February 2, 2020

ANUJA dixit

student id: 94489489

Section-A

**Predicting Super Bowl 54**

FORECASTING PROJECT

**BANA 290: Super Bowl 54 Forecasting Project**

**Team Members: Anuja Dixit, Kathleen Sebastian, Supriya Shahane, Tatiksha Singh**

**Sunday, February 2, 2020**

**Anuja Dixit:**

I used 2 different datasets to build my models to predict the scores of Kansas City Chiefs and San Francisco 49ers in the upcoming Super Bowl Finals.

**1st Data Sheet:** NFL 2018-19 Regular Season Data.csv -- This is the consolidated datasheet consisting of 2018 and 2019 regular season datasets scrapped by Michael. I used R-Studio to build models to predict the scores and predict the winner.

*Multivariate Regression Model Scores:* Predicts Kansas City Chiefs to be the winners with a R-square of 65%. Match Scores would be:

Kansas City Chiefs: 28.26

San Francisco 49ers: 24.64

*Random Forest Classification Model:* Predicts Kansas City Chiefs to be the winners.

**2nd Data Sheet:** 2019 Season Data with Playoffs.csv -- Scrapped data provided by Micheal.

I used Alteryx to build 4 models to predict the final scores.

*Random Forest:* Predicts San Francisco 49ers to be the winners with R-square of 35%. Match Scores would be:

Kansas City Chiefs: 23.037

San Francisco 49ers: 27.75

*Neural Networks:* Predicts a tie between the 2 teams. Match Scores would be:

Kansas City Chiefs: 22.077

San Francisco 49ers: 22.077

*Support Vector Machine Model:* Predicts Kansas City Chiefs to be the winners with a very negligible difference in scores and a very low R-square of 65%.. Match Scores would be:

Kansas City Chiefs: 22.772995

San Francisco 49ers: 22.772987

*Linear Regression Model Scores:* Predicts Kansas City Chiefs to be the winners with a R-square of 72%. Match Scores would be:

Kansas City Chiefs: 32.64

San Francisco 49ers: 29.21

Hence from the 2nd data-set I will consider the Linear Regression Model scores and from the 1st dataset I will consider the Multivariate Regression Model Scores to predict my ensemble model scores. Both these models have a reasonably high R-square. I am using averaging technique to find the final scores of my ensemble model. So the final Scores for the finals will be:

**Kansas City Chiefs: 30.45 ~ 30**

**San Francisco 49ers: 26.93 ~ 27**

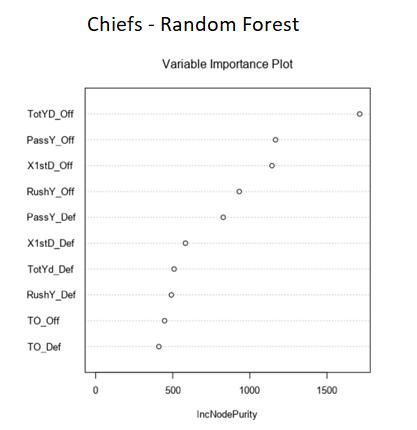
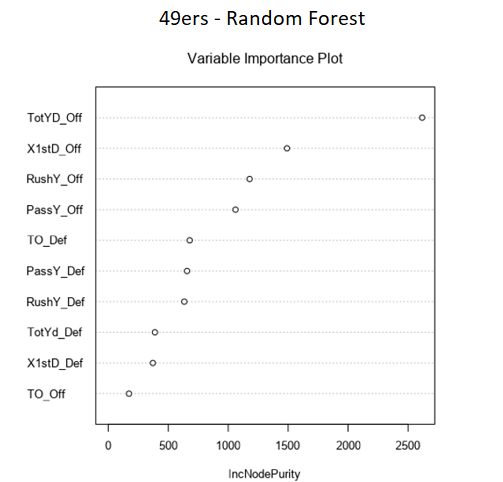
**Kansas City Chiefs will be the Super Bowl 54 Winners.**

**Tatiksha Singh:**

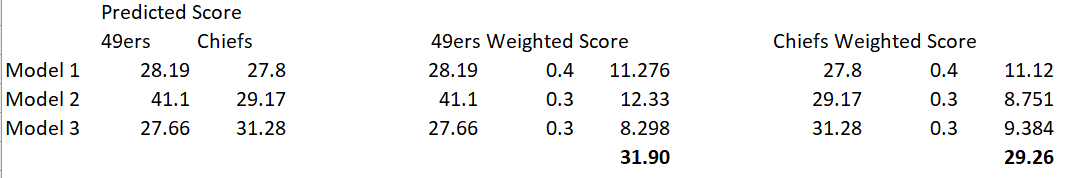
Model 1: Trained on 2017 & 2018 Data from Pro Football Reference's Website and Tested with 2019 Data. Used 10 Variables (Offense and Defense Stats) in the Multivariate Linear Regression Model. Used the average for all years as the basis for variables to help predict the score for the Superbowl. 49ers→ R2= 0.64 (Adj R2= 0.55); Chiefs→ R2=0.73 (Adj R2=0.62)

Model 2: Trained with 80% of 2019 Data. Tested with the remaining 20% of Data. Usd the same variables for Multivariate Linear Regression. Used the Median as the basis for all the variables to help predict the score for the Superbowl. 49ers→ R2= 0.83 (Adj R2= 0.013); Chiefs → R2=0.82 (Adj R2=-0.096)

Model 3: Trained with 2018 Data.Tested with 2019 Data. Used the same variables for Multivariate Linear Regression. Used the 2019 Median as the basis for all the variables to help predict the score for the Superbowl. 49ers→ R2=0.66 (Adj R2=0.004); Chiefs→ R2=0.67 (Adj R2=0.207)

I ran a random forest as well as a decision tree model to see the most important variables. (Total Yards, 1st Down, and Rushing Yards - All Offense).

But when I modelled the data using those variables I found the score did not change too much and the R2 value was also similar. I also tried using data from 2014 till 2019 to train and test my model, however since there was a lot of variability in the data, the model did not perform very well and I chose not to include it

in my overall ensemble model 

**Kansas City Chiefs: 29.26**

**San Francisco 49ers: 31.90**

**Prediction: San Francisco 49ers will be the Super Bowl 54 Winners**

**Kathleen Sebastian**

**Model 1: Linear Regression**

Utilized data from the Pro Football Reference website, extracting stats and scores for the 2019 season games for the Kansas City Chiefs and San Francisco 49ers, and creating two distinct datasets for each respective team. Then, a linear regression model was generated for each team’s dataset. To acquire input values to run the regression model and predict the scores for the upcoming Super Bowl, the average values for each predictor variable was used. The model was trained with 80% of the data, and tested with the remaining 20%. For the Kansas City Chiefs, the linear regression’s output indicated an R^2 value of 0.632 (R^2 adjusted of -0.03). On the contrary, for the 49ers’, it generated a slightly higher R^2 value of 0.672 (R^2 adjusted of 0.082). The linear regression models predict that the San Francisco 49ers are going to win.

**Kansas City Chiefs: 30.11**

**San Francisco 49ers: 31.26**

**Model 2: Random Forest**

Using the same datasets and initial methodology of implementing the average values for each predictor variable, a random forest model was generated. Similar to the linear regression model, the random forest model was trained with 80% of the data and tested with the remaining 20% of the data. The “First\_Down” predictor variable appears to be the most important feature in the model for the Kansas City Chiefs. However, we notice that the total yards predictor variable is the most important feature for the San Francisco 49ers. The results from the random forest model also illustrates that the San Francisco 49ers are also likely to win, and by a wider margin in comparison to the linear regression model.

**Kansas City Chiefs: 28.40**

**San Francisco 49ers: 31.77**

**Model 3: Neural Network**

When a neural network model was generated, the model also predicts that San Francisco 49ers are likely to win Super Bowl 54. In this particular model, the number of nodes in the hidden layer was reduced from 10 nodes to 8 nodes. As a result, we notice an even wider range in the score differences between the two competing teams.

**Kansas City Chiefs: 29.17**

**San Francisco 49ers: 43.96**

**Weighted Scores**

**San Francisco 49ers Score Prediction**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Model Score** | **Assigned Weight** | **Weighted Score** |
| **Model 1** | 31.26 | 0.10 | 3.13 |
| **Model 2** | 31.77 | 0.50 | 15.89 |
| **Model 3** | 43.96 | 0.40 | 17.58 |
| **Final Predicted Score for San Francisco 49ers** | | | 36.60 |

**Kansas City Chiefs Score Prediction**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Model Score** | **Assigned Weight** | **Weighted Score** |
| **Model 1** | **30.11** | 0.10 | 3.01 |
| **Model 2** | **28.40** | 0.50 | 14.2 |
| **Model 3** | **29.17** | 0.40 | 11.67 |
| **Final Predicted Score for Kansas City Chiefs** | | | 28.88 |

**Super Bowl 54 Winner:** San Francisco 49ers

**Supriya Shahane**

For the prediction, have used two datasets:

**Dataset 1 - 2019RegularSeasonData.xlsx** Used the 2019 Regular Season data provided as part of the assignment

**Dataset 2** - **NFLdata2014to2019.xlsx** Taken team statistics data for both Kansas Chiefs and SF 49ers from pro-football-reference.com from 2014 to 2019.

All the models are trained on 80% of the data and tested on the remaining 20%.

**Pre-modelling analysis**

Used Python to look at the descriptive and correlation/linear relation analysis for the datasets.

The notebooks captures observations based on the analysis.

**Models used -** Linear regression, Random Forest

**Dataset 1 (2019RegularSeasonData) - Modelling**

* Model 1 - Used the **2019** Regular season data and started with Linear Regression by using all the variables. The adjusted R2 value came out to be **60.7%**.
* Model 2 - Based on the results, ran another Linear Regression model with fewer number of variables. The Adjusted R2 value for the second linear regression model was also around ~ **60%**.
* Model 3 - I also ran a Random Forest model. **Team1\_Passing and Team1\_Rushing** were the important variables.

**Dataset 2 (NFLdata2014to2019) - modelling**

**Scenario 1** - Used the pro-football-reference team statistics data from **2014 to 2019**

* Model 4 - Started with Linear Regression by using all the variables. The adjusted R2 value came out to be **51.3%**
* Model 5 - Based on the results, ran another Linear Regression model with fewer number of variables. The Adjusted R2 value for the second linear regression reduced to **48.2%**

**Scenario 2** - At this point decided to use just 3 years of data that is from **2017 to 2019**

* Model 4 - Started with Linear Regression by using all the variables. The adjusted R2 value came out to be **50.4%**
* Model 5 - Based on the results, ran another Linear Regression model with fewer number of variables. The Adjusted R2 value for the second linear regression reduced to **45.1%**

Not much difference in both of the models.

* Model 6 - Also ran random forest model for both the scenarios. Off\_TotYards, Off\_RushY, Off\_PassY were the important variables.

**Predicted Scores**

To predict the scores, have taken the average of the variables f**or the year 2019** for both Chiefs and 49ers. Have not used Model 2, Model 5 for the datasets as the adjusted R2 was less. Weights are given below based on judgement.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Kansas Chief** | **SF 49ers** | **Weights** |
| Dataset 1 (Regular) | **Model 1 - LR** | 24.52 | 27.18 | 30% |
|  | **Model 3 - RF** | 24.16 | 25.28 | 30% |
| Dataset 2 - scenario 1 | Model 4 - LR | 26.47 | 22.25 | 0 |
|  | Model 6 - RF | 25.99 | 24.28 | 0 |
| Dataset 2 - scenario 2 | **Model 4 - LR** | 26.82 | 28.71 | 40% |
|  | Model 6 - RF | 28.54 | 27.1 | 0 |

**Ensemble model score**

Kansas Chief - 25.33 ~ 25

SF 49ers - 27.22 ~ 27

Prediction: SF 49ers will win the Super Bowl 54.

**Group Ensemble Model**

As a group we decided to equally weigh our scores when calculating our ensemble model since we all had similar R2 results. Even though there were some models that had higher accuracies, we chose to take each individual ensemble model score as our starting point to calculate the group ensemble score. Ultimately, we predict the San Francisco 49ers will win the Super Bowl with a score of **~31** (30.6635) and the Kansas City Chiefs will lose with a score of ~ **28** (28.4775) points.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Weight** | **49ers** | **Chiefs** |
| Anuja | 0.25 | 26.93 | 30.45 |
| Tatiksha | 0.25 | 31.904 | 29.25 |
| Kathleen | 0.25 | 36.6 | 28.88 |
| Supriya | 0.25 | 27.22 | 25.33 |
| **Group Score** | | **30.6635** | **28.4775** |

**END OF DOCUMENT**