Capstone Project - 2
Bus Tickets Sale Prediction

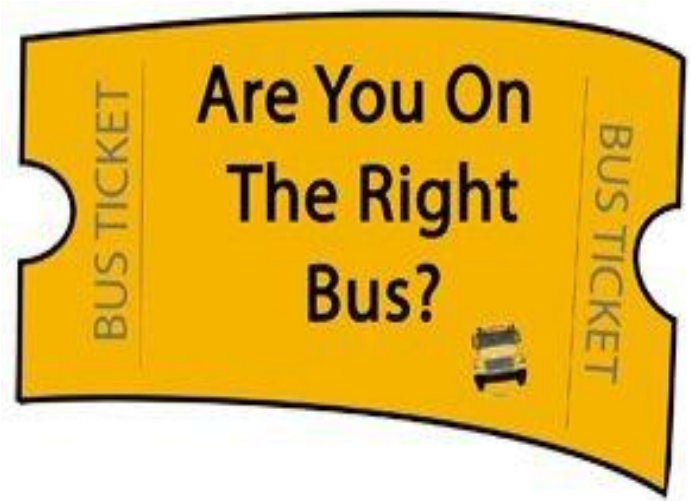
Suraj Pandey

Content

- Problem Statement
- Data Summary
- Ride Origination Towns
- Travel time
- Quarterly Trend
- Month wise booking trends
- Feature Engineering
- ML Models and Metrics
- Challenges
- Conclusion
- Q & A

Problem Statement

Exploring 14 different towns to the North-West of Nairobi towards Lake Victoria and using the data provided by bus ticket sales from Mobiticket, predicting the number of tickets that will be sold for buses that ends into Nairobi.

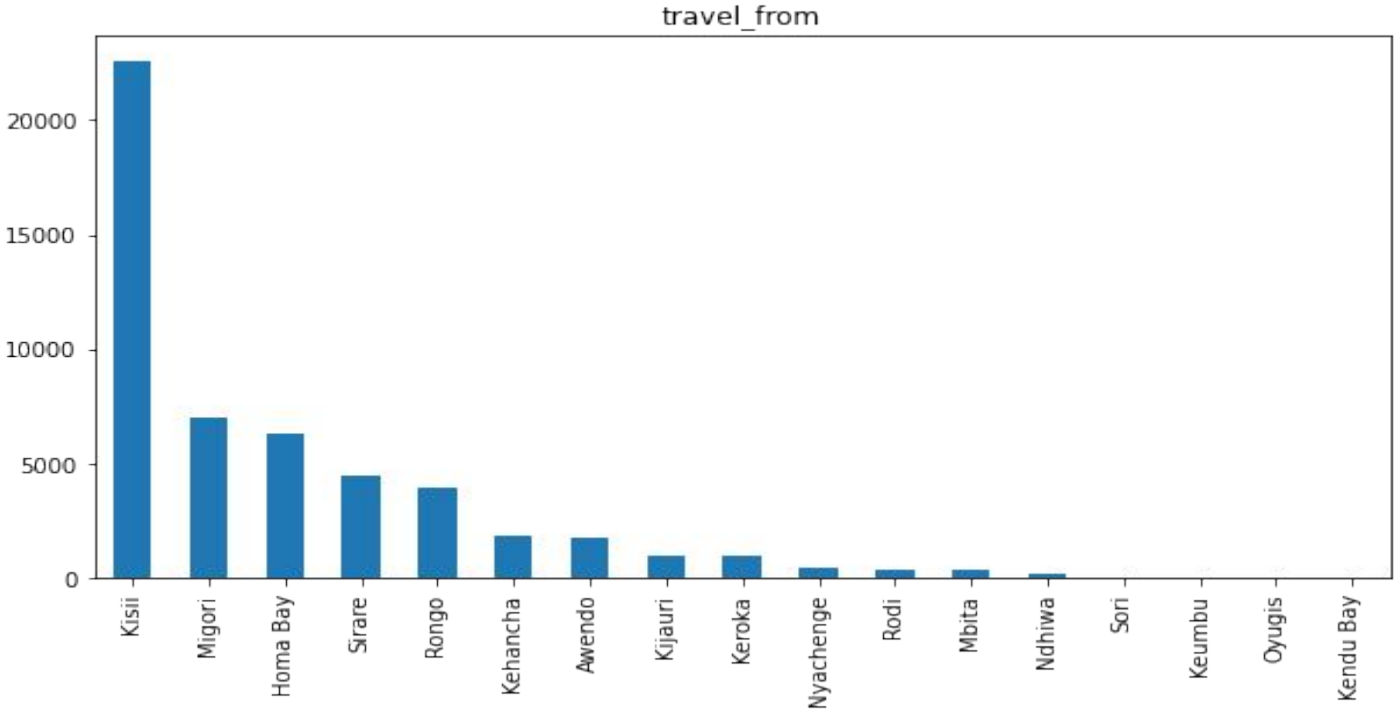


Data Summary

This dataset includes the variables from 17 October 2017 to 20 April 2018

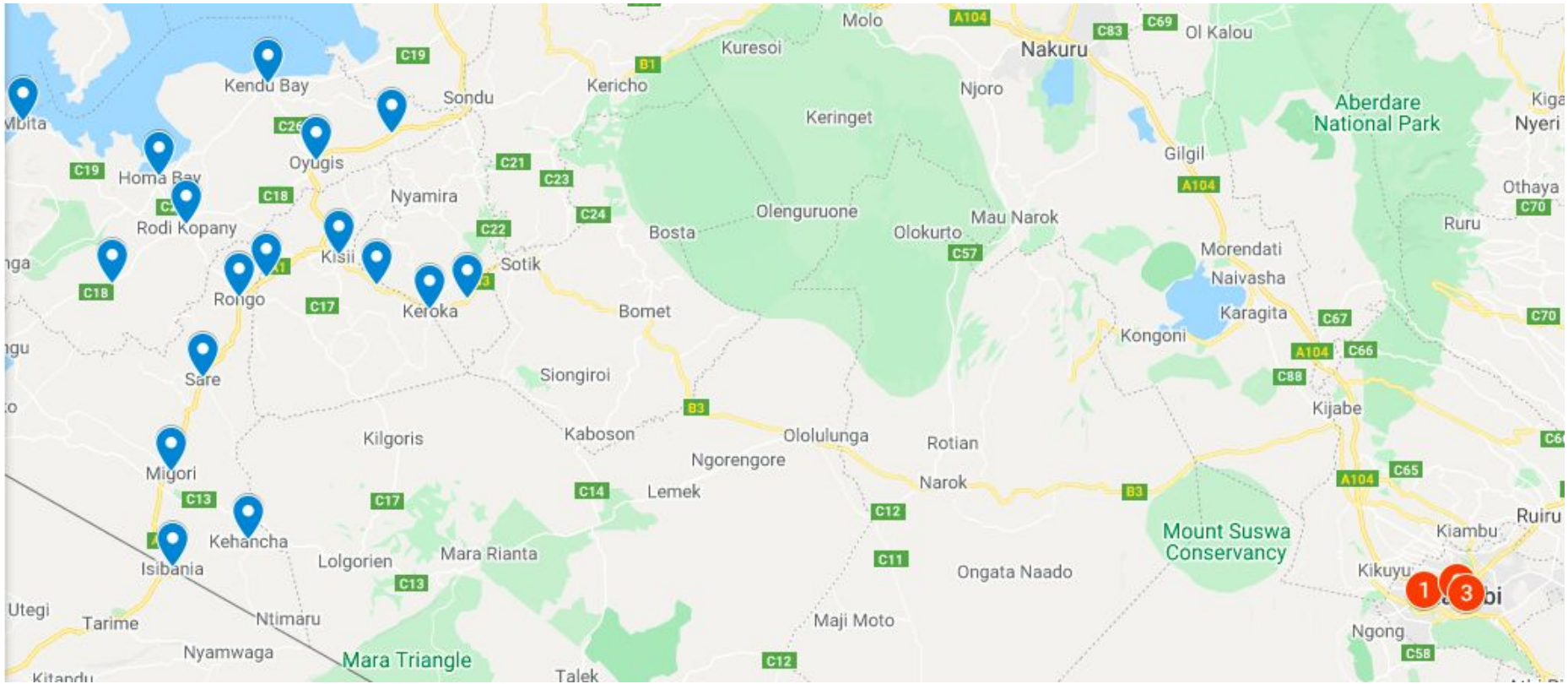
- **ride_id**: unique ID of a vehicle on a specific route on a specific day and time.
- **seat_number:** seat assigned to ticket
- **payment_method:** method used by customer to purchase ticket from Mobiticket
- **payment_receipt:** unique id number for ticket purchased from Mobiticket
- **travel_date:** 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

Ride Origination Towns

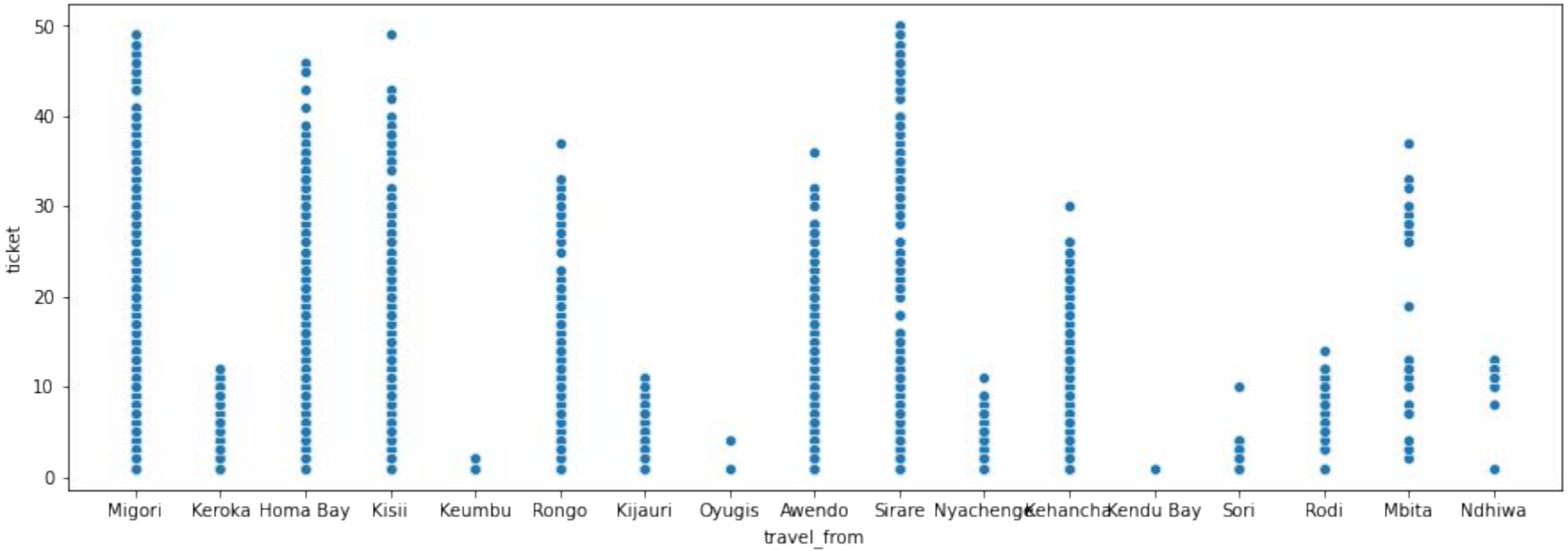


Kisii is the top place from where the most number of rides originate.

Map

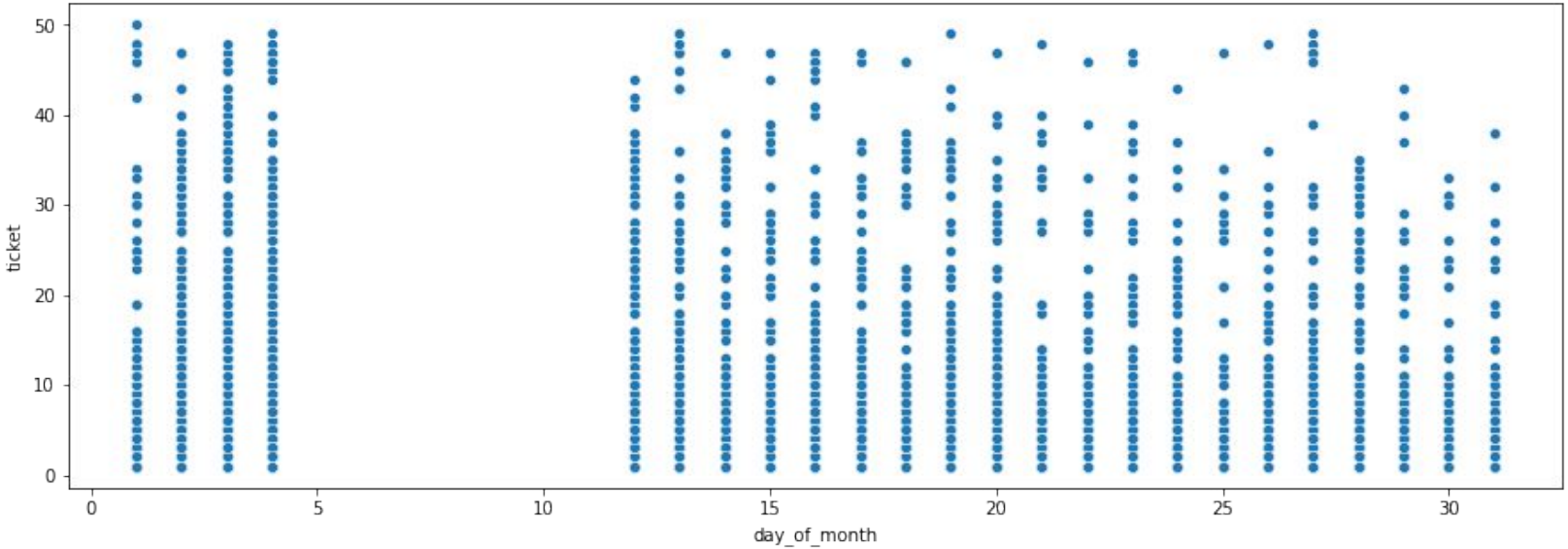

Kendu Bay
Homa Bay
Oyugis
Rodi Kopany
Nyamira
Kisii
Sondu
Kericho
Kuresoi
Molo
Nakuru
Njoro
Keringet
Ol Kalou
Aberdare National Park
Gilgil
Nyeri
Othaya
Ruru
Olenguruone
Mau Narok
Olokurto
Bosta
Sotik
Morendati
Naivasha
Karagita
Kongoni
Rongo
Keroka
Bomet
Sare
Siongiroi
Kijabe
Kilgoris
Kaboson
Ololulunga
Rotian
Ngorengore
Migori
Lemek
Narok
Kehancha
Isibania
Lolgorien
Mara Rianta
Mount Suswa Conservancy
Kiambu
Ruiru
Kikuyu
Ongata Naado
Utegi
Tarime
Ntimaru
Maji Moto
Ngong
Nyamwaga
Mara Triangle
Talek
1
3

EDA



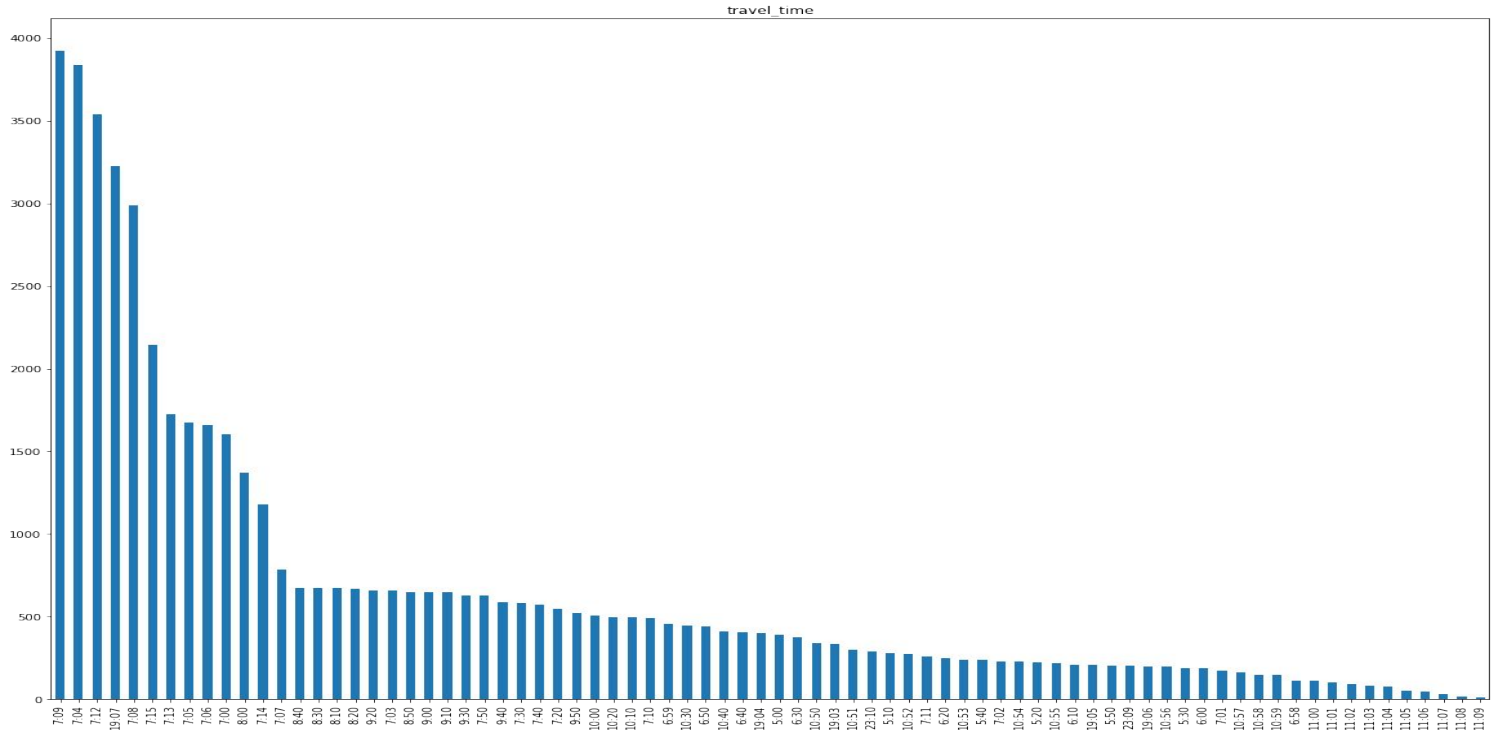
- **Scatter plot of travel_from by number of tickets**

Day wise Travel Trend



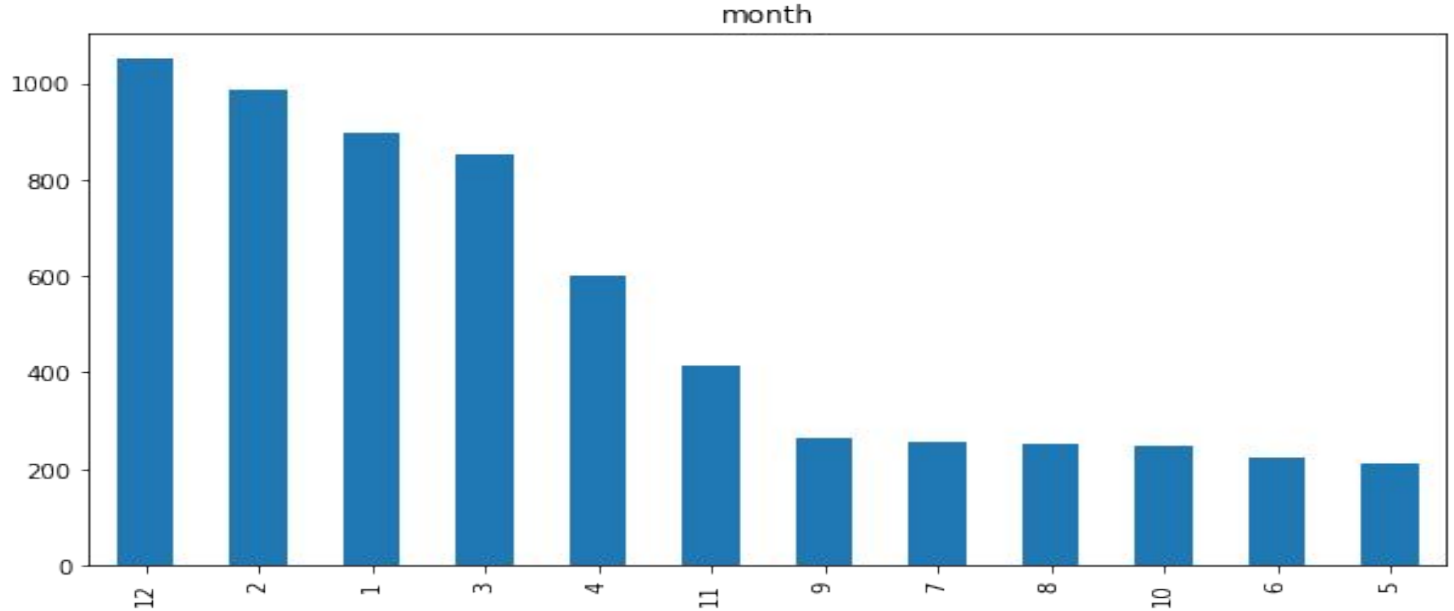
The density of the rides are almost similar among the days of the month, There are no rides between 5th to 10th of every month,but this might be because of missing data

Departure Time



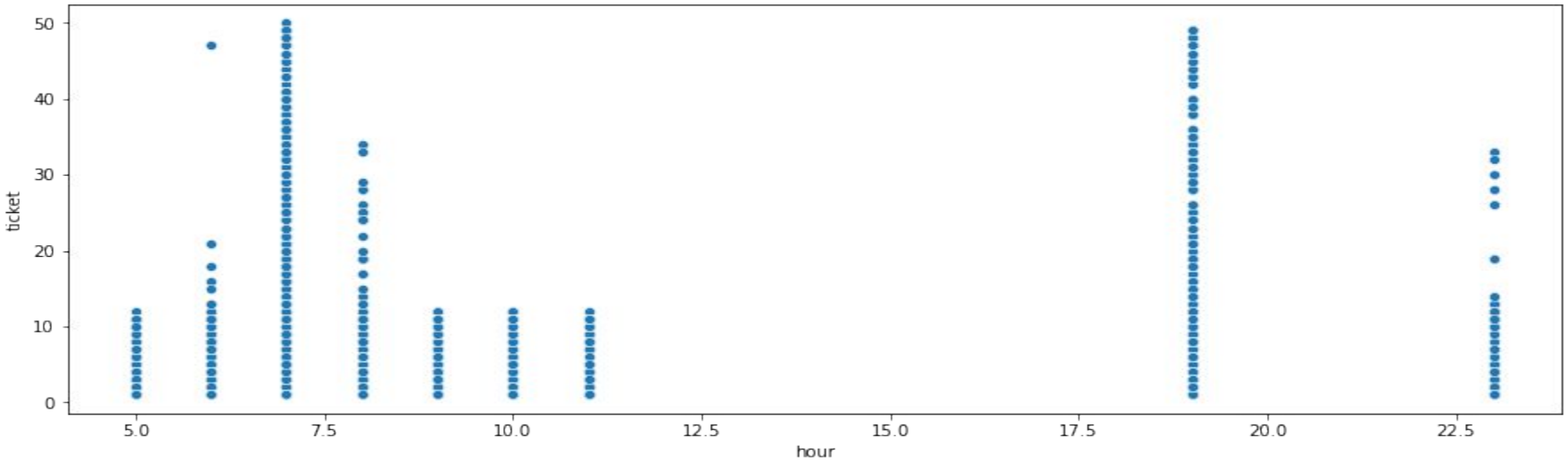
Highest number of buses depart at around 7 AM in the Morning

Month-wise Rides Trends



During the month of December,February and January there are more number of rides, and least during the months of May and June

Hourly Travel Trend

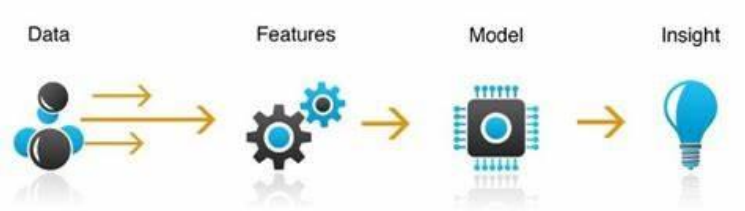


The frequency of rides are more in the Morning hours and during the night times

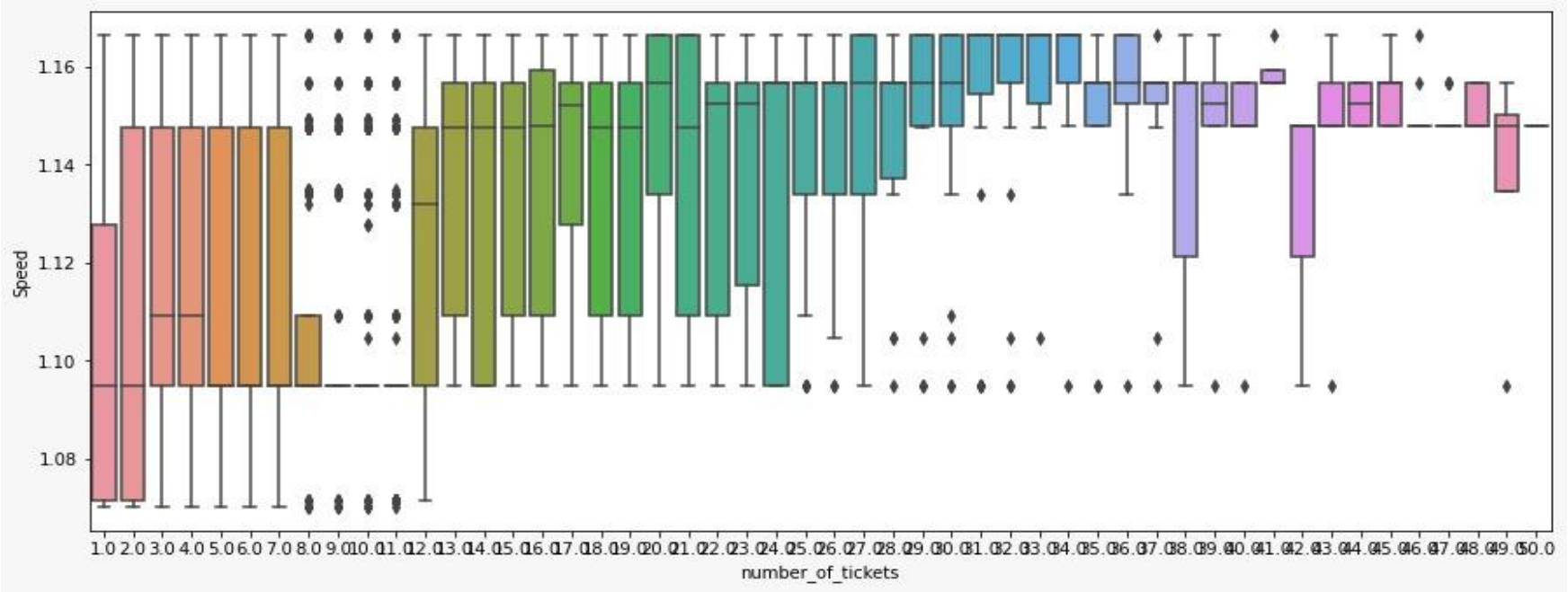
Feature Engineering

Using domain knowledge to extract features from raw data, the performance of the model can be improved.

- Speed
- Travel_month
- No_of_tickets
- travel_day
- hod_arrived_date
- Is_rush_hour
- Travel_from
- Time_gap_between_buses
- Travel_from_distance
- hourly_travelers
- daily_travelers



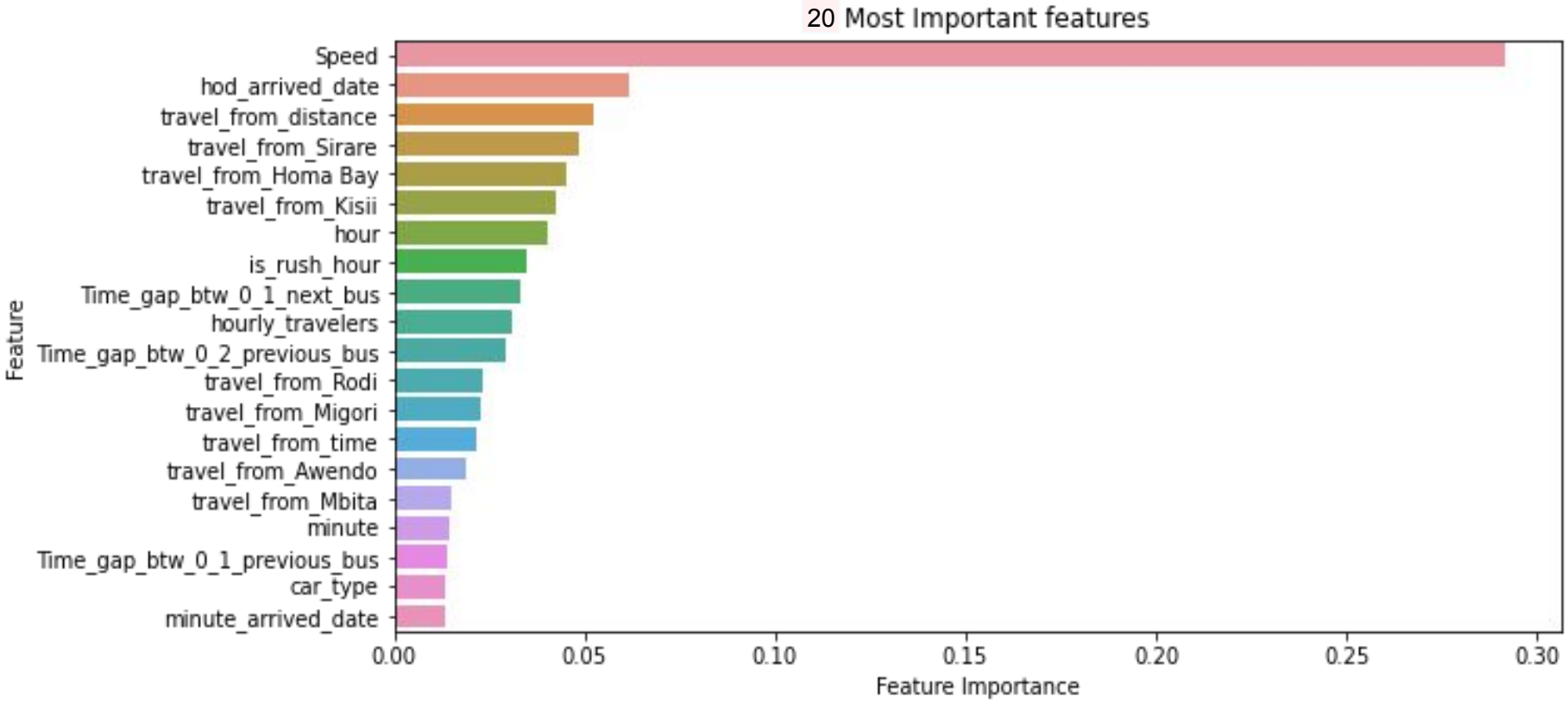
Variation of Number of Tickets with Speed



ML Models and Metrics

TYPE OF REGRESSION	Train Score	Test Score	R2 SCORE	ADJ_R2	MAE	MSE
LINEAR	0.41531	0.354621	0.354679831	0.3476561	4.7474791	48.4351195
LINEAR-LASSO	0.293599	0.343606	0.355067	0.3487478	4.7417715	48.4241544
LINEAR-RIDGE	0.405354	0.3553535	0.3550673	0.3481087	5.026478	48.4015719
GRADIENT BOOSTING	0.676331137	0.60851	0.6085084	0.6046721	3.540035	29.3904512
RANDOM FOREST	0.62637829	0.623421	0.6234206	0.6152057	3.4301030	28.2619184
XGBOOST	0.84559453	0.84211254	0.84211254	0.8386682	2.2667203	11.8493008

Feature Importance



Challenges

- To find the dependent variable
- Feature engineering
- Feature selection
- Model Training and performance improvement.

Conclusion

This resulting model can be used by Mobiticket and bus operators to anticipate for the tickets for certain rides. We have compared the performance of six different regression models. XGBoost regression model performed the best among them including the ensemble model proposed with the lowest error rate. We pre-processed data to apply regression models for forecasting the speed of vehicles and distance between the source and destination.

Q & A