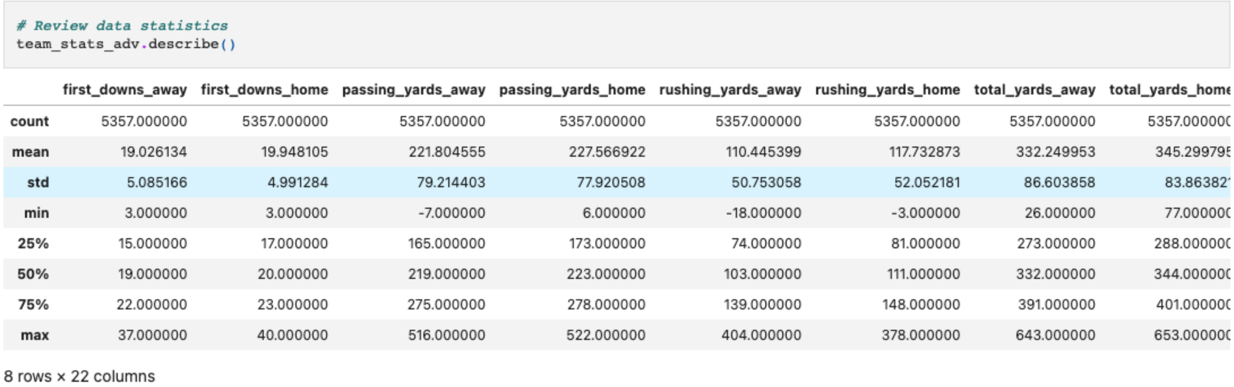
Super Bowl Prediction Final Report

Business aligned goal

The hypothesis I formed in the problem statement for this report was “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 2021 season?” To answer that question I performed a detailed data analysis while building a model for deployment.

Summarized Findings

Initially when exploring the data, we could see that the shape of the data was 5357 observations and 39 features. Immediately, we could see that score away and score home could be used as a target feature for the other features. It was intuitive to create a small data frame as a representation for a simple model that just included the team names, dates, and scores away and scores at home. However it was not a robust enough model for any true predictions. When creating an advanced model we could see what features directly contributed to the scores of each team and how important that feature was.



```
# Review data statistics
team_stats_adv.describe()
```

	first_downs_away	first_downs_home	passing_yards_away	passing_yards_home	rushing_yards_away	rushing_yards_home	total_yards_away	total_yards_home
count	5357.000000	5357.000000	5357.000000	5357.000000	5357.000000	5357.000000	5357.000000	5357.00000
mean	19.026134	19.948105	221.804555	227.566922	110.445399	117.732873	332.249953	345.29979
std	5.085166	4.991284	79.214403	77.920508	50.753058	52.052181	86.603858	83.86382
min	3.000000	3.000000	-7.000000	6.000000	-18.000000	-3.000000	26.000000	77.00000
25%	15.000000	17.000000	165.000000	173.000000	74.000000	81.000000	273.000000	288.00000
50%	19.000000	20.000000	219.000000	223.000000	103.000000	111.000000	332.000000	344.00000
75%	22.000000	23.000000	275.000000	278.000000	139.000000	148.000000	391.000000	401.00000
max	37.000000	40.000000	516.000000	522.000000	404.000000	378.000000	643.000000	653.00000

8 rows × 22 columns

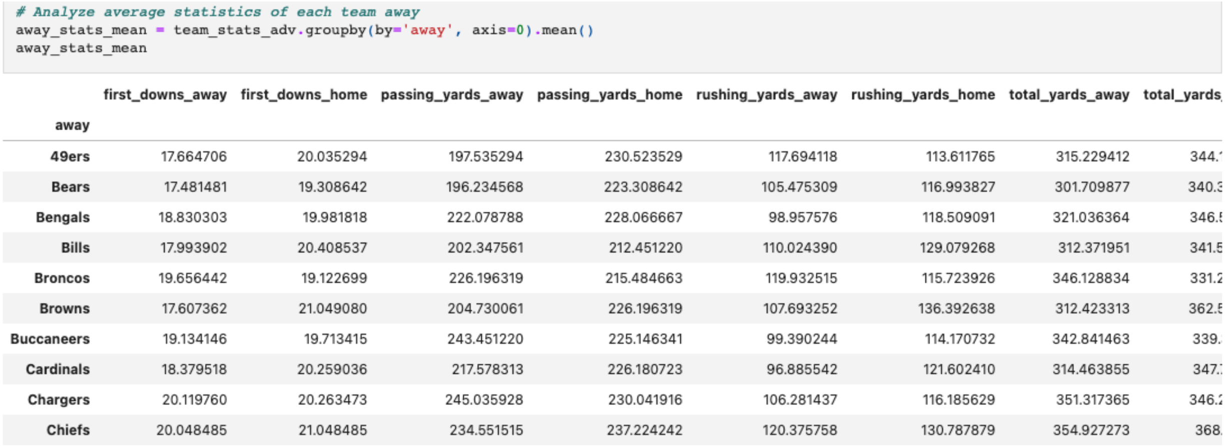
From the image above we can see the statistics for each feature were widely different indicating that standardization would be very important when using these features in a model. Some interesting observations after reviewing the statistics for the dataframe:

- Both the max score away and the max score at home were relatively close with a point difference of 3.
- Interceptions at home and away are exactly the same maximum.
- Some statistics, like the minimum, passing and rushing yards can be negative.
- On average, teams at home scored at least 2 more points then teams away, this makes sense.
- Home teams scored an average of 23 points per game at home, with lowest score of '0' and highest score of '62'.

Super Bowl Prediction Final Report

- Away teams scored an average of 21 points per game at home, with lowest score of '0' and highest score of '59'.

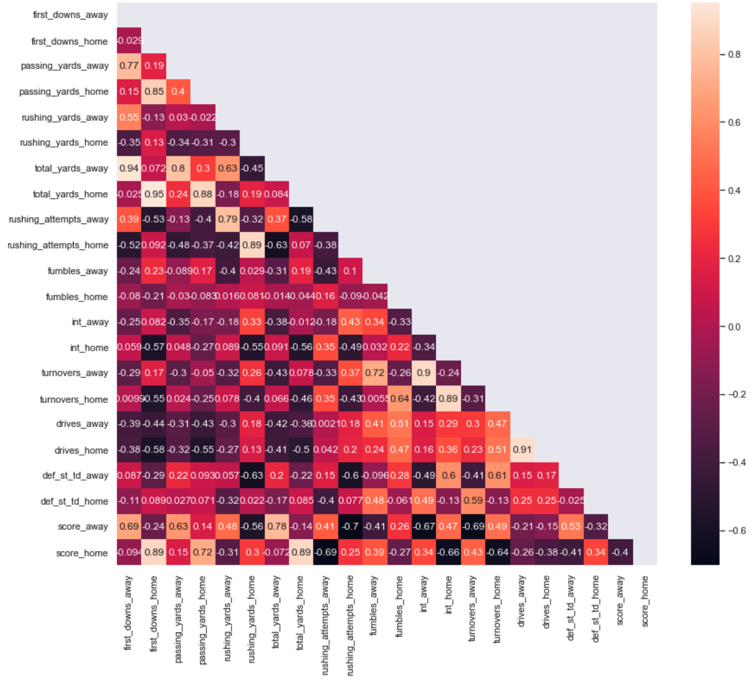
A glance at teamwise statistics at home and away show a clear trend among the teams that ended up getting the highest score.



```
# Analyze average statistics of each team away
away_stats_mean = team_stats_adv.groupby(by='away', axis=0).mean()
away_stats_mean
```

	first_downs_away	first_downs_home	passing_yards_away	passing_yards_home	rushing_yards_away	rushing_yards_home	total_yards_away	total_yards
away								
49ers	17.664706	20.035294	197.535294	230.523529	117.694118	113.611765	315.229412	344.
Bears	17.481481	19.308642	196.234568	223.308642	105.475309	116.993827	301.709877	340.3
Bengals	18.830303	19.981818	222.078788	228.066667	98.957576	118.509091	321.036364	346.5
Bills	17.993902	20.408537	202.347561	212.451220	110.024390	129.079268	312.371951	341.5
Broncos	19.656442	19.122699	226.196319	215.484663	119.932515	115.723926	346.128834	331.2
Browns	17.607362	21.049080	204.730061	226.196319	107.693252	136.392638	312.423313	362.5
Buccaneers	19.134146	19.713415	243.451220	225.146341	99.390244	114.170732	342.841463	339.
Cardinals	18.379518	20.259036	217.578313	226.180723	96.885542	121.602410	314.463855	347.
Chargers	20.119760	20.263473	245.035928	230.041916	106.281437	116.185629	351.317365	346.2
Chiefs	20.048485	21.048485	234.551515	237.224242	120.375758	130.787879	354.927273	368.

- Teams at home scored more points overall, approximately over 25%.
- Teams away scored less points overall, approximately less than 25%.



Super Bowl Prediction Final Report

When determining the home team’s score we can see there is significant correlation with the following features in decreasing order: A home team's score has significant correlation with the following features in decreasing order:

1. first_downs_home
2. total_yards_home
3. passing_yards_home
4. turnovers_away
5. fumbles_away

The defining assumption for the model is that the data includes a feature as date and was also a contributing feature to the dataset’s other features. This is due to the date or week of the football season being a difference that matters when it comes to the scores. Therefore a time-series analysis was performed. A Dickey-Fuller test was augmented in order to check for stationarity in particular.

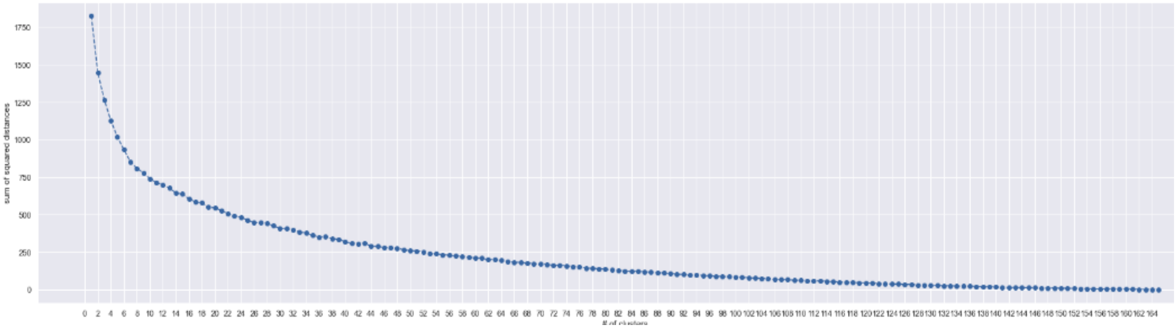
Model

For the Model, our target is to predict the winner of the Super Bowl game between Los Angeles Rams & Cincinnati Bengals based on the score. The prediction is derived from the listed features of a team. It is important to point out that values of features for both teams on that day is not included. Thus, the prediction model will involve three steps:

1. Estimate performance statistics of the two teams playing Super Bowl LVI.
2. Develop a regression model that accurately determines the team from previous games played in the season.

3. Input the estimated feature values in the model to predict the score for each team.

The first step was done using clustering analysis. We classified the Rams' opponents into several clusters. Then, we calculated the average performance of the Bengals in their last three games as 'test data'. This data was labeled using the cluster model. We then considered the teams that have the same label as 'test' data for further analysis.



The data was then split into features (all features) and the target variable (score at home) to prepare for standardization of a linear regression model.

Recommendations

Since the data frame contained 51 features and 5327 entries (after creating dummy variables), “Ridge” and “ElasticNet” regressors were used to punish large weights of the features along with Random Forest Regressor. We intend to predict the next week's data using previous weeks' entries.

Based on the aligned business goal previously established regarding the following question:

Super Bowl Prediction Final Report

> "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 2021 season?"

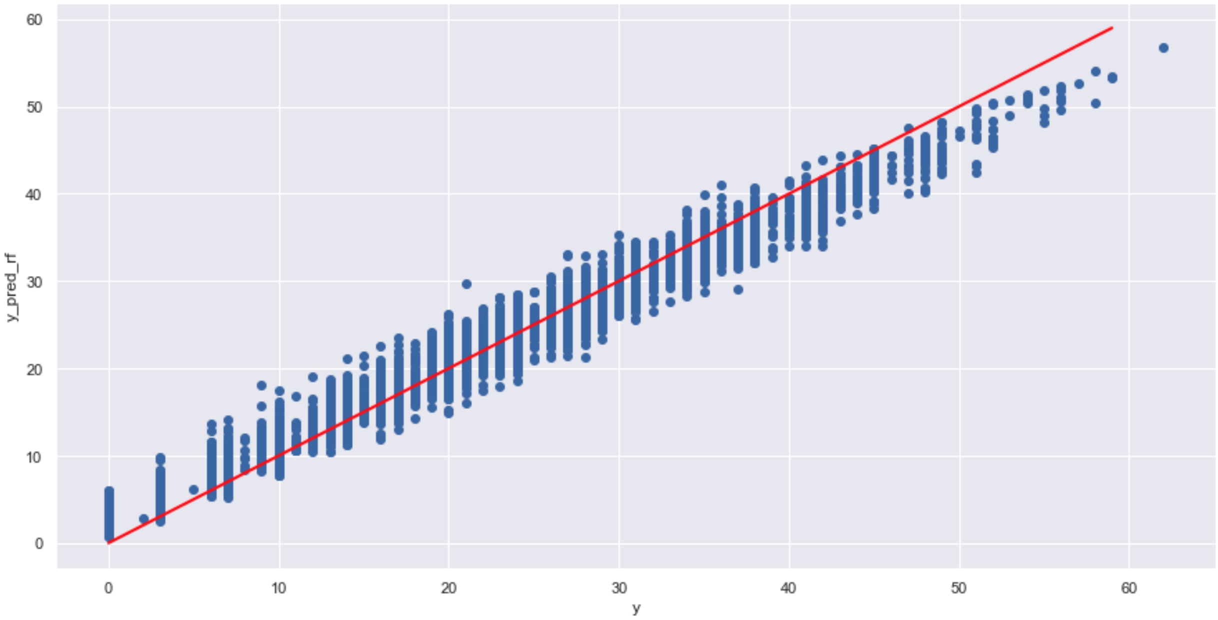
The answer is yes and a few options are recommended.

There can be many use cases for this type of prediction especially for an audience of millions. '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. Therefore, having the ability to predict the NFL champion team accurately and ahead of time can be financially 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. Using historical stats of teams for a business case, I 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.

Based on the Model Metrics File, it is recommended to use the Random Forest Regressor due to the lowest metric score (using Mean Squared Error). For all model algorithms new hyperparameters were tested and different modeling methods chosen.

Super Bowl Prediction Final Report

The Random Forest Regressor Model showed the best performance with an emphasis on high dimensional/complex data.



Potential deployment and maintenance

*Three potential scenarios that can be represented using our model are the following:

1. Focusing on what teams are more valuable than others and being able to plan for those higher ticket players or teams in the future.

2. Using the current stats for training purposes. Adjusting or changing that model to match factors or predictors is a critical tool that can be used in the future by any audience interested in understanding football statistics better.

3. Selling highlighted features to investors who are interested in football or wagering large amounts of money on games or seasons and basing a sales model around that concept.

For future improvements, I recommend running a linear regression model because it is simple but also leads to similar reliability as Random Forests models when it comes to this data.

Testing can also be used by attempting to use FIFO or First In First Out historical data always to observe how accurate model results are to actual results to avoid potential valuation loss.

It is highly critical the statistics are always updated and accurate. It is also important to use the most recent and updated data as the data is time-sensitive.