Capstone Project - 2

Demand Prediction for Public Transport

Pranay Umredkar

Content

- Problem Statement
- Data Summary
- Ride Origination Towns
- Month Wise trends
- Day wise Travel Trend
- Travel time wise trends
- Feature Engineering
- Encoding Categorical variables
- Model Training
- ML Models and Metrics
- Challenges
- Conclusion

Problem Statement

To analyze data provided by Mobiticket regarding bus ticket sales of 14 routes end in Nairobi and originate in towns to the North-West of Nairobi towards Lake Victoria.

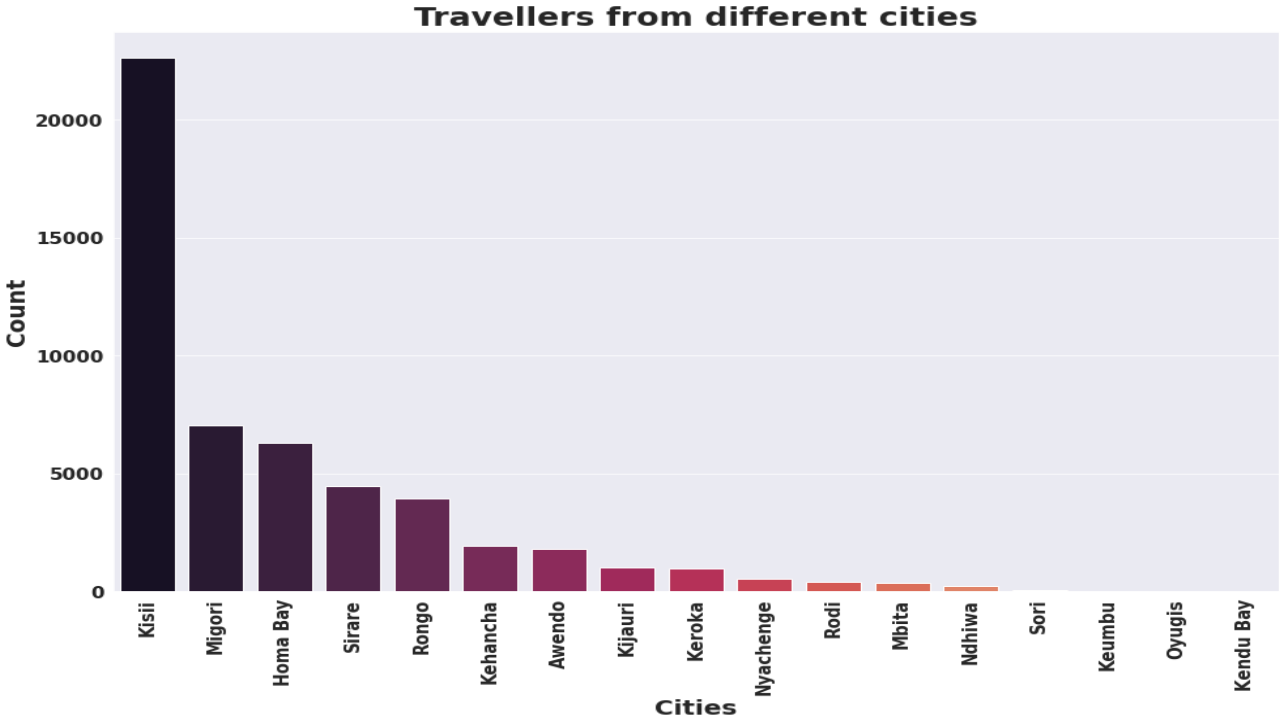
To build a model that predicts the number of seats that Mobiticket can expect to sell for each ride, i.e. for a specific route on a specific date and time

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

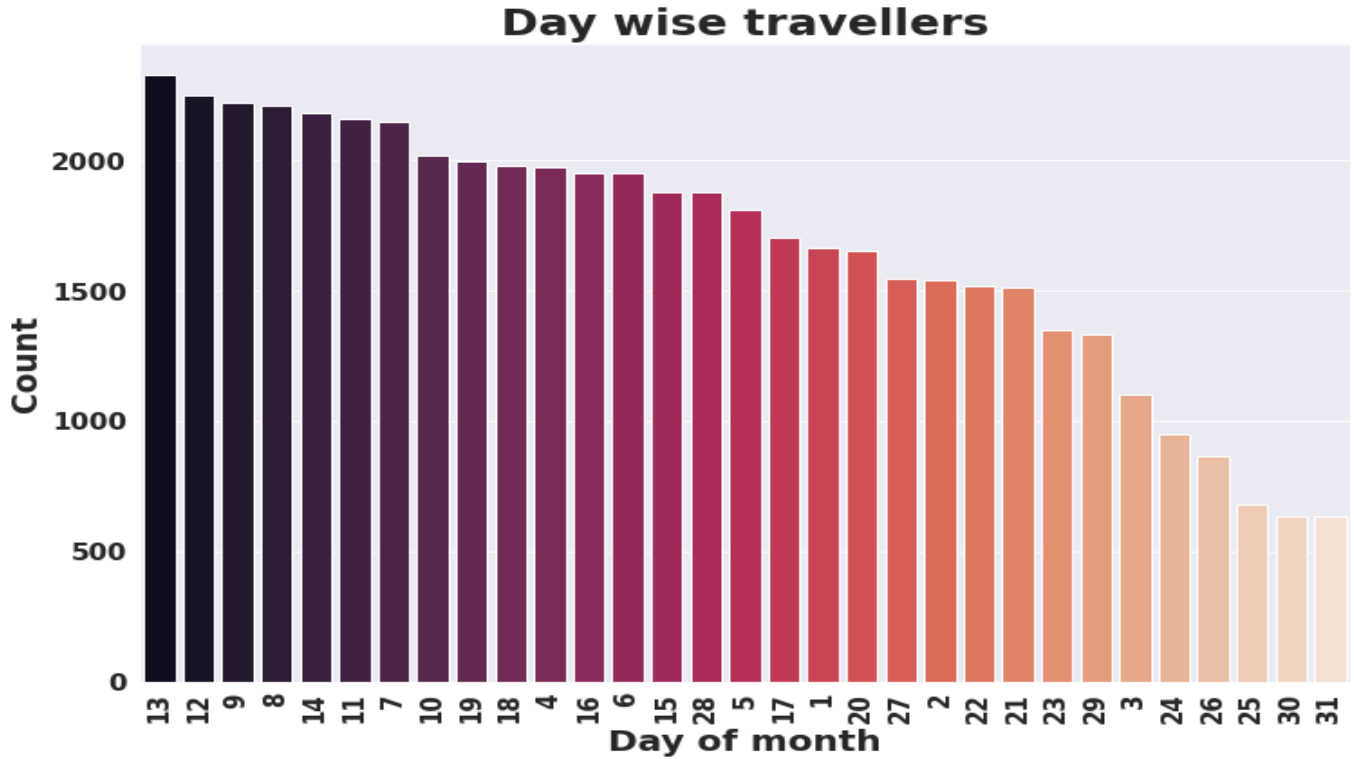


Kisii is the top place from where the most number of rides originate.

Month wise Rides Trends

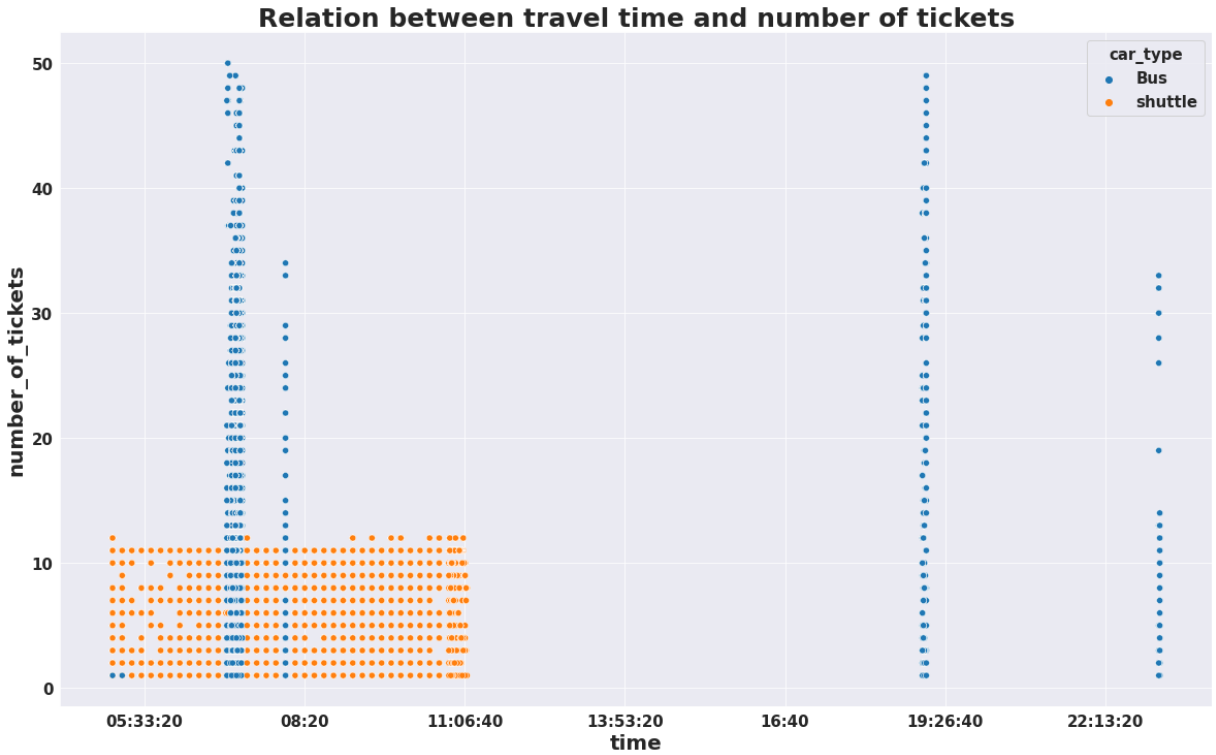
During the month of December there are more number of rides, because of festive season and people prefer to travel in summer season between December to march.

Day wise Travel Trend



The frequency of the rides are almost similar among the days of the month

Travel Time wise Trends



Mostly people travels from early morning to morning time and evening in afternoon there are no tickets booking.

Feature Engineering

Feature Engineering is a machine learning technique that leverages data to create new variable that aren't in the training set . We produce new features with the goal of simplifying and speeding up data transformation while also enhancing model accuracy.

- Travel month
- No of tickets
- Travel Hours/Minutes
- Day of month
- Month
- Weekday
- Is weekend
- Time slot

Encoding categorical variables

In simple words encoding means converting data into required format. Since ML models takes only numerical data to do computation we will convert all cat variable into numerical data.
We used two methods to encode data.

- **Label Encoding:** Label Encoding refers to converting labels to numeric form.
- **One Hot Encoding:** It is also the process of converting categorical data into numeri data but here we don't give labels to each category instead we create new columns for each category and gives binary values

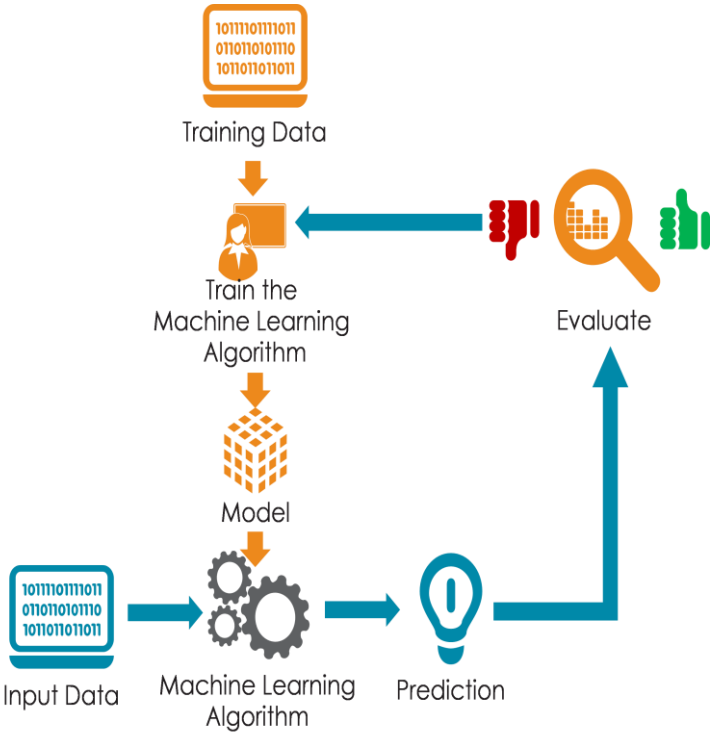
Type
AA
AB
CD
AA

Onehot encoding →

Type	AA_Onehot	AB_Onehot	CD_Onehot
AA	1	0	0
AB	0	1	0
CD	0	0	1
AA	0	0	0

Model Training

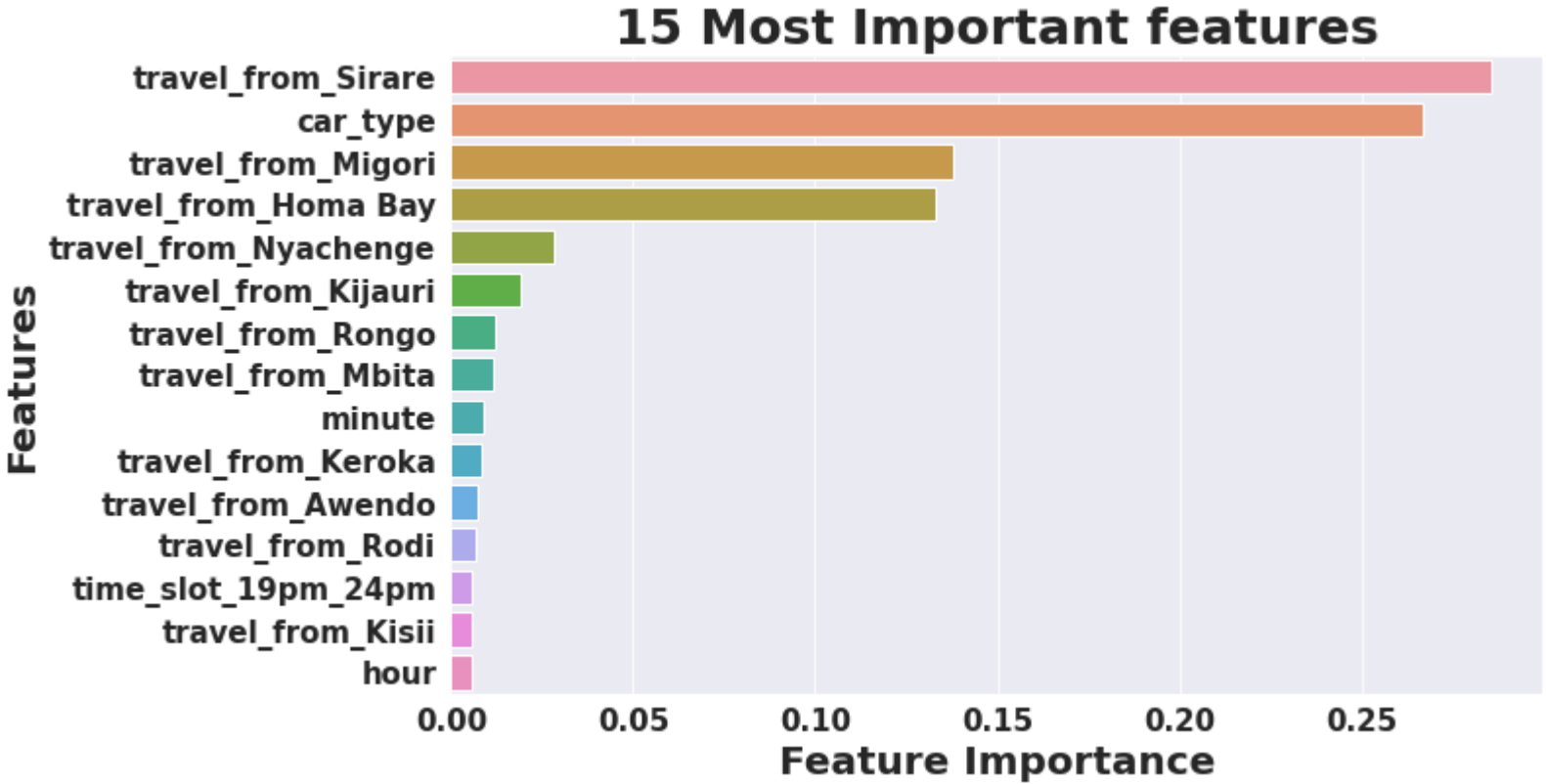
- Model training is the process of fitting a data into machine learning model from which model learns the patterns in data to predict the dependent variable. Model do it so by assigning a weight to each variable.
- After our model is trained we test our model on test data to check how our model is performing.



ML Models and Metrics

TYPE OF REGRESSION	Train Score	Test Score	R2 SCORE	ADJ_R2	MSE	RMSE
LINEAR	0.5971	0.5885	0.5885	0.5868	63.75	7.98
LINEAR-LASSO	0.5971	0.5885	0.5885	0.5868	63.75	7.98
LINEAR-RIDGE	0.5971	0.5885	0.5885	0.5868	63.75	7.98
GRADIENT BOOSTING	0.7113	0.7036	0.7036	0.7024	45.92	6.77
RANDOM FOREST	0.7651	0.7597	0.7597	0.7587	37.22	6.10
XGBOOST	0.9393	0.9313	0.9313	0.9310	10.64	3.26

Feature Importance



Challenges

- To find the dependent variable
- Feature Engineering
- Selecting a features to train a model.
- Model training, tuning and performance improvement.

Conclusion

We used different types of model to train and test performances and compared their performances like Linear regression, Regularized linear regression(Ridge and Lasso), Gradient Boosting Regressor, Random Forest Regressor and XGBoost regressor. XGBoost model performs best among all after hyperparameter tuning. This resultant model can be used by Mobiticket and travel operators to anticipate for the tickets of certain rides.

Thank You