Updated_Cap_Two_Proposal (March, 2022)

Problem Identification

Problem Statement

Are there more important factors and qualities of a NFL Super Bowl team and can all of these factors be used to predict a winner accurately, from the 2022 season?

Context

The 'Super Bowl' is the championship game of the National Football League (NFL), that happens annually. Two leagues compete, the National Football Conference (NFC) and the American Football Conference (AFC), until there is one team from each league (two in total) elected with the highest winnings from one entire season. According to Rolfe (2022), "For winning Super Bowl LVI (or Super Bowl 56), every member of the team receives $150,000. For a rookie on the NFL minimum salary, that is more than $100,000 over their normal game check during the 2021-22 season (approximately $39,000)." Consequently, having the ability to predict the NFL champion team in advance or accurately can be a financial game changer and very lucrative. Developing a robust model to predict the winner can also highlight the most important factors that make a particular football team successful and therefore profitable. It may help teams in devising better game plans and improving recruitment strategy. Using historical stats of teams for a business case, we can use a dataset to understand the most important factors related to winning the 'Super Bowl'. There can be many use cases for this type of prediction especially for an audience of either the public, football enthusiasts, investors, or just anyone who is looking to understand the fundamentals of the sport better.

Criteria for success

Given the teams' records in regular season and playoff games, accurately predict the winner of the Superbowl LVI (between the Los Angeles Rams and Cincinnati Bengals).

Scope of solution space

The focus of the solution space will be on identifying at least 39 features or indicators located in the Data Sources section below to make accurate predictions for each observation. Although there is data available, ranging from 2002, the data that will be used for this case study will be the year 2021-22 data.

Constraints

- It must be assumed that the data represents the true statistics/observations as documented accurately to the best possible (no cheating).
- It must be assumed that all data accessible has been included and no vital information has been excluded.
- There is no direct implication of the data based on its current state.
- Any vital information based on seasonality or rare events has been excluded due to simplicity and out of the scope of this case-study.
- This case-study does not include any athletes our for injuries.
- Special teams stats and performance are possibly split between offensive (e.g., field goal kicker), defensive (ball possession turnover), or not included at all (statistics of punters).

Updated_Cap_Two_Proposal (March, 2022)

Stakeholders

- Investors
- Data Analysts
- Football enthusiasts
- Businesses
- Gamblers
- Consumers
- Retail and Marketing industries

Data Sources

The dataset has one CSV file that includes all observations for the stats from all regular season and playoff games received from the NFL Team Stats 2002-2021 (ESPN) via Kaggle database. The following link is to a data snippet along with metadata column descriptions that identify important columns:

Download NFL dataset (304 kB).csv

This data has been extracted from Kaggle and it is free to use the data.

All data is scraped from ESPN's Team Stats page for each game. Seasons include all regular season games plus all playoff games, with the exception of 3 games that are missing from ESPN's site:

- DAL@WAS 12-30-2007
- CAR@PIT 12-23-2010
- TB@ATL 1-1-2012

Any errors or quirks in ESPN's data will be present in this dataset. For example, redzone conversions are missing prior to the 2006-07 season.

Potential Applications

Some of the questions that were also inferred from this dataset are:

- How did the various types of stats change over time?
- How accurate would predicting the future super bowl LVII be?
- Which stats have more spread and which ones are more consistent/reliable?
- How do the stats correlate with each other, is it important?
- Are seasonal trends or any other factors important and can cause fluctuations (e.g., weather)?

Project strategy:

1. Estimate statistical performance of the two teams playing in the Super Bowl LVI.
2. Develop a regression model that accurately determines the scores from previous games played in the season.
3. Input the estimated feature values in the model to predict the score for each team.
4. The team that's estimated to score more points is the projected winner.

Deliverables

The deliverables for this case study is the following:

❏ 1-2 page google document (commentable by mentor)

❏ The final submission will be a PDF document submitted to the associated Github repo.

References

Rolfe, B. (2022, February 16). *Super Bowl payouts: How much do players get for playing and winning the Super Bowl?* Pro Football Network. Retrieved November 13, 2022, from https://www.profootballnetwork.com/super-bowl-payouts-how-much-do-players-get-for-playing-and-winning-the-super-bowl/