Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

NAMES	E-MAIL	CONTRIBUTION
Charan	charanbzact9865@gmail.com	All project

Please paste the GitHub Repo link.

GitHub link:- https://github.com/Charanpoojary/TRANSPORT-AND-DEMAND-PREDICTION

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

This Project summary is entitled as "**Demand Prediction for Public Transport".** Demand forecasting is a key component of demand management that can be used to support demand planning and decision-making.

Nairobi is one of the most heavily congested cities in Africa. Each day thousands of Kenyans make the trip into Nairobi from towns such as Kisii, Keroka, and beyond for work, business, or to visit friends and family. The journey can be long, and the final approach into the city can impact the length of the trip significantly depending on traffic. How do traffic patterns influence people's decisions to come into the city by bus and which bus to take?

It is the dataset of tickets purchased from Mobi ticket 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. This resulting model can be used by Mobi ticket and bus operators to anticipate customer demand for certain rides, to manage resources and vehicles more efficiently, to offer promotions and sell other services more effectively or even improve customer service by being able to send alerts and other useful information to customers.

The aim is to create a predictive model using traffic data provided from Uber Movement and historic bus ticket sales data from Mobi ticket to predict the number of tickets that will be sold for buses into Nairobi from cities in "up country" Kenya.

Here we have used libraries like Pandas and NumPy to analyze the data, and Matplotlib, Seaborn for visualizing the data. and the supervised regression algorithms to train and predict the model.

Firstly, the data is loaded using the pandas library, then we use shape () and describe () methods for the basic understanding of the database. Before jumping to the exploration stage, we perform some basic data pre-processing methods like null value imputation and removal of unwanted or constant values columns and drop the unwanted columns.

The analysis of variables provided a beautiful pictorial representation of our observations which include bar plots, line charts, histograms and heatmaps and maps.

By analysis we came to know that, Travelers have used 2 types of payment method. They are Mpesa and cash. And most of the people have used Mpesa to pay for their ticket. There are totally 61 unique seats in this dataset. The record of 149 unique days is present in this dataset out of 2 years. There are 2 different types of cars used (shuttle and bus) and most of them are buses.

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. Most customers travel from Kisii town to Nairobi. Most of the tickets were sold at 7 AM and 8 PM. And that seems true because in the morning most of the people go to work and office. From the above we can say that there is no ride between 12pm to 5.30Pm.

We used different types of regression algorithms to train our model like Linear Regression, Regularized linear regression (Ridge and Lasso), GBM, Random Forest Regressor, XGboost regressor. And also, we tuned the parameters of Random Forest regressor and XGboost regressor and also found the important features for training the model.

After the model is built, if we see that the difference in the values of the predicted and actual data is not much, it is considered to be a good model and can be used to make future predictions.

Few metric tools we used to calculate error in the model they are:

MSE (Mean Squared Error), RMSE (Root Mean Squared Error),. MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), R2(R – Squared) and Adjusted R2

As we used different types of regression algorithms to train our dataset, out of them XGboost with tuned hyperparameters gave the best result.

So, the accuracy of our best model is 79% which can be said to be good for this large dataset. This performance could be due to various reasons like no proper pattern of data, too much data, and not enough relevant features.