TTC Delay Classification

Elena Wang

December 11, 2018

Outline

- Problem Definition
- Literature Review
- Dataset
- Approach
- Results
- Summary
- Future Work

Problem: Determine whether TTC trains will be delayed or not

Delays have 2 major impacts:

1. Inconvenience to the passenger
2. Cost to the service provider

Literature Review

Railway Passenger Train Delay Prediction
By Yaghini & al

- Iranian Railway Dataset from 2006 to 2009
- Develop a highly accurate neural network for scheduling
- Artificial Neural Networks (ANN), Classification & Regression Trees (CART), Multinomial Logistic Regression

A Comparative Analysis of Models for Predicting Delays in Air Traffic Networks
By Gopalakrishnan and Balakrishnan

- Bureau of Transportation Statistics from 2011 to 2012
- Classification (Delay/No Delay) & Regression (Length of Plane Delay & Length of Airport Delay)
- Models: Markov Jump Linear System (MJLS), CART, ANN

Predicting Flight Delay Based on Multiple Linear Regression
By Yi Ding

- www.umetrip.com (flight tracking) November 3, 2015 to March 5, 2016
- Classification & Regression (Delay/No Delay, Length of Delay)
- Model: Multiple Linear Regression

Dataset

TTC Subway Delay Data

- CIty of Toronto Open Data Catalogue
- Updated Monthly
- Excel spreadsheet >> R Studio
- Date range for analysis: January 1st to December 31st, 2017
- 18,885 rows of data
- 10 initial attributes
  - 2 Quantitative
  - 8 Categorical
- Additionally used an associated data log for the delay codes

Field	Description
Date	Date (YYYY/MM/DD)
Time	Time (24h clock)
Day	Name of the day of the week
Station	TTC subway station name
Code	TTC delay code
Min Delay	Delay (in minutes) between trains
Min Gap	Time length (in minutes) between trains
Bound	Direction of train dependent on the line
Line	TTC subway line i.e. YU, BD, SHP, and SRT
Vehicle	TTC train number

Data Cleaning & Processing

Data Collection & Pre-Processing	Data Cleansing	Data Processing

Data Collection & Pre-Processing

- Consolidated data into one file containing data from 2017
- Data possibly has some fields where data was manually inputted

Data Cleansing

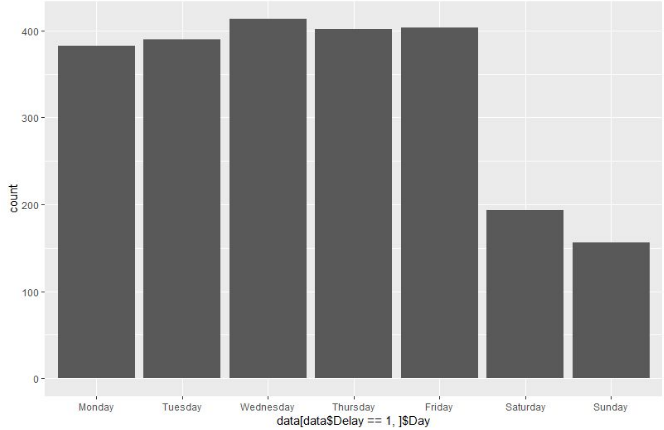
- All attributes were checked for completeness and consistency
  - Data formats
  - Missing values
  - e.g. Stations, Bound

Data Processing

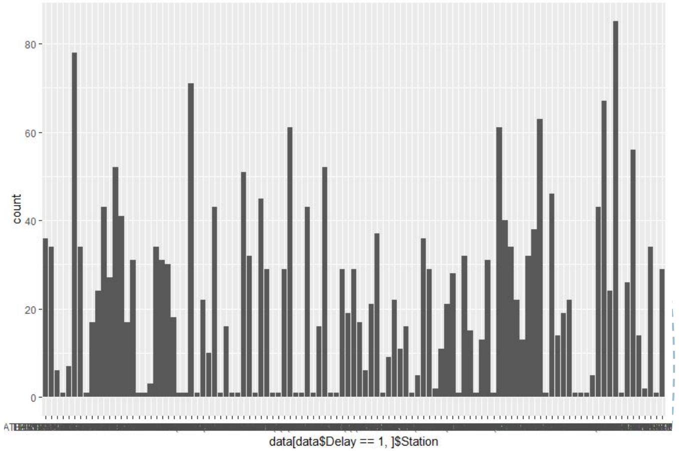
- Created additional category (Daypart) as buckets for day & time
- Created attribute for prediction class (delayed/not delayed) based on TTC Schedule

Exploratory Data Analysis

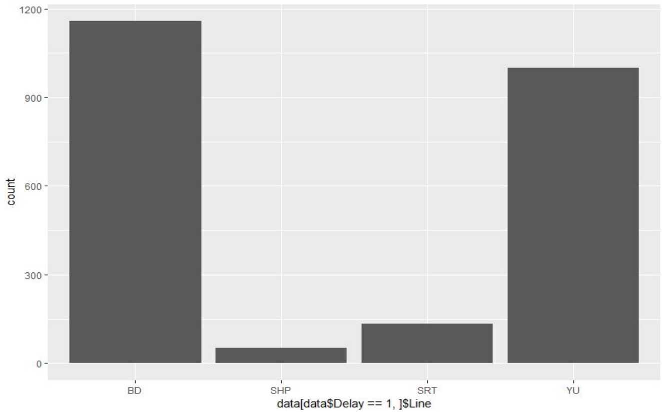
Day



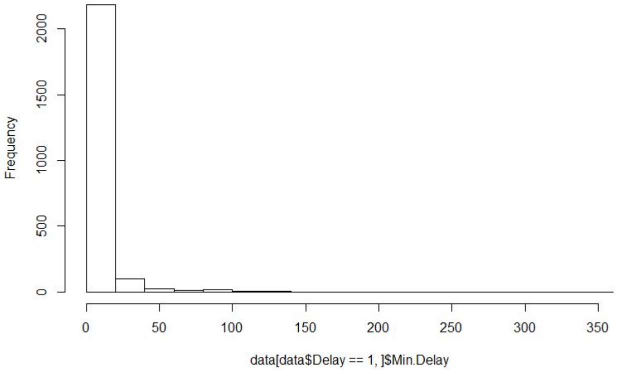
Station



Line

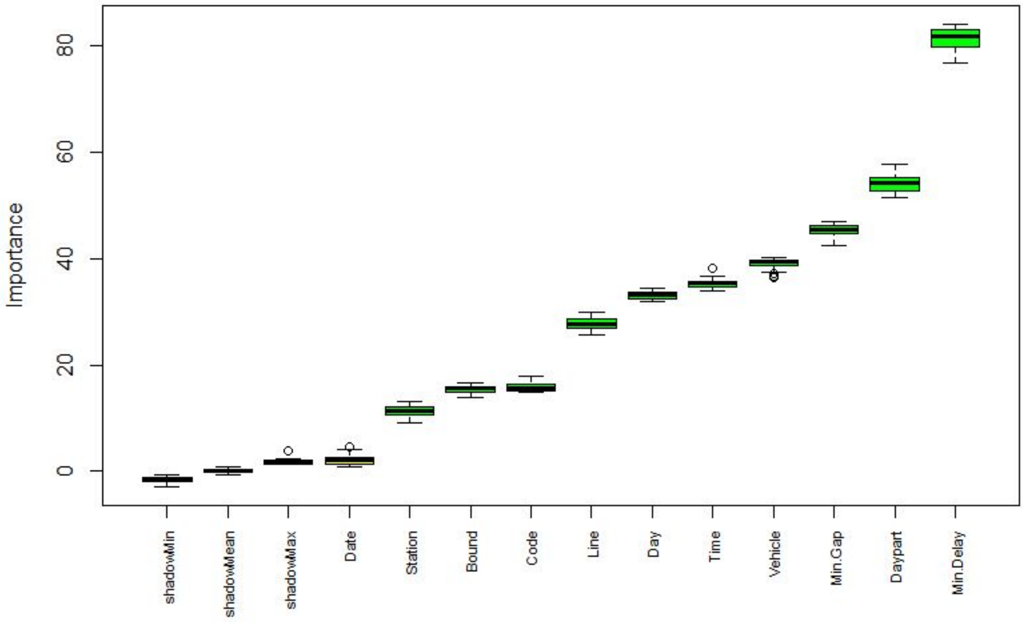


Min Delay



Feature Selection

- 80% Training (with Cross-Validation), 20% Testing
- All: Wrapper Method >> Boruta Algorithm
  - Random Forest
  - Variable Importance Measure (VIM)
- Numeric: Spearman Correlation
- Excluded date and Min Gap attributes



Modeling

Decision Tree

K-fold Cross Validation
with 5-folds

caret package
rpart & rpart.plot packages

- Takes categorical inputs
- Quick, simple, robust
- Handles messy data relatively well

Naive Bayes

K-fold CV with 5-folds
manually applied

e1071 package for Naive Bayes analysis

- Takes categorical inputs
- Quick, simple method
- Low training time

Logistic Regression

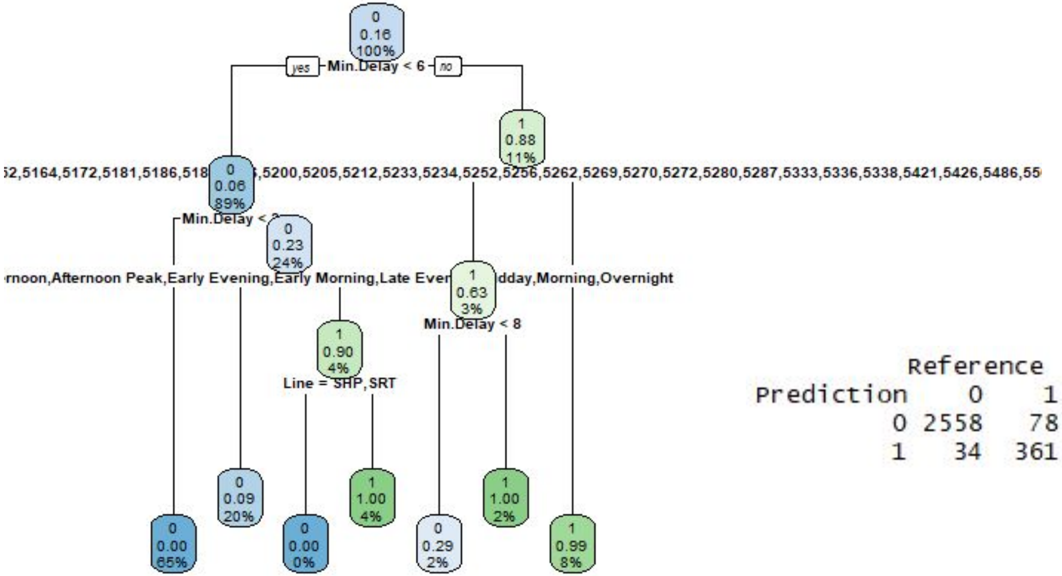
Forward selection
AIC Comparison

glm package

- Takes categorical inputs
- Quick, simple, robust

Decision Tree

- Used the Complexity Parameter (CP) from the CV stage to prune the decision tree
- Inputs: Minutes delayed, Line, Daypart, Vehicle



	Reference	
Prediction	0	1
0	2558	78
1	34	361

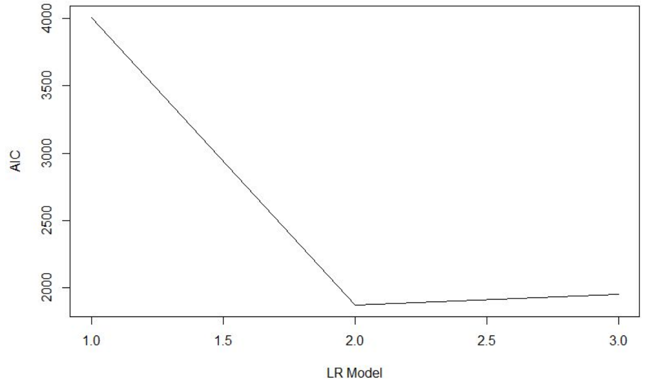
Naive Bayes

◎ Used all inputs; no additional feature selection/reduction

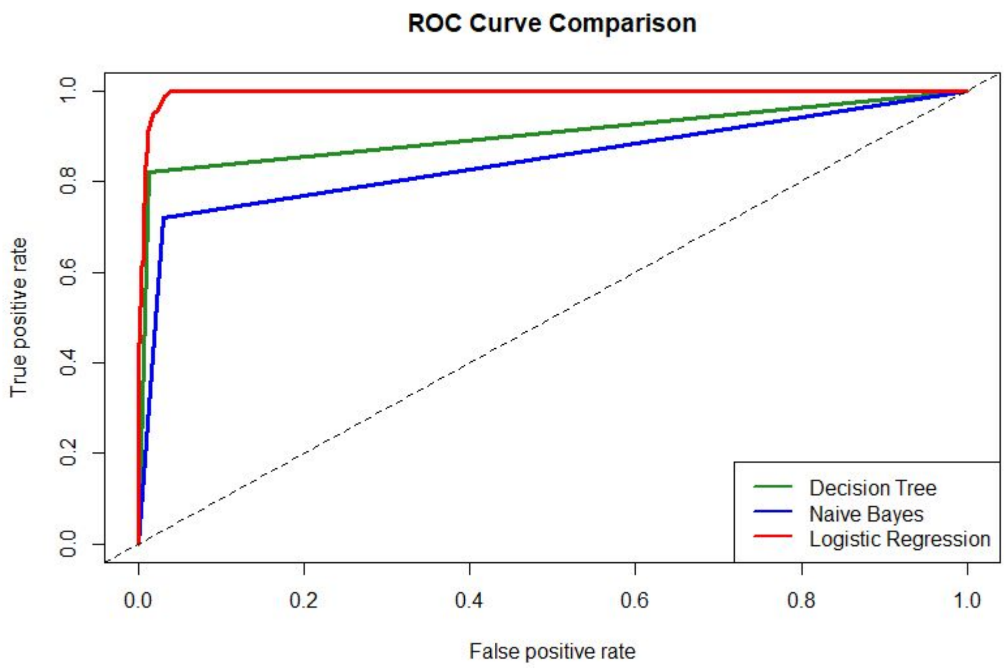
	Reference	
Prediction	0	1
0	2514	123
1	78	316

Logistic Regression

◎ Fit determined by lowest AIC
◎ Inputs: Min Delay, Daypart



	Reference	
Prediction	0	1
0	2561	41
1	31	398



Model Evaluation

Receiver Operating Characteristic

Model Evaluation

	Decision Tree	Naive Bayes	Logistic Regression
Accuracy	96	93	98
Precision	97	95	98
Recall	99	97	99
F1 Score	0.98	0.96	0.99

```
          Reference
Prediction    0    1
         0 2558   78
         1   34  361
```

```
          Reference
Prediction    0    1
         0 2514  123
         1   78  316
```

```
          Reference
Prediction    0    1
         0 2561   41
         1   31  398
```

Summary

- Large, messy dataset from the TTC
- 80%/20% Training-Testing, 5-fold CV
- Used Boruta algorithm and measured correlation for feature selection
- Created & compared three models
- The Logistic Regression model yielded the best results

Future Work

- Recommendation for standardization in the data inputs in order to refine model
- Additional regression analysis for time and/or location of delays

Thanks!

Any questions?