National College of Ireland

Comparative analysis of data mining versus human intuition in the prediction of horse race outcomes

MSc Research Project
Data Analytics

Sean Dwyer
Student ID: 22124314

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sean Dwyer
Student ID:	22124314
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	12/08/2024
Project Title:	Comparative analysis of data mining versus human intuition in the prediction of horse race outcomes
Word Count:	6223
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	15th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Comparative analysis of data mining versus human intuition in the prediction of horse race outcomes

Sean Dwyer
22124314

Abstract

Research into how machine learning models can be applied to predict the outcome of horse races. An investigation into the use of different data and learning model approaches to find the most effective and profitable strategies. Compare these against industry experts, and incorporate their knowledge to gain an edge.

Various regression and classification models were employed to predict the race outcome. Pitting the best prediction model against its human counterpart yielded higher profits for the machine learning models. In conclusion, these results show that machine learning is better at predicting horse race outcome than human experts. These results could be further improved and optimized over time as more data becomes available for reiterative model retraining.

1 Introduction

The widespread availability of broadband, and internet accessibility in general, has completely changed the landscape of sports betting. From a previously laborious physical activity of either attending in person race meetings, or bookmakers shops, the convenience of placing a bet online is why the industry in Ireland is worth billions to government coffers. In fact, it would be correct to say that betting in general, whether online or in person, far outweighs the other revenue streams in the horse racing in particular. Along with each horse race meeting, there are a plethora of experts, tipsters, and pundits, all biding for the attention of the punter. Couple this with online forums, groups, and tipping websites, places a multitude of choice in the hands of the individual. Making a bet selection benefits from being able to analyse these sentiment data sources, as well as forming ones own opinion through careful analysis of horse racing data, or form, that is widely available online, and in newspapers.

The research herein seeks to apply data mining, machine learning, techniques and algorithms to raw physical form and historic data, and compare against an analysis of group, and individual sentiment data. Secondly, the research take the optimal output from these first analytical steps, and combine with research into betting market use as a predictive tool. This can be somewhat aligned to financial market analysis where investors look at history, sentiment, and market performance of financial instruments.

The main data mining and machine learning research data is garnered from historic horse race, horse breeding, and race meeting betting markets data sets. For sentiment analysis, data is retrieved via web scraping of various pundit and tipster sites, and using application program interfaces (API) where available. Feature analysis, reduction, and

optimization is then applied, and a variety of machine learning models trained, tested, and validated, to determine the optimum model for each data type. All of which is used to predict race outcome.

Research Question. How does human intuition compare with machine learning when predicting outcome of horse racing? Sub-question : Can machines make more profitable decisions than humans when betting on horse races, using predictive data and current market analysis?

The research presented, although seemingly novelty in nature, can be considered to contribute to a number of areas in data analytics, and can be applied to real-world industry - not solely betting and bookmakers, but other financial sectors such asset management, pensions and investments, and other areas where predictive analytics can be utilised for commercial gain.

The remainder of this paper is presented in the following sections:

- Section 2 discusses related work in the field of horse race prediction.
- Section 3 highlights the research method/methodology used in completing this research, including data sets, data methodology, data models and process.
- Section 4 presents the design and implementation of the framework used to model and evaluate the research
- Section 5 details the evaluation and comparative analysis of the results of the machine learning models/algorithms, sentiment analysis, and betting markets prediction model.
- Section 6 provides a summary of the research and concludes with future work possibilities.

2 Related Work

The information age has opened the door to sports prediction. Whether it is football, baseball, soccer, or horse racing, there is a myriad of information, sites, schemes, systems all pertaining to the art of prediction.

There has always been focus on betting markets in the horse racing industry. Stekler et al. (2010).

The availability of data and the capabilities available on the web that support machine learning, and artificial intelligence has prompted much interest in the science of prediction.

Stalwart work has been done on establishing a foundation for sports prediction using an extensible framework Bunker and Thabtah (2019) upon which to build models. Building on the various model previously proposed by Gupta and Singh (2024), or Selvaraj (2017) a number of models will be assessed and employed herein.

There has also been foundational work done by Davoodi and Khanteymoori (2010) in the usage of neural networks in particular in the prediction of horse racing outcomes. This can be coupled with the excellent comparative analysis by Lyons (2016) in the area of greyhound race prediction. This is closely aligned with the proposed research.

The subject of expert analysis covered by Lyons (2016) and Song et al. (2007) involves sentiment analysis and crowd or group thing Brown and Reade (2019) which may be apt

for certain sports is somewhat lacking in the horse racing field. It is preferable to opt for the more quantitative measurements such as tipsters and pundits.

Betting market efficiency is at the heart of many studies such as Gonçalves et al. (2019) or Hubá\vcek et al. (2019) on deep learning models (albeit economics), but could easily be ported to sports betting (or stock markets). While not extensively making use of the betting market, our research will involves calculations around pricing and various statistical calculations ARIMA, and rolling averages.

Pricing is essential to successful betting strategies, and there will be allusion to work done by Zhang and Thijssen (2022).

2.1 Research Niche

The proposed project will employ novel research elements to machine learning models, and betting markets through the use of uniquely featured datasets, augmented sports analysis frameworks, in an attempt to evaluate the efficiency of said models against a human counterpart.

3 Methodology

The industry standard Cross Industry Standard Process for Data Mining (CRISP-DM)defined by Chapman (2000) is adopted as the steadfast framework upon which the research will be built. However, as the subject is sports related, a slight variation is followed. The framework, called Sport Result Prediction CRISP-DM (Bunker and Thabtah (2019)) , has more of a sports focus so more apt to our research. 1.

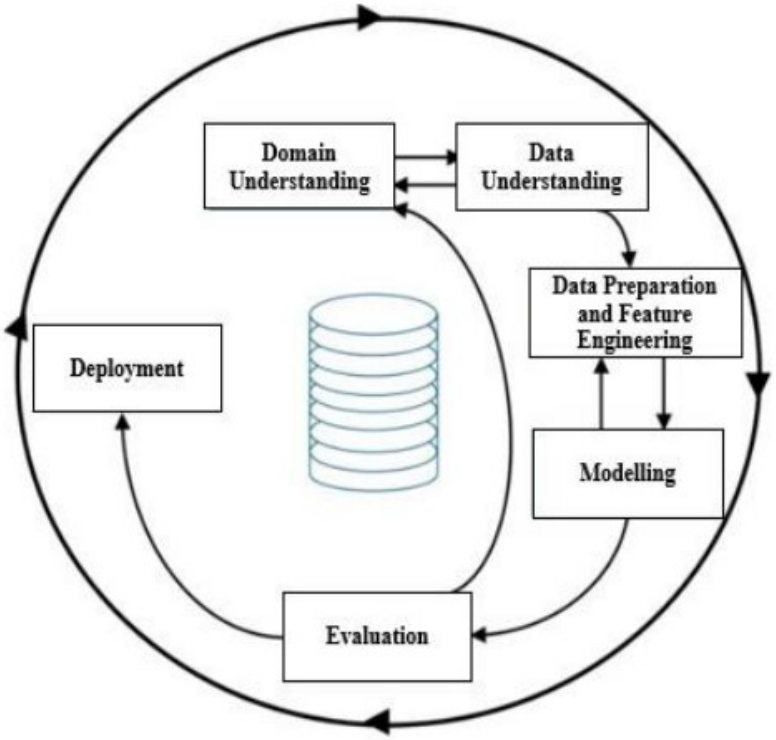


Figure 1: SPR-CRISP-DM - Bunker et al. (2019)

3.1 Domain Understanding

Our research domain is specifically focused on horse racing prediction in UK and Ireland. This is primarily because there are more than 10,000 horse races ran in the UK and Ireland each year. As mentioned, the betting industry in UK and Ireland is worth billions to the exchequer, and the availability of not only statistical, but breeding , and other data for UK and Ireland is second to none.

There are two main seasons of racing in focus of this research. These are flat and national hunt races. Flat racing is that with no hurdles, or jumps and is run on flat (or turf), and is considered to run annually from late March to October. National hunt racing has chases, hurdles, and bumpers which have varying heights of fences, and hurdles - this runs from July to April, so overlaps with flats season.

Any data that is used should span over an entire calendar year from March to April (at least 13 months), as horses are trained, and entered into races based on these schedules. This ensures that the data can be garnered and used for prediction within a defined seasonal space.

3.2 Data understanding

As mentioned, the data ideally should span the two seasons, to give an accurate representation of the horses performance. Also, the surrounding industry of pundits, tipsters, and the general racing community are generally active around these times with betting market, and commentary data most active and accurate around these schedules.

We will consider two distinct datasets for training predictive models, and a separate

data set for tipster (sentiment and tip) analysis for comparison with the selected, most efficient, models.

All datasets will be loaded into a PostgreSQL database (horseracedata) to allow cleaning and preparation of data, prior to the data modelling phase.

3.2.1 Model datasets

For model training, testing and validation, the data sets contain horse and race data, and are from two distinct sources.

3.2.1.1 Modelling DataSet 1 From Kaggle, we are using *Horse Racing Data from 1990 -2020* (2014) and extracting the data for 2015 and 2016. For each year, there is a races and a horses CSV file. These are extracted and put into the horseracedata database, race and horses tables respectively. The tables are linked using the race_id key and contain data from many different countries.

Country	No. Of Races
United Kingdom	20,075
Ireland	5099
France	3761
United States	1181
Others	2733

Table 1: Races per country.

Only UK and Ireland data are used, as they are common to both modelling datasets. For UK and Ireland, there are 25,174 races over 90 courses for the years 2015/2016 with a total number of 33,072 horses competing in these races.1.

Races The races table contains the data for each of the seasons.

FIELD NAME	DESCRIPTION
rid	Race id
course	Course of the race
time	Time (GMT) of the race in hh:mm format
date	Date of the race
title	Title of the race
rclass	Race class
band	Band
ages	Ages allowed
distance	Race distance
condition	Surface condition
hurdles	Hurdles, their type and amount
prizes	Places prizes [array-comma separated]
winningTime	race time for 1st horse
prize	Prizes total (sum of prizes column)
metric	Distance in meters
countryCode	Country of the race (UK and IE only)
ncond	condition type (created from condition feature)
class	class type (created from rclass feature)

Table 2: Races table.

Horses For each race, linked by race id, this table contains a list of horses that ran in said race, with valuable data about the horses race statistics and breeding information.

FIELD NAME	DESCRIPTION
rid	Race id
horseName	Horse name
age	Horse age
saddle	Saddle - where horse starts
decimalPrice	1/Decimal price (probability)
isFav	Was horse favorite before race start? can be more than one.
trainerName	Trainer name
jockeyName	Jockey name
position	Finishing position, 40 if horse didn't finish
positionL	how far behind horse in front this horse has finished
dist	how far this horse has finished behind winner
weightSt	Horse weight in stone
weightLb	Horse weight in pounds
overWeight	Overweight code
outHandicap	Handicap
headGear	Head gear code
RPR	Racing Post Rating
TR	Top speed
OR	Industry Official Rating
father	Horse's Father name
mother	Horse's Mother name
gfather	Horse's Grandfather name
runners	Runners total
margin	Sum of decimal Prices for the race
weight	Horse weight in kg
res_win	Horse won or not
res_place	Horse placed or not

Table 3: Horses table.

3.2.1.2 Modelling DataSet 2 From Racing-bet-data, we are using *Racing Bet Data* (2024). This dataset is a series of daily excel spreadsheets, which are individually uploaded to the RBD-Results table in the horseracedata database. This dataset is a combined race and horse results table, with a single row containing both race and horse information for that race. Additionally, there is added data pertaining to betting market horse pricing(betting odds) for each horse for the 15 minutes up to the start of the race. Because this data is sourced from a commercial data provider, there are no preparation steps required, as the data integrity and cleanliness is guaranteed by the data provider. However, there are some fields for which there is no data available, such as Official rating - where a horse has not been rated. These will need to be defaulted in the preparation process. There will be feature engineering required, if some categorical data fields are to be used.

3.2.2 Tipster dataset

This is the dataset that will be used to compare the tipster (tips) with the models to research which performs better. This is the basic premise of the research presented in

this paper. From Kaggle, we are using *Horse Racing - Tipster Bets* (n.d.) and extracting the data into the tips table in the horseracedata database. This dataset contains approx. 39,000 bets using 31 tipsters. The names have been changed for ethical reasons. In addition to the Kaggle data in this table for 2015/2016, it will also contain tipster information for 2023/2024 that will be scraped from the gg.co.uk data source, and possibly an additional source (irishracing.com) if data is made available.

3.2.2.1 Tipster table The tips table has the following fields

FIELDNAME	DESCRIPTION
UID	Unique Id
ID	Id
Tipster	Name of Tipster
Date	Date of Race
Track	Track/Course name
Horse	Horse tipped
Bet Type	Win or Place (Each way) bet
Odds	Prices of the horse at race time
Result	actual result
Tipster Active	is tipster still active

Table 4: Tips table.

3.3 Data Exploration

Having procured relevant datasets, a more in-depth exploration of the data set themselves is required. A standard exploratory data analysis (EDA) is undertaken to better determine the use of the data in the modelling exercises.

For brevity, only a single dataset will be discussed (the other so closely related). The 'tipster' dataset does not warrant further analysis, as it is used only in the comparison phase of the research.

3.3.1 Modelling Dataset 1

This dataset was introduced in the section above and consists of horse and race data for UK and Ireland for the years 2015 and 2016. This range appropriately spans both major seasons of the horse racing calendar (April to April). The two separate datasets/dataframes will be merged into a single usable dataframe.

3.3.1.1 Races The races dataset is quite static, in that it is simply a list of dates, times, and venues, for each race of the season. It is linked to the horse data by means of a race_id.

Null values a number of columns were found to have a high volume of null values, so have been removed from the data set. These are hurdles(15791) and band (11476). rclass has 5099 null values - in this case NULL denotes an race in Ireland (class is not applicable here). Therefore rclass nulls will be relabelled Class-Ire and encoded.

Numeric values metric is used for distance in place of imperial mile/furlong. winningTime is the time recorded as the winner crossed the finish line. prize is the winning prize money for this race. ncond and class seem to be pre label-encoded values categorical variables and will be discarded (and re-encoded later).

Categorical values rclass will be label-encoded in place of using class feature It can be seen here that Irish races total approx. 5000 (class = 0).

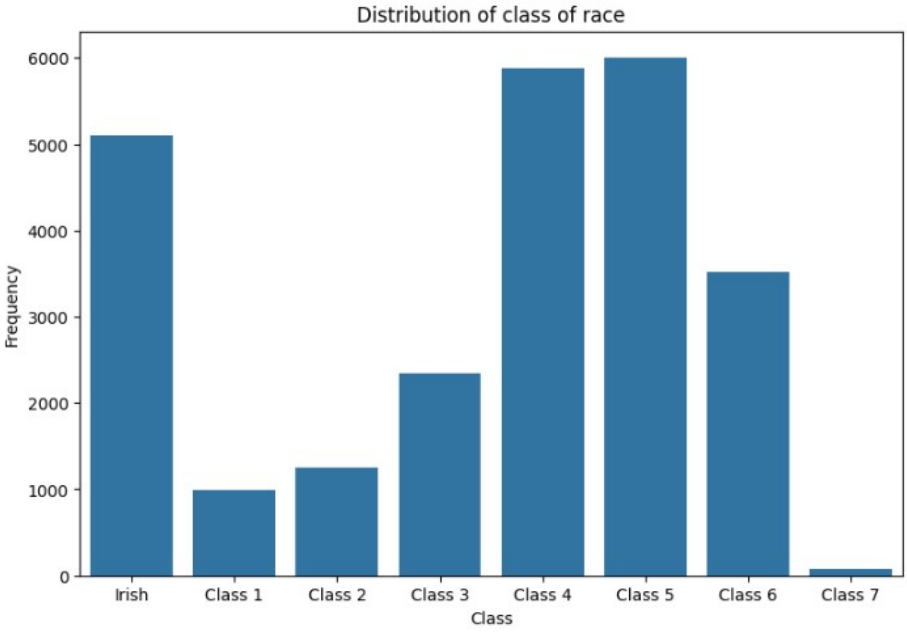


Figure 2: Distribution of class of race

Similar to rclass, **ncond** is a numeric counterpart value of condition. this will not be used, but condition will be label-encoded. The goal is to minimize dimensionality, so this encoding will only be used if unavoidable.

3.3.1.2 Horses This dataset is linked to **races** the rid (race_id) variable. It contains valuable performance information about the horses. Number of Horses: 33072 Number of Jockeys: 1607 Number of Trainers: 1816

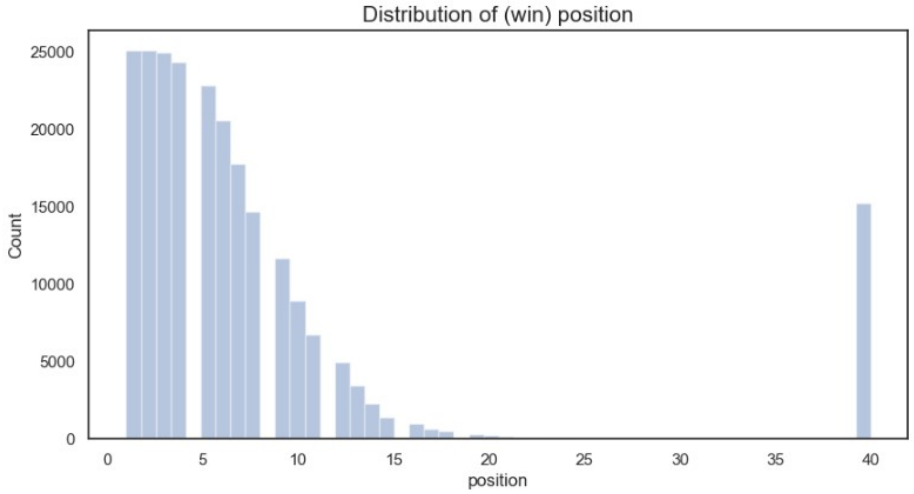
Null values There were a number of null values identified.

Field	Null count	discarded
overWeight	231120	Y
outHandicap	226953	Y
headGear	50429	Y
dist	65299	N
OR	61255	N
TR	51260	N
positionL	40203	N
RPR	20447	N

Table 5: null in horses dataset

The null values that are retained, have some correlation, and will be further analysed in the data preparation phase. For example OR and RPR are horse ratings from 'Official' and 'Racing Post' respectively.

Numeric values some numeric values appeared in the null values section. One important value (position) was found to have a particular outlier, as can be seen from the graph below.



Since position 40 means that the horse did not finish the race, these entries will be removed from the dataset. Once removed, the analysis of other numerical fields makes sense.

For horses weight, the field, **weight**, denoting weight in kilos will be used instead of weightSt and weightLb (which would have to be combined).

After preliminary data cleaning there now remains - horses: 32193, jockeys: 1569, and trainers: 1730 in the dataset.

3.4 Data Preparation & Feature engineering

Having explored the dataset(s) in the previous section, the data will be further embellished with additional features as well as dimension reduction using domain knowledge and further analysis of variables for commonality and collinearity. This is in preparation

for the model implementation phase wherein the dataset will be fitted to each model, and feature importance measured.

3.4.1 Modelling Dataset 1

The two datasets races and horse are combined to for a single manageable dataset, merging races data [metric (distance), winningTime, rclass, and condition] with horses data.

3.4.1.1 Feature analysis and engineering Having added race variables to the dataset, we look to analyse the correlation matrix below.

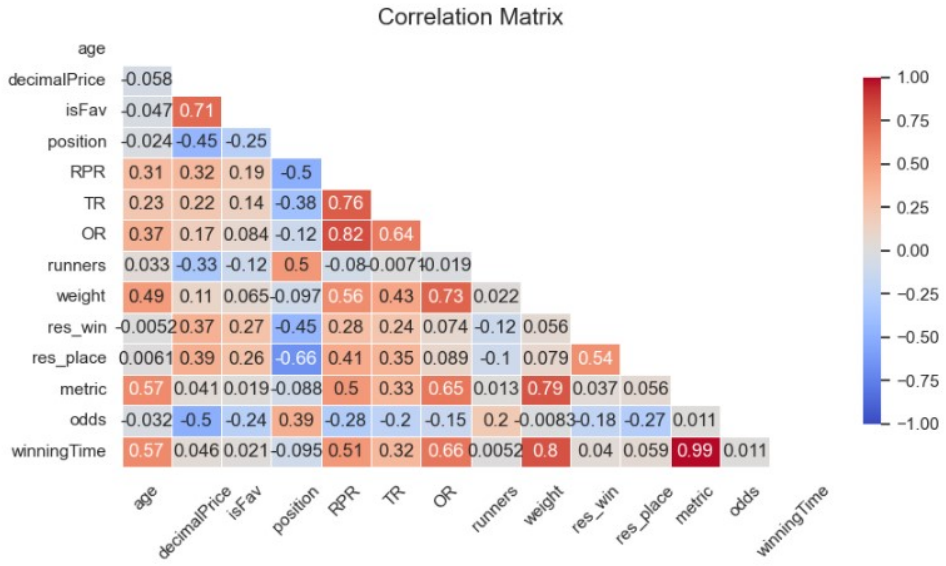


Figure 3: Correlation of dataset 1

Updates - The odds feature has been added - this represents the odds or 'price' of the horse ,converting and rounding from decimalPrice . i.e. $0.06667 = 1/0.066667 = 15$ or '15 to 1' in betting parlance. Plotting this against the actual winning horses shows a clear correlation - i.e. the lower the odds (price) the more horses won. This is analogous to the initial plot of probability against finish position.

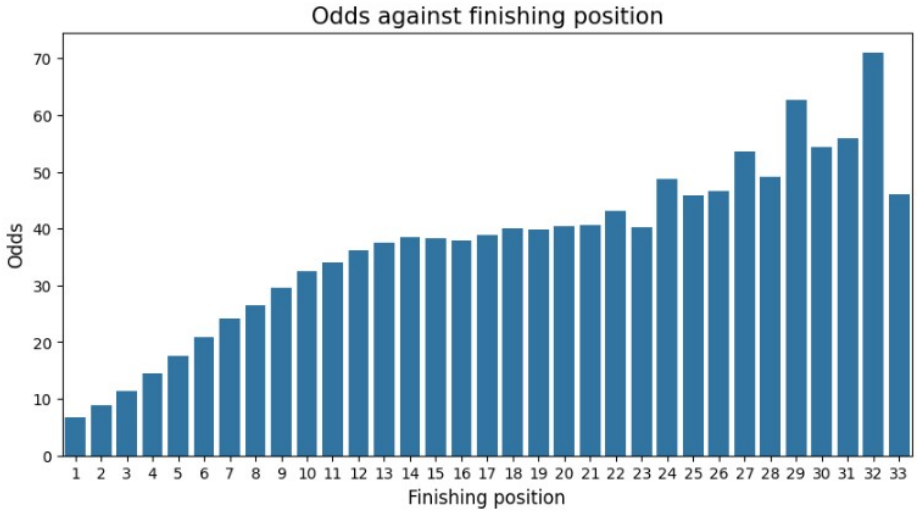


Figure 4: odds and finishing position

The horse ratings features RPR and OR correlate most closely to the RPR rating. This is not surprising as the rating systems are closely aligned. The close correlation figures for each rating against other features means RPR can be selected as the 'rating' feature, the others dropped. From previous analysis, RPR has the least amount of NULL values, so is used.

The feature positionL and its associated **dist** feature denote the distance finished behind horse in front, and cumulative distance (respectively) from winning horse. the industry term values for head, shoulder, neck (hd, sh, nk) are converted to 0.25 and 30 for value of 'dist'. Since, only one feature is required, the cumulative feature **dist** (distance from winning horse) is used, because it better represents the intent and measure for the prediction - which would otherwise have to be calculated (from positionL).

Engineered features - As per previous analysis done by Schumaker (2013) on horse-racing (harness) data regarding the optimal amount of historic information required to predict finish_position we add a new feature (AvgPosLast5) that averages the finish position of each horse for the past **5 races**. This is a rolling average and was garnered from work also done by Lyons (2016) albeit greyhound rather than horse racing.

Added a feature (HorseRankTop50Percent) that denotes the horse is ranked in the top 50% of winning horses.

3.5 Feature selection

This section describes the feature selection process. Analysis is done on the datasets to discern which of the features to use to achieve best results. An iterative approach is taken, in that the output from the implementation section may result in a reiteration over the feature selection for a particular model.

In order to use the feature_importance functionality, training and test sets must be created from our data. An 80 / 20 train/test divide is created to use in data modelling. A further month of data is also retained as a validation for each model. This will be used to assess the predictive capabilities of each model in the evaluation section.

The racing seasons span the calendar from late-March to following April, so the training and test split data can be conveniently split into dataframes of 75:25 ratio, i.e. 9 months and 3 months respectively - March - December inclusive, and January - March inclusive. Prediction/validation dataframe will contain April race and horse information.

In each of the dataframes, we need to add new features denoting trainer and jockey average finish positions. These are done at this stage, because the values have an effect on only the trainer and jockeys in each set.

3.5.1 Classification models

Taking the feature set from a feature importance analysis has been undertaken using Logistic Regression and RandomForestClassifier models. the features are all the numeric features in training, and prediction is res_win . These will assist in selecting feature to use in model.

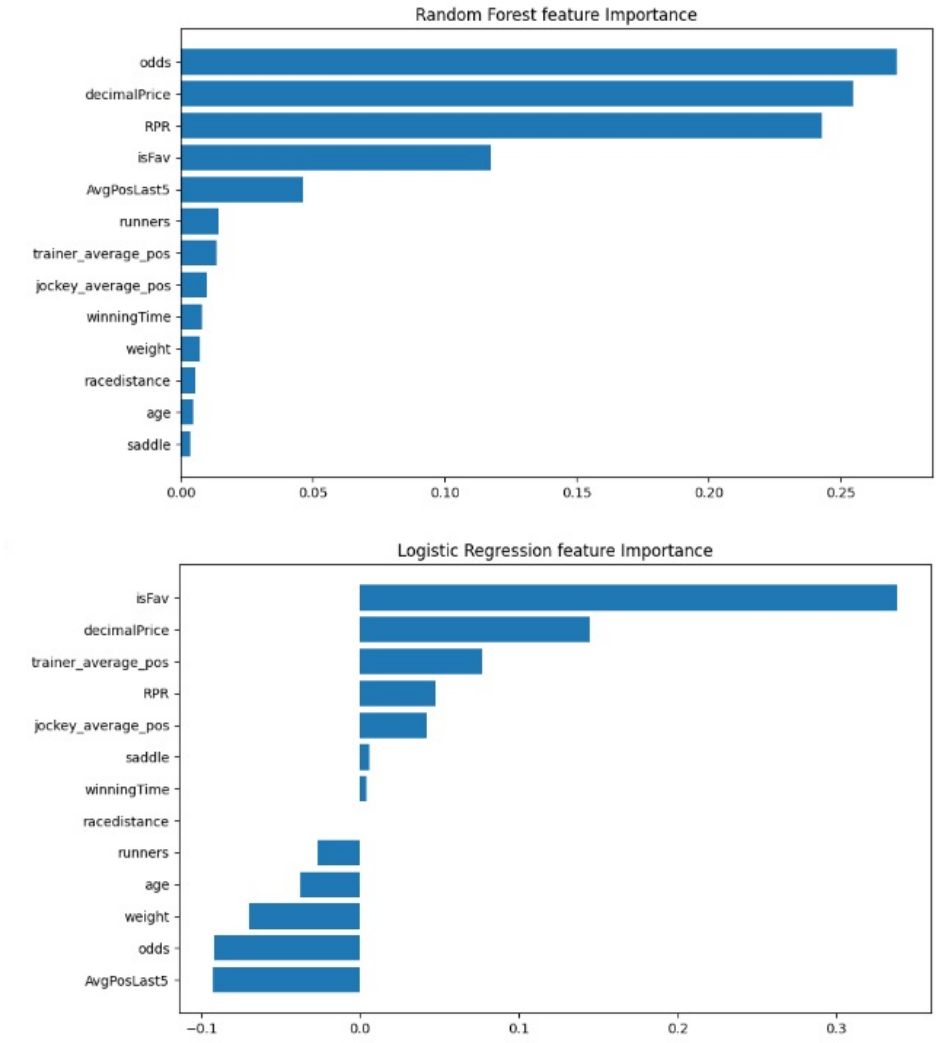


Figure 5: feature importance for classification model

3.5.2 Regression models

Similarly feature importance was run for regression models.

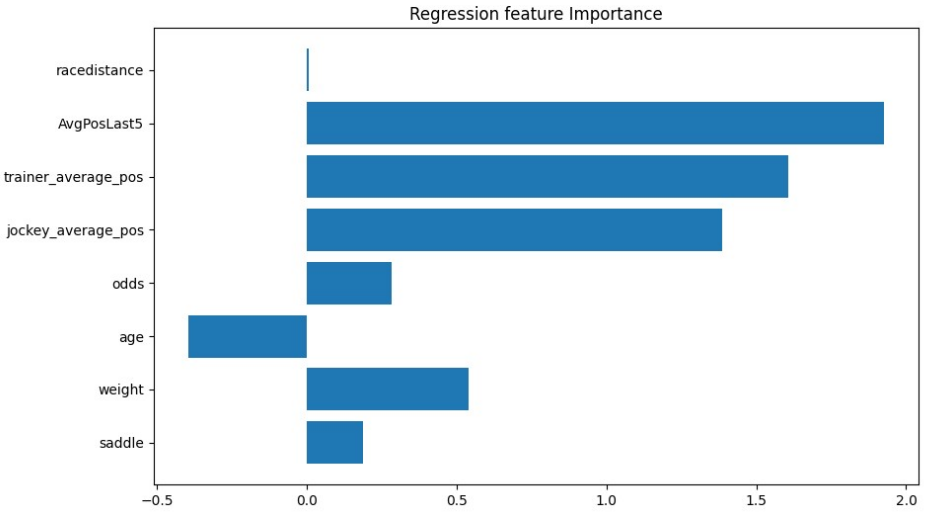


Figure 6: feature importance for regression model

Now we have a fundamental dataset we can experiment with different models to see which are the most effective.

It is worth noting that the second dataset will be used for modelling exercise. As it is from a commercial provider there was minimal updates or cleaning required.

4 Design Specification

Having taken the dataset through the initial phases of cleansing, feature engineering, and feature selection. In order to effectively reiterate across multiple datasets, and multiple experimental scenarios, a framework or process must be defined. The technical details and more detailed descriptions of the features presented below can be found in the accompanying configuration manual document.

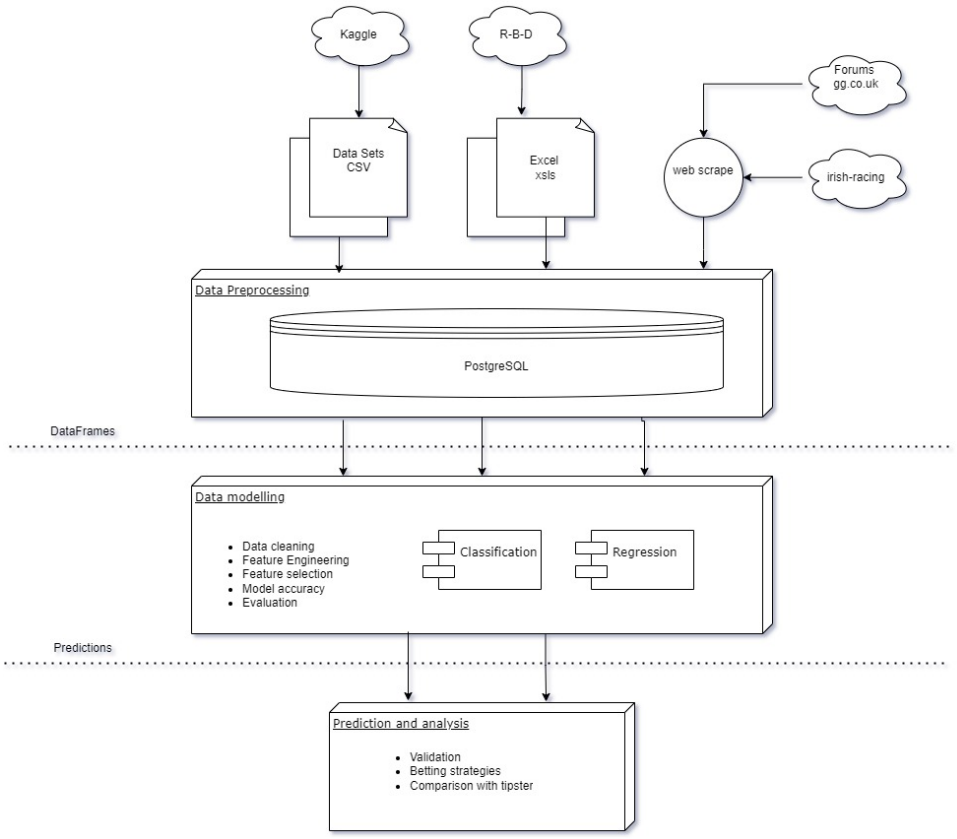


Figure 7: Project framework

The flow is from top-down. Raw data sourced from Kaggle, Racing-Bet-Data and scraped from websites. Run through an ETL (Extract, Transform, Load) process, each data are loaded into the database (PostgreSQL). Some data require transformation, such as date formatting. This is primarily done using Pandas DataFrame objects to read from CSV and Excel, and insert data into database tables. Some cleaning takes place in the database tables, through the use of SQL. data embellishment, such as meaningful values instead of NULLs, and removal of corrupt or unusable rows.

The data modelling phase extracts data from the database, into dataframes for ML model creation. Further cleaning of data is necessary here. Additional features are added to the dataframes. In this phase a number of classification and regression models are created.

This is the crux of the research, wherein selected models are used to evaluate the performance of the human ('tipsters') against the performance of the machine learning models. this is done using validation data (a small percentage (1 months) from the original datasets)).

5 Implementation

This is the most intensive and iterative part of the research process. The data that has been prepared is used to create and evaluate performance of a number of machine learning models. Both classification and regression modelling is used, in order to cover a variety of approaches.

5.1 Tools

As aforementioned in the design phase, there are a number of tools and paradigms used to implement the process defined in the design section.

Across all functional areas is the usage of the Python language (3.7+).

5.1.1 Data

Initial ETL is performed using Python to load CSV and Excel files. These are loaded into a PostgreSQL database hosted using Docker technology.

5.1.2 Modelling

Creation of models is implemented in Python using a number of libraries with sklearn being the most widely used in the solution.

5.2 Classification and Regression model creation

Data feature engineering and selection a feature_importance was performed using basic regression and classification model baselines. These models are further enhanced, and extended using hyperparameter tuning, to allow creation of predictions that will be used in further evaluation and comparison against the human data sets (namely tipster information).

In this part of the process, each model is created, trained, and tested using the datasets. Each models performance is evaluated here using the standard measurements for the model category (classification or regression).

It is important to note here, that the output of this part of the process is not solely the standard measures of model efficiency and fit. There is a 'predicted' data set output during the training and testing of each model. Both classification and regression models use this **predict** dataset to output prediction for each model type. This is used in the subsequent evaluation section to compare predictive performance of each model in betting strategies, and also performance against the tipsters.

The top performant models are saved (to pickle) for use in the deployment phase to predict inputs from a web client.

[The outputs from each contain exactly the same amount of rows to allow merging in the next phase - these match the predicted datasets.]

The regression output has only a single column (distance from winner), the classification models predict two classes for win or place (first three positions).

5.2.1 Classification

Because our binary classification is sensitive to the imbalance in out data, we use the Precision-Recall AUC to measure effectiveness of the model. The focus is more more

about the positive class, so using PR AUC, which is more sensitive to the improvements for the positive class, is a better choice. The fraction of positive classes res_win and res_place, is small. Each of the regression models is tested and standard model evaluation metrics are below.

Classification is trying to classify the outcome as win or place - each model is fitted and predicts these separately.

Model	Prediction	CV-F1	F1-Score	PR-AUC	Recall	Precision
LogisticRegression	res_win	0.845	0.873	0.300	0.317	0.665
LogisticRegression	res_place	0.676	0.734	0.488	0.512	0.639
RandomForestClassifier	res_win	0.838	0.872	0.321	0.504	0.509
RandomForestClassifier	res_place	0.691	0.736	0.492	0.594	0.603

Table 6: Standard classification measurements

Below shows the confusion matrix for logistic regression on res_win (winning horse) classification. there is a significant imbalance in the negative vs positive results evident in the high figures for true negative / predicted negative quadrant.

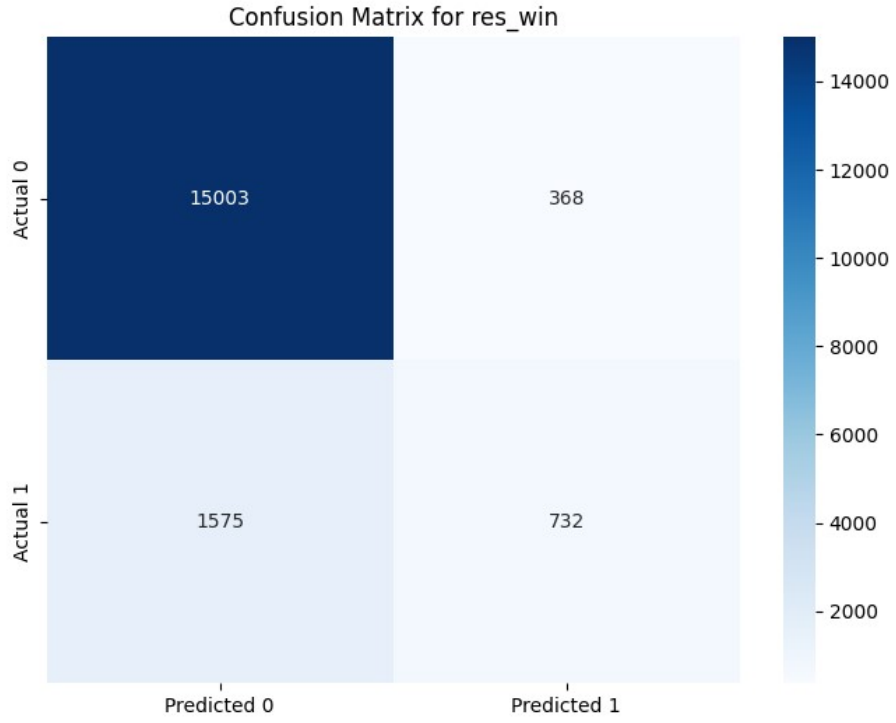


Figure 8: Confusion matrix for Win result (Logistic regression)

5.2.2 Regression

The target for regression models is distance from winner. Obviously, the lower this figure, the better chance the horse has of winning.

Model	RMSE_train	RMSE_test	Generalization %.
LinearRegression	15.175	18.211	20.007
KNNRegressor	14.040	17.220	22.650
RandomForestReg	14.002	16.893	20.647
SVR	16.636	20.577	23.69

Table 7: Standard regression measurements

Model	win_Tr_Ac.	win_Tst_Ac	place_Tr_Ac	place_Tst_Ac
LinearRegression	0.230	0.806	0.507	0.918
KNNRegressor	0.239	0.800	0.516	0.923
RandomForestReg	0.220	0.892	0.516	0.998
SVR	0.223	0.882	0.505	0.941

Table 8: Accuracy measurements

6 Evaluation

The output from the implementation process is a number of prediction datasets. In this part of the process a new prediction dataset is introduced for each of these models. Use is made of here, of the tipster dataset also.

The prediction dataset is validated against each model, and used in a 'betting' strategy to compare winnings firstly against each model, and secondly and simultaneously against the 'human' tipster prediction.

The resulting comparison is two-fold, in that firstly the evaluation of the best model i.e. that produces the largest profit is found, and secondly, the comparison of said model against the human tipster.

To keep calculations simple, a standard unit of wager is 1 (pound or euro).

The validation set used is a months worth of racing data. This contains the actual results for all races within the month. In each scenario the predicted values from classification and regression models for this month is compared against the actual values, wagers are made, and a cumulative total is taken.

6.1 Baseline

The baseline scenario runs through the dataset and places a bet, in each race, on the horse with the lowest odds, the favourite. The results shown, confirm a well-known adage that betting on favourites does not make financial sense.

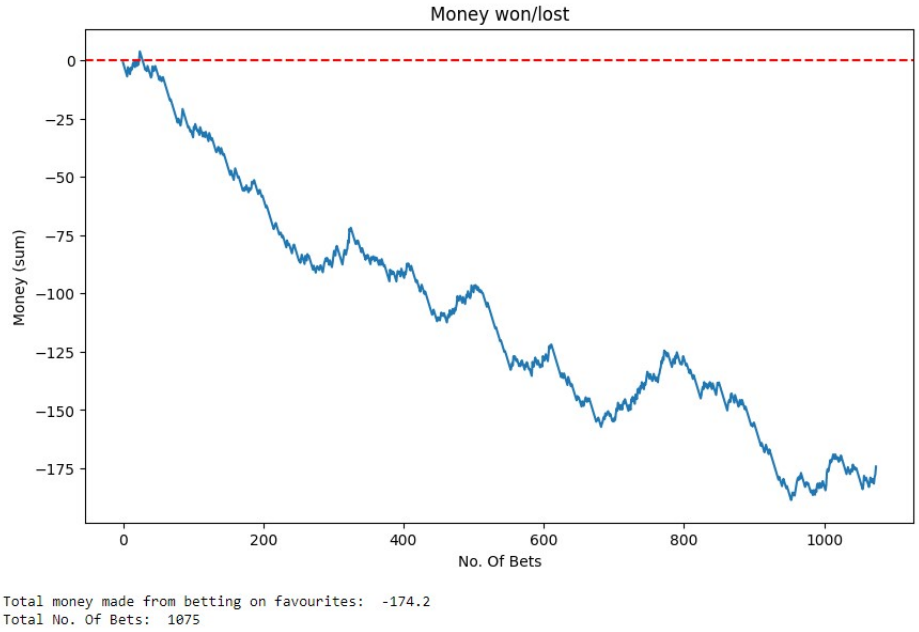


Figure 9: baseline scenario - betting on favourites

6.2 Classification models

In order to evaluate the efficiency of the classification models the output of the implementation phase,i.e. the predicted data set, containing the predicted values is merged with the actual results, and the bets are wagered (units of 1 euro/pound), and the overall returns profit or loss are calculated based on a win , and the winning price. Below is the result of the Logistic Regression classifier for the validation period for win (or res_win target). It immediately outperforms the baseline, in terms of profitability.

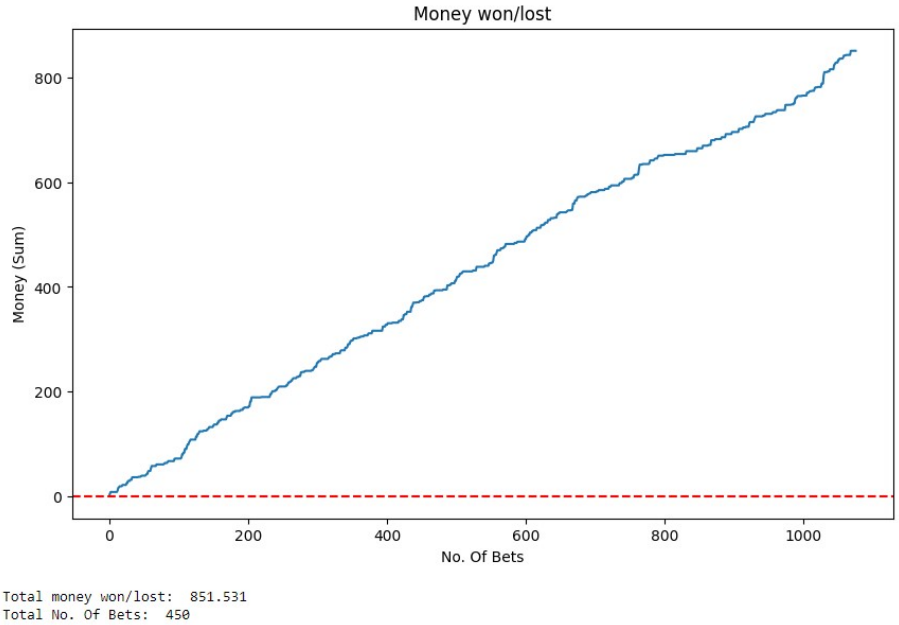


Figure 10: Classification betting scenario

the res_win was also run for the Random Forest Classifier, and output is below:

Model	Target (y)	Money won/lost	No. Of Bets
Logistic Regression	Win	851.53	450
Random Forest Classifier	Win	1327.82	761

Table 9: Classification (Win) measurements

Classification models were also run for the Place (or top three) target. This yielded the following results.

Model	Target (y)	Money won/lost	No. Of Bets
Logistic Regression	Place	2755.92	1975
Random Forest Classifier	Place	3133.08	2334

Table 10: Classification (Place/top 3) measurements

6.3 Regression models

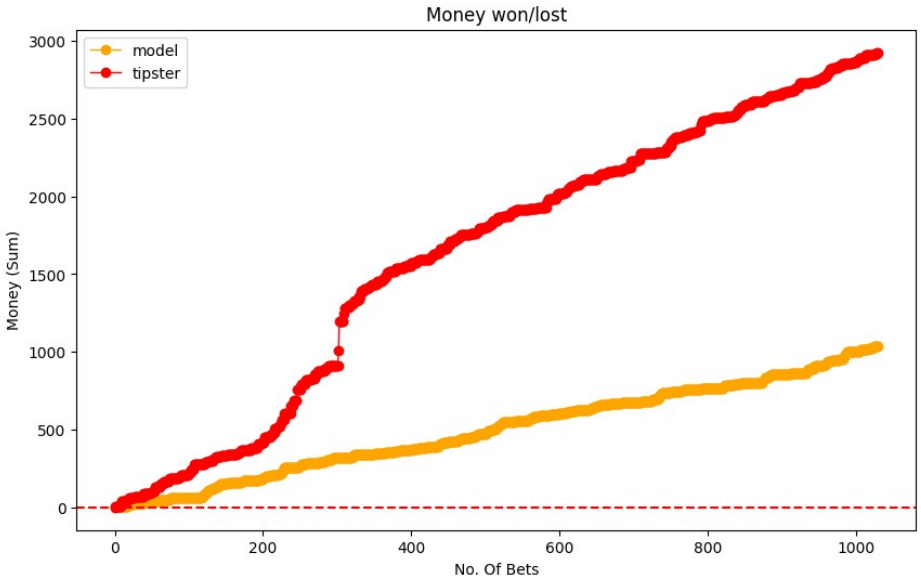
Regression models targeting the distance from winner were evaluated using the validation/prediction data. Each race predicted is measured against the actual results, producing the following output.

Model	Money won/lost	No. Of Bets
Linear Regression	4371.89	1075
KNN Regression	4344.19	1075
Random Forest Regressor	4914.64	1076

Table 11: Regression measurements

6.4 Comparison of classification and tipster performance

This section is the crux of the research, in that it compares human prediction (tipster) with the models. Since the tipster will be trying to predict the winner (or place) of a race, the classification Win (res_win) model is most apt. To keep things fair, only races from the validation set in which the tipster actually predicted will be compared with the models counterpart prediction. The results, are shown below.




```
Total money won/lost:  1036.671
Total No. Of Bets:  439

Total tipster money won/lost:  2926.44
Total no. of winning tipster bets:  282
Total no. of losing tipster bets (1 euro):  928
```

Figure 11: classification vs tipster

6.5 Discussion

From the above experiments, it can be seen that models are more efficient that simply 'following the herd', i.e. betting on favourites. Surprisingly, this is a popular strategy. Using models to predict the outcome of races seems quite profitable. However, it is worth noting that the sheer amount of bets made, sometimes in the thousands yield only small returns.

The total number of bets is just that, the total of winning AND losing bets placed. It would have been better to calculate the win/loss percentages, too. This would give a more accurate assessment of the proficiency of the models.

In the main classification vs tipster evaluation, it is evident that the human is more profitable. This is possibly due to some random 'high odds' wagers. This can be seen in the graph. However, it is also evident that the tipster had a significantly higher number of losing bets. the classification model has a low number of bets in comparison, however, the win/lose ratio is unfortunately lacking here. Only conjecture can suggest that the model is more efficient.

It would have been beneficial to have another tipster set to compare, but this was unobtainable from the source originally intended. However, given that the racing industry is known for its longevity and non-change, similar results would be expected even with newer data.

7 Conclusion and Future Work

The main premise of the research was to assess the ability of machine learning models to compete with the human intellect and intuition when predicting the outcome of horse races. A lot of time was spent assessing and cleaning the data to the detriment of the overall objective. Although the question was somewhat answered (in the Implementation/Evaluation) section, there should be more focus on these comparisons and not on the models themselves.

In future research, it would be beneficial to get more data from the human side. The original intent was to use semantic analytics to garner information from various social media, but budgetary restrictions prevailed. Further use of the 'tipster' datasets would prove useful in that regard and the information therein is more quantitative, therefore easier to compare, and possibly incorporated in future models.

Acknowledgements

I would like to thank Dr. Catherine Mulwa for invaluable help and guidance at project inception. I would like also to thank Jorge Basilio for advice and guidance throughout the journey - many thanks.

Without the support of my family and friends, I would not have been able to complete this project, and for that I am eternally grateful.

References

Borowski, P. and Chlebus, M. (2021). *MACHINE LEARNING IN THE PREDICTION OF FLAT HORSE RACING RESULTS IN POLAND*.

Brown, A. and Reade, J. J. (2019). The wisdom of amateur crowds: Evidence from an online community of sports tipsters, *European Journal of Operational Research* **272**(3): 1073–1081.
URL: *https://www.sciencedirect.com/science/article/pii/S0377221718306209*

Bunker, R. and Thabtah, F. A. (2019). A machine learning framework for sport result prediction, *Applied Computing and Informatics* .
URL: *https://api.semanticscholar.org/CorpusID:49582784*

Chapman, P. (2000). Crisp-dm 1.0: Step-by-step data mining guide, *CRISP-DM*.
URL: *https://api.semanticscholar.org/CorpusID:59777418*

Colle, P. (2022). *What AI can do for horse-racing ?*

Davoodi, E. and Khanteymoori, A. (2010). Horse racing prediction using Artificial Neural Networks, pp. 155–160.

Gonçalves, R., Ribeiro, V. M., Pereira, F. L. and Rocha, A. P. (2019). Deep learning in exchange markets, *The Economics of Artificial Intelligence and Machine Learning* **47**: 38–51.
URL: *https://www.sciencedirect.com/science/article/pii/S0167624518300702*

Gupta, M. and Singh, L. (2024). Horse Race Results Prediction Using Machine Learning Algorithms With Feature Selection, *International Journal of Intelligent Systems and Applications in Engineering* **12**(2s): 132–139. Number: 2s.
URL: *https://www.ijisae.org/index.php/IJISAE/article/view/3565*

Hong Kong Horse Racing Results 2014-17 Seasons (n.d.).
URL: *https://www.kaggle.com/datasets/lantanacamara/hong-kong-horse-racing*

Horse Racing Data from 1990 -2020 (2014).
URL: *https://www.kaggle.com/datasets/hwaitt/horse-racing/data*

Horse Racing - Tipster Bets (n.d.).
URL: *https://www.kaggle.com/datasets/gunner38/horseracing*

Hubá\vcek, O., Sourek, G. and Železný, F. (2019). Exploiting sports-betting market using machine learning, *International Journal of Forecasting* .
URL: *https://api.semanticscholar.org/CorpusID:88507081*

Li, Y.-M., Hsieh, C.-Y. and Fan, S.-N. (2024). A social selection mechanism for sports betting market, *Decision Support Systems* **178**: 114119.
URL: *https://www.sciencedirect.com/science/article/pii/S016792362300194X*

Lyons, A. (2016). *Man v Machine: Greyhound Racing Predictions*, Master's thesis, Dublin, National College of Ireland.
URL: *https://norma.ncirl.ie/2527/*

Montone, M. (2021). Optimal pricing in the online betting market, *Journal of Economic Behavior & Organization* **186**: 344–363.
URL: *https://www.sciencedirect.com/science/article/pii/S0167268121001487*

Ng, W. W. Y., Liu, X., Yan, X., Tian, X., Zhong, C. and Kwong, S. (2023). Multi-object tracking for horse racing, *Information Sciences* **638**: 118967.
URL: *https://www.sciencedirect.com/science/article/pii/S0020025523005364*

Peeters, T. (2018). Testing the Wisdom of Crowds in the field: Transfermarkt valuations and international soccer results, *International Journal of Forecasting* **34**(1): 17–29.
URL: *https://www.sciencedirect.com/science/article/pii/S0169207017300754*

Racing Bet Data (2024).
URL: *https://www.racing-bet-data.com/results/*

Sameerchand Pudaruth, Nicolas Medard, Z. B. D. (2013). Horse Racing Prediction at the Champ De Mars using a Weighted Probabilistic Approach, *International Journal of Computer Applications* **72**(5): 37–42. Place: New York, USA Publisher: Foundation of Computer Science (FCS), NY, USA.
URL: *https://ijcaonline.org/archives/volume72/number5/12493-9048/*

Satopää, V. A., Salikhov, M., Tetlock, P. E. and Mellers, B. (2023). Decomposing the effects of crowd-wisdom aggregators: The bias–information–noise (BIN) model, *International Journal of Forecasting* **39**(1): 470–485.
URL: *https://www.sciencedirect.com/science/article/pii/S0169207021002168*

Schumaker, R. P. (2013). Machine learning the harness track: Crowdsourcing and varying race history, *Decision Support Systems* **54**(3): 1370–1379.
URL: *https://www.sciencedirect.com/science/article/pii/S016792361200379X*

Schumaker, R. P., Jarmoszko, A. T. and Labedz, C. S. (2016). Predicting wins and spread in the Premier League using a sentiment analysis of twitter, *Decision Support Systems* **88**: 76–84.
URL: *https://www.sciencedirect.com/science/article/pii/S0167923616300835*

Selvaraj, P. (2017). *Predicting The Outcome Of The Horse Race Using Data Mining Technique*, Master's thesis, Dublin, National College of Ireland.
URL: *https://norma.ncirl.ie/3094/*

Smith, M. A. and Vaughan Williams, L. (2010). Forecasting horse race outcomes: New evidence on odds bias in UK betting markets, *International Journal of Forecasting* **26**(3): 543–550.
URL: *https://www.sciencedirect.com/science/article/pii/S0169207009002155*

Song, C., Boulier, B. L. and Stekler, H. O. (2007). The comparative accuracy of judgmental and model forecasts of American football games, *International Journal of Forecasting* **23**(3): 405–413.
URL: *https://www.sciencedirect.com/science/article/pii/S0169207007000672*

Stekler, H., Sendor, D. and Verlander, R. (2010). Issues in sports forecasting, *Sports Forecasting* **26**(3): 606–621.
URL: *https://www.sciencedirect.com/science/article/pii/S0169207010000099*

Sung, M.-C., McDonald, D. C., Johnson, J. E., Tai, C.-C. and Cheah, E.-T. (2019). Improving prediction market forecasts by detecting and correcting possible over-reaction to price movements, *European Journal of Operational Research* **272**(1): 389–405.
URL: *https://www.sciencedirect.com/science/article/pii/S0377221718305575*

Teall, J. L. (2023). Chapter 12 - Market Efficiency, *in* J. L. Teall (ed.), *Financial Trading and Investing (Third Edition)*, Academic Press, pp. 359–402.
URL: *https://www.sciencedirect.com/science/article/pii/B9780323909556000124*

Thaler, R. H. and Ziemba, W. T. (n.d.). Parimutuel Betting Markets: Racetracks and Lotteries.

Wunderlich, F. and Memmert, D. (2020). Are betting returns a useful measure of accuracy in (sports) forecasting?, *International Journal of Forecasting* **36**(2): 713–722.
URL: *https://www.sciencedirect.com/science/article/pii/S016920701930233X*

Zhang, C. and Thijssen, J. (2022). On sticky bookmaking as a learning device in horse-racing betting markets, *Journal of Economic Dynamics and Control* **144**: 104525.
URL: *https://www.sciencedirect.com/science/article/pii/S0165188922002299*

Štrumbelj, E. (2014). On determining probability forecasts from betting odds, *International Journal of Forecasting* **30**(4): 934–943.
URL: *https://www.sciencedirect.com/science/article/pii/S0169207014000533*