Allay Airway Delay!

Predicting Flight Delays to Reduce Wasted Customer Time

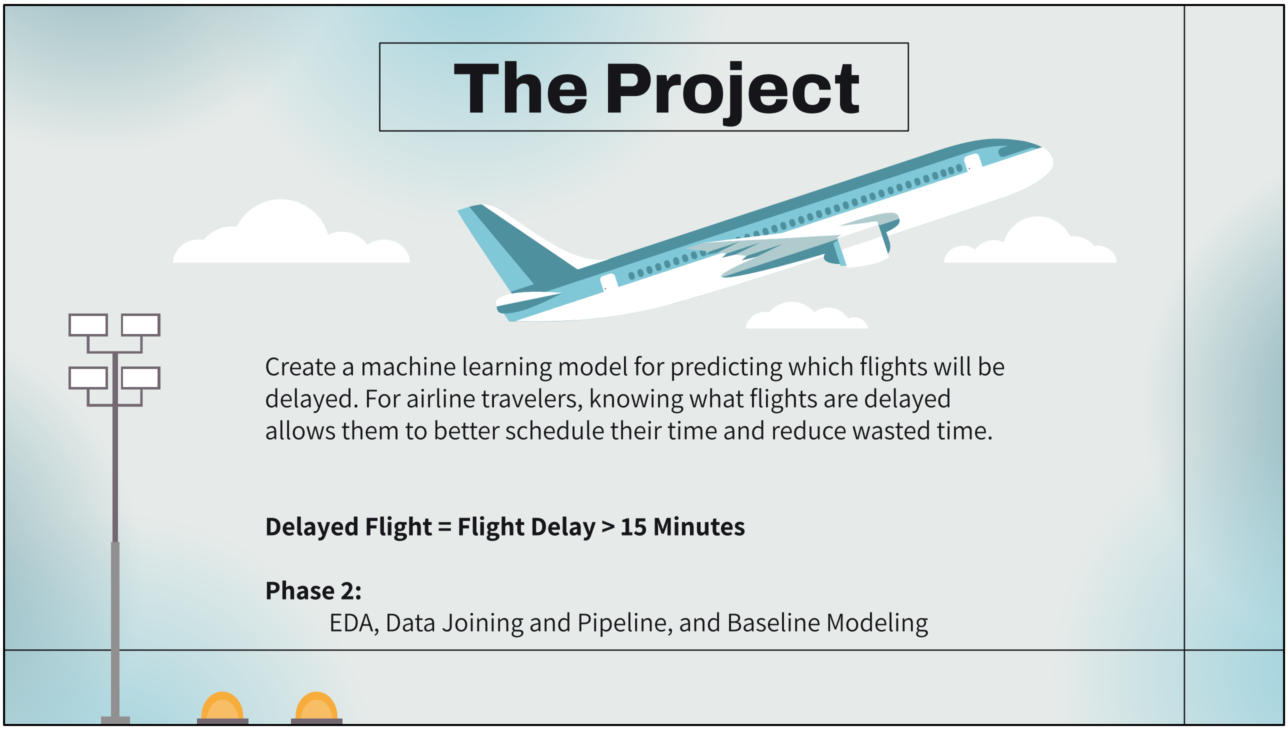
Phase 2 Check-In

Team 13:

Sparks and Stripes Forever

Nashat Cabral
Deanna Emery
Nina Huang
Ryan S. Wong (Current Speaker)

Welcome to the Phase 2 Check-In for Team 13, also known as Sparks and Stripes Forever.



The Project

Create a machine learning model for predicting which flights will be delayed. For airline travelers, knowing what flights are delayed allows them to better schedule their time and reduce wasted time.

Delayed Flight = Flight Delay > 15 Minutes

Phase 2:

EDA, Data Joining and Pipeline, and Baseline Modeling

As a reminder, the goal of our project is to create a machine learning model that will predict whether or not a flight will be delayed by 15 minutes or more. Our aim is improve the experience of air travellers, reducing the amount of time they waste waiting for a delayed flight with high precision.

In Phase 2, our work was primarily focused on EDA, Data Joining and Pipeline Creation, and Baseline Modelling, all of which we will go over now.

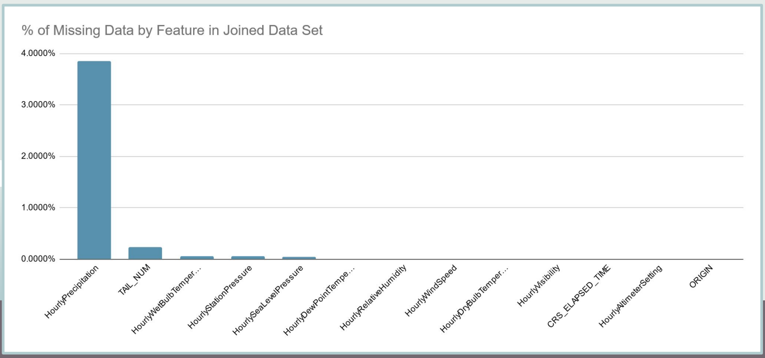
EDA

25 Features Retained
3 Features Created

28 Total Features Used

# of rows in joined data set

30,820,796



% of Missing Data by Feature in Joined Data Set

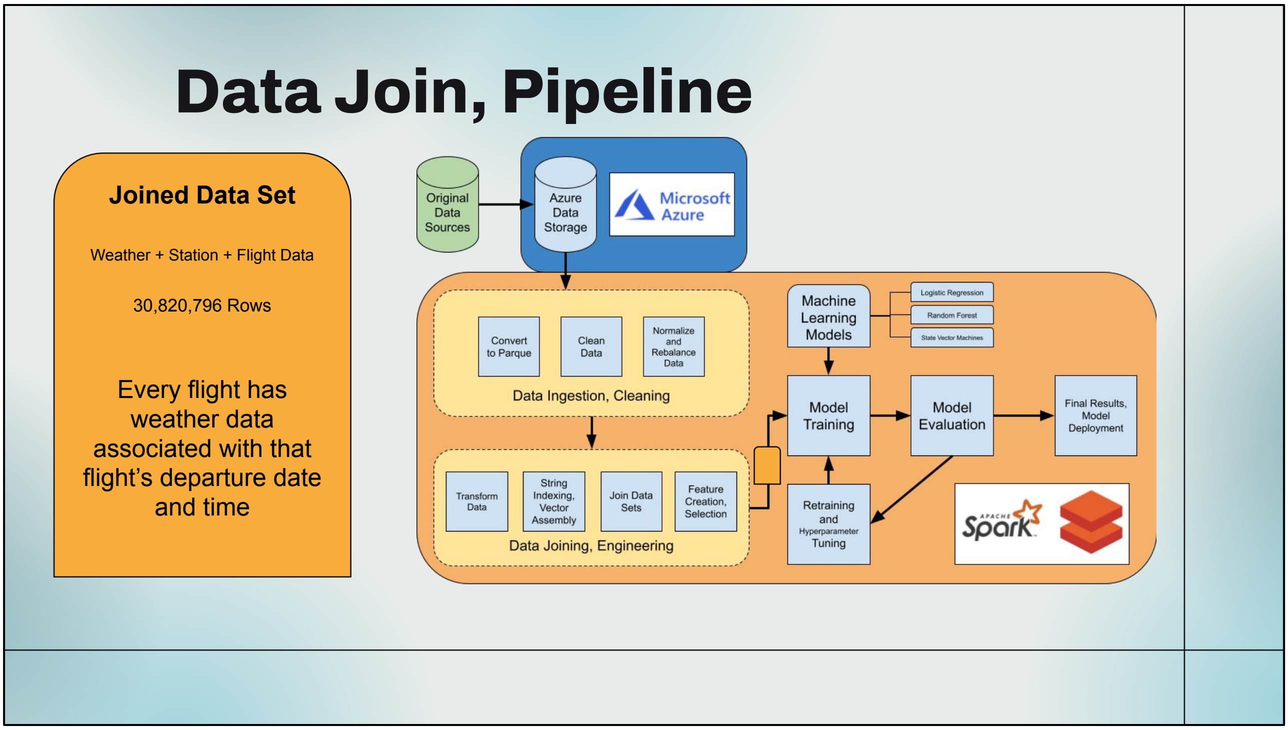
% of rows retained after join

41.55%

% of post-join data without errors

96%

Our original data set was over 70 million rows across three data disparate data sets. Our initial EDA on the raw data showed that each of which was rife with missing data and extraneous features too numerous to state here. As a result, we picked or created 28 relevant features to use in our joined data set. Ultimately, our joined data set was both sufficiently large and remarkably clean: 30.8 million rows with 96% of them being completely free of missing values.



Data Join, Pipeline

Joined Data Set

Weather + Station + Flight Data

30,820,796 Rows

Every flight has weather data associated with that flight's departure date and time

Original Data Sources

Azure Data Storage

Microsoft Azure

Convert to Parque

Clean Data

Normalize and Rebalance Data

Data Ingestion, Cleaning

Transform Data

String Indexing, Vector Assembly

Join Data Sets

Feature Creation, Selection

Data Joining, Engineering

Machine Learning Models

Logistic Regression

Random Forest

State Vector Machines

Model Training

Model Evaluation

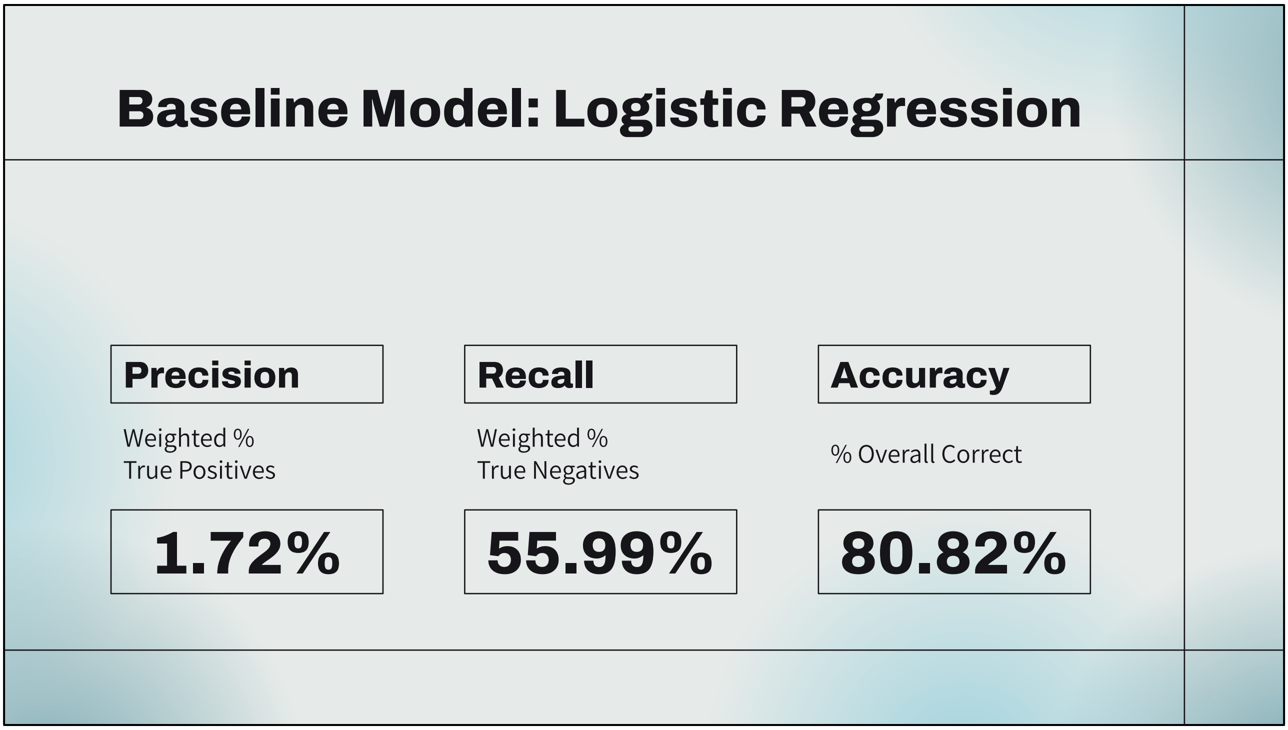
Final Results, Model Deployment

Retraining and Hyperparameter Tuning

Apache Spark

Having that joined data set, combining the flight data, station data, and weather data, was achieved by our current data pipeline.

The diagram on the right shows the data pipeline we have created thus far. This pipeline can successfully take and store raw data, ingest and clean it, combine and transform it into a joined data set, and then apply that data towards training our models.



Baseline Model: Logistic Regression

Precision	Recall	Accuracy
Weighted % True Positives	Weighted % True Negatives	% Overall Correct
1.72%	55.99%	80.82%

The first model that we will be applying to our joined data set is a logistic regression model, which will act as our baseline model. After being trained, cross-validated, and evaluated against our test data, this baseline model has an astounding precision of… 1.72%. Seeing as our goal is to minimize false positives, this is a surprisingly low precision level. Perhaps some of our features are introducing a lot of unintended noise, or maybe logistic regression just does not perform well with this categorization problem. Further investigation will be done as we move to Phase 3.