“Nairobi Public Transport Demand Prediction”

By: - Abhishek Kirar, Ganesh Subramanian,
Mayur Chulbhare, Syed Aquib
Data science trainees, Alma Better, Bangalore

★ INTRODUCTION

One of Africa's cities with the worst traffic is Nairobi. Numerous Kenyans go to Nairobi every day for work, business, or to see friends and family. These travelers come from cities like Kisii, Keroka, and other nearby towns. The drive can take a while, and depending on traffic, the last approach into the city may have a big impact on how long it takes. Which bus to take while entering the city and how do traffic patterns affect those decisions? Does understanding Nairobi's traffic patterns assist predict when there may be a need for specific routes? In order to anticipate the amount of tickets that will be sold for buses arriving in Nairobi from "up country" Kenyan cities, a predictive model will be developed utilizing traffic information provided by Uber Movement and historical bus ticket sales information from Mobiticket.

★ OBJECTIVE

- Data 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 goal of the project is to develop a predictive model using the traffic information that has been provided to us and historical bus ticket sales data from Mobiticket to forecast the amount of tickets that will be sold for buses travelling into Nairobi from nearby cities.

★ PROBLEM STATEMENT :-

The 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 route ends in 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 take 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:

- Kawangware: the first stop in the outskirts of Nairobi
- Westlands
- 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 Wrangling:-

The dataset has 51645 observations having 10 columns, and the dataset contains zero 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. 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 where the ride originated
- travel_to: Destination of ride. All rides are to the Nairobi.
- car_type: Vehicle type (Shuttle or Bus)
- max_capacity: Number of seats in the vehicle

Approach:-

In order to achieve better results, we applied the Outliers treatment and normalized the characteristics. To train the dataset to forecast future supply, I employed supervised learning regression analysis models such Linear regression, Lasso regression, and Ridge regression, as well as Gradient boosting regressor, XG boost, and Random forest.

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 table below is showing the description of the data (object data) such as count, unique, top values, and frequency.

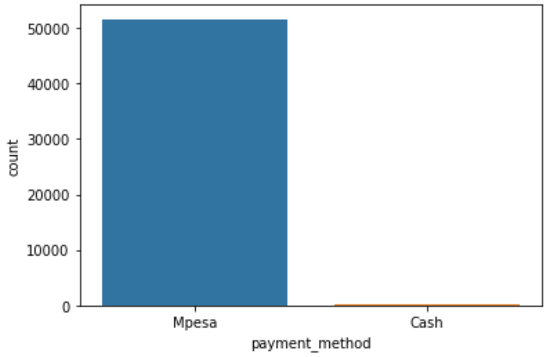
	Seat number	Payment method	Payment receipt	Travel date	Travel time	Travel from	Travel to	Car type
Count	51645	51645	51645	51645	51645	51645	51645	51645
unique	61	2	51645	149	78	17	1	2
Top	1	Mpesa	UZUEHCBUSO	10-12-17	7:09	Kisii	Nairobi	Bus
Freq	2065	51532	1	856	3926	22607	51645	31985

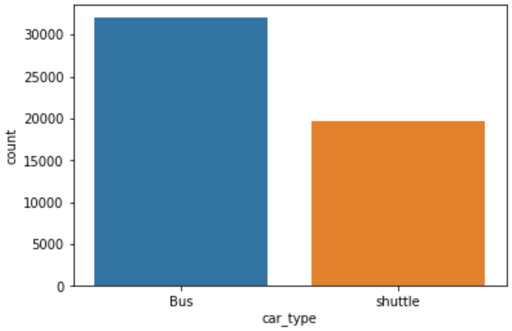
Exploratory Data Analysis

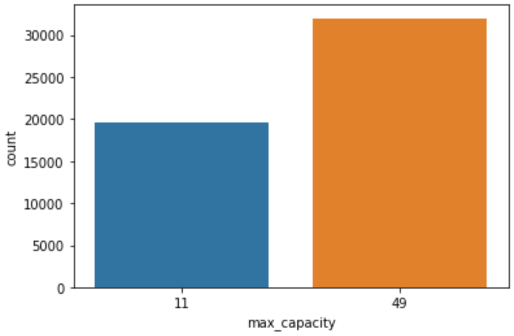
In this EDA, an analysis is done on a dataset which helps us to visualize the different factors in data which help us to conclude the different insights, predictions of the data,

In this project I have clean the data at the very beginning, 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

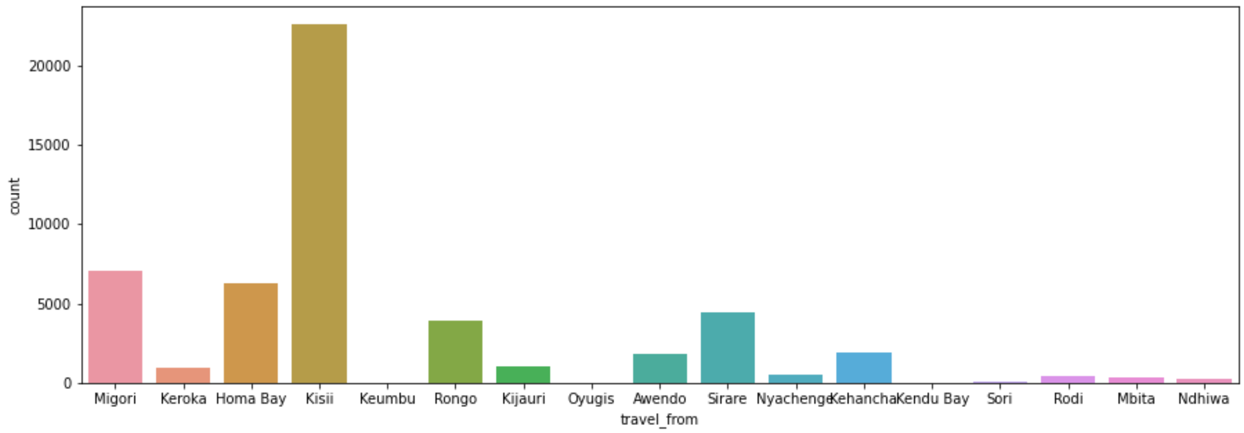






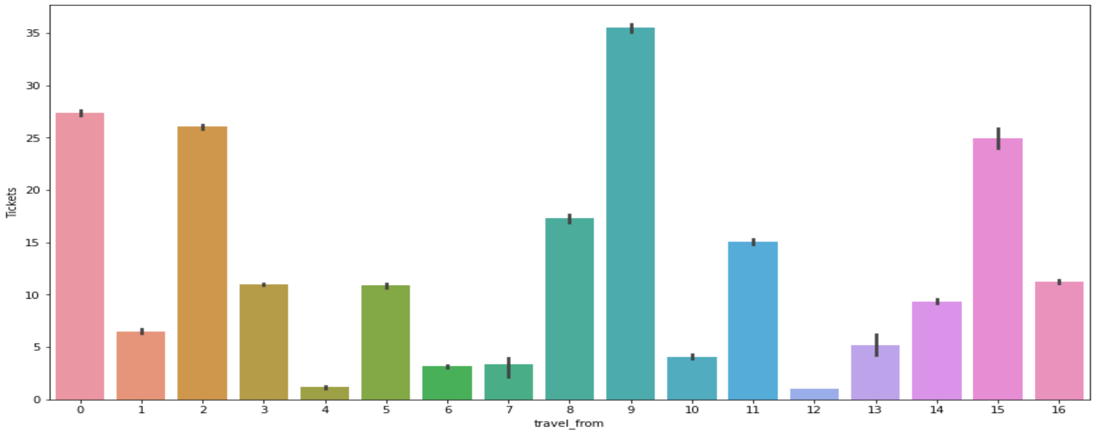
- There are two types of payment methods- Mpesa & Cash, from which the Mpesa is mostly used to buy tickets.
- Car type with bus and shuttle having 49 and 11 max capacity respectively,

2) Towns from which these routes originated



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

3) Travel from v/s Number of ticket



Maximum Tickets were from Sirare(9) area

ML MODEL AND METRICES

TYPES OF REGRESSION	Train Score	Test Score	R2 score	ADJ_R2Score	MAE	MSE
LINEAR REGRESSION	0.3771862	0.3835700	0.383570	0.382793150	--	94.20727
LASSO REGRESSION	0.374013	0.3799886	0.3834121	0.3826350	7.00606	94.23167
RIDGE REGRESSION	-95.6779	-94.21755	0.3802801	0.379807	--	94.217555
XGBOOST	0.970131	0.9642049	0.964204	0.964198	1.52775	5.470452
GridSearchCV	0.918179	0.919089	0.9190896	0.919027	2.46956	12.30100

Challenges Faced:-

The data collection was examined for recognizable statistical trends and patterns in order to make it tenable for comprehension and further analysis. Following preliminary analysis, the following actions were made to turn the data into a systematically usable 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

Conclusion:-

The resulting model can be used by bus operators and Mobiticket to predict customer demand for specific trips, manage resources and vehicles more effectively, offer promotions and sell other services more successfully, such as micro-insurance, or even enhance customer service by being able to send alerts and other helpful information to customers. Our model was trained using a variety of regression algorithms; however XG Boost with customized hyperparameters produced the best results.

References:-

- Transport Demand Data- Nairobi

https://zindi.africa/competitions/traffic-jam-predicting-peoples-movement-into-nairobi/data

- Python Pandas Documentation

https://pandas.pydata.org/pandas-docs/stable

- Python Sklearn Documentation

https://scikit-learn.org/stable/

- Python MatPlotLib Documentation

https://matplotlib.org/stable/index.html

- XGBoost Documentation

https://xgboost.readthedocs.io/en/stable/

THANK YOU!!!!...