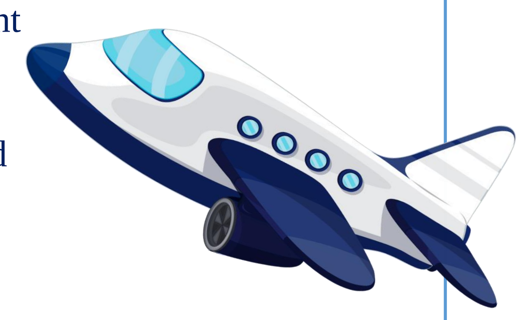
Prediction of Flight Delays

Ancheng Deng, Ruijia Sun, Shuran Yu, Jiawen Zang, Yuchen Zhou

550.636 Data Mining

INTRODUCTION

Project purpose



The purpose of our project is to analyze different features that may affect the departure of flights and predict the fight delays. We choose two airports that are close to us: BWI and DCA, and hope to help people to form a reasonable expectation of possible delays in their next trip.

Data Source

We get our data from the US Department of Transportation's Bureau of Transportation Statistics website. We select 1-year data of flights departing from BWI or DCA in 2017. All variables we think may affect the departure of flights are downloaded first and then processed differently based on their properties.

VARIABLES SELECTED

Numerical Variables:

- Departure Time
- Arrival Time
- Wheels Off Time
- Wheels On Time(Land)
- Delayed Time of Departure
- Number of Cancelled Flight
- Number of Diverted Flight
- Weather Score
- Taxi-in
- Taxi-out
- Distance

Categorical Variable:

- Airline
- Original Airport
- Destination Airport
- Destination City Name

Target Variable:

- Delay Index
  - 1 = Delayed over 15min
  - 0 = Delayed within 15min

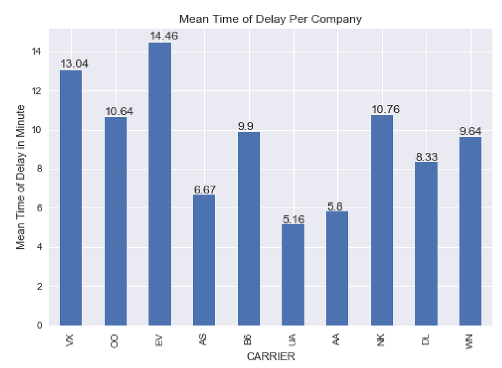
DATA EXPLORATION

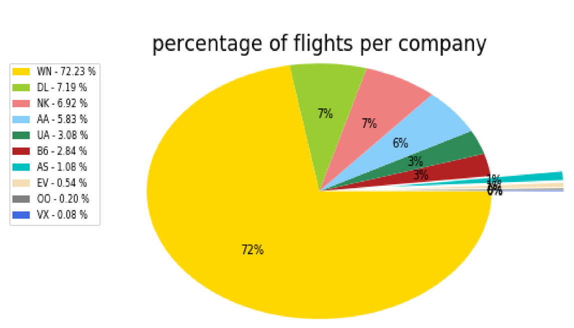
Impact of destination airports

The two figures below show the delaying rate with regard to different destination airports and months, and the departure airport is BWI and DCA respectively.

Impact of airlines

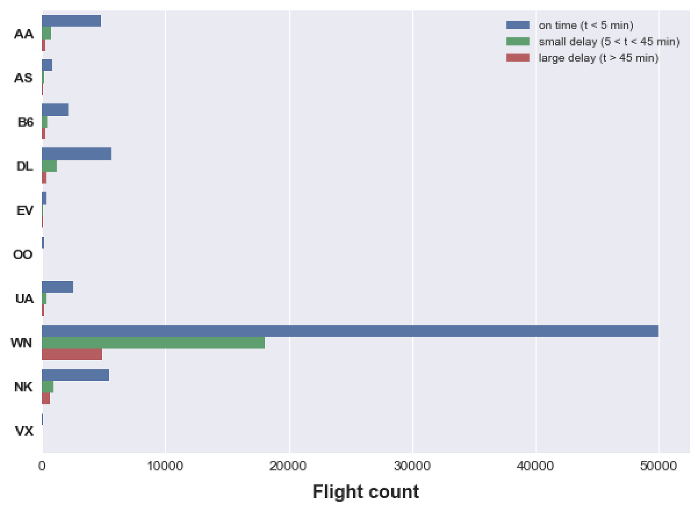
The figures below show the percentage of mean delay per company and percentage of flights per company with BWI as the departure airport.





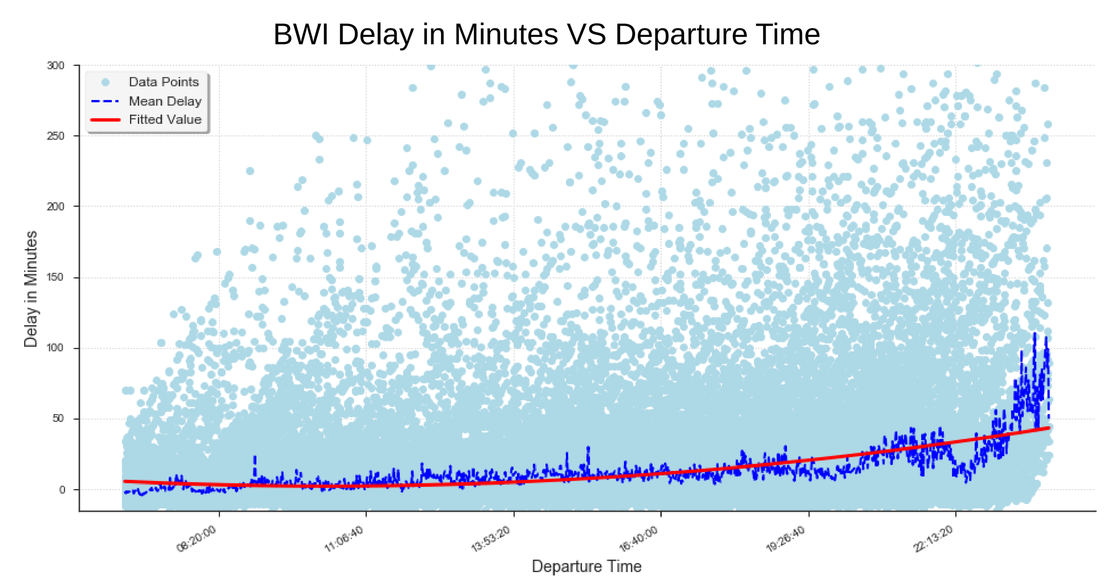
The bar chart shows the number of delays with different degrees in each airline respectively.

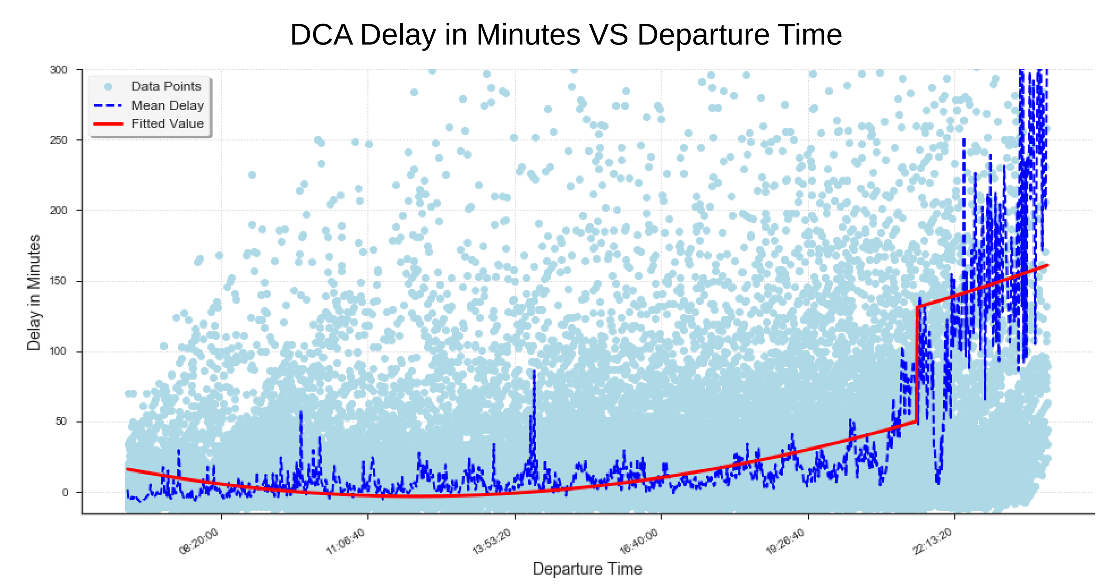
IATA_CODE	AIRLINE
UA	United Air Lines Inc.
AA	American Airlines Inc.
US	US Airways Inc.
F9	Frontier Airlines Inc.
B6	JetBlue Airways
OO	Skywest Airlines Inc.
AS	Alaska Airlines Inc.
NK	Spirit Air Lines
WN	Southwest Airlines Co.
DL	Delta Air Lines Inc.
EV	Atlantic Southeast Airlines
HA	Hawaiian Airlines Inc.
MQ	American Eagle Airlines Inc.
VX	Virgin America



Impact of departure time

The two figures show the delay in minutes with respect to departure time.





DATA PROCESSING

Missing Data

The first step of data preprocessing is to investigate on missing data. Fortunately, both of our dataset has less than 3% missing values, corresponding to flight cancellation and diversion. We are safe to just remove those NAs and obtain 99101 observations for BWI and 72755 for DCA.

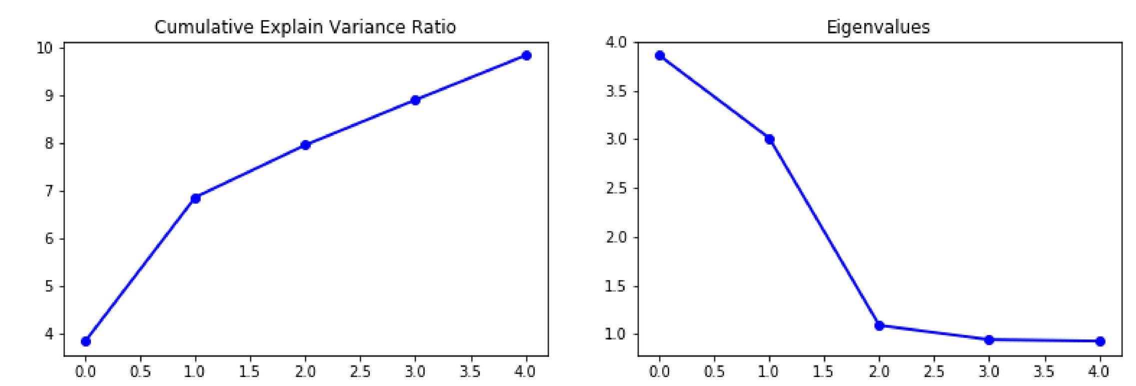
Date/Time Conversion & Train/Test Split

Next, we'd like to consider the date/time variables in our dataset. Given that we only have time information within a day for variables like 'Departure Time', we convert them into minutes since 00:00, which is both convenient for storage and downstream processing. Training and testing sets are obtained with 6:4 ratio.

Standardization & PCA

After dropping variables that are redundant, we take 10 potential variables and standardize them in order to eliminate unit effects. After PCA transformation, 5 components accounting for 95% variance are selected.

The figures show the cumulative variance ratio and eigenvalues when doing PCA.



CALSSIFICATION

The accuracy scores of prediction of flight delays using different classification methods are listed below. Generally, our models perform better wrt. DCA Airport than BWI Airport. Among them, we find Decision Tree method achieves the best performance after PCA.

BWI Airport

	Accuracy Score Before PCA	Accuracy Score After PCA
KNN	0.814	0.825
NaiveBayes	0.749	0.798
DecisionTree	0.808	0.825
RandomForest	0.81	0.813
LogitRgression	0.6399	0.815
QDA	0.813	0.801
LDA	0.815	0.808
RBF-SVM	0.628	0.599
Poly-SVM	0.808	0.808
Sigmoid-SVM	0.586	0.617

DCA Airport

	Accuracy Score Before PCA	Accuracy Score After PCA
KNN	0.835	0.85
NaiveBayes	0.754	0.832
DecisionTree	0.839	0.853
RandomForest	0.842	0.851
LogitRgression	0.839	0.843
QDA	0.726	0.832
LDA	0.844	0.839
Poly-SVM	0.838	0.838

The prediction accuracy of the multi-layer perceptron (MLP) classifier using neutral network method with different parameters.

(Set hidden layer sizes = (2,).)

BWI – before PCA

activation \ solver	lbfgs	sgd	adam
logistic	82.69%	80.81%	80.81%
identity	82.27%	82.05%	82.62%
tanh	82.26%	80.81%	82.42%
relu	82.38%	82.53%	81.13%

BWI – after PCA

activation \ solver	lbfgs	sgd	adam
logistic	80.81%	80.81%	80.81%
identity	81.31%	81.30%	81.39%
tanh	80.81%	81.20%	80.81%
relu	80.81%	80.81%	80.81%

DCA – before PCA

activation \ solver	lbfgs	sgd	adam
logistic	86.54%	83.75%	84.88%
identity	85.01%	85.10%	84.98%
tanh	85.87%	84.40%	85.81%
relu	85.85%	84.59%	84.72%

DCA – after PCA

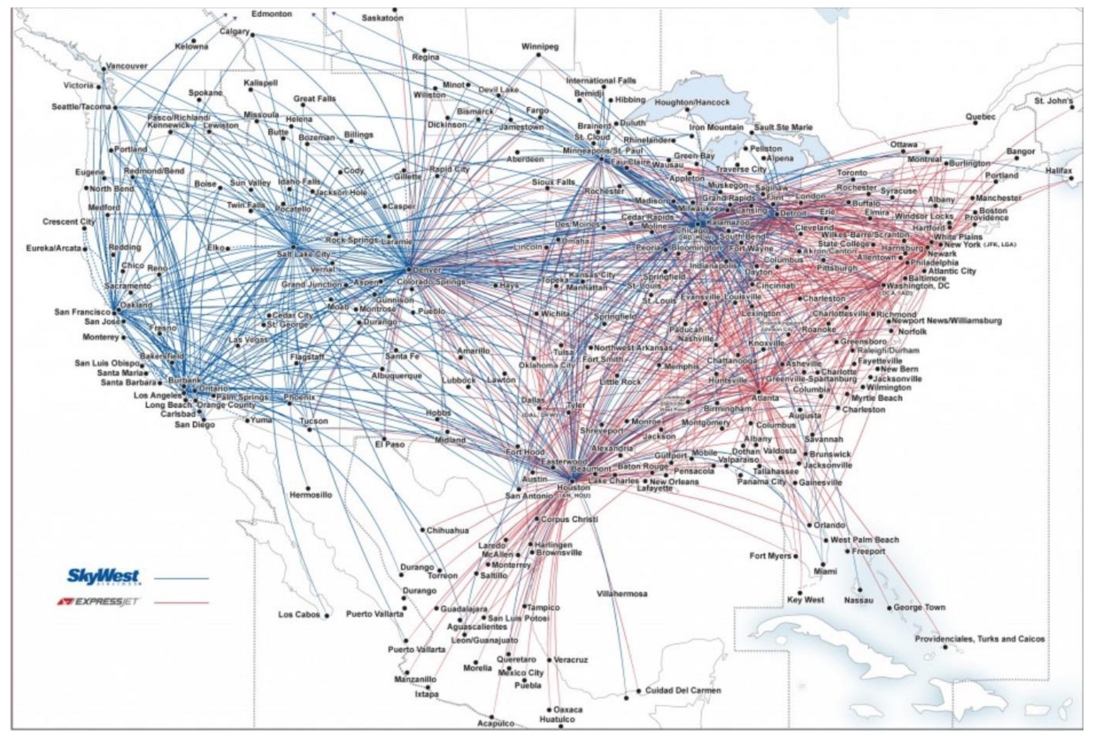
activation \ solver	lbfgs	sgd	adam
logistic	85.91%	83.75%	83.75%
identity	84.41%	84.42%	84.39%
tanh	86.15%	83.75%	84.97%
relu	86.65%	83.75%	84.44%

Logistic Regression Analysis

Using a pure logistic regression for original data before PCA, we find the following four explanatory variables are significant in both BWI and DCA: Departure time, Wheels off, Arrival time, and Weather. The results also show there is a positive correlation between extreme weather and delay which also coincides with our intuition. The worse the weather, the more likely a delay may take place. What`s more, the positive coefficient in departure time means that a larger depart time implies more likely a delay is going to take place. However, the negative correlation between wheels off and delay is not able for us to further explain which means our logistic model may be further improved.

SUMMARY

- We analyze the effects of different features on flights' departure on-time status for BWI and DCA airports.
- Several types of classifiers are trained before and after PCA transformation to original datasets.
- Among these classifiers, Decision Tree provides relatively better prediction results for both BWI and DCA airports.
- Given conditions, like weather, destination city and airline, we could predict flights' on-time status with over 82% accuracy, which may help passengers form reasonable expectation of their flights' departure time.



FUTURE WORK

Further, we can consider to discuss the Flight Delays of main International Airports, such as JFK, ORD, IAD, in USA. Using the principle components from PCA as Predictor Variables, Delay Index as Response Variables.

- Plot the relationship of Response Variables and Predictor Variables of each airport, to see if there is any random effect between each airport or within airport.
- Fit a LMM model, to further discuss the effect of each parameter has on the probability of Delay
- By observing the plots such as Residuals vs Fitted Value, we can consider further, to fit a Semi Parameter Model or Generalized Additive Model, to make improvements.
- Then we can try to predict the future delay rates in each Airports using data such as Weather forecast, Scheduled Departure Time and so on.

REFERENCE

- Understanding the Reporting of Causes of Flight Delays and Cancellations (https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations)
- Airports and Airlines Data (https://www.bts.gov/topics/airlines-and-airports)
- 550.436 Data Mining Lecture Notes by Professor Tamas Budavari