**Introduction**

Predicting flights delay is always an interesting topic to Data Scientists. Since the issues of flights being late costs both airplane operations and consumer money and time, flight delay prediction plays a key role in aviation industry. There is much research have tried to develop and deploy different Machine Learning models to increase the accuracy of prediction.

In this one-week Midterm Project, we are supposed to perform a Data Science tasks which involving Preprocessing/Clean Data, Exploratory Data Analysis (EDA), Model Selection, Feature Engineering, Feature Selection, Hyperparameter Tuning, and Prediction

**Project Workflow**

The timeline for the project is a week. Therefore, the team decided to utilize an iterative minimal viable product (MVP) process, where we go through the entire process from data collection to model prediction deployment. We had a total of 3 iterations taking us from model 0 to model 3. Every iteration’s goal was to increase prediction power using different techniques and combining the best ones. Some of the tools used to increase predictive power were feature engineering (weather conditions, route traffic, datetime dates), dimensionality reduction using variable selection as well as changing categorical variables into dummy variables.

**Data**

The data was stored in a Postgress database and accessed using writing SQL queries in VSCode exported as CSV files. For the MVP (model 0) we used 14k rows randomly selected samples of the data. The trade-off here was smaller sample size (space) for time much faster run times throughout our project. We made the trade off because we figured we can always go back extract more rows and feed it to a ready model once the pipelines of the entire project were layed out.

**Data Preparation**

The techniques and methodologies used for preprocessing are summarized below, for code and more details please check both EDA and modeling files in the repository.

1. Handling missing values: Columns with high percentage were dropped, lower percentage columns were replaced with the mean.
2. Formatting times: Initially the times in the dataset are in the form of 4 digit numbers which are not very useful, using a transformed function, the times were transformed into HH:MM format
3. Feature selection: Some of the features were not needed for the prediction of delays, so the following were dropped for base model this was done in a sense manually using common sense. As we progressed in our process more analytical dimensionality reduction techniques were used such as filter feature selectors, based on two criteria small variance and high correlation.
4. Dummy variables & label Encoding: transforming categorical variables into numerical values to make our data machine learning model friendly, as generally model tend to give better results when data is presented numerically.
5. Normalize the values and scale: Used standardised scaling for X values fed before feeding it into our models.

**Modeling Process**

As mentioned above, we used an MVP iterative approach, and we had a total of 3 iterations taking us from model 0 to model 3. Every iteration’s goal was to increase prediction power using different techniques and combining the best ones. Each model is summarized below:

\*\* note for code and more details please check both EDA and modeling files in the repository\*\*

MVP (base model 0): Only numerical categories, with no feature engineering, variable selection, or

Model 1: Feature engineered Weather conditions using data pulled from weather API, Route Traffic using total flights from city to city, Dates to day month day of week. All changed to dummy variables.

Model 2: Same features from model 1 but instead of dummy variables we used Lab Encoder to reduce the features and potentially increase predictive power.

Model 3: model 2 kept unchanged, Variable selection to reduce number of features (with small variance and high correlation)

**Actual Machine learning models used**

We treated this as a regression problem, we want to give back an actual estimate in minutes for each flight given a set of features we know before the flight takes place, therefore we used regression models, as well as stacking which takes base models and uses a meta model which increases complexity and can increase predictive power to a great degree. The regression models we decided to use are listed below:

* Models Model 1: Linear Regression
* Model 2: SVR
* Model 3: Random Forest Regressor
* Stacked Linear regression with 5 cross folds

**Results**

Using stacked Linear Regression as final estimator increases the result 1%, but increases the cost associated by 10 times. After consideration, we decided to use Random Forest Algorithm to predict our arrival delay targets.



