Capstone Project

Odds Fluctuations in Australian Horse Racing to predict a winner

An Analytical Report by Billy Bingham



# Report Contains:

* **Summary**
* **Hypothesis**
* **Risks & Assumptions**
* **Findings**
* **Modelling**
* **Conclusion**

**Summary**

Using sportsbet.com.au, information relating to 820 individual Australian horse races has been collected. Statistics collected include; race winner, location of race (town/suburb), location of race (State), race number, opening price (odds when the race was initially put online – approximately 3 days prior to race start), starting price (odds when the race starts), percentage of odds fluctuation between opening price and starting price, whether the winning horse was the biggest or second biggest fluctuation in the race, whether the winning horse was the favourite at the start of the race, jockey name, trainer name, and ground type. A full Data Dictionary will be included in this report.

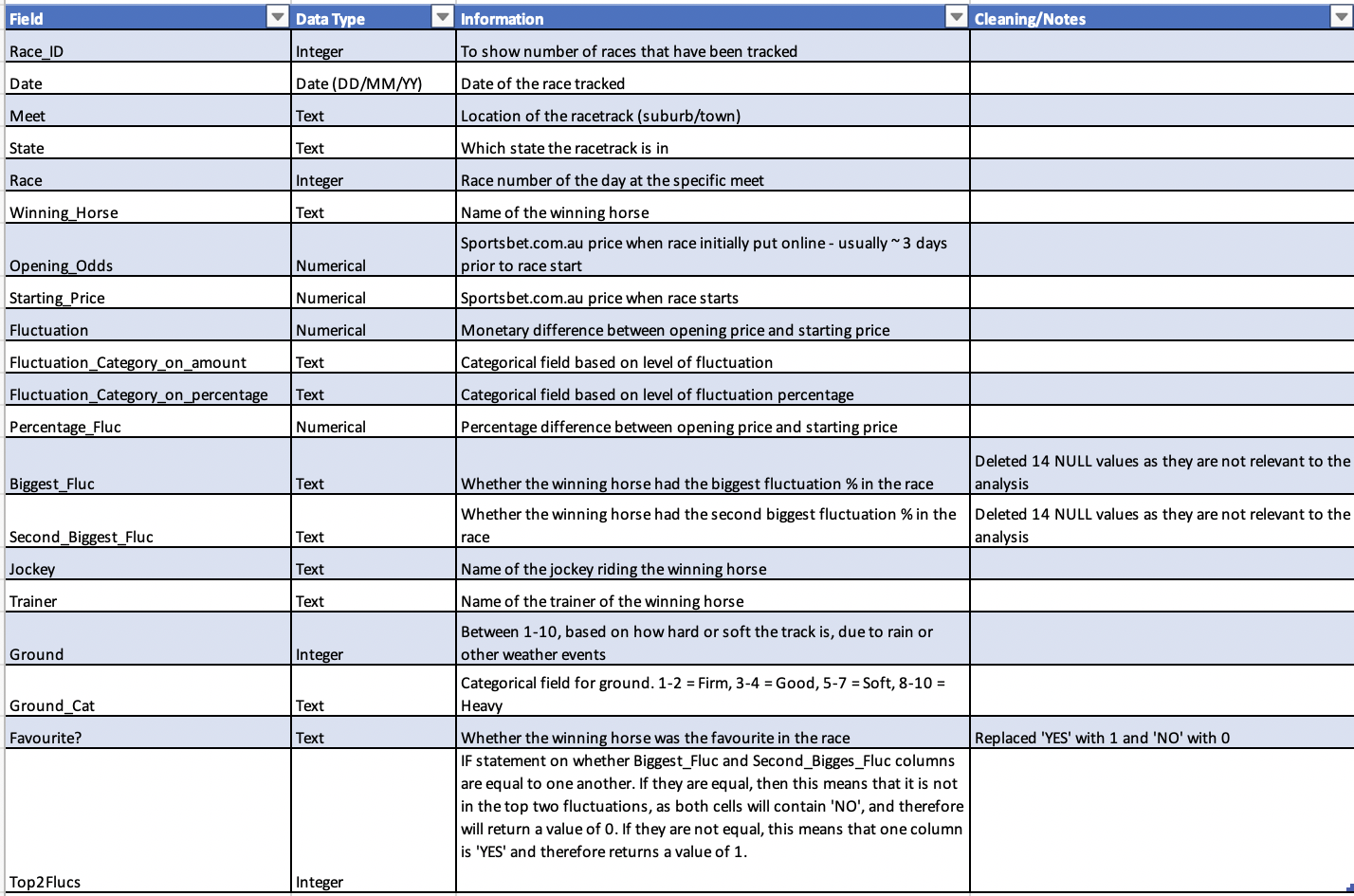
14 of the 820 races had no information post-race relating to the opening price. These observations have been deleted from the dataset as they would not be useful for the analysis as it was not possible to tell whether or not they would fall within the biggest fluctuations or not.

In total, 806 individual races had full data available, and thus is the final dataset that has been used for this analysis.

This analysis will look at the viability of looking at the odds fluctuations of winning horses to be able to predict future winners, and whether or not this is more or less likely at a specific location. This has been done using the data above and analysing it in Tableau, as well as using a logistical regression model in python.

Of the three success metrics chosen for the hypothesis, one of these three was met – to have an overall average odds fluctuation of 10%+ across all races (**10.85%**). A hit rate of 40%+ winning horses being in the top 2 fluctuations of its race was missed (**36.5%**), whilst a total of 70%+ of winning horse’s odds reducing between opening price and starting price was also missed (**66.63%**). Digging deeper into separate locations did find some results where all three success metrics were met.

The logistic model that has been created, along with cross-validation against the dataset, could correctly identify with 80% (+/- 5%) accuracy, whether the horse was likely to be in the top 2 biggest fluctuations of its race.

**Data Dictionary**

**Hypothesis**

The main hypothesis of this analysis is that the more the odds reduce between opening price and starting price, the more likely that it will be in the top two biggest fluctuations of the race, and thus have a high chance of winning the race.

With regards to success metrics used to measure this hypothesis, the following statistics will show whether this is a valid hypothesis:

* **A hit rate of 40%+ winning horses having the biggest or second biggest fluctuation in the race.**

The average number of horses per race in Australia is 12, if we assume all runners are equal, this gives each horse an 8.33% chance of winning the race – or a 16.66% chance of finishing in the top 2. As the metric for this analysis is whether the horse is in the top 2 biggest fluctuations, to create consistency we will use the likeliness of the horse finishing in the top 2 positions. A hit rate of 40%+ in this analysis would provide almost 2.5x the number of winners using this method, than the 16.66% likelihood of any non-specific horse finishing in the top two.

* **An overall average odds fluctuation of 10%+**

Across the 804 races that data has been collected on, an average odds fluctuation of over 10% would show a significant pattern of winning horses odds reducing between their opening price and starting price.

* **A total of 70% of winning horse’s odds reducing between opening price and starting price.**

This would show a similar significant pattern as the average odds fluctuation being over 10%.

**Risks & Assumptions**

There are a few factors to consider for this analysis on Australian horse racing.

1. **Percentage of odds fluctuation can have some limitations that can skew the statistics.**

When the odds increase, the percentage of the increase would be higher than if the odds decreased at the same numerical value. An example of this is as follows: If a horse has an opening price of 10 and the starting price reduces to 5, the odds have halved, however have only decreased 50%. If this same horse went from an opening price of 10 to an opening price of 20 and doubled, this would be an increase of 100%. Also, it is possible for a horse to increase in odds by over 100%, whilst it is mathematically impossible for this to happen when the odds reduce, as the price cannot go below 0.

1. **Horses at much bigger opening prices tend to fluctuate more, whilst still being very unlikely to win the race.**

An example of this is a horse’s opening price starting at 150/1, and then reducing to 50/1 as the race starts. This would be an odds reduction of 66% and make it likely to be in the top 2 biggest fluctuations, despite having relatively little chance of winning (2%). This is partially the reason that the top 2 biggest fluctuations in a specific race was used as a metric, as opposed to the outright biggest in a race.

1. **The odds tend to reduce more on heavy ground.**

On heavy ground (after a lot of rain), many horses get scratched from the race, meaning there are less horses running than what was initially planned. As there are now less runners, the likeliness of any non-specific horse winning the race increases, and thus the odds are reduced to reflect this. For this reason, the ground type has been included for each race, to see the difference that this can make between the level of odds fluctuation.

1. **Only the winners have been tracked**

The model has been used to predict whether a horse is likely to be in the Top 2 Fluctuations of it’s race – should this analysis be repeated, it would be beneficial to include all runners in the race, along with a column stating whether it was the winner or not. The model could then be used to try to predict the winner.

**Analytical Findings**

Over the course of this analysis, multiple charts, graphs and tables have been created to see if there is any relationship between a horse having a high level of odds fluctuation and winning a race, as well as modelling created in python.

The chart below shows the total percentage of races where the winner was in the Top 2 biggest fluctuations of the race.



The total number of winners that were in the Top 2 Biggest fluctuations of their race was 294/806 – or **36.5%.** This result falls just short of meeting the hypothesis. However, we can drill down further into the analysis using filters on specific locations, to see if any locations are more likely to support this hypothesis than others.

Out of the 49 different meets that had been tracked, 25 of them had over 40% of races with winners that fell into the Top 2 biggest fluctuations. Therefore, whilst the overall hypothesis was not quite reached based on the success metric of 40%+, **over 51% of the locations tracked did support the hypothesis**.

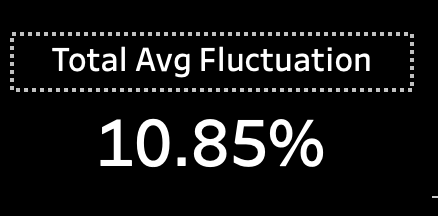
Of these locations, some were far above this 40% threshold – below are the four best performing locations that support the hypothesis.

|  |  |  |  |
| --- | --- | --- | --- |
| Meet Location | State | % of winners in Top 2 Biggest fluctuations | Number of Races Tracked |
| Darwin | Northern Territory | 80% | 6 |
| wagga | New South Wales | 87.5% | 8 |
| NARROGIN | Western Australia | 71.4% | 7 |
| Ipswich | Queensland | 64% | 17 |

There were also some locations significantly below the 40% benchmark – below are the bottom four locations:

|  |  |  |  |
| --- | --- | --- | --- |
| Meet  Location | State | % of winners in Top 2 Biggest fluctuations | Number of Races Tracked |
| hamilton | Victoria | 12.5% | 8 |
| Gosford | New South Wales | 14.3% | 14 |
| Murray Bridge | South Australia | 21.4% | 14 |
| Randwick | New South Wales | 21.3% | 75 |

Another success metric to test the hypothesis, is the overall average fluctuation of winners being over 10%. This is important for the analysis as it can show a significant pattern of winning horses odds reducing between their opening price and starting price.

The below right graphic shows the average overall fluctuation:

The overall average fluctuation was measured to be at **10.85%**, resulting in the success metric being met.

Again, similar to the previous hypothesis, some locations supported this hypothesis better than others, whilst different ground types also performed differently to one another.

On the right, we can see that the ‘heavy’ ground type has a larger average odds fluctuation than that of ‘soft’ and ‘good’. This had been previously mentioned in the risks/assumptions page and could be due to the number of runners being reduced after a large volume of rain at the racetrack. Regardless, all three ground types had an average fluctuation of above 10%.

Of the 806 races tracked, only 21 races were on ‘heavy’ ground, 183 races were on ‘soft’ ground, whilst the remaining 602 races were on ‘good’ ground.

With regards to locations hitting this metric – **26 out of the 49 recorded meets had an overall average of above 10%, whilst 23 meets fell under this benchmark**.

Below are two tables showing the best three locations/meets that support this hypothesis, as well as the worst three locations:

**Top 3:**

|  |  |  |  |
| --- | --- | --- | --- |
| Meet  Location | State | Average % Fluctuation | Number of Races Tracked |
| Kembla Grange | New South Wales | 32.4% | 14 |
| Kyneton | Victoria | 31.5% | 10 |
| wyong | New South Wales | 31.9% | 7 |

**Bottom 3:**

|  |  |  |  |
| --- | --- | --- | --- |
| Meet  Location | State | Average % Fluctuation | Number of Races Tracked |
| Gosford | New South Wales | 17% | 14 |
| Cranbourne | Victoria | 15.1% | 8 |
| Alice Springs | Northern Territory | 11.5% | 3 |

The final success metric for the hypothesis is having a total of 70% of winning horse’s odds reducing between opening price and starting price. Similarly, to the metric above, this is important for the analysis as it can show a significant pattern of winning horses odds reducing between their opening price and starting price.

The chart on the right shows the percentage of races that fell into a specific fluctuation category. The categories were as follows:

* **None/Negative** – 0% or negative fluctuation
* **Minimal** – Between 1% and 19% fluctuation
* **Decent** – Between 20% and 39% fluctuation
* **Big** – Between 40% and 59% fluctuation
* **Huge** – 60% or greater fluctuation

33.37% of the winners had either no fluctuation, or a negative fluctuation, meaning that the remaining **66.63% of the winners had some form of odds reduction between the opening price and the starting price**. Whilst this falls just short of the benchmark, it is still shows a very strong relationship between a horse winning and the odds reducing between opening price and starting price.

One final finding in the analysis that is interesting and important to note, is that **the favourite horse in a race only won 33.13% of the time** – 267 races out of a possible 806 were won by the favourite.

An interactive dashboard was created (see below), so that any additional data collection can be analysed to increase the sample size and accuracy:



**Modelling**

The model used to predict whether a horse will be in the Top 2 fluctuations in the race was a Logistical Regression model, built in Python. The variables included in the prediction model were; ‘Opening Price’, ‘Starting Price’, ‘Percentage Fluc’, ‘Ground’ and ‘Favourite’.

Using train and test sets, the accuracy of the model correctly predicting whether the horse was in the top 2 fluctuation was **80% (+/- 5%).** This used a test size of 0.2.

As there was no secondary dataset to test this model on new data, cross-validation was used with 10 folds. This means the data was randomly split into ten groups, with one group being the train set, and the remaining 9 being the test sets. Each of the ten groups is be used as a train set to test against the test sets, and then calculating the mean accuracy between all iterations of the model. The mean accuracy for this model was **81% (+/- 2%**).

**Base Logistic Regression Model:**

Accuracy: 0.8209876543209876

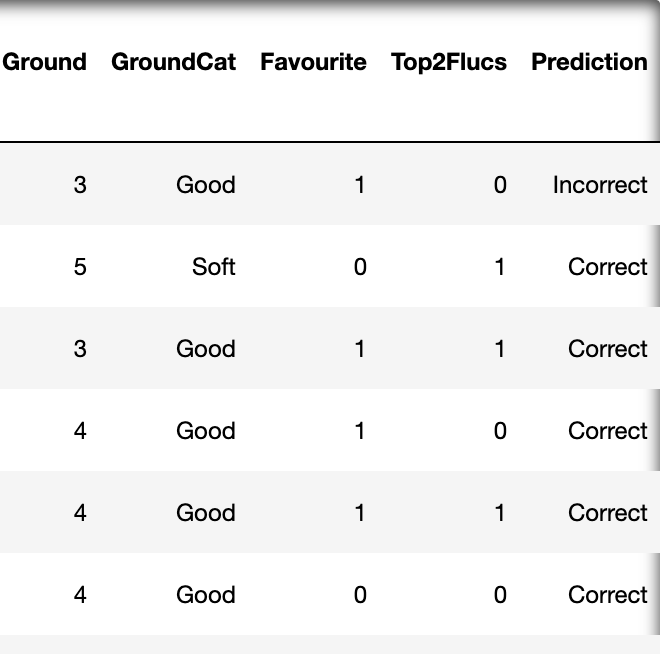
Confusion matrix:

[[97 15]

[14 36]]

**Cross-Validation Logistic Regression:**

Mean accuracy score: 0.8214903846153845

A ‘Prediction’ column was added to the dataset, in order to see which observations were predicted correctly or incorrectly – sample of this below:

**Conclusion**

Whilst the hypothesis did not come to fruition with only one of the three success metrics being met, there is still a reasonable relationship between odds fluctuations between the opening price and starting price.

This can be seen by the fact that 25 of the 49 meets/locations tracked had over 40% of the winners being in the top 2 fluctuations in the race, and 26 out of 49 meets/locations having an overall average fluctuation of over 10%.

The cross-validated predictive model also had a good accuracy level of 82% (+/- 2%), meaning that the model correctly predicted whether or not the horse would be in the top 2 fluctuations in its race.

Should this analysis be repeated, it would be wise to include all horses in a race, as this model could then be used to try to predict a winner.