Project Proposal for CS334
Zhou Fang, Ray Chen

Please Please Don't Delay!
-- Airline Delay and Cancellation Prediction

Objective & Motivation:

The project aims to predict the delay time of airlines based on a series of airline information, specifically, during the COVID-19 pandemic. We are interested in this problem because annoying flight delays and cancellation happen everyday, only more often during the pandemic. These unexpected changes will disrupt people's travel plans and potentially cause financial losses. If we develop a robust flight delay & cancellation predictive model, it will give much more convenience to airports, flight companies, and passengers to timely adjust the schedule.

Dataset:

- **Source:**
  The dataset called "Reporting Carrier On-Time Performance (1987-present)" was extracted from the Bureau of Transportation Statistics in the United States Department of Transportation (1). Similar extraction from this source is also available on Kaggel but that dataset only includes data until 2018 (2; see References 3 and 4 for pre-existing milestones). Since we wish to focus on the airline status during the COVID-19 pandemic, we collected the whole-country data for the latest 12 available months, from Aug. 2020 to Jul. 2021, from the data source. The entire dataset we extracted consists of approximately 6 million rows of flights and 27 columns of features.
- **Target Features:**
  There are five target features that we are interested in. We will primarily focus on the first three binary features. If time allows, we will also experiment with the two continuous features.

Feature Name	Property	Description
DepDel15	Binary	If the airline has a departure delay for over 15 minutes (1 = Yes)
ArrDel15	Binary	If the airline has arrival delay for over 15 minutes (1 = Yes)
Cancelled	Binary	If the airline is cancelled (1 = Yes)
DepDelay	Continuous	Differences in minutes between scheduled and actual departure time. (Negative for early departure)
ArrDelay	Continuous	Differences in minutes between scheduled and actual arrival time. (Negative for early arrival)

- **Potential Predictor Features:**
  The rest of 22 features will be evaluated whether they are valid predictors. Some valid ones include day of month, day of week, origin and destination state, scheduled elapsed time of flight, flight distance, etc. Some invalid ones will be excluded either because they are uniform (e.g., year) or because they have direct association with the target features (e.g., actual departure or arrival time).

Working Plans:

- **Data Preprocessing:**
  Several steps will be done for preprocessing the data.
  1. Some features will be excluded as mentioned in the last section.
  2. Categorical features such as airline and state will be converted into one-hot encoding. If some features turn out to have too many categories (e.g., city), these features may be excluded for the initial prediction attempts.
  3. The numerical features will be evaluated with a pair-wise correlation matrix and all except one features that highly correlate with each other will be excluded.
  4. All numerical features will be centered to zero and scaled to unit variance.
  5. Considering that the airline status might undergo changes when increasing numbers of people were vaccinated, data since May 2021 will be compared with the rest to see if there are noticeable differences. Since our ultimate goal is to predict delay time in the future but not just within the past year, if any significant difference is found, we will only train and test with data since May 2021, which is still sufficiently large with approximately 1.5 million rows.
- **Model Selection:**
  Multiple models will be tuned and compared to predict the binary target features, including kNN classifier, decision tree classifier, logistic regression, and Naive Bayes classifier. If time allows, we will also tune and compare linear regression and support vector machines for regression to predict the continuous target features.
- **Model Evaluation:**
  In our case of predicting flight delay, false positives are less important because one's less likely to be impacted by an on-time flight predicted as delayed, compared to a delayed flight predicted as on time. Therefore, we will primarily evaluate the F-2 score - a weighted harmonic mean of precision and recall with more weight on recall - via 5-fold cross-validation. Other metrics will also be calculated just for references.

References:

1. **Reporting Carrier On-Time Performance (1987-present)**
   Source of our data.
   https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FGJ
2. **Airline Delay and Cancellation Data, 2009 - 2018**
   A dataset from Kaggle that was extracted from the same source (1) but of different timespan.
   https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018
3. **Classifying a Flight with naive Bayes Classifier**
   The project from Kaggle with the best accuracy to predict the delay of flight in the dataset (2). The author preprocessed the data, made exploratory data analysis to examine the attributes and applied the Naive Bayes Classifier for prediction. As a result, the classifier got an accuracy of 81%.
   https://www.kaggle.com/gokhanegilmez/classifying-a-flight-with-naive-bayes-classifier
4. **Predictive model for on-time arrival flight based on weather data**
   This paper used a dataset of flights in Japan and predicted on-time flights. Different from (3) and our project, their predictor features were mainly weather attributes. As a result, a Random Forest Classifier achieved the best accuracy of 77%.
   https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0251-y