Predicting Flight Delays @ SFO

Applications of Data Mining

Andy Chen, Daniel Zhengyu Qin, Ted Samore

Data Set

- Source: Transtats; Bureau of Transportation Statistics Flight On-Time Performance
- 12 month period used for training (total of ~1,200,000 rows)
- December 2013 used as test set (~100,000 rows)
- Attributes included:
  - Carrier (nominal)
  - Destination (nominal)
  - Scheduled Time (numerical)
  - Delay Time (Class to predict; nominal)
  - Distance of Flight (numerical)

Pre-Processing

- All flights that met the follow criteria were pruned:
  - Cancelled/Diverted
  - non-SFO outbound

- Size of the pruned data sets:
  - Training: > 168,000 instances
  - Test: > 16,000 instances

Generating Models

- Models generated on Weka 3.7.10
  - Several methods used to generate models (C4.5, Random Forest, Logistic Regression, KNN, SVM, kStar, MLP, CART Tree)

- Most time consuming to build: CART Tree and kStar
- Fastest to build: Random Forest and C4.5 Tree

Model Evaluation

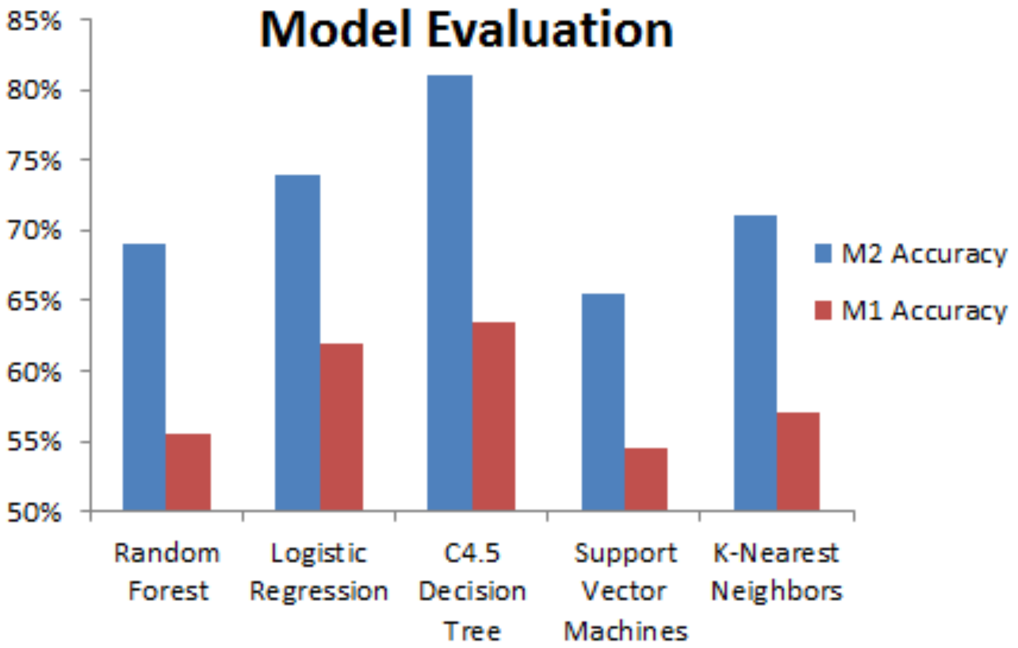
for $X,\ Y\ in\ \omega_1, \omega_2$ **do**
    $\alpha =$ X.DEP_DELAY;
    $\gamma =$ Y.DEP_DELAY;
    if $|\alpha - \gamma| \leq \tau_1$ **then**
        C1++;
    end
    if $|\alpha - \gamma| \leq \tau_2$ **then**
        C2++;
    else
end

$$M_1 = \frac{C_1}{|\omega|}$$

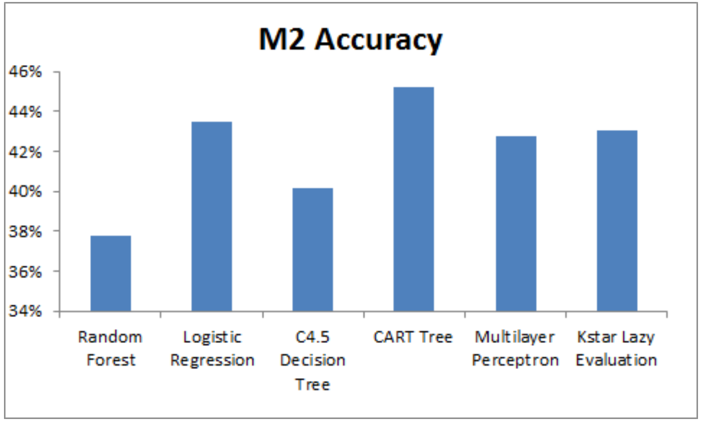
$$M_2 = \frac{C_2}{|\omega|}$$

$X =$ Each instance in the original test set
$Y =$ Each instance in the predicted test set
$\alpha =$ Original delay time
$\gamma =$ Predicted delay time
$\omega_1 =$ Original test set
$\omega_2 =$ Predicted test set
$\tau_1 =$ 3 minutes, first class tolerance
$\tau_2 =$ 5 minutes, second class tolerance

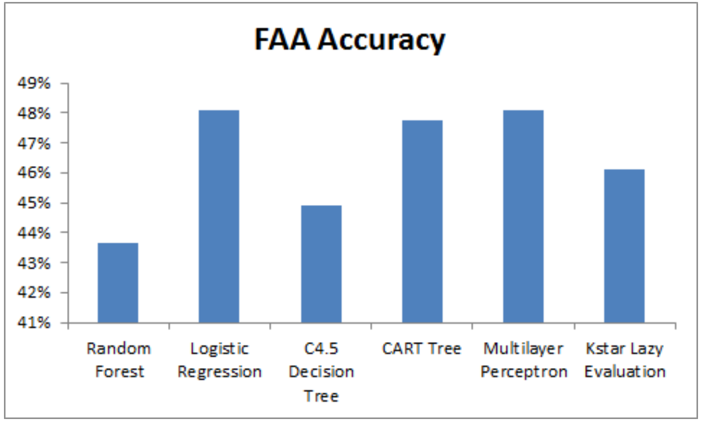
Model Evaluation



Test Set



M2 Accuracy
46%
44%
42%
40%
38%
36%
34%
Random Forest
Logistic Regression
C4.5 Decision Tree
CART Tree
Multilayer Perceptron
Kstar Lazy Evaluation



FAA Accuracy
49%
48%
47%
46%
45%
44%
43%
42%
41%
Random Forest
Logistic Regression
C4.5 Decision Tree
CART Tree
Multilayer Perceptron
Kstar Lazy Evaluation

Post-Processing

- Weka results parser written in Python 2.7

- This parser would feed in data into the model evaluation pipeline where it would then return us and metric on model accuracy.

Results

- Models performed worse on test data (compared to training)
- Sampled training set may have not been a good representative for Dec. 2013 sampled set.
- Models were most likely overfitted despite attempts to avoid them via reservoir sampling

Interesting Trends

- Regional airlines tend to have a much greater number of delays. Potential reasons :
  - operational procedures differ from those of larger aircraft.
  - contractual stipulations

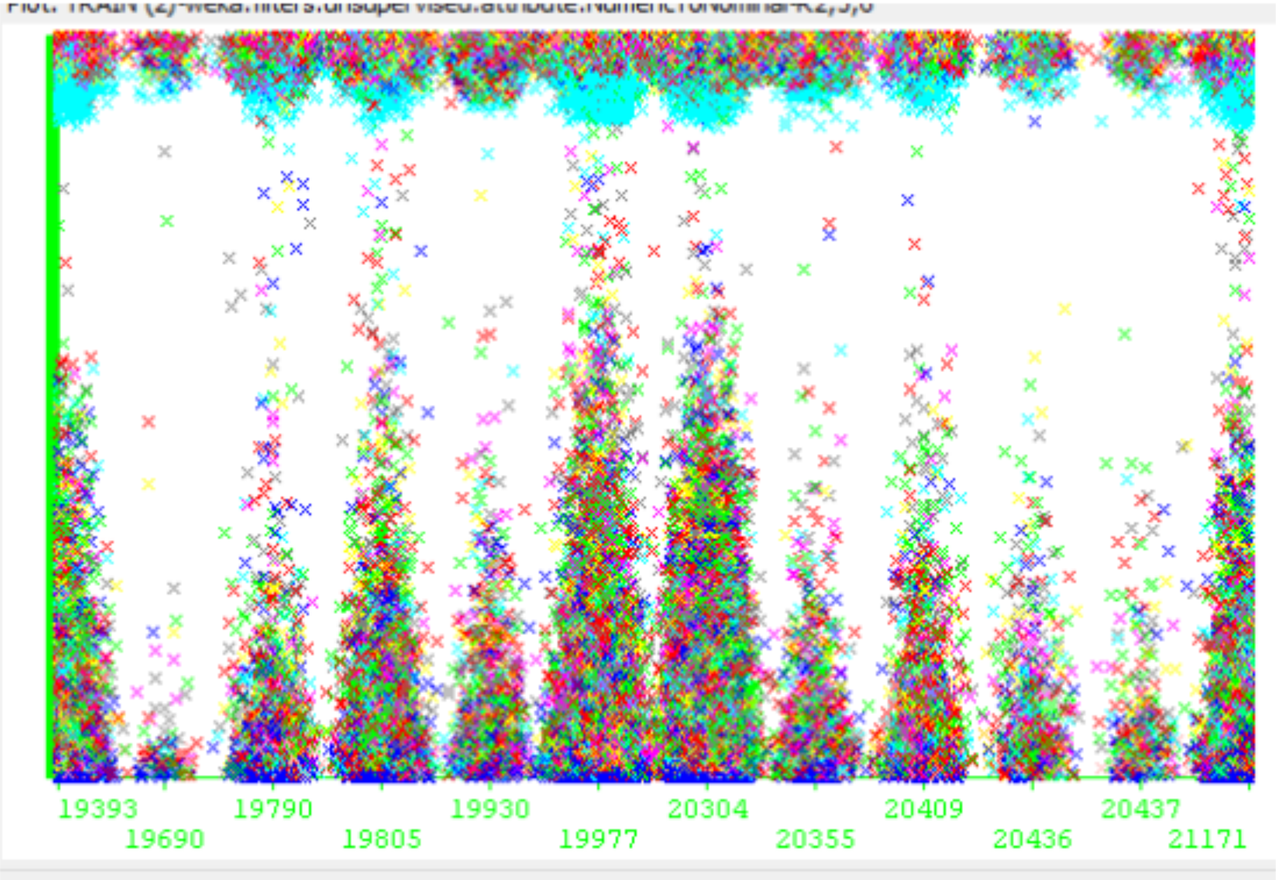
UNITED EXPRESS

AmericanEagle®

DELTA CONNECTION

- Worst legacy carrier (delays) : United Airlines
- Best legacy carrier (delays) : US Airways

U·S AIRWAYS


19393
19690
19790
19805
19930
19977
20304
20355
20409
20436
20437
21171

Interesting Trends (cont.)

- Best Value-Segment Carrier (delays): jetBlue AIRWAYS®
- Highest Potential for Longer Delays (Time) : 11:00-15:00
- Major hub airports tend to have more delays.

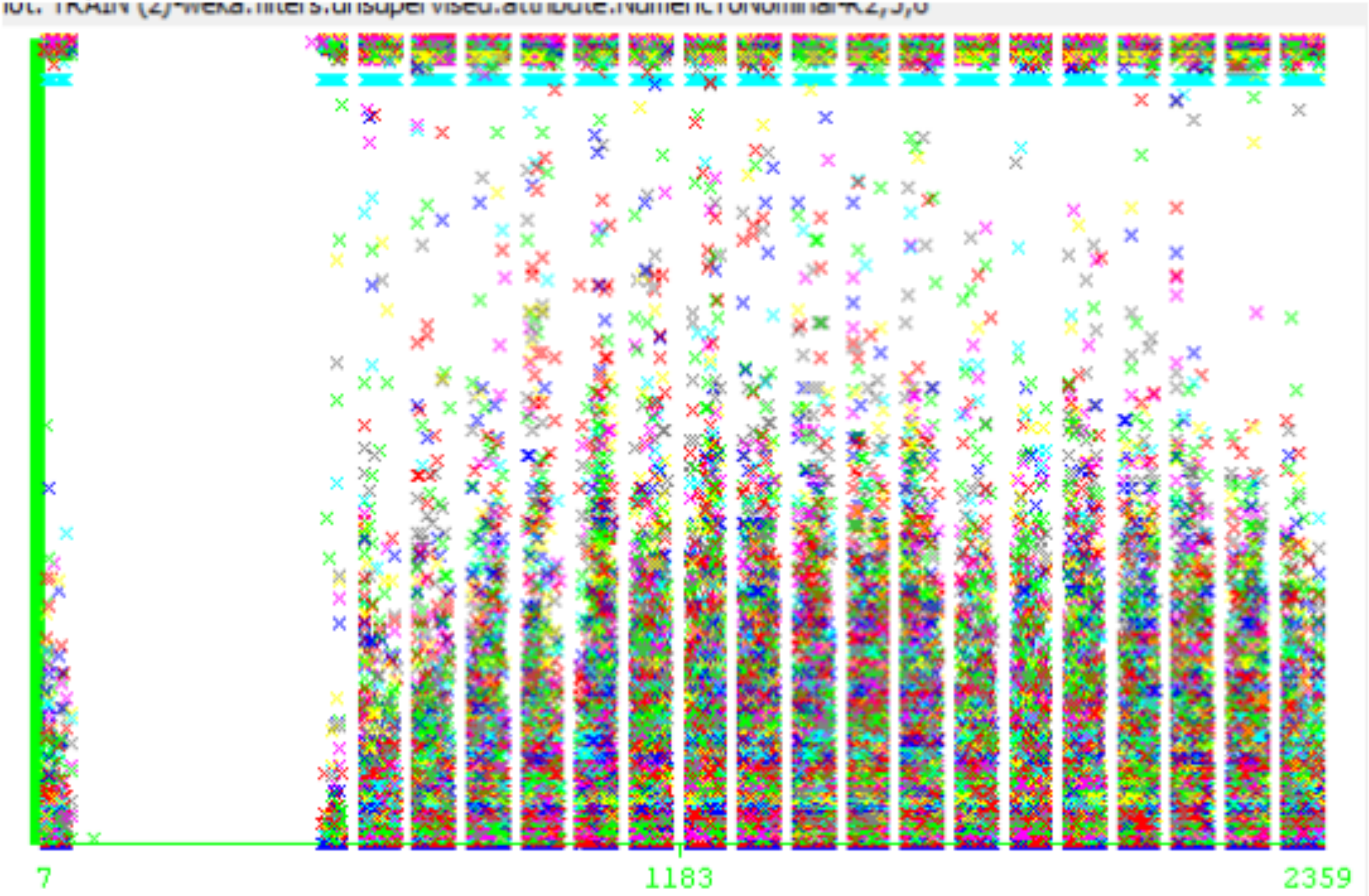
Hartsfield-Jackson Atlanta International Airport™

O'HARE INTERNATIONAL AIRPORT

DFW

DALLAS/FORT WORTH INTERNATIONAL AIRPORT

Delay vs Scheduled Times:



Demo

goo.gl/IEUYIQ

Lessons Learned

- Predicting flight delays remains to be a difficult problem.
  - Past performance is not always indicative of future performance.

- There is much room for improvement if data was expanded. (e.g including weather, ATC data, aircraft registration number, aircraft type.)