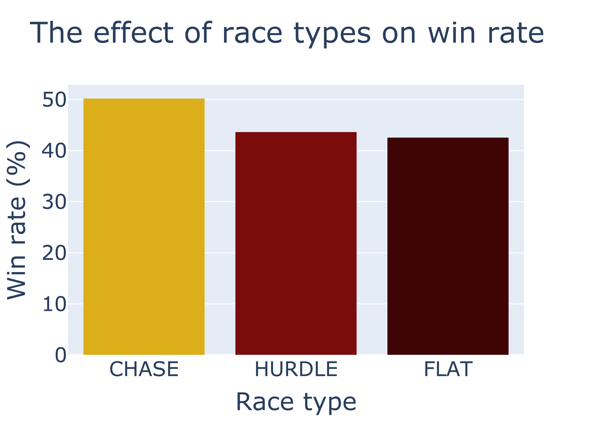
Report on Horse Betting Strategy

# An initial look into the data

The dataset provided contains information on 3208 bets placed on horse races by Mustard Systems. None of the 8 features nor the target column contain any missing values, however the results column does contain some ‘NON-RUNNER’ entries resulting in the bet being cancelled refunded. These will be removed before creating the model as they provide no useful information on the outcome of race and do not affect the profits.

# Exploratory Data Analysis



Chart, bar chart

Description automatically generatedWin rate is defined as the percentage of bets placed with a positive profit returned. Although a bet placed on a chase race tends to win more often (50.2% of bets), Figure 1 shows there the type of race the bet is place on does not affect the win rate to a great extent. One possible explanation for the chase race to have the highest win rate is due to the fact that it is generally considered to be the most technical race out of the three, and so there is less likely to be an ‘upset’. This feature may be removed when creating the model as it may generate unnecessary noise during training leading to underfitting.

Whilst one would assume that the greater the bet strength, the more likely a bet is to be successful, Figure 2 demonstrates that this is not the case. A bet strength of 4 is the most likely to win, at a calculated average of 48.7%.

Chart, bar chart

Description automatically generated

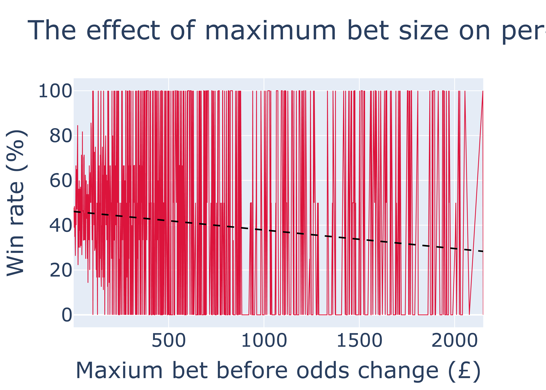
There is no surface condition (going) which provides a substantially better win rate over the others, however a heavy graded surface has by far the lowest win rate at 35.5%. This may be because very wet conditions lead to an increase in unpredictability of the winner, and therefore to optimise the betting strategy, Mustard Games should avoid bets on heavy graded surfaces. Bets placed on good, firm, good/firm and fast surfaces should be the focus, with all resulting in a win rate of above 45%.

Figure x shows the effect the number of runners in each race has on the win rate, which has been fitted with a polynomial line of best fit of order 2. From this plot it is clear to see that Chart, line chart

Description automatically generatedthe value with the highest win rate is 14 runners at a value of 51.8%. Races with 12-15 runners inclusive should be the focus of Mustard Games as they all have win rates of more than 50%.

Chart, line chart

Description automatically generatedSeveral pieces of information can be understood from Figure x. To start with, the least squares trendline demonstrates a negative coefficient meaning that in general, the more days since a horse last raced the less likely a bet is to be successful. Whilst the variance in win rate is large for all values, it becomes more pronounced after roughly 65 days since the last race. As the coefficient of the trendline gradient is of a small magnitude, this feature is not too important when considering what bets Mustard Games should place in the future, but a good strategy would be to avoid races where the horse has not run for more than 65 games.



Similar to the days since a horses last race, the maximum amount that can be staked before the odds change shows a similar negative gradient. Figure x shows the least squares trend line predicting a win rate of 46.1% at a maximum bet of £2, and all the way to 28.3% at £2150. As a result of this, bets with a smaller maximum bet should be taken as they will be successful more often.

# Most profitable betting opportunities

From the plots in Section 2, it seems that the most likely bets to be successful are those with 14 runners in a chase race on a good surface with a bet strength rating of 4. Bets with a smaller value of days since a horses last race and maximum available bet before the odds change should be preferred.

# Creating the model

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedFirst the data is split into train, validation and test sets in the ratio of 80:10:10 before a logistic regression model is fitted in order to calculate the baseline score to be improved upon. The classification ensembles Random Forest and XGBoost will be used as a comparison to the baseline model and then the best model selected and used for the remainder of the project. After tuning the hyperparameters of each model using GridSearchCV with a KFolds cross validator, the accuracy and macro average F1 score (which takes both precision and recall into account) of each model can be compared. These metrics will be saved in a .json file in the model directory in case of the need for inspection at a later date. As is demonstrated in Figure X, using the XGBoost Classifier has both a greater accuracy and F1 score than the other two models, and so will be used going forward.

## Testing the model