Time Series Analysis and Forecasting of In-Play Odds on a Betting Exchange

Research Project

Andrew Bunyan
MSc in Data Analytics (FT)
School of Computing
National College of Ireland
Dublin

***Abstract*—One of the newest phenomenon's in the world of gambling, exchange betting has often been compared to the fluctuations and unpredictability of the financial markets. Within this unique technology is perhaps a more unambiguous form of gambling – in-play betting. This live betting scenario offers a rare insight into the peaks and troughs of betting odds throughout the duration of various sporting events. Horse racing is a sport unanimously associated with the punter and gambling and coupled with its in-play options, is an ideal area to analyse. In-play betting provides a plethora of potential analysis but perhaps the most intriguing of which is time series analysis. From start to finish, time stamped data of racing odds can be analysed to better understand the nature of the odds fluctuations and may even lay the foundations for potential race prediction and forecasting. This project will research horse racing data through time series analysis across a multitude of races whilst attempting to forecast a winning event (horse) based on the previous odds during the race, through the implementation of the R programming language and WEKA data mining interface.**

Keywords—Betting Exchange, In-Play, Live Betting, Time Series, Forecasting

I. INTRODUCTION

The focus of numerous research papers and theories, betting exchange has seen a consistent rise in its use over the last number of years. With its potential to provide analysts with a financial gain, the betting industry as a whole has become a focal point of exploration, particularly coinciding with the recent advancement in data mining techniques. Applying such techniques to sport in general can help in predicting event outcomes, none more so than horse racing, where race prediction has become a central point of interest for many. Betting exchanges in particular are often compared to the fluctuations of the stock market and it is here where we find the unambiguous topic of in-play exchange betting, the cornerstone of the research for this project. In-play betting is a unique 'live' insight into the odds market and this project aims to analyse this in-play dynamic through use of functions such as time series analysis and regression to better understand the trends and tendencies of this high octane environment. In order to do so, time stamped data must be sourced and cleaned for analysis through two well-known data mining tools and programming languages, R and WEKA, with the goal to be successfully implemented in the domain of time series analysis and trends.

II. LITERATURE REVIEW

Each article plays its own individual part in highlighting the significance of the key foundations for this project and may also provide evidence of potential future workings.

Let's first look at the close association between exchange betting and the stock markets. Haahr (2011) likens the behaviours of the betting exchange to the stock market and more specifically, the odds market. Focusing on stock market experience coupled with behaviour finance theory, the author highlights a number of significant research questions based on human behaviour including trading volume, extreme values, trading patterns and strategies, but perhaps the most relevant to this project is when the author investigates 'anchoring' and its ability to restrict the fluctuation of odds. Haahr (2011) concludes that anchoring does indeed occur during in-play betting but more specifically around odds of under 1.2. This is worth noting for the research in this project as we will look at various market fluctuations with a view to ultimately finding the winner of each the race, whilst now considering the threat of anchoring to the final results. Fundamental information such as extreme values and trading volume can be the catalyst for further research into areas such as profit attainment and betting strategies.

On a similar line, exchange betting and stock markets are known to unearth long range similarities in their respective

market returns as Hardiman et al (2010) compare the statistics of Betfair exchange market to that of the Iowa Political Stock Market (Majumder et al (2009)). In-play betting on the exchange has been likened to the financial exchange by other researchers such as Franck et al (2010) and it is now widely regarded that the dynamics of a betting market may provide an insight into the dynamics of its financial counterpart and vice versa, also initially highlighted some years ago by De Bondt (1993) in attempting to forecast prices and exchange rates.

In 2003, Klassen & Magnus (2003), proposed a method to forecast the winner of a tennis match. Principally based on in-play, point per point events, the authors implemented their research through a computer-based program, TENNISPROB. In conjunction with this probabilistic calculator were two basic assumptions, the probability that one of the two players wins the match, both before and during the match. These two theories have given the authors a foundation on which to work for in-play exchange betting, written in a paper by Easton & Uylangco (2010) who again took to forecasting outcomes in tennis matches using in-play betting markets. Here we see the implementation of a model developed by Klassen & Magnus (2003), providing a point-by-point correlation with that model. Although based on a much smaller data set, these statistics had their advantages, the fundamental of which is named below. Easton & Uylangco (2010) concluded that "The significance of service breaks and service being held is anticipated up to four points prior to the end of the game." Moreover the authors established that "there is no evidence of a biased reaction to a player winning a game on service", endorsing the thoughts of Klassen & Magnus (2003) with regard to there being a high level of efficiency within individual betting markets, ultimately demonstrating the theory that betting markets have a high correspondence to the outcome of tennis matches. An element worth mentioning in terms of the previous studies is its accuracy. Klassen & Magnus (2003) have based their research on a data set taken from the five 'show' courts at Wimbledon. As a result this will generally only include the top ranked players in the world. This causes an under-representation in the data set of matches involving weaker players, or more importantly a miss-representation of the overall patterns of play. While smaller in size, the data employed by Easton & Uylangco (2010) included a variety of matches and was not limited to the top players. Both relevant papers, their limitations provide for extensive future studies, particularly around the prediction of in-play market fluctuations, a key hypothesis to this paper.

In terms of time series analysis itself, there are a number of papers that concentrate on an array of different topics such as medication quality (Warner et al, 2002) and financial time series analysis (Tsay, 2005). However, there is a distinct lack of research in any kind of sport, let alone a sport which incorporates in-play betting. This presents an acceptable research gap for this project.

III. Research Background

Known as the stock market of sports, in-play betting is a gambling phenomenon greatly on the increase. In-play or live betting is essentially any bets placed after an event has started and up to its conclusion. The research for this project has been carried out based on raw data obtained from the Betfair exchange market – an online marketplace for punters to bet against themselves - and not the traditional bookmaker. This research will attempt to unearth trends and tendencies within this live data with a view to better understanding the "stock market-like" fluctuations of in-play betting and its nature. Offering data on a multitude of different sports, horse racing was chosen both out of personal interest and that during a standard UK race, over £1500 per second is matched on the exchange highlighting the sport's popularity to punters. Historical time-stamped data was sourced and a focused hypothesis established.

Given that in-play betting has two key characteristics – time series and matched odds, the focus of this research was firmly on both. Coupled with the related work above, it was decided the research question should entail time series analysis whilst the distinct popularity in horse racing prediction called for the project to necessitate this topic in some form likewise. Time series analysis is unique in that it has a natural materialistic order. As a result time series analysis is far more palatable from other more common analysis, in which there is a lack of natural order within observations. The general perception with a time series model is that they reflect the fact that observations close together in time will be more closely related than observations further separated. One of the most insightful characteristics of time series is that they are often governed by trends and tendencies and it is this characteristic that has focused the research for this project as it attempts to gain a better understanding of the data whilst potentially predicting future values.

IV. Formulating a Research Question

A. Introduction to the Data Set

Apart from its unique betting exchange environment, one of the main reasons for choosing Betfair for this project was its ability to offer a varied and unique source of data. Sister sites such as the *Betfair developer program* and *Betfair data* offer users the opportunity to integrate Betfair data and betting services into any kind of application, on various platforms. Betfair's historical data is offered in two platforms, with and without detailed time-stamped data. Whilst both provide the in-play data required for analysis, to fully understand the nature of in-play fluctuations, the data must be fully time-stamped and ordinal. This time-stamped historical data scheme was in conjunction with a third party vendor called Fracsoft. Subsequent contact was made and after a small fee was paid,

they were able to produce an excel spreadsheet of this time-stamped data for a user defined period of March 2014. As a result, the data obtained contained all matched bets post and during each horse racing meeting in the UK for the month of March 2014. The breakdown was as follows:

Variable	Description
Event ID	Betfair Reference
Name of Market	Individual Event Name (incl. Distance)
Market ID	Betfair Reference
Date	
Course	UK Race Track
Race Time	Time race is due off
Time Stamp	Time bet placed ($1/10^{th}$ Second)
Inplay Flag	Pre-race or In play bet
Market Status	Highlighted any suspended markets
Selection ID	Betfair Reference
Selection	Name of horse backed
Total Matched	Total money matched on selection
Last Price Matched	Final Odds selection is matched
Back Price 1	Odds of Selection at Time Stamp

Indeed to be calling this data set large is an understatement as in total there were over 150 race meetings incorporating well over 1000 races and 7500 horses. This meant a data set that in total contained well over 100,000,000 rows and 17.5GB of storage. Whilst providing a potentially colossal supply of analysis, the sheer size of the data can potentially cause problems within the research development.

The key project variables in the data set are:

- Time Stamp
- Selection
- Back Price 1

Time Stamp – A significant variable in that coupled with the odds, the time the bet is matched is critical to any in-play analysis. Stamped at $1/10^{th}$ of a second the time series values represent the minutes and seconds once the race has entered live (in-play) status, right to its conclusion and subsequent market suspension.

Selection – The name of the horse backed. Of no real statistical relevance other than to label the data.

Back Price 1 – The most significant variable for this research as it encompasses the live odds that the selection was matched at a given second. On Betfair's exchange market, odds range between 1-1000, where 1.0 is absolute certainty and 1000.0 (999/1 to traditionalists) is maximum.

B. Formulating Research Hypotheses

Given the nature of the data obtained coupled with the key parameters taken from it, the research question was based on time series analysis. Authors have often discussed the familiarity between exchange betting and the stock exchange and while there are a number of well cited papers written on time series analysis of the stock market (see literature review), there are little or none relating to time series analysis of betting exchange odds. There is a clear gap in the research of this topic and something with which this paper can begin to fill. Taking all into consideration the research question was finalised as *Time Series Analysis and Forecasting of In-Play Odds on a Betting Exchange*. In order to fully evaluate the research there was an emphasis put on three key hypothesis that would accumulate to an overall understanding of the research question and its development:

1. *Is there any relational trends/tendencies among variables?*
2. *Is there a common trend amongst 'winning' selections?*
3. *Can we forecast odds through Time Series Analysis?*

Each one will be analysed and evaluated throughout the project with a few to having a better understanding of the in-play fluctuations and their trends.

C. Introduction to R

R is a programming language and software environment for statistical computing and graphics and in recent years has become one of the most important tools in the application of statistics in numerous areas such as business and IT. An implementation of the S programming language, R provides a wide variety of statistical and graphical techniques, including linear and non-linear modelling, time-series analysis, classification, clustering, and machine learning amongst others. Its ability to perform the aforementioned functions with relatively low computational memory makes the R language the perfect package for this research.

With the focus of the research on time series analysis and statistical modelling there were a number of libraries in use across different programming scripts. These included the *TTR* package for technical trading rules and the *e1071* package for miscellaneous functions such as regression analysis and support vector machines. Standard functions also included *plot.ts* and *SMA* to name a few.

D. Introduction to WEKA

WEKA is an easy to use graphical user interface which provides pre-processing of data and algorithms for classification, clustering, association and regression. Developed and written in Java, WEKA is a free software license available under the GNU General Public License. It supports some of the most standard data mining techniques including the pre-processing of the data and with the more current versions (3.7) one also has access to WEKA's extension packages where a package manager allows for the seamless addition of a number of add-ons, such as time series analysis and forecasting.

V. PROJECT DEVELOPMENT

With our data set obtained and our hypotheses established and reasoned, the attention now focuses on developing the data into an analytical experiment. As with any analytical experiment or research the process of development can follow a unique strategy to optimise findings and this project is no different. Having determined a research question and formulated key hypotheses, the development process will now follow the highly popular analytical practice of data preparation, data mining & modelling followed by optimisation of results for interpretation.

A. Data Preparation

Data preparation and the transformation it entails is one of the most important steps in analytics. The challenge of pre-processing the data source allows for a multitude of analysis given your data is now accurate and it is one which can establish a foundation for project development and evaluation. This cleansing process must require some knowledge of the final objectives of the project. For example, we cannot delete columns (fields) when cleaning if we need them later in the analysis. Thus this modified ETL (Extract, Transform & Load) process must be carried out with the research question and project deliverables (hypotheses) in mind, a requirement which can be fragile and time consuming but is important groundwork for any data-based research.

As mentioned previously, the enormous size of this data set meant cleansing and transformation was never going to be an easy task. Excel can only handle a maximum of 1,048,576 rows and this was encompassed over 14 races. Attempts to create a dataframe in SQL or Open Refine proved futile and as a result our data consisted of just this one spreadsheet. Highlighted already were the key parameters to analysing this data, so the first objective was to remove any redundant columns of data. This resulted in the deletion of the following columns:

- EventID
- MarketID
- Date
- Course
- Race Time
- Market Status
- SelectionID
- Total Matched
- Last Price Matched
- Back Price2
- Back Price3
- Back Volume1
- Back Volume2
- Back Volume3

Whilst certain parameters such as the total volume matched and lay'd may prove useful in other research areas, they are not required for this particular research question. Likewise date course and time have no influence on the analysis of the odds. Once the key parameters are established you can then work on transforming the data to suit the analysis.

First and foremost the data contained all bets matched post-race and in-play until market suspension. Therefore our research into in-play betting necessitated filtering out all bets matched prior to the race commencement. As mentioned one of the key parameters was the time stamp column and this was also key to the data preparation. In its raw form, the values represented the time after the race had begun in the format *mm:ss*. For example, if the race was off at 13:10, a time stamp value would be 10:44.5 indicating the bet was matched at 13:10:44.5. In terms of implementing this data into R and WEKA, this format would not suffice. Therefore the first change was to increase the time stamp from $1/10^{th}$ of a second to a second flat. This involved rounding values and a number of duplicate values. These duplicate values were then removed to create the ideal time series from 1:N, where N is the duration of the race in seconds, at a frequency of 1 second.

Once the time stamp column was cleaned, the focus was to split the data into categories which can easily be analysed and compared. Once again, the research question and hypotheses must first be considered before making a decision. The overall idea is to analyse each race individually and then look at comparing races on a similarity basis. Therefore, the data was split by race, resulting in 14 individual categories for analysis. Also accumulated were the time stamped odds on all 14 winners of these races to compare the analysis of these 'winning' selections as per the hypothesis outlined earlier.

B. Data Mining

With the data cleaned, filtered and transformed the real analysis of the project could begin. With the project hypotheses taken into consideration, data mining techniques were researched and tested with a view to determining a best fit scenario for this data set. Initial approaches were very much focused on classifying

the data with a view to producing a support vector machine (SVM) model that could potentially predict the outcome of a race based on analysing odds from a training data source of races. After some unfruitful attempts at prediction along with alternative considerations and a review of the related work, the focus turned to time series analysis for unearthing trends and tendencies within the data, with a final approach on forecasting based on the initial analysis.

With the data filtered by race, each one was individually analysed from a time series perspective. Focusing on the first of the research hypotheses, there was no form of prediction required at this point, therefore the entire race was analysed – start to finish. This was primarily done through regression analysis.

A main stay of simple component analysis, regression analysis is the process of fitting an approximated continuous function to a set of independent data points. Considering the objective is to analyse different races as a whole, multiple regression was used as the author looks to gauge a better understanding of the race given the multitude of horses running in each race. Given its more generalised functions, a multiple linear regression model was implemented on this data. This was based on the model:

$$(y \sim x1 + x2 + x3 \ldots xN)$$

Where *y* acts as the time stamp series and *x1:xN* represent the odds of the horses in the races. This was implemented through the R programming language through fitting linear models (lm), a model which lays the foundations for various statistical analysis such as regression. The significance of the *x1+….xN* indicates all the odds of the all horses in a race together (duplicates removed previously). Following on from the model fit, the data was plotted as a multiple line graph and statistical time series analysis such as regression and ANOVA carried out. Other tests undertaken, but not necessarily useful were predicted values and the covariance matrix for model parameters.

C. Data Modelling

Described as "the analysis of data objects that are used in a business or other context and the identification of the relationships among these data objects"(TechTarget,2015), data modelling is the process of creating a data model for an information system. Data models will take different sources of data and provide a structure through specific formatting highlighted by the project goals. However with the project deliverables implemented here it is necessary to introduce and use the WEKA data mining application. As mentioned in the hypotheses, there will be an attempt at forecasting the winner of a race through its previous odds. This required the training and testing of data through various different WEKA-based models.

As time series was the predominant factor in the overall analysis, the best fit model for prediction was the time series forecast model in WEKA. This is an add-on package easily accessible through the *package manager* in the WEKA GUI chooser.

Once installed on the program, there were a number of considerations to test, most notably which function would be best suited to predicting the winner. The three considered and tested were:

- Linear Regression
- Gausiann Processes
- SMOreg (Support Vector Machine)

As forecasting is essentially attempting to predict the winner of the race, not all the race data could be used and indeed there was a requirement for a cut-off point in the race that allowed for efficient forecasting from the model. Again there were three time stamped cut-off options for consideration - 70%, 75% and 80% of the race duration. The final choice was to be 80% cut-off purely based on a stat sourced from data.betfair.com that states "within the last 20% of a race, there is 80% of the total volume matched."

So the idea was to cut the data at 80% of race duration, run it through the three forecasting models and analyse the results compared to the actual winner.

RESULTS

Whilst all 14 races were analysed, there is little point discussing all results here. Therefore the results of four races will be discussed. These four will be of different race distance as to provide a broader range of results with the possibility to compare across different time stamp lengths. Below is a table of race distance and duration:

	Race Distance (Furlongs)	Time Stamp Length (secs)
Race 1	6	75
Race 2	16	256
Race 3	21	321
Race 4	24	362

Results of the four races will be presented in the following manner: Residuals, Coefficients, Analysis of Variance (ANOVA), ANOVA Comparison of Winner vs. Rest of Field.

A. Race 1 (14:15 Lingfield)

Residuals:

Min	1Q	Median	3Q	Max
-18.4542	-7.8966	0.6988	7.0842	15.1044

Coefficients:

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	60.3412207	5.2681944	11.454	< 2e-16 ***
Absolute.Bearing	0.0231214	0.0170371	1.357	0.1794

```
Diamond.Vine        0.0056605  0.0098826   0.573   0.5688
Ghost.Wing          0.0157627  0.0067937   2.320   0.0235 *
Hinton.Admiral     -0.0025110  0.0086721  -0.290   0.7731
Johnny.Splash      -9.8323666  1.3118821  -7.495  2.3e-10 ***
Metropolitan.Chief -0.0069864  0.0113592  -0.615   0.5407
Senora.Lobo         0.0072885  0.0124172   0.587   0.5593
Waterloo.Dock      -0.0008995  0.0066802  -0.135   0.8933
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.256 on 65 degrees of freedom
Multiple R-squared:  0.8351, Adjusted R-squared:  0.8148
F-statistic: 41.13 on 8 and 65 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: Time.Stamp
                   Df  Sum Sq  Mean Sq  F value    Pr(>F)
Absolute.Bearing    1 12058.5 12058.5 140.7433 < 2.2e-16
Diamond.Vine        1  3667.1  3667.1  42.8010 1.108e-08
Ghost.Wing          1  7196.4  7196.4  83.9947 2.541e-13
Hinton.Admiral      1     0.3     0.3   0.0033    0.9544
Johnny.Splash       1  5235.6  5235.6  61.1084 6.166e-11
Metropolitan.Chief  1     5.7     5.7   0.0668    0.7968
Senora.Lobo         1    28.4    28.4   0.3317    0.5667
Waterloo.Dock       1     1.6     1.6   0.0181    0.8933
Residuals          65  5569.0    85.7

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Analysis of Variance Table(Winner vs. Rest of Field)

  Res.Df    RSS Df  Sum of Sq F Pr(>F)
1     72 7624.1
2     66 10381.7  6   -2757.6
```

Looking at the analysis of coefficients in this race, we begin with highlighting that of the 8 horses in the race, there is only one of high significance(***), which happens to be the winner of the race – Johnny Splash. Therefore we can interpret that it is highly likely there is a relationship between the winner (Johnny Splash) and the time stamped duration of the race. Likewise when looking at the p-values of each horse, the winning selection is again highly relevant with all other runners of a similar relevance, while a 0.82 R-squared value indicates a strong correlation between the odds over the time series.

From the analysis of variance (ANOVA) table we can see that for the winner, along with three other horses (Absolute Bearing, Diamond Vine & Ghost Wing) the null hypothesis is rejected (p>0.0001).

B. Race 2 (14:30 Kelso)

```
Residuals:
    Min      1Q  Median      3Q     Max
-55.989  -7.240   0.237   8.146  43.735

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     248.687810   3.658290  67.979  < 2e-16 ***
Ballycool         0.023846   0.008022   2.973 0.003234 **
Dartford.Warbler -0.017501   0.008318  -2.104 0.036349 *
Desgrey           0.148602   0.043198   3.440 0.000679 ***
Inoogoo           0.006902   0.016030   0.431 0.667122
Jet.Master      -33.472122   0.766504 -43.669  < 2e-16 ***
Makbullet        -0.144834   0.041682  -3.475 0.000600 ***
Smadynium         0.006277   0.009988   0.628 0.530252
Surprise.Vendor   0.021771   0.005677   3.835 0.000158 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.99 on 256 degrees of freedom
Multiple R-squared:  0.9629, Adjusted R-squared:  0.9618
F-statistic:   831 on 8 and 256 DF,  p-value: < 2.2e-16

Analysis of Variance Table
Response: Time.Stamp
               Df Sum Sq Mean Sq   F value    Pr(>F)
Ballycool       1 957580  957580 4262.9627   < 2.2e-16
Dartford.Warbler  1  37461   37461  166.7675 < 2.2e-16
Desgrey         1   1420    1420    6.3200 0.0125527
Inoogoo         1    265     265    1.1815 0.2780653
Jet.Master      1 485665  485665 2162.0884 < 2.2e-16
Makbullet       1   7283    7283   32.4210 3.395e-08
Smadynium       1    299     299    1.3297 0.2499254
Surprise.Vendor  1   3303    3303   14.7065 0.0001582
Residuals     256  57505     225

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Analysis of Variance Table(Winner vs. Rest of Field)
  Res.Df    RSS  Df Sum of Sq  F Pr(>F)
1    263  93149
2    257 485857  6   -392707
```

Again looking at the coefficients initially, we see there are far more horses of significance with only three selections showing little or no significance (Dartford Warbler, Inoogoo & Smadynium). Interestingly again we see the race winner (Jet Master) with a low p-value indicating a high level of relevance with an R-squared value approximately 96% correlated.

The ANOVA table highlights that four selections reject the null hypothesis, including the winner.

C. Race 3 (13:30 Newbury)

```
Residuals:
    Min      1Q  Median      3Q     Max
-93.900 -23.508   2.111  24.705  82.370

Coefficients:
                 Estimate        Std. Error      t value    Pr(>|t|)
```

```
(Intercept)        129.241812  12.703467  10.174     < 2e-16 ***
Bangkok.Pete         0.007610   0.034081   0.223      0.82347
Castletown           0.110251   0.008009   13.765     < 2e-16 ***
Kings.Bayonet        0.012756   0.025055   0.509      0.61103
Minella.Definitely   0.005710   0.022159   0.258      0.79683
Pilgreen             0.009983   0.020469   0.488  0       .62608
Premier.Portrait     0.001566   0.035406   0.044       0.96475
Red.Devil.Lads       0.002882   0.028994   0.099       0.92089
Torero              -0.037243   0.018290   -2.036      0.04257 *
Westaway            -9.044703   2.816718   -3.211      0.00146 **
What.an.Oscar        0.083426   0.008602   9.698     < 2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.88 on 309 degrees of freedom
Multiple R-squared:  0.8015, Adjusted R-squared:  0.7951
F-statistic: 124.8 on 10 and 309 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: Time.Stamp
                  Df  Sum Sq Mean Sq  F value    Pr(>F)
Bangkok.Pete       1 1022791 1022791 583.1674 < 2.2e-16
Castletown         1  918233  918233 523.5511 < 2.2e-16
Kings.Bayonet      1    3583    3583   2.0428    0.1539
Minella.Definitely 1    1544    1544   0.8805    0.3488
Pilgreen           1     312     312   0.1779    0.6735
Premier.Portrait   1     228     228   0.1302    0.7185
Red.Devil.Lads     1       5       5   0.0030    0.9565
Torero             1    4624    4624   2.6362    0.1055
Westaway           1   72429   72429  41.2970 4.945e-10
What.an.Oscar      1  164950  164950  94.0499 < 2.2e-16
Residuals        309  541941    1754
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Analysis of Variance Table (Winner vs. Rest of Field)
     Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    318 1973561
2    310  560025  8   1413536 97.807 < 2.2e-16

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This is the first race we see that the race winner (Westaway) is not one of the more significant selections in the race, although its two stars indicate there is some form of significance there. There are three horses with high relevance (p>0.0001), the winner however is not one of them. Again there is a high R-squared value with approximate correlation of 80%.

From the ANOVA table we see that four horses reject the null hypothesis (including the winner)

D. Race 4 (13:45 Doncaster)

```
Residuals:
    Min      1Q   Median      3Q     Max
-224.200  -71.210    5.834   77.788  167.287

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    144.39090   32.61691   4.427 1.27e-05 ***
Cowards.Close  -0.21990    0.04734  -4.645 4.77e-06 ***
My.Dads.Horse   0.42939    0.04618   9.299  < 2e-16 ***
Victor.Hewgo    0.77781   15.78035   0.049    0.961

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 88.06 on 362 degrees of freedom
Multiple R-squared:  0.3129, Adjusted R-squared:  0.3072
F-statistic: 54.94 on 3 and 362 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: Time.Stamp
                Df    Sum Sq    Mean Sq    F value    Pr(>F)
Cowards.Close   1     462381  462381  59.6219 1.136e-13 ***
My.Dads.Horse   1     815835  815835 105.1981 < 2.2e-16 ***
Victor.Hewgo    1      19      19   0.0024    0.9607
Residuals       362 2807392    7755

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Analysis of Variance Table (Winner vs. Rest of Field)
 Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    364 3641457
2    363 2807411  1    834046 107.84 < 2.2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The fourth and final race highlighted only has three runners. Interestingly enough, two are highly significant and it’s the two losers. Perhaps as a result of the low numbers in the race it is harder to determine the significant/relevant horses. This is backed up by a somewhat low correlation value of R-squared approximately equal to 30%.

Similarly in the Analysis of Variance, the two losing selections reject the null hypothesis whereas the winner returns a p-value greater than 0.0001.

As per the second research hypothesis, the same time series regression analysis was carried out on the 14 winners of each race. The results were as follows:

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.2487 -1.0686 -0.2658  1.3769  3.6747

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -129.7351    55.2442  -2.348 0.023637 *
Billy.Twyford        6.4544     3.3964   1.900 0.064267 .
Catalinas.Diamonds   0.2485     0.7684   0.323 0.748012
Clever.Cookie       17.2754    16.3512   1.057 0.296769
```

```
Desoto.County       -79.9293    24.6030  -3.249 0.002284 **
Dishy.Guru          -0.7628     0.9408  -0.811 0.422034
Doeslessthanme      -2.0270     1.1168  -1.815 0.076679 .
Jet.Master           1.9422     2.3686   0.820 0.416844
Johnny.Splash       -0.5873     0.7703  -0.762 0.450053
Jumps.Road           8.3414     3.4198   2.439 0.019030 *
Mayfair.Music       -7.7165     3.2710  -2.359 0.023050 *
Summery.Justice      2.6337     0.9173   2.871 0.006383 **
Trust.the.Wind       3.9578     1.9046   2.078 0.043859 *
Victor.Hewgo        33.8256     9.1629   3.692 0.000636 ***
Westaway            15.3911     4.4951   3.424 0.001389 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.119 on 42 degrees of freedom
Multiple R-squared:  0.9878, Adjusted R-squared:  0.9837
F-statistic: 242.4 on 14 and 42 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: Time.Stamp
                  Df  Sum Sq Mean Sq  F value    Pr(>F)
Billy.Twyford      1 13170.6 13170.6 2932.4207 < 2.2e-16 ***
Catalinas.Diamonds 1    30.6    30.6    6.8190  0.012454 *
Clever.Cookie      1   264.1   264.1   58.8039 1.634e-09 ***
Desoto.County      1   513.8   513.8  114.4042 1.449e-13 ***
Dishy.Guru         1    12.1    12.1    2.6997  0.107832
Doeslessthanme     1    84.7    84.7   18.8492 8.735e-05 ***
Jet.Master         1    43.7    43.7    9.7304  0.003270 **
Johnny.Splash      1   285.5   285.5   63.5768 6.093e-10 ***
Jumps.Road         1   472.9   472.9  105.3008 5.160e-13 ***
Mayfair.Music      1    52.3    52.3   11.6506  0.001433 **
Summery.Justice    1   117.0   117.0   26.0544 7.581e-06 ***
Trust.the.Wind     1    39.0    39.0    8.6755  0.005240 **
Victor.Hewgo       1   100.2   100.2   22.3108 2.600e-05 ***
Westaway           1    52.7    52.7   11.7234  0.001389 **
Residuals         42   188.6     4.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Of the 14 winners, only one was highly significant (Victor Hewgo). The most interesting analysis here is that when compared to the losers in that race, the selection was classed as insignificant. As they are all winning selections you would expect a high correlation between there odds and that is identified with the R-squared value approximately 98%.

Analysing all 14 winning horses from the ANOVA table returned some interesting concepts. 8 of the 14 rejected the null hypothesis and were considered highly significant.

Finally in relation to the third and final hypothesis, there was an attempt to predict the race outcome through time series forecasting in WEKA. As mentioned there were three predictive functions to test in an attempt to find the winner of the race at an 80% cut-off point during the race. Each race was tested with all three functions and again for the purpose of this paper, the results of the four previous races will be only be discussed. The results were based on an attempt to predict the odds over the final 20% of the race, with the winner determined by the first horse whose odds drop below 1.0 (absolute certainty). These results are tabled below:

Race 1:

Model	Predicted Winner	Actual Winner
Linear Regression	Metropolitan Chief	Johnny Splash
Gaussian Processes	Metropolitan Chief	Johnny Splash
SMOreg	Metropolitan Chief	Johnny Splash

Race 2:

Model	Predicted Winner	Actual Winner
Linear Regression	Desgrey	Jet Master
Gaussian Processes	Jet Master	Jet Master
SMOreg	Jet Master	Jet Master

Race 3:

Model	Predicted Winner	Actual Winner
Linear Regression	Kings Bayonet	Westaway
Gaussian Processes	Kings Bayonet	Westaway
SMOreg	Kings Bayonet	Westaway

Race 4:

Model	Predicted Winner	Actual Winner
Linear Regression	Cowards Close	Victor Hewgo
Gaussian Processes	Cowards Close	Victor Hewgo
SMOreg	Cowards Close	Victor Hewgo

From the tables above it is clear that the only two of the twelve predictions were correct, a mere 17% accuracy. Perhaps the most interesting result was that in three of the four races all three models produced the same predicted winner. Although these were not the actual winners, it does say a lot about the fluctuation of odds.

PROJECT CONCLUSION

With our process complete, results declared, the project must finally refer back to its hypotheses and research question. The three research hypotheses were established to highlight the aims and the scope of the project. Having determined these potential

opportunities, the data set was implemented into various technologies designed to produce an efficient outcome. These results as portrayed above, have been represented primarily through statistical analysis and modelling, and can now try to associate them to our hypotheses.

1. *Is there any relational trends/tendencies among variables?*

Perhaps the most interesting of all conclusions was the significance the number of runners in each race had on evaluating the goodness of fit of the model. The more runners in the race the more accurate the correlation between the odds and time series is, therefore concluding that in terms of analysing in-play odds, more horses is good, while few is bad.

2. *Is there a common trend amongst 'winning' selections?*

While it's not true 100% of the time, the majority of races see the winner have a high level of significance, highlighting that it's unlikely no relationship between the winning horses and the time stamped odds exists. This can be classed as confirmation that as the race nears completion, the odds of the winning horse converge towards absolute certainty (1.0). Whilst proving the accuracy of the odds, this is based on analysis once the race has been run and the result is known. The initial conclusion in the first hypothesis can also apply here as prior knowledge of this trend may certainly aid future analysis.

3. *Can we forecast odds through Time Series Analysis?*

The last and final hypothesis to the research question is based solely on the attempt to predict the winning selection (horse) from a timeframe 80% into the race. As shown through the results, there was a less than 20% overall accuracy with an even less score for any one of the three functions used. Therefore in relation to our hypothesis, we can say that forecasting the race result through previously matched odds is inadequate and unfulfilling.

Overall this research project was a struggle in terms of first sourcing the correct data and then trying to perform sufficient analysis in a short timeframe. However, this initial work has highlighted some interesting areas that could potentially see research developed in the future. For example, Time series analysis is much more than regression analysis and various models such as the ARIMA and autocorrelations offer a different perspective to both the analysis and the potential forecasting. Likewise, other data mining techniques may be applied to the data set in hope of more accurate predictions. Support Vector Machines is one area that would certainly provide for interesting testing.

The data itself can provide for further research. As mentioned only 14 races were analysed out of a potential 1000 or so. Future analysis may well need the involvement of a data frame in an application such as SQL Management or Oracle.

REFERENCES

De Bondt, W. P. M. (1993). Betting on trends: Intuitive forecasts of financial risk and return. *International Journal of Forecasting*. doi:10.1016/0169-2070(93)90030-Q

Easton, S., & Uylangco, K. (2007). An Examination of In-Play Sports Betting Using One-Day Cricket Matches. *The Journal of Prediction Markets*, *1*, 93–109. doi:10.2139/ssrn.948013

Easton, S., & Uylangco, K. (2010). Forecasting outcomes in tennis matches using within-match betting markets. *International Journal of Forecasting*, *26*, 564–575. doi:10.1016/j.ijforecast.2009.10.004

Franck, E., Verbeek, E., & Nüesch, S. (2010). Prediction accuracy of different market structures - bookmakers versus a betting exchange. *International Journal of Forecasting*, *26*, 448–459. doi:10.1016/j.ijforecast.2010.01.004

Haahr, F. G. (2011) Market Efficiency: An Analysis of the Internet Betting Exchange Market [Online] CBS. Available from: http://studenttheses.cbs.dk/ [Accessed 9th July 2015]

Hardiman, S. J., Richmond, P., & Hutzler, S. (2010). Long-range correlations in an online betting exchange for a football tournament. *New Journal of Physics*, *12*. doi:10.1088/1367-2630/12/10/105001

Klaassen, F. J. G. M., & Magnus, J. R. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research*, *148*, 257–267. doi:10.1016/S0377-2217(02)00682-3

Majumder S. R., Diermeier D., Rietz T. A. and Amaral L. A. N. (2009) *Proc. Natl. Acad. Sci. USA* 106, 679-684

Tsay, R. S. (2005). *Analysis of Financial Time Series*. *Technometrics* (Vol. 48, pp. 316–316). doi:10.1198/tech.2006.s405

Wagner, A. K., Soumerai, S. B., Zhang, F., & Ross-Degnan, D. (2002). Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics*, *27*, 299–309. doi:10.1046/j.1365-2710.2002.00430.x

Zhao, Y. (2013). R and Data Mining: Examples and Case Studies [Online] Available From: www.RDataMining.com [Accessed 8th August 2015]

Zucchini, W. Nenadic, O, *Time Series Analysis with R – Part 1*