Big Data & Predictive Analytics

Coursework 1

CO3093

Author: Mohamed Mubaraak
Department of Informatics
Date of Submission: February 2018

Part 1

***Objective**: Using the given dataset, we would like to build up a model that can predict the winning team of the next premier league match between Manchester United and Manchester City by using simulation and a historical dataset.*

The initial unfiltered dataset **PremierLeague1718.csv** consists of 240 rows and 68 columns containing data relating to games played by teams in the Premier League during the season between 2017 and 2018. It also contains further data relating to team division, game referee and betting odds which are, although interesting pieces of information,not directly relevant in determining the in-game performance and predicting the better between two teams as it to be outlined in the first part of this report and so would need to be removed or filtered in order to proceed

Findings

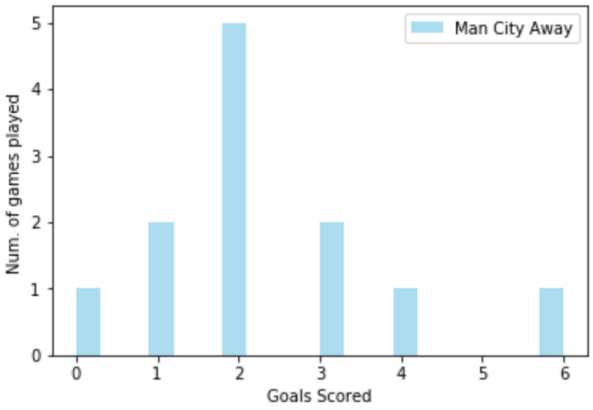
Upon filtering the data by removing columns not relating to the objective of predicting the winning team and dropping columns with null data, the next step is to use pandas' "describe" method to summarise the relevant data.

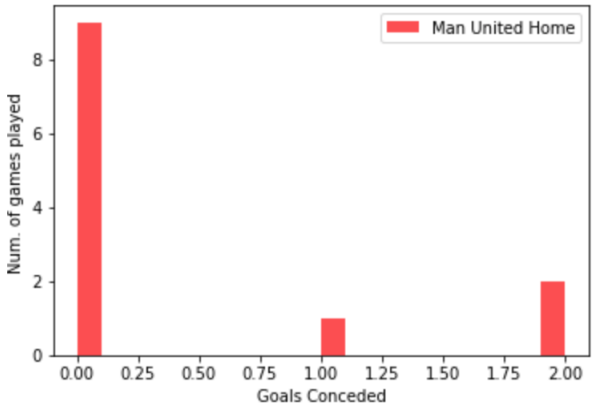
	FTHG	FTAG	HST	AST
count	240.000000	240.000000	240.000000	240.000000
mean	1.491667	1.183333	4.554167	3.791667
std	1.344419	1.250830	2.781661	2.385086
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	2.750000	2.000000
50%	1.000000	1.000000	4.000000	3.500000
75%	2.000000	2.000000	6.000000	5.000000
max	7.000000	6.000000	15.000000	14.000000

The descriptive analysis of the dataset is as follows:

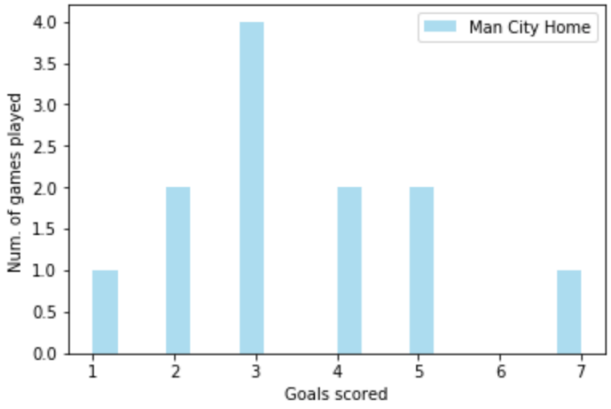
- The mean is larger than the median in every column with Full-Time Home Goals (FTHG) showing the greatest difference from the mean compared to the remaining columns and this means that the distribution favours larger values than the median.
- The standard deviation gathered for each column is significantly smaller than the range (*found by subtracting the max value from the min value*) which indicates that the distribution displays "long tails". This means that the occurrences within the distribution are more spread out to either side of the median and the number of outliers/anomalies is greater.
- In addition to the first two points, further proof indicating that the distribution is skewed towards larger values is evident in the majority of columns where the first quartile is closer to the median compared to the third quartile. Where the first quartile isn't closer to the median than the third quartile, the difference in the range between the first and third quartile to the median is equal. (e.g. in FTAG and AST)

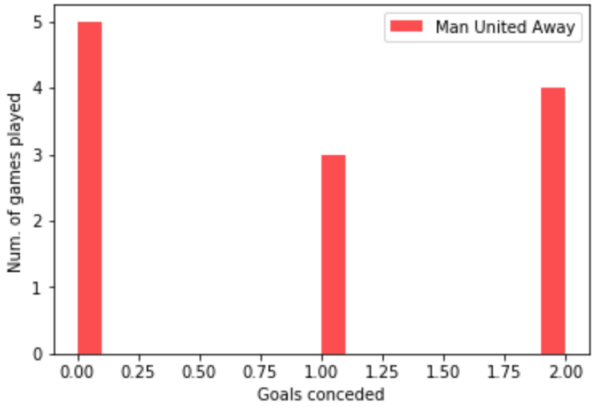
Comparing Manchester City's Away offence with Manchester United's Home defence.





Comparing Manchester City's Home defence against Manchester Uniteds away offence.





Simulation

In order to predict the winner between Manchester United and Manchester City, games comparing the away offensive and home defensive performance of each team were simulated as fantasy games as a Poisson Distribution.

But why use a Poisson Distribution?
The data can be displayed as a Poisson Distribution since we are counting the number of occurrences an event occurs. In this case the event is the number of times a team wins in the simulated games as a probability of all the games they play. Also a poisson distribution is useful for calculating the probability of an event occuring (i.e. Man United winning Home) for a single event is independent of other events occuring priot to it. For example, just because ManUnited won their 30^{th} game at home in a simulation out of 100 games, they are not going to win the 31^{st} game at home too.

What can be gathered from the data from the simulated games?
Out of 500 games simulated where Man City played Home and Man United Played Away, the following was derived:

```
Games Man City Won HOME:  343
Games Man United Won AWAY:  91
Draws:  66
```

Compared to the results of 500 simulations where Man City played Home and Man United played Away:

```
Games Man City Won AWAY:  216
Games Man United Won HOME:  201
Draws:  83
```

For comparison purposes, the probability of games won by each team is displayed in the following table:

Team	Probability
ManCity Home	68.6%
ManUtd Away	18.2%
Draws from ManCityHomevsManUtdAway	13.2%
ManCity Away	43.2%
ManUnited Home	40.2%
Draws from ManCityAwayvsManCityHome	16.6%

It is clear from the probabilities gathered from games simulated between both teams that ManCity plays significantly better when they play Home compared to when they play Away. This can be further backed by data displayed in the histogram presenting the offensive performance of ManCity in their home games during season 2017/18 which shows that they won the majority of the games that they played scoring as high as 7 goals in a game against Stoke in October 2017.

This is contrary to the offensive performance of ManCity simulated in games when they played Away vs. ManUnited playing Home which shows a distribution of probability of games won by either team as being more closer to 50% each. This is because based on the provided dataset, the offensive performance of ManCity Away and ManCity Home is deemed fairly equal with, in this particular set of simulation, resulting in ManCity being slightly better.

Part 2

The initial dataset for this section consists of data relating to the stock prices for Apple. It contains relevant information about the date, opening price and closing price, highest and lowest stock price for the day along with adjusted close and volume. It is key for this second part to note the definition of the adjusted close which is the **closing price of a stock which includes the changes that may have occurred prior to the following days' opening.** This information can be visualised as a time series below.

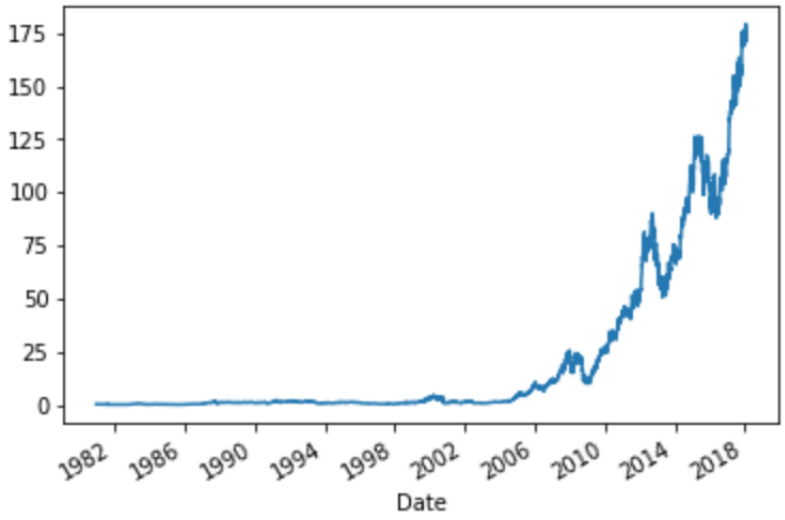


Figure 1: Plotting the change in adjusted close over time

From this Time Series the following can be gathered based on patterns in the trends in the data (particularly from 2006 onwards):

- A significant increase in stocks starting from 2006 can be credited to the release of the first iPhone on June 29th 2007
- From this day onwards, a gradual increase in the adjusted close (Y axis) can be seen until around 2012/2013 where it peaks at around a value of 80.
- We then see a drastic drop in the adjusted close from beginning of 2013 till the end of that year. The same pattern of a gradual rise in the beginning of the year followed by a rapid drop is repeated again in 2015,2016 and 2017.
- This odd 'seasonal'-esque pattern is most likely due to the unveil/release cycle Apple has where they release new products like the Mac, and iPhone towards the end of the year resulting in more people buying stocks and shares in Apple depending on how well the new Apple products are received by the general public.
- For example, with the increase of the cost of these new products, the general public may feel they are becoming over-priced and the enthusiasm for the product may die resulting in people pulling out their money invested into AAPL stocks, thus resulting in a lower adjusted close and vice versa in cases where a product may gain the interest of the general public. (Noticeable with the trend in 2017 onwards)

Predictive Model

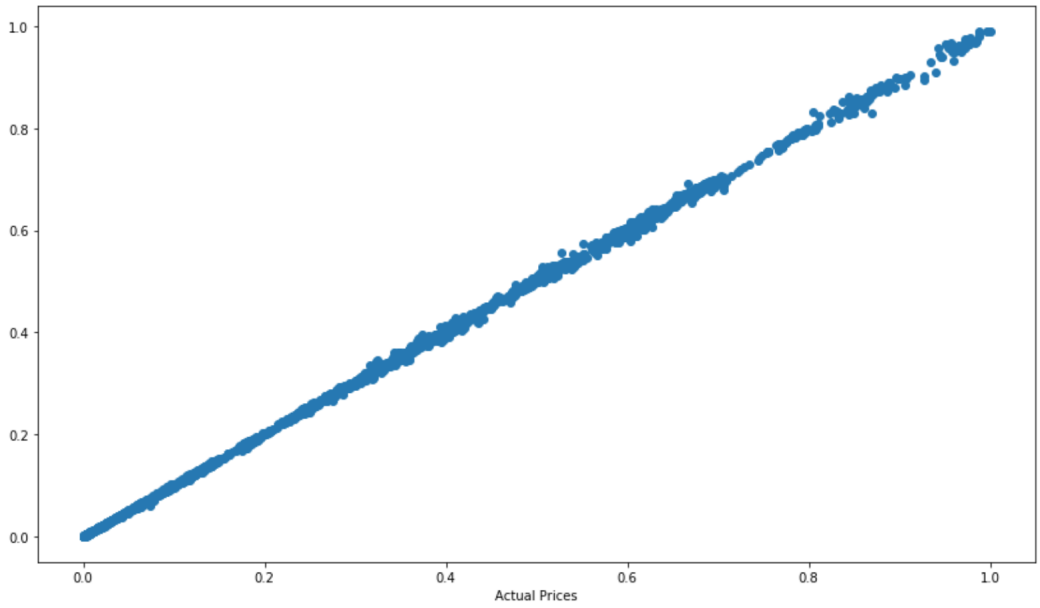
The relevant predictors or input features for making a prediction for the future adjusted close on this data set are the following:

- Open
- High
- Low
- Close
- Volume
- Moving Average

For this model the initial dataset was split into two sections, where 0.25 or 25% comprised the test set and the remaining 75% made up the training set. It is vital to have a large enough training set in order to ensure that predictions on unseen data can be as accurate as possible, thus strengthening the accuracy of the model. This results in a model with the following data for the model:

```
Y-axis intercept -7.210861394468804e-05
Weight coefficients:
                  Open: -0.4096018925338409
                  High: -0.5368370239621666
                   Low: 0.2310753146096059
                 Close: 0.862317174479047
                Volume: -0.0035423569065977517
        Moving Average: 0.8509342794825155
R squared for the training data is 0.9998145599496179
Score against test data: 0.999837025752675
```

The accuracy of the model can be determined in how close it is to the value of 1. In the case of this predictive model, the model is extremely accurate since 1-the R-squared value results in a very very small number, i.e. the likelihood of inaccuracy is exactly **0.00018544005.** This results in a Scatter Graph comparing the predicted Adjusted Price and the Actual Price as follows:



This model is appropriate to use because the points on the scatter graph are closely clustered together showing a positive correlation with just a handful of data points slightly above or below as anomalies.This means that for the majority of predictions, the predicted adjusted close is very close or equal to the actual adjusted close. The anomalies are a result of the R squared value for the model being close to, but not exactly 1, resulting in slight inaccuracies in the predictions of the adjusted close.

The mathematical equation for the fitted model is:

$$\text{Adj} = -7.211 + -0.410 * \text{Open} + -0.537 * \text{High} + 0.231 * \text{Low} + 0.862 * \text{Close} + -0.004 * \text{Vol} + 0.851 * \text{MA}$$

this allows us to compare the accuracy of the model further as it allows to see how far the accuracy of the adjusted close is away from the actual adjusted close. This predictive model is accurate but it is not accurate enough to be used in real life day to day use, to make money off, or make predictions for in the future. The reason for this is that although the R-squared value is 0.9998 (4 decimal places) there is still a 0.0002 chance for an anomaly to occur in the prediction and too much confidence in a high risk, high reward model like this can prove to be detrimental to finances if a case where an anomaly exists occurs.