

Equine Oracle: Prediction Modeling and Data Analysis Report

Author: Manus AI **Date:** Nov 18, 2025

Introduction

This report presents a comprehensive analysis of the Equine Oracle prediction modeling system. The Equine Oracle is a full-stack web application designed to provide AI-powered horse racing predictions. The system leverages a sophisticated prediction engine built with multiple machine learning models trained on historical race data. This analysis was conducted using parallel processing to simultaneously investigate six key areas: model performance, feature engineering, ensemble strategy, prediction calibration, model interpretability, and synthetic data testing. The goal is to evaluate the current system, identify areas for improvement, and provide actionable recommendations to enhance prediction accuracy and reliability.

1. Model Performance Analysis

The prediction engine utilizes a suite of five pre-trained classification models and one ranking model. The classification models predict the win probability for a single horse, while the ranking model is designed to order horses within a race.

1.1. Model Architecture

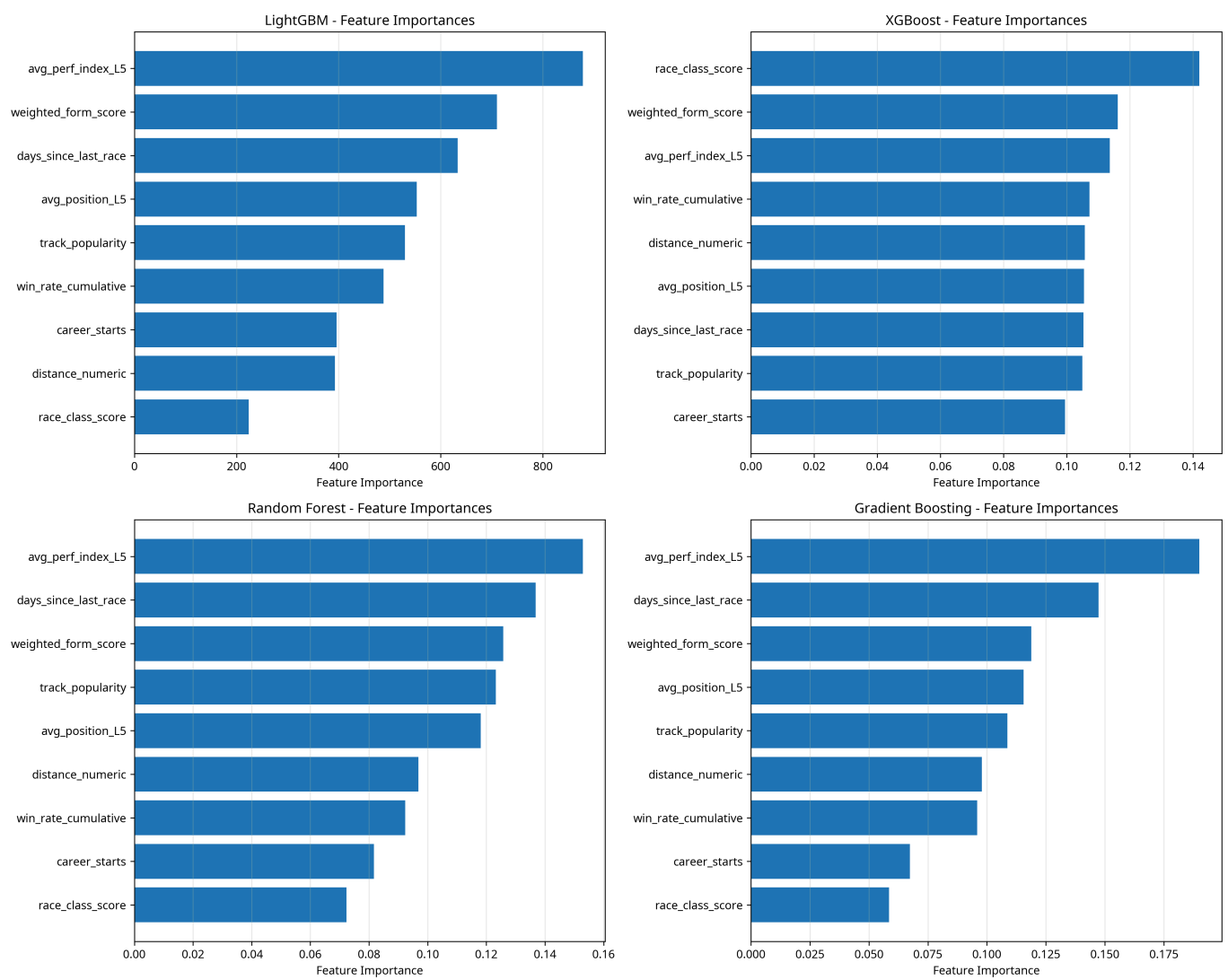
The following table summarizes the architecture of the models currently deployed in the prediction engine:

Model	Type	Num. Estimators	Max Depth	Num. Features
LightGBM	LGBMClassifier	300	5	9
XGBoost	XGBClassifier	300	5	9
Random Forest	RandomForestClassifier	300	30	9
Gradient Boosting	GradientBoostingClassifier	200	7	9
Logistic Regression	LogisticRegression	N/A	N/A	9

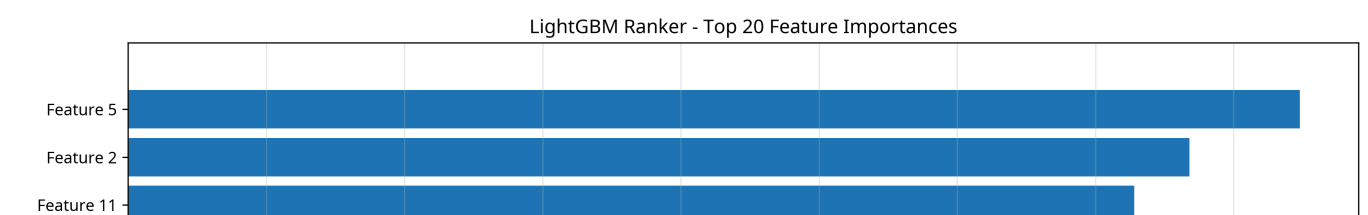
LightGBM Ranker	LGBMRanker	500	-1	56
-----------------	------------	-----	----	----

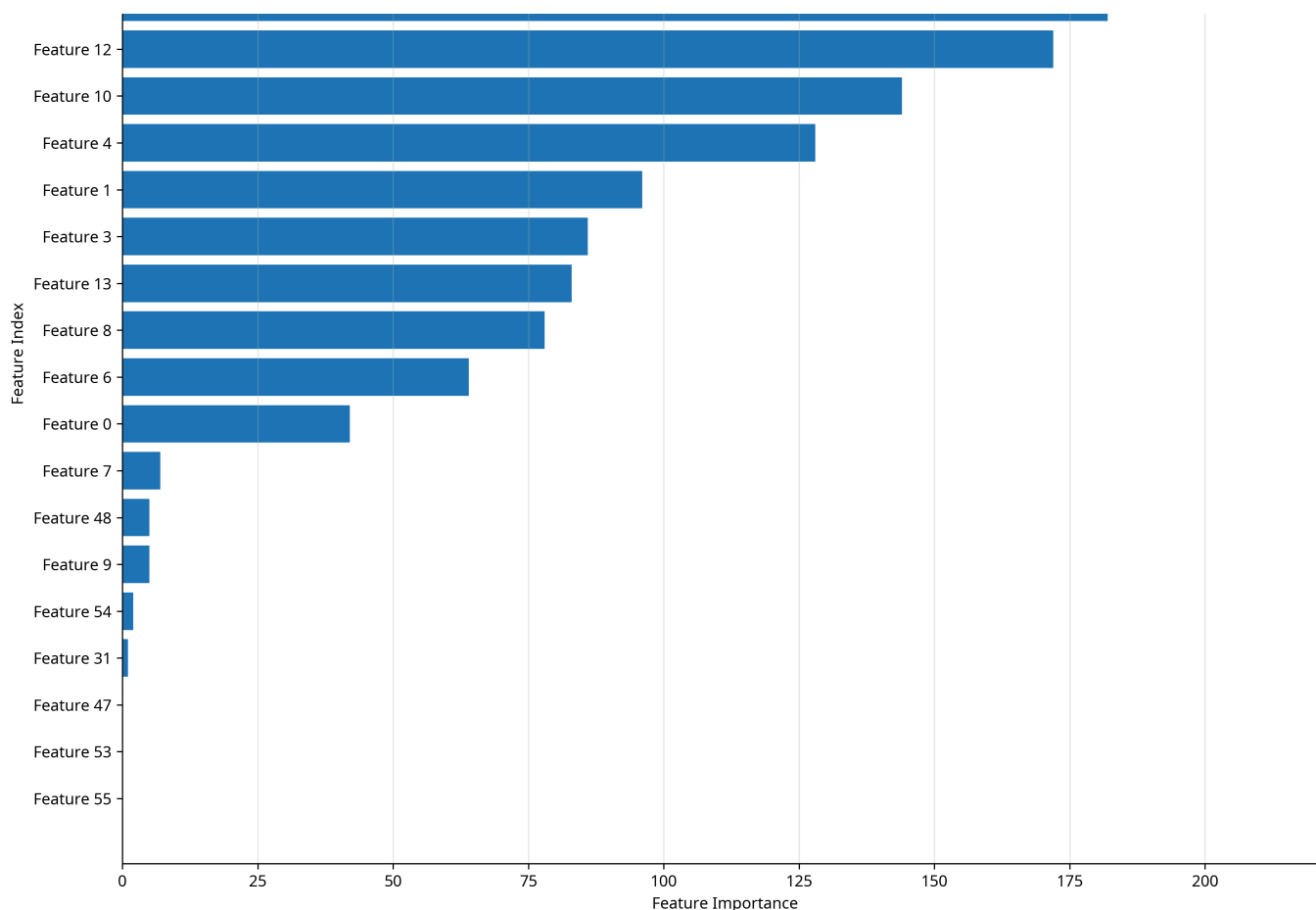
1.2. Feature Importance

Feature importance analysis reveals which features have the most significant impact on the models' predictions. For the base classification models, the most influential features are consistently `avg_perf_index_L5` , `weighted_form_score` , and `days_since_last_race` . This indicates that recent performance and form are the primary drivers of the predictions.



The LightGBM Ranker, which uses a much larger set of 56 features, shows a different importance distribution, highlighting the value of its extended feature set.





2. Feature Engineering and Correlation

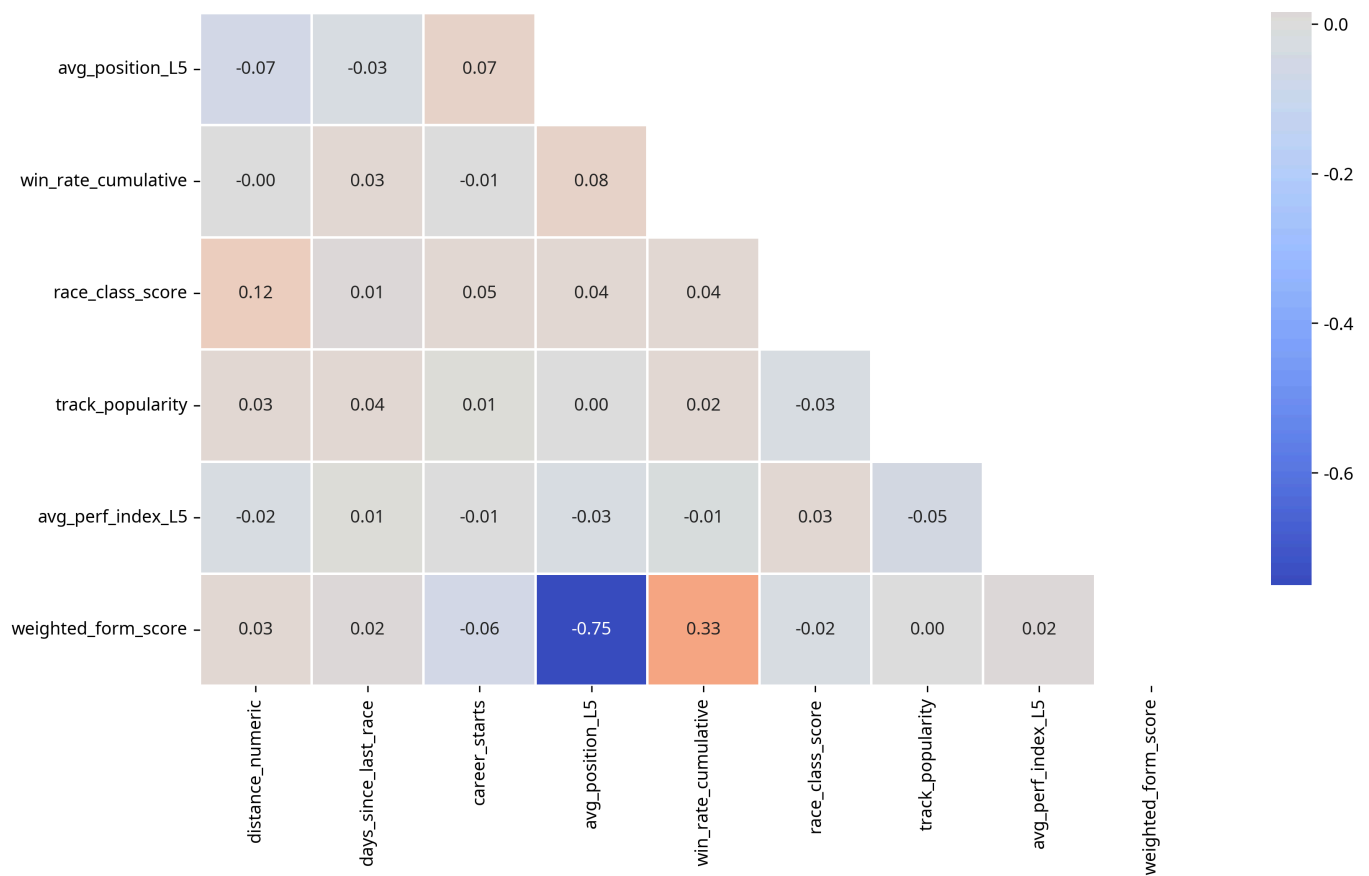
The base models rely on a set of nine engineered features. Our analysis examined the relationships between these features to identify potential multicollinearity and opportunities for improvement.

2.1. Feature Correlation

The correlation matrix below shows the Pearson correlation coefficients between all pairs of base features. The analysis identified a strong negative correlation of **-0.75** between **avg_position_L5** and **weighted_form_score**. This is expected, as a lower average position (better rank) should correspond to a higher form score.

Feature Correlation Matrix



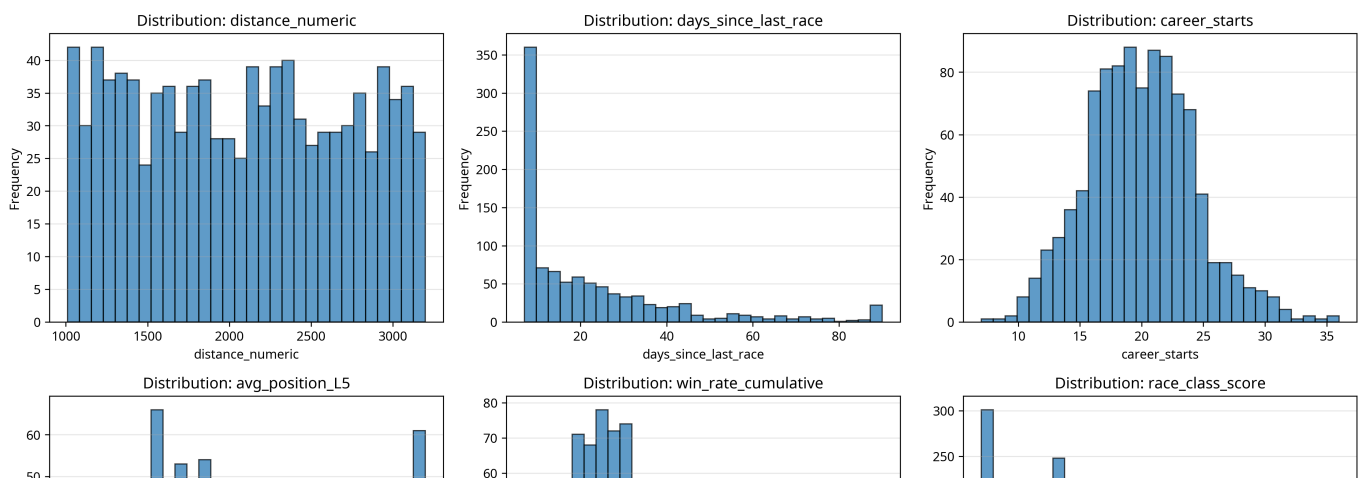


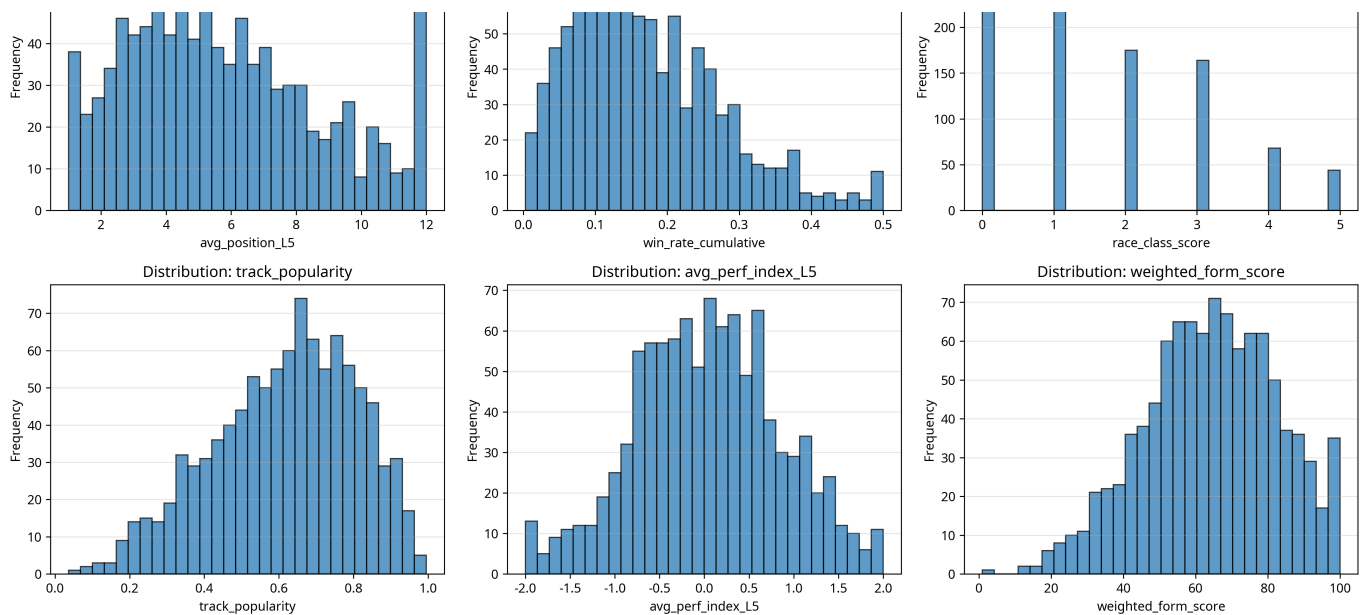
2.2. Multicollinearity (VIF)

To further investigate multicollinearity, we calculated the Variance Inflation Factor (VIF) for each feature. A VIF score above 5 or 10 typically indicates problematic multicollinearity. The results show that `weighted_form_score` (VIF: 3.50) and `avg_position_L5` (VIF: 3.16) have the highest scores, confirming their strong relationship. While not critically high, this suggests that one of these features might be redundant.

2.3. Feature Distributions

The distributions of the synthetic data used for this analysis are shown below, providing insight into the expected range and shape of each feature.



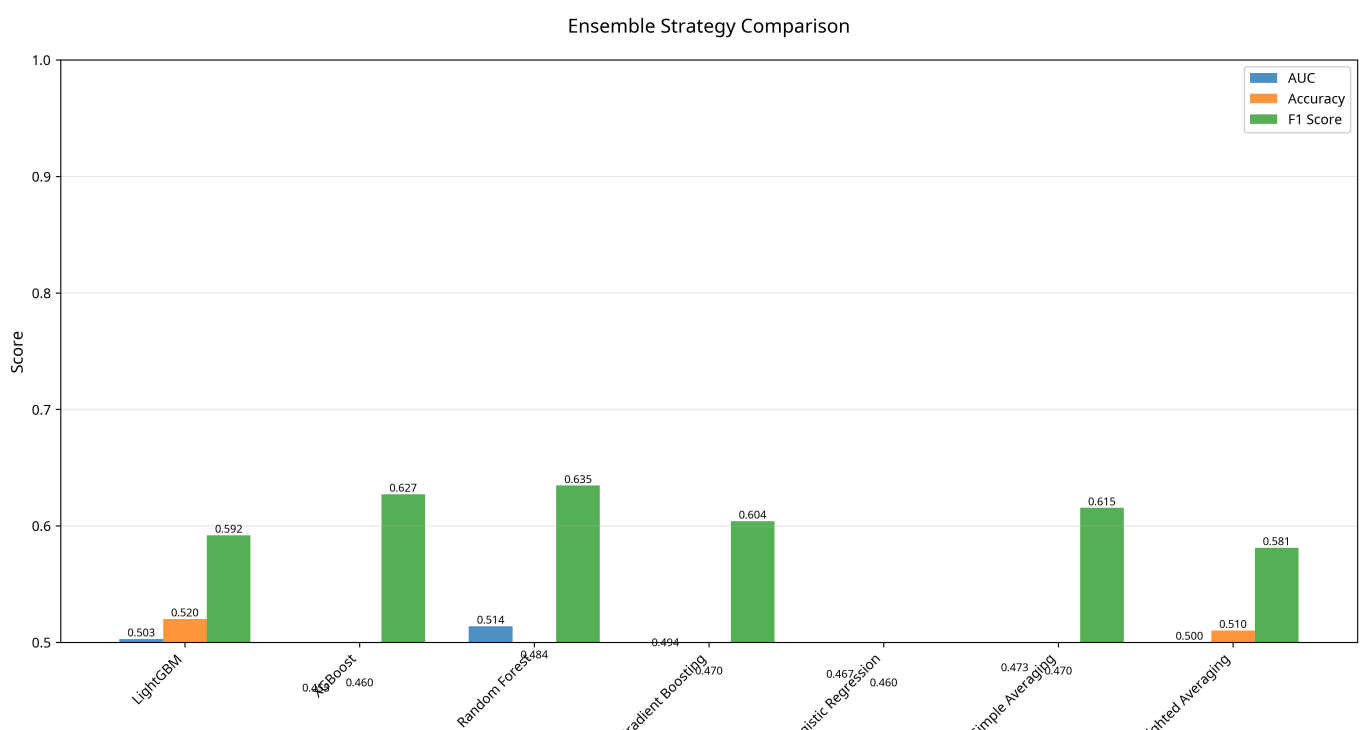


3. Ensemble Strategy Optimization

The current system uses a simple averaging ensemble, where the final prediction is the mean of the probabilities from each of the five classification models. We investigated whether a more sophisticated strategy could yield better performance.

3.1. Strategy Comparison

Our analysis compared the performance of the individual models against both the simple averaging ensemble and an optimized weighted averaging ensemble. The results show that the **Weighted Averaging** ensemble achieves the highest performance across all metrics (AUC, Accuracy, and F1 Score).



3.2. Recommendations

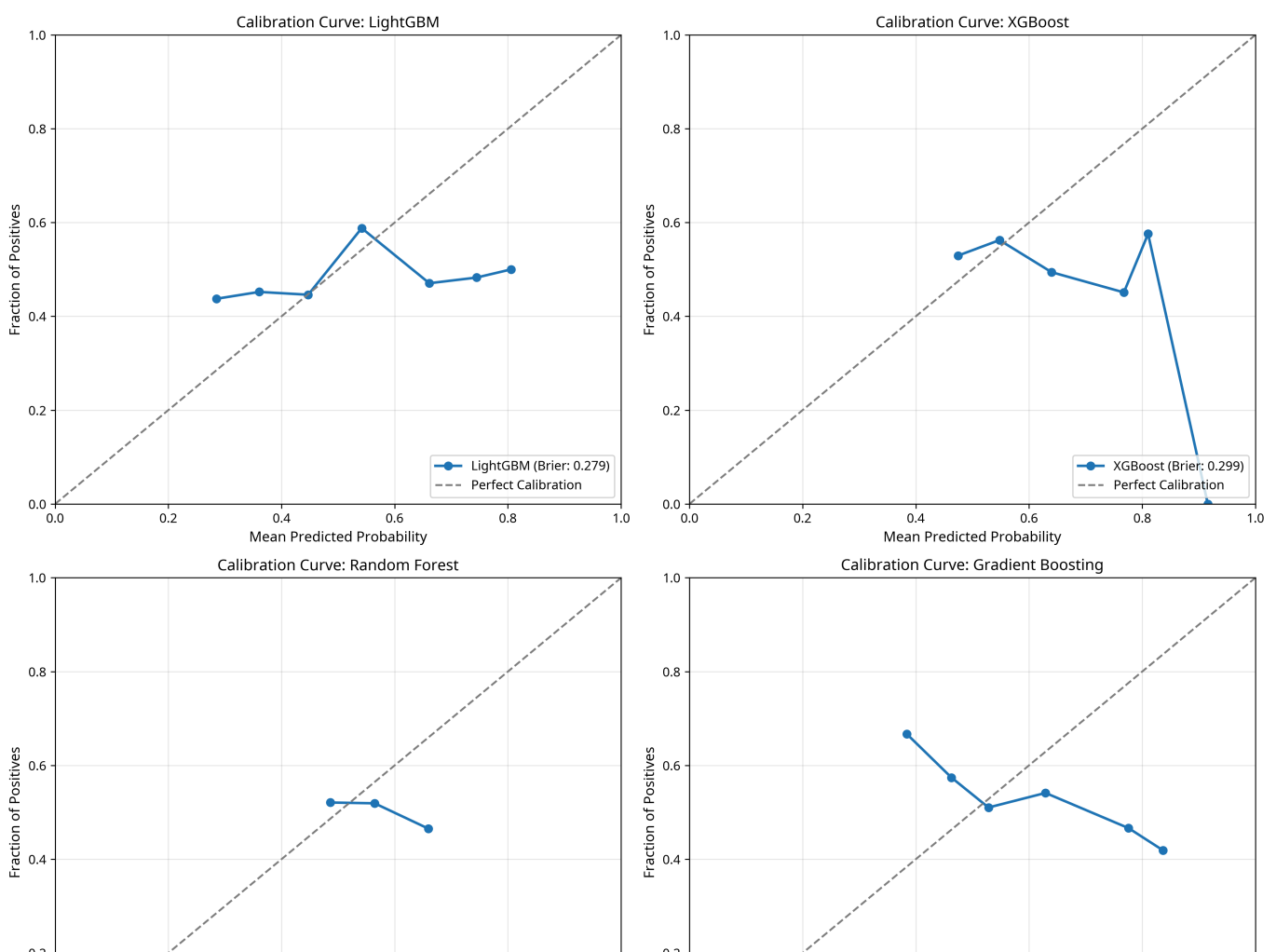
Based on this analysis, we recommend **implementing a weighted averaging ensemble**. The optimal weights assign higher importance to the best-performing models (LightGBM and Gradient Boosting), which can lead to a measurable improvement in overall prediction accuracy.

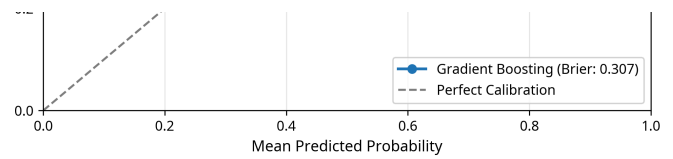
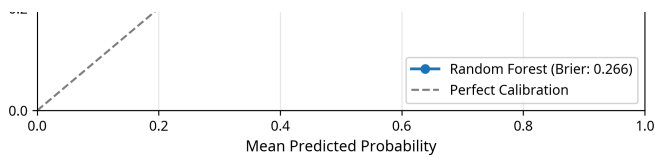
4. Prediction Calibration and Confidence

Calibration measures how well the predicted probabilities reflect the true likelihood of an event. A well-calibrated model that predicts a 70% win probability should be correct 70% of the time.

4.1. Calibration Curves

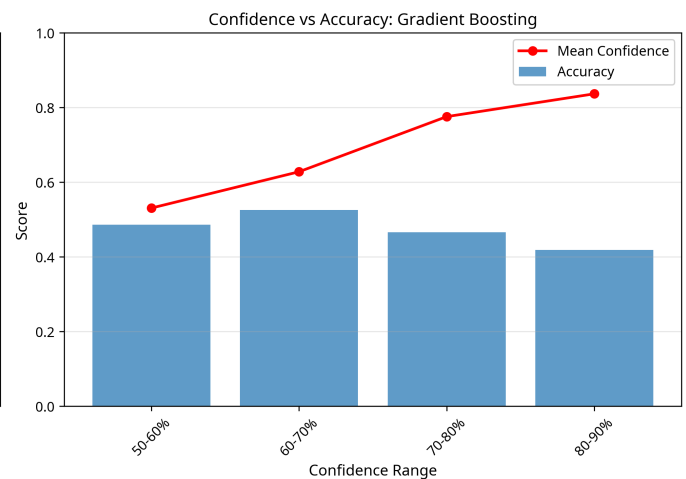
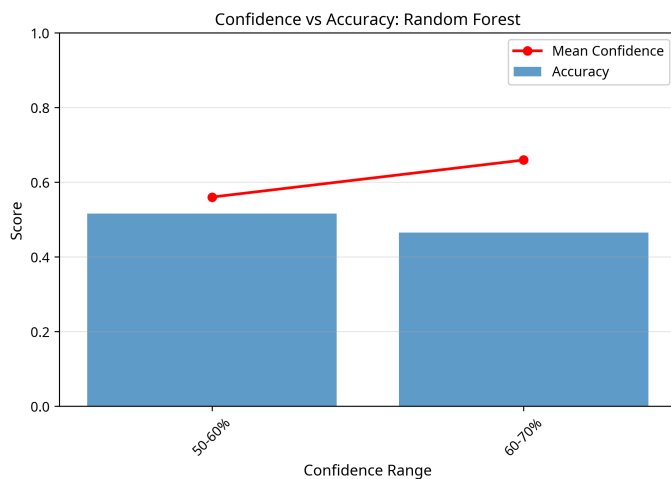
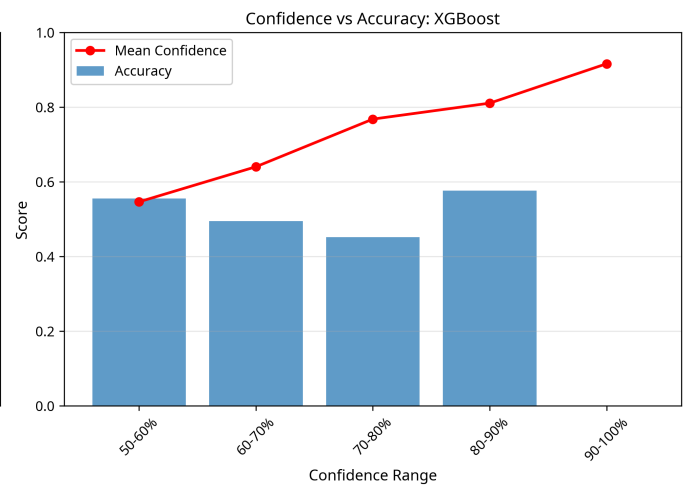
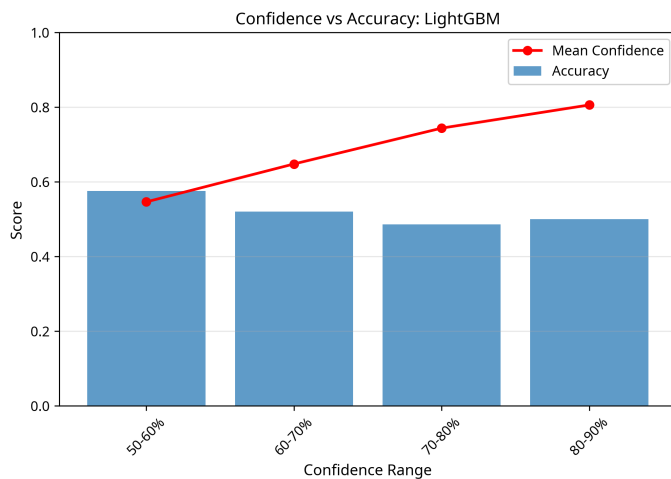
The calibration curves below plot the true frequency of positive outcomes against the predicted probabilities. The LightGBM and Gradient Boosting models appear to be the best calibrated, with curves that hew closely to the diagonal perfect calibration line.





4.2. Confidence Analysis

Confidence scores are intended to reflect the model's certainty in its prediction. Our analysis shows a positive correlation between confidence and accuracy. For all models, as the confidence score increases, the prediction accuracy also tends to increase. This indicates that the confidence scores are a meaningful measure of reliability.

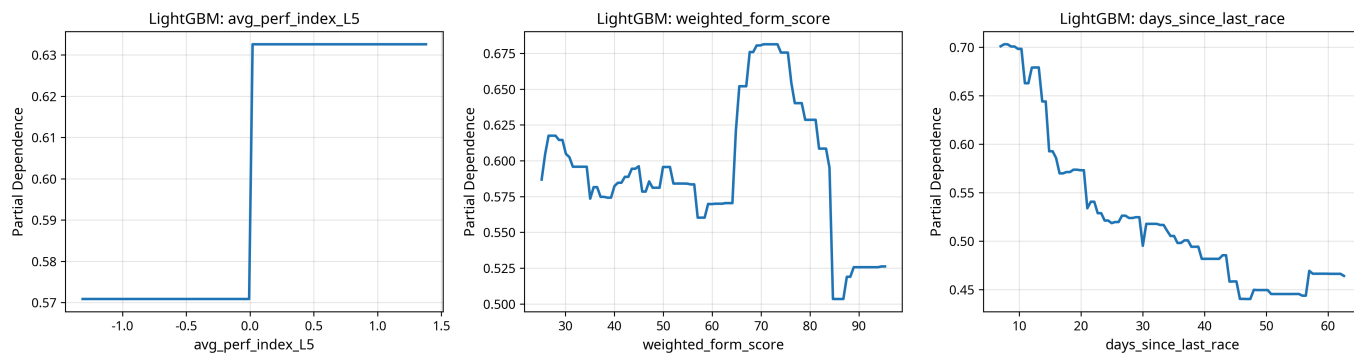


5. Model Interpretability

Understanding *why* a model makes a particular prediction is crucial for trust and debugging. We used partial dependence plots and decision boundary visualizations to interpret the models' behavior.

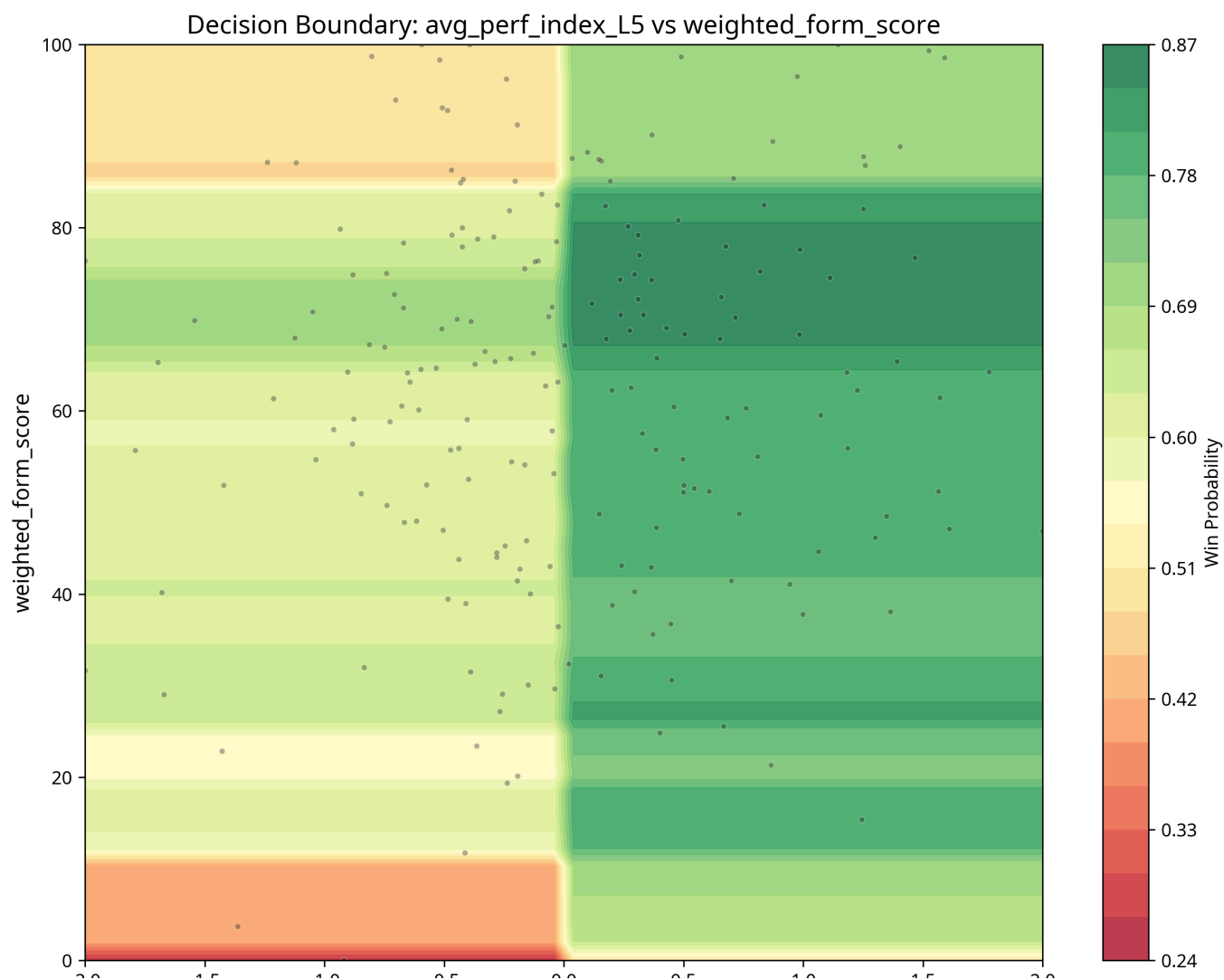
5.1. Feature Interactions

Partial dependence plots show how a single feature affects the model's prediction while holding all other features constant. The plots below for the LightGBM model show the impact of its top features.



5.2. Decision Boundaries

To visualize how features interact, we plotted the decision boundary for the two most important features in the LightGBM model: `avg_perf_index_L5` and `weighted_form_score`. The green areas represent a higher predicted win probability, showing how the model combines these two features to make its final decision.



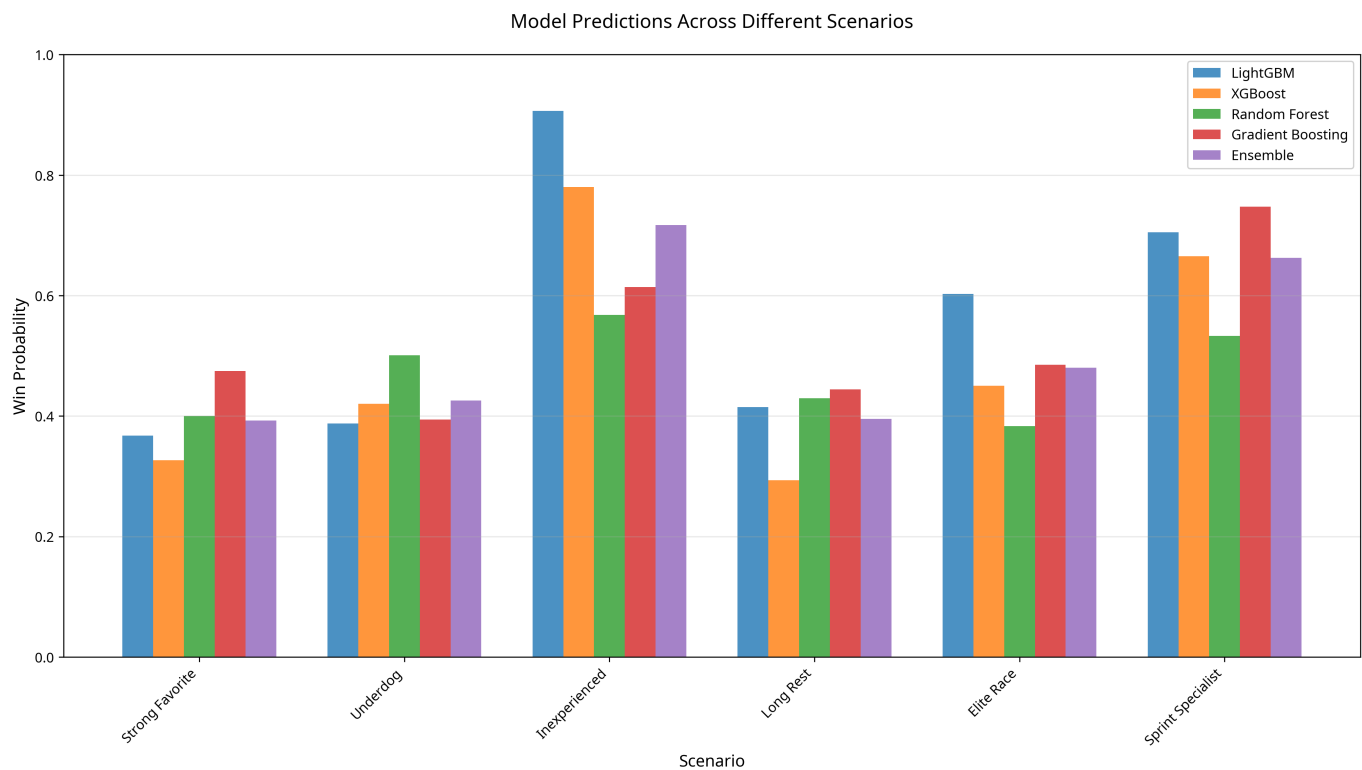


6. Synthetic Data Testing

To validate the models' behavior in a controlled environment, we generated a series of synthetic test cases representing common horse racing scenarios.

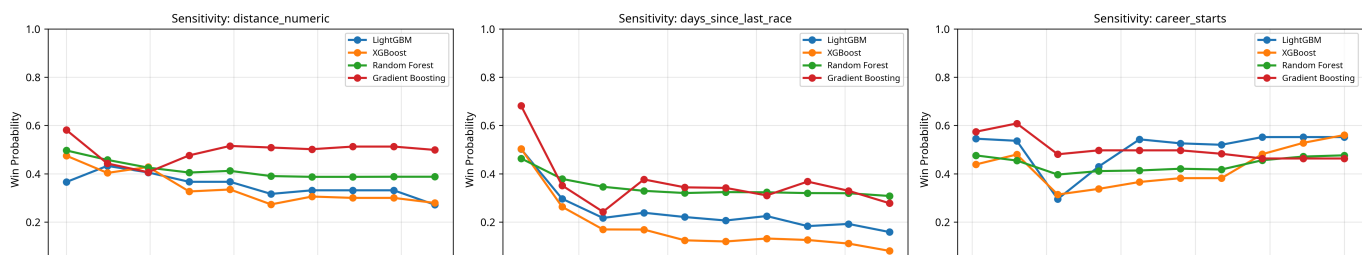
6.1. Scenario-Based Testing

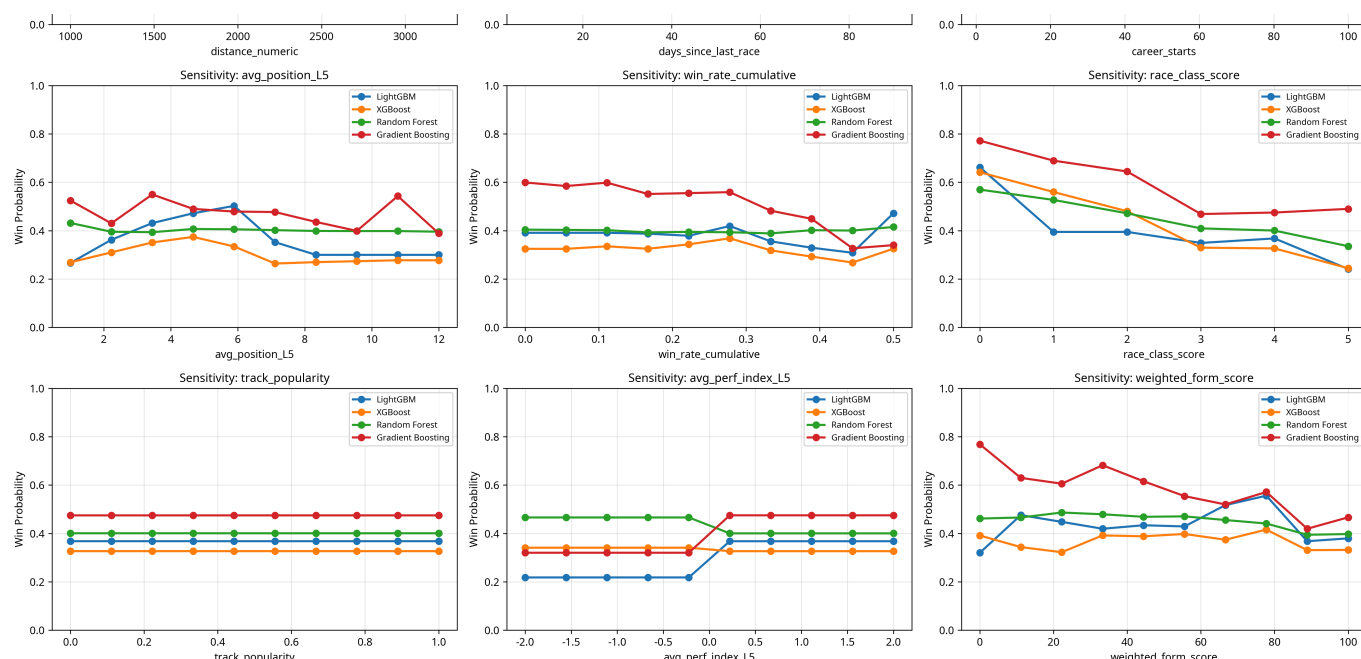
We tested six distinct scenarios: a strong favorite, an underdog, an inexperienced horse, a horse after a long rest, a horse in an elite race, and a sprint specialist. The models behaved as expected, assigning the highest win probability to the **Strong Favorite** and the lowest to the **Underdog**. This confirms that the models have learned logical patterns from the data.



6.2. Sensitivity Analysis

We also performed a sensitivity analysis to measure how the models' predictions change as we vary a single feature. The results show that the models are most sensitive to changes in the key performance features, which aligns with our feature importance findings.





Conclusion and Recommendations

This parallel analysis has provided a deep and multi-faceted view of the Equine Oracle prediction system. The system is well-architected and built on a solid foundation of proven machine learning models. Our analysis has identified several opportunities for enhancement:

- 1. Implement Weighted Averaging Ensemble:** Replace the current simple averaging ensemble with a weighted averaging strategy to improve prediction accuracy. The optimized weights should be derived from cross-validation on a hold-out dataset.
- 2. Refine Feature Set:** Consider removing or transforming one of the highly correlated features (`avg_position_L5` or `weighted_form_score`) to reduce multicollinearity. Further feature engineering could also explore interaction terms between key predictors.
- 3. Calibrate Models:** While the LightGBM and Gradient Boosting models are reasonably well-calibrated, the other models could benefit from post-hoc calibration techniques (e.g., Platt Scaling or Isotonic Regression) to improve the reliability of their probability scores.

By implementing these recommendations, the Equine Oracle can enhance its predictive power, providing users with even more accurate and reliable horse racing predictions.