Predicting Significant Flight Delays using Supervised Learning

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
    - We are trying to predict if a given flight will be significantly delayed (by at least 15 minutes).
- What industry/realm/domain does this apply to?
    - This problem applies to the airline industry, or to a firm who is seeking to purchase airline tickets for business related travel with lowest probability of delay.
- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)
    - Our hope is that our model will help ticket purchasers make more informed decisions about their purchases. Specifically, this model could help a purchaser estimate the risk of a flight being significantly delayed.

Data Understanding

- What data will you collect?
    - We will collect data pertaining to flights spanning between August 2021 and July 2022.
- Is there a plan for how to get the data (API request, direct download, etc.)?
    - We intend to download relevant data from Kaggle that is available for use by the public domain.
- Are the features that will be used described clearly?
    - Yes - the features are described in https://www.kaggle.com/datasets/whenamancodes/flight-delay-prediction.

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?
    - We will select a few key features to use in our model given the several dozen features in the original dataset.
    - We will consider using a random sample from the data for our model to reduce the high number of rows, pertaining to hundreds of thousands of flights per month.
    - We may use one-hot encoding as the dataset has several categorical features.
    - We might be interested in engineering our own features, such as if a flight departs during a holiday week.
- What are some of the cleaning/pre-processing challenges for this data?
    - The high number of features could make it difficult to investigate multicollinearity.
    - One-hot encoding could quickly increase the number of features in our model.

Modeling

- What modeling techniques are most appropriate for your problem?
    - We are interested in techniques for supervised classification such as logistic regression, decision tree and random forest models, K-nearest neighbors, or naïve Bayes. Depending on the model, we will be tuning hyperparameters or applying regularization as appropriate to improve model performance.
- What is your target variable? (remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)
    - The target variable is a binary indicator which is set to 1 if a flight is delayed by at least 15 minutes and 0 if it is not.
- Is this a regression or classification problem?
    - Due to the categorical nature of the target variable, this is a classification problem.

Evaluation

- What metrics will you use to determine success (MAE, RMSE, etc.)?
    - We will use a confusion matrix along with several related metrics including accuracy, precision, and recall to evaluate each classification model.

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?
    - We are planning to build, tune, and evaluate a logistic regression model and a decision tree model.