ML Project

By

Kushal(IMT2021035)

Raghunadh(IMT2021042)

Chokshi(IMT2021012)

Introduction

The data you have pertains to whether a bus arrives on schedule or experiences a delay. Your task is to predict the number "0" when the bus is on time and "1" when it is running behind schedule. Features in the given data set:

1. **Index**-index
2. **Bus_ID**- The identification or number assigned to each bus currently in service.
3. **DepartureTime**- Departure time of the bus in minutes.
4. **Journey_Time**- Total time taken in minutes for the bus to reach its destination.
5. **Bus_Operator**- The company that owns/operates a given bus.
6. **Departure_Bus_Stop**- Location where bus starts journey
7. **Arrival_Bus_Stop**- This refers to the bus stop where the bus is expected to arrive.
8. **Day**- which day it is.
9. **Target**- 0 if the bus is on time to its destination. 1 if the bus is delayed.

Preprocessing and EDA

Removing Duplicate Rows

We checked the number of duplicate rows in the dataset and removed them.

Dealing with null values

We have to remove null values because they are redundant and are not needed.

There are null values in 2 columns of the dataset. They are **Journey_Time** and **Day.**

For **Journey_Time** column,

We observed that journey times of a bus with same Bus_ID are nearly same. We grouped data based on Bus_ID and found the mean of each group. We replaced null values of that column by mean of that group. Even in test_data, we did the same to replace null values of this column.

For **Day** column,

We observed that most buses having same Bus_ID have one a particular day's frequency more than others. We grouped data based on Bus_ID and replaced null values with mode of that data.

Encoding

We use encoding to change any categorical columns to numerical data as many machine learning algorithms run on numerical data only.

Label encoding

We concatenated training and test data, performed label encoding and then separated them. We did label encoding to all columns having string values.

The following columns required label encoding:

- Arrival_Bus_Stop
- Departure_Bus_Stop

One hot encoding

We did one hot encoding to the column **Day** because there are only 7 unique values and **Bus_Operator** column as it has only 16 unique rows.

Models used

We used the following models to train data and predict on test data. The accuracy scores of each model provided insights into how well the models have classified data.

- **Logistic regression:**

We did logistic regresssion and got an accuracy of 54.32%. This suggested us to use a complex model when compared to it.

- **KNN:**

We did KNN to the dataset and got an accuracy of 56.45% when k is equal to 6. We are not able to increase accuracy beyond that. So, we thought of trying a little more complex model and used boosting techniques.

- **Random Forest and Grid Search:**

We used random forest classification and grid search. The accuracy we obtained is 63.78%. So, we decided to try boosting techniques as we got less accuracy.

- **XG Boost with Grid Search:**

At last, we did XG Boost and achieved an accuracy of 66.78%. We also used grid search and did hyperparameter tuning to get this result.

We got the best result with XG Boost with grid search method after doing hyper-parameter tuning.

Making prediction on test data

So, we trained data using XG Boost and predicted values on test data.