Football Match Score Prediction and Classification using multiple regression models and KNN

Abstract

The goal of this project is to predict the total number of goals in football matches and whether the matches are interesting to watch. As a goal of this project, I attempted to predict number of goals in matches and classify whether the matches are interesting.

In attempting to solve this goal prediction problem, I used Linear Regression as well as Polynomial Regression on the training data set, with and without the use of Principle Components Analysis. On the other hand, regarding the classification problem, I applied K Nearest Neighbours as well as Logistic Regression. At the end, the best regression model to predict the total number of goals is Linear Regression. Polynomial Regression degree 2 model follows suit with slightly worse MSE measure on the test data even though it fits better compared to Linear Regression on the training data. The best classification model to predict whether the match is interesting is Logistic Regression slightly better than KNN.

At the moment the report was written, the predictions are not very accurate yet and leave rooms for improvement, due to the number of features are still quite limited. It remains that football matches are affected by countless factors as well as a high degree of randomness, which makes it impossible to achieve perfect prediction accuracy in my opinion. However, I believe that given more data sources, it is certainly possible to further increase the performance of the model.

Introduction

Football's recent trend of player recruitment using data analytics has proved the significance of using data in football industry. Previously, clubs recruit players by spending huge amount of money on global scouting networks. However, data analytics has evened the playing field for smaller clubs with limited budget against big clubs with much more significant spending. This is one in many roles data analytics has contributed to football. Furthermore, advanced technology over the past years has allowed us to collect more sophisticated data such as player's positions in real time through GPS technology and large amount of data per game. And the emergence of new Machine Learning techniques and tools made it possible to analyse those data with better performance.

Despite such advancements, match prediction still hasn't translated into an easy task, thanks to a lot of random and unpredictable factors that might affect the game result. However, this research aims to test different models on predicting outcomes of football matches, compare their performances and limitations, and attempt to improve prediction accuracy.

Data Analysis

Data Preprocessing

Data Type and Null values check

Data types of all the features in the data set are all integers, thus reducing the amount of preprocessing work that we need to do. Afterwards, we need to check whether we have null values in our dataset. There are no null values in the database as well.

Conclusion: The data is clean and can be used straight away for analysis purpose. This rarely happens in real life, where we need to spend 80% of the time to clean the data.

Data Exploration

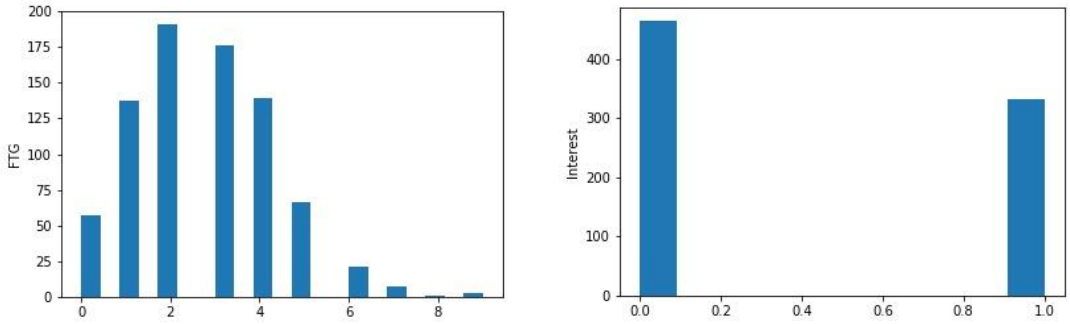
Data Review

First, we are going to quickly review the data. There are 13 features in "training_data_x.csv", HomeTeam and AwayTeam being the ids of two teams in one match. There are 798 rows in total.

There are 2 features in "training_data_y.csv", representing the labels we're trying to classify (Interest – whether the game is interesting) and regress (FTG – the total number of goals in a game).

Histogram

We're going to look at the distribution of the two labels, "Interest" and "FTG".



Histogram of FTG and Interest

As can be seen from the histograms, the FTG label is right-skewed, which will mess up our predictive model, affecting the regression intercept, coefficients associated with the model. (Vasudev R., 2017)

Correlation Matrix

Correlation matrix shows the correlation coefficients between variables. Each cell shows the relationship between variables. The correlation value ranges from -1 to +1, with -1 meaning the 2 variables having an inversely relation to each other (one proportionally increases as another decrease) to +1 being 2 variables having positive linear relationship to each other (one proportionally increase as another increase). We combined the training_data_x and training_data_y to plot the correlation matrix as below:

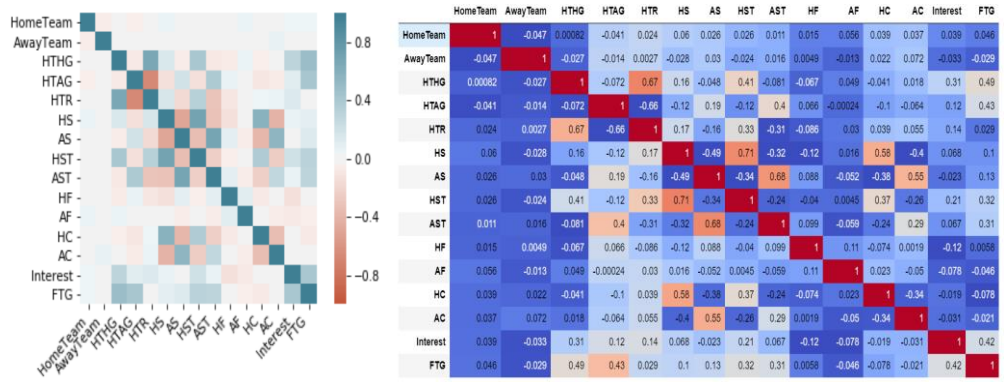


Figure 1 Correlation Matrix with Correlation Coefficients

After plotting the correlation matrix, we can see strong correlation between several pair of variables, whose squares are more blue

- **HST - HS**. Cor("HST","HS") = 0.71
- **AST – AS**. Cor("AST","AS") = 0.68
  Explanation: The team who have more shots is more likely to have more shot on target.
- **HS – HC**. Cor("HS","HC") = 0.58
- **AS – AC**. Cor("AS","AC") = 0.55
  Explanation: The team who have more shots is more likely to have more corner kicks.

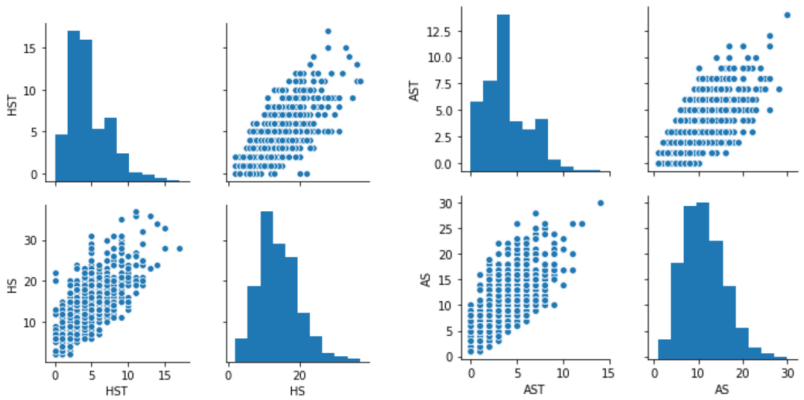
As well as inverse relation between:

- **AS – HS**. Cor("AS","HS") = -0.49
- **HS – AC.** Cor("HS","AC") = -0.4
- **HC – AS.** Cor("HC","AS") = -0.38
  Explanation: As one team has more shots, one can relate that this team is more dominant in the match, which results in the reduced amount of shots as well as corner that the opponent receives.

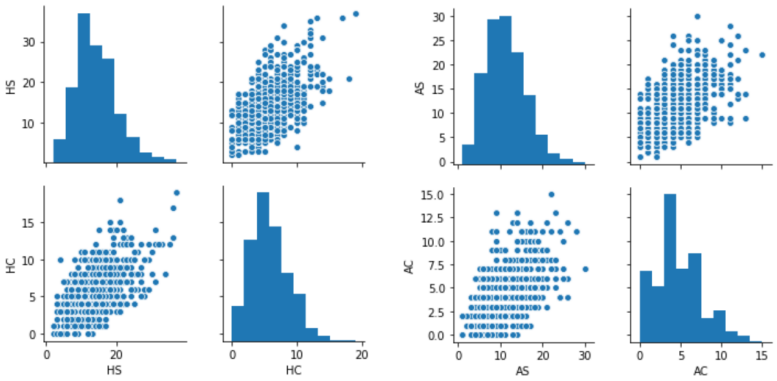
Pair Plot

We attempted to plot the pair plots for the above-mentioned pairs of variables to support our hypothesis regarding their relationships to each other.

- **HST – HS and AST – AS**

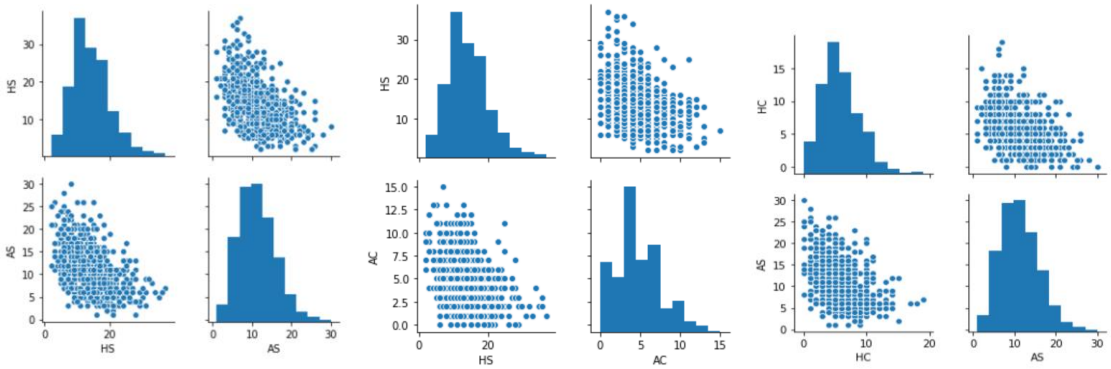


- **HS – HC and AS – AC**



As can be seen from the pair plot, the 4 pairs of variables have a positive linear relationship to each other.

- **AS – HS, AC – HS, AS - HC**



The inverse relationships between these 3 pairs of variables can be seen clearly from the pair plots as well.

Principle Component Analysis

According to Wikipedia, Principle Component Analysis (PCA) is "convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called **principal components**". This means "extracting" and combining existing features into several most representative features for the dataset. We have several pairs of correlated variables as mentioned in the previous chapter, which will come in handy when we do PCA on the data.

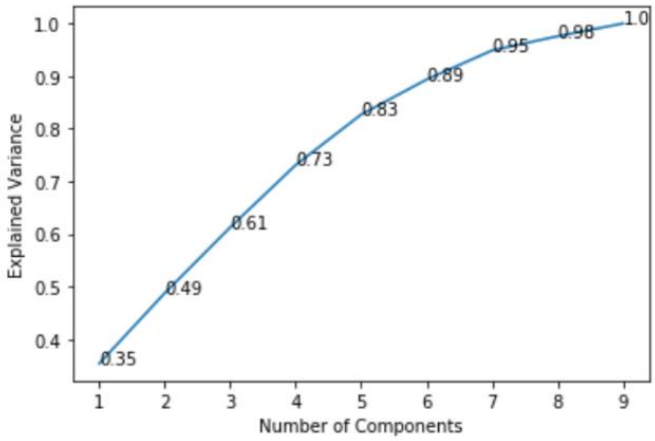


Figure 1 Cumulative Explained Variance

Figure 3 shows that PCA with 2 components give the cumulative explained variance of 49%. Which means that the 2 components merely explain half of the variance and taking 2 components results in losing a lot of information when we're doing this transformation. According to Hair, J. F. (2014, p.107)., the explained variance ratio should be at least 60%. This threshold depends a lot on the context and the data we're working with. But in this case, we're taking 60% as a floor value. Thus 2 components are not good enough.

We're going to investigate the correlations between variables and see how they affect the principle components. If we look at the first and second component's contribution by all variables, we have the following graph:

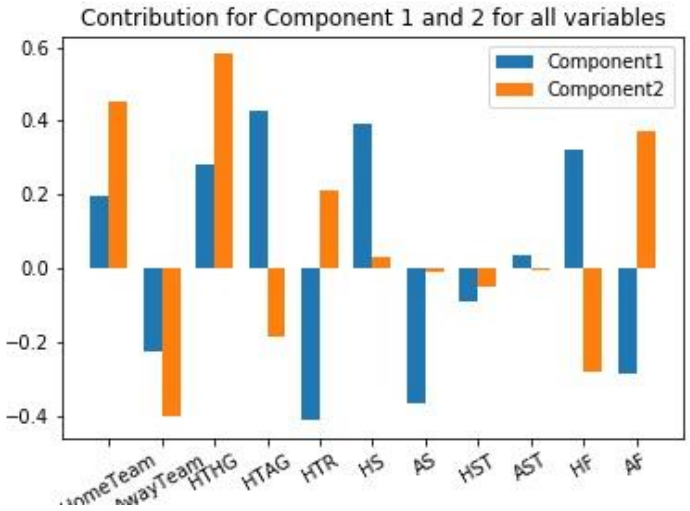


Figure 2 Contribution of Variables to Components

For the first component's contribution, we can see HS and HST contributions. Since they have strong correlation, the weight of HS in the first component is considerable and thus reduce the need to use HST, which explains the much smaller contribution to the first component. Similarly, when we look at AS and AST, AS's weight to the first component is also large, when AST's weight is much smaller in comparison.

Methods

Explain your whole approach (you can include a block diagram showing the steps in your process). What pre-processing was done (removing some observations, normalization, standardization, removing stop words, etc ...). say what methods/algorithms, you are using. explain them mathematically. Tell why you chose these methods.

Regression

We try to predict the value of FTG using linear regression model and polynomial regression with different sets of features, along with PCA. We're going to measure the accuracy of models using coefficient of determinations (R squared) and mean squared errors (MSE).

According to Stattrek, R squared is used to measure how much variability of the dependent variables can be explained by the independent variables. This ranges from 0 to 1. An R2 of 0 means that the dependent variable cannot be predicted from the independent variable. An R2 of 1 means the dependent variable can be predicted without error from the independent variable. Therefore, we aim to maximize R Squared for our models.

On the other hand, we use a second measure of accuracy, which is mean squared error – "average squared difference between the estimated values and the actual value" (Wikipedia). Therefore, we aim to minimize MSE for our models.

Classification

We aim to classify whether each game is interesting using K-nearest neighbours (KNN), Logistic Regression and Random Forest Classifiers. We measure the accuracy of our predictions using the confusion matrix of models. The accuracy ranges from 0 to 1, calculated from the confusion matrix by:

Accuracy = (True Positive + True Negative) / (True Positive + True Negative + False Positive + True Negative).

Regarding KNN, we determine the optimal number of neighbours based on the elbow method as well as Silhouette method. We also try different algorithms such as Logistic Regression to compare performance against K-nearest neighbours.

Experiments and Results

Regression

Linear Regression

This is the accuracy plot using Linear Regression models with and without the use of PCA.

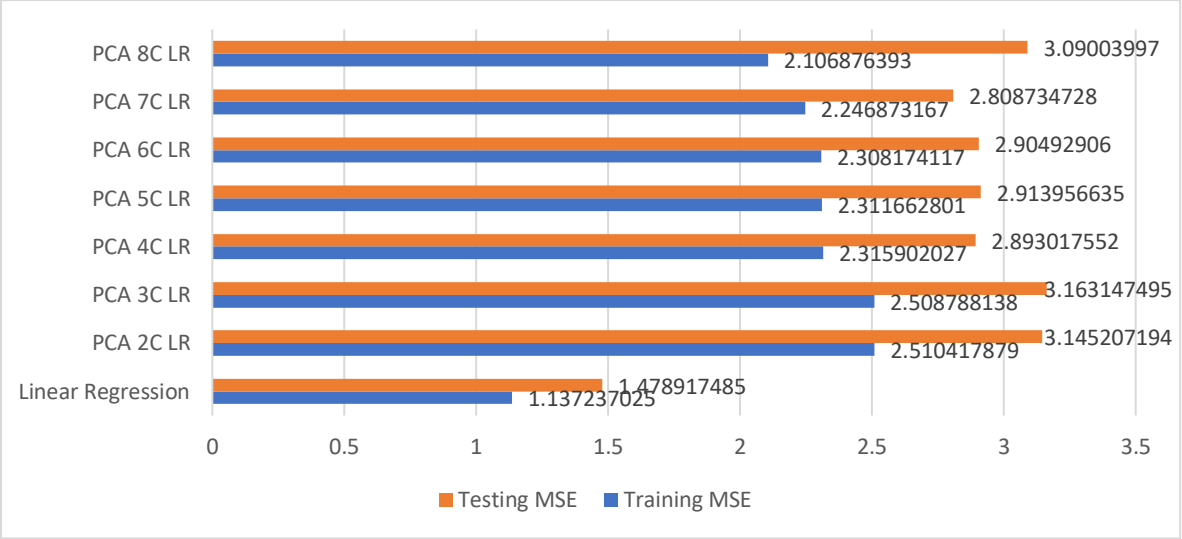


Figure 3 Linear Regression Performance

Notice that the MSE measure is more optimal in the case of Linear Regression without PCA. This can be explained by the fact that Principal Components do not necessarily have any correlation to classification accuracy. According to Stackoverflow, there could be a situation where the first PC takes up 99% of explained variance but that PC has no relation to the underlying classes in the data. Whereas the second PC (which only contributes significantly smaller part of the variance) is the one that can separate the classes. If you only keep the first PC, then you lose the feature that provides the ability to classify the data.

Polynomial Regression

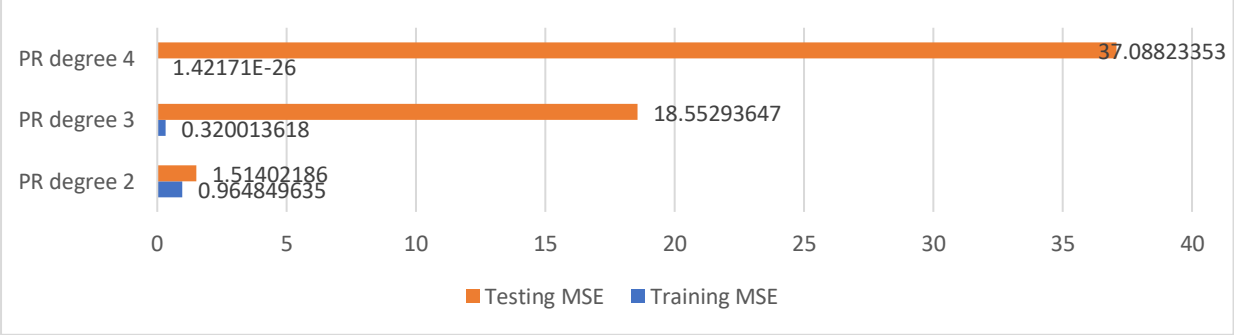


Figure 7: Polynomial Regression MSE Measurements

Polynomial Regression fits better with degree 2 compared to degree 3 or 4. This can be because polynomial regression models with degree more than 2 overfit on the training data and perform poorly on the testing data.

Classification

K-Nearest Neighbours

For KNN, we firstly determine the optimal number of neighbours using the elbow as well as Silhouette method. Using the elbow method, we tested on a range of k values and use the inertia attribute to identify the sum of squared distances of samples to the nearest cluster centre. As k increases, the sum of squared distance tends to 0. Below is a plot of sum of squared distances for k in the range specified above. If the plot looks like an arm, then the elbow on the arm is optimal k.

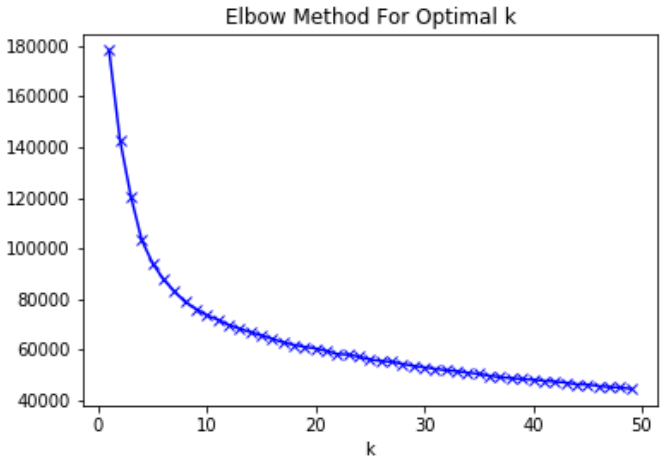


Figure 4 Elbow Method for Optimal k

Unfortunately, we do not have such clearly clustered data. In this case, the elbow may not be clear and sharp. Therefore, we also use the Silhouette Method. Silhouette method is an alternative to determine the number of neighbours for K Nearest Neighbours, which calculates the Silhouette score. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1] (Scikit-learn). A value of +1 indicates that the sample is far away from its neighbouring cluster and very close to the cluster its assigned. On the other hand, -1 indicates that the point is close to its neighbouring cluster than to the cluster its assigned. And, a value of 0 means it's at the boundary of the distance between the two cluster. Thus, the higher the score, the better. In our case, clustering by 2 and 4 nearest neighbours seem to be the most optimal.

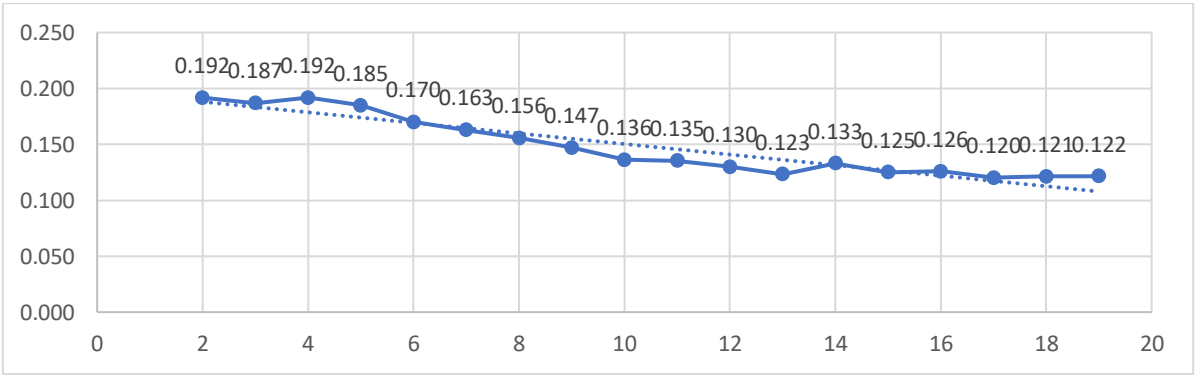


Figure 5 Silhouette

The accuracy result of KNN on the testing data can be seen below:

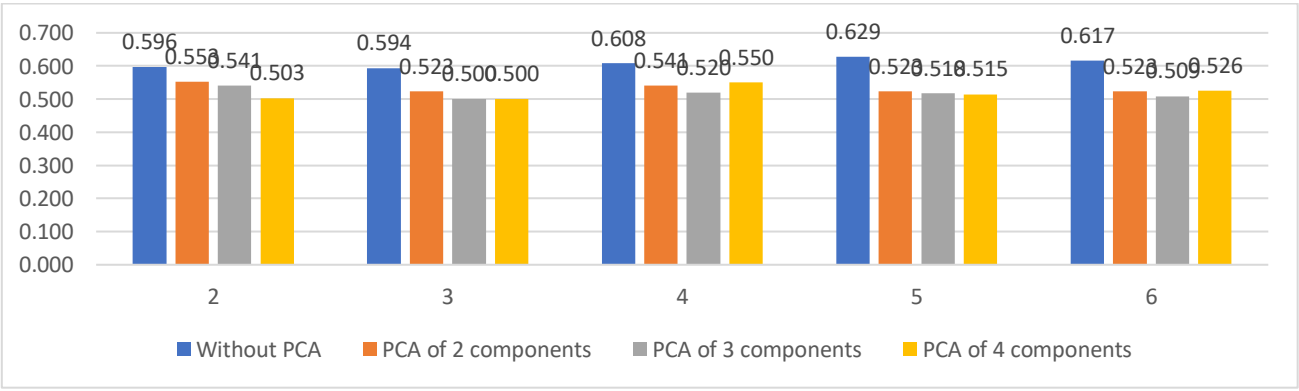


Figure 6 Accuracy of KNN from 2 to 6 neighbours

The maximum accuracy can be obtained as 62.9% using KNN with k = 5. It is not shown on the graph but using 14 nearest neighbours can up the rate to 64.6%. Using PCA doesn't provide any accuracy benefit, similarly to what was observed using Linear Regression.

Logistic Regression

Another algorithm for classification is Logistic Regression. We try to include the most meaningful variables into the model by fitting a Logit model and choosing variables with p-values larger than 0.05.

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
HomeTeam	0.0205	0.0108	1.9036	0.0570	-0.0006	0.0417
AwayTeam	-0.0047	0.0103	-0.4557	0.6486	-0.0248	0.0154
HTHG	0.6086	0.1875	3.2449	0.0012	0.2410	0.9761
HTAG	0.5323	0.1944	2.7384	0.0062	0.1513	0.9132
HTR	0.2413	0.2490	0.9691	0.3325	-0.2467	0.7293
HS	-0.0421	0.0219	-1.9216	0.0547	-0.0851	0.0008
AS	-0.0458	0.0253	-1.8094	0.0704	-0.0954	0.0038
HST	0.1733	0.0469	3.6963	0.0002	0.0814	0.2652
AST	0.1345	0.0507	2.6520	0.0080	0.0351	0.2340
HF	-0.0771	0.0217	-3.5548	0.0004	-0.1196	-0.0346
AF	-0.0496	0.0214	-2.3103	0.0209	-0.0916	-0.0075
HC	-0.0178	0.0325	-0.5468	0.5845	-0.0814	0.0459
AC	-0.0122	0.0350	-0.3488	0.7272	-0.0809	0.0564

Figure 7 Logit Model with all variables

We choose variables with p-values smaller than 0.05, thus obtaining a list of meaningful variables as "HTHG","HTAG","HST","HF" and "AF". The accuracy obtained from this streamlined logistic regression model is 67.8%, which is slightly improved from the 66.9% accuracy from the logistic regression model with all the original variables included.

Conclusion and discussion

In summary, the results of my project reflect a solid first effort in prediction match outcomes in football using machine learning techniques. The best regression model to predict the total number of goals is Linear Regression, with MSE 1.47 on the test data. Polynomial Regression degree 2 model follows suit with slightly worse MSE 1.51 measure on the test data even though it fits better compared to Linear Regression on the training data.

The best classification model to predict whether the match is interesting is Logistic Regression with 67.8% accuracy score, slightly better than KNN with 64.9% accuracy.

Overall, football matches are affected by countless factors as well as a high degree of randomness that no amount of data can account for, which makes it impossible to achieve perfect prediction accuracy. However, given more features and data, it is possible to further increase the performance of the model.

References

DATA PRE-PROCESSING. DHAIRYA KUMAR.2018. https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d

PCA. https://en.wikipedia.org/wiki/Principal_component_analysis

ACCEPTABLE PCA VARIANCE RATIO. HAIR ET ALL. 2012. MULTIVARIATE ANALYSIS.

SILHOUETTE ANALYSIS. https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

LOGISTIC REGRESSION. https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8