment\_receipt, travel\_date, travel\_time, travel\_from, travel\_to, car\_type, max\_capacity.

Uber Movement traffic data can be accessed [here](https://movement.uber.com/). Data is available for Nairobi through June 2018. Uber Movement provided historic hourly travel time between any two points in Nairobi. Any tables that are extracted from the Uber Movement platform can be used in your model.

Variables description:

* ride\_id: unique ID of a vehicle on a specific route on a specific day and time.
* seat\_number: seat assigned to ticket
* payment\_method: method used by customer to purchase ticket from Mobiticket (cash or Mpesa)
* payment\_receipt: unique id number for ticket purchased from Mobiticket
* travel\_date: date of ride departure. (MM/DD/YYYY)
* travel\_time: scheduled departure time of ride. Rides generally depart on time. (hh:mm)
* travel\_from: town from which ride originated
* travel\_to: destination of ride. All rides are to Nairobi.
* car\_type: vehicle type (shuttle or bus)
* max\_capacity: number of seats on the vehicle

Challenges faced-

To make the data tenable for understanding and further analysis , the data set was analyzed for identifiable statistical trends and patterns. After preliminary analysis, the following steps were undertaken to transform the data into a systematically workable dataset:

The following are the challenges faced in the data analysis.

To filter the given data

Feature engineering – to get the more required features that will ease the further analysis

Feature to be selected to get the required output

Model implementation

APPROACH

We performed the Outliers treatment and normalized the features for better results. I have used supervised learning regression analysis models like linear regression, implementing lasso regression and ridge regression, training gradient boosting regressor, training xg boost , random forest for the purpose of training the dataset to predict future supply

Hyperparameter tuning plays an important role to predict the best model among the above regression models

Tools used

The whole project was done using python, in google colaboratory. Following libraries were used for analyzing the data and visualizing it and to build transport demand model.

* Numpy: For some math operations in predictions.
* Pandas: Extensively used to load and wrangle with the dataset.
* Matplotlib: Used for visualization.
* Seaborn: Used for visualization.
* Datetime: Used for analyzing the date variable.
* Sklearn: For the purpose of analysis and prediction.
* Math:
* Xgboost:
* Warnings: For filtering and ignoring the warnings.

The below table shows the dataset in the form of Pandas DataFrame.

ride\_id, seat\_number, payment\_method, payment\_receipt, travel\_date, travel\_time, travel\_from, travel\_to, car\_type, max\_capacity

pandas dataframe

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Ride id | Seat number | Payment method | Payment receipt | Travel date | Travel time | Travel from | Travel to | Car type | Max capacityy |
| 0 | 1442 | 15A | Mpesa | UZUEHCBUSO | 17-10-17 | 7:15 | Migori | Nairobi | Bus | 49 |
| 1 | 5437 | 14A | Mpesa | TIHLBUSGTE | 19-11-17 | 7:12 | Migori | Nairobi | Bus | 49 |
| 2 | 5710 | 8B | Mpesa | EQX8Q5G19O | 26-11-17 | 7:05 | Keroka | Nairobi | Bus | 49 |
| 3 | 5777 | 19A | Mpesa | SGP18CL0ME | 27-11-17 | 7:10 | Homa bay | Nairobi | Bus | 49 |
| 4 | 5778 | 11A | Mpesa | BM97HFRGL9 | 27-11-17 | 7:12 | Migori | Nairobi | bus | 49 |

Description

The below table shows the description of the data (object data) such as count,unique, top, frequency.

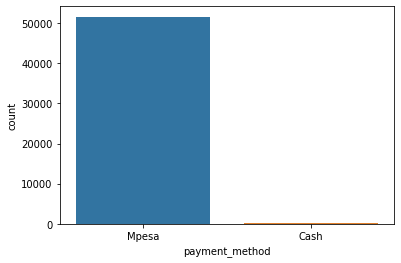
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Seat number | Payment method | Payment receipt | Travel date | Travel time | Travel from | Travel to | Car type |
| **count** | 51645 | 51645 | 51645 | 51645 | 51645 | 51645 | 51645 | 51645 |
| **unique** | 61 | 2 | 51645 | 149 | 78 | 17 | 1 | 2 |
| **top** | 1 | Mpesa | UZUEHCBUSO | 10-12-17 | 7:09 | Kisii | Nairobi | Bus |
| **freq** | 2065 | 51532 | 1 | 856 | 3926 | 22607 | 51645 | 31985 |

**Exploratory Data Analysis**

In this EDA, an analysis is done on a dataset which helps us to visualize the different factors in data which help us to conclude the different insights, predictions of the data,

In this project I have 1st clean the data, make the data as much we can utilize each factor to its maximum limit, then I have done the EDA to obtain the perception from the given data.

1. Values count for payment mode, car type , maximum capacity

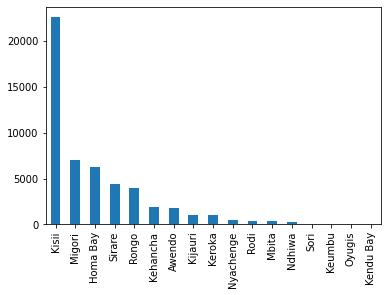




There are two types of payment ,ethods in which the mpesa is mostly used to buy tickets.

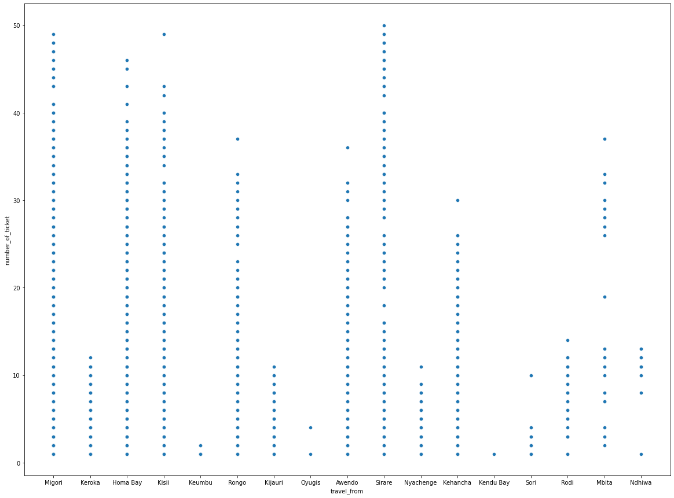
Car type with bus and shuttle having 49 and 11 max capacity

1. Towns from which these routes originate.



Kisii is the place from where most number of rides originates.

Travel from v/s number of ticket



**ML MODEL- AND METRICES**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TYPE OF REGRESSION** | **Train Score** | **Test Score** | **R2 SCORE** | **ADJ\_R2** | **MAE** | **MSE** |
| **LINEAR** | **0.41531** | **0.35462** | **0.35467** | **0.34765** | **4.74747** | **48.435119** |
| **LINEAR-LASSO** | **0.41406** | **0.35476** | **0.35476** | **0.347804** | **4.74177** | **48.42415** |
| **LINEAR-RIDGE** | **0.4302051** | **0.4838009** | **0.35535** | **0.34129** | **4.74177** | **48.42415** |
| **GRADIENT BOOSTING** | **0.67633** | **0.608508** | **0.608508** | **0.60467** | **3.54003** | **29.39045** |
| **RANDOM FOREST** | **0.62769** | **0.62338** | **0.623387** | **0.615172** | **3.375124** | **5.31642** |
| **XGBOOST** | **0.845594** | **0.842112** | **0.842112** | **0.838668** | **2.2667203** | **11.8493008** |

**Challenges faced**

Feature engineering – to get the more required features that will ease the further analysis

What should be the dependent variables

To filter the given data

Feature to be selected to get the required output

**Conclusion**

We used different type of regression algorithms to train our model like, linear regression, regularized linear regression (Ridge and lasso), GBM, Random Forest Regressor, XG Boost regressor, and also we tuned the parameters of Random forest regressor and XG Boost regressor and also found the important features for training the model. Out of them XG Boost with tuned hyperparameters gave the best result.

**THANK YOU !!!!...**