Transport Demand Prediction

Jayalaxmi Mekap

Data science trainees,

AlmaBetter, Bangalore

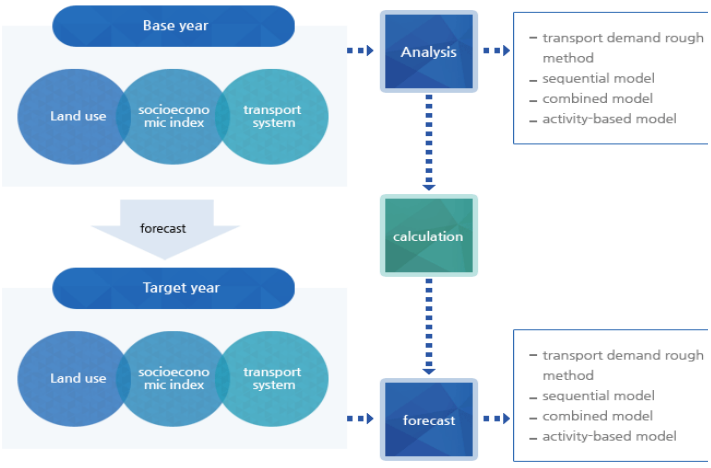
Abstract:

Transport demand forecasting is **to predict future transport demand when establishing transport plans within a given budget**. Since transport demand closely

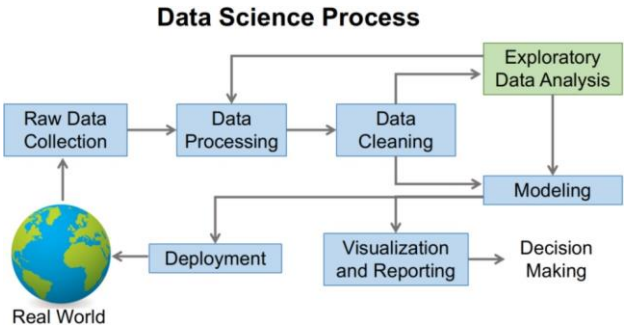
interacts with socioeconomic environment and land use, future socioeconomic indexes and land use patterns need to be estimated first. As Public Transport (PT)

becomes more dynamic and demand-responsive, it increasingly depends on predictions of transport demand. But how accurate need such predictions

be for effective PT operation



- loading the data into data frame
- cleaning the data
- extracting statistics from the dataset
- exploratory analysis and visualizations
- questions that can be asked from the dataset
- conclusion



1.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 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 Kisii Mbita Migori Ndhiwa Nyachenge Oyugis Rodi Rongo Sirare Sori

The routes from these 14 origins to the first stop in the outskirts of Nairobi takes approximately 8 to 9 hours from 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 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.

2. Introduction

train_revised.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.

test_questions.csv is the dataset on which you will apply your model to estimate number of tickets sold by Mobiticket per unique ride. This dataset contains all of the rides offered

on the same 14 routes during the two weeks following train.csv, i.e. 21 April 2018 to 9 May 2018. The variables included in this dataset: ride_id, travel_date, travel_time,

travel_from, travel_to, car_type, max_capacity.

sample_submission.csv is a table to provide an example of what your submission file should look like. This table has two columns: ride_id, number_of_ticket.

Uber Movement traffic data can be accessed at movement.uber.com. Data is available for Nairobi through June 2018. (If the data for April-June are not up yet, they will be shortly.)

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.

- **Seat number**: seat assigned to ticket
- **Rider id**: unique ID of a vehicle on a specific route on a specific day and time
- **Payment method**: method used by customer to purchase ticket from Mobiticket (cash or Mpesa)
- **Travel date**: date of ride departure. (MM/ DD/YYYY)
- **Travel time**: time of the ride
- **Travel from**: from where rider travel to Nairobi
- **Day of the week**: in which day of the week he/she travelling
- **Day of the month**: ride date of the month
- **Day of year**: which date of the year he/she is travelling
- **Is weekend or not**: riding day is occur in weekend or not
- **Quarter**: quarter of the year
- **Period**: am or pm
- **Number of tickets:**

3. Data Cleanings and validations

In this step removing faulty data and filling in gaps. The task to be crucial and important thus validating by following steps

- Removing extraneous data
- Handing in missing values.
- Data shifting in respective columns
- Conforming data to a standardized pattern.

```
[ ] dataset.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51645 entries, 0 to 51644
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   ride_id         51645 non-null  int64
 1   seat_number     51645 non-null  object
 2   payment_method  51645 non-null  object
 3   payment_receipt 51645 non-null  object
 4   travel_date     51645 non-null  object
 5   travel_time     51645 non-null  object
 6   travel_from     51645 non-null  object
 7   travel_to       51645 non-null  object
 8   car_type        51645 non-null  object
 9   max_capacity    51645 non-null  int64
dtypes: int64(2), object(8)
memory usage: 3.9+ MB
```

```
dataset.describe(include='object')
```

	seat_number	payment_method	payment_receipt	travel_date	travel_time	travel_from	trave
count	51645	51645	51645	51645	51645	51645	5
unique	61	2	51645	149	78	17	
top	1	Mpesa	UZUEHCBUSO	10-12-17	7:09	Kisii	N
freq	2065	51532	1	856	3926	22607	5

- There are total 61 unique seats in this dataset
- travelers have used 2 types of payment method and most of the people have used Mpesa to pay for their ticket.

- The record of 149 days out of 2 year is present in this dataset.
- There are 2 different types of car and most of them are bus.

4.DATA

```
[ ] import pandas as pd
    import numpy as np
    import seaborn as sns
    import matplotlib.pyplot as plt
    %matplotlib inline
```

```
dataset = pd.read_csv('/content/drive/MyDrive/SUPERV
dataset.head()
```

	ride_id	seat_number	payment_method	payment_receipt	t
0	1442	15A	Mpesa	UZUEHCBUSO	
1	5437	14A	Mpesa	TIHLBUSGTE	
2	5710	8B	Mpesa	EQX8Q5G19O	
3	5777	19A	Mpesa	SGP18CL0ME	
4	5778	11A	Mpesa	BM97HFRGL9	

```
[ ] dataset.tail()
```

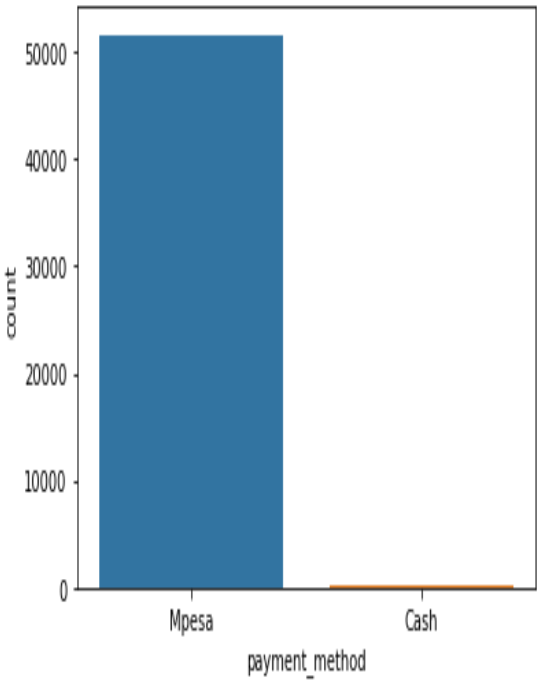
	ride_id	seat_number	payment_method	payment_receip
51640	13826	9B	Mpesa	8V2XDDZR6
51641	13809	18A	Mpesa	4PEBSVJSN
51642	13809	17A	Mpesa	LVN64LZDN
51643	13796	16B	Mpesa	REYBSKTYW
51644	14304	7	Mpesa	AQN7FBUSG

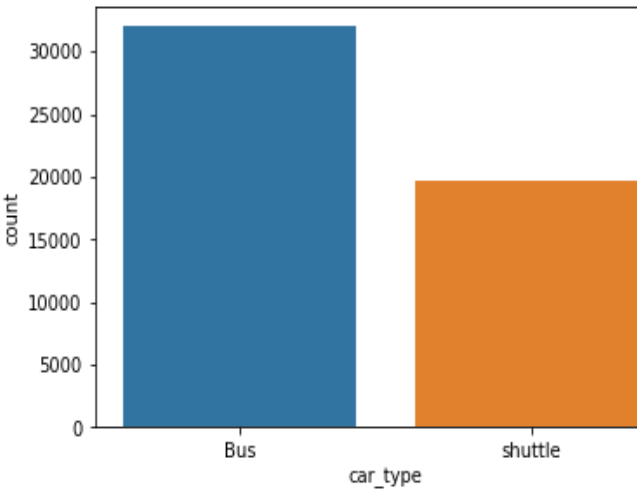
4. Exploratory Analysis and Visualization

Exploratory data visualizations are the type of visualizations we assemble when we do not have a clue about what information lies within our dataset.

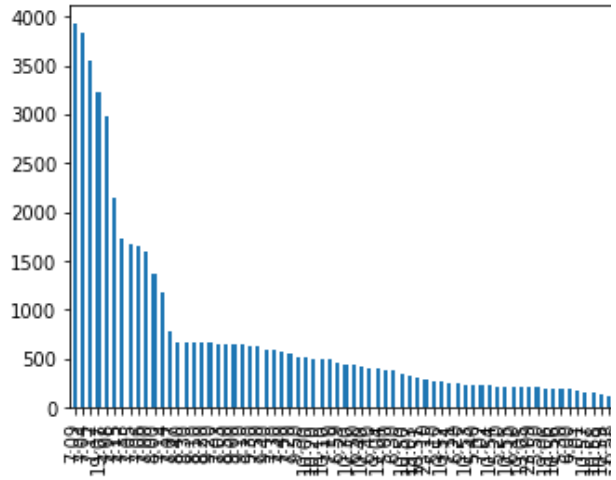
. maximum rider used mpesa method for their payment

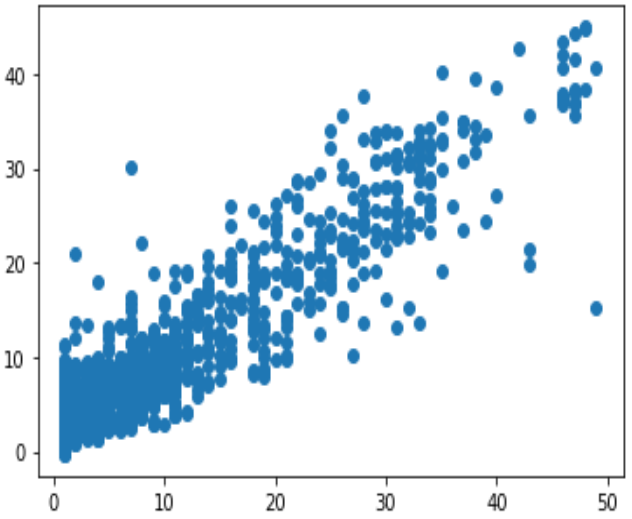
. approximately 35000 travelled by bus

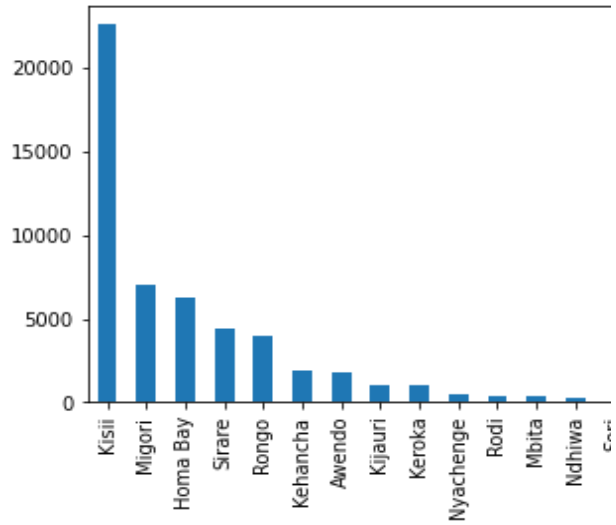


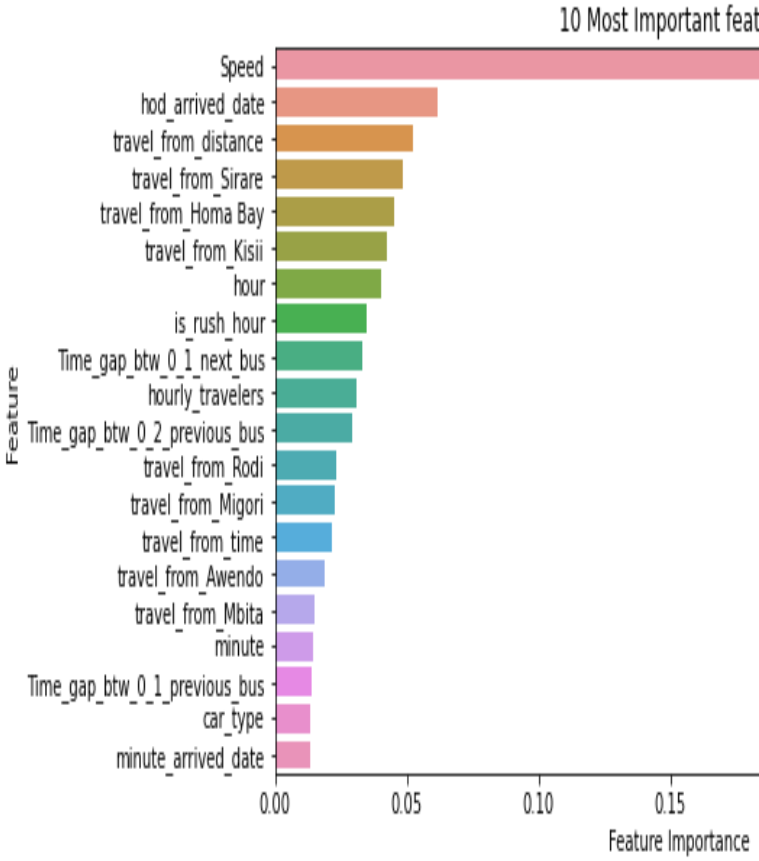


. majority of the rider travel from kisii and lowest number of rider travelled from kendu bay

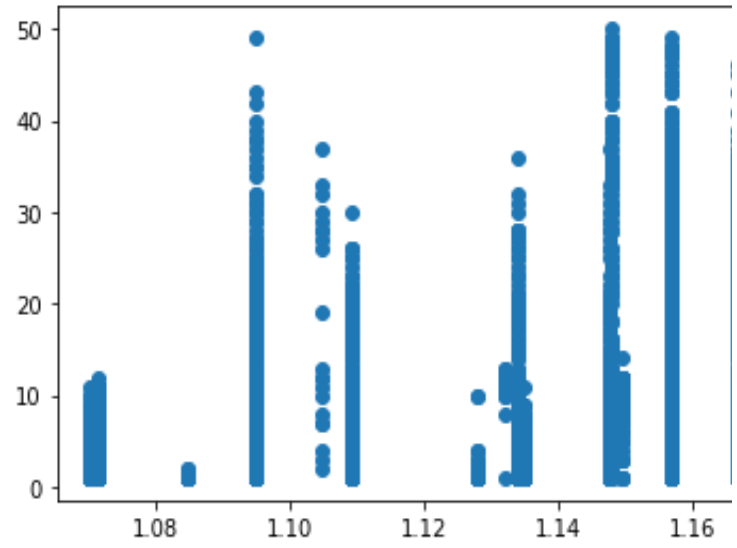








. this plot shows speed vs number of tickets this data concludes if the speed is high the number of tickets also sold higher vise versa



. Above plot shows some important features and their frequency

6.Contributions

- Introduction to Data
- Data cleaning
- Correlation
- Understanding the data
- Payment method across each rider
- Car or bus type across each rider
- Max capacity in car and bus
- Max rider travel from
- Rider spending time in bus or car
- Number of tickets Vs rider
- Travel from vs number of tickets
- Day of the month Vs number of tickets
- Number of tickets vs hours
- Merging of data frames on apps
- Speed vs number of tickets
- Encodings categorical features
- Training the model
- Linear regression
- Implementing lasso regression
- Ridge regression
- Training gradient boosting regressor
- Training XGboost
- Random forest regressor
- Grid search cv on random forest
- Grid search cv on XGboost algorithm

7. Programing Language:

We have used python programing Language and used below library for EDA

Numpy

Pandas

seaborn

WordCloud

matplotlib

warnings

8. Conclusion:

We used diffent type of regression algorithms to train our model like, Linear Regression, Regularized linear regression (Ridge and Lasso), GBM,Random Fores Regressor,

XGboost regresssor. and Also we tuned the parameters of Random forest regressor and XGboost regressor and also found the important features for training the model.

Out of them XGboost with tuned hyperparameters gave the best result.

- There are two type of payment methods people have used to buy the tickets.
- There are two type of cars Bus and shuttle and the maximum capacity of the bus is 49 while shuttle can contain 11 travelers.
- There might be many ways of finding the target variable but here I am using one way that is I will find the count of each ride_id and that will be the number_of_ticket as our target variable.
- We can see that most of the ticktes were sold at 7 AM and 8 PM. And that seems true because in the morning most of the people go to the work and office.
- From the above we can say that there is not ride between 12pm to 5.30Pm
- 99% rider used mpesa payment method for their ride.
- 60% people used bus and 40% people used shuttle.
- Maximum rider travel from Kisii.