Flight Delay Predictions

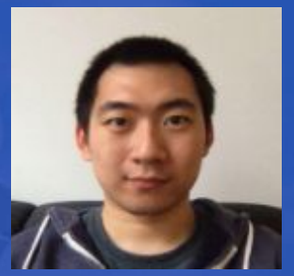
w261 Machine Learning at Scale – Final Presentation

TEAM 5

Esther Chen

Trevor Johnson

Michelle Shen



Yi Zhang

Overview

I. Business Case

II. Evaluation Metrics

III. Dataset and Join

IV. EDA

V. Feature Engineering

VI. Algorithms

VII. Final Prediction Pipeline

VIII. Project Leaderboard and Gap Analysis

IX. Performance and Scalability

X. Limitations, Challenges, Future Work

I. Business Case

Problem:

Airline executive reports about unforeseen flight delays with negative downstream impacts

Objective:

Create a predictive machine learning classifier to give adequate advance notice to airline and airport staff of potential flight delays

Primary Stakeholders:

Airline executives

Secondary Stakeholders:

Airline/airport staff

II. Evaluation Metrics and Cross-Validation

$$F_2 = \frac{5 \cdot Precision \cdot Recall}{4 \cdot Precision + Recall}$$

5-fold rolling CV

Fold	Training Set	Development (Test) Set	Test Set
1	2015 Q1 to 2017 Q3	2017 Q4	---
2	2015 Q2 to 2017 Q4	2018 Q1	---
3	2015 Q3 to 2018 Q1	2018 Q2	---
4	2015 Q4 to 2018 Q2	2018 Q3	---
5	2016 Q1 to 2018 Q3	2018 Q4	---
Final evaluation	2015 Q1 to 2018 Q4	---	2019 Q1 to 2019 Q4

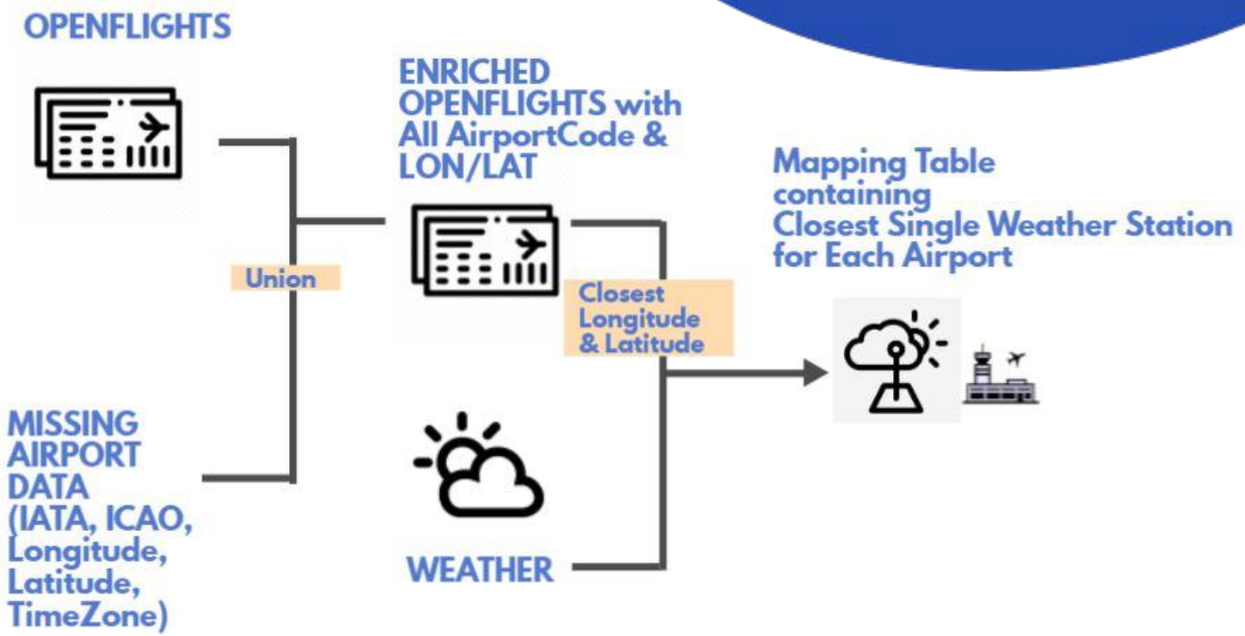
III. Datasets and Join

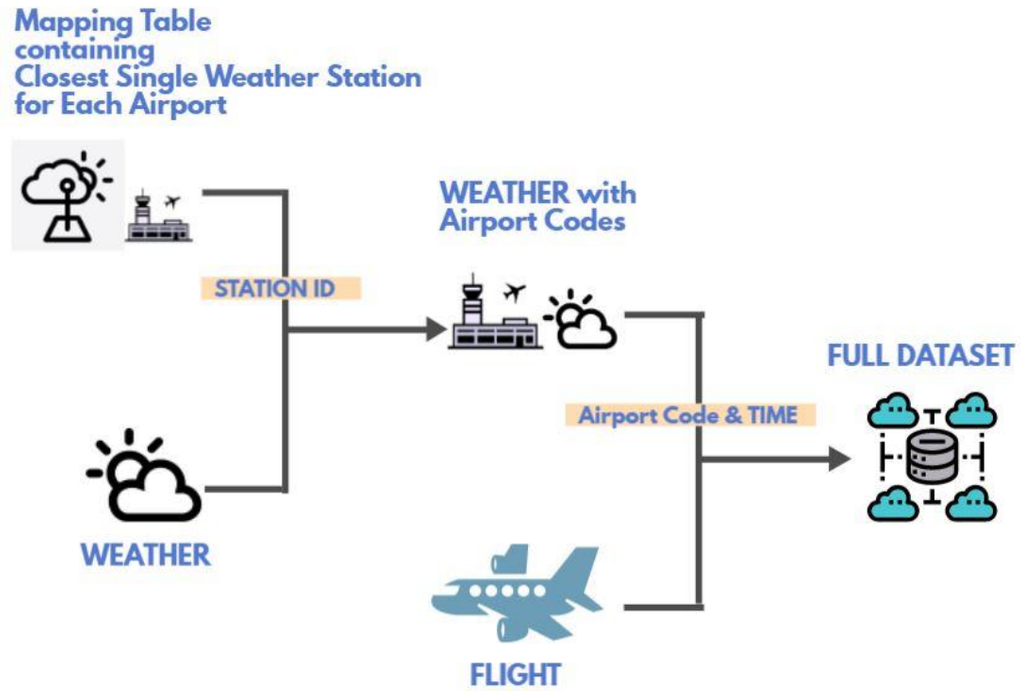
Primary Data Sources (provided):

- Flights: 109 columns (2015 to 2019)
- Weather: 177 columns (2015 to 2019)

Supplemental dataset:

- Airport data from openflights.org consisting of 14 columns that helps connect our airlines dataset to weather stations





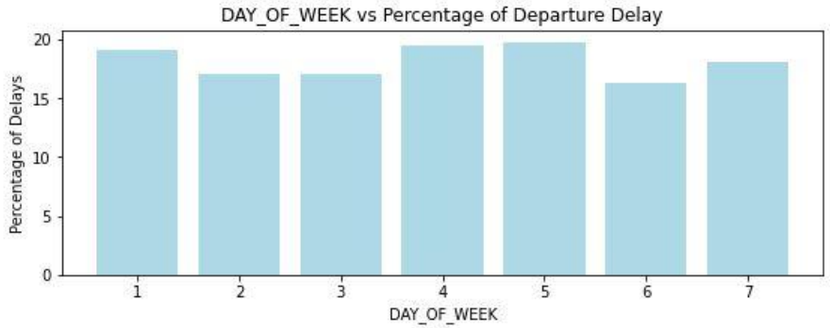
IV. EDA - Flight Dataset

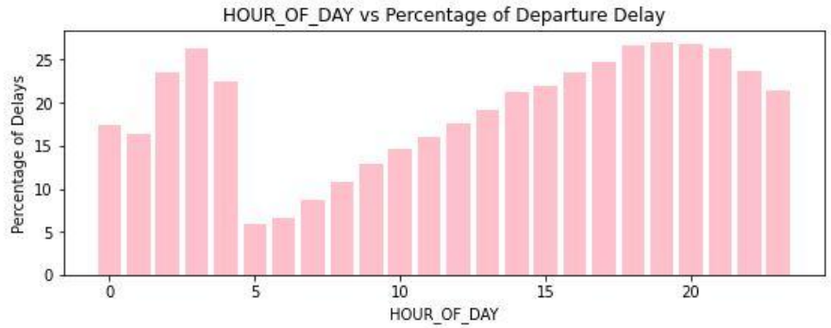
Property	Data
Flight count in total	63493682
Flight delayed by 15 minutes or more	11387082
% delayed flight	17.93%
% on-time flight	80.56%
% cancelled flight	1.54%
% diverted flight	0.25%
Duplicate records	31746841 (50%)

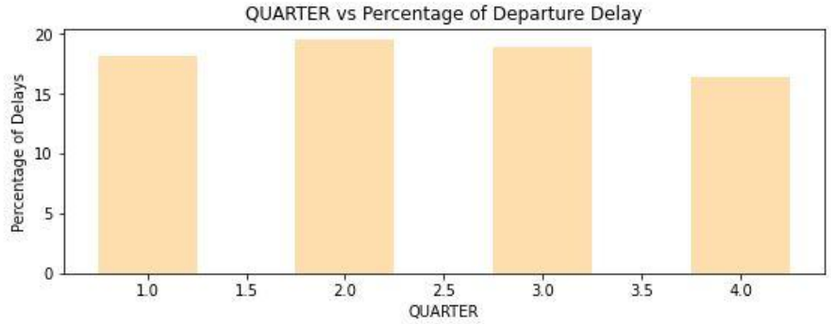
Factors that can cause Flight Delays:

- Seasonality and Time-Related Factors
  - hour of the day, day of the week, month of year, holiday
- Airline Carrier Factor
- Flight Distance Factor
- Airport Factor

- Class Imbalance
- Diverted Flights
- Cancelled Flights



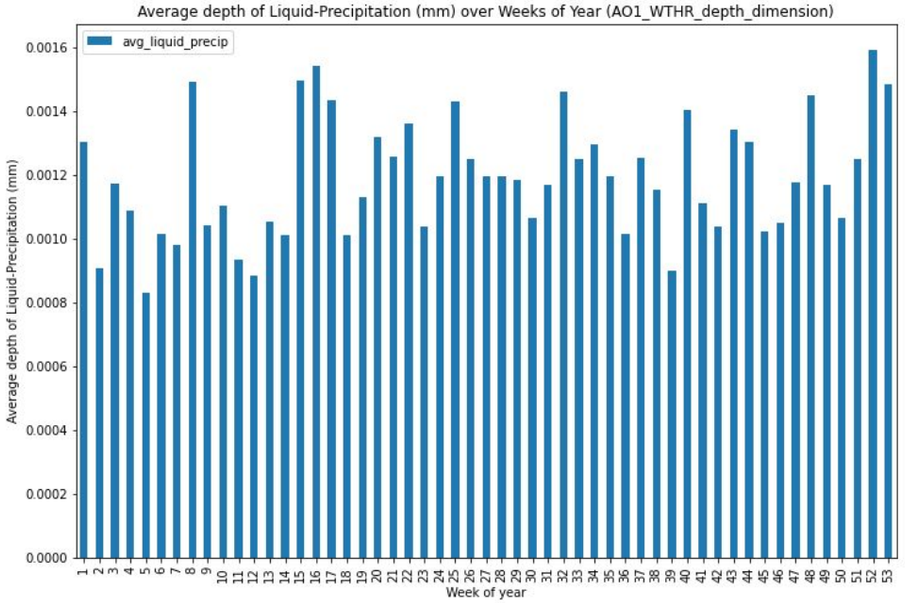


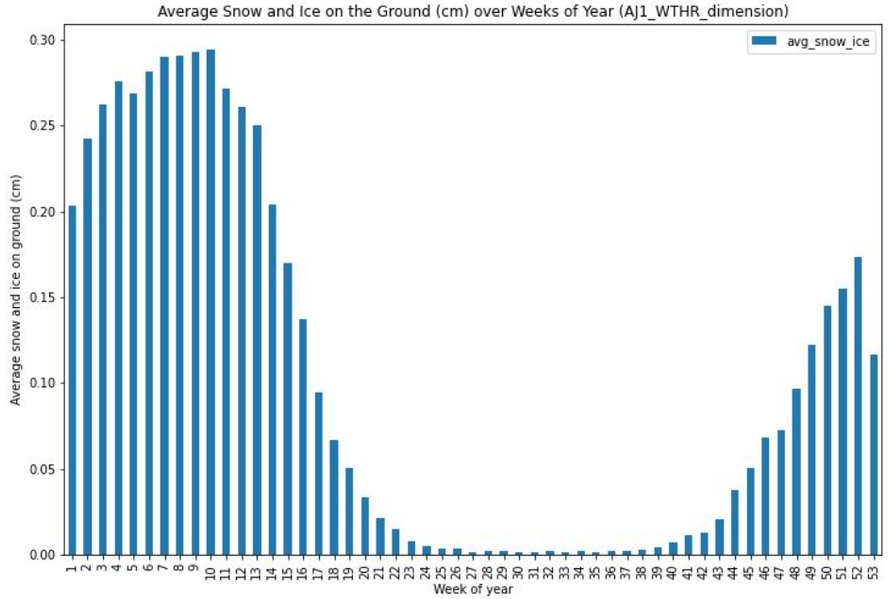


IV. EDA - Weather

Variable Name	Percentage of Data Filled	Variable Name	Percentage of Data Filled
WND	100.00	GF1	57.27
CIG	100.00	GA1	50.67
VIS	100.00	AA1	23.32
TMP	100.00	A01	12.74
DEW	100.00	AU1	4.74
SLP	100.00	AA2	3.15
MA1	78.95	AJ1	1.16
GF1	57.27	AT1	0.32

Property	Data
Total Records	630904436
Variable Count	177





V. Feature Engineering

Flight Dataset - Derived Features:

- Extraction of Local Departure Hour
- Holiday Indicator
- Previous Flight Delay & Other Related Features
  - Previous Flight Delay 15 mins or more
  - Enough Time between Estimate Arrival Time of Previous-Flight and Planned Departure Time of Current Flight
  - Poor Scheduling Indicator (Check if Scheduled Arrival Time of Previous-Flight is later than Scheduled Departure Time of Current-Flight)

Weather Dataset:

- Unwrapping weather sub-variables
- Arbitrary imputation of numerical variables
- One hot-encoding categorical variables

'WND_WTHR'
'190,1,N,0015,1'

→

WND_WTHR_direction_angle	WND_WTHR_direction_quality_code	WND_WTHR_type_code	WND_WTHR_speed_rate	WND_WTHR_speed_quality_code
190.0	'1'	'N'	15.0	'1'

WND_WTHR_type_code
'N'
'R'
'Q'

→

WND_WTHR_type_code_N	WND_WTHR_type_code_R	WND_WTHR_type_code_Q
1	0	0
0	1	0
0	0	1

VI. Algorithms

Baseline Logistic Regression

Refined Logistic Regression

Random Forest & XGBoost

1. Baseline Logistic Regression

$$p(x) = \frac{1}{1 + e^{-\beta x}} = \frac{e^{\beta x}}{e^{\beta x} + 1}$$

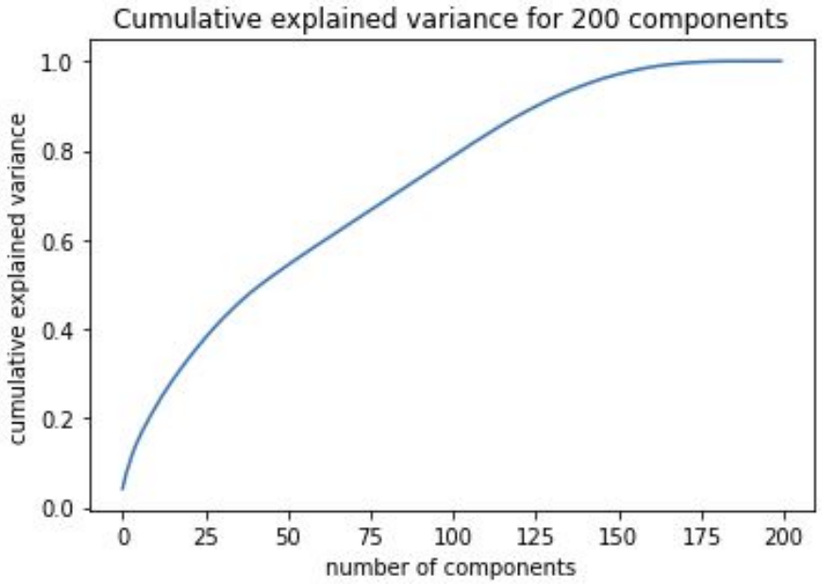
```
df_train_6 = df_train_5 \
    .select(['features_transformed', 'DEP_DEL15_AIRLNS']) \
    .withColumnRenamed("DEP_DEL15_AIRLNS", "label") \
    .withColumnRenamed("features_transformed", "features")

lr = LogisticRegression(maxIter=15, regParam=0.05, elasticNetParam=0.05)
lrModel = lr.fit(df_train_6)
```

- Under-sample majority class
- Preliminary PCA, limited feature space to 40 (9.5 min)
- Trained regression model (6.4 min)
- Evaluated model on hold-out test set (3.1 min)
- No hyper-parameter tuning

- F-2 score: 0.613
- Precision: 61.39%
- Recall: 61.33%
- Overall accuracy: 61.33%

1.1 Refined Logistic Regression

- MLLib string indexer and one-hot-encoder: added in categorical and derived features prepared from feature engineering process
- PCA: 125 features
- Performed more thorough hyper-parameter tuning on the model, e.g. regularization parameters (4 hours)
- Experimented with oversampling minority class



- F-2 score: 0.702
- Precision: 70.64%
- Recall: 70.31%
- Overall accuracy: 70.31%

```
final_params = {'maxIter': 25, 'regParam': 0.1, 'elasticNetParam': 0.0}

final_train = final_train.drop('FL_DATE_AIRLNS')
final_test = final_test.drop('FL_DATE_AIRLNS')

f2, final_pca, final_model, predictions = train_and_predict(
  oversample(final_train, over_sample_ratio=2.0, label_col='label'),
  final_test,
  final_params,
  return_model=True
)
```

2 & 3. Random Forest & XGBoost

Random Forest

```
RandomForestClassifier(
  featuresCol="features",
  labelCol=”DEP_DEL15_AIRLNS”,
  maxBins=370,
  maxDepth=12,
  numTrees=50,
  featureSubsetStrategy='sqrt',
  subsamplingRate=.8)
```

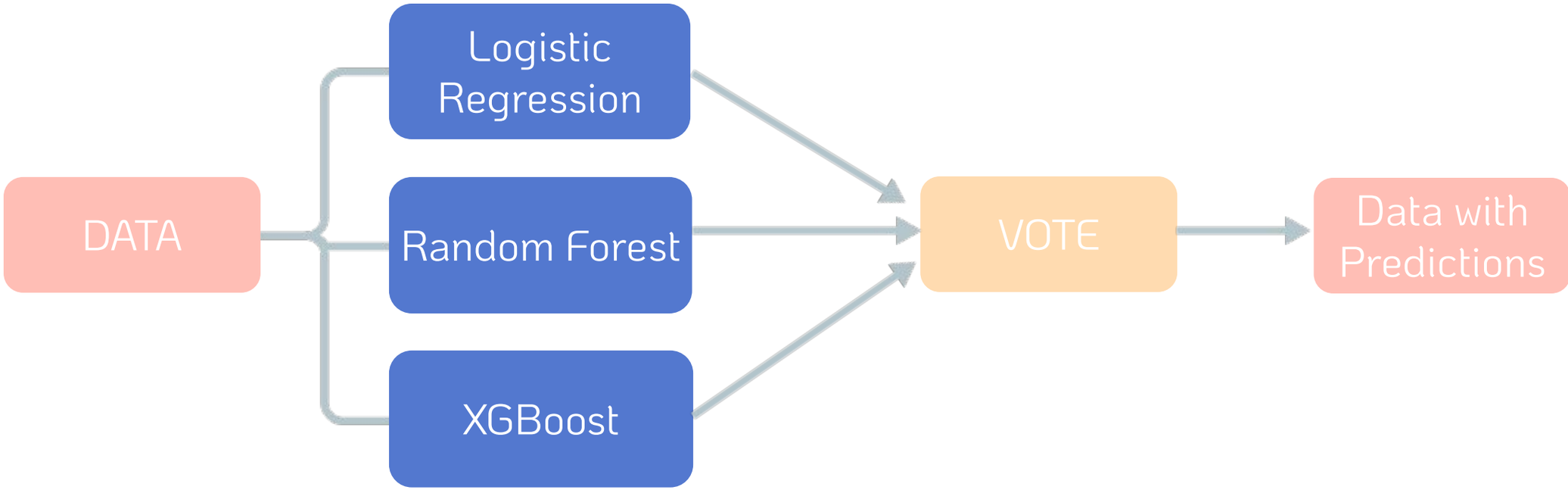
- F1 Score Train: 0.801
- F1 Score Test: 0.806
- F2 Score Train: 0.792
- F2 Score Test: 0.799

XGBoost

```
XgboostClassifier(
  labelCol=”DEP_DEL15_AIRLNS”,
  missing=0.0,
  max_depth=10,
  n_estimators=60,
  learning_rate=.1,
  colsample_bytree=.8,
  gamma = .05,
  reg_lambda = .1)
```

- F1 Score Train: 0.811
- F1 Score Test: 0.811
- F2 Score Train: 0.802
- F2 Score Test: 0.804

VII. Final Prediction Pipeline



VIII. Project Leaderboard and Gap Analysis

Final Classification Metrics on 2019 test data:

- F1 Score: 0.713
- F2 Score: 0.714

Comparison to Leaderboard and Gap Analysis:

- Of those reporting F2 scores, our model did well
- Of those reporting F1 scores, our model was not #1

IX. Performance and Scalability

Performance

- 6 workers - limited hyperparameter tuning over all parameters
- Stalling after period of time
- ~1.5 hr to train ensemble model
- ~10 min to make predictions

Scalability

- 1.5 hr reasonable training time
- Remove older data (seasonality and shifts over time)

X. Limitations, Challenges, Future Work

Challenges:

- Less hyper-parameter tuning than we would have liked
- We did not check-point data as much as we could have; more time running unnecessary compute

Future work:

- Use different probability threshold to make delayed/not-delayed prediction
- More thorough hyper-parameter searching, including some form of random search
- Try out having an imbalanced training set, in the hope to further increase recall and the F-2 score

XI. Conclusion

Summary:

- Performed an end-to-end Machine Learning project, including data cleaning, feature engineering, modeling, and its analysis
- Achieved a peak F-2 score of 0.804 with XGBoost, and precision/recall both > 80%, which we believe can significantly help with mitigating the flight delays in real time

Learnings:

- Exploratory data analysis is very important. We derived many insights from it
- Dealing with data at scale is difficult, and requires careful thought and planning
- Ensemble model is not necessarily always better. Performance tuning is still required for model building

THANK YOU!