Model Comparison

Mikelino Chatigia

September 2024

1 Introduction

The analysis of sporting outcomes has long been a domain of interest, not only for enthusiasts but also for statisticians and data scientists. In recent years, machine learning models have become invaluable in predicting game results, allowing for more informed decisions in various areas, including sports betting. This paper presents a detailed analysis of two predictive models applied to Major League Baseball (MLB) data, focusing on the seasons of 2020 and 2021. Our goal is to evaluate the performance of these models using a variety of performance metrics to assess their ability to predict game outcomes under different conditions.

The 2020 MLB season was heavily disrupted by the COVID-19 pandemic, resulting in a significantly reduced number of games. Specifically, the 2020 season included only 60 games compared to the typical 162-game schedule due to health and safety concerns. In contrast, the 2021 season saw a return to a more standard format, with 162 games played by each team. These two periods allow us to investigate how predictive models perform under "stress" conditions (during the pandemic-shortened season) compared to a regularized season. As such, breaking down our analysis into two parts—season-specific evaluation and combined analysis across both seasons—provides a comprehensive understanding of model robustness and accuracy.

Our approach includes turning raw American odds into implied probabilities to make the predictions comparable, followed by normalization. We then used several evaluation metrics, each progressively increasing in complexity, to measure the accuracy and quality of the models' predictions. These include basic accuracy, ROC-AUC, Brier score, and cross-entropy, all of which highlight different aspects of the models' performance, from simple correctness to the calibration of predicted probabilities.

In this paper, we will discuss the transformation of raw odds, explain the metrics we employed in detail, present results for the individual seasons and their aggregate, and provide an in-depth discussion of our findings.

Contents

2 Evaluation Metrics

In this section, we explore the four key evaluation metrics used in this study: Accuracy, ROC-AUC, Brier Score, and Cross-Entropy (Log Loss). Each of these metrics assesses different aspects of the model's performance, from its simple predictive correctness to the calibration of its predicted probabilities. Below, we provide a detailed explanation of each metric, along with its mathematical formulation and its significance for our analysis.

2.1 Accuracy

Accuracy is the simplest and most intuitive metric for evaluating a model's performance. It is defined as the proportion of correct predictions made by the model out of the total number of predictions. Mathematically, accuracy is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where:

- TP = True Positives (correct positive predictions),
- TN = True Negatives (correct negative predictions),
- FP = False Positives (incorrect positive predictions),
- FN = False Negatives (incorrect negative predictions).

Accuracy is straightforward and easy to interpret, but it has some limitations, particularly when dealing with imbalanced datasets. In our case, since baseball games have a binary outcome (win/loss), accuracy provides a general overview of model performance. However, it does not account for how well the model is calibrated, which makes it necessary to complement this metric with others such as ROC-AUC and Brier Score.

2.2 ROC-AUC (Receiver Operating Characteristic - Area Under the Curve)

The ROC-AUC metric evaluates the model's ability to distinguish between two classes. To achieve this in our case, we construct a binary variable, where if a model scores a probability higher than 50%, it is classified as win (1), else as a loss (0). The ROC curve plots the True Positive Rate (TPR, also known as sensitivity or recall) against the False Positive Rate (FPR) at various threshold levels.

The TPR (True Positive Rate) is the proportion of actual wins that are correctly predicted by the model as wins. It is defined as:

$$\text{TPR} = \frac{TP}{TP + FN} \tag{2}$$

Where:

- TP = True Positives (games where the model correctly predicted a win),
- FN = False Negatives (games where the model predicted a loss, but the actual outcome was a win).

For example, if the model predicts a win with a probability greater than 0.5, and the actual outcome is a win, this is counted as a True Positive.

The FPR (False Positive Rate) is the proportion of actual losses that are incorrectly predicted by the model as wins. It is defined as:

$$\text{FPR} = \frac{FP}{FP + TN} \tag{3}$$

Where:

- FP = False Positives (games where the model predicted a win, but the actual outcome was a loss),
- TN = True Negatives (games where the model correctly predicted a loss).

In this context, a False Positive occurs when the model predicts a win (probability > 0.5), but the actual result is a loss.

ROC Curve and AUC: The ROC curve is created by plotting TPR on the y-axis against FPR on the x-axis across various threshold settings. The Area Under the Curve (AUC) provides a single scalar value that summarizes the model's ability to discriminate between wins and losses across all thresholds. The value of AUC ranges from 0 to 1, where:

- An AUC of 0.5 represents a model that is no better than random guessing,
- An AUC of 1.0 represents a perfect model,
- Values closer to 1.0 indicate a model with better discriminatory power.

Interpreting the ROC Plot: Figure 1 illustrates how ROC curves represent models with varying performance. The red dashed line indicates the performance of a random classifier, with an AUC of 0.5, where TPR = FPR across all thresholds. A model with an ROC curve above this line (green or blue) indicates better performance. The curve in blue represents a "perfect classifier," where TPR = 1 and FPR = 0, corresponding to an AUC of 1.0. This model correctly identifies all wins and losses. The green line represents a model with intermediate performance, performing better than random, but not perfectly.

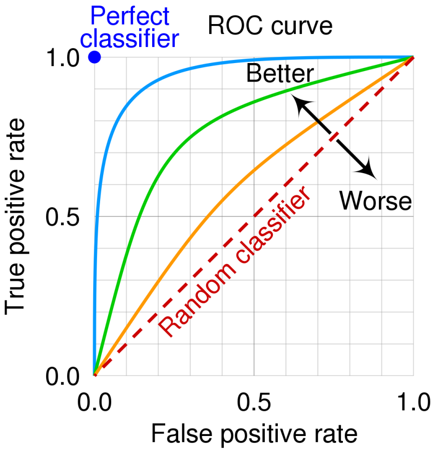

Figure 1: Example of ROC Curves. The blue line represents a perfect classifier with AUC = 1, while the dashed red line represents a random classifier with AUC = 0.5.

When comparing two models, the one with a higher AUC is considered to perform better in terms of distinguishing between predicted wins and losses. For instance, if one model has an AUC of 0.85 and the other has 0.75, the former is better at differentiating between wins and losses. The ROC curve also visually compares models; a model whose curve is consistently above another model's curve performs better.

ROC-AUC is particularly useful in binary classification problems, as it evaluates how well the model can separate the two possible outcomes (win vs. loss). A higher AUC value indicates better discriminatory power—this means that the model is better at distinguishing between games that result in wins and those that result in losses. This metric is especially important because it is not affected by class imbalance. For example, if there were significantly more wins than losses (or vice versa) in the dataset, accuracy could be misleading, as the model might simply predict the dominant class to achieve a high score. ROC-AUC, however, provides a more holistic view by evaluating the balance between TPR and FPR across all possible thresholds.

Moreover, the ROC curve allows for the comparison of models at different classification thresholds, providing flexibility in performance evaluation. This is particularly useful in sports predictions, where depending on betting strategies, one might prefer a model that is more conservative (fewer false positives) or more aggressive (more true positives).

By analyzing the ROC-AUC, we gain insights into the model's capability to handle uncertainty and borderline cases, where it is difficult to classify a game as either a win or a loss. In our evaluation, a higher AUC suggests that the model is robust in distinguishing between the two outcomes, even in cases where the predicted probabilities are close to the decision boundary.

2.3 Brier Score

The Brier Score measures the accuracy of probabilistic predictions, specifically how far the predicted probabilities are from the actual outcomes. It is a proper scoring rule, meaning it rewards models for making well-calibrated predictions. The Brier Score is defined as:

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \tag{4}$$

where:

- p_i = predicted probability of the outcome,
- o_i = actual outcome (1 for win, 0 for loss),
- N = total number of predictions.

It ranges from 0 to 1, where a score of 0 indicates perfect probabilistic predictions, and 1 represents the worst possible prediction. A lower Brier Score reflects better performance. Unlike accuracy, which only considers whether a prediction is correct, the Brier Score accounts for the confidence of the predictions, making it particularly useful when dealing with probabilistic forecasts. It penalizes both incorrect predictions and overconfident but incorrect predictions. For instance, if a model predicts a high probability (e.g., 0.9) for a win but the actual outcome is a loss, the Brier Score penalizes this mistake more than if the model had predicted a lower probability (e.g., 0.6). This makes the Brier Score especially valuable in evaluating how well-calibrated a model is, as it assesses not just whether the model got the prediction right, but also how confident it was in making that prediction. This ability to penalize overconfidence makes the Brier Score a critical tool for assessing model calibration, especially in contexts like sports betting, where predicted probabilities hold significant importance.

2.4 Cross-Entropy (Log Loss)

While the Brier Score provides a useful assessment of the accuracy and calibration of probabilistic predictions, it treats all errors proportionally, penalizing incorrect predictions based on how far the predicted probability is from the actual outcome.
However, Cross-Entropy, also known as Logarithmic Loss (Log Loss), takes this a step further by imposing a heavier penalty on confident but incorrect predictions. While both metrics assess the accuracy of probabilistic forecasts, Cross-Entropy is stricter because it magnifies the cost of overconfidence, particularly when the predicted probability is close to 1 or 0 and the outcome is opposite. For instance, if a model predicts a probability of 0.9 for a win but the actual outcome is a loss, Cross-Entropy penalizes this mistake much more severely than the Brier Score would. This is due to the logarithmic nature of the loss function, which disproportionately penalizes predictions that are both wrong and highly confident. Thus, Cross-Entropy serves as a valuable tool in ensuring not only that the model is correct but also that it remains cautious in its predictions when uncertainty is high. Cross-Entropy is defined as:

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}[o_i \log(p_i) + (1 - o_i)\log(1 - p_i)] \tag{5}$$

where:

- p_i = predicted probability of the outcome,
- o_i = actual outcome (1 for win, 0 for loss),
- N = total number of predictions.

A lower Cross-Entropy value indicates better performance, with 0 being the ideal. In our case, this metric is critical because it emphasizes the importance of calibrated probabilities—predictions that match the true likelihood of an outcome occurring. This is especially important in sports betting, where accurate probability estimation is key.

3 Analysis of the First Season (2020)

In this section, we present the performance results of our two models evaluated on the first season. The season contains fewer data points due to the shortened schedule caused by the COVID-19 pandemic. Below, the results for both models are tabulated:

3.1 Results for First Season

Metric	Model 1	Model 2
Accuracy	0.61	0.63
ROC-AUC	0.66	0.68
Brier Score	0.27	0.25
Cross-Entropy	0.67	0.65

Table 1: Performance Metrics for First Season (Model 1 vs. Model 2)

As seen from the results, Model 2 slightly outperforms Model 1 in every metric. While both models show relatively similar accuracy, the ROC-AUC of Model 2 is slightly better, indicating a better discriminatory power. The Brier Score for both models is quite close, but Model 2 is marginally better calibrated. Cross-Entropy values, which penalize overconfidence more heavily, indicate that Model 2 handles uncertainty better than Model 1.

3.2 Plots and Analysis

ROC Curve:

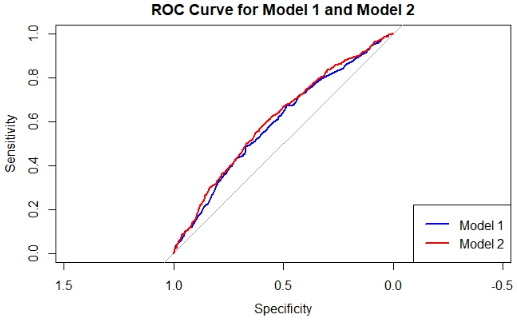

Figure 2: ROC Curve for Model 1 and Model 2 - First Season

The ROC curve for both models (Figure 2) illustrates the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) across various thresholds. The ROC curve of Model 2 is slightly higher than that of Model 1, reflecting the superior ROC-AUC score. This means that Model 2 is better at distinguishing between wins and losses across all thresholds.

However, the curvature of both ROC curves is not very pronounced, indicating that neither model has particularly strong discriminatory power. In other words, while both models are better than random guessing, their ability to correctly classify wins and losses is only moderately better than average. A more pronounced curvature would suggest a stronger ability to differentiate between true positives (wins) and false positives (losses) with higher confidence across a wider range of thresholds.

Brier Score Distribution:

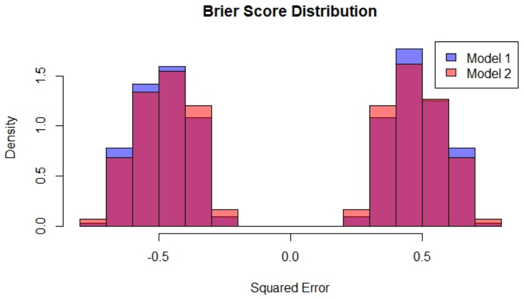

Figure 3: Brier Score Distribution - First Season

Figure 3 shows the Brier Score distribution for both models. The Brier Score measures the squared error between the predicted probabilities and the actual outcomes, where a lower score indicates better calibration. As the plot demonstrates, both models exhibit similar distributions, but Model 2 performs slightly better. This is evident because Model 2's squared errors are more concentrated around zero, indicating that its predicted probabilities are closer to the actual outcomes. In contrast, Model 1 shows slightly larger squared errors, suggesting that its predictions are less well-calibrated. Thus, Model 2 is better calibrated in this instance, as its predictions are generally more accurate in terms of probability estimates.

Calibration Plot:

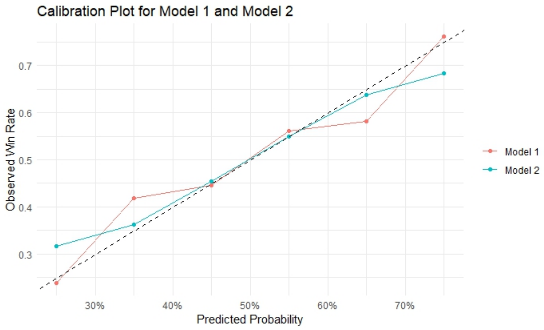

Figure 4: Calibration Plot for Model 1 and Model 2 - First Season

The calibration plot (Figure 4) shows how well the predicted probabilities align with the actual observed win rates. A well-calibrated model will have its points lie on or near the diagonal reference line, meaning that if the model predicts a 70% chance of a win, then approximately 70% of those games should result in a win. In the plot, both Model 1 and Model 2 perform similarly within the 35% to 65% probability range, indicating that both models are well-calibrated when they predict probabilities in this middle range. This suggests that both models can handle moderately uncertain predictions reasonably well.

However, noticeable differences arise in the lower and higher probability intervals. Model 2 shows significantly better calibration between 30% and 40% and between 60% and 70%, where its predicted probabilities are much closer to the observed win rates compared to Model 1. This means that Model 2 is better at estimating the win likelihood when it predicts probabilities in these ranges. For instance, when Model 2 predicts a win probability of around 65%, it aligns more closely with the actual outcome compared to Model 1.

On the other hand, Model 1 performs slightly better when it is highly confident, particularly for win probabilities above 70%. This suggests that when Model 1 is certain about a win (with a probability over 70%), its predictions tend to be more accurate than Model 2's in this region. Overall, while both models are reasonably calibrated, Model 2 outperforms Model 1 in key ranges, particularly in the intermediate probability intervals, while Model 1 excels when it is highly confident about an outcome.

Cumulative Accuracy Over Time:

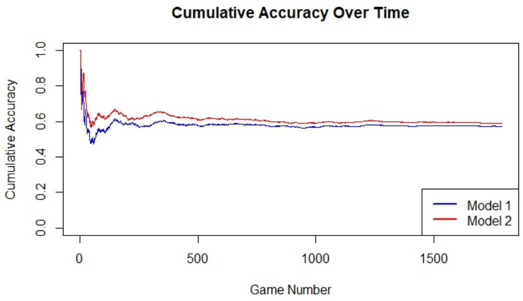

Figure 5: Cumulative Accuracy Over Time - First Season

Figure 5 presents the cumulative accuracy of both models over time. Initially, the accuracy fluctuates as fewer games are played, but as more data is accumulated, the accuracy stabilizes. Model 2 consistently maintains higher accuracy over most of the season compared to Model 1, with no overlaps in the graph, indicating that Model 2 performs slightly better throughout the entire season.

However, as the season progresses, the gap between the two models narrows, with their cumulative accuracy values converging by the end of the season. This convergence suggests that both models eventually achieve similar predictive performance as they gather more data and refine their predictions.

Predicted Probability vs Actual Outcome:

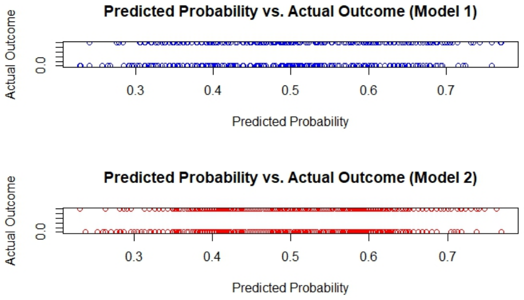

Figure 6: Predicted Probability vs. Actual Outcome - First Season

Lastly, the scatter plots in Figure 6 compare predicted probabilities with actual outcomes. For both models, the closer the predictions are to the top and bottom of the graph, the more confident the model is. While both models have clusters near 0.5, indicating uncertainty, Model 2 demonstrates slightly better alignment with the actual outcomes. This reinforces its slightly better performance across other metrics.

In summary, Model 2 consistently outperforms Model 1 in most aspects, albeit by a small margin. Both models demonstrate reasonable predictive power, but the performance differences highlight the importance of evaluating probabilistic predictions using a combination of metrics such as ROC-AUC, Brier Score, and Cross-Entropy.

4 Analysis of the Second Season (2021)

In this section, we present the results of Model 1 and Model 2 for the 2021 season. This season provided a more typical schedule with a larger number of games, allowing for a more robust evaluation of the models' performance.
ROC Curve:

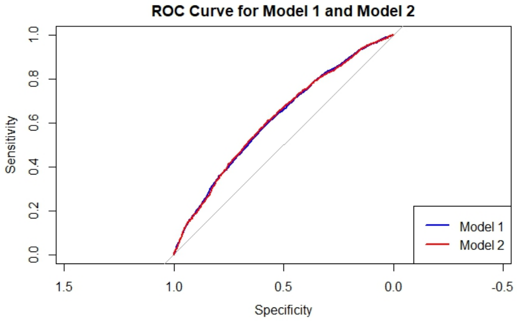

Figure 7: ROC Curve for Model 1 and Model 2 - Second Season

As seen in Figure 7, both models exhibit similar performance, with Model 2 slightly outperforming Model 1. The ROC curves remain close together, and the difference between the two models in terms of discriminatory power is minimal, reflecting a small advantage for Model 2.

Brier Score Distribution:

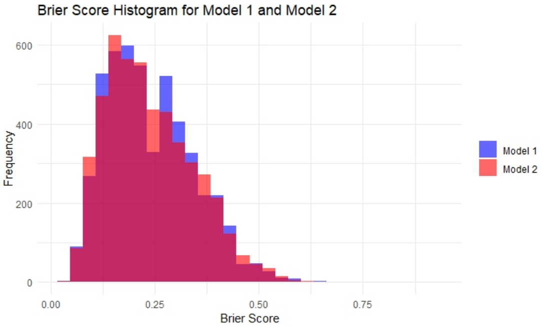

Figure 8: Brier Score Distribution - Second Season

Figure 8 shows the Brier Score distributions for both models. Once again, Model 2 demonstrates slightly better calibration, as its scores are more concentrated near lower squared errors. However, the difference between the two models is not as pronounced as in the previous season, with both models performing similarly in terms of their Brier Scores.

Calibration Plot:

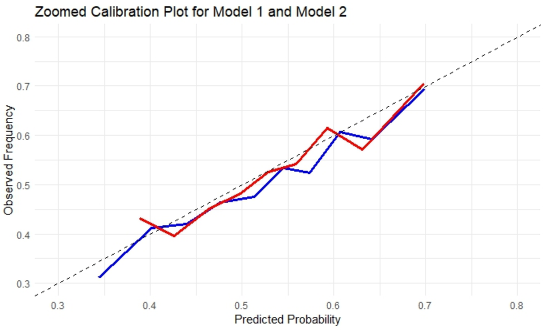

Figure 9: Calibration Plot for Model 1 and Model 2 - Second Season

The calibration plot for the 2021 season (Figure 9) shows a similar pattern to the first season. Both models are relatively well-calibrated, with Model 2 again slightly outperforming Model 1 in the probability ranges of 30-40% to 60-70%. The overall difference between the models remains minimal, with both models performing quite closely.

Cumulative Accuracy Over Time:

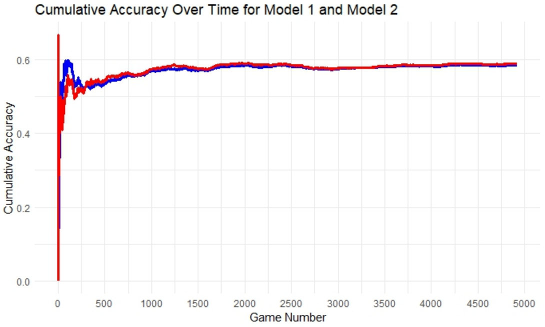

Figure 10: Cumulative Accuracy Over Time - Second Season

Figure 10 presents the cumulative accuracy of both models over time. The gap between the models is now almost negligible, with some overlaps possibly occurring too. Although Model 2 seems to be still slightly more reliable, the difference between the two models is almost non observable.

Predicted Probability vs Actual Outcome:

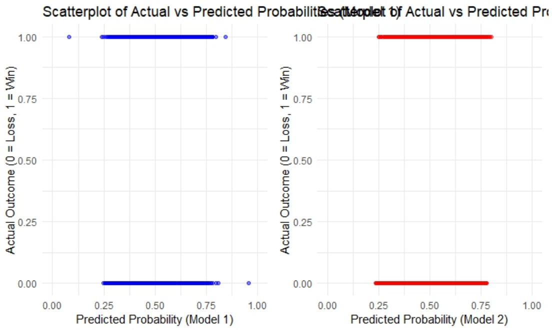

Figure 11: Predicted Probability vs Actual Outcome - Second Season

Figure 11 presents the scatter plots comparing predicted probabilities to actual outcomes for both models. As in the first season, both models demonstrate similar distributions, with no model outstanding.

4.1 Summary of Results for Second Season

In summary, while Model 2 continues to slightly outperform Model 1 across most metrics, the difference between the two models is much smaller compared to the first season. This could be attributed to the larger number of games and more stable conditions in 2021, as the season was not affected by disruptions like the COVID-19 pandemic. With a more typical schedule, the performance gap between the two models converges, suggesting that Model 1 performs better with a larger dataset, even though Model 2 still maintains a slight edge in overall performance.

Furthermore, while both models show slight improvements compared to the previous season, the improvement is negligible, which might be surprising given the increased sample size of data available in 2021. One might have expected a more noticeable improvement in the models' performance, but the results suggest that the larger dataset may not significantly enhance the models' calibration or predictive accuracy beyond a certain point.

5 Conclusion

In this study, we evaluated two predictive models across two Major League Baseball (MLB) seasons, focusing on various performance metrics including accuracy, ROC-AUC, Brier Score, and Cross-Entropy. The results showed that Model 2 consistently outperforms Model 1 across all metrics in both the COVID-affected 2020 season and the more regular 2021 season. While the differences between the models were more pronounced in the 2020 season, these gaps narrowed significantly during the 2021 season. The larger sample size and more stable conditions in 2021 likely contributed to this convergence, as both models were able to stabilize their predictions over time with more data.

Despite slight improvements in model performance from 2020 to 2021, the overall improvement is negligible, which was unexpected given the larger data set. This suggests that while more games may contribute to better model performance, it is not always the most decisive factor, as calibration and model structure could play a more significant role in prediction accuracy.

Overall, Model 2 shows slightly better discriminatory power, calibration, and accuracy across both seasons, making it the more reliable model. The fact that both models perform similarly as more data accumulates points to the importance of continued evaluation with larger datasets and perhaps further optimization of model features.

Further plots and tables summarizing the integrated model performance across both seasons are presented below, providing a comprehensive look at how the models performed over the entire dataset.

Appendix: Model Performance Summary

Tables

Table 2: Performance Metrics for Second Season (2021)

Metric	**Model 1**	**Model 2**
Accuracy	0.611	0.616
ROC-AUC	0.637	0.643
Brier Score	0.233	0.227
Log Loss	0.688	0.681

Table 3: Performance Metrics for Both Seasons (2020-2021)

Metric	**Model 1**	**Model 2**
Accuracy	0.602	0.610
ROC-AUC	0.631	0.637
Brier Score	0.237	0.232
Log Loss	0.702	0.694

Plots

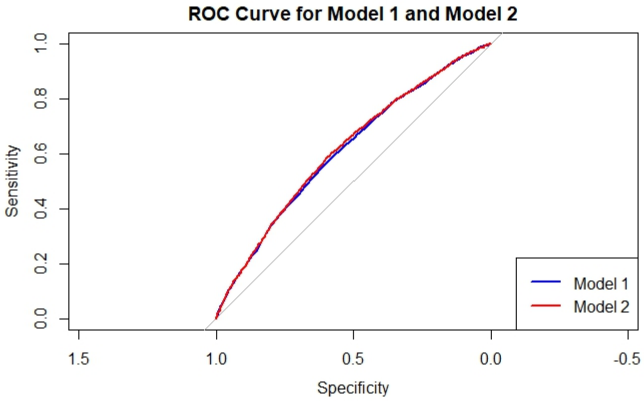

Figure 12: ROC Curve for Both Models Across Both Seasons

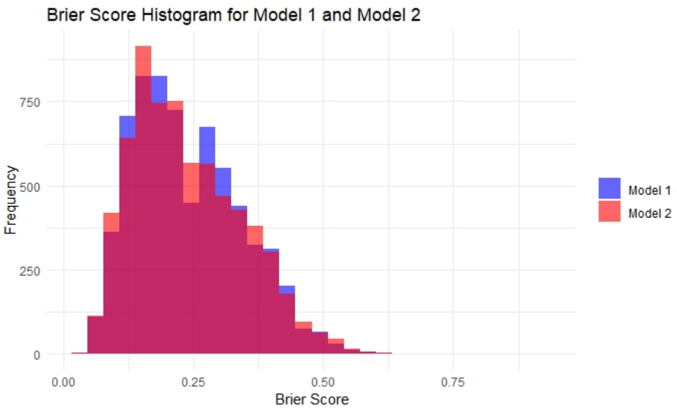

Figure 13: Brier Score Distribution for Both Models Across Both Seasons

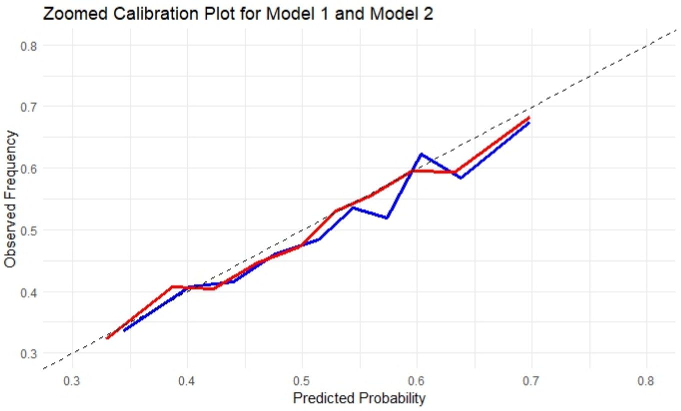

Figure 14: Calibration Plot for Both Models Across Both Seasons

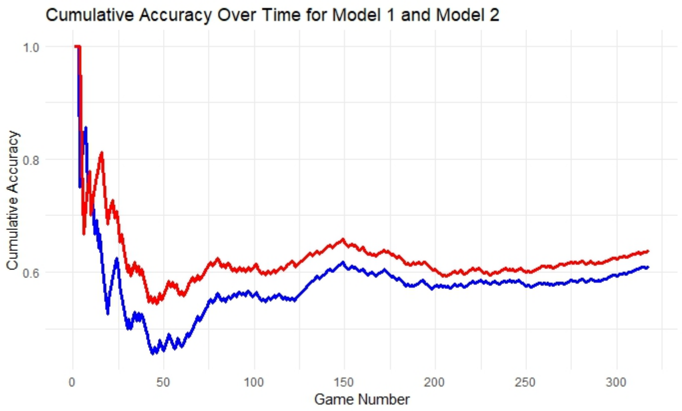

Figure 15: Cumulative Accuracy for Both Models Across Both Seasons