REPORTING AND VISUALIZATION

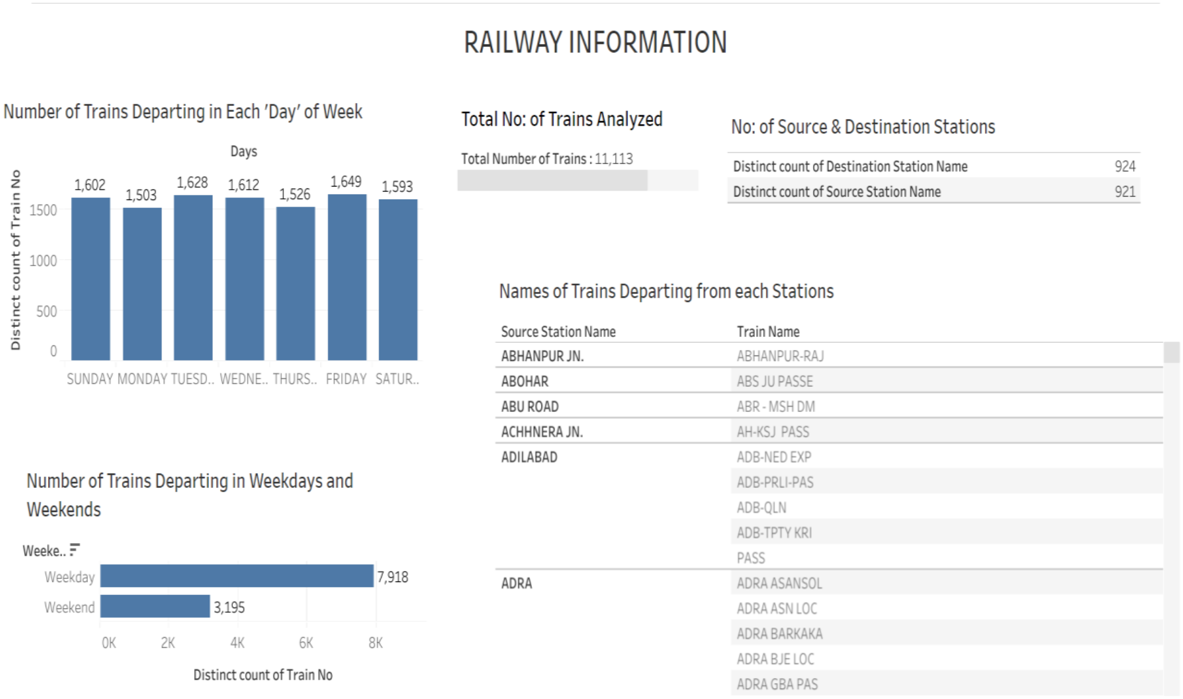
1. SUMMARY OF FINDINGS:

The following are the key findings obtained during Data Analysis:

- Total Number of Trains Analysed: 11, 113
- Distinct Number of Source Station Names: 921
- Distinct Number of Destination Station Names: 924
- Number of Trains Departing on Weekdays: 7918
- Number of Trains Departing on Weekends: 3195
- Friday has the Highest Number of Train Departures i.e. 1649
- Monday has the Lowest Number of Train Departures i.e. 1503
- 'CST-MUMBAI' is the most common station name among source and destination stations.

The below figure shows the screenshot of Tableau visualization of analysis of data regarding Railway Information.

The link of complete Viz of Tableau Public is : https://public.tableau.com/app/profile/sinu.s.mariam/viz/Railway_Information/Dashboard1



Total No: of Trains Analyzed

Total Number of Trains : 11,113

No: of Source & Destination Stations

Distinct count of Destination Station Name	924
Distinct count of Source Station Name	921

Names of Trains Departing from each Stations

Source Station Name	Train Name
ABHANPUR JN.	ABHANPUR-RAJ
ABOHAR	ABS JU PASSE
ABU ROAD	ABR - MSH DM
ACHHNERA JN.	AH-KSJ PASS
ADILABAD	ADB-NED EXP
	ADB-PRLI-PAS
	ADB-QLN
	ADB-TPTY KRI
	PASS
ADRA	ADRA ASANSOL
	ADRA ASN LOC
	ADRA BARKAKA
	ADRA BJE LOC
	ADRA GBA PAS

Fig 1.1: Screenshot of TABLEAU Public Visualization Report

2. DATA INSIGHTS

2.1 FEATURE IMPORTANCE PLOT

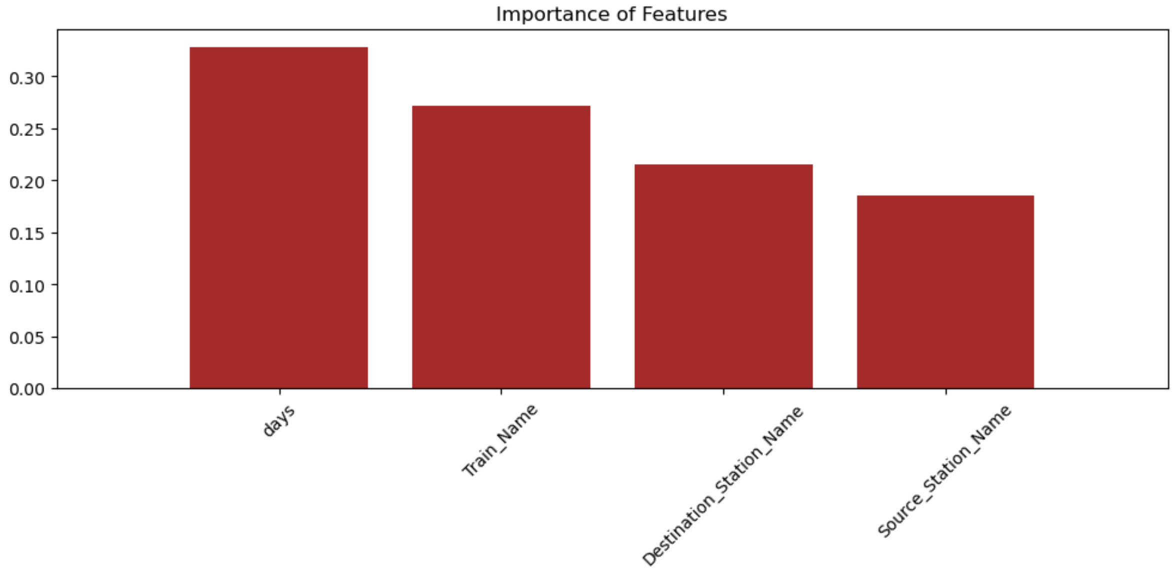


Fig 2.1 : Feature Importance Plot of Feature Variables Affecting the Prediction of Target Variable 'Frequency of Operation'

- Random Forest Regressor is used to predict Target Variable 'Frequency of Operation' of Trains based and the hyperparameter tuning technique Grid Search is used
- Best Hyperparameters obtained for Random Forest Regressor after Grid Search are: *{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}*
- The variable 'days' which represents the Day of Departure has the highest Impact on the predictive model followed by 'Train_Name'

2.2 POTENCIAL ASPECTS / ISSUES IDENTIFIED DURING DATA ANALYSIS

When the 'Frequency of Operation' of Trains are analysed, only one unique value was identified and that is '1'. Some issues with the data are identified with the data that may affect the analysis are:

- Only one value in the Target variable 'Frequency of Operation' will cause overfitting of data and may cause inaccurate predictions when new data is scored during production phase of code.

- When analysed with predictive models such as Linear regression, Decision Tree and Random Forest Regressor for predicting frequency of operation, Mean Squared Error remained 0.0 and Coefficient of Determination or R Squared Value (r2) , Accuracy Score, Precision Score and Recall value remained 1.0 because of only one unique value in target variable.
- So, since we are unable to identify the best model out of three models*, we assumed that Random Forest Regressor,* and use it for developing the final model pipeline for Deployment

Addressing the above issues will help in more accurate and reliable results.

3. ACTIONABLE RECOMMENDATIONs

- Including columns regarding 'Train Timings' can help to analyse train delays.
- Including data regarding customer feedbacks and ratings in the data can help in analysing the overall customer sentiments through NLP- Sentiment Analysis and it may help in understand the aspects that affect customers satisfaction.

PRESENTED BY,

SINU S MARIAM
Machine Learning Intern,
COGNIFYZ TECHNOLOGIES