Ninad Dixit **Capstone Project Report**

Predicting the Super Bowl LVI Winner

Summary:

A linear regression model was developed to predict the Super Bowl LVI winner between the Los Angeles Rams and the Cincinnati Bengals. Exploratory data analysis showed time-dependence of several performance features of both teams. However, the degree of dependence for each team varied significantly. 'Ridge', 'Elastic Net', and 'Random Forest' regressors were employed to predict the points scored by each team. Random Forest regressor made the most accurate prediction, which matched the actual score (Rams – 23, Bengals – 20). Several factors were identified that may improve the model performance and prediction accuracy.

Problem: **How can we accurately predict the winner of NFL Super Bowl LVI?**

'Super Bowl' is the annual championship game of the National Football League (NFL). It is played between the winners of National Football Conference (NFC) and American Football Conference (AFC) leagues. It is estimated that 23.2 million people wagered nearly 4.3 billion USD on 'Super Bowl 2020' played between the Tampa Bay Buccaneers and Kansas City Chiefs.[1] Thus, predicting the NFL champion accurately can be financially quite beneficial. Developing a robust model to predict the winner can also highlight the most important factors that make a particular football team successful. It may help teams in devising better game plans and improving recruitment strategy.

Data Wrangling:

The dataset was created from the information available online.[2] The resulting Excel file consisted of 32 sheets, each sheet corresponding to the records each NFL team over the 2021 season (regular season and playoff games).

Attribute information:

The following features were considered for further analysis:

- Week
- Date
- Start_time
- Overtime
- Home_game
- Opponent
- 1^{st}_downs
- Total_yards
- Rushing_yards
- Turnovers
- Opp_1^{st}_downs
- Opp_Total_yards
- Opp_Rushing_yards
- Opp_Turnovers

For this project, the target feature was *'Score'*, which is the points scored by the team of interest. There were several features in the data frame that did not provide any valuable information related to *'Score'*. These columns were: *'Boxscore'*, *'Record'*, and subsets

of *'Expected Points'* i.e., *'Offense_pts'*, *'Defense_pts'*, and *'Sp_teams_pts'*. All these columns were dropped from further analysis.

There were few more columns that were dropped as well:

- *'Passing_yards'* and *'Opp_passing_yards'* - these can be calculated by subtracting the rushing yards from total yards. Thus, these columns don't provide any additional information. On the other hand, a significant correlation between the total and passing yards can affect the model performance. Theoretically, either passing or yards columns can be removed. However, I decided to keep rushing yards since a rushing play takes more time off the clock and has more impact on time management.
- I dropped the *'Day'* column as a detailed timestamp information is provided in the *'Date'* column.
- Since *'Result'* column depends on the difference between points scored by both teams, I dropped it as well.
- *'Opp_score'*, i.e. points scored by the opponent was also removed. We are only interested in the points scored by the team of interest.

NFL games are played under various weather and other conditions depending on the location and month. E.g. southern states usually have warmer temperatures. Thus, games played in these states may be more comfortable for the players. On the other hand, places like Denver are at high altitudes where players can have difficulty in breathing. This may limit offensive play options and, potentially, scoring opportunities. Another prime example can be playoff games at Green Bay, Wisconsin in January that are often played under extremely cold weather and thus can be low scoring. Thus, 'Date' and game venue (depends on the opponent and home/away game) were retained.

Null values in *'Overtime'*, *'Turnovers'*, and *'Opp_Turnovers'* columns were replaced with '0' and the other response was replaced with '1'.

Exploratory Data Analysis:

Some interesting observations from game wise statistics:

- About 8% of the total games went into overtime.
- Teams earned an average of 20 first downs per game.
- Teams committed an average of one turnover per game. However, they did not commit a turnover in at least 25% of the games.
- About 1/3rd of total yards covered by teams came from rushing plays.
- Teams scored an average of 23 points per game, with the lowest score of '0' and highest score of '56'.

Team wise statistics show a clear trend among the teams that qualified for the playoffs:

- Team offense covered more than 350 yards/per game.
- Team defense allowed for covering 330 yards or less per game by the opponents' offense.

Correlation analysis:

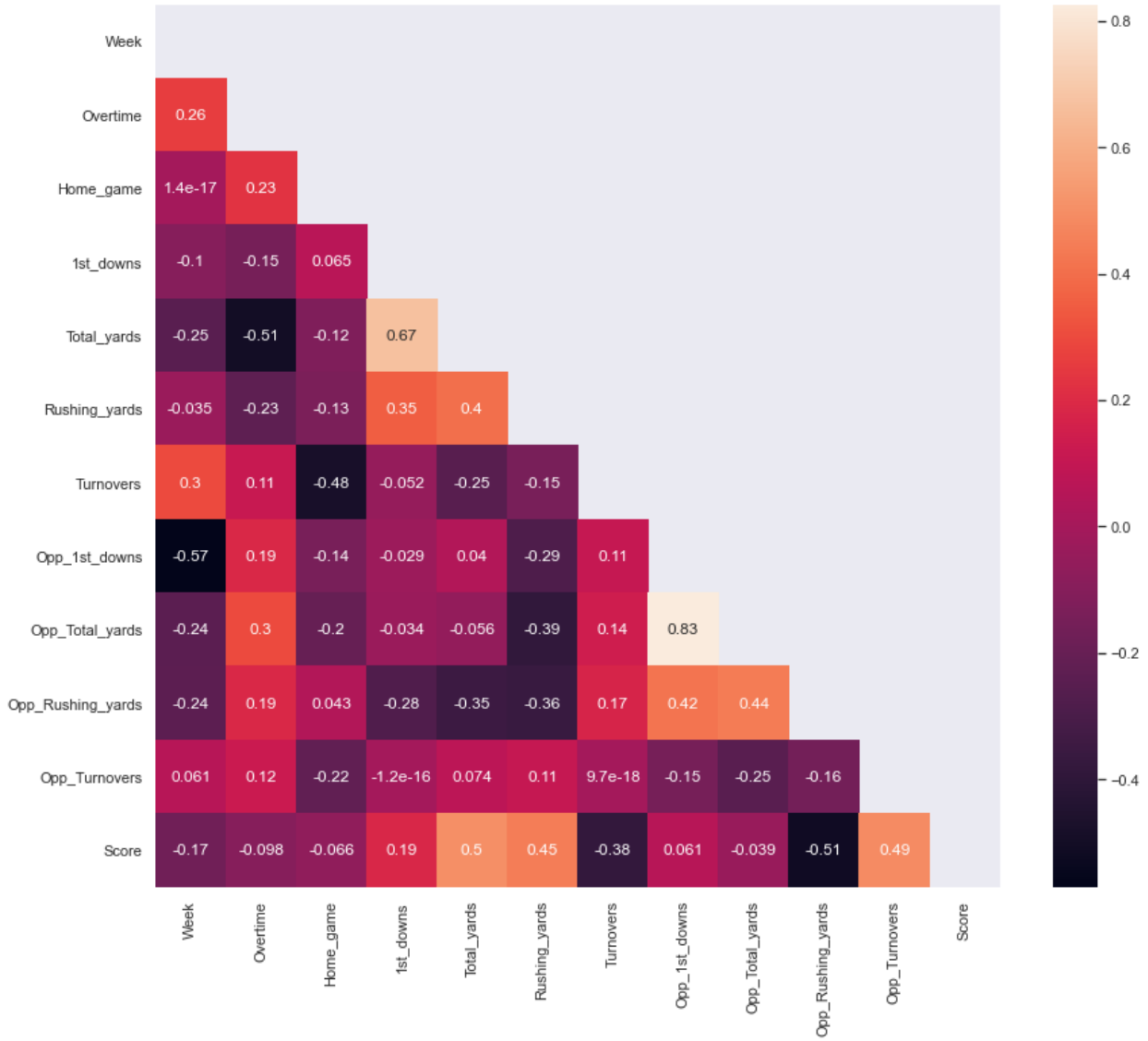

Fig. 1. Heatmap describing correlations between the Rams' numerical features.

Total points scored by the Los Angeles Rams showed significant correlation with the following features:

- *Total_yards (0.5)*
- *Opp_Rushing_yards (-0.51)*
- *Rushing_yards (0.45)*
- *Opp_Turnovers (0.49)*
- *Turnovers (-0.38)*

The Rams also showed some correlation between 'Week' and the following features:

- *Opp_1st_downs* - this is a significant negative correlation. The Rams caused the opposition to earn less first downs later in the season. Since the number of first downs is significantly correlated with the opposition's score, the Rams were more likely to outscore their opponent and win the game.
- *Turnovers* - as weeks passed the Rams tended to commit more turnovers, which are negatively correlated with the points scored. Thus, limiting the number of turnovers is a concern for them.
- *Overtime* - Rams were slightly more likely to play games that go into overtime during the later stages of the season. This trend was expected because they would encounter stronger teams during the playoffs. Such teams may be difficult to outscore.
- *Total_yards* and *Opp_Total_yards* - both features show a weak negative correlation. Negative correlation with total yards suggests that the Rams encountered teams with stronger defenses as the season progressed. On the other hand, the Rams' defense got better at restricting their opponents as shown by the negative correlation with oppositions' total yards.

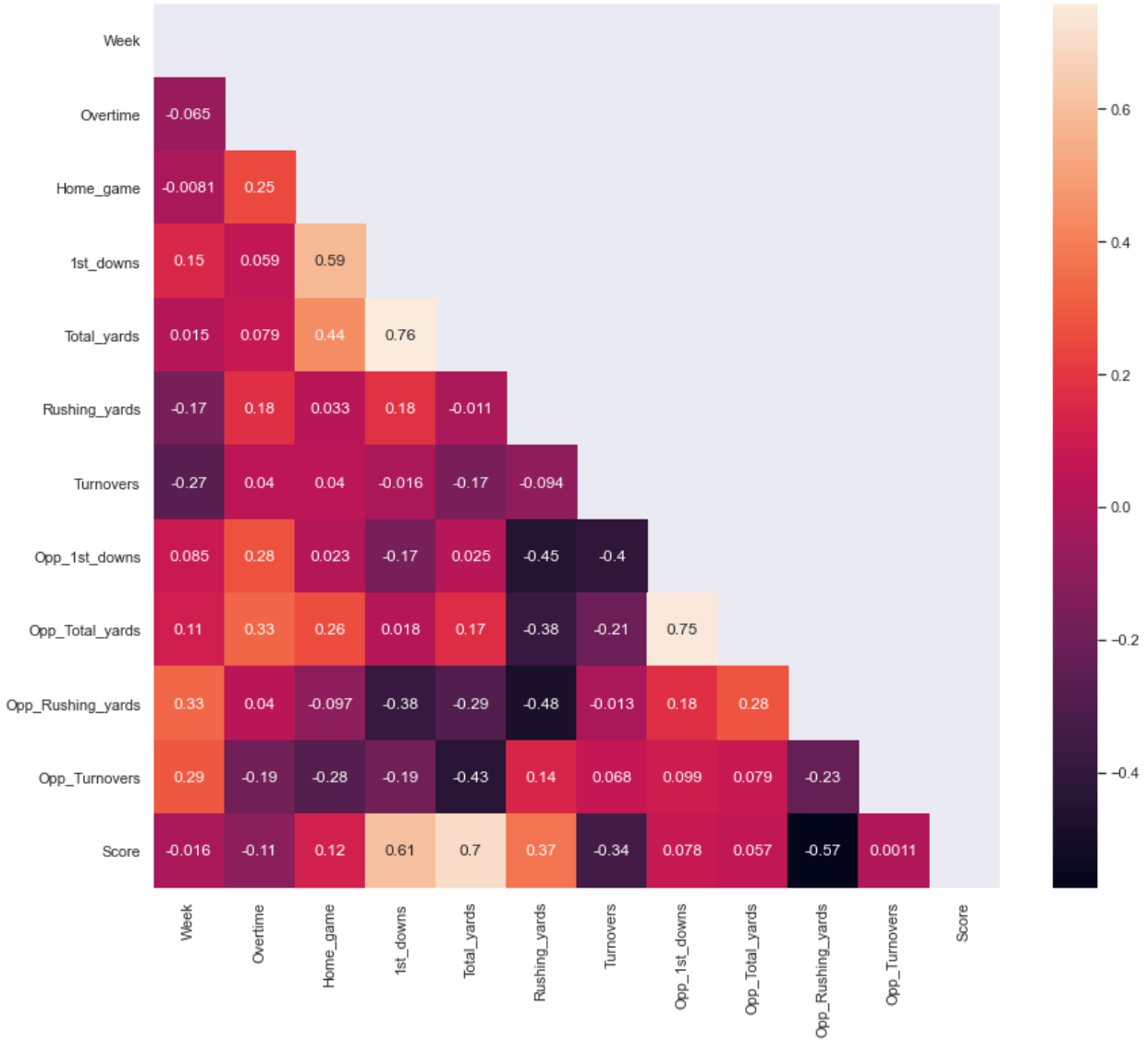

Fig. 2. Heatmap describing correlations between the Bengals' numerical features.

Total points scored by the Cincinnati Bengals showed significant correlation with the following features:

- *Total_yards (0.7)*
- *1st_downs (0.61)*
- *Opp_Rushing_yards (-0.57)*
- *Rushing_yards (0.37)*
- *Turnovers (-0.34)*

The Bengals also showed some correlation between 'Week' and the following features:

- *Opp_Rushing_yards* - the Bengals allowed more rushing yards in a game as the season progressed.

- *Opp_Turnovers* - the Bengals defense caused increasing number of turnovers from their opponents in the later stages of the season.
- *Turnovers* - On the other hand, the Bengals were likely to commit less turnovers as the season progresses as shown by the negative correlation with *'Week'*.

Both teams showed some time-dependency of their performance statistics. Thus, time-series analysis was performed on both data sets.

Time-series analysis:

Autocorrelation and partial autocorrelation function plots for features such as *'1st_downs'* and *'Total_yards'* did not show statistically significant time-dependency. It appears that each team has a distinct set of features that are in correlation with time. Therefore, following another approach for data modeling was necessary.

Data Pre-processing and Model Performance:

Performance statistics for a team consist of first downs earned, total yards covered, and turnovers committed by the team and its opponent.

Thus, the Super Bowl LVI score prediction model involved three steps:

- Estimate performance statistics of the two teams playing the game.
- Develop a linear regression model that accurately determines the scores from previous games played in the season.
- Input the estimated feature values in the model to predict the score for each team.

Estimate performance statistics:

Since the Rams did not play against the Bengals in regular season, a reliable estimate for both teams' performance statistics in the Super Bowl was not available. Thus, I compared the average performance of one team with that of the opponents of the other team to find the closest match. E.g., I took the average performance of Bengals in their three recent games and compared it against the performances of the Rams' opponents in regular season.

This estimation was done using K-means clustering analysis. The Rams' opponents from previous weeks were classified into several clusters. Then, average performance of the Bengals in their last three games was input as 'test data'. This data was labeled using the cluster model to identify the closest match to the Bengals against Rams. Finally, the average of Bengals' last three performance and the Rams' closest matching opponent was used as test data in regression analysis. These steps were also carried out for the Bengals' opponents to estimate their performance statistics.

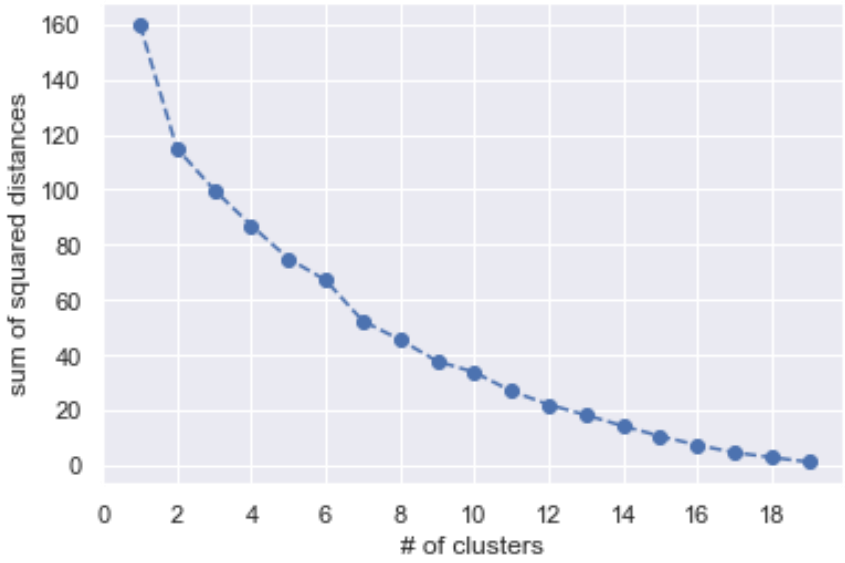

Fig. 3. K-means clustering analysis of the Rams' opponents during regular season.

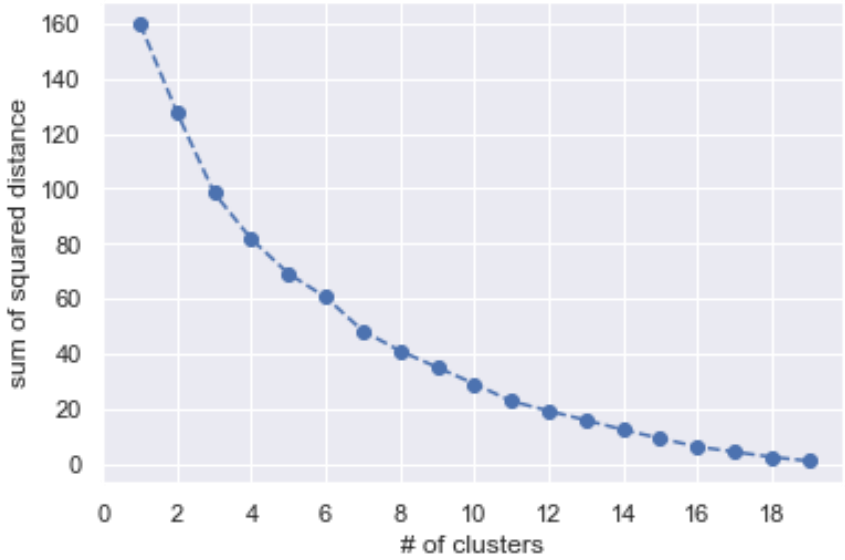

Fig. 4. K-means clustering analysis of the Bengals' opponents during regular season.

'Sum of squared distances' represents distribution of the points in the corresponding cluster. In both cases, as the number of clusters increased, the distribution consistently got tighter. Thus, I chose 19 clusters for each team to assign the label to respective test data. This means only one of the Rams' opponent from regular season was identified as the closest match to the Bengals' average performance over their last three games and vice versa.

The Bengals' performance in the last three games was closest to the Arizona Cardinals' performance against the Rams. The Rams' performance in the last three games was closest to the Baltimore Ravens' performance against the Bengals.

Model selection, training, and cross-validation:

The only categorical feature would be used in this analysis was 'Team', i.e. team names. A data frame with dummy variables corresponding to each team was constructed using Pandas *'get_dummies'* function. Also, the data frame rows were sorted using 'Week'. Since the pre-processed data frame consisted of 45 features and only 568 entries, 'Ridge' and 'ElasticNet' regressors that punish large weights of the features were used. A Random Forest meta-regressor was also employed to estimate the points scored by both team in Super Bowl LVI.

Since the game statistics are generated over many weeks, this data must be considered time-dependent. For cross-validation, it was necessary to make sure that the training set consisted of older data and validation set consisted of the recent data. This was achieved by employing 'TimeSeriesSplit' as the cross-validator. 'TimeSeriesSplit' is similar to 'KFold' cross-validator but uses first 'K' folds as the training set and $(K+1)^{th}$ fold as the validation set.

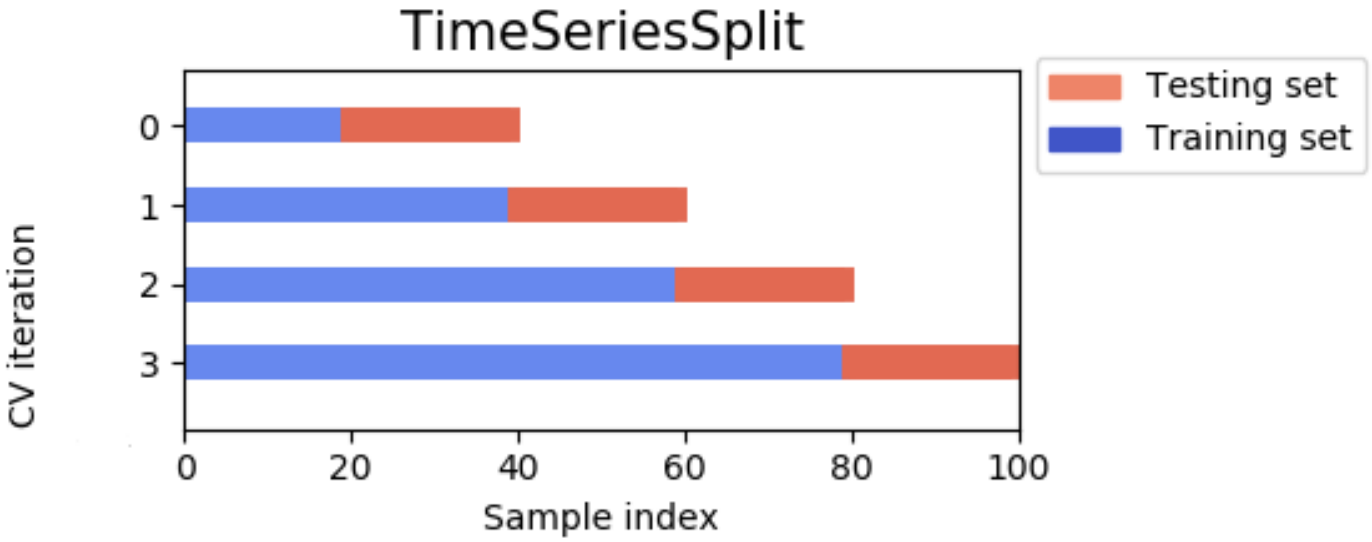

Fig. 5. 'TimeSeriesSplit' cross-validation scheme.[3]

Choosing the number of splits was also critical. Not every team plays a game every week during regular season. Thus, if we chose slices of length 32 (total number of teams in NFL), some slices would feature one or more teams twice (the slice spans across two weeks). On the other hand, if we chose much smaller slices (e.g., 15 entries), we would end up using the performance of some teams to predict the performance of other teams that played in the same week (K^{th} and $(K+1)^{th}$ folds are in the same week).

Therefore, the minimum number of teams that played during the regular season, i.e. 26 was chosen as the slice length. This length ensured that a team appeared only once in a given slice during regular season. Since $568/26 \approx 21.846$, the data frame was split into 22 slices during the cross-validation step. 'GridSearchCV' was also used to tune the hyperparameters of algorithms. Mean squared error of the actual scores and predicted scores was used as the scoring metric. Random Forest regressor was clearly the best performer based on the cross-validation results shown in the table below:

Regressor	Mean Squared Error	Tuned Hyperparameters
Ridge	34.74	*alpha*: 1
Elastic Net	36.53	*alpha*: 0.1, *l1_ratio*: 0.05
Random Forest	5.71	*n_estimators*: 500

Table 1. Regressor performance and hyperparameter values.

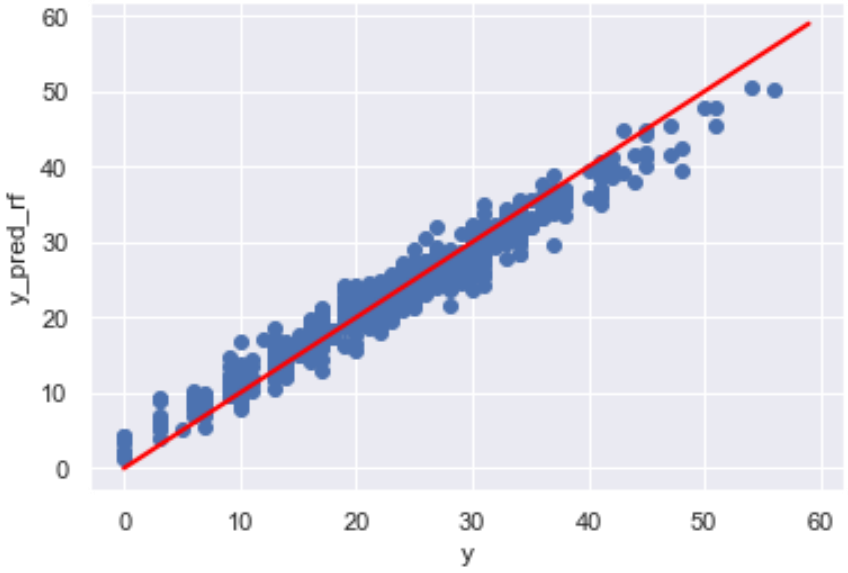

Fig. 5. Random forest regressor performance; *'y'*-axis represents the predicted values of *'y'*.

Predict the scores:

The scores predicted by each regressor are shown in the table below:

Regressor	Los Angeles Rams	Cincinnati Bengals
Ridge	25	23
Elastic Net	23	21
Random Forest	23	20
Actual score	***23***	***20***

Table 2. Super Bowl score prediction by regressors.

Random Forest regressor predicted the correct score of the game, while Elastic Net also predicted a score quite close to the actual one.

Conclusions and Improvements:

A linear regression model that accurately predicted the Super Bowl LVI score was developed. Two of the three regressors shortlisted for this study performed reasonably well. However, the model may be further improved by including the following features:

- Special teams' performance - kicker and punter statistics, punt return statistics, field goal kicker accuracy, etc. E.g., if a team's punter and special team performs well, their opponent will have to cover more yards to score the same number of points. Their starting field position for each drive would also determine the type of schemes they play.
- Penalties - penalties can have a huge impact on the game as well. E.g., penalties for pass interference foul committed by the opponent can award a team 25+ yards. Such penalties greatly improve the scoring chances of a team.
- Time of possession – time management is another important issue that needs to be considered. Having sufficient time to run certain plays during the last few minutes of a game can result in win or loss for a given team.

Client Recommendations:

Clients can use this model for several purposes:

- Predicting Super Bowl winner accurately can be very beneficial in betting.
- Correlation between various aspects of teams' performance can help identifying stronger and weaker spots of the teams.

References:

1. https://www.thelines.com/super-bowl-how-much-money-bet/
2. https://www.pro-football-reference.com/
3. https://datascience.stackexchange.com/questions/41378/how-to-apply-stacking-cross-validation-for-time-series-data

Acknowledgement:

I would like to thank my mentor, Ramkumar Hariharan, for his valuable guidance.