Flight delay prediction

Tom Chau

Dataset description

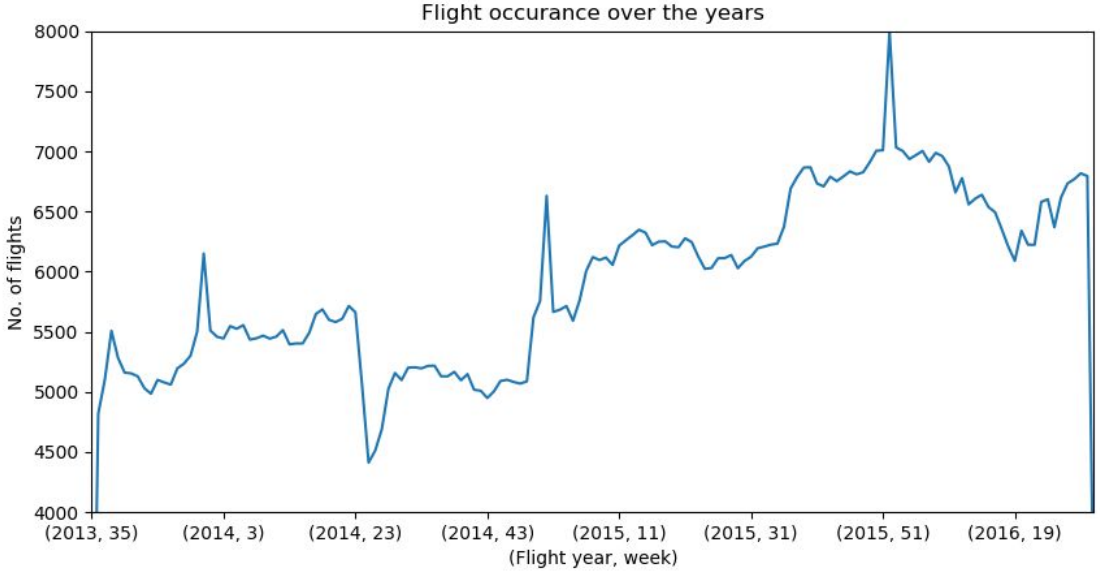
- Range: 2013/01 to 2016/07 with 899114 flight records
- All records are departed from HKG, and arrived to 163 different airports from 122 airlines
- is_claim only consists of two values in the dataset: 0 or 800
- When is_claim equals to 800, delay_time is either a real value >= 3.0, or marked as "Cancelled"
- **There are only 39413 flight records (i.e. 4.4% of all flights) that are claimed in the dataset.**

Data study

- Treat as time series, view metrics from different perspectives to see if any significance to denote potential flight delays/cancellations.
  - Perspectives
    - Departure
    - Arrival
    - Airline
  - Metrics
    - Delay time average
    - Delay count
    - Cancel count

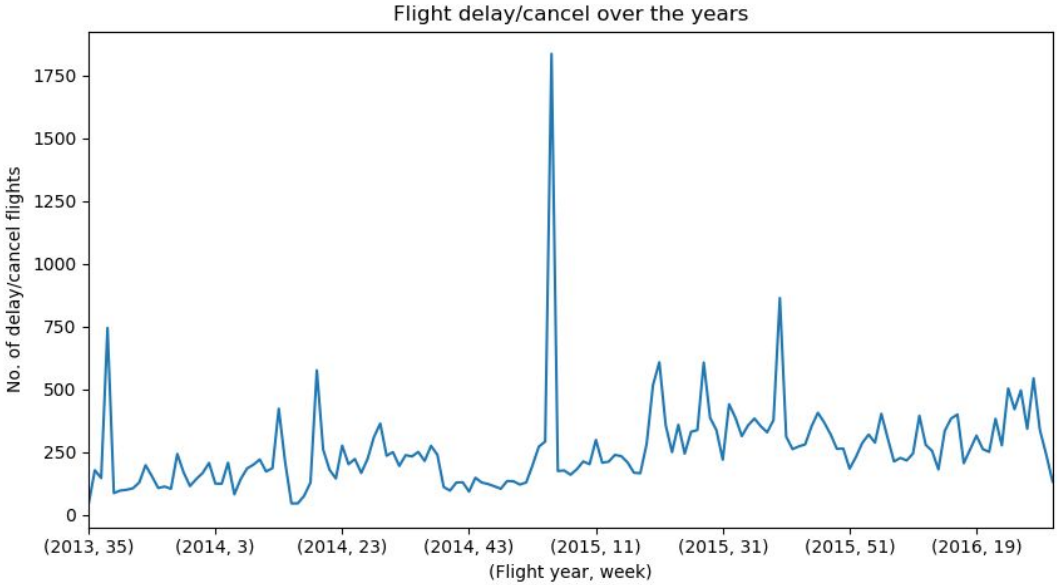
Data study

- Flight occurance trend
  - Non-stationary pattern, with seasonal effect (e.g. “Peaks” in holiday season) and increased number of flights over the years



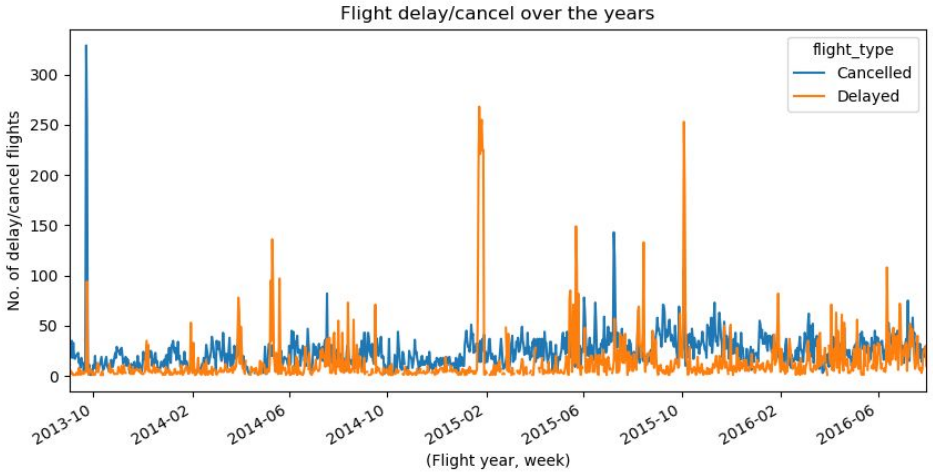
Data study

- Flight delay/cancel count trend
  - Looks more like residual in comparison to flight occurance trend
  - Multiple “spikes” appeared, those denoted high claim risk for any flights at those periods



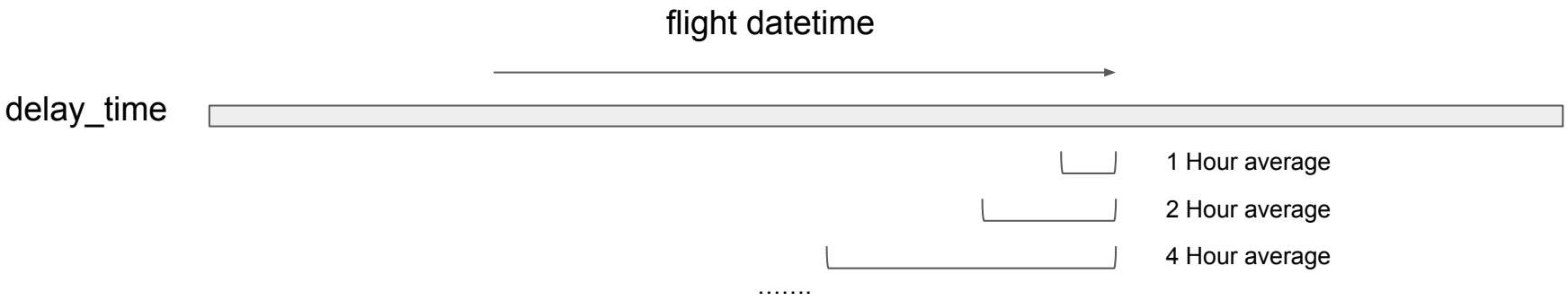
Data study

- Flight delay/cancel count trend
  - Detailed breakdown shows there are local effects for delay/cancel trend (i.e. trends to go up and down for delay/cancel count)
  - In the assessment, decided to go for feature engineering approach such that prediction model could determine potential flight claim by change in local delay/cancel trends



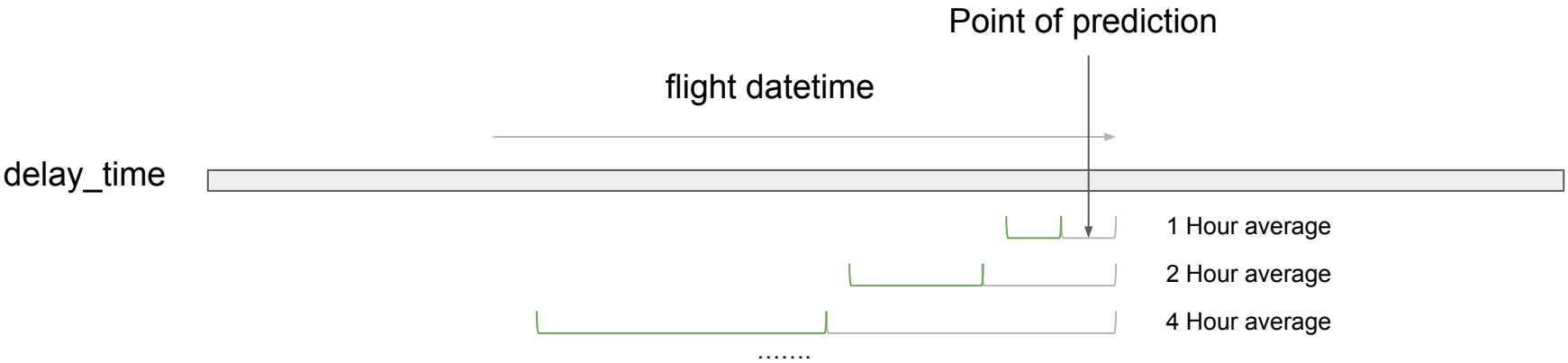
Feature engineering

- Using metrics of different perspectives found in data study
  - Departure/Arrival/Airline + delay time / delay count / cancel count
- Quantization of flight datetime into different time bins:
  - Hour/Day/Week
- Varied time bin periods as to capture local trend change:
  - e.g. 1 Hour / 2 Hour / 4 Hour



Feature engineering

- Upon prediction time, use last available time bin
  - … since current time bin statistics would not be available until the time bin period passes
  - Assumption: Historical flight records can be access with current prediction
  - Illustration below (Green denotes the time bin values used)



Model training

- ML technique used: Gradient Boosting
  - Justification: Produces good prediction model with relatively little tweaking (as opposed to neural network)
- Bagging (e.g. Random forest) could also be used, but not tested in current scope
- For the model, both classifier and regressor are tested for training
  - Justification: is_claim only appears as two values, thus could be treated as binary classification of needed to claim of not
- Compares naive approach (with only raw record columns used) and feature engineering approach

Model training

- Result (using classifier):

Approach	Naive	Feature engineering
Training set accuracy	0.965	0.998
Training set accuracy (on claimed case only)	0.262	0.955
Testing set accuracy	0.958	0.971
Testing set accuracy (on claimed case only)	0.180	0.425
Training set Q1 error	27.717	1.570
Training set Q2 error	22173.984	1255.841
Testing set Q1 error	33.657	23.392
Testing set Q2 error	26925.390	18713.478

Model training

- Result (using regressor):

Approach	Naive	Feature engineering
Training set Q1 error	44.232	16.187
Training set Q2 error	16876.900	1628.388
Testing set Q1 error	53.125	44.611
Testing set Q2 error	23500.604	15465.921

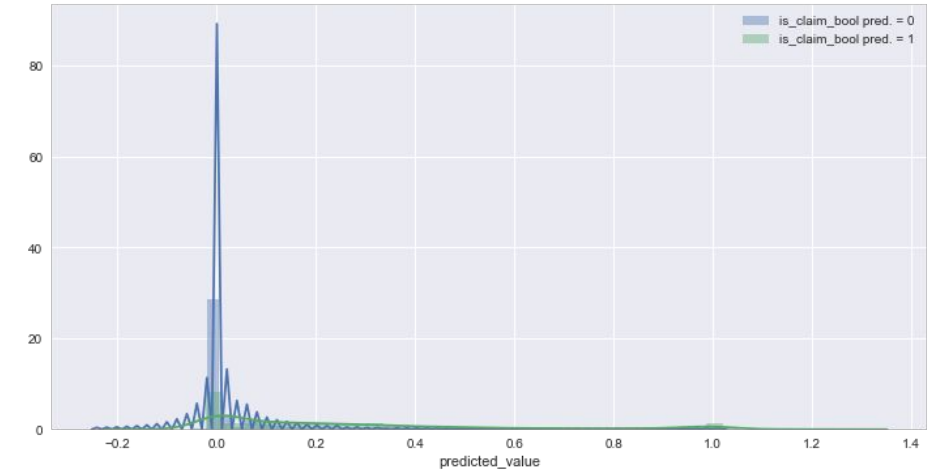
Model training

- Discovery from model training:
  - Highly-skewed dataset easily mis-interpret the prediction model's accuracy
    - Only 4.4% of the whole dataset would have flight claims
    - We can achieve more than 95% accuracy even if we just predict all flight records with no claim needed, which is unreasonable
  - Feature engineering approach helps with model's predictive power
    - With classifier-based model, it showed a jump from 18% to 42.5% accuracy upon flight claim case
    - Still have rooms of improvement for better accuracy
  - Feature engineering approach shows overfitting with training dataset
    - Per tested, the testing set accuracy does not show improvement with less num_leaves
  - The model has Q2 error that is significantly larger for all models. This denotes high variance of predicting wrong values, which is quite match with the classifier+accuracy view.

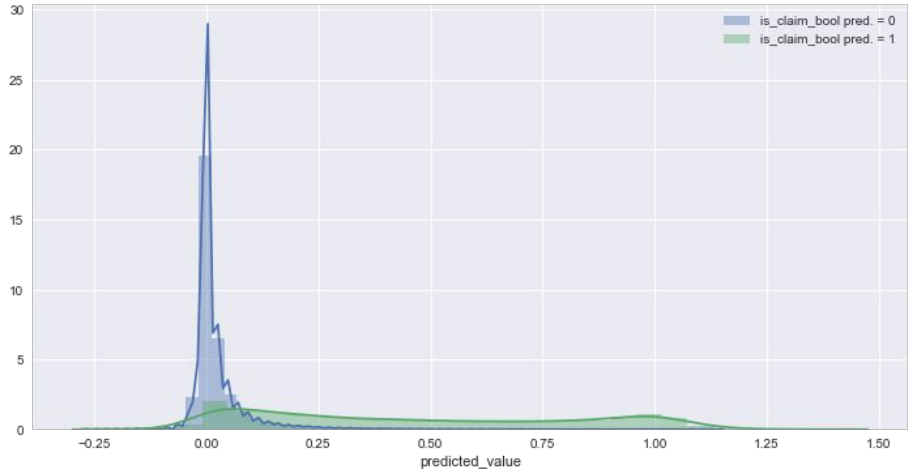
Model training

- Discovery from model training:
  - Shift of predicted is_claim probability distribution in naive/feature engineering approach using classifier model

Naive



Feature engineering



Thoughts of next steps

- Better strategy for preparing training data
  - Avoid model biasness on non-claimed flight records
- Use rolling average time periods instead of quantized time bins
  - Using last time bin might not well reflect metrics of to-be-predicted flight record
  - Needs engineering effort to generate rolling average per flight record in the dataset
- Use of sequence model
  - e.g. Recurrent Neural Network to transform sequences of metrics into high-dimensional context vector, that can be followed by a deep neural network for classification/regression on flight claim amount.
  - Requires much more effort for deducing such model; including modelling toy example; network structure and parameter tuning, etc.

Thoughts of next steps

- Linkage with external features
  - Hard to collect all features that infers the decision on flight claims, and the feature might not be a strong indicator of the decision
  - e.g. Hong Kong typhoon effect on flights
    - “Even when the No. 8 signal is in force, if the winds were blowing in nearly the same direction of the runways, they generally have a minor impact on aircraft operations. On the other hand, if the winds were blowing across the runways, particularly the southeasterly winds through the mountain gaps of Lantau Island, even though the No. 3 signal has replaced the No. 8 signal for the departing storm, the crosswinds may still exceed the operating limit of aircraft, and flights would be delayed or cancelled. Apart from meteorological factors, airlines may also change their flight schedules due to other factors.”
    - Reference: https://www.hko.gov.hk/education/article_e.htm?title=ele_00459