Super Bowl Prediction Analysis

DSC 680 Project 2 Proposal

Isabella Sturm

2022 January 9

# Topic

The topic of this project is Super Bowl Teams. The goal is to predict which two teams will make it to the Super Bowl.

# Business Problem

Many Americans, from avid football fans to the laisse-faire type, get ready for the one Sunday per year they can get even the least interested family members to sit in front of the TV to watch the game.

Year after year, interest in football and the Super Bowl spikes around January and February as the big game approaches. The big questions everyone wants to know, will my team make it to the Super Bowl and who will win?

Some people bet their money they know who will make it and they know who will win. Others just want to know if they really want to tune in or if they will only be interested in the commercials and the half-time show.

For professionals, this project will provide insight into what it takes to make a good team – a team good enough to make it to the Super Bowl. For fans, this project will tell them if their team (or their backup, backup team) will be in the Super Bowl. And for fantasy football players, it will give them both.

# Data

The plan for the data is to use Beautiful Soup to scrape a couple key websites for team statistics. The first site I will use is NFL.com. From here, I will pull offensive, defensive, and special teams data for all teams since 1970. Then I will join this data will data scraped from pro-football-reference.com, which will include data on wins and losses for each team per season. From this second site, I should also be able to determine teams that were in the Super Bowl (and who won and the score). Once I have all the data, I will clean it and organize it in a way that can be used for data analysis and modelling.

|  |  |  |
| --- | --- | --- |
| **NFL.com Team Stats** | | |
| **Offensive Stats** | **Defensive Stats** | **Special Teams** |
| Passing | Passing | Field Goals |
| Rushing | Rushing | Scoring |
| Receiving | Receiving | Kickoffs |
| Scoring | Scoring | Kickoff Returns |
| Downs | Tackles | Punting |
|  | Downs | Punt Returns |
|  | Fumbles |  |
|  | Interceptions |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pro-football-reference.com Schedule & Game Results – Kansas City Chiefs 2019** | | | | | | | | |
| **Week** | **Day** | **Date** |  | **Rec** |  | **Opp** | **Tm** | **Opp** |
| **1** | Sun | 8-Sep | W | Jan-00 | @ | Jacksonville Jaguars | 40 | 26 |
| **2** | Sun | 15-Sep | W | Feb-00 | @ | Oakland Raiders | 28 | 10 |
| **3** | Sun | 22-Sep | W | Mar-00 |  | Baltimore Ravens | 33 | 28 |
| **4** | Sun | 29-Sep | W | Apr-00 | @ | Detroit Lions | 34 | 30 |
| **5** | Sun | 6-Oct | L | 1-Apr |  | Indianapolis Colts | 13 | 19 |
| **6** | Sun | 13-Oct | L | 2-Apr |  | Houston Texans | 24 | 31 |
| **7** | Thu | 17-Oct | W | 2-May | @ | Denver Broncos | 30 | 6 |
| **8** | Sun | 27-Oct | L | 3-May |  | Green Bay Packers | 24 | 31 |
| **9** | Sun | 3-Nov | W | 3-Jun |  | Minnesota Vikings | 26 | 23 |
| **10** | Sun | 10-Nov | L | 4-Jun | @ | Tennessee Titans | 32 | 35 |
| **11** | Mon | 18-Nov | W | 4-Jul | @ | Los Angeles Chargers | 24 | 17 |
| **12** |  |  |  |  |  | Bye Week |  |  |
| **13** | Sun | 1-Dec | W | 4-Aug |  | Oakland Raiders | 40 | 9 |
| **14** | Sun | 8-Dec | W | 4-Sep | @ | New England Patriots | 23 | 16 |
| **15** | Sun | 15-Dec | W | 4-Oct |  | Denver Broncos | 23 | 3 |
| **16** | Sun | 22-Dec | W | 4-Nov | @ | Chicago Bears | 26 | 3 |
| **17** | Sun | 29-Dec | W | 4-Dec |  | Los Angeles Chargers | 31 | 21 |
|  |  | Playoffs |  |  |  |  |  |  |
| **Division** | Sun | 12-Jan | W | 13-4 |  | Houston Texans | 51 | 31 |
| **Conf. Champ.** | Sun | 19-Jan | W | 14-4 |  | Tennessee Titans | 35 | 24 |
| **SuperBowl** | Sun | 2-Feb | W | 15-4 | N | San Francisco 49ers | 31 | 20 |

# Methods

For this project I will use web scraping, web APIs if available and accessible and flat file exports to retrieve data on the NFL teams since 1970. I will clean and prepare the data for data analysis by joining datasets and removing data features I don’t need.

I will attempt to train and test three models for this project: Logistic Regression, Random Forest and Support-Vector Machine. Unlike other projects have typically outputted a single prediction (True or False, or some numeric prediction), the predictions for this project will output the top two teams predicted to make it to the playoffs. I am still figuring out exactly how to do this and what it will really look like – it might be a list of all teams and their likelihood or percent favoritism for the big game.

If I have time or if I can set up my project as I code properly, I would like to build my model behind a client application that can accept inputs and/or a file upload of current season statistics so predictions can be made with new input data.

# Ethical Considerations

My main concerns are with data collection. I am not exactly sure what the policies are for the sites I plan to get my data. I started reading the policies but there is some wording I am still working through understanding. I do not want to infringe on copyright licenses or anything like that. If I run into any issues, I will find an existing dataset whether it be for a different sport (such as baseball or basketball) or switch leagues (from NFL to College level football, i.e.).

# Challenges

Along with the considerations for if I can legally scrape the data from the sites I plan to use, the next concern will be with collecting the data. For both sites, I will have to pull the data for each team separately. Once I figure out how to get one team’s data, I should be able to all teams’ data easily by scripting out a function and running it for each team.

# References

*Official site of the National Football League*. NFL.com. (n.d.). Retrieved January 14, 2022, from https://www.nfl.com/stats/team-stats/

*Pro Football Statistics and history*. Pro. (n.d.). Retrieved January 14, 2022, from https://www.pro-football-reference.com/