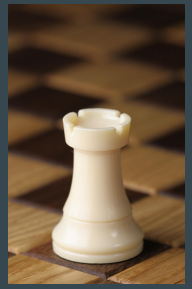


Team: RooKeys
Anudeep Arora, Oussama Landoulsi,
Lalit Yadav

Predicting Flight Departure Delay

The Erdős Institute, Fall 2022 Bootcamp

Problem Statement & Business impact

Problem: Given a 4-hour period horizon and the scheduled flight data together with the weather data, **will a flight be delayed by more than 15 minutes**?

Business position: According to the Federal Aviation Administration, the total flight delays cost is $22 billion yearly.

Target audience: Airline companies.

Goals: Predicting flight delays to improve airline operations by:

- Saving on delays costs and providing quality service to their customers.
- Reducing potential crowding situation at the airport.

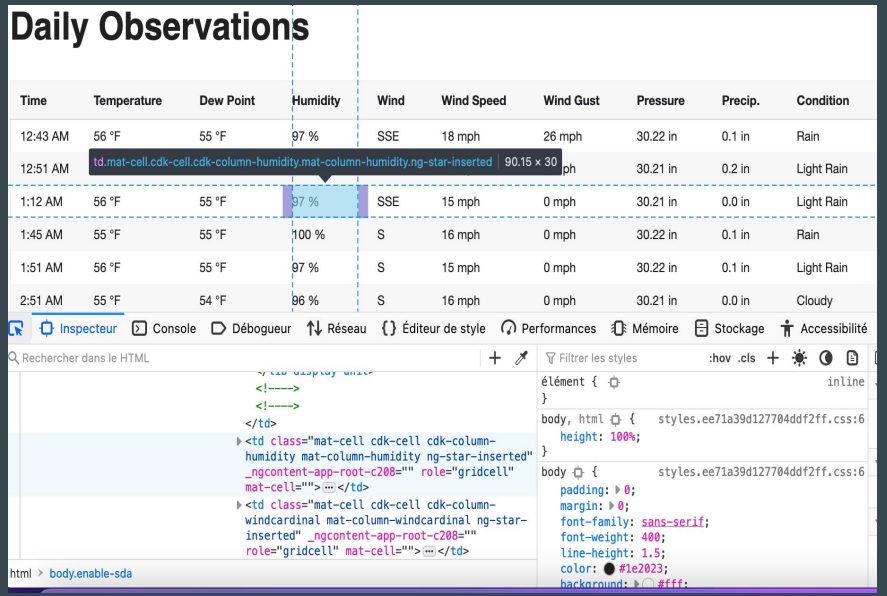
Data gathering

Airlines Dataset: Jan 2018 - Aug 2022

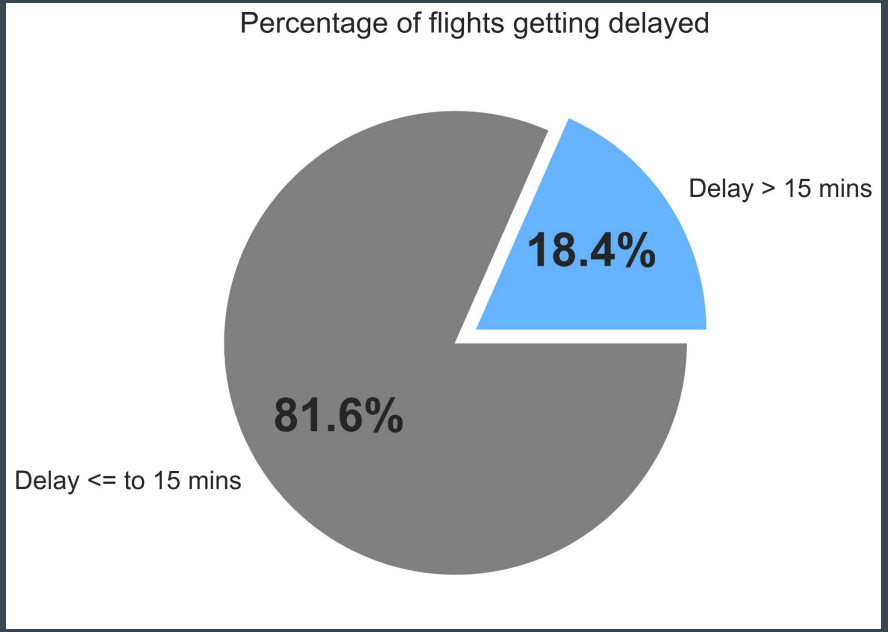
- Reporting Carrier On-Time Performance - Bureau of Transportation Statistics.
- 109 features in total. Extracting the data for JFK as an origin.
- Reducing usable features to 17.

Weather Dataset: Jan 2018 - Aug 2022

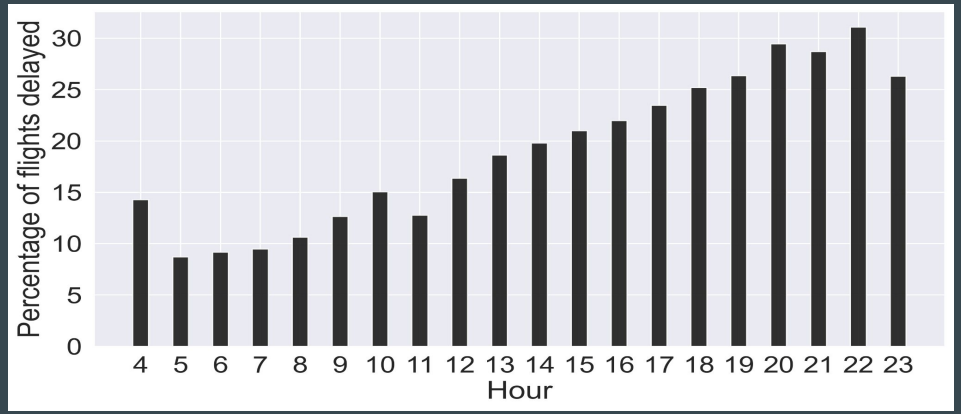
- Web Scraping daily/hour weather data from weather-underground website.
- Selenium library with WebDriver.



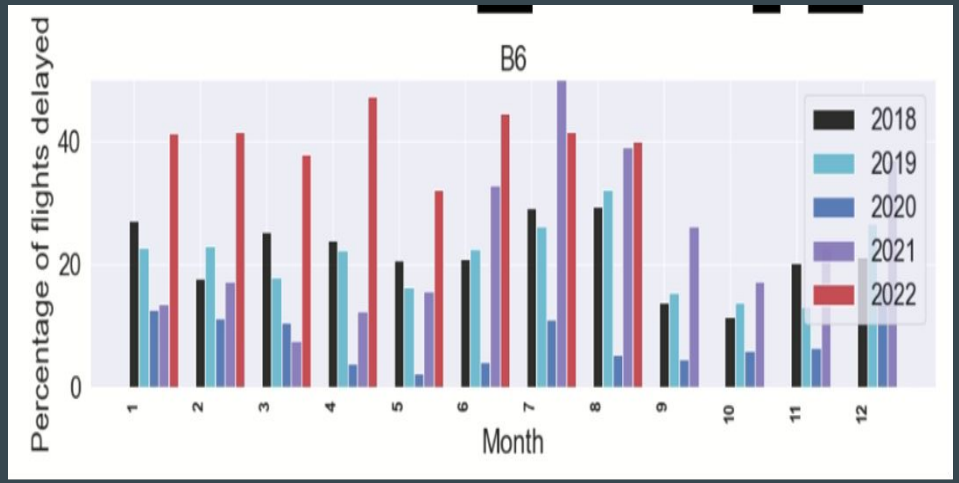
- Highly imbalanced dataset.
- Merge flight and weather data with non uniform time stamps.
- Capturing seasonality and drift for the carrier specific delay.

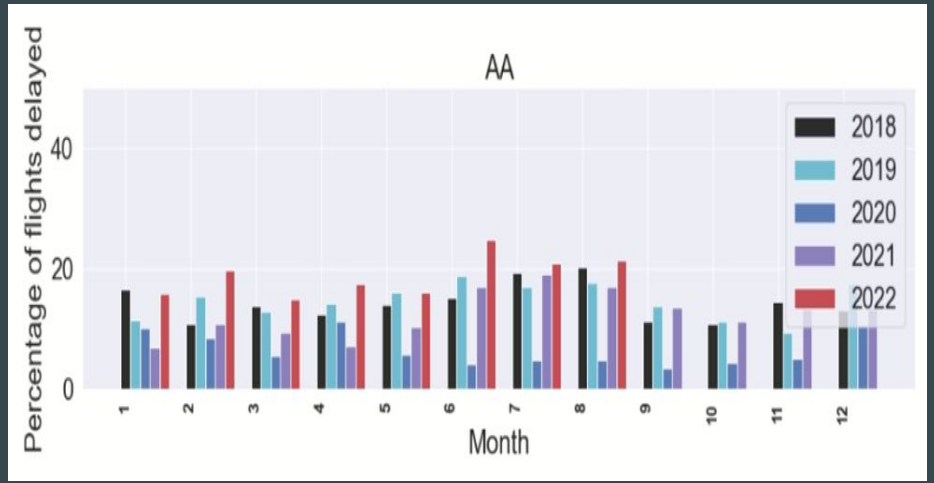


Exploratory Data Analysis

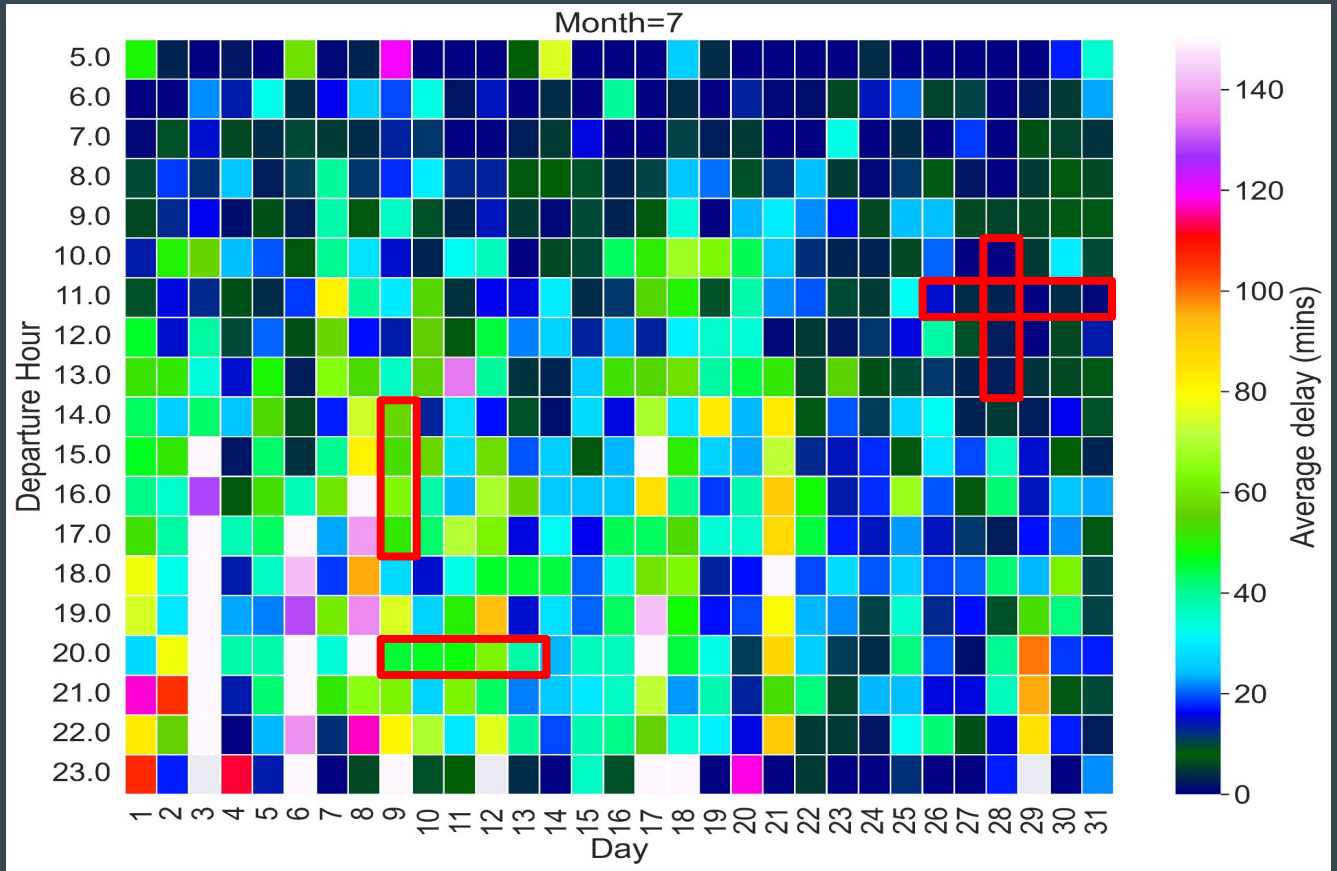


Carrier delay exhibits **seasonality** and **drift**

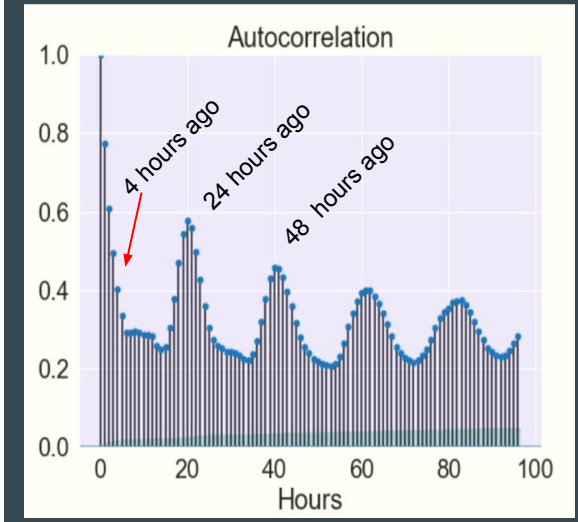




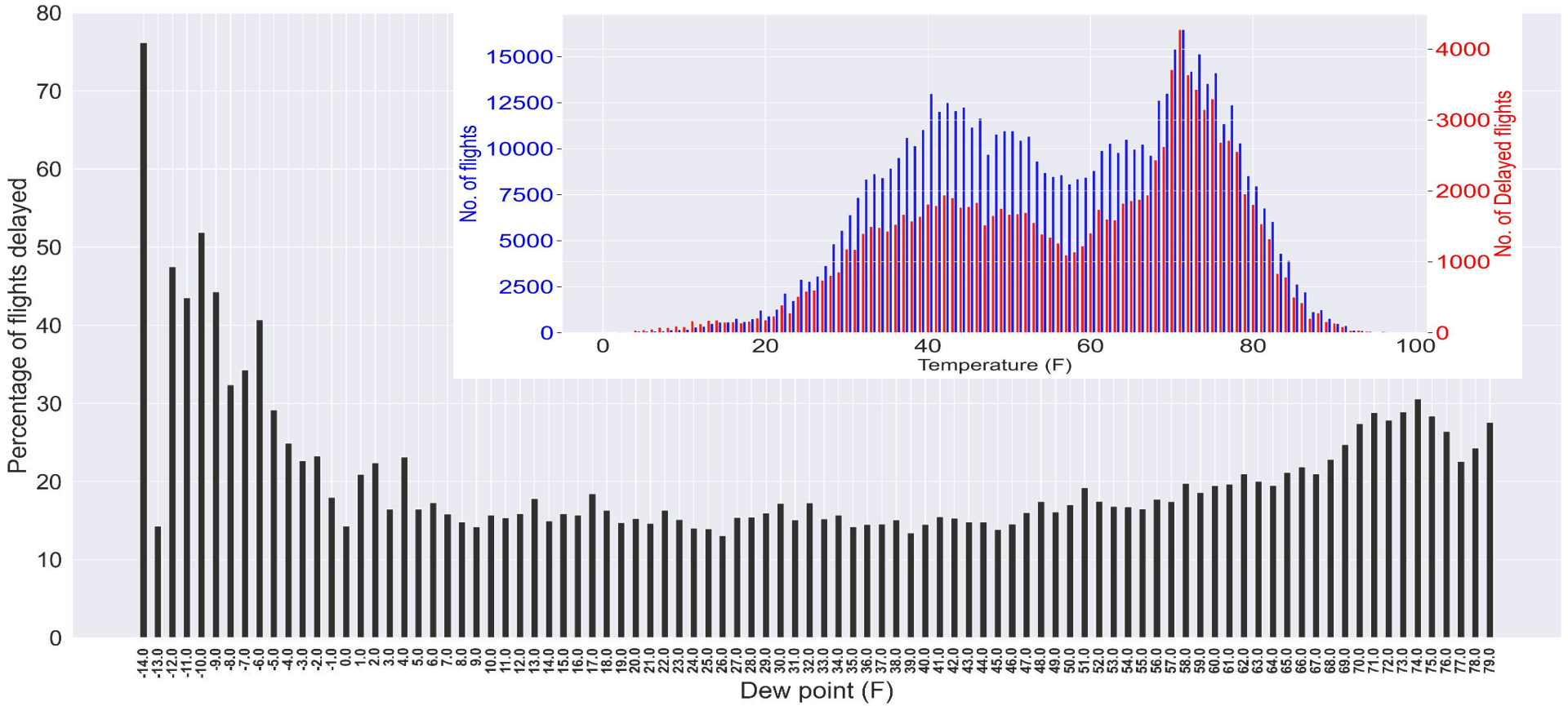
Exploratory Data Analysis (contd.)



Cluster formation and **temporal correlation**



Exploratory Data Analysis (contd.)



Feature Engineering

- Predetermined information:
  - Hour, Day and Month of the flight's scheduled departure.
- Weather information:
  - temperature, dew point, humidity, wind, wind speed, wind gust speed, pressure, precipitation and condition.
- Historical information:
  - Delayed flights at the same hour, a day before the target flight's departure.
  - Probability of delay for a given carrier (past performance).
  - Probability of delay for a given destination (historical information).
  - Delayed flights (%) in 1 hours time window at T - 4 hour for a scheduled departure at time T

Feature Engineering (contd.)

Encoding:

- Normalization using min-max scaler
  - Delayed flights (%) at T - 4 hour, Temperature, Dew Point, Humidity, Wind Speed, Wind Gust, Pressure, and Precipitation.
- Cyclic embedding
  - Day of Week, Month, Hour, and Wind Direction.
- One-Hot encoding
  - Weather Condition

$$F_i(t) \Rightarrow [\sin\left(\frac{2\pi t}{T}\right), \cos\left(\frac{2\pi t}{T}\right)]$$

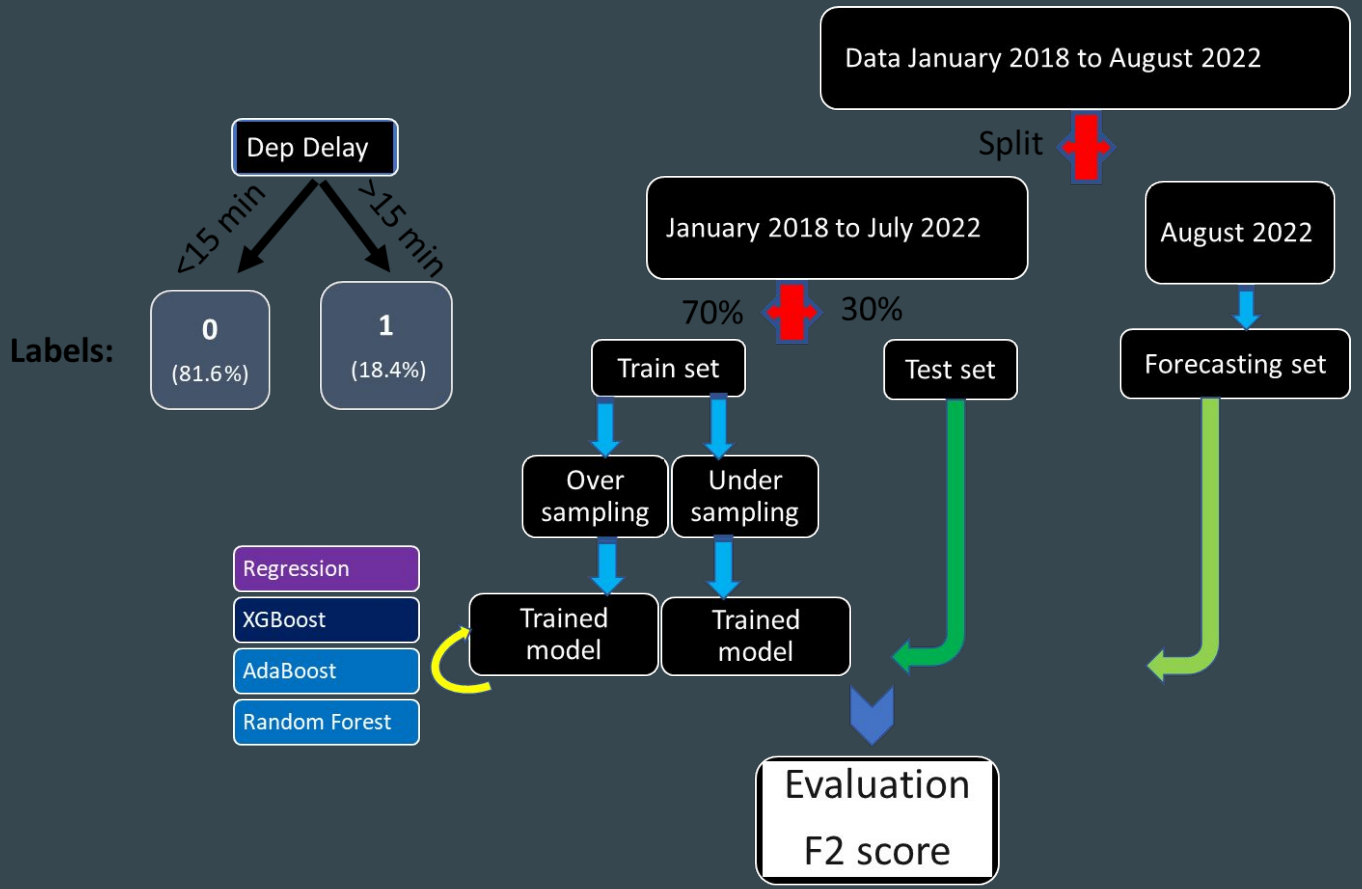
{"N","NNE","NE" …}→degrees→cyclic coordinates

Total features to pass into the model = 71

Modeling Framework

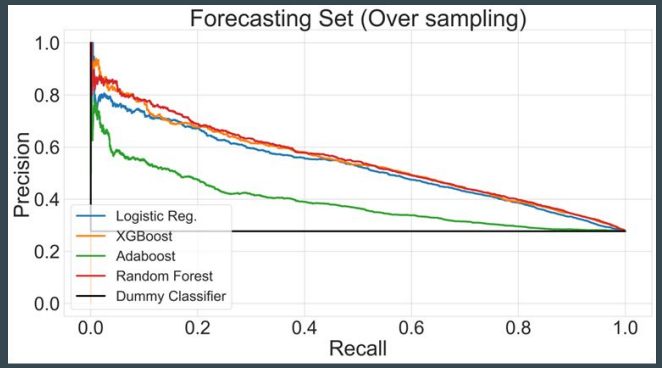
KPI for stakeholders:

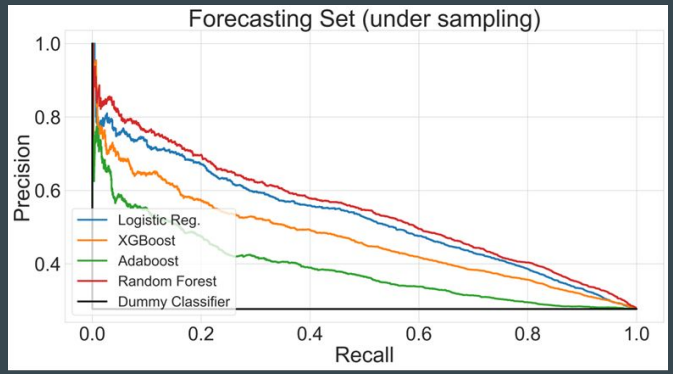
- F2 score (weighing recall over precision)
- Not classifying a delay correctly is costlier



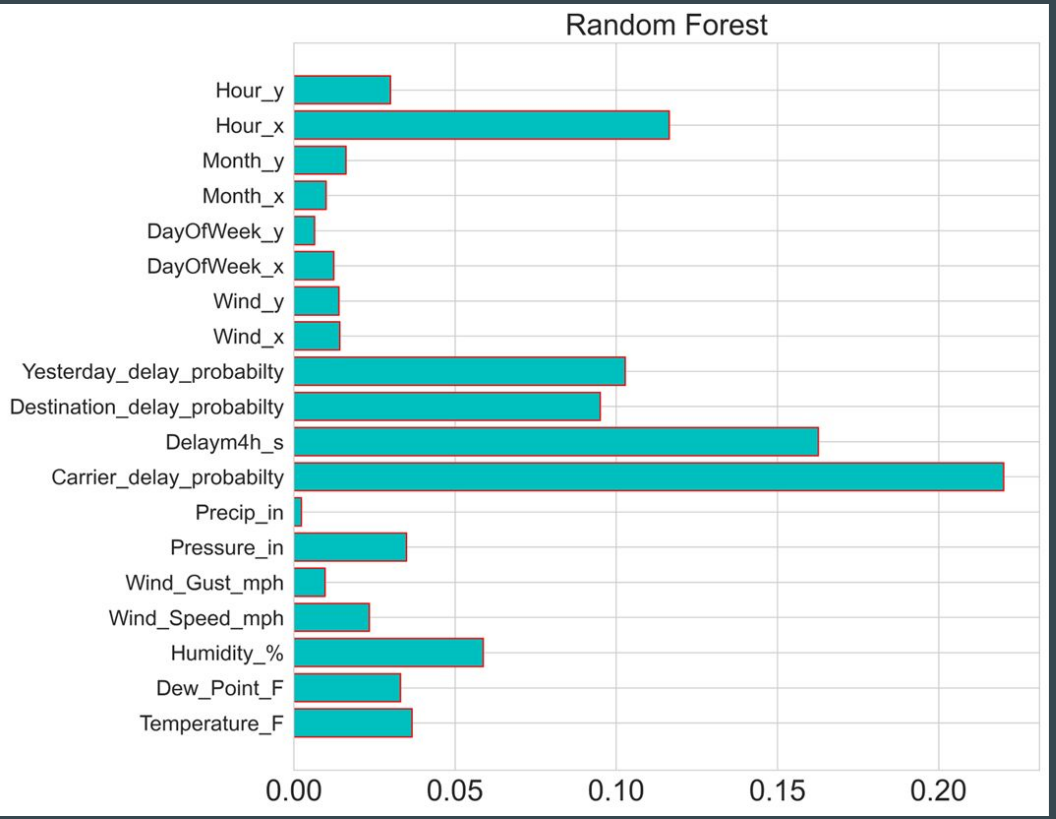
Model Performance

		Train Set	Test Set	Forecasting Set
Over sampling	Logistic Regression	Prec:0.68 Recall: 0.63 F2: 0.64	Prec: 0.33 Recall: 0.64 F2: 0.53	Prec: 0.39 Recall: 0.80 F2: 0.66
	XGBoost	Prec:0.69 Recall: 0.70 F2: 0.70	Prec: 0.33 Recall:0.70 F2: 0.57	Prec: 0.36 Recall:0.86 F2: 0.67
	AdaBoost	Prec: 0.88 Recall:0.90 F2: 0.89	Prec: 0.48 Recall: 0.66 F2: 0.55	Prec: 0.35 Recall: 0.43 F2:0.41
	Random Forest	Prec: 0.73 Recall: 0.71 F2: 0.68	Prec: 0.37 Recall: 0.64 F2: 0.55	Prec: 0.39 Recall: 0.82 F2:0.67
Under sampling	Logistic Regression	Prec:0.68 Recall:0.63 F2 : 0.65	Prec: 0.33 Recall: 0.64 F2: 0.53	Prec: 0.39 Recall: 0.80 F2: 0.66
	XGBoost	Prec: 78 Recall:0.96 F2: 0.92	Prec: 0.33 Recall:0.86 F2: 0.65	Prec: 0.34 Recall:0.77 F2: 0.62
	AdaBoost	Prec: 0.91 Recall:0.91 F2: 0.91	Prec: 0.38 Recall: 73 F2:0.62	Prec: 0.34 Recall: 56 F2: 0.50
	Random Forest	Prec:0.73 Recall:0.74 F2:0.68	Prec:0.36 Recall:0.64 F2:0.56	Prec:0.39 Recall:0.83 F2:0.68





Feature Importance



If we remove “Delaym4h_” feature, the prediction horizon would increase to 24 hours.
What would be the performance of model?

- Prediction Horizon: 24 hour
- Model: Random Forest
- Precision: 0.39
- Recall: 0.82
- F2 score: 0.67

Conclusion & Future Work

Conclusion:

- Model achieved 80% ability in predicting a delay.
- Feature engineering is critical for model performance.

Future Work:

- More general model including all airports as an origin.
- Delay linked to the airlines organization:
  - Understanding the spread of the delay for all origin-destination pairs.
  - Predict the evolution of the delay in real time.
  - Probabilistic Graphical Models (PGMs), e.g. Bayesian Network.
  - Incorporate weather information at Destination airport.

Special thanks to Roman, Lindsay, Matt, Shuvra, Alec and everyone in the Erdős Institute.

Thank you for listening!