**Introduction**

For this project, I was hired as a data analyst for a large sports betting company. I was responsible for producing predictions regarding which NFL teams are most likely to make the playoffs this season. Highly accurate predictions will allow the company to offer bets with odds that mos t favor the company themselves. The analysis has produced key predictions that will allow the company to avoid non-ideal offers in the betting marketplace.

According to the American Gaming Association (AGA), sportsbook revenue hit $7.5 billion USD in 2022. The total addressable US betting market is estimated to be worth more than $40 billion USD by 2030, according to Flutter. To best capitalize on this growing market, our company must be able to consistently offer favorable bets. Rather than relying only on human opinion, I have designed a model to classify teams into either “making the playoffs” or “not making the playoffs” to leverage the large amount of statistical data available on the NFL. This model, combined with the expert opinion of our company’s analysts will allow us to offer favorable bets more consistently to ourselves.

**Section 1: Data Overview**

For this project, I needed to source my own data to build my models. I started with ESPN’s NFL statistics. The ESPN data consisted of offensive, defensive, special team, and turnover statistics for each of the 32 teams in the NFL. Their data ran back to 2014 and showed which teams made the playoffs each year.

In addition to the NFL statistics, I wanted to include expert opinions as variables to capture their sentiment in my models. To do this, I used the Massey ratings. The Massey ratings are an open-source computer rating of all 32 teams in the NFL. They are rated on a variety of statistics including power, defense, record, etc. The rating system has been in use (with regular improvements and updates) since August 1999 and uses regression and Bayesian inference to derive its statistics. Overall, the Massey rating system proved a useful input to represent an expert’s opinion on each team’s season performance.

To create the dataset, I copied each team’s performance across all the above measures into an excel document for the years 2014-2022 (2023 was made separately). I first copied each section (four ESPN sections and one Massey rating section) into separate sheets. Then, I used INDEX(MATCH()) to combine all five sheets into one final dataset.

I repeated this process separately for the 2023 data, which is what I ultimately will be predicting over.

**Section 2: Model Overview and Performance**

I tested 3 different models: a cross-validated Naïve Bayes Classifier, a probit linear regression, and a pruned classification decision tree. Below are the accuracy results of each model:

|  |  |
| --- | --- |
| Model Type | Accuracy |
| Naïve Bayes (10-fold CV) | ﻿60.42% |
| Probit Regression | 66.67% |
| Classification Tree (ccp=0.012) | 84.40% (improved from 75.00%) |

The table above illustrates the effectiveness of our final model, the pruned classification tree. Naïve Bayes likely did not perform as well as there were more teams that did not make the playoffs than did make the playoffs. I elected to not try KNN due to the high dimensionality of the data, which had 33 features after cleaning. The probit regression performed decently; however, many of the variables were not normally distributed and thus the model’s inferential power is severely diminished. Indeed, this data, being high dimensional and varied, is a perfect fit of a tree-based model. The pruned classification tree was able to pick out the most important features including the Massey record statistic, the Massey power statistic, and total defensive yards per game, which all had some of the strongest correlations with making the playoffs.

After running our classification tree on the 2023 data, we are left with 4 definite “yes” results for teams expected to make it to the playoffs. The model predicted a 1 for the LA Chargers, LA Rams, Cleveland Browns, and Cincinnati Bengals. The model predicted a 0 for all the other teams. This may tell us that only the above four teams have the statistics at the current moment to safely predict them entering the playoffs. It would be prudent to continue to update the 2023 data and rerunning the model as the playoffs approach to get a better idea of which teams are most likely to make it. However, these early predictions may still prove useful as the company, in conjunction with expert analysis, may flag the above teams early to have offers on the table.