PREDICTING OUTCOMES OF NFL MATCHUPS

Louie Bafford

PROBLEM CONTEXT

- NFL has a huge market – 8 Billion in revenue in 2017
- Top experts ~67% accurate

Predicting NFL Outcomes

- Sports gambling platforms
- NFL Analysts and other analysis platforms

Understanding important components of a winning team

- Owners, GMs, Coaches, Talent Scouts, Agents

DATA COLLECTION

Online API - https://profootballapi.com/

- Flexible data collection and feature engineering
- Leverages domain knowledge

- Aggregate statistics at a yearly level for each team
- Compare statistics for each team matchup
- Predict winner based upon matchup

FEATURES

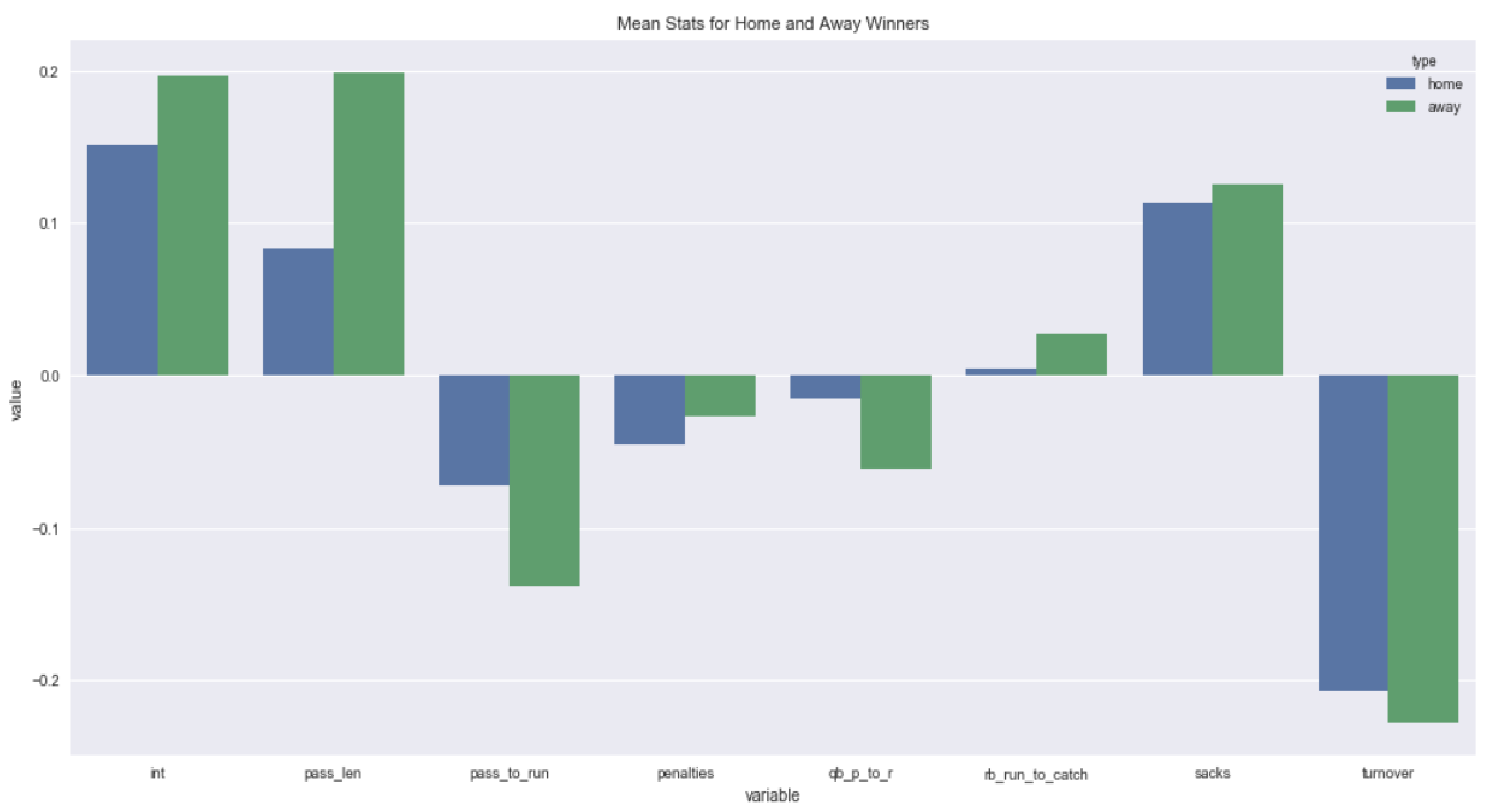
- Data from 2011 to 2017 season
- Features should cover many aspects of gameplay
- Overall ~40 features were chosen (~20 for each team per matchup)
- Focus on features which can influence team decisions, examples below

Feature	Value
QB Pass/Run ratio	Effectiveness of QB playstyle
Average Pass Length	Effectiveness of play call styles
Turnovers	Impact of turnovers influence risky play calling

EXPLORATORY ANALYSIS

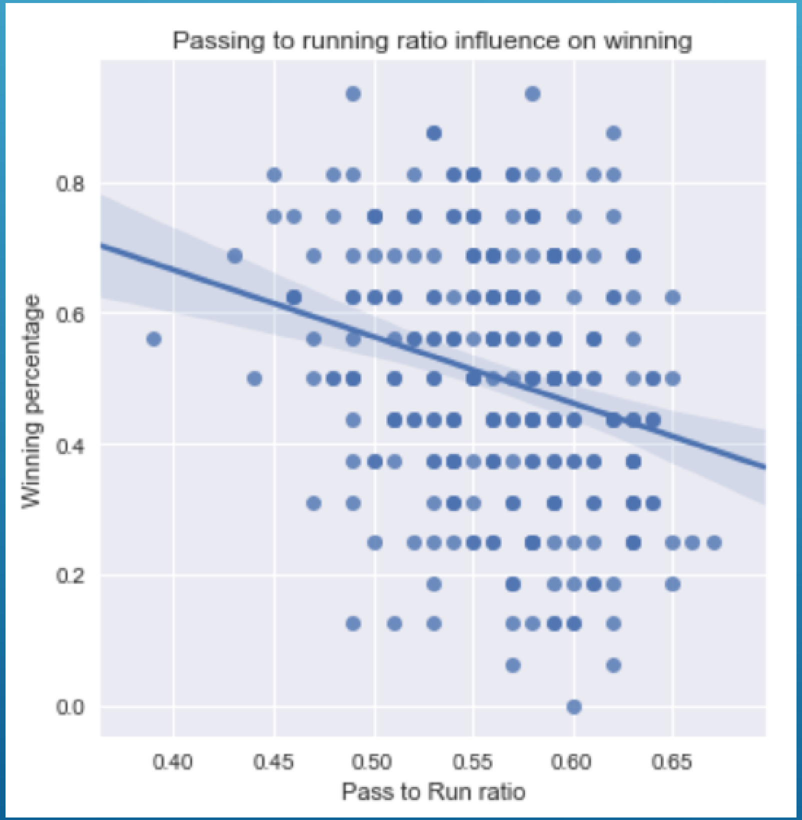
Average statistics of winning teams

- Turnovers are most significant
- Features are more extreme for away teams

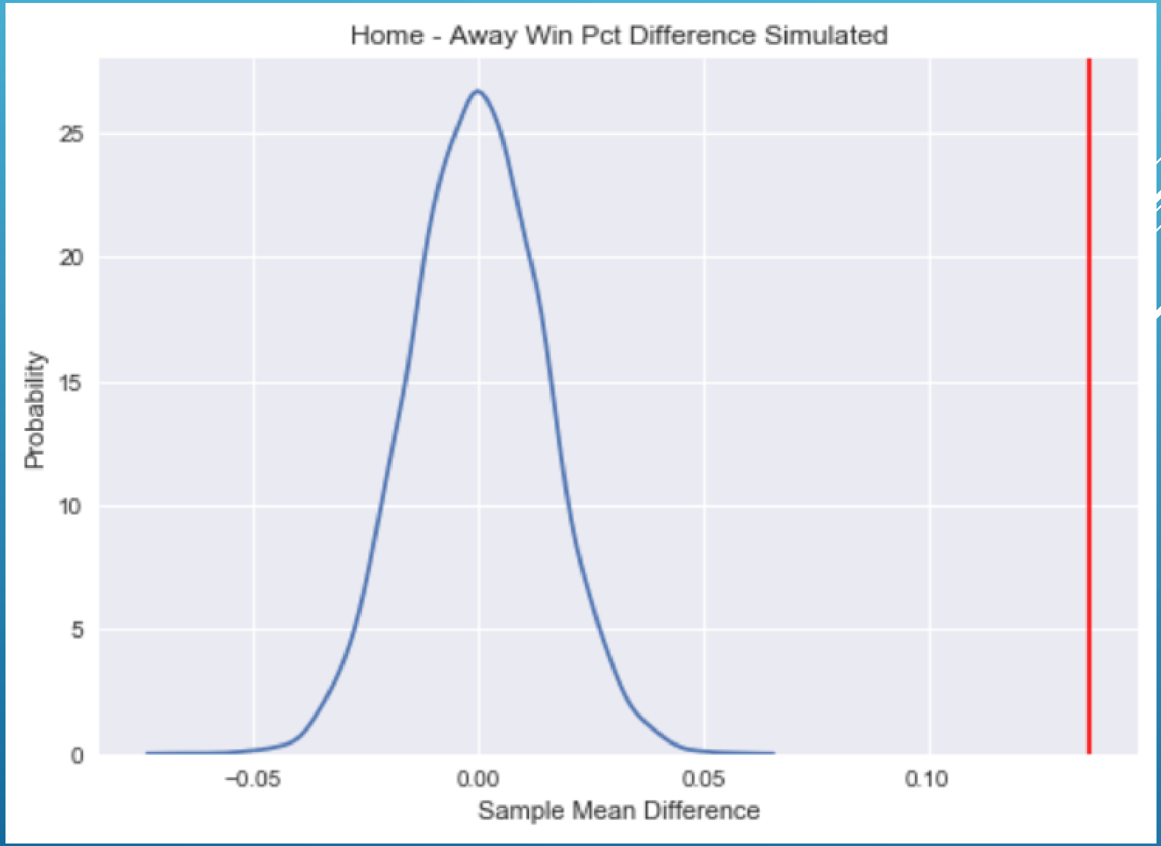


STATISTICAL ANALYSIS

- Statistically significant correlation between low passing ratio and winning percentage
- P-value < .01



- Hypothesis test statistically validates home field advantage
- P-value < .01



BUILDING THE MODEL

Explore and tune a list of applicable models

- Linear Regression
- Logistic Regression
- SVM
- Random Forrest

Choose top model and explore for insights

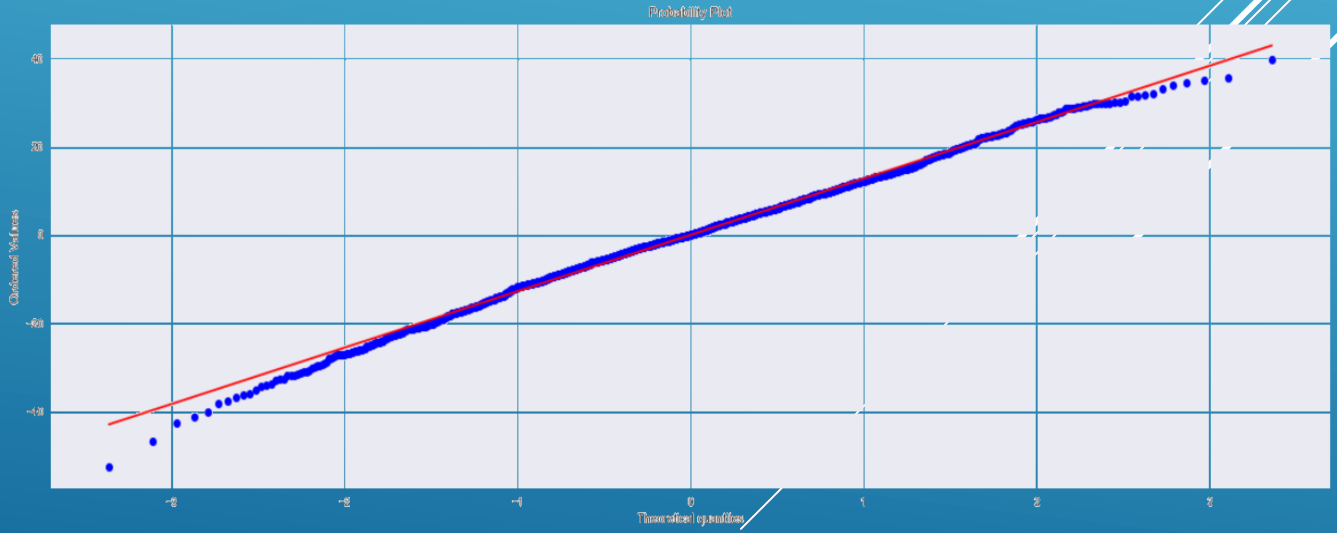
LINEAR MODEL

- Train model on the score difference (an integer outcome)
- Convert to binary classification to tune using accuracy as the performance metric

Cross Validation Accuracy	69.2%

Assumption – predictions are normally distributed around correct value

QQPlot shows non-normal distribution weakening the validity of the model



CLASSIFICATION MODELS

Logistic Regression

- Top performance among models
- Feature importance visibility

Random Forest/Gradient Boosted Forest

- Poor model performance
- Feature importance visibility
- Gradient Boosted improved performance but model is still weak

SVM

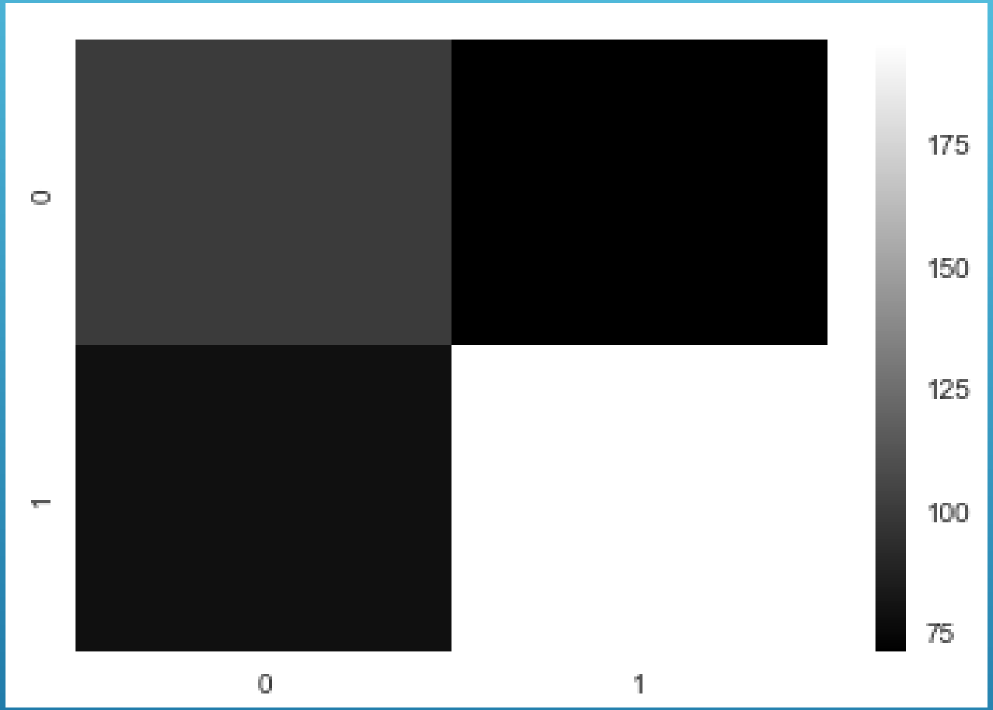
- Top performance among models
- No feature visibility

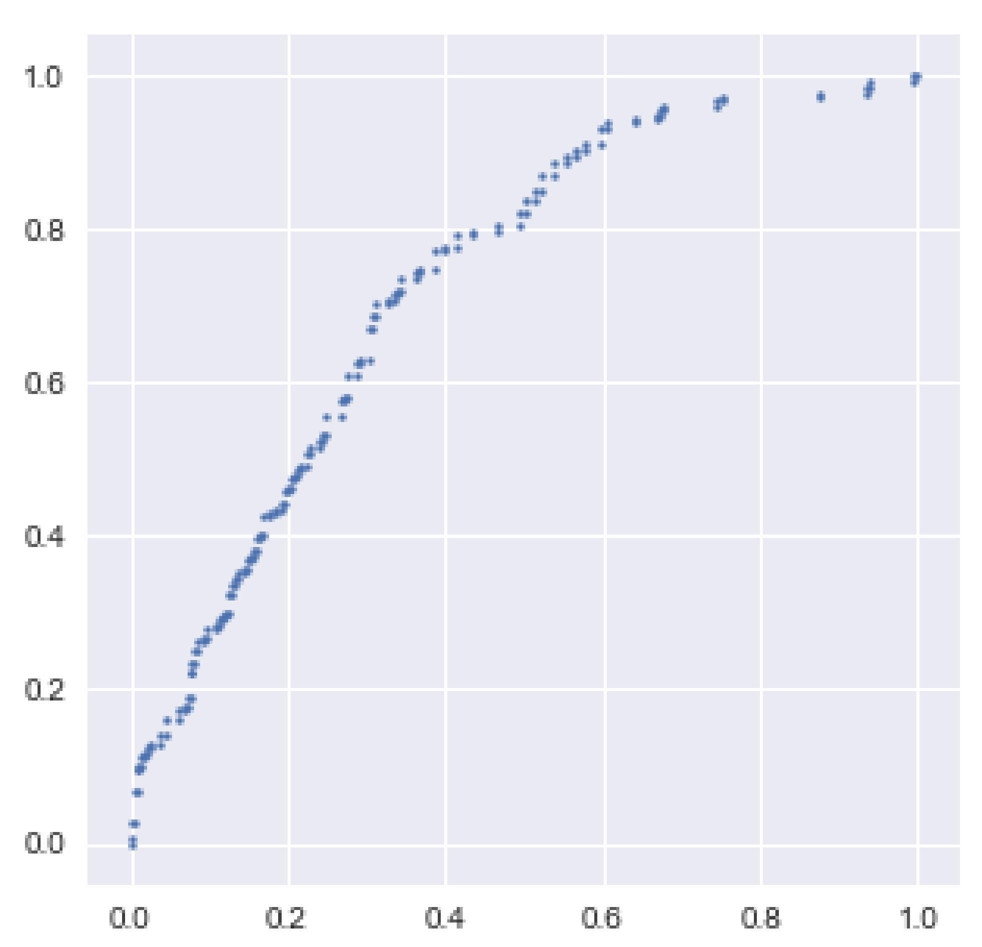
Model	CV Accuracy
Logistic	70.28%
Random Forest	63.2%
Boosted RF	67.5%
SVM	70.34%

Logistic Regression selected as final model

Test Accuracy: **66.36%**

MODEL INSPECTION

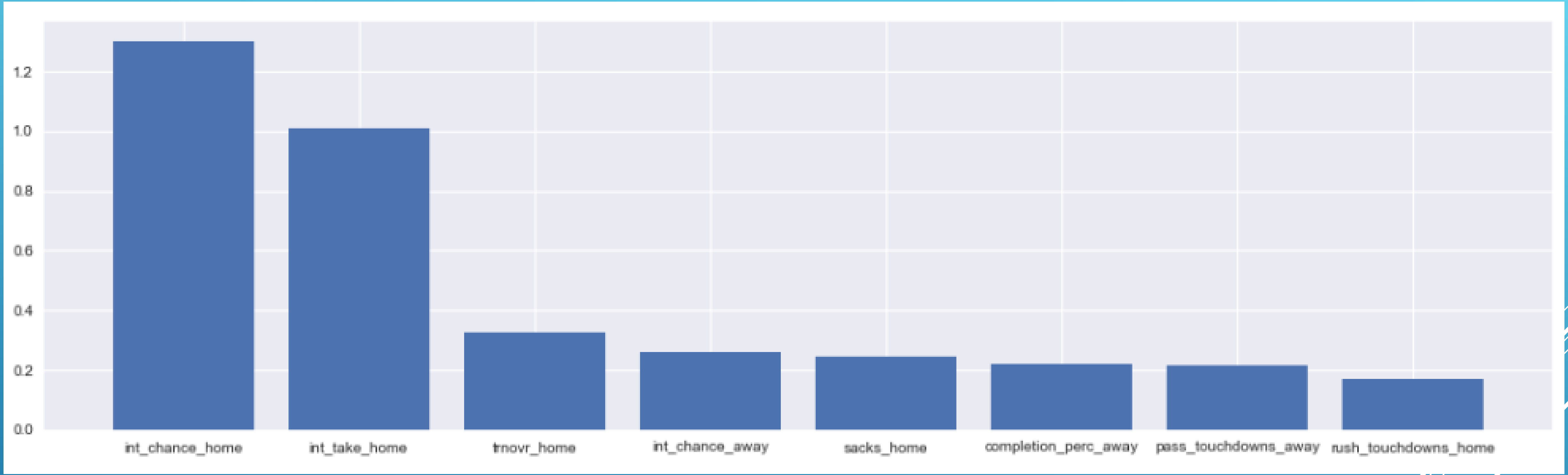




- Confusion Matrix behaving normally
- Even counts of misclassification

- ROC Curve is also normal
- Low curve shows room performance increase

FEATURE ANALYSIS



** INT CHANCE = (PASS-RUN RATIO OF OPPONENT * INTERCEPTION TAKEAWAYS) **

- Pass to run ratio important to both home and away teams
- Turnovers and takeaways (turnover differential) very important
- Sacks for home team
- Completion percentage for away team

IMPROVEMENTS

Very Noisy Data

- Adding features caused overfitting
- Model performed best with limited features

- Decrease noise with more relevant features
- Further explore feature Engineering

Hierarchical Models?

- Lots of correlation between features
- Build relationships between features for a structured model

SOURCES

Revenue

https://qz.com/1383416/amid-controversy-the-nfl-is-still-thriving-financially/

Analyst Predictions

https://www.fantasyfootballnerd.com/nfl-picks/accuracy