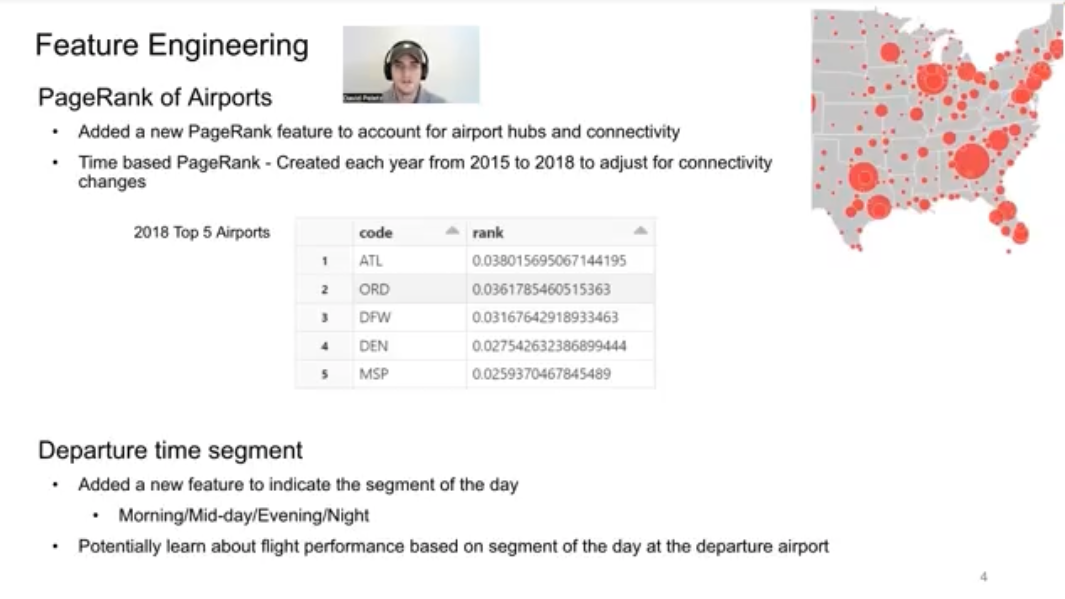
Gap Analyses

# Phase III

* Calculated averages for null values
* Hyperparameter tuning – max depth
  + F1 score of 0.32 on 2019 data
* Handling imbalance
* Gradient boosted trees
* PageRank feature - one feature for each year - Team 29
* 
* Team Super Mario – F\_beta score of 0.86
  + Gaps: implementing random forest
* Cross validation implementation
  + Different strategies/more folds
  + Hyperparameter tuning

A gap analysis is a process that compares actual performance or results with what was expected or desired (versus other teams in this case). The method provides a way to identify suboptimal or missing strategies, structures, capabilities, processes, practices, technologies or skills, and then recommends steps that will help the company meet its goals.

**Phase III Gap Analysis**

When comparing our project with others on the leaderboard, we observe several key next steps that we can pursue for our project. Most notably, Team Super Mario’s model has an F\_beta score of 0.86, which outperforms our current model. Team Super Mario uses a Random Forest model to achieve this score, and we plan to implement this model next. Other teams implement additional strategies that we will pursue in the coming week, including hyperparameter tuning, cross validation with multiple folds, and additional feature engineering (ex. PageRank, filling in null values).

In terms of teams performing better than us – Team Super Mario has implemented a Random Forest model, which we plan to pursue.

In terms of just looking at what other teams are doing, there are gaps in hyperparameter tuning, cross validation with multiple folds, and additional feature engineering (ex. PageRank, filling in null values)

In this phase, we have a fully consolidated dataset that includes flight and weather data from 2015 to 2021. We have completed a baseline model with initial results. We have a pipeline in place that we can use to experiment and tune models and algorithm. This includes implementing cross-validation in ways that take into account the temporal nature of data. We are including support for weights and oversampling of the minority class to take into account the imbalanced nature of the dataset. We have performed more data analysis and processing to facilitate the feature selection/engineering work that we plan to complete next week. We implemented a F\_beta metric as part of our CustomCrossValidator.

Next, we will explore additional models, expand upon feature engineering and selection, tune hyperparameters, test various options for Cross Validation, make final determination on using aircraft data for tailnumbers, and determine a preferred model for the final solution and setup for final testing in preparation of the final wrap up.