Project Group 20
Tianyi Li, Ruifeng Jiang, Meidi Peng

EECS 349 Final Project Report

Initial Motivation

Airplanes have become a main transportation tool since the last decades. Becoming more and more common to sit in a flight and reach your destination in 2-3 hours, strengthens the value of travel time saving in air travel tremendously. As time saving is one of the most important reasons why a customer would choose to planes as their transportation tool, it would just too frustrating if you reach an airport finding that your flight has been canceled or delayed when you are actually having an really important meeting at the destination.

Furthermore, for airport, one arrival delay may incur a chain of departure delays, making it important to predict the promising delay ahead of time. With better predicting the possible delays, airport can manage their resources better and can reduce their loss that delays may incur, which will benefit the customers in turn.

So the purpose of our project is to give people a detailed prediction of cancelled or delayed flights based on past year trends. So that the passenger can plan his trip in advance.

Task Description

Our task is to predict the possible delays(departure/ arrival), cancellation or diversion of a future flight in a specific area during a specific time period based on features of flight data.
The most interesting aspect about this task is that, we think we can learn a lot interesting trends about flights. For example, the difference of delay rate according to the popularity of cities, like tourist cities or major transportation airports. For individual customers, the importance of this task is also obvious: to reduce the probability of stucking in an airport with surprise delay.

Data Set Description (number of examples in training/test sets, specific features employed)

The data set contains flight data from 2014 to Feb 2017.
(source: https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time)
We preprocessed the data set:

a. Divided the data set into different sub-data-sets, according to month and location(state). The reason why we decided to do this is because we decided not to consider specific weather conditions that may affect each specific flight. The main purpose of our project is to predict the probability of flight delays for future flights, and weather is actually a really uncertain factor for future events. Instead of considering weather of a particular day as a factor, we are going to consider the climate in general, according to location(state) and a period of time(month).

b. We first defined 6 output classes: i) no departure delay and no arrival delay; ii) no departure delay but arrival delay; iii) departure delay but no arrival delay; iv) departure delay and arrival delay; v) diversion; vi) cancellation. And we re-managed the data on these four features and created the output class.

During the investigation we decided not to stick to so many output classes and we combined all delay, cancel, diverted flights to one class and we ended up with a data set that has two output classes.
The feature we have for now:
1. month
2. day of month
3. day of week
4. scheduled departure
5. scheduled arrival time
6. departure airport
7. arrival airport
8. distance
9. unique carrier (Carrier ID)
10. class

The training set we used are flight data from 2012-2015 of each July (291K rows of data) and the test set we used is flight data of July 2016 (61K rows of data).

Investigation Steps

1. For the baseline we chose to use ZeroR and got 54% percent of accuracy, which also described the data set that of all data half are on time flights and half are delayed or canceled flights.
2. Then we used different methods to train on the data sets, we listed 6 methods and their accuracies in the result chart. We first applied 10 cross validation to build each model, then we applied the test data set and get the second accuracy to compare on.
3. After comparing each accuracy, we started exploring the models, by changing some parameters, and compare the accuracy in between the same methods.

Result and Analysis

Classifier	Accuracy of 10-fold cross validation	Accuracy of test set
ZeroR	54.5352	54.2656
Naive Bayes	62.009	60.2154
Bayes Net	63.2402	60.6915
AdaBoost	61.2474	60.6753
Decision Tree	63.8545	59.4089
Random Tree	58.2736	54.0696
REP Tree	62.7479	58.4049

Table 1 Accuracy of classifiers

According to the table, Bayes Net and Decision Tree have best performance on 10-fold cross validation. So we explore more on the results of these two methods.

```
=== Confusion Matrix ===

    a     b   <-- classified as
20063 13446 |     a = a
10827 17414 |     b = b
```

Figure 1 Bayes Net evaluation on test

According to the confusion matrix of Bayes Net evaluated on test set, there are 30860 examples predicted as delay and the accuracy is 56.429. And there are 30890 examples predicted as on time, the accuracy is 64.9498. Therefore, the total accuracy is 60.6915. Besides, there are 28241 delayed records and 17414 of them are correctly predicted. The hit rate is 61.6621.

```
=== Confusion Matrix ===

    a     b   <-- classified as
22495 11014 |     a = a
14051 14190 |     b = b
```

Figure 2 Decision Tree evaluation on test

According to the confusion matrix of Decision Tree evaluated on test set, there are 25204 examples predicted as delay and the accuracy is 56.3006. And there are 36546 examples predicted as on time, the accuracy is 61.5526. Therefore, the total accuracy is 59.4089. Besides, there are 28241 delayed records and 14190 of them are correctly predicted. The hit rate is 50.2461.

Conclusion

Based on the result from our experiments, BayesNet has best performance of flight delay prediction in both training and test set. Decision Tree has highest cross validation accuracy but not the best performance on test set. AdaBoost has a good performance in training and testing.

According to decision tree, the most important attributes that split the data are scheduled departure time, scheduled arrival time and carrier.
For AdaBoost, the accuracy of training set and test increases as the iteration goes up, then reaches a limit where training accuracy is capped by 64% and test accuracy is capped by 61%.
For Naive Bayes and Bayes Net, changes of parameters have little effect on the accuracy.

Future Work

1. Train the data in different locations. Pay special attention to big cities and major airports.
2. Expand the time period to the whole year.
3. Explore on more features that may affect the prediction.

Group Work Distribution

Tianyi Li: Data collection, data training, report

Ruifeng Jiang: Data processing, data training, report

Meidi Peng: Website building, data collection data training