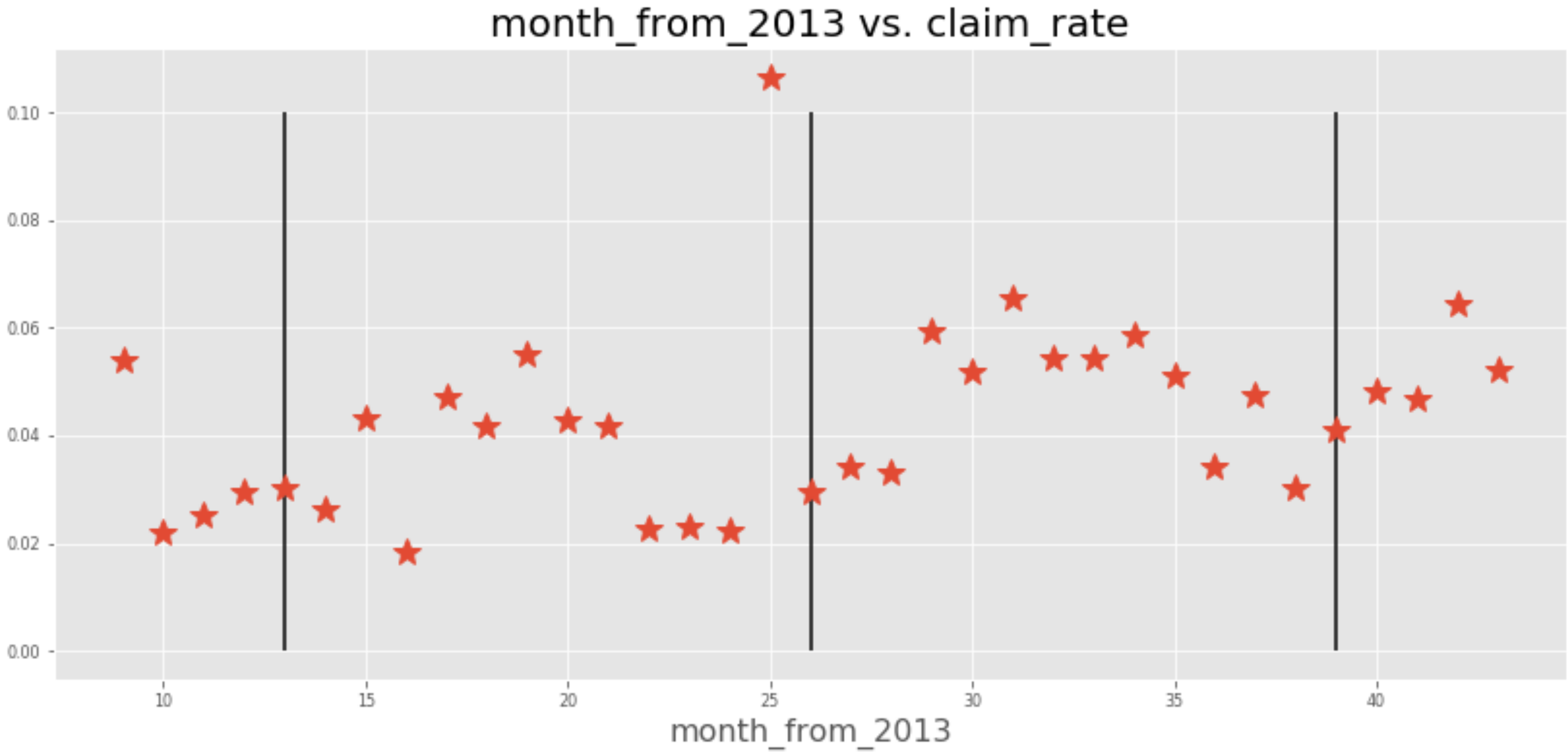
Contents:

Add Data: airport information

- Mostly Geography information.

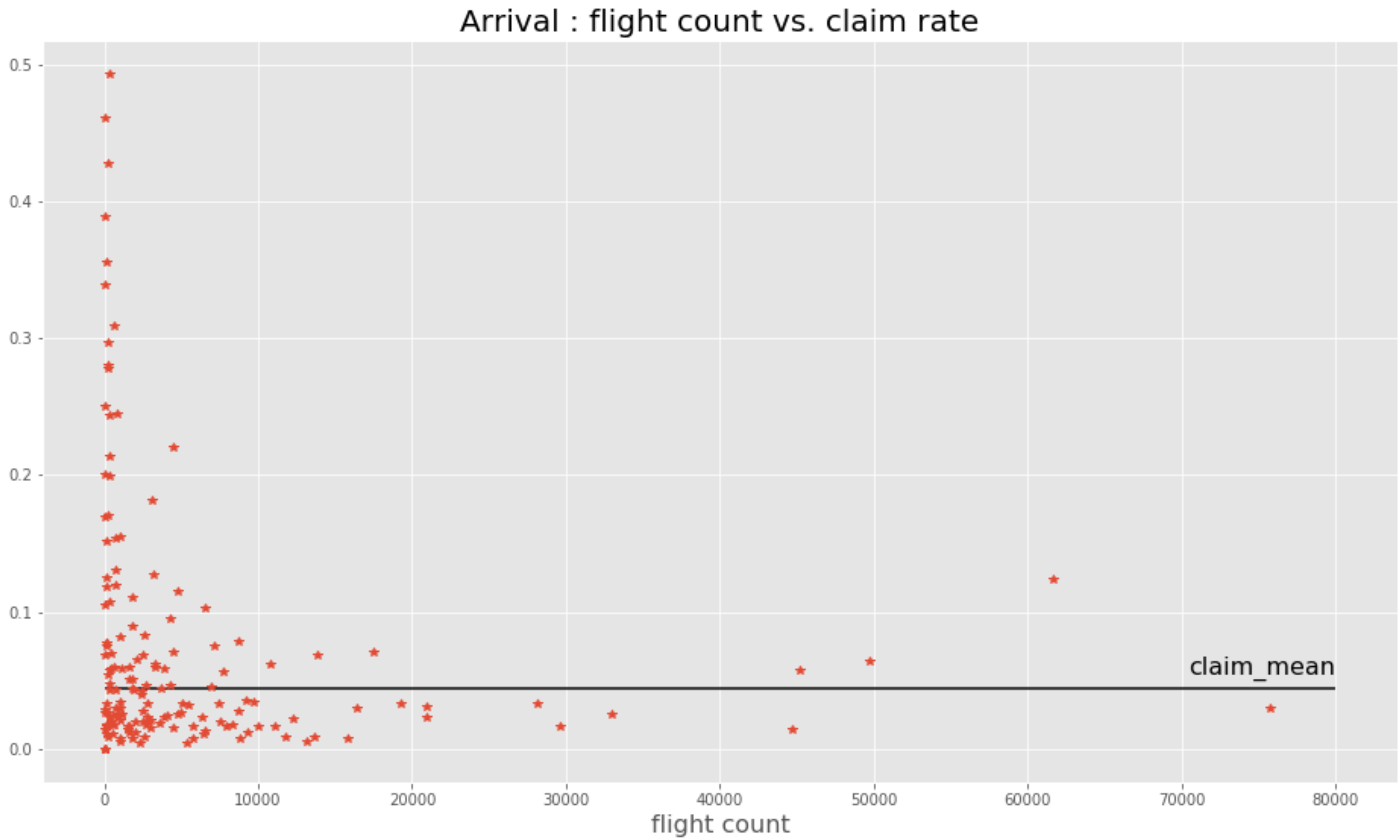
Exploratory Data Analysis

- Most numerical variable doesn’t have linear or quadratic tendency.
- For category variable, number of certain category might be useful.



Exploratory Data Analysis

- Most numerical variable doesn’t have linear or quadratic tendency.
- For category variable, number of certain category might be useful.



Feature engineering

- Geography:
  - Altitude, latitude, longitude information.
  - Distance from one airport to HKG(haversine) is highly correlated to **timezone difference**.

    -> Only use timezone difference since calculation is much cheaper.
- Time/Datetime:
  - Add weekday/week/year/month, all treat as category variable.

Feature engineering

- Encoding:
  - Most category variable have more than 10 category (up to thousand).
    -> Apply **Target Encoding** and **Count Encoding** instead of One-Hot encoding.
  - Encoding numeric variables too.
    -> For float variables, split to 10 group by 11 quantiles.

Modeling : Select Model

Here I choose Random Forest

- Fast/ Strong /Robust
- Easy to interpret
- If proven to be useful, we can further train boosting model to improve acc.

Modeling : Up-sampling

Tradeoff between accuracy (precision) and recall.

- Metrics based on 9/1 validation set.
- choose proper sampling rate based on business objective.
- Here I choose '0.3' for a reasonable recall and acc.

Positive/Negative rate	Acc	Prec	Recall	ROC_AUC
No sampling	0.956	Nan(all predict 0)	0	0.5
0.1	**0.957**	1	0	0.5
0.2	0.955	0.434	0.11	0.553
0.3	0.903	0.202	0.422	0.673
0.4	0.84	0.155	0.601	0.729
0.5	0.770	0.125	**0.723**	**0.747**

Modeling : Grid Search

After up-sampling, further use grid search to decide model parameters.

- Grid search on **number of trees** and **max tree depth**.
- Best one is number of tree = 200 and max tree depth =5

Positive/Negative rate	Acc	Prec	Recall	ROC_AUC
Model	0.896	0.203	0.479	0.670

Modeling : Feature Importance

Following are most important features.

- Based on this result. Target and count encoding are proved to be useful.

Features	Feature Importance
flight_no (target encoding)	0.388
flight_no (count encoding)	0.119
Arrival (target encoding)	0.112
Airline (target encoding)	0.106
tz (target encoding)	0.045
tz (count encoding)	0.037
longitude_q10 (target encoding)	0.023
Altitude (as numeric)	0.022

Conclusion

We train a Random Forest model to get a strong baseline model.
To further improve model, we can:

- Dataset:
  - Add weather data for airport.
- Features Engineer:
  - Try interaction variables on category variables (e.g. airline x flight_no)
  - get dummy variable from variables with many category.
    -> e.g. select only 5 category from a variable to do one-hot encoding.
- Modeling:
  - Ensemble: use two or more model on different target.
    -> create a model predict 'cancelled' ones.
    Then predict whether claimed on those predicted not-cancelled
  - Choose more powerful model : boosting ones.