Predicting Horse Racing Results

Metis Project 3

By: Patrick Bovard

February 9, 2021

Agenda

- Project Introduction
- Methods and Feature Engineering
- Final Model and Results
  - Use Example
  - Challenges
- Future Improvements

Horse Racing is a multi-billion-dollar industry

- Prize money from races is over $3.5 billion per year *
- Over $100 billion wagered annually on horse races*
- Almost $16 billion wagered with the Hong Kong Jockey club in the last year alone**

Can machine learning be utilized to make informed race predictions?

*https://www.americasbestracing.net/the-sport/2017-american-horse-racing-vs-the-world-whats-the-same-whats-different
**https://www.drf.com/news/hong-kong-jockey-club-handles-158-billion-2019-20-season

Project Goal: Build a model that can successfully predict horse racing results in Hong Kong

- Data Set
  - Graham Daley's Hong Kong Racing Data from Kaggle
- Features:
  - Physical Horse Data
  - Previous Racing Stats
  - Comparison to the Field
- Classification Classes:
  - Show (Place 1st, 2nd, or 3rd)
  - Not Show (Place 4th or lower)
- Evaluation Metrics:
  - Precision
  - Fbeta (beta=0.5)

Tools such as Pandas, Python, PostgreSQL, and Scikit-Learn were used to iteratively build and improve a classification model

Precision: 28.67%
Fbeta: 0.218

- Included horse specific physical data
- Horse rating, weight, age, sex, etc.

Precision: 54.4%
Fbeta: 0.299

- Added Previous Race Stats: Rolling Averages for Finish Time, Speed, etc.
- Added comparison stats to field for weight, rating, etc.

Precision: 60.5%
Fbeta: 0.464

- Added odds to show, race distance, and horses in the race after observing residuals
- Added top jockey/trainer, home country

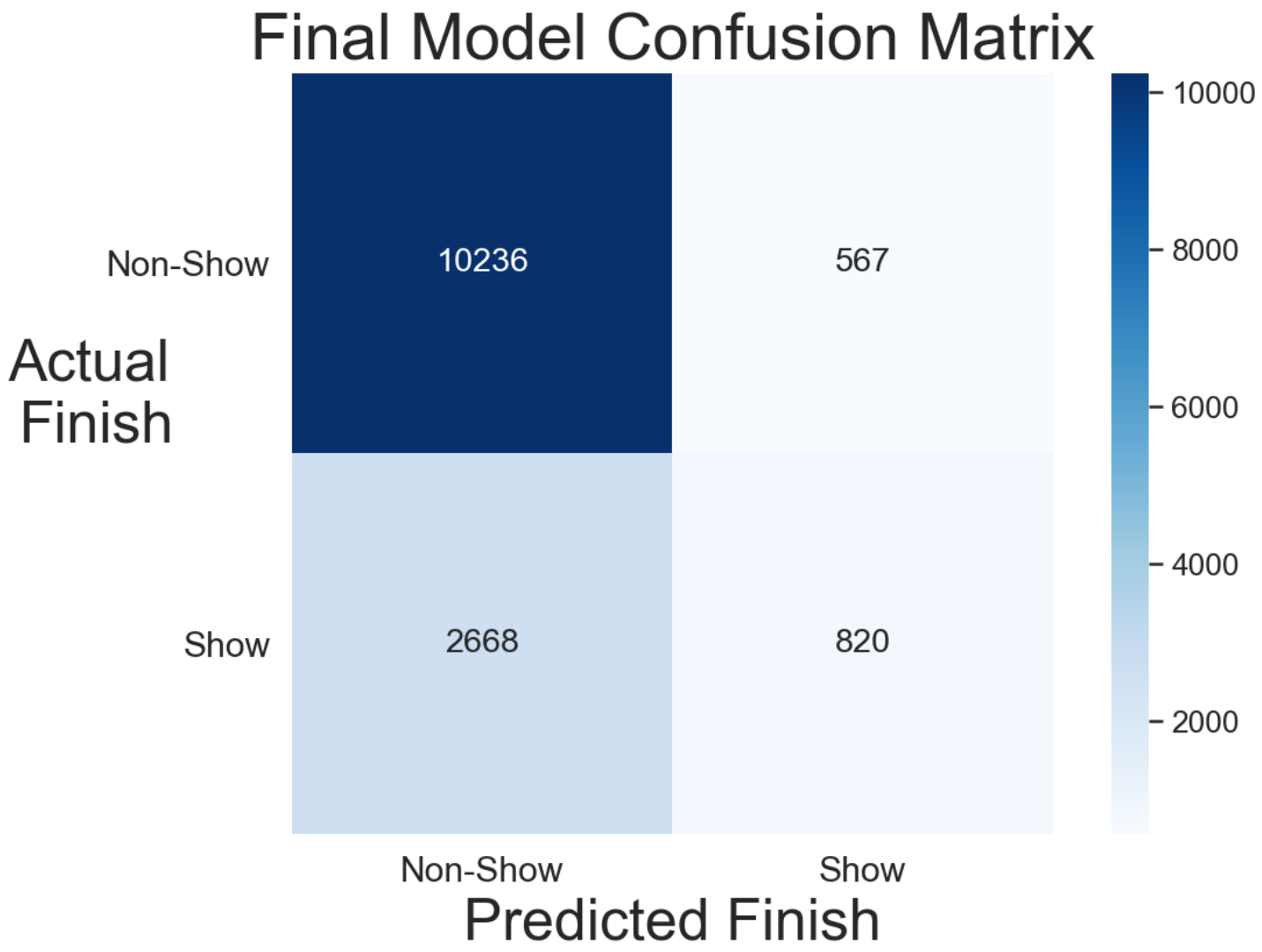
Precision: 60.7%
Fbeta: 0.474

- Selected Model: XGBoost
- Added Days Since Last Race, In Race Rolling Averages
- Added strong interaction terms

TUNING HYPERPARAMETERS

Final Model: XGBoost Classifier

- **Test Precision: 59.1%**
- Test Fbeta: 0.454
- Features w/ most importance (gain):
  - Place Odds (several interaction terms)
    - Horse Rating
    - Applied Weight
  - Rolling Averages: Lengths Behind Leader, Speed, Time Drop
  - Career Show Rate
  - Top 10 Jockey/Trainer



The show probabilities from the model can be utilized to evaluate a racing field

- By assigning probabilities, the model can go beyond hard classification

Race 4659 - Card

Horse	Show Odds	Show Probability	Race Result
1	1.9	43%	8
2	2.5	42%	2
3	2.5	38%	4
4	2.2	35%	11

The show probabilities from the model can be utilized to evaluate a racing field

- By assigning probabilities, the model can go beyond hard classification
- Case 1: Deciding between horses with similar odds

Race 4659 - Card

Horse	Show Odds	Show Probability	Race Result
1	1.9	43%	8
2	2.5	42%	2
3	2.5	38%	4
4	2.2	35%	11

The show probabilities from the model can be utilized to evaluate a racing field

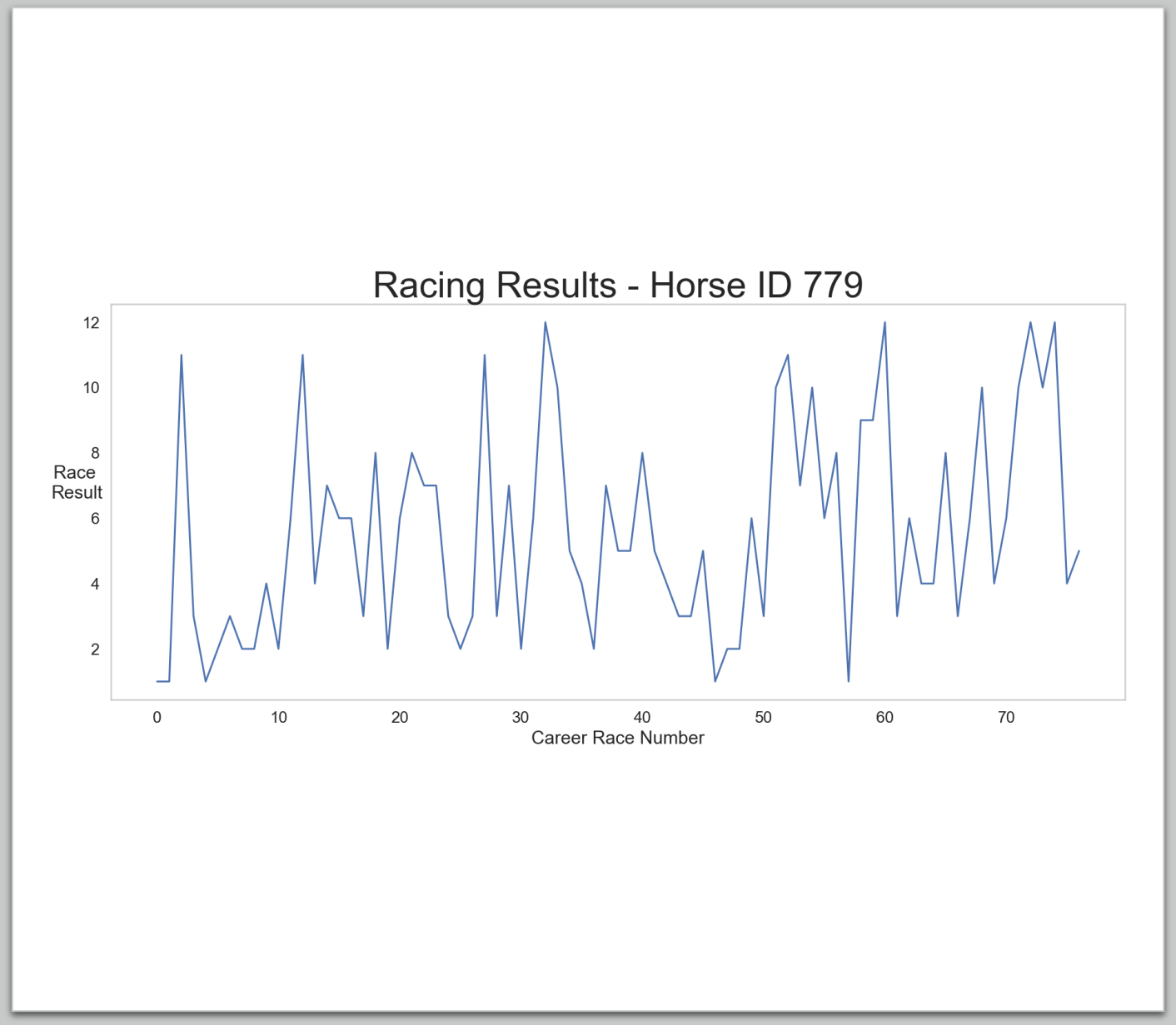
- By assigning probabilities, the model can go beyond hard classification
- Case 1: Deciding between horses with similar odds
- Case 2: Finding advantages against the odds

Race 4659 - Card

Horse	Show Odds	Show Probability	Race Result
1	1.9	43%	8
2	2.5	42%	2
3	2.5	38%	4
4	2.2	35%	11

The model struggled on horses with lower career races, shorter distance races, and races with no dominant horse

- False Positives:
  - Shorter distance – less time between show and non-show
- False Negatives:
  - Heavy “underdogs” with less races
  - No clear favorite in the race
- General Challenges
  - Very small differences between 3rd place and lower places
  - Trying to predict a variable result



Future Improvements

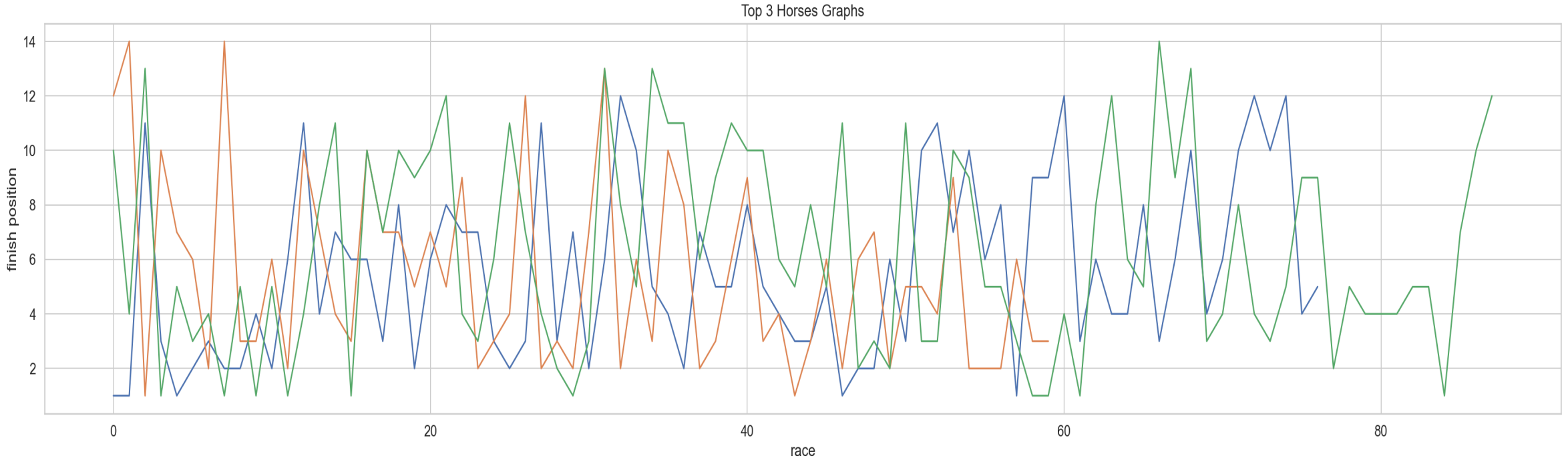
- More specialized models to reduce false negatives
  - Track
  - Track Configuration
  - Distance
- Similar model for different markets – United States

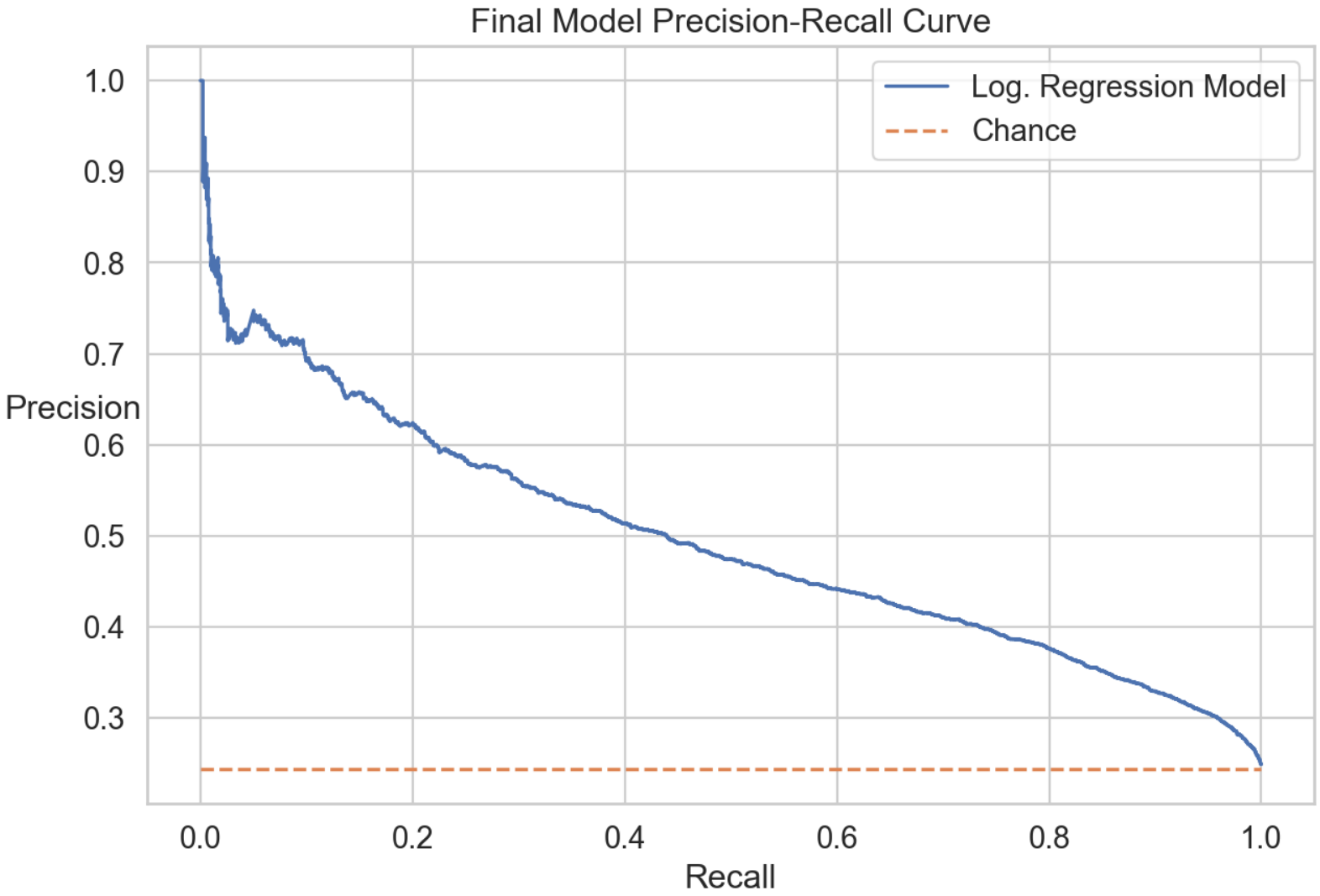
Thank you!
Any questions?

Appendix

Horse Racing a highly variable sport to predict.

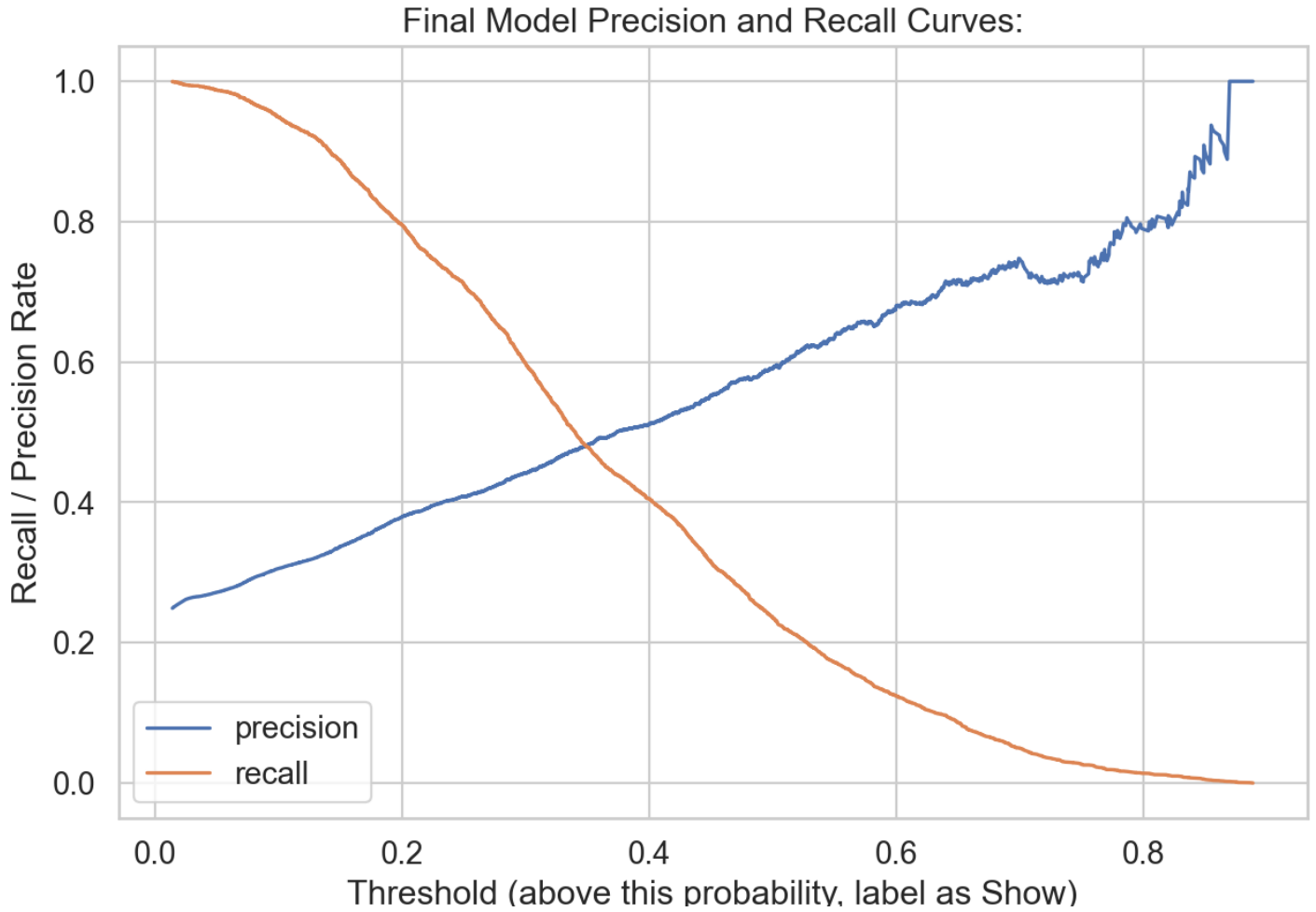
- Above: top 3 horses in terms of number of shows
- There tend to be “waves” in performance, with a lot of volatility
- Led to adding some moving average metrics to the features

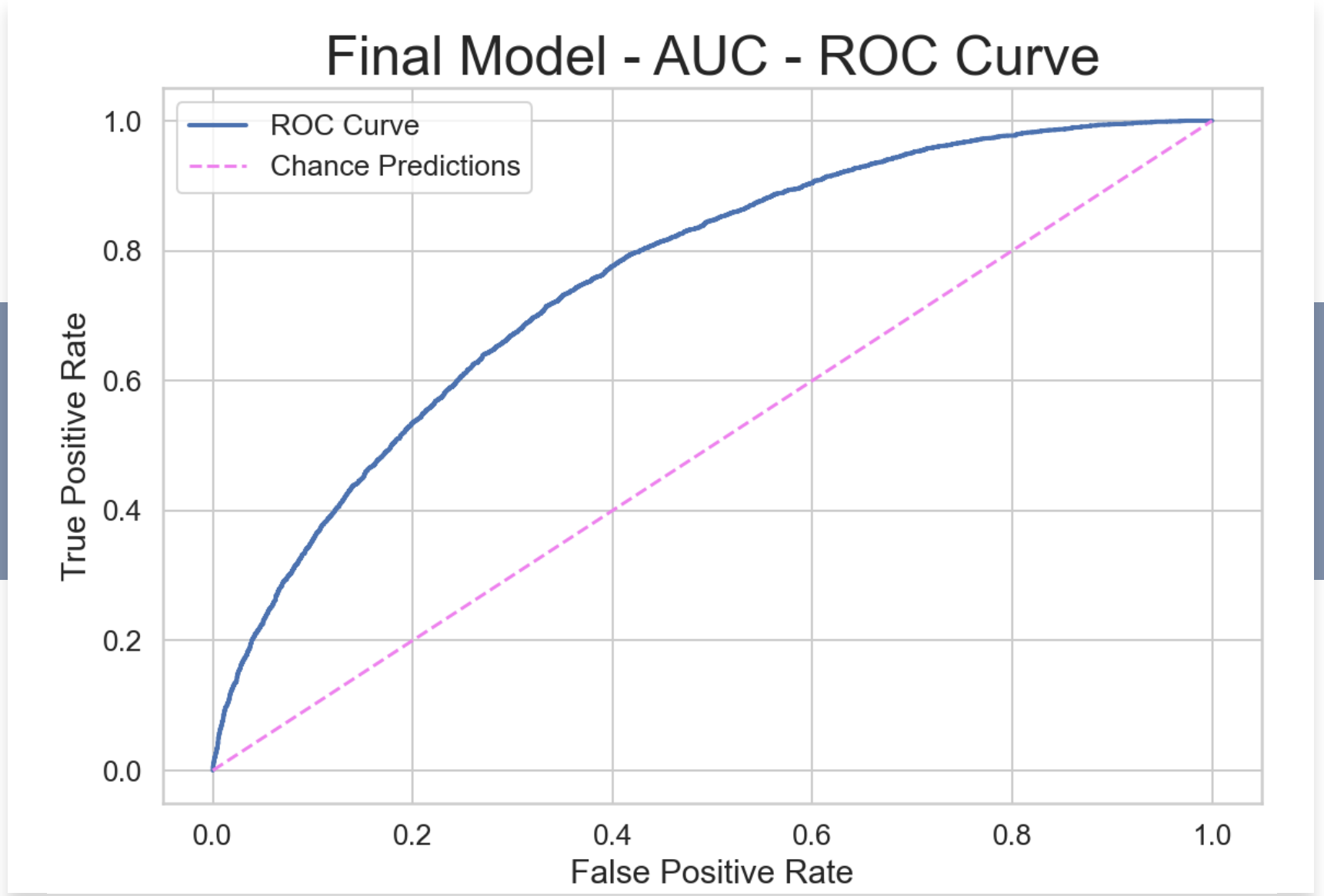




Final Model Precision-Recall Curve

Final Model Precision and Recall Curves





Final Model AUC ROC Curve

Final Model – False Positive Example

- About 0.3 seconds between 5th and 1st
- Less class differentiation

Race 4849 – 1200 m

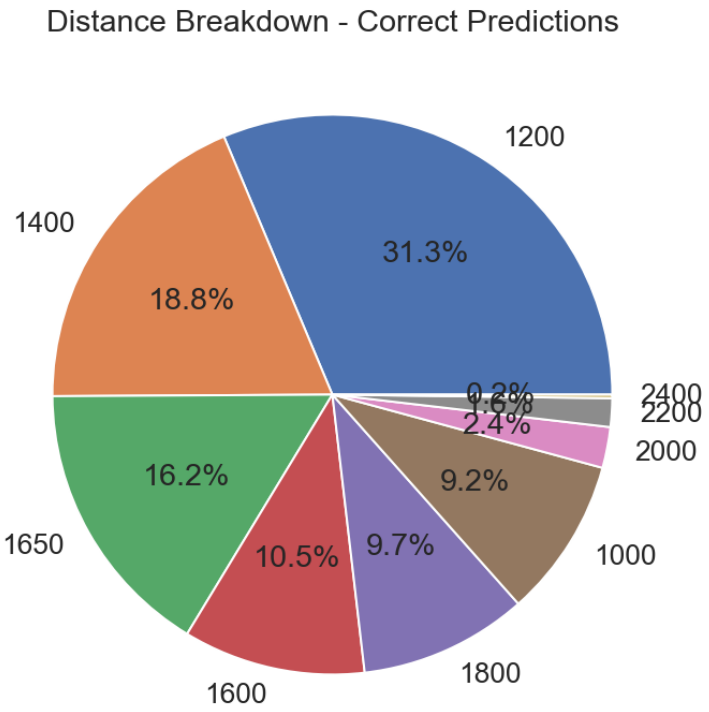
Place	Finish Time	Show Probability
1	70.45	25%
2	70.48	56%
3	70.55	34%
4	70.71	87%
5	70.78	30%

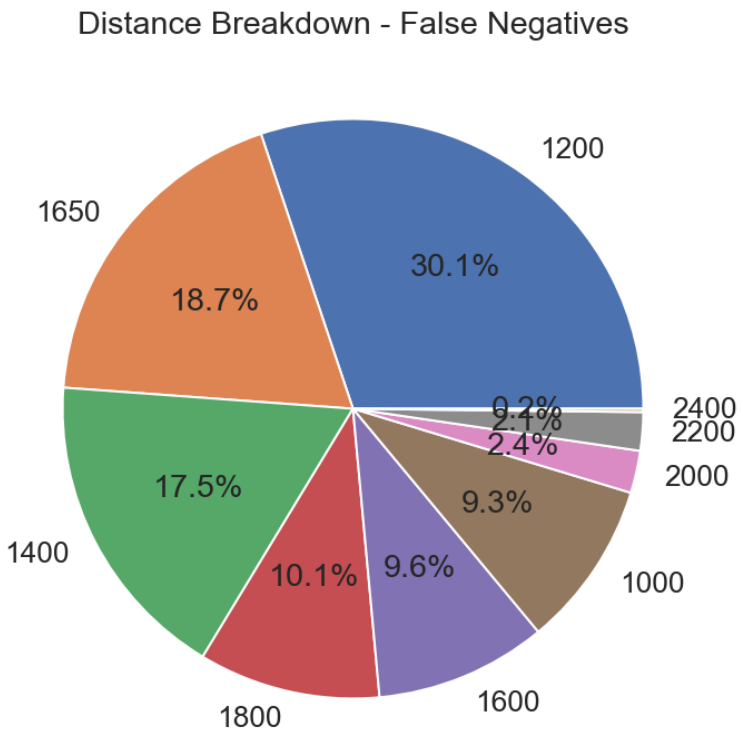
Final Features List:

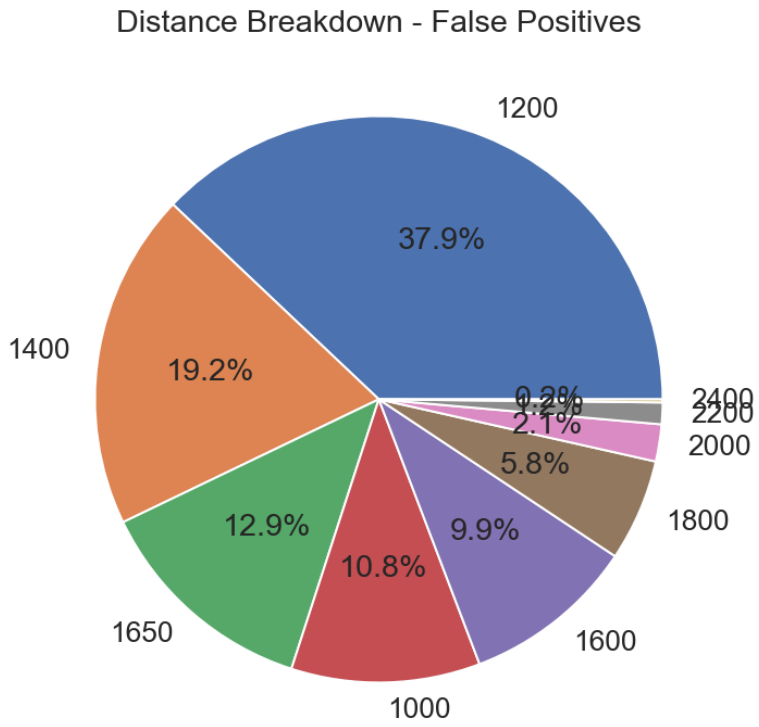
- 'actual_weight * place_odds', 'place_odds * three_race_rolling_average_distance_per_time', 'horse_rating * place_odds', 'place_odds + three_race_avg_mid_time', 'place_odds + three_race_rolling_average_distance_per_time', 'AUS', 'NZ', 'IRE', 'GB', 'USA', 'Other', 'horse_age', 'horse_rating', 'declared_weight', 'actual_weight', 'draw', 'place_odds', 'distance', 'three_race_rolling_avg_finish', 'three_race_rolling_average_lengths', 'three_race_rolling_average_time', 'three_race_rolling_average_distance_per_time', 'horses_in_field', 'field_rating_rank', 'diff_from_field_rating_avg', 'field_age_rank', 'diff_from_field_age_avg', 'diff_from_field_declared_wgt_avg', 'field_dec_wgt_rank', 'diff_from_field_handicap_wgt_avg', 'field_handicap_wgt_rank', 'career_races', 'career_shows', 'shows_in_last_5_races', 'three_race_rolling_average_len_len1', 'float_days_since_last_race', 'three_race_avg_mid_time', 'top_10_jockey', 'top_10_trainer', 'three_race_avg_mid_len_gain', 'career_show_rate', 'horse_type_Filly', 'horse_type_Gelding', 'horse_type_Horse', 'horse_type_Mare', 'horse_type_Rig'

Race by Distance: Residuals (Final Model)

- False Positives occurred in a higher proportion of races below 1600 m than the dataset as a whole.
- False Negatives occurred on a more similar proportion compared to the dataset as a whole.



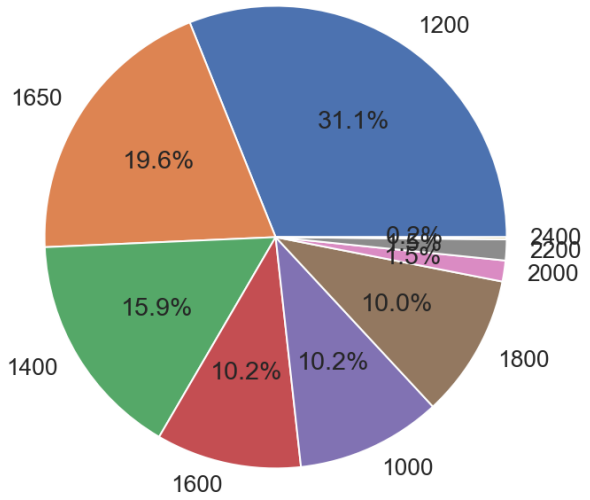




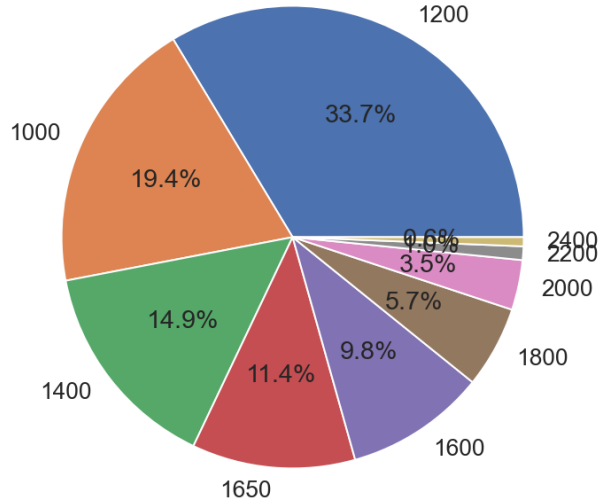
Race by Distance: Residuals (Model 3)

- False Positives occurred in a higher proportion of races below 1600 m than the dataset as a whole.
- False Negatives occurred on a more similar proportion compared to the dataset as a whole.

Distance Breakdown - False Negatives



Distance Breakdown - False Positives



Distance Breakdown - Whole Set