Predicting Flight Delays and Cancellations with Machine Learning

Andrew Wang, David Kawashima, Steven Brotz
CS 141: Hack Society

Dataset Specifications

- 5,819,079 domestic flights from 2015
- Final set of features:
  - *Time:* month, day of week, day of year, departure/arrival hour
  - *Flight specifics:* airline, origin/destination airport, distance of flight(miles)
  - *Weather:* temperature, dew point, humidity, wind speed, precipitation, altitude, visibility, sky forecast
  - *Airplane model*: certificate date of airplane, model of airplane, manufacturer

Initial Data Analysis

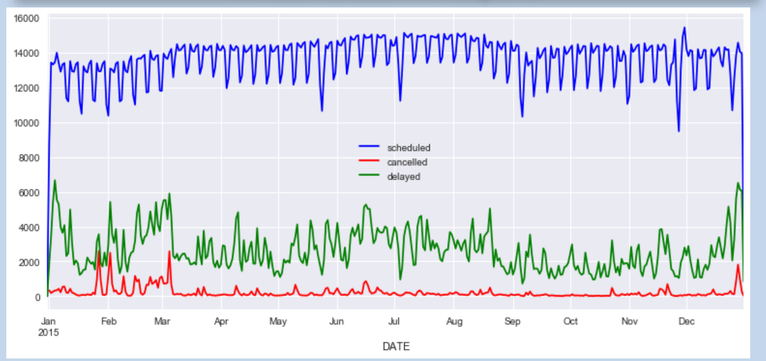

Figure 1: Scheduled, cancelled, and delayed flights throughout 2015. Interesting to note the spike in cancellations and delays during the winter months.

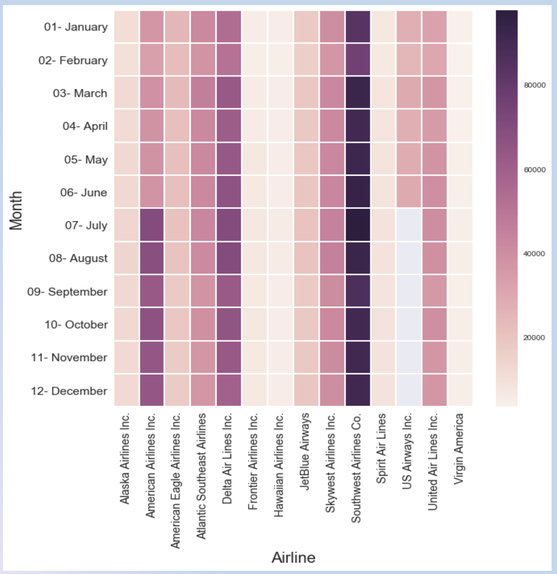

Figure 2: Heat map of scheduled flights per airline.

General statistics

- Mean arrival time: +4.36 min.
- For delays: 58.8 mins.
- 18% flights delayed
- 1.6% cancelled

Cancellation Classification

Top 5 Most Cancelled By Airport, Airline

- Origin Airports: SUN (8.88%), ASE (7.63%), MKG (6.83%), LAW, CMX
- Destination Airports: CMX (6.85%), MKG (6.56%), LAW (6.54%), TXK, DBQ
- Airlines: American Eagle/Envoy (5.11%), ExpressJet (2.66%), US Airways (2.07%), Spirit, SkyWest

Top Features: *Day of year, departure visibility/ wind speed/ air pressure/ temperature/ dew point, arrival temperature/ visibility/ air pressure, day of week*

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Model	**Classification**	**Precision**	**Recall**	Max Depth	Estimators	Balanced
Decision Tree	97.6%	24.23%	30.22%	None	1	False
Random Forest	98.55%	89.12%	5.2%	10	10	False
Random Forest	85.7%	7.03 %	68.81%	10	10	True
Random Forest	98.55%	90.36%	4.96%	10	50	False
Random Forest	85.83%	7.21%	70.15%	10	50	True

Figure 3: Models used to predict flight cancellations. Balanced: Weights inversely proportional to class frequencies. Other models, such as Naive Bayes, had suboptimal results.

Delay Classification

Top 5 Most Delayed by Airport, Airline

- Origin Airports: ASE (34%), PBG (33%), OTH (35%), BPT, LGA
- Destination Airports: ASE (36%), OTH (35%), BPT (32%), PBG, GUC
- Airlines: Spirit (29%), JetBlue (25%), American (22%), Alaska, Delta

Top Features: Distance, departure dew-point / temp / altitude / humidity / hour, arrival temp / altitude / humidity / hour

Model	**Classification**	**Precision**	**Recall**	Max Depth	Estimators	Balanced
Decision Tree	82.18%	n/a	0	None	10	False
Random Forest	83.43%	67.52%	13.64%	20	10	False
Random Forest	83.07%	58.56%	19.46%	30	10	False
Random Forest	83.13%	58.82%	18.66%	40	10	False
Naive Bayes	82.11%	40.97%	00.38%	n/a	10	False

Figure 4: Models used to predict whether or not a flight is delayed. Max Depth is a method of early stopping for decision trees, therefore n/a for Naive Bayes models. Precision and Recall ill-defined for Decision Tree model because it predicts all negative.

Arrival +/- Delay Regression

Features used: Airline, departure temperature/ humidity/ altitude/ visibility, arrival humidity, day of year, departure/arrival hour

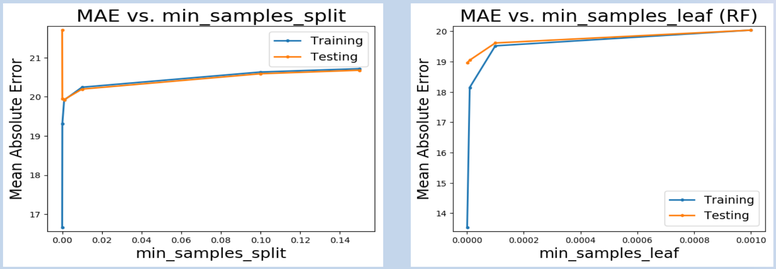

Figure 5: Varying parameters of single tree regressor and random forests (100 estimators). min_samples_split refers to the minimum number of samples required to split an internal node. min_samples_leaf refers to the minimum number of samples to be at a leaf node.

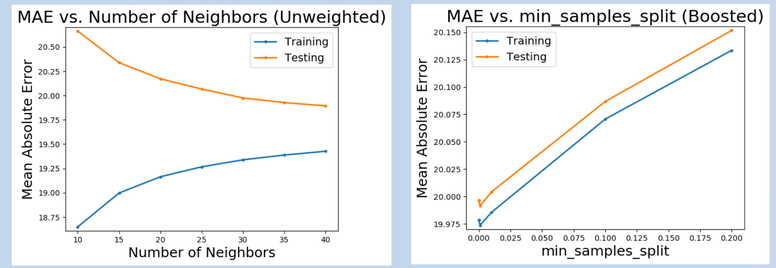

Figure 6: Results for using KNN and AdaBoost regressor (50 estimators). .

Summary of Results

- **For cancellation classification: the minimum classification error achieved was 1.45%.** Max. precision: 90.36%, recall: 70.15%
- **For delay classification, the minimum classification error achieved was 16.6%,** slightly less than the naive 17.8%. Max precision: 67.52%, recall: 19.46%.
- For regression, **the least mean absolute error achieved was 19.05 minutes.**

Current Work

- Further optimization of models
- Integrate international flight data into dataset
- Website and mobile application

Acknowledgements

- Claire Ralph, TAs, and fellow students of CS141 for guidance
- Data courtesy of:
  Kaggle, U.S. Department of Transportation
  Iowa Environmental Mesonet
  Federal Aviation Administration