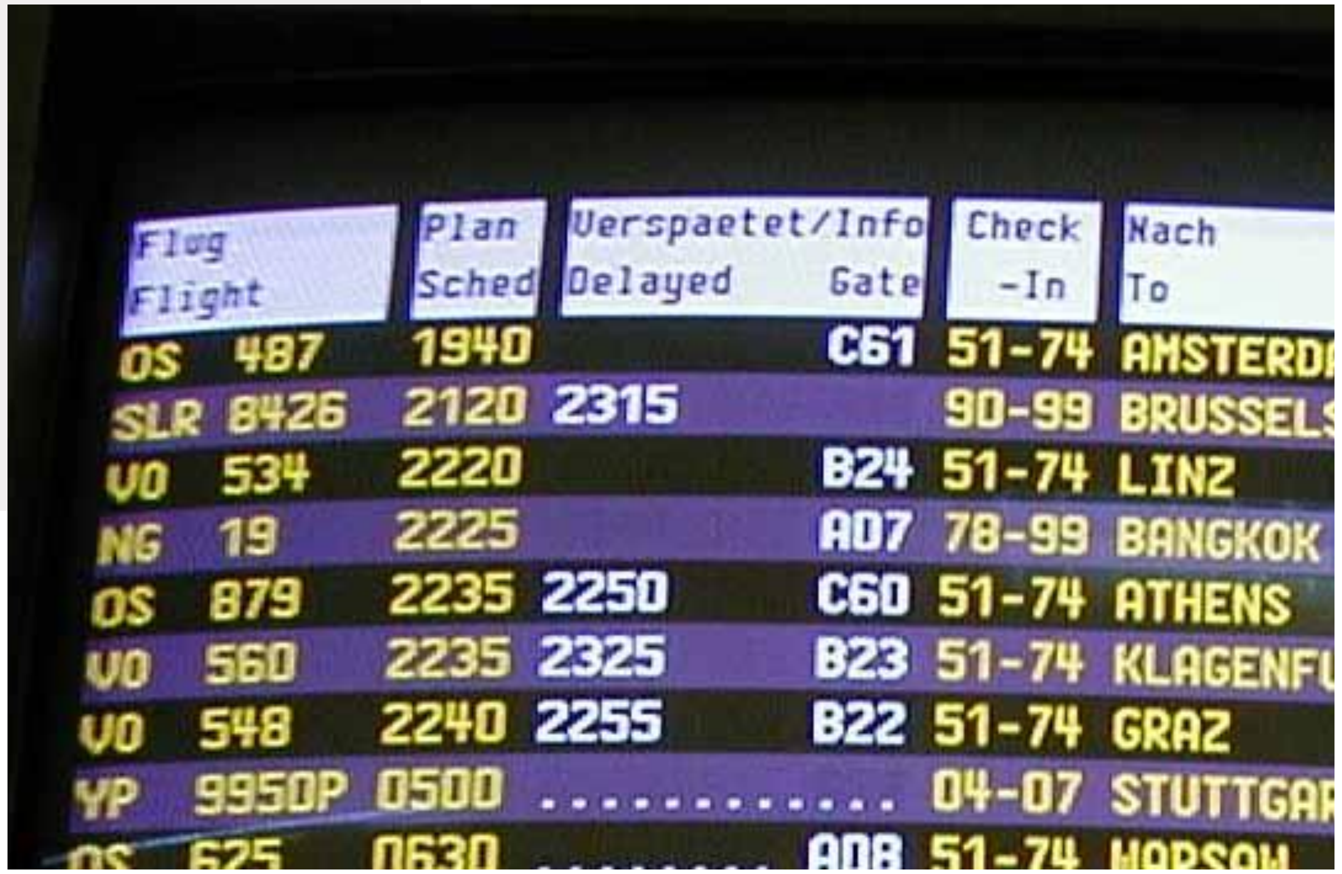
Big Data for Finance Final Project

Predicting Airline Delays



Jason Nam, Shantanu Bisht, Muhammad Shahzad, Jiale Guan, Benjamin Rowley

Problem Description

Problem Overview

- The airline business is a significant component of the transportation sector and has a significant impact on the world economy.
- **Flight delays** are one of the single biggest difficulties the airline industry must overcome. Weather, technical problems, air traffic congestion, crew scheduling, and other reasons can all contribute to flight delays. Flight delays can harm travelers schedules, airlines, and the entire supply chain as a whole.

Problem Description

Overview of Dataset

- Our dataset comes from the US Department of Transportations, which tracks the on-time performance of domestic flights
- The dataset provides data of on-time, delayed, canceled, & diverted flights for 10 years, from 2009 to 2019
- The Airline Delay file contains 8 columns and 539383 rows
- Columns of the Airline Delay file consist of: Time (Length of flight), Length (Length of flight), Day of Week, Airline, Airport From (Origin airport), and Airport To (Destination Airport
- Our goal in analyzing the data set was to outline any specific patterns in airline delays

	Flight	Time	Length	Airline	AirportFrom	AirportTo	DayOfWeek	Class
0	2313	1296	141	DL	ATL	HOU	1	0
1	6948	360	146	OO	COS	ORD	4	0
2	1247	1170	143	B6	BOS	CLT	3	0
3	31	1410	344	US	OGG	PHX	6	0
4	563	692	98	FL	BMI	ATL	4	0

Problem Description

Objectives

- After careful analysis of our dataset, we concluded that we can use delay classification parameter as labelling to implement a **supervised predictive delay model**.
- Supervised machine learning is a type of machine learning where a model is trained on labeled data, with the goal of making accurate predictions on new, unseen data. In supervised learning, the input data (also called the features or predictors) and the desired output data (also called the target or labels) are provided to the model during training.
- Our primary objective is to use relevant flight related data to produce increasingly accurate predictive models in order to predict potential flight delays with a high degree of accuracy for a variety of airline companies.
- By successfully developing highly accurate predictive models, we can help airline companies better manage their flight schedules and reduce the impact of delays on passengers. This will ultimately lead to improved customer satisfaction and more efficient airline operations.

Solution Summary

Overview of Predictive Models Results

- Classification algorithms were used to: **Predict flight delays, Identify delay causes, Categorize delays, Analyze delay trends.**
- The prediction model's results were disappointing, with an accuracy score of just 64%. This indicates that accurately predicting flight delays with the given parameters is challenging.
- Solutions can help to:
  - Reduce airline delays.
  - Save costs generated by delays.
  - Improve customer satisfaction.
  - Enhance airport management ability.
- Several tools were used to test and implement the solutions, including Python, Jupyter Notebook, and Excel. Python libraries such as NumPy, Pandas, and Scikit-learn were also utilized.
- The projected timeline for each step varies, with predicting delays taking the longest time, while identifying delay causes and categorizing delays requiring similar durations. Analyzing delay trends, on the other hand, will take less time.

Solution Details

Solution Details for Predictive Model

- First, the data was cleaned up and normalized using K-Nearest Neighbor Imputer, Standard Scalar, and Simple Imputer.
- The flight number parameter was deemed irrelevant, and was cleared from our working dataset.
- The categorical data were further transformed and converted to numerical data using One Hot Encoder.
- Some key assumptions were made: ***the data is accurate, the data is representative, the features are relevant, the data is normally distributed.***
- The data was split into training data and testing data, 70% and 30% respectively from the entire dataset.
- The model was run using 4 predictive models:
  - **Decision Tree:** A decision tree classifier is used with a maximum depth of 4.
  - **Random Forest:** A random forest classifier is used with a maximum depth of 4.
  - **K-Nearest Neighbors:** A K-Nearest Neighbors classifier is used with k=5.
  - **Logistic Regression:** A logistic regression classifier is used.
- From the results of the models implemented, it was determined that the logistic regression had the best accuracy score. However, the other models had accuracy score not far behind logistic regression.
- Overall, all models exhibited poor prediction test accuracies.

Solution Details

Data Modeling

```
Accuracy score of the Decision Tree model is 0.6342737076290826
Accuracy score of the Random Forest Tree model is 0.6258628680901029
Accuracy score of the K-Nearest Neighbors model is 0.6334023421808855
Accuracy score of the Logistic Regression model is 0.6418873404814139
```

```python
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.pipeline import Pipeline
from sklearn.impute import KNNImputer
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
```

```python
y = df.iloc[:,7]
x = df.iloc[:,1:7]
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 42)
```

```python
# Pre-process data: Numerical Xs

step_1a = Pipeline(
    [
        ("1A_i", KNNImputer(n_neighbors = 6)),
        ("1A_ii", StandardScaler())
    ])
```

```python
# Pre-process data: Categorical Xs

step_1b = Pipeline(
    [
        ("1B_i", SimpleImputer(strategy = 'most_frequent')),
        ("1B_ii", OneHotEncoder())
    ])
```

```python
# Pre-process data: Column transforms

num_x = ['Time', 'Length', 'DayOfWeek']

cat_x = ['Airline', 'AirportFrom', 'AirportTo']

step_1c = ColumnTransformer(
    [
        ('1C_i', step_1a, num_x),
        ('1C_ii', step_1b, cat_x)
    ])
```

```python
# Decision Tree

max_depth = 4

model_dt = Pipeline(
    [
        ('1C', step_1c),
        ('2_RF', DecisionTreeClassifier(max_depth = max_depth, criterion = "entropy"))
    ])

model_dt.fit(x_train, y_train)
```

```python
# Random Forest Tree

max_depth = 4

model_rf = Pipeline(
    [
        ('1C', step_1c),
        ('2_RF', RandomForestClassifier(max_depth = max_depth))
    ])

model_rf.fit(x_train, y_train)
```

```python
# K-Nearest Neighbors

n_neighbors = 5

model_knn = Pipeline(
    [
        ('1C', step_1c),
        ('2_RF', KNeighborsClassifier(n_neighbors = n_neighbors))
    ])

model_knn.fit(x_train, y_train)
```

```python
# Logistic Regression

n_neighbors = 5

model_lr = Pipeline(
    [
        ('1C', step_1c),
        ('2_RF', LogisticRegression())
    ])

model_lr.fit(x_train, y_train)
```

```python
print("Accuracy score of the Decision Tree model is " + str(model_dt.score(x_test, y_test)))
print("Accuracy score of the Random Forest Tree model is " + str(model_rf.score(x_test, y_test)))
print("Accuracy score of the K-Nearest Neighbors model is " + str(model_knn.score(x_test, y_test)))
print("Accuracy score of the Logistic Regression model is " + str(model_lr.score(x_test, y_test)))
```

Results and Recommendations

Recommendations

Unfortunately, our predictive analysis failed to predict flight delays accurately.

We recommend to the airlines industry and airports around the US to

- **Weather severity metric:** Weather can impact flight operations, and a severity metric can help in quantifying the impact.
- **Delay Analytics:** Complement real-time analytics with historical “flight delay” data
- **Airline Revenues:** Revenues of different airlines can help us categorize different airlines into small, medium, and large.
- **Airport Foot Traffic**: High foot traffic can lead to congestion, which can affect airport operations, such as delays.

Keep record of every significant data that can be analysed and then utilized to be used in predictive analysis.