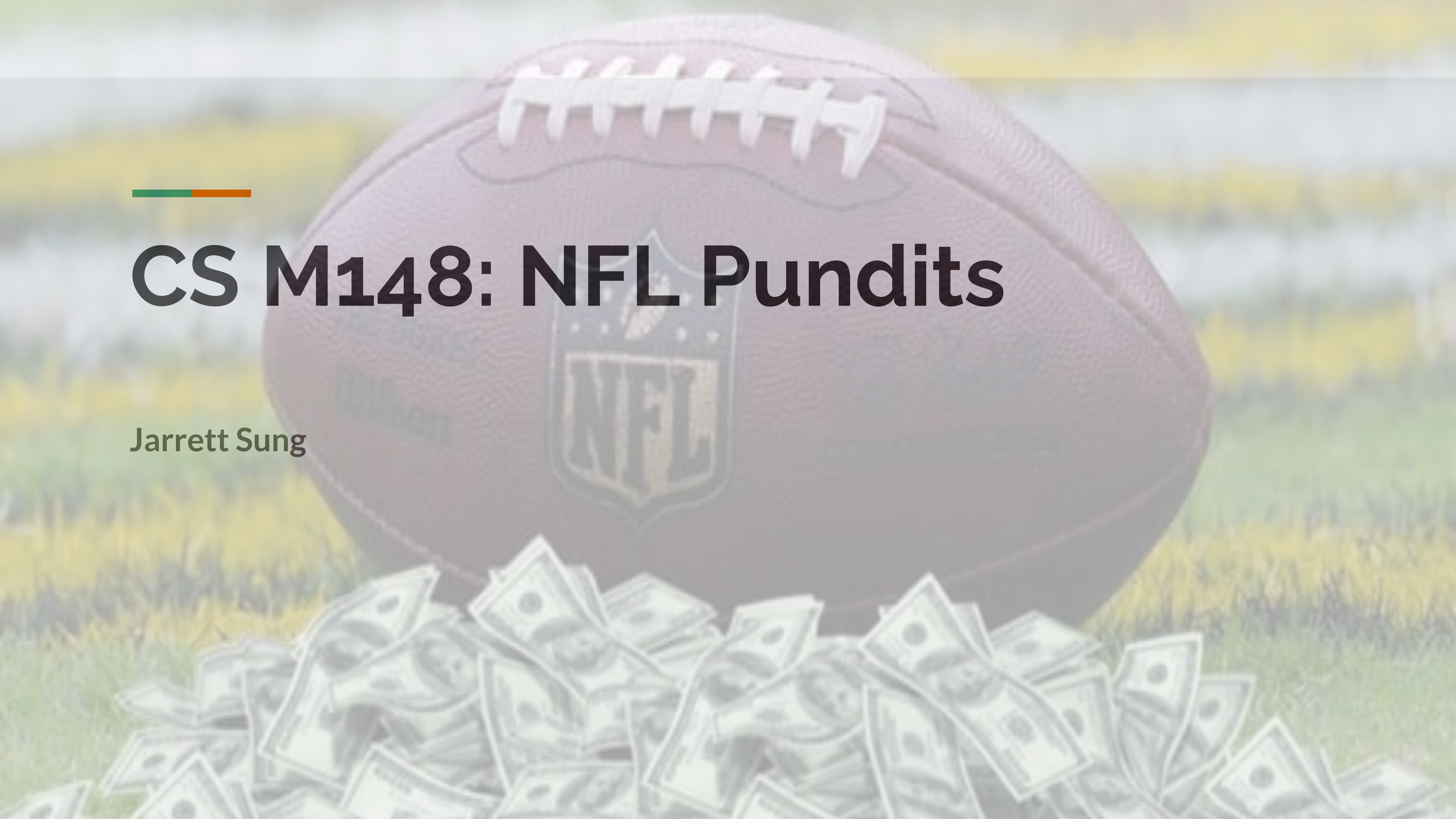


CS M148: NFL Pundits

Jarrett Sung

What and Why?

With this project, I aim to predict the win percentages of each team competing in the National Football League (NFL), likelihood of advancing to the playoffs and the Super Bowl for the years 2021 and 2022, using the statistics of the teams like points scored, allowed, yard differential, etc for the years 2015-2020.

As football enthusiasts, this project is especially important to me because determining win percentages is a problem that lies at the core of the NFL betting market, which is predicted to consist of as many as 73.5M Americans in 2023

Data Collection and Cleaning

Data Description:

Our data consists of eight years (2015-2022) worth of statistics from NFL statistics. Our training data is the first six years: 2015-2020, and our training data is 2021 and 2022. All of our data has been scraped from pro-football-reference.com.

Statistics used:

I used the following stats to base our win predictions on:

Points for, points allowed, total yards, rush yards, pass yards, win percentage, and distance covered

Using all of these stats for the both the team and it's opponents gives me 22 stats in total

Questions to be Answered:

Through our analysis, I aim to answer the following questions:

1. How accurately am I able to predict win percentages for each of the 32 teams in the NFL in the years 2021 and 2022, based on the last 5 years data ?
2. Which factors are most useful in making these predictions? (using feature selection)
3. Which teams are predicted to make it into the playoffs?
4. Which teams are predicted to advance per round?
5. Which team is predicted to win the superbowl?

1. How accurately am I able to predict win percentages for each of the 32 teams in the NFL in the years 2021 and 2022, based on the last 5 years data ?

Method used: Linear Regression (predictor variable: win percentage)

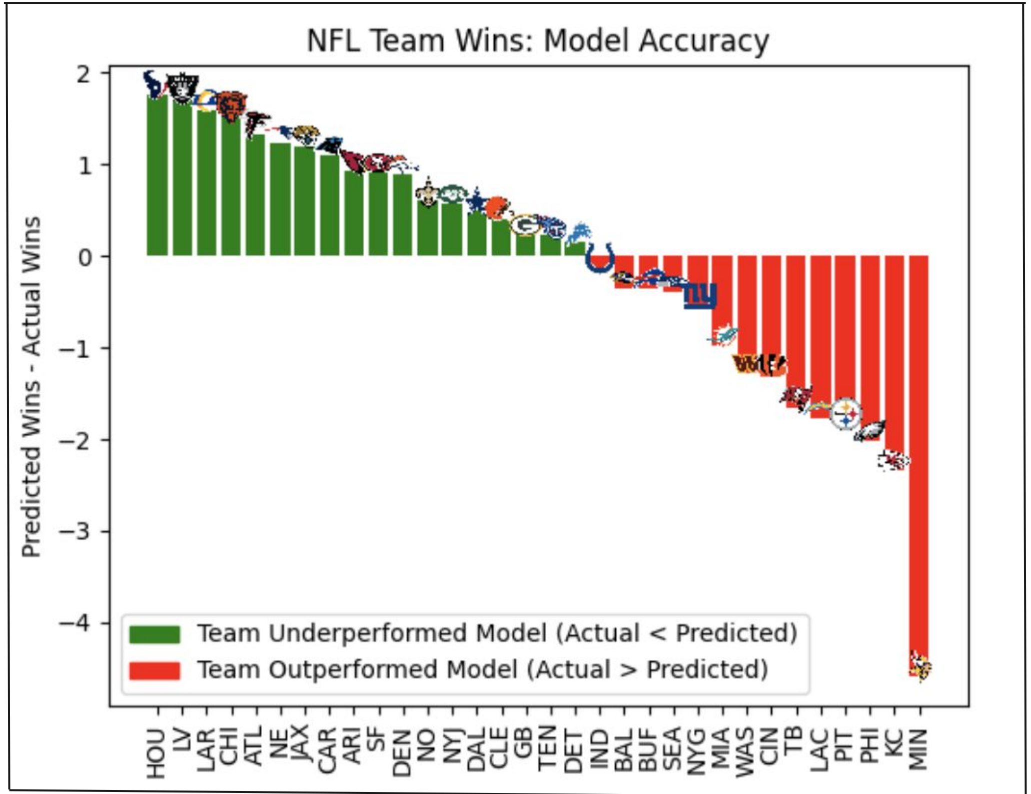
Libraries used: matplotlib, sklearn, pandas, numpy, nfl_data_py

Results: I was able to predict the win percentages for all the teams in the years 2021 and 2022 with an average Mean Squared Error (MSE) of 2.53 games, and an R squared value of 0.726. 76% of the predictions were within 1 win of the actual win totals.

Inference: The win percentages for the league can only be predicted with 72.6% accuracy, even after considering all possible outcomes

MSE:
2.532

R-squared:
0.726


NFL Team Wins: Model Accuracy
Predicted Wins - Actual Wins
2
1
0
−1
−2
−3
−4
🟩 Team Underperformed Model (Actual < Predicted)
🟥 Team Outperformed Model (Actual > Predicted)
HOU
LV
LAR
CHI
ATL
NE
JAX
CAR
ARI
SF
DEN
NO
NYJ
DAL
CLE
GB
TEN
DET
IND
BAL
BUF
SEA
NYG
MIA
WAS
CIN
TB
LAC
PIT
PHI
KC
MIN

2. Which of the 22 Stats are Most Useful in Making Predictions?

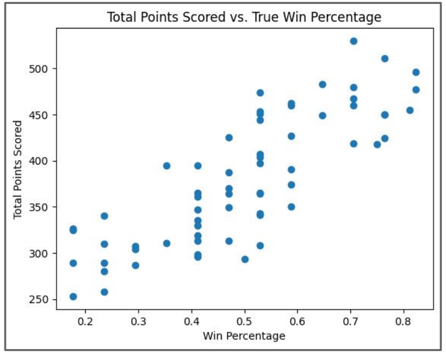
Method used: L1 and L2 Regularization, Feature Selection

Libraries used: matplotlib, sklearn, pandas

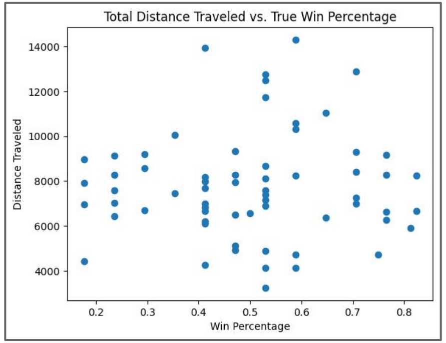
Results: Using feature selection, I analysed the importance of all the stats in predicting win percentages, and found that the top three most useful stats are: Points Scored, Points Allowed, and Total Yards, while the least useful feature is Distance Traveled

Lasso Regression model (α=0.01) was able to achieve similar MSE and R-squared using only 11 of original 22 features

Inference: Anyone looking to place bets or make predictions about the future performance of a team should look at these top three stats for any team.



Total Points Scored vs. True Win Percentage



Total Distance Traveled vs. True Win Percentage

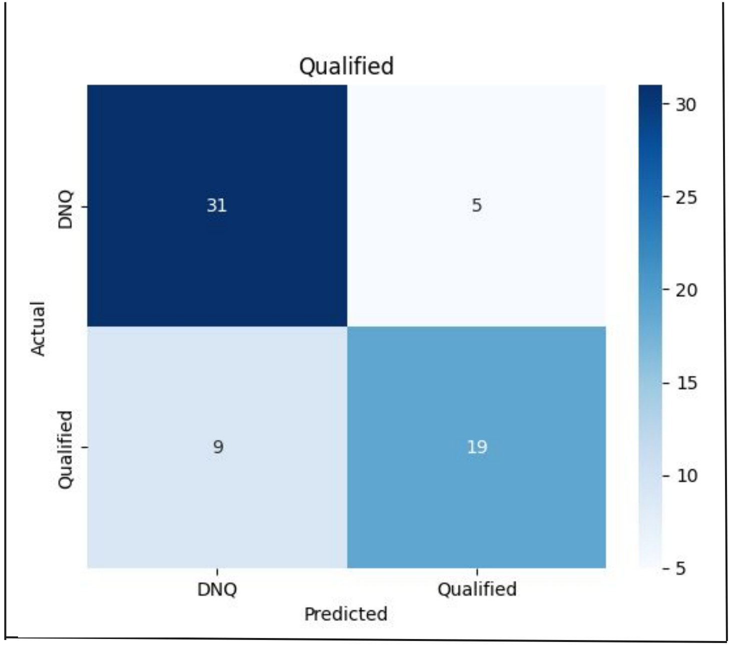
3. Which teams are predicted to make it to the Playoffs?

Method used: Logistic Regression

Libraries used: matplotlib, sklearn, pandas, seaborn

Results: Out of the 64 teams (32 for each year I'm making predictions for), I managed to predict the playoff status of 50 teams correctly, giving me an accuracy of 0.781 power of 0.679, and specificity of 0.861.

Inference: Our model can predict whether or not a team will make it to playoffs roughy 78% of the times.



4. Which teams are predicted to advance per round?

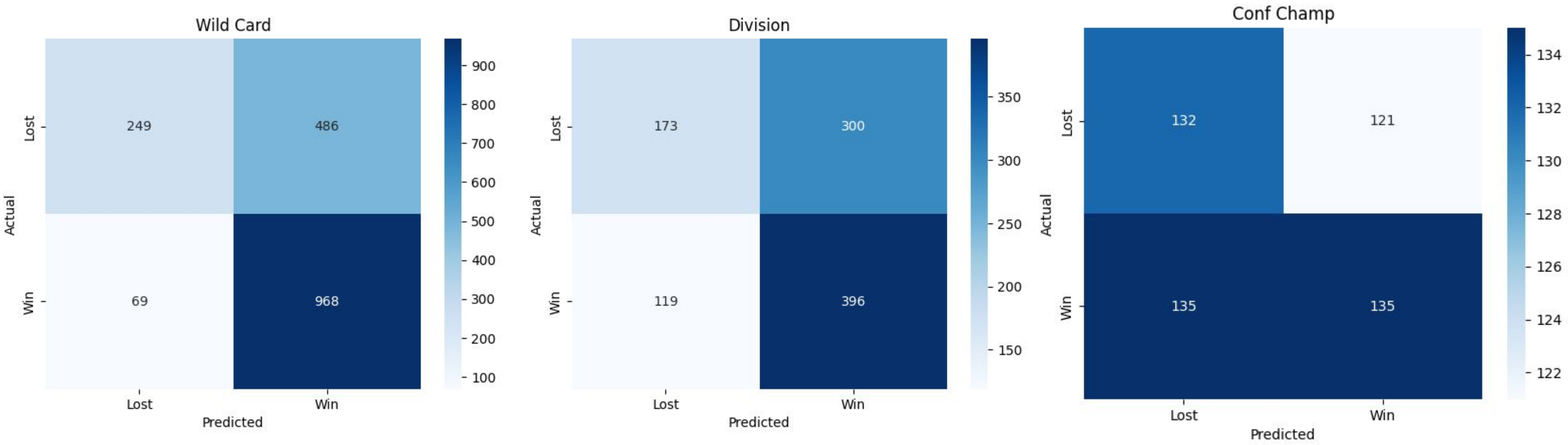
Method used: Logistic Regression, Bootstrapping

Libraries used: matplotlib, sklearn, pandas, seaborn

Results: As the rounds progressed, it became harder to predict winners. In the first round, it predicted about 68.7% of the winners. In the divisional it predicted about 57.6% of the winners, then 51.1% of the winners in the semi-finals.

Inference: Our model can predict the earlier rounds accurately, but later rounds generally become harder to predict due to the smaller disparity between team talent (in the wild card good teams play against weaker teams, eliminating lower seeded teams early on and matching similarly matched teams together later).

4. Which teams are predicted to advance per round?



5. Which Team is Predicted to Win the Super Bowl?

Method used: Logistic Regression

Libraries used: matplotlib, sklearn, pandas, seaborn

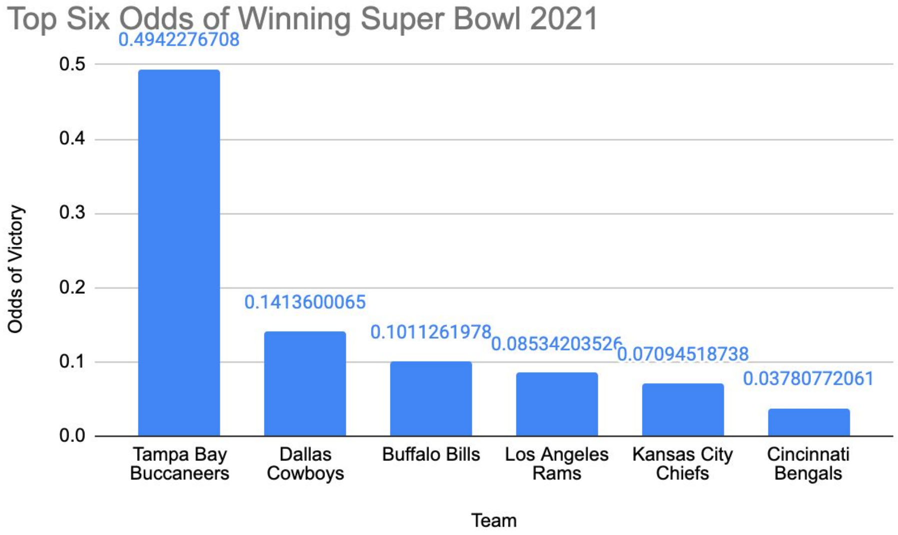
Results: The top six teams with the highest probability per year won a playoff game (except the 2021 Cowboys), all of the semifinalists made the top six (except for the 2021 49ers), the top six contained all finalists, and the Super Bowl winners were within the top six.

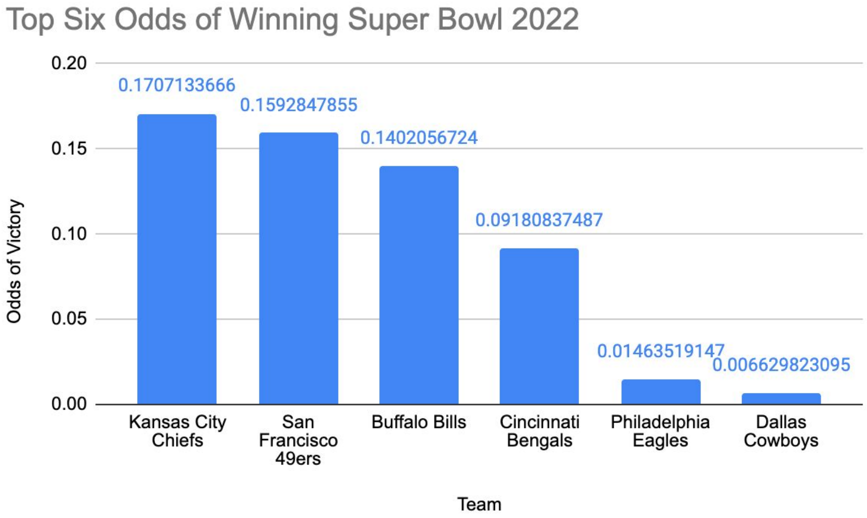
Inference: While it's hard to find a definitive Super Bowl winner, the model found the contenders more accurately than what conventional wisdom and what many pundits predicted.

Top feature: Point Differential

Least Useful: Opponent Yards

5. Which Team is Predicted to Win the Super Bowl?


Top Six Odds of Winning Super Bowl 2021


Top Six Odds of Winning Super Bowl 2022

Relevance and Practical Application

If you are a sports enthusiast, here is how you our model helps you out:

1. Next time you are arguing in favor of your favorite team, the following three statistics are the most useful to make your case: 1. Points Scored 2. Points Allowed 3. Total Yards
2. Between predicting the win percentages and the playoffs, I was able to predict the win percentages with greater accuracy, thus for people looking to place bets in the sports betting market, bets about the win percentages might be safer ones.
3. After running a complex model which built on 5 years’ worth of data, I was still only able to predict the win percentages with 72.6% accuracy. Thus, given the unpredictable nature of the game, I advise everyone to bet responsibly, if at all.

Thank you!