FLIGHT DELAY CLASSIFICATION

VENKAT BIYYAPU

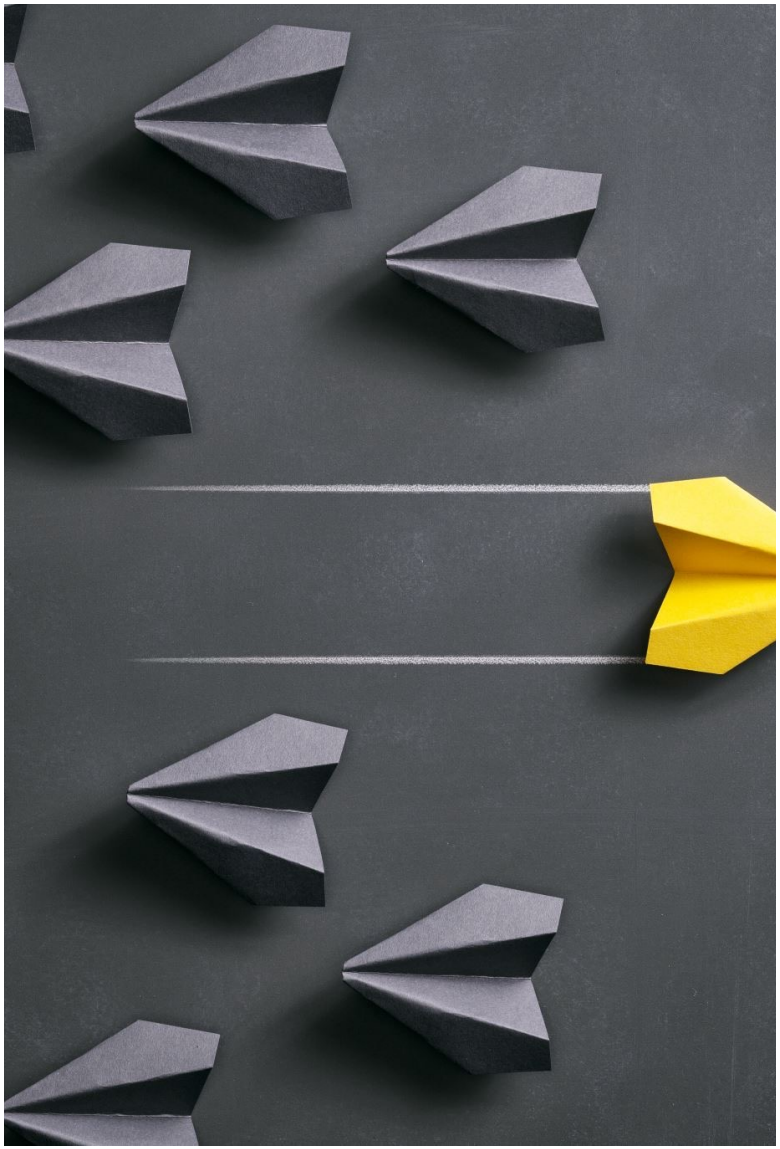
VXB220005

AGENDA

- Problem Statement
- Concept and Motivation
- Dataset Description
- Explore Data Analysis
- Methodology
- Results
- Conclusion

PROBLEM STATEMENT

To classify the Jan_2019 Ontime Flight dataset into Delayed or Not Delayed category(Flight Delayed or Not) by using various Machine Learning classification models and evaluating them using different evaluation metrics.



CONCEPT AND MOTIVATION

- Today, the aviation sector is vital to global transportation, and many companies depend on different airlines to connect them to other regions of the globe. However, due to various reasons there may be flight delays.

- To solve this issue, accurately classifying these flight delays allows passengers to be well prepared.

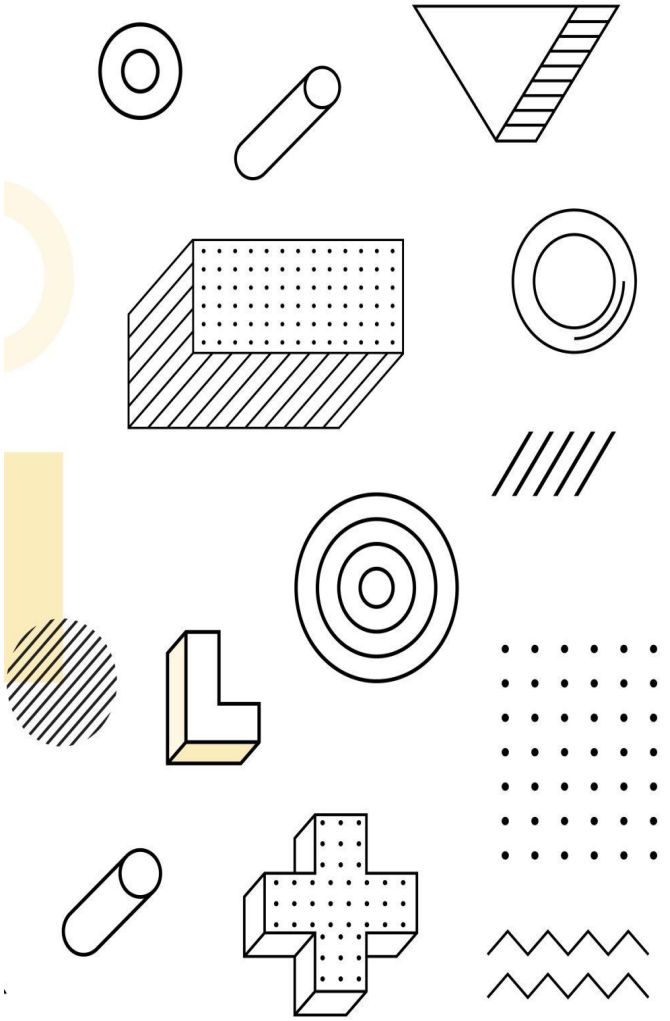
- The goal is to classify the dataset using sklearn and keras python module to build a classification model for flight delay.

DATASET DESCRIPTION

The Jan_2019 Ontime Flight Dataset from Kaggle was used for this project.

Based on several input features, this dataset is used to decide if a flight is delayed or not.

Relevant details regarding the flight are provided in each row of the data.

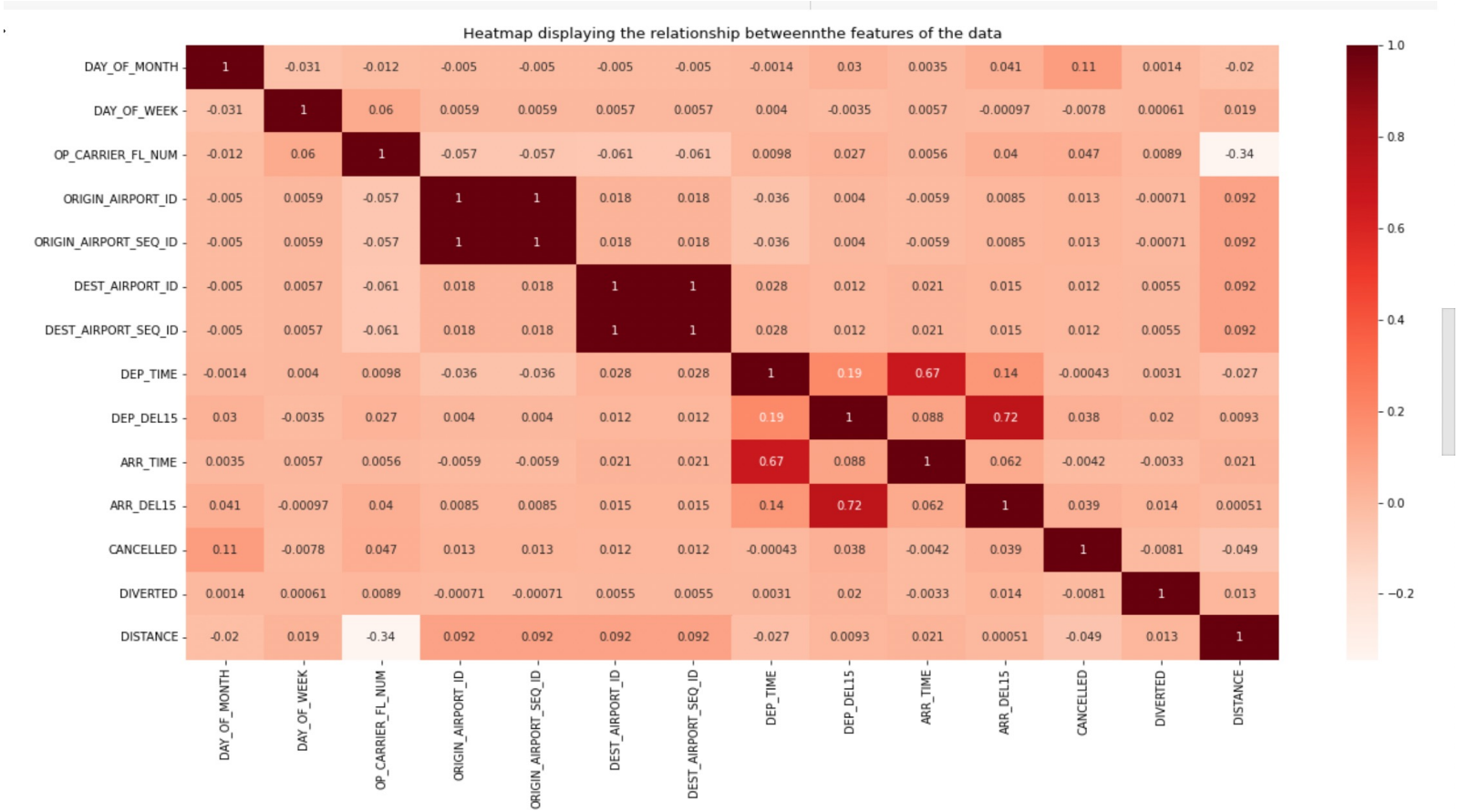


DATASET DESCRIPTION

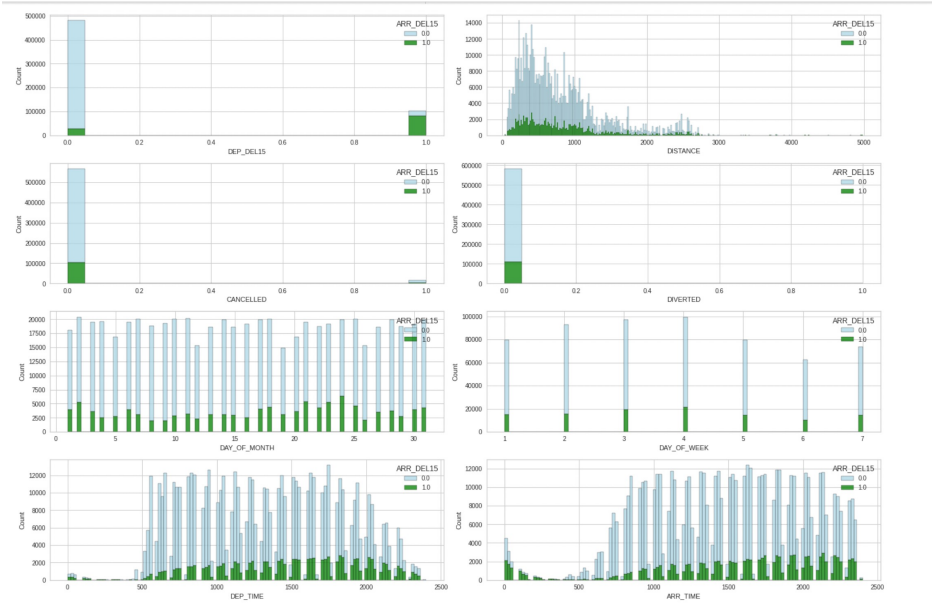
- The dataset contains 583985 rows with 22 features set.
- Features are:
  - 'DAY_OF_MONTH', 'DAY_OF_WEEK', 'OP_UNIQUE_CARRIER', 'OP_CARRIER_AIRLINE_ID', 'OP_CARRIER', 'TAIL_NUM', 'OP_CARRIER_FL_NUM', 'ORIGIN_AIRPORT_ID', 'ORIGIN_AIRPORT_SEQ_ID', 'ORIGIN', 'DEST_AIRPORT_ID', 'DEST_AIRPORT_SEQ_ID', 'DEST', 'DEP_TIME', 'DEP_DEL15', 'DEP_TIME_BLK', 'ARR_TIME', 'ARR_DEL15', 'CANCELLED', 'DIVERTED', 'DISTANCE', 'Unnamed: 21'

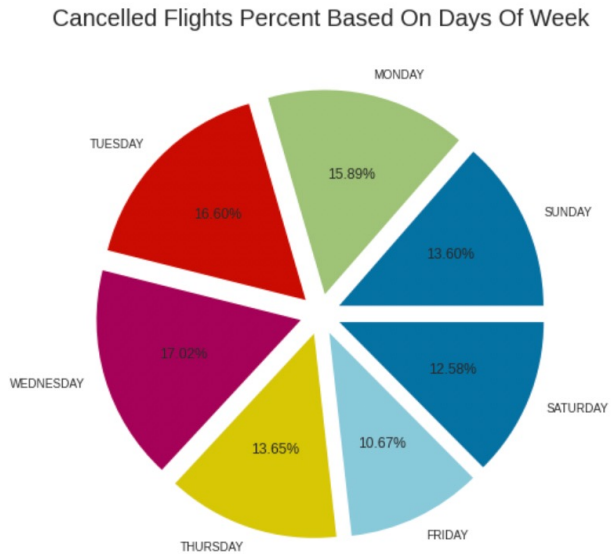
- Description about some important features are below:
  - **DAY_OF_MONTH*:*** indicates day of the month
  - **DAY_OF_WEEK*:*** indicates day of the week
  - **ORIGIN_AIRPORT_ID*:*** indicates unique origin airport id
  - **DEST_AIRPORT_ID*:*** indicates unique destination id
  - **DEP_DEL15**: indicates there is departure delay or not(1=yes , 0=no)
  - **CANCELLED*:*** indicates whether flight is cancelled or not(1=yes , 0=no)
  - **DIVERTED*:*** indicates whether flight is diverted or not(1=yes , 0=no)
  - **DISTANCE** : indicates the distance between airports in miles
  - **ARR_DEL15*:*** indicates if the flight is delayed or not (1 = yes, 0 = no)(Target feature)

EXPLORE DATA ANALYSIS



EXPLORE DATA ANALYSIS





METHODOLOGY

- Initially I performed Data Cleaning .
- Then, I did some exploratory analysis on the data, like performing Uni and Bi- variate analysis on features and also explored more about individual features that affecting the target feature.
- I trained the data set using 8 different classification models and also did tunning hyper parameters in order obtain best model. Below are the models that are used :
  - **K-Nearest Neighbor** - which classifies or predicts how an individual data point will be grouped using proximity.
  - **Naive Bayes -** simple probabilistic classifier
  - **Decision Trees Classifier -** simple decision rules inferred from the data features
  - **Logistic Regression -** uses the sigmoid function to return the probability of a label.
  - **GBDT -** a meta-estimator that averages the results of several decision tree classifiers fitted to different dataset sub-samples to increase the predicted accuracy.
  - **Random Forest** - uses bagging and feature randomness.
  - **Neural Networks -** A NN Sequential model is appropriate for a plain stack of layers where each layer has exactly one input tensor and one output tensor.
  - **SVM** – used for classification, outlier detection and regression

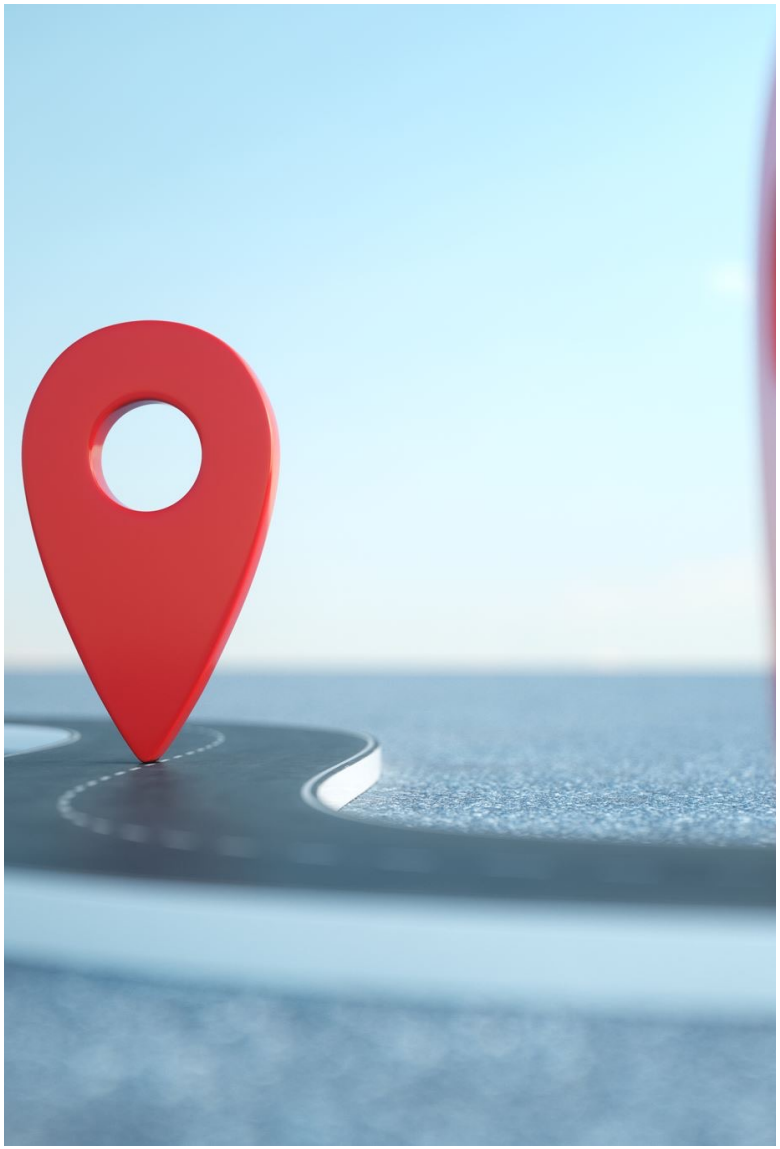
RESULTS

Model	F1-Score	Cross_Val_Score (ROC_AUC)
KNN	0.8679	0.784
Naïve Bayes	0.8977	0.851
Decision Trees	0.9163	0.879
Logistic Regression	0.9158	0.858
GBDT	0.9194	0.926
Random Forest	0.9227	0.937
Neural Network	0.9158	-
SVM	0.9173	0.890

CONCLUSION

Successfully able to classify the dataset into delayed and not delayed and also, we can observe that Gradient Boosting Decision Trees, Random Forest and Neural Networks were able to classify the dataset with highest accuracy and cross validation score.

Moreover, with the help of EDA I was able to select the relevant features for the classification which provided me utmost accuracy



REFERENCES

- https://www.kaggle.com/code/godswayjd/flight-delay/data
- https://www.geeksforgeeks.org/
- https://towardsdatascience.com/