Super Bowl Prediction Analysis

DSC 680 Project 2 Final

Isabella Sturm

2022 February 6

Business Problem

The NFL is a business, and the players and coaches make a lot of money, with salaries ranging depending on position and years of experience. For teams that make it to the Super Bowl, there is additional income that goes to both the winning and losing teams.

To be one of the two teams in the Super Bowl, it is important to understand what really impacts a team's chances of making it. Then using past and current season statistics, chances of making it to the Super Bowl can determined and teams can focus on weak points to increase their chances of making it to and winning the Super Bowl.

Background

The Super Bowl started in 1966 when the NFL Commissioner and the owner of the AFL's Kansas City Chiefs negotiated a merger agreement stating the two leagues, National Football League (NFL) and American Football League (AFL), would formally join together in 1970 but until then, the two champions from the two leagues would play head-to-head in a "Super Bowl".

Unlike other major sporting games (such as in MLB and NBA), whose viewership is largely dependent on who is playing or what big stars are in the game, the Super Bowl is not affected by these factors. Year over year, the Super Bowl has millions of viewers regardless of who makes it into the big game.

The purpose of this project is to predict which two teams will make it into the Super Bowl. While the two teams in the game do not really change how many people will watch the game, it is still a major question that comes up every year with many different businesses and

sources producing algorithms to predict not only who will make it to the Super Bowl but who will win.

Part of the question this project aims to answer is not only who will make the Super Bowl or win the Super Bowl but also determine which factors greatly influence the teams that make it.

Data Explanation

Data for this project came from two sources, NFL.com and pro-football-referece.com. Using Python and the package pandas, I used read_html to pull data from the two sites for each team.

For the NFL data I had to pull data from every team, year available (1970 through 2021), for each team type (offense, defense, and special teams) and statistic types (ie. passing, rushing, fumbles, field goals, etc.). For each year there is a single row for each team; I concatenated each statistic onto the team's data frame for that year then appended the year data frame to the main data frame. Once the entire data frame was constructed, I exported it to CSV to have a historical record of the data since it seems the date range available shifts by one year every time a new season year is added.

For the Pro Football Reference data, I had to read data for each team for each year, but only from one table on the site – the Schedule & Game Results. I added the team table to a current year data frame then appended the current year data frame to a main data frame then exported the main data frame to a CSV.

To prepare the data, I renamed the values in the Team column in the imported NFL data source so the two sources matched – the NFL name duplicated the Team Name (ex. Bills Bills). I removed unneeded columns from both sources and then created a new data frame to summarize the statistics in the Pro Football Reference dataset so there would only be one row per year per team to match the already summarized data frame from NFL.

From there, I was able to merge the two data frames and start exploring the data.

NFL Data

Imported data has 1528 rows and 178 columns, several of which have been removed. Remaining are listed below. The date range of the NFL data is from 1970 through 2021.

COLUMN	DESCRIPTION
TEAM	NFL Team Name
OFF_PASSING_ATT	Offensive Passing Attempts
OFF_PASSING_CMP	Offensive Passing Completions
OFF_PASSING_CMP %	Offensive Passing Completion Percentage
OFF_PASSING_YDS/ATT	Offensive Passing Yards per Attempts (avg)
OFF_PASSING_PASS YDS	Offensive Passing Yards
OFF_PASSING_TD	Offensive Passing Touchdown passes
OFF_PASSING_INT	Offensive Passing Interceptions Thrown
OFF_PASSING_RATE	Offensive Passing Rate
OFF_PASSING_LNG	Offensive Passing Longest Run
OFF_PASSING_SCK	Offensive Passing Sacks
OFF_PASSING_SCKY	Offensive Passing Sack Yards Lost
OFF_RUSHING_ATT	Offensive Rushing Attempts
OFF_RUSHING_RUSH YDS	Offensive Rushing Yards
OFF_RUSHING_YPC	Offensive Rushing Yards Per Carry
OFF_RUSHING_TD	Offensive Rushing Touchdowns

OFF_RUSHING_LNG	Offensive Rushing Longest Completed Pass
OFF_RUSHING_RUSH FUM	Offensive Rushing Fumbles
OFF_RECEIVING_REC	Offensive Receiving Receptions
OFF_RECEIVING_YDS	Offensive Receiving Yards
OFF_RECEIVING_YDS/REC	Offensive Receiving Yards per Reception
OFF_RECEIVING_TD	Offensive Receiving Touchdowns
OFF_RECEIVING_LNG	Offensive Receiving Longest Reception
OFF_RECEIVING_REC FUM	Offensive Receiving Fumbles
OFF_SCORING_RSH TD	Offensive Scoring Rush Touchdowns
OFF_SCORING_REC TD	Offensive Scoring Receiving Touchdowns
OFF_SCORING_TOT TD	Offensive Scoring Total Touchdowns
OFF_SCORING_2-PT	Offensive Scoring 2-point Conversions
DEF_PASSING_ATT	Defensive Passing Attempts
DEF_PASSING_CMP	Defensive Passing Completions
DEF_PASSING_CMP %	Defensive Passing Completion Percentage
DEF_PASSING_YDS/ATT	Defensive Yards per Attempt
DEF_PASSING_YDS	Defensive Passing Intercepted Returned Yards
DEF_PASSING_TD	Defensive Passing Interception Touchdowns
DEF_PASSING_INT	Defensive Passing Interceptions
DEF_PASSING_RATE	Defensive Passing Rate
DEF_PASSING_LNG	Defensive Passing Longest Interception Returned
DEF_PASSING_SCK	Defensive Passing Total Sacks
DEF_RUSHING_ATT	Defensive Rushing Attempts
DEF_RUSHING_RUSH YDS	Defensive Rushing Yards
DEF_RUSHING_YPC	Defensive Rushing Yards per Carry
DEF_RUSHING_TD	Defensive Rushing Touchdowns
DEF_RUSHING_LNG	Defensive Rushing Longest Interceptions Returned
DEF_RUSHING_RUSH FUM	Defensive Rushing Fumbles
DEF_RECEIVING_REC	Defensive Receiving Receptions

DEF_RECEIVING_YDS	Defensive Receiving Yards
DEF_RECEIVING_YDS/REC	Defensive Receiving Yards per Reception
DEF_RECEIVING_TD	Defensive Receiving Touchdowns
DEF_RECEIVING_LNG	Defensive Receiving Longest Interception Returned
DEF_RECEIVING_REC FUM	Defensive Receiving Fumbles
DEF_RECEIVING_PDEF	Defensive Receiving Pass Defended
DEF_SCORING_FR TD	Defensive Scoring Forced Fumble and Recovery that lead to Touchdown
DEF_SCORING_SFTY	Defensive Scoring Safety
DEF_SCORING_INT TD	Defensive Scoring Interceptions that Lead to Touchdown
DEF_TACKLES_SCK	Defensive Tackles Sacks
DEF_TACKLES_COMB	Defensive Tackles Combined Tackles
DEF_TACKLES_ASST	Defensive Tackles Assisted Tackles
DEF_TACKLES_SOLO	Defensive Tackles Unassisted Tackles
DEF_FUMBLES_FF	Defensive Forced Fumbles
DEF_FUMBLES_FR	Defensive Forcred Fumbles and Recovery
DEF_FUMBLES_FR TD	Defensive Forced Fumble and Recovery that Lead to Touchdown
DEF_FUMBLES_REC TD	Defensive Fumbles Reception Touchdowns
DEF_FUMBLES_REC FUM	Defensive Reception Fumbles
DEF_FUMBLES_RUSH FUM	Defensive Rush Fumbles
DEF_INTERCEPTIONS_INT	Defensive Interceptions
DEF_INTERCEPTIONS_INT TD	Defensive Interceptions that Lead to Touchdown
DEF_INTERCEPTIONS_INT YDS	Defensive Interception Yards Gained
DEF_INTERCEPTIONS_LNG	Defensive Longest Interception
YEAR	Year/Season

Pro Football Reference Data

There are 1381 rows with 36 features, two of which are target variables, depending on what is being predicted (make the Super Bowl or win the Super Bowl). The date range for the Pro Football Reference data is from 1976 through 2021.

COLUMN	DESCRIPTIONS
TEAM	NFL Team Name
YEAR	Year/Season
NUM_WINS	Number of Wins for the season (incl playoff games)
NUM_LOSSES	Number of Losses for the season (incl playoff games)
MIN_WIN_SCORE	Minimum score that resulted in a win
MAX_WIN_SCORE	Maximum score that resulted in a win
AVG_WIN_SCORE	Average score that resulted in win
MIN_LOSE_SCORE	Minimum score that resulted in a loss
MAX_LOSE_SCORE	Maximum score that resulted in a loss
AVG_LOSE_SCORE	Average score that resulted in a loss
MIN_WIN_OFF_PASS_YDS	Minimum offensive passing yards that resulted in a win
MAX_WIN_OFF_PASS_YDS	Maximum offensive passing yards that resulted in a win
AVG_WIN_OFF_PASS_YDS	Average offensive passing yards that resulted in a win
MIN_WIN_OFF_RUSH_YDS	Minimum offensive rushing yards that resulted in a win
MAX_WIN_OFF_RUSH_YDS	Maximum offensive rushing yards that resulted in a win
AVG_WIN_OFF_RUSH_YDS	Average offensive rushing yards that resulted in a win
MIN_WIN_DEF_PASS_YDS	Minimum defensive passing yards that resulted in a win
MAX_WIN_DEF_PASS_YDS	Maximum defensive passing yards that resulted in a win
AVG_WIN_DEF_PASS_YDS	Average defensive passing yards that resulted in a win
MIN_WIN_DEF_RUSH_YDS	Minimum defensive rushing yards that resulted in a win
MAX_WIN_DEF_RUSH_YDS	Maximum defensive rushing yards that resulted in a win
AVG_WIN_DEF_RUSH_YDS	Average defensive rushing yards that resulted in a win
MIN_LOSE_OFF_PASS_YDS	Minimum offensive passing yards that resulted in a loss

MAX_LOSE_OFF_PASS_YDS	Maximum offensive passing yards that resulted in a loss
AVG_LOSE_OFF_PASS_YDS	Average offensive passing yards that resulted in a loss
MIN_LOSE_OFF_RUSH_YDS	Minimum offensive rushing yards that resulted in a loss
MAX_LOSE_OFF_RUSH_YDS	Maximum offensive rushing yards that resulted in a loss
AVG_LOSE_OFF_RUSH_YDS	Average offensive rushing yards that resulted in a loss
MIN_LOSE_DEF_PASS_YDS	Minimum defensive passing yards that resulted in a loss
MAX_LOSE_DEF_PASS_YDS	Maximum defensive passing yards that resulted in a loss
AVG_LOSE_DEF_PASS_YDS	Average defensive passing yards that resulted in a loss
MIN_LOSE_DEF_RUSH_YDS	Minimum defensive rushing yards that resulted in a loss
MAX_LOSE_DEF_RUSH_YDS	Maximum defensive rushing yards that resulted in a loss
AVG_LOSE_DEF_RUSH_YDS	Average defensive rushing yards that resulted in a loss
SUPERBOWL	T/F, whether the team made it to the Super Bowl
SUPERBOWLWINNER	T/F, whether the team won the Super Bowl

The two datasets are related on column Team.

Methods

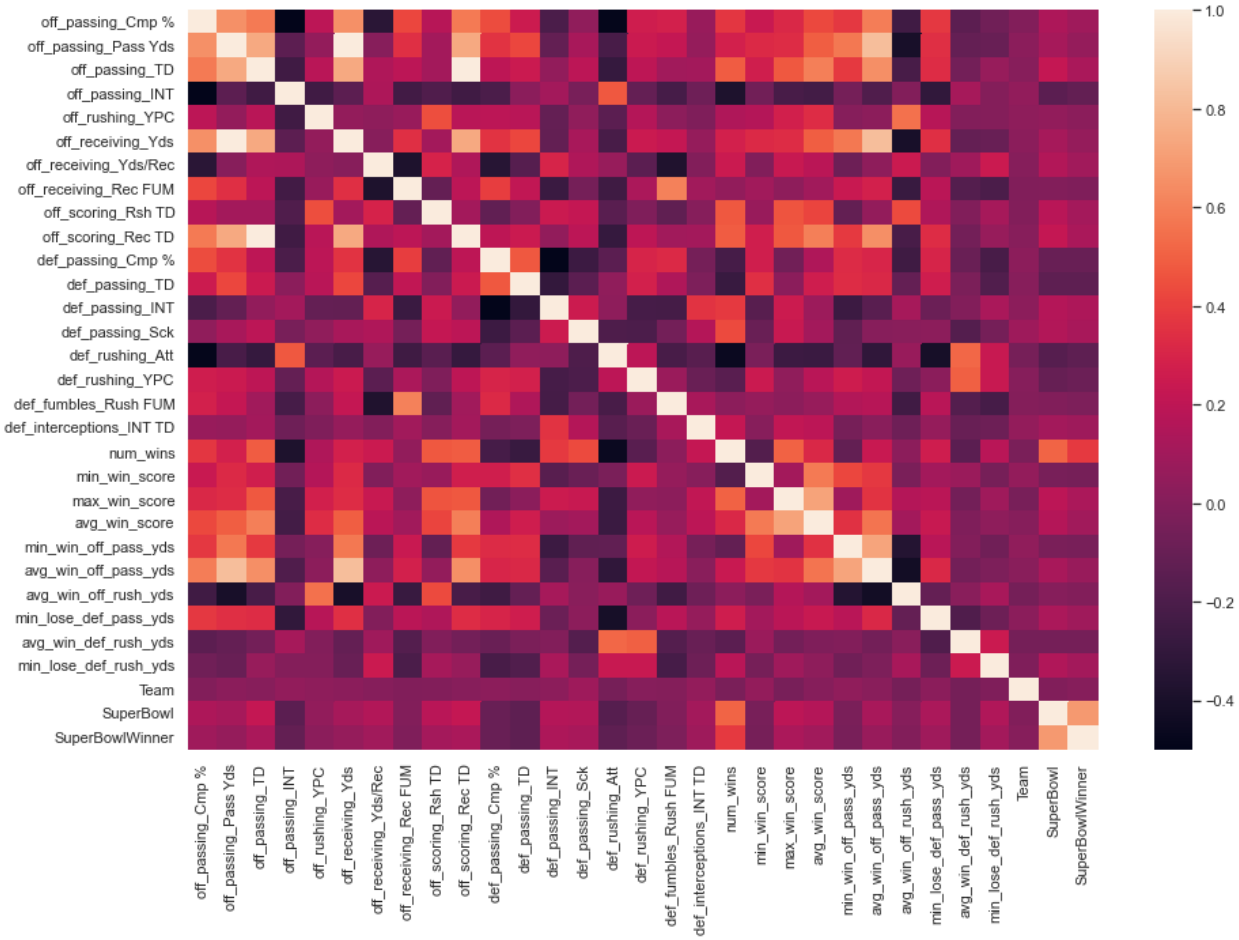
For this project I used web scraping to retrieve data on the NFL teams since 1970 from nfl.com and data since 1976 from pro-football-reference.com. I removed columns from the data sets I didn't need, aggregated the Pro Football Reference data to summarize game statistics (number of wins/losses, min/max/mean score, etc), and then joined the two datasets on Team and Year.

I trained and tested 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

model predictions for who will play in the Super Bowl and include all teams with the model's probabilities for each team.

Analysis

First, I had to determine which features I would use for the models. After joining the two datasets, I had 104 columns (102 features – two of the columns are targets: one for who will play in the Super Bowl which was used for this project and the other is who will win the Super Bowl which could be used for a future related project). Due to my limited knowledge of football, I could not rely on my business acumen to help determine features. Instead, I used SelectKBest to determine the top 30 features from the merged dataset.



Then, I began modelling. For each model, I used StandardScaler to scale and standardize my data before fitting the data to each model. All three models had high accuracy – 94% and above. The Support Vector Machine (SVM) gave an accuracy of 94.026%; Logistic Regression resulted in a model with 95.796%; and Random Forest gave 96.681% accuracy.

To put the three models to the test, I pulled data for the current season (which was not used to train the model) and prepared it to be inputted into each model for predictions. Below are the results for each model: the output of the model includes each team and the model prediction for if they will play in the Super Bowl (Prediction), whether or not the model predicted accurately (Score), the probability the model gave for not making it to the Super Bowl (Probability_False), and the probability the model gave for the team to make it to the Super Bowl (Probability_True).

Support Vector Machine

Team	Prediction	Score	Probability_False	Probability_True
Rams	0	0	0.137367	0.862633
Bengals	0	0	0.606978	0.393022
Cowboys	0	1	0.815357	0.184643
Buccaneers	0	1	0.85498	0.14502
49ers	0	1	0.94965	0.05035
Bills	0	1	0.949708	0.050292
Raiders	0	1	0.949734	0.050266
Chiefs	0	1	0.957197	0.042803
Packers	0	1	0.966025	0.033975
Titans	0	1	0.968072	0.031928
Falcons	0	1	0.974661	0.025339
Patriots	0	1	0.976233	0.023767
Chargers	0	1	0.977285	0.022715
Jaguars	0	1	0.987106	0.012894
Eagles	0	1	0.988964	0.011036
Steelers	0	1	0.989526	0.010474
Cardinals	0	1	0.991595	0.008405
Seahawks	0	1	0.992942	0.007058
Lions	0	1	0.993341	0.006659
Ravens	0	1	0.994035	0.005965
Jets	0	1	0.994364	0.005636
Giants	0	1	0.994835	0.005165
Saints	0	1	0.994989	0.005011
Colts	0	1	0.995064	0.004936
Panthers	0	1	0.996558	0.003442
Bears	0	1	0.996669	0.003331
Vikings	0	1	0.996775	0.003225
Football Team	0	1	0.997838	0.002162
Texans	0	1	0.998011	0.001989
Dolphins	0	1	0.998583	0.001417
Browns	0	1	0.999193	0.000807
Broncos	0	1	0.99973	0.00027

Logistic Regression

Team	Prediction	Score	Probability_False	Probability_True
Rams	1	1	0.21602	0.78398
Buccaneers	0	1	0.77397	0.22603
Chiefs	0	1	0.84532	0.15468
Bengals	0	0	0.90277	0.09723
Cowboys	0	1	0.92483	0.07517
Bills	0	1	0.94581	0.05419
Packers	0	1	0.97687	0.02313
Titans	0	1	0.97707	0.02293
Raiders	0	1	0.98277	0.01723
49ers	0	1	0.99191	0.00809
Cardinals	0	1	0.99458	0.00542
Patriots	0	1	0.99745	0.00255
Eagles	0	1	0.99782	0.00218
Chargers	0	1	0.99842	0.00158
Saints	0	1	0.99908	0.00092
Dolphins	0	1	0.99929	0.00071
Ravens	0	1	0.99964	0.00036
Falcons	0	1	0.99969	0.00031
Colts	0	1	0.99971	0.00029
Steelers	0	1	0.99972	0.00028
Vikings	0	1	0.99978	0.00022
Seahawks	0	1	0.99978	0.00022
Broncos	0	1	0.99988	0.00012
Football Team	0	1	0.99993	0.00007
Browns	0	1	0.99996	0.00004
Texans	0	1	0.99998	0.00002
Panthers	0	1	0.99999	0.00001
Giants	0	1	0.99999	0.00001
Bears	0	1	0.99999	0.00001
Jets	0	1	0.99999	0.00001
Lions	0	1	1	0
Jaguars	0	1	1	0

Random Forest

Team	Prediction	Score	Probability_False	Probability_True
Rams	1	1	0.22	0.78
Bengals	1	1	0.36	0.64
Cowboys	0	1	0.81	0.19
Chiefs	0	1	0.82	0.18
Bills	0	1	0.91	0.09
Buccaneers	0	1	0.92	0.08
Patriots	0	1	0.95	0.05
Cardinals	0	1	0.95	0.05
Chargers	0	1	0.97	0.03
Packers	0	1	0.97	0.03
Browns	0	1	0.98	0.02
49ers	0	1	0.98	0.02
Seahawks	0	1	0.99	0.01
Texans	0	1	0.99	0.01
Vikings	0	1	0.99	0.01
Raiders	0	1	0.99	0.01
Saints	0	1	0.99	0.01
Colts	0	1	0.99	0.01
Dolphins	0	1	0.99	0.01
Ravens	0	1	1	0
Panthers	0	1	1	0
Falcons	0	1	1	0
Lions	0	1	1	0
Football Team	0	1	1	0
Giants	0	1	1	0
Eagles	0	1	1	0
Titans	0	1	1	0
Jaguars	0	1	1	0
Bears	0	1	1	0
Jets	0	1	1	0
Steelers	0	1	1	0
Broncos	0	1	1	0

Conclusion

Overall, all three models ended up with high accuracies for their models. When looking at the results of the Super Bowl LVI predictions and probabilities it becomes clearer that the models, even if predicted inaccurately at times still had some degree of confidence. For instance, the SVM model didn't predict any team for the Super Bowl, but looking at the probabilities it shows ~83% for the Rams, which is one of the teams that will be playing in the Super Bowl 2022.

There are few things I would like to do to improve the model and the overall ease of using this model. Balancing the data could improve the models further and ensure the accuracies are a good representation of the models' performances. I would also like to clean up the data and add a front-end client for uploading data with current season statistics and outputting the predictions in the table format with the probabilities.

Assumptions

There are a few assumptions made to accomplish this project. First, I do not consider home games versus away games. Also, I do not compare scores per game or stats/scores between teams that have played. All the data is summarized to represent how each team did overall per season.

Limitations

The main limitation I ran into for this project is the lack of business acumen. Working with more than 25-50 features makes it difficult to know what to look at especially without

knowledge about football to use a guide. I had to rely on scikit-learn and the SelectKBest class to help determine the best features in the dataset as it relates to the target SuperBowl (whether a team made it to the Super Bowl).

Challenges

The first challenge in this project was collecting the data. Not only did I have to pull data from two sources, I had to filter and summarize the data and join the two datasets for analysis. I had some challenges getting a summarized data frame of the pro football reference data.

Another challenge I had was working with so many features – this is the first project I have worked on that has had more than 25 features.

Lastly, a third challenge is figuring out how to create a model that will answer the question "who will play in the Super Bowl" instead of "who will win". I have never had more than one output for a model.

Future Uses/Additional Applications

The initial goal of this project was to predict who will play in the Super Bowl. Other applications could also be to predict who would win the Super Bowl. You might even be able to use this data to predict who will win a particular game, however that might require using some of the game data I removed.

Recommendations

Because there are a lot of features in the dataset, it would be nice to have a front-end interface for either inputting data needed to make a prediction. Alternatively, the front-end

could allow an upload of a file that contains either the data as is or that can be parsed and aggregated if needed in a way that the data could be used as input for the model.

Implementation Plan

The implementation plan for this project involves a few steps:

1. Yearly, we can retrain and tune the model with new data from the two sites (NFL and Pro Football Reference)
2. New functions would need to be created in order to have a systematic way of taking in and formatting new data to be used in training the model
3. A front-end would be created to upload data for predictions and the backend would run model predictions
4. New functions would need to be put in place to handle the inputted data, whether that be manual entries or a file upload.

Ethical Assessment

I was initially concerned with if I was going to be able to scrape the data from the two sites I used. I reviewed their policies but had some difficulty understanding some of the wording. On the Pro Football Reference site, they encourage people to pull the data themselves, so that added some reassurances.

Other than that, there is no player data so there is no personal information at risk of causing any ethical issues. Also, because this type of analysis is done every year by multiple different people/companies/businesses, and the predictions made cannot inherently change the outcomes of games played, I don't see any ethical concerns of this project and analysis.

References

ESPN Internet Ventures. (2005, August 8). *NFL statistics glossary*. ESPN. Retrieved February 6, 2022, from https://www.espn.com/nfl/news/story?id=2128923

The history of the Super Bowl. The History of the Super Bowl | The American Historian. (n.d.). Retrieved February 6, 2022, from https://www.oah.org/tah/other-content/the-history-of-the-super-bowl/

Official site of the National Football League. NFL.com. (n.d.). Retrieved February 6, 2022, from https://www.nfl.com/stats/team-stats/

Team encyclopedias and records. Pro. (n.d.). Retrieved February 6, 2022, from https://www.pro-football-reference.com/teams/