Predicting Horse Racing Results

BA REPORT

August 2020

Andrew Woon

Word Count: 4571

Table of Contents

Abstract

Since the advent of horse racing, gamblers have tried to "beat the odds". This report seeks to develop a model to predict winners and top three finishers (placers) in handicap horse races. Four years of data from Hong Kong Jockey Club was collected using data scraping technologies and analysed with features inspired by the multinomial logit model specification Bolton and Chapman (1986) developed. A ranked ordinal classifier defined by Frank and Hall (2001) was developed and proved unsuccessful with predictions only generated for the first and last values in the ranked set. It is hypothesised horse races cannot be treated as a ranked ordered set, due to inherent unexplained variance forbidding classification into well defined ranked classes. Further attempts were made, with Random Forest binary classifiers generated to predict winners and placers achieving a precision of 23.85%, which was deemed unsuccessful being outperformed by a naïve estimator choosing favourites as winners.

1. Introduction

1.1 Background

Since the advent of horse racing, bettors have sought to improve their chances of winning through careful analysis of data. With the advancement of machine learning and computer processing, it is thought that previously hidden patterns can be uncovered with the aim to beat the races.

Horse racing operates on a pari-mutuel betting system, where all bets are pooled, and winners receive an amount proportional to their wager after a percentage has been taken by the house. As a result, each bettor influences the final odds and profiting requires being able to capitalise on differences between these observed odds and the actual odds. Wagering systems have been designed specifically to exploit these differences and consist of "two components: a model of the horse race process and a wagering strategy" (Bolton and Chapman 1986). This report focuses on developing a predictive model of the former with the latter considered out of scope as many good wagering strategies having been identified such as the Kelly criterion (Kelly, 1956) specifically in the context of horse racing (Metel, 2017).

Handicap horse races are a specific type of flat racing, which aims to create an equal race amongst all participants by handicapping better performing horses with additional weights. The Hong Kong Jockey Club (HKJC) implements handicap horse races and will be the primary focus of this study. In addition to weight changes, HKJC implements a ratings system based on each horse's recent performance, which in turn determines their race class further ensuring horses of similar calibre are racing together.

HKJC consists of three distinct racecourses Sha Tin (ST) Turf, Sha Tin All Weather Track (AWT) and Happy Valley (HV) Turf. Races are grouped into three distance group brackets of short, middle, and long with all races ranging from 1000 to 2400 metres.

1.2 Literature Review

Research into modelling the horse racing process has a long history with Ali (1977) finding bettors behaving as risk lovers who on the whole consistently under-bet on horses with high winning probabilities and over bet on horses with low winning probabilities. Therefore, leading to discrepancies between the objective odds and the final odds influenced by this subjectivity, and a potential opportunity to capitalise and profit on these differences. Additionally, Hausch, Ziemba and Rubinstein (1981) took a different line, and instead of looking only at the win odds, found there are significant inefficiencies between the win, place and show pools for handicap races to the extent that substantial profits can be made.

Both Ali (1977) and Hausch et. al (1981) demonstrated success in developing wagering systems exploiting observed differences in odds generated by the bettor's preferences across the pool. Bolton and Chapman (1986) approached the problem differently using a multinomial logit model to estimate the winning probability for horses in each race. Benter (1994) applied this same methodology to handicap races in Hong Kong reporting net profits in four out of five seasons, and his success was further publicised in Bloomberg (Bloomberg, 2018).

One of the significant differences in approach discovered by Bolton and Chapman (1986) was to treat each race as a ranked ordered set. Originally developed for marketing research, Chapman and Staelin (1982) found ranked preference sets can be exploited by exploding the set such that second choice becomes the winner of a new set with the first choice removed and so forth. This technique enables rapid duplication of data for training and thus improvement of model performance. However, an important caveat is that the explosion process should be limited to a depth where choice remains meaningful. Bolton and Chapman (1986) suggests a partial explosion strategy for the top three finishers, as these jockeys will all finish in the purse.

A limitation of the methodology described by Bolton and Chapman (1986) is that the multinomial logit model is generated from a linear combination of features and fails to capture non-linearities or complex relationships between features, which one would expect in horse racing. Frank and Hall (2001) demonstrated, that similar to the explosion strategy for ranked ordered sets, ranked ordinal data can be classified through a series of n-1 binary classifiers (where n is the number of classes). The results from each of the binary classifiers can then be combined and used to calculate associated probabilities for each class in the ranked set.

2. Data acquisition

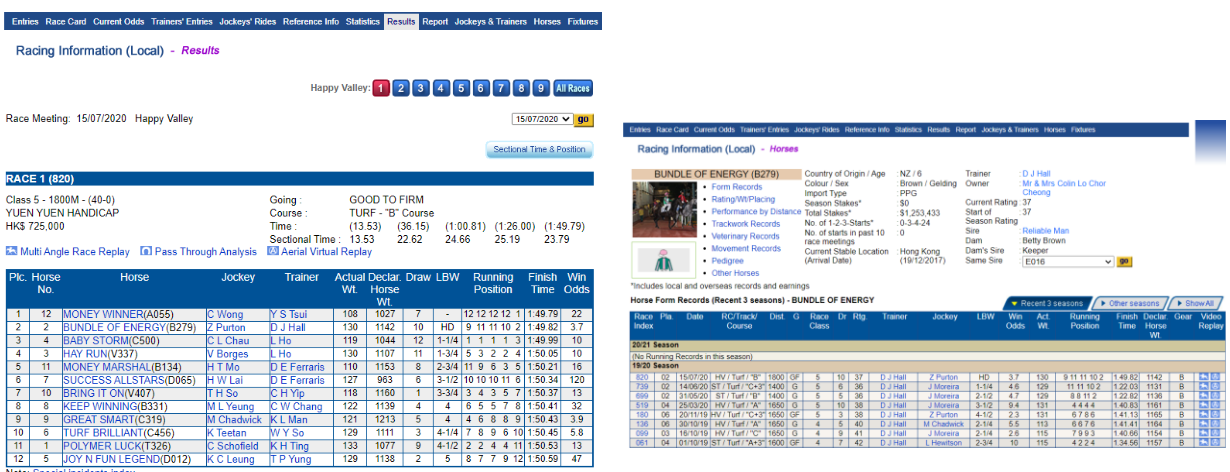
Horse racing results are readily available from the Hong Kong Jockey Club website (Hong Kong Jockey Club, n.d.), which contains historic data on local races, horses, jockeys etc. However, HKJC does not provide an API to request data or readily formatted datasets to download, therefore data scraping technologies were employed to acquire the necessary data. The scrapy library in python was used as the primary data scraping tool and was used in tandem with the scrapy_splash extension and docker container, which were required as the HKJC websites are populated using JavaScript.

The range of data to scrape was chosen to be four years starting from the 2016-2017 season. As each season contains 88 race days with approximately 10 races per day, the four-year data sample would contain ~3500 races with ~50,000 racehorse participants. This data sample would also allow the final 2019-2020 season to be held out as the test sample with the prior three seasons used for training, giving a good 75:25 training testing split.

Investigating the HKJC website three key data sources were identified as shown in Figure 1:

1. Local horse racing results including sectional times and dividend pay-outs
2. Horse information
3. Overseas horse racing results

Other additional information on jockeys and trainers were not scraped, because they only provided aggregate statistics from the prior season, which could be reverse engineered from the horse racing results. Trackwork records providing information on recent workouts and barrier trials may be useful but was not included in this analysis.



a) b)

Figure 1 – Reference of Hong Kong Jockey Club website including a) Local horse racing results, b) horse information with link to overseas racing results (Hong Kong Jockey Club, n.d.)

Scrapy uses 'spiders' to crawl unique webpages, and for this use case three spiders were built to crawl the two webpages shown in Figure 1 as well as the "Sectional Time & Position" linked from Figure 1a. One of the key lessons learnt from the data scraping was to provide the scrapy splash server enough time to load the webpage, as the HKJC website would frequently timeout. A wait time of 10 seconds was allowed for each webpage, which meant that data scraping had to be performed efficiently and not blindly for all dates spanning across four years.

The data scraping approach consisted of first manually collecting the race dates from preseason announcements, which would be used in the first pass for the local racing results. The horse information was collected next from the URLs obtained from the first pass. As the horse information also included the horse's prior racing results, the dates from these form records were cross checked against the first pass with any discrepancies run again in a second pass of the local racing results. However, even with introducing a longer wait time, some webpages would return unknown errors through timeouts or issues connecting with the scrapy splash server, so the data scraping process involved manual intervention to validate and verify crawled results.

Finally, the overseas racing results were collected separately using the tabula library from python to parse the result tables from the PDFs into CSV format and required manual adjustments due to incorrect formatting. Investigations into the overseas racing results revealed only active horses were returned, suggesting full overseas records are not available from the website.

3. Exploratory data analysis

Exploratory data analysis was conducted to evaluate potential features to include in the modelling process. The selection of features was determined through evaluating correlations of the newly generated feature versus placing number, win and place (any horse finishing in the top three). Features with a correlation exceeding a magnitude of ~0.2 were considered to have predictive power and chosen for modelling.

Through the data scraping process, a total of 3,258 local races were collected since the start of the 2016/2017 season with a total of 55,162 participating racehorses. Additionally, each horse's full career was captured including all races prior to this season, therefore data for horses at different stages in their careers are fully available allowing for direct comparison.

Bolton and Chapman (1986) developed a multinomial logit model using the following specification:

$$\begin{aligned} U_h = {} & \theta_1 LIFE\%WIN_h + \theta_2 AVESPRAT_h + \theta_3 W/RACE_h + \theta_4 LSPEDRAT_h \\ & + \theta_5 JOCK\%WIN_h + \theta_6 JOCK\#WIN_h + \theta_7 JMISDATA_h + \theta_8 WEIGHT_h \\ & + \theta_9 POSTPOS_h + \theta_{10} NEWDIST_h + \epsilon_h \end{aligned} \quad (1)$$

Where the data attributes can be separated into three categories:

1. Horse related attributes including LIFE%WIN, AVESPRAT, W/RACE, LSPEDRAT and WEIGHT representing career win rate, average speed in recent races, winnings per race, track adjusted speed rating and handicap added weight respectively
2. Jockey characteristics including JOCK%WIN, JOCK#WIN and JMISDATA representing win percentage, number of wins and missing data respectively
3. Race related attributes including POSTPOS and NEWDIST representing post position (draw) and race distance indicator by horse previous races

The model listed above formed the initial basis for exploratory data analysis and in general most features were found to be useful displaying predictive power when compared to win and place.

When handling the horse racing data, a key factor that needed to be considered was data timeliness. As the races are a series of sequential events, all new features must be generated considering the specific time snapshot of the given race. This must be done to ensure the training data is valid and we are not peeking into future events which have yet to occur.

The horse's overall win rate and place rate prior to their next race was assessed, with the hypothesis that successful horses remain successful. Additionally, a further detailed grouping of these rates on specific courses also showed good correlation and was included for future modelling. Similar behaviour was observed for jockey win and place rates. However, in contrast to the horse data, jockey data prior to the 2016/2017 season was not fully captured so the start of this season was used as the cut-off point for all jockeys.

Each horse's recent form was assessed using a simple moving average (SMA) and an exponential weighted average (EWA) across various time windows. The EWA decay used the span formulation, which shows similar properties to an N-day SMA:

$$\alpha = \frac{2}{(span + 1)} \tag{2}$$

Overall, it was found that EWA outperformed SMA in all cases largely due to the null handling. This is because 50% of horses participate in 20 or fewer races across their careers, therefore choosing an SMA window of 5 allows data to only become available from the 6th race onward, whilst the EWA formulation allows the exponential factor to be associated to the data as it becomes available. The data attributes used to perceive horse form were prior placing numbers, length behind winner and win odds. Each of these data attributes showed strongest correlations for span of ~3 suggesting performance in the most recent 3 races best represented the horse's current form. Additionally, the moving average method was applied to jockey placings, and a higher optimal span of 10 was observed. Minor correlation improvements were further observed at higher spans, but at these higher spans the moving average tended to perfect inverse correlation with the jockey win rates, so a lower span was selected to balance against this.

Features related to a horse's average speed were calculated. Initial investigations calculated average speed by horse across different breakdowns such as distance and course type and were observed to have no significant impact when compared to win and place. The speed features were further transformed by calculating the difference between the horse's speed and the average speed across all horses participating in the same race. The SMA and EWA were applied to this speed difference across several different groupings and the EWA of the speed difference grouped by distance contained the most predictive power.

Draw or post position determines the horses starting position with lower draws being advantageous on courses with turns as these horses have a shorter distance to the inside lane. Figure 2 shows that there are significant differences between average win and place percentages for different draws. Features were generated to combine draw and course group by comparing differences between win percentages within these course groups, but it was found that draw values alone provided more information, so the original values were selected as modelling features.

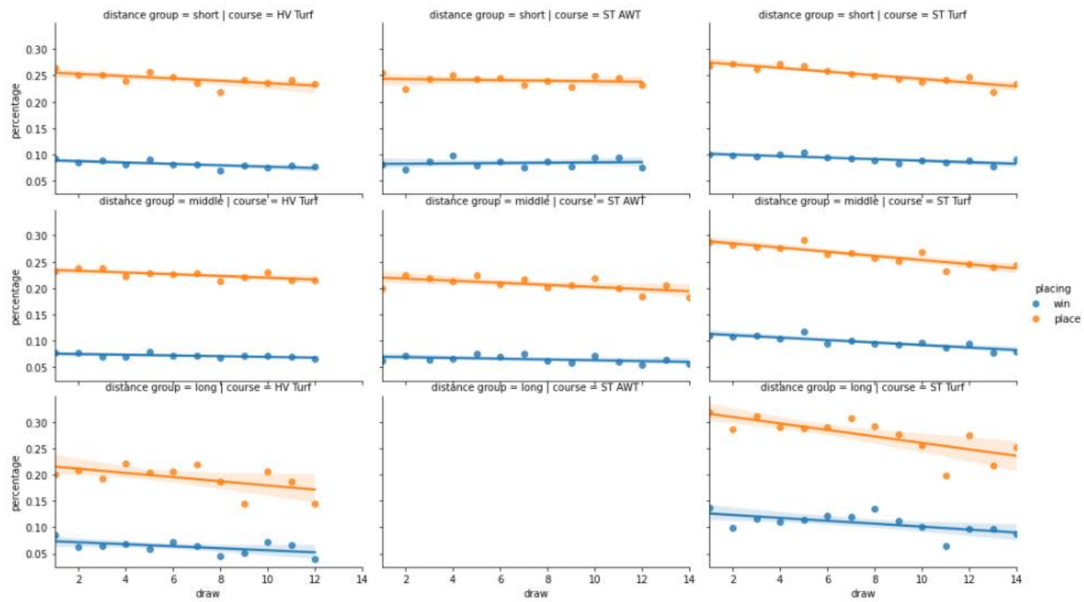


Figure 2 – Horse performance by draw by course type (columns) and distance groups (rows)

A horse's pedigree was evaluated by comparing win rates for horses with the same sire. However, it was observed that there are minimal differences in mean offspring performance by sire, especially for sire's selected for heavy breeding. As no clear differences are observed, pedigree was not considered an important feature to include for modelling. In addition to pedigree, horse age was investigated, but insufficient data was found as birth year was not recorded with only the current age of active horses being displayed on the website. No direct relationship could be identified between registration year and birth year. Therefore, horse age was also unable to be added to the model features.

The final series of features compared race to race differences including changes in ratings, race class, added weight and date of last race. The first three data attributes listed are related to the handicapping system, however only changes in rating impacted predictive power. Contrary to the feature suggested by Bolton and Chapman (1986), changes in weight were observed to not provide any significant information.

The final set of features selected for modelling are shown in Figure 3 and it can be observed that these features show good correlations across the win and place target variables. It can also be seen that aside from draw most features have some correlation to others, most obviously across the various win and place rates. Additionally, other features are natural inverses to one another such as relative speed difference and length behind winner.

All data features are required to be shifted to the next race so that results of target race are not exposed in the features. As a result, null values are generated for the first race in each shift grouping, which will be handled by filling with 0 values for rate and difference columns, while the moving average columns will be filled with the mean value across their grouping. Assessment of feature selection and importance will be determined during the modelling process.

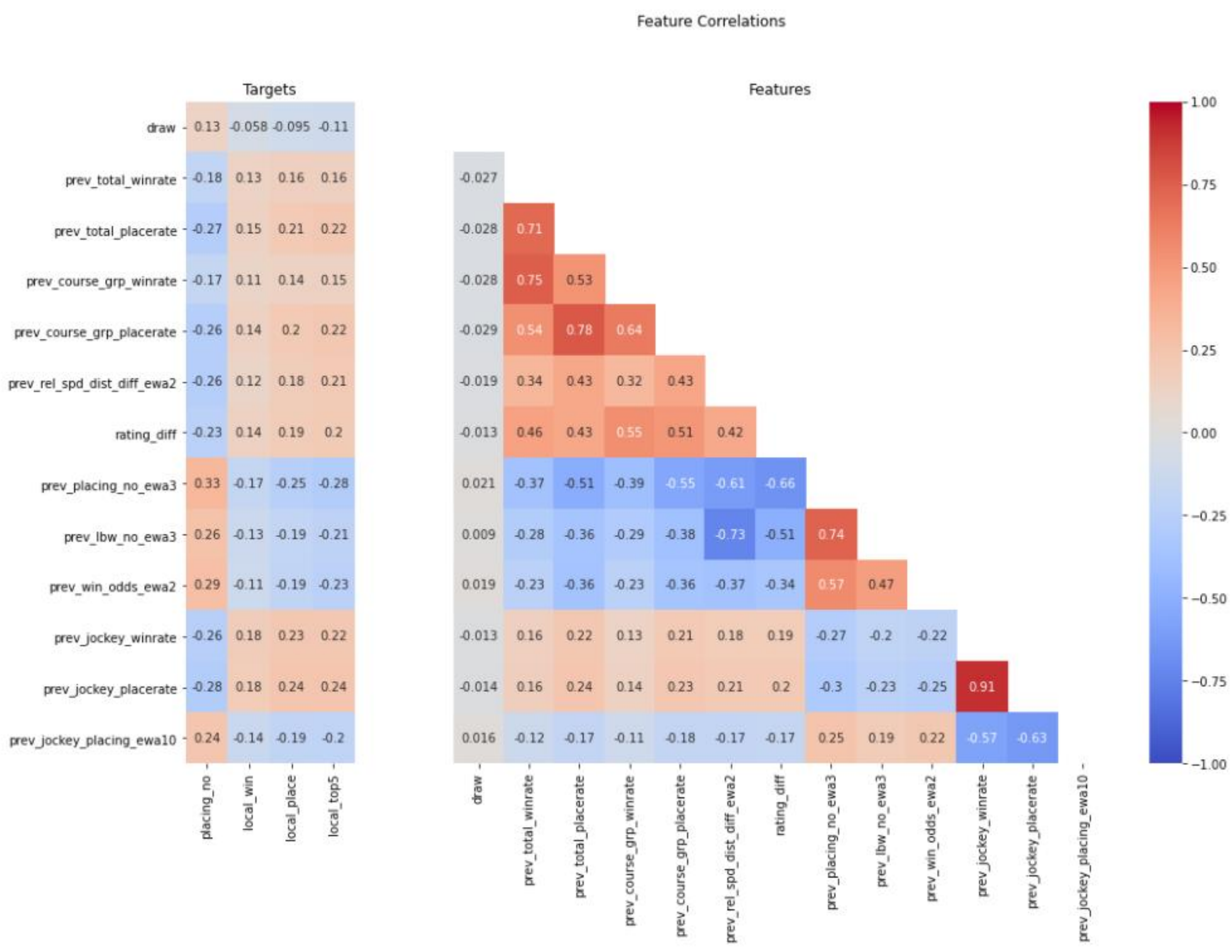


Figure 3 – Correlation plot of features selected for modelling

4. Predictive modelling

A class data attribute was generated to group race finishing placings into four distinct classes, which are given ranking properties and ordered as follows:

$$Win > Place > Top\ 5 > Other$$

The classes were split using top 5 as the cut-off point as top 5 placings across all handicap races hosted by HKJC (Hong Kong Jockey Club, 2019) finish in the purse. Therefore, applying a similar logic that Bolton and Chapman (1986) used to choose an explosion depth of 3 in their multinomial logit model.

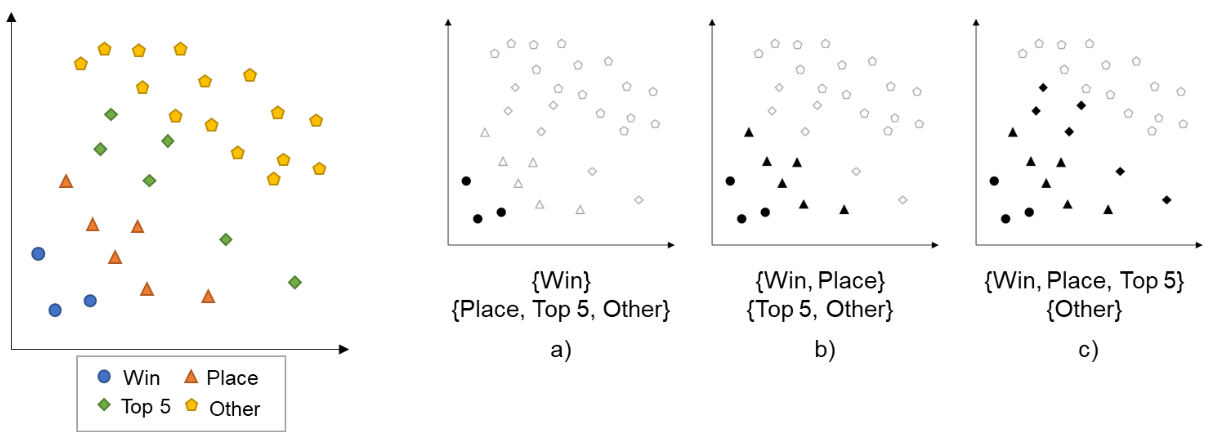


Figure 4 – $\mathbb{R}^2$ representation of class data attribute split into three binary classifiers a) win vs. rest b) place (with win) vs. rest c) top 5 (with win and place) vs. rest

Figure 4 illustrates the three binary classifiers used to model the four ranked ordinal classes applying the methodology as described by Frank and Hall (2001). The resulting probabilities $\{p_{win}, p_{place}, p_{top5}\}$ generated from the binary classifiers will interact and generate a unique probability for each of the four classes $\{C_{win} \ldots C_{other}\}$, using the following formulation as mentioned by Frank and Hall (2001):

$$\begin{aligned} \Pr(C_{win}) &= p_{win} \\ \Pr(C_{place}) &= p_{place} - p_{win} \\ \Pr(C_{top5}) &= p_{top5} - p_{place} \\ \Pr(C_{other}) &= 1 - p_{top5} \end{aligned} \tag{3}$$

Cardoso and Costa (2007) explains one of the shortcomings of the above formulation is that negative probability estimates may be generated and suggests the following alternative:

$$\begin{aligned} \Pr(C_{win}) &= p_{win} \\ \Pr(C_{place}) &= (1 - p_{win}) \cdot p_{place} \\ \Pr(C_{top5}) &= (1 - p_{place}) \cdot p_{top5} \end{aligned} \tag{4}$$

$$\Pr(C_{other}) = 1 - p_{top5}$$

Prior to building the models data was split into training and testing datasets by season with the 2019/2020 season held out for testing and the remaining data from 2016/2017 to 2018/2019 used for training. The two evaluation metrics used to determine model performance are balanced accuracy, as classes are heavily skewed 1:12, 1:4, 1:2 for win, place and top 5 classes respectively, and precision as from the perspective of a bettor we would want to optimise our return from our predictions. Figure 5 shows the preliminary results for the win classifier with most models returning a balanced accuracy of ~65%, which shows the selected features are providing information to the model. However, precision is poor averaging ~16% across the five models. Both the place and top 5 classifiers produced similar balanced accuracies with precisions rising to an average of ~36% and ~55% respectively as expected due to the different class ratios mentioned above.

Across all three binary classifiers, Random Forest, Logistic Regression and XGBoost Classifiers all consistently performed well with Random Forest performing marginally better on both balanced accuracy and precision. Linear support vector classifiers often showed results with higher precision and lower balanced accuracy, but after hyperparameter tuning they were also observed to show similar performance when optimising for balanced accuracy. Some experiments to optimise models on precision were run, but it was found these models dropped balanced accuracy to 50% and were generating higher precisions from random guessing.

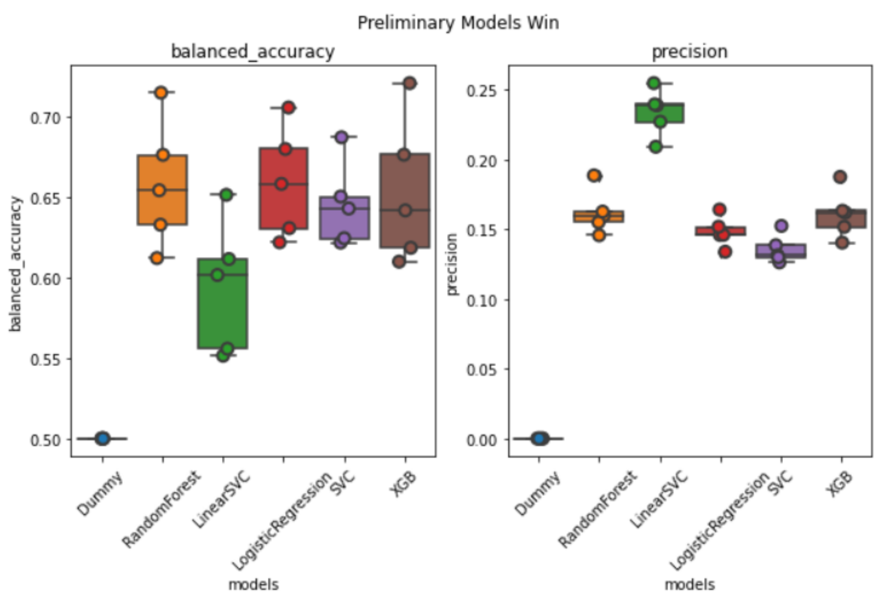


Figure 5 – Preliminary model results for win classifier against Random Forest, Linear Support Vector Classifier, Logistic Regression, Gaussian SVM, XGBoost Classifier using 5-fold cross validation

Feature importance was investigated by evaluating the feature coefficients of a Lasso classifier run against the three binary classes. Features related to win rates had zero coefficient suggesting low importance, however dropping these features did not have any observable impact on the preliminary models. As Random Forest slightly outperformed other models across all three binary classes, the features were not dropped so that there is a sufficient pool of features for the Random Forest classifier to subset from. Grid search algorithms were performed on all three binary classes varying maximum depth and maximum number of features in subset using 5-fold cross validation. This hyperparameter tuning only had minor impact to the overall model performance and the best estimator from each grid search algorithm was selected for each binary class prediction.

The final tuned Random Forest models were retrained on a subset of the training data, with 25% randomly split for validation. Figure 6 shows the feature importance on the validation set, with all binary classes having the same top 5 most important features two related to jockey performance and the other three generated from the overseas and local races moving averages.

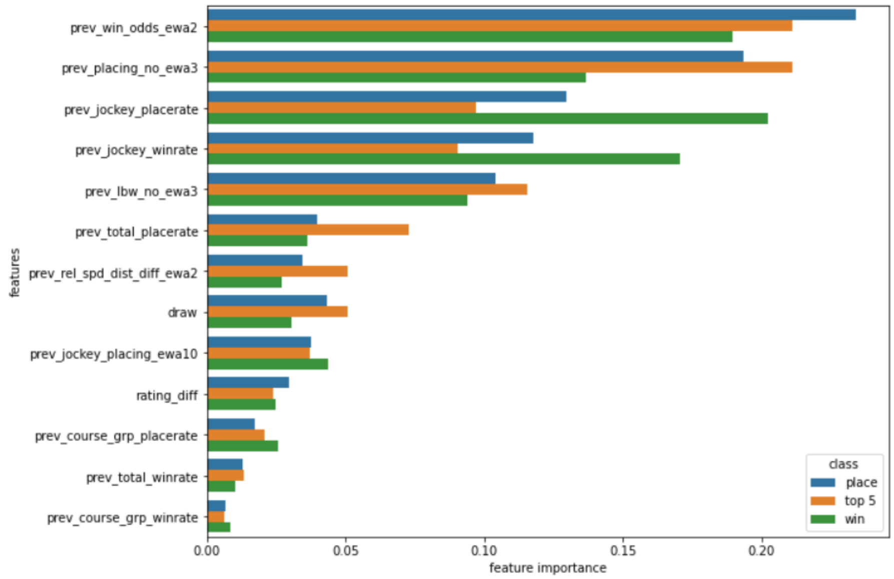


Figure 6 – Random Forest feature importance across three binary classes on validation set

Combining the results across the three Random Forest classifiers, a single ranked ordinal classifier is generated by applying the probability functions in equations 3 and 4 for Frank and Hall (2001) and Cardoso and Costa (2007) respectively. From Figure 7, both matrices are seen to be identical and only predict two classes win and other. Transforming the combined classifier to use the maximum probability generated by the three classifiers did not show any improvement to model performance and instead the expected negative performance associated with higher multiclass problems was observed.

Further investigation into the probabilities from the binary classifiers reveals extreme similarities between the generated probabilities across all three classifiers. These similarities cause the middle classes, namely place and top 5, to be suppressed due to the formulation in equations 3 and 4 forcing resulting probabilities of these classes to be near zero. This suggests the models are unable to distinguish accurately distinguish between finishers placing in the top 5. With this result, it is inconclusive whether the ranked ordinal classification suggested by Frank and Hall (2001) can be applied in the context of horse racing.

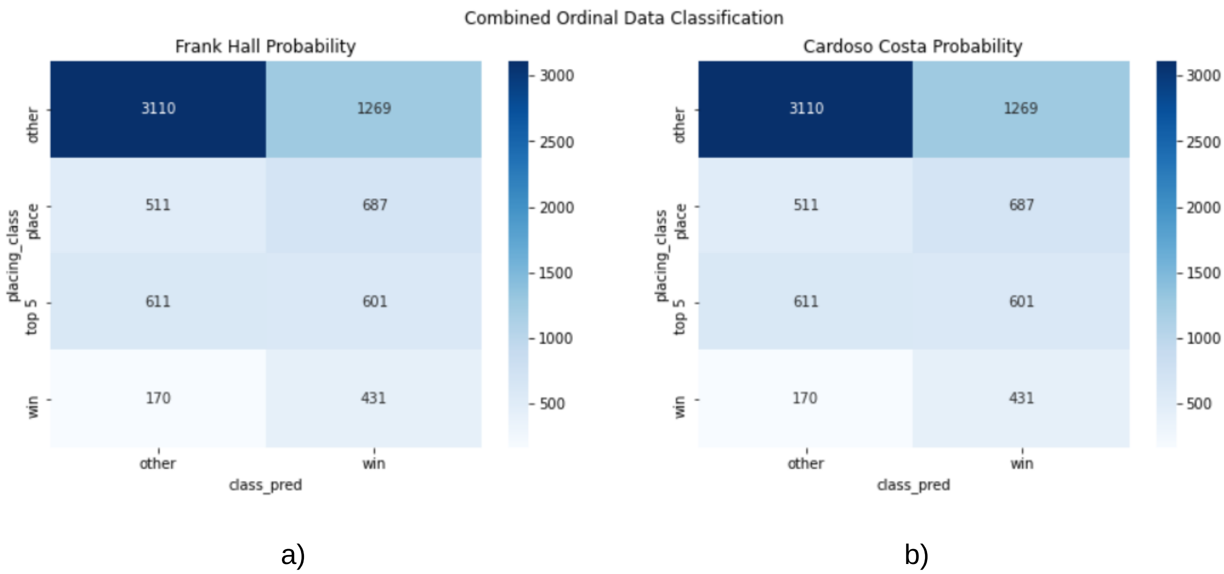
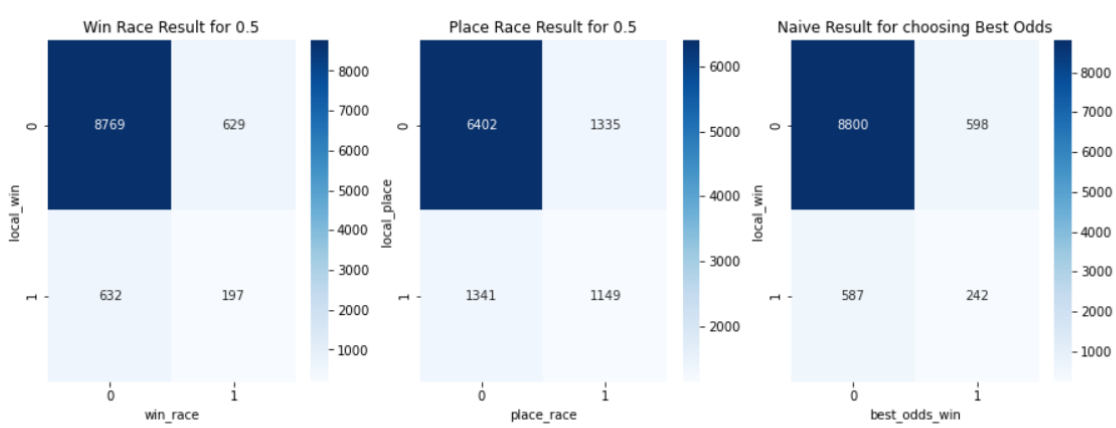


Figure 7 – Confusion matrices of ranked ordinal classifier generated using a) Frank and Hall (2001), b) Cardoso and Costa (2007) probability functions

Random Forest classifier models for the three binary classes were retrained on the full training dataset and evaluated against the hold out testing dataset. Figure 8, shows the final model results on a per race basis where predictions were assigned using largest probabilities. The final win prediction achieves a precision of 23.85% suggesting approximately one in four of our wagers would pay-out. When compared to a random guess, the prediction would perform 3 times better (assuming 12 horses participate in a race). However, a better naïve estimate would be a comparison against the race favourite, which achieves a higher precision of 28.81%.

Further evaluation may show the prediction would provide a better monetary return than choosing the favourite. However, as the objective of this report was to model the horse racing process, disregarding the wagering strategy, the final model is deemed to be unsuccessful.



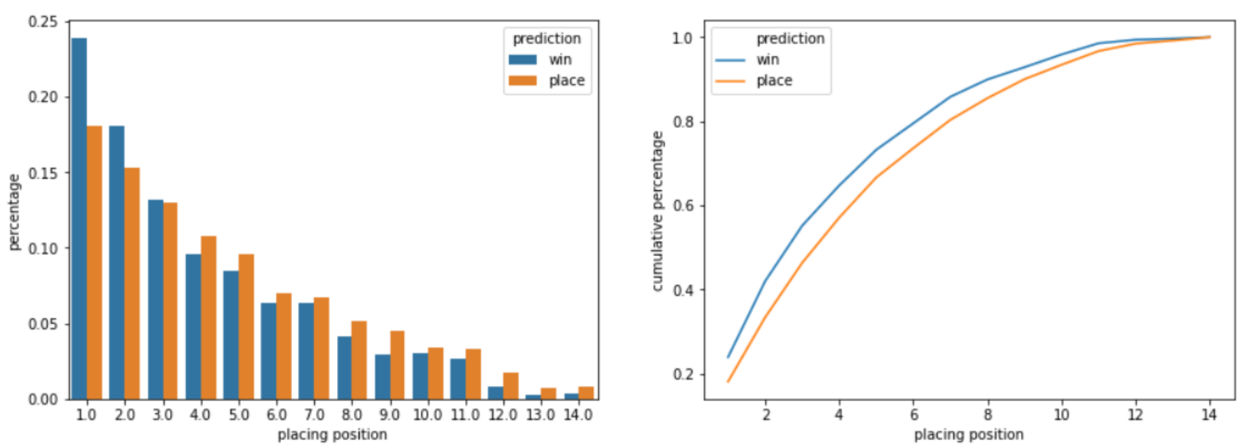
	Win prediction	Place prediction	Naïve prediction
Balanced Accuracy	58.54%	64.44%	61.41%
Precision	23.85%	46.26%	28.81%

Figure 8 – Confusion matrices of results on testing data per race a) win classifier, b) place classifier, c) naïve classifier assigning winner to odds favourite

Figure 9 shows that both win and place binary predictions display an exponential decay relationship across placing number, proving the model can identify horses with a higher likelihood of success. Looking at the win classifier, it can be seen ~40% of predictions finish in the top two places and shows the inherent difficulty required by the model to differentiate these horses.

Additionally, the win prediction overall performs better than the place prediction with a higher percentage of horses finishing in the top three places. This may suggest a place classifier does not provide any additional value, but instead the top three probabilities generated by the win model should be used. It is hypothesised that there is significant unexplained variance leading to noise in the model, hence using a target variable only containing the winners surfaces the top three finishers as they are perceived as having a similar likelihood to win. Whereas, using place as the target variable induces extra noise from the lower placing horses who may have a perceived similar likelihood to finish in third place.

Given the similar distributions observed in Figure 9, it is believed horse racing cannot be treated as a rank ordinal set, and hence the rank ordinal classifier specification defined by Frank and Hall (2001) cannot be applied in this context. This does not contradict with the finding by Bolton and Chapman (1986) as the ordering was used to exploit existing race data to generate “new races” so that intra-race comparisons could be better assessed. Due to the nature of horse racing there cannot exist a dataset which can explain all variance, such that each horse can be separated into ranked classes to the required level of granularity of identifying a single winner. Therefore, a ranked ordered set of classes with well defined boundaries can not exist in the context of horse racing.



a) b)

Figure 9 – Binary classification results on testing dataset for win and place predictions against actual placing number a) by percentage of prediction, b) cumulative percentage

5. Conclusion and recommendation

This report set out to predict horse racing results using data collected from Hong Kong Jockey Club, with the primary focus of correctly identifying winning horses as opposed to maximising potential profits through optimising a wagering system. Through exploratory data analysis, the generated features were found to be similar to those in the multinomial logit model specification defined by Bolton and Chapman (1986) in equation 1. Given this consistency over a thirty-year difference in research, incorporating more data may potentially enhance trends and patterns eventually boosting overall model performance.

Three separate binary classifiers were developed each predicting a class in the ordered set as illustrated in Figure 4. The results were combined using both the Frank and Hall (2001) and Cardoso and Costa (2007) formulations to generate a ranked ordinal classifier. The resulting predictions only returned the first and last values in the ranked set, win and other respectively, largely due to the similar predictions generated by the individual binary classifiers. It is hypothesised that horse racing data cannot be treated as a rank ordered set due to inherent unexplained variance leading to an impossibility of assigning predictions into ranked classes with well defined boundaries.

In conclusion, the final Random Forest model predicting winning horses produces a precision score of 23.85%, which is a significant improvement over a random guess, but fails to outperform the naïve estimate of choosing the race favourite. As a result, the original objective to accurately predict horse racing results was not achieved.

6. Limitations and future research

Throughout this report, several limitations were encountered primarily due to data availability. It was observed that HKJC removes data on horses once they have retired, this includes but is not limited to horse age, overseas races and trackwork records. As with all statistical analyses, gaps in data skews potential patterns and trends, which are picked up by the models. Missing data on prior overseas races is particularly damaging to model performance as most of the identified features relate to in-race horse performance. A potential mitigation to this issue would be to identify other data sources or providers who have been collecting and archiving this data over a longer period.

Further research into the topic area would require revisiting the exploratory data analysis process to further assess and identify potential features to include for modelling. Additional data collection via data scraping or through data providers to generate a larger sample size would be advantageous to further support the exploratory data analysis. With the success of Betner (1994) on the same horse racing dataset a potential strategy would be to replicate the specified model as closely as possible. This would first allow us to identify if the model can still be successful within a modern environment, and second provide a good set of once successful features to investigate.

The analysis conducted in this report used binary classification to predict winners from all horses, but alternative approaches can be investigated such as developing a ranking system i.e. ELO rating or predicting finish times. The current analysis was unable to model the inter-race competition aspect of horse racing, through either modelling or via the features. An attempt was made by implementing a ranked ordinal classifier, which was found to be unsuccessful. Further investigation would be required to identify if there are any other available methodologies to model and choose a winner from a choice set.

Many modern applications of modelling the horse racing process attempts to use neural networks (Williams & Li, 2008) and (Davoodi & Khanteymoori, 2010) as the machine learning algorithm due to their ability to pick up complex non-linear relationships. A brief investigation was conducted as an extension during the modelling process using a 3-layer architecture with 32, 16 and 1 nodes. The initial results were in line with the preliminary models in terms of both balanced accuracy and precision when using the same feature set. This investigation proves the feasibility of using a neural network on the horse racing dataset, but as with the other models it is expected new findings in feature development would provide more significant performance gains over finetuning this model with the existing feature set.

Finally, to operationalise the horse racing prediction into a wagering system a wagering strategy such as the Kelly criterion (Kelly, 1956) would need to be investigated and implemented. Variations of the Kelly criterion have been used, notably by Betner (1994) who implemented a fractional Kelly betting strategy.

7. References

Ali, M. (1977) *Probability and Utility Estimates for Racetrack Betting*. Journal of Political Economy, 85, 803-815.

Benter, W. (1994) *Computer Based Horse Race Handicapping and Wagering Systems: A Report.* in Hausch, D.B., Lo, V.S.Y., and Ziemba, W.T. ed. (1994) Efficiency of Racetrack Betting Markets, Academic Press, p.183-198.

Bolton, R.N. & Chapman, R.G. (1986) *Searching for Positive Returns at the Track, A Multinomial Logit Model for Handicapping Horse Races*. Management Science, 32, 1040-1059.

Cardoso J.S & Costa J.F.P (2007) *Learning to Classify Ordinal Data: The Data Reciprocal Method.* Journal of Machine Learning Research, 8, 1393-1429

Chapman, R.G. & Staelin R. (1982) *Exploiting Rank Ordered Set Data within the Stochastic Utility Model.* Journal of Marketing Research, 19 (3), 288-301

Chellel K., 2018. *The Gambler Who Cracked the Horse-Racing Code.* Available from: https://www.bloomberg.com/news/features/2018-05-03/the-gambler-who-cracked-the-horse-racing-code [Accessed 29 August 2020]

Davoodi, E. & Khanteymoori, A.R. (2010). Horse racing prediction using artificial neural networks.

Frank, E. & Hall, M. (2001) *A simple approach to ordinal classification*. In Proceedings of the 12th European Conference on Machine Learning, volume 1, 145–156

Hausch, D.B., Lo, V.S.Y., & Ziemba, W.T. ed. (1994) *Efficiency of Racetrack Betting Markets*, Academic Press, 237-247

Hong Kong Jockey Club. (n.d.) Available from: https://racing.hkjc.com/racing/information/English/racing/LocalResults.aspx [Accessed 29 August 2020]

Hong Kong Jockey Club. (2019) Available from: https://racing.hkjc.com/racing/english/racing-info/rules_pdf/2019-20/r162.pdf [Accessed 29 August 2020]

Kelly, J. (1956), *A New Interpretation of Information Rate.* Bell System Technical Journal. 35, 917-926.

Metel, R.M. (2017) Kelly betting on horse races with uncertainty in probability estimates. [Preprint] Available from: https://arxiv.org/abs/1701.02814v2. [Accessed: 29 August 2020].

Williams J. & Li Y. (2008) *A case study using neural networks algorithms: horse racing predictions in Jamaica*. In: ICAI 2008: International Conference on Artificial Intelligence, 14-17 Jul 2008, Las Vegas, NV. United States.