Flight Quest: A better flight profile system for pilots

Ashwini Ramamurthy, Kishore Rajasekar, Ravindra Dangar
Computer and Information Science and Engineering Department
UF UNIVERSITY of FLORIDA
The Foundation for The Gator Nation

INTRODUCTION

Flights have become commonplace in our lives but the efficiency is not as good as it can be. Gate conflicts, operational challenges, air traffic management, dynamics of a flight can change quickly and lead to costly delays. Airlines spend around 22 billion dollars on improving the efficiency. Real time big data analysis is the solution to this problem. Pilots could augment their decision making process with this real time information given to them.

To reduce the delays and to make the decision making more accurate and efficient, we have come up with a module which will predict the delays in flight considering various factors. Large real time data related to flights, airports, weather are combined, analyzed and made into a model .Future predictions are made based on this model.

OBJECTIVES

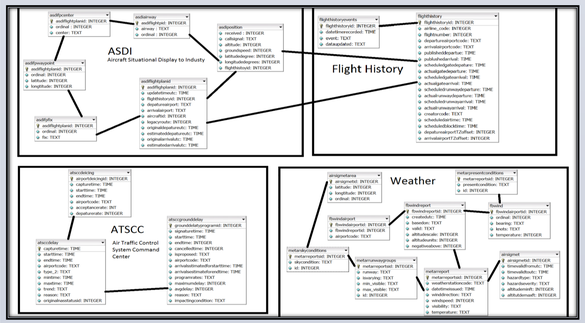
Main Objective

- Help pilots in decision making by providing real time business intelligence by predicting delays for a given flight plan.

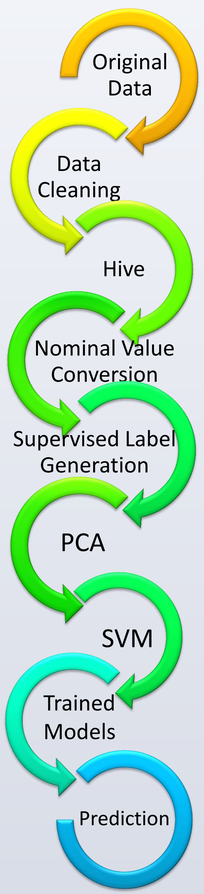
Secondary Objective

- Remove noise from the data.
- Collect and combine relevant flight data to build a model.
- Extract the important parameters that contribute to delays.
- Train the model by classifying the data into different classes defined by their delay time.

RELATIONAL MODEL OF DATA



DATA PIPELINE



- Data from FlightStats,Inc.
- U.S Department of Transportation's Volpe Transportation Center
- Weather data from METAR reporting system
- Google Refine to observe anomalies
- Rule based cleaning via a shell script
- Give a tabled structure to data
- Reducing / combining features to obtain significant macro-features.
- Combining relevant data from scattered data efficiently
- Mapping every unique string value to a unique numeric value, and reverse mapping to obtain results
- Classification into groups based on delay calculated from given scheduled and actual arrival time.
- Reducing the feature vector set to avoid overtraining of SVM
- 23 features reduced to 13 features with a threshold of 95%
- Multi-way classification using 11 separate models, using RBF kernel
- Data scaled and centered and convexity enforced using square transform.
- Input data is locally grouped, so input order is randomized
- Highly skewed dataset for fringe classes.
- Model Trained on 80% of input data
- Testing average accuracy of 67% of correct classification
- High error rates for extreme classes

CHALLENGES

- Understanding and cleaning the data
- Building a relational structure into the data
- A Robust mapping algorithm to handle numerous nominal values
- Transforming data based on PCA
- SVM not settling down due to non-convex nature of input data
- Fringe classes with <1% positive samples

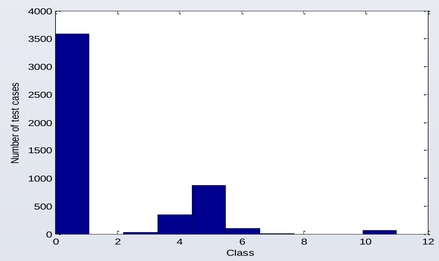
RESULTS

- Original data size : 211,532 samples and 35 features
- After cleaning data: 165,021 samples and 35 features
- After joining related data : 165,021 samples and 23 features
- After feature extraction using PCA : 165,021 samples and 13 features

Feature Extraction Results:

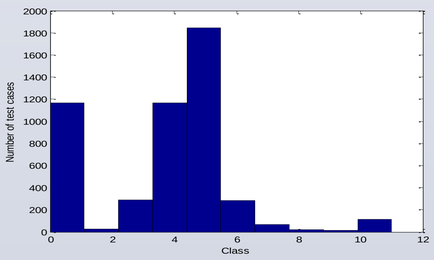
Input data : 165021 samples with 23 features
Output data: 13 features extracted with a threshold of 95%

Classification Results:



Input data
165021
samples
13 features
(20% Training
+ 80% Test)

- Accuracy was only 28.30%.
- The classification on test data was very poor, because training data was locally grouped (example: data specific to one airport was stored together)



Input data
165021
samples
13 features
(20% Training
+ 80% Test)

- Accuracy improved to 65.56%
- Random shuffling of the training sample to break the local grouping present in data.
- Fringe classes still have high inaccuracy rates due to highly skewed training data (example: class 1 had <1% positive samples)

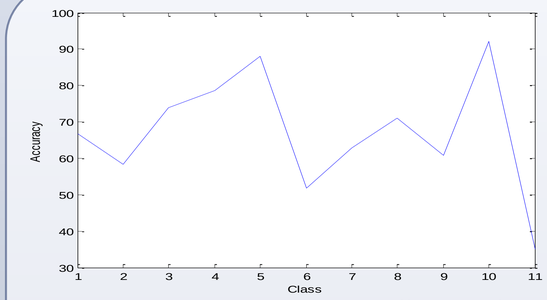


Figure : Accuracy variation among classes (final result)

- Improvements can be achieved by increasing training data size, as this will lead to increase in feature set and reduction in chances of over-fitting.
- Other possible routes can be calculated to reduce delay.

CONCLUSIONS

After collecting the flight related data over a long period of time and analyzing the data, we have constructed an approximate model to predict delays on particular route given the current air traffic, weather and other conditions.

This is a definite improvement over current systems and the real-time information produced is valuable in re-calculating routes for critical flights and to inform the passengers of an almost exact amount of delay if there is one.

REFERENCES

- https://www.gequest.com/c/flight/
- http://www.kaggle.com
- http://hive.apache.org/
- https://cwiki.apache.org/Hive/gettingstarted.html
- http://docs.aws.amazon.com/ElasticMapReduce/2009-03-31/GettingStartedGuide/Welcome.html
- http://www.mathworks.com/help/stats/princomp.html
- A tutorial on Principal Components Analysis Lindsay I Smith, February 26, 2002.
- http://www.mathworks.com/help/stats/svmclassify.html

ACKNOWLEDGEMENT

- Daisy Zhe Wang.
  Assistant Professor, University of Florida
- Christan Grant.
  Database Research Center, University of Florida