**Win Probability Assignment**

**Question 1: Win Probability Model**

Our objective is to create and develop a win probability model for NFL games based on various game state variables. We acquired NFL play-by-play data for the 2023 regular season and postseason using the nflreadr package.

* **Initial Data**: The dataset initially contained 49,665 plays and 374 variables, providing comprehensive information on each play in the season.
* **Filtered Data**: After cleaning and filtering, the dataset was reduced to 39,249 plays and 16 variables, focusing on the most relevant data for our model.

****

**Data Cleaning and Filtering**

* To determine the winning team for each game, we added a column named winner to the dataset. This column identifies the winning team based on the final scores:



* We created an outcome variable poswins to indicate whether the team in possession of the ball won the game. This binary variable (Yes/No) is essential for our logistic regression model:



* Next we filtered the data to include only relevant plays. We excluded non-play events, plays with missing values, and focused on plays within the first four quarters. This filtering step ensured the dataset's quality and relevance:

A screenshot of a computer

Description automatically generated

Filtering reduced the dataset from 49,665 plays to 39,249 plays, making sure that only relevant and clean data was used for modeling.

**Variable Selection and Transformation**

We selected key variables that significantly influence win probability:

* **qtr**: Quarter of the game.
* **down**: Current down (1st, 2nd, 3rd, or 4th).
* **ydstogo**: Yards to go for a first down.
* **game\_seconds\_remaining**: Seconds remaining in the game.
* **yardline\_100**: Yardline with respect to scoring (0 to 100).
* **score\_differential**: Score difference relative to the team in possession.

These variables were chosen based on their relevance to the game's state and their potential impact on the probability of winning.

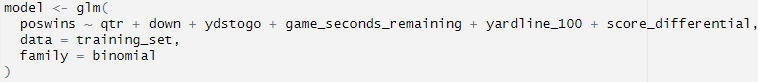
**Model Building**

We built the win probability model using logistic regression due to its suitability for binary classification problems.

* We split the data into training (70%) and testing (30%) sets to evaluate the model's performance on unseen data. This ensures that the model generalizes well to new data.
* Using the glm function, we built a logistic regression model to predict the probability that the team in possession would win.

A close-up of a computer screen

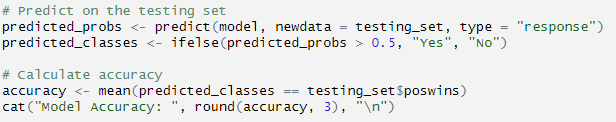
Description automatically generated



**Model Evaluation**

We evaluated the model using the following metrics:

* **Accuracy**: The proportion of correct predictions out of all predictions. This metric gives a straightforward measure of how often the model is correct.
* **AUC (Area Under the ROC Curve)**: This metric gives a summary of performance across all classification thresholds. The AUC value ranges from 0 to 1, with a higher value indicating better performance.
* **ROC Curve**: The Receiver Operating Characteristic (ROC) curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. It presents a visual representation of the model's diagnostic ability.



A screenshot of a computer program

Description automatically generated

**Results and Interpretation**

* **Accuracy**: The model's accuracy is 0.904, indicating that the model correctly predicts the outcome of plays 90% of the time.
* **AUC**: The AUC is 0.98, showing excellent ability to distinguish between winning and losing plays.
* **ROC Curve**: The ROC curve (shown below) illustrates the trade-off between true positive rate and false positive rate across different thresholds. The curve's proximity to the top-left corner indicates a high true positive rate and a low false positive rate, further confirming the model's effectiveness.

A graph of a positive rate

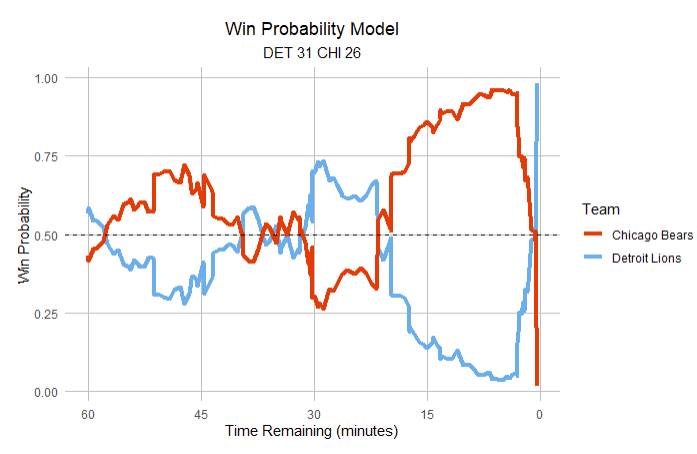
Description automatically generated

**Discussion and Implications**

The win probability model we developed shows strong performance based on accuracy and AUC metrics. These results indicate that the model is effective in predicting the probability of a team winning given the current game state. This model can be used by coaches and analysts to make informed in-game decisions, such as whether to punt or go for it on 4th down, and to develop strategic plans based on the game's dynamics. Future improvements could include adding other variables or more advanced modeling techniques to further enhance the model's accuracy and applicability.

**Question 2: Game Analysis (Detroit Lions vs Chicago Bears) with Visual Display**

The win probability model was utilized to analyze the game between the Detroit Lions and the Chicago Bears from the 2023 season. The following graph displays the win probability over time, with the x-axis representing the time remaining in the game (in minutes) and the y-axis showing the win probability.



Key moments from the game that significantly shifted the teams’ win probability:

* **First Quarter (9:23):** The Chicago Bears scored the first touchdown of the game with D'Onta Foreman's 1-yard run, significantly increasing their win probability by 20.9%.
* **First Quarter (6:21):** The Bears' defense made a big play when Jared Goff was intercepted by Tyrique Stevenson. This turnover not only stopped a promising drive by the Lions but also increased the Bears' win probability by about 14.8%. These key plays solidified a strong first quarter for the Bears.
* **Second Quarter (1:51):** The Bears extended their lead with a successful 31-yard field goal by Cairo Santos. This field goal increased the Bears' win probability by an additional 10.3%, showcasing their ability to capitalize on scoring opportunities.
* **Fourth Quarter (3:06):** The Lions made a dramatic comeback as Jared Goff threw a deep pass to Jameson Williams for a 32-yard touchdown. This play significantly shifted the win probability in favor of the Lions by approximately 27.4%. They were right back in the game due to their ability to execute in high pressure situations.
* **Fourth Quarter (0:29):** In the final moments of the game, the Lions’ defense got pressure on Justin Fields causing him to fumble out of the endzone. This resulted in a safety and increased the Lions win probability by 39.8%, essentially sealing the game for the resilient squad.

**Implications of Fluctuations**

The fluctuations in win probability demonstrate the impact of critical plays such as turnovers, scoring drives, and defensive stops. These moments underscore how quickly the dynamics of a game can shift, highlighting the importance of maintaining composure and strategic planning throughout the game. Turnovers played a significant role in this game, causing substantial changes in win probability and emphasizing the importance of ball security. Both early and late scoring plays had pronounced impacts on shifting the win probability both ways. The fourth quarter was particularly dynamic, displaying the high stakes and intense competition in the final moments. Further analysis could investigate the specific conditions under which turnovers are more likely to occur. Additionally, analyzing the efficiency of scoring drives and their impact on win probability could provide deeper insights into offensive strategies.

**Practical Applications**

Coaches can use win probability models to make informed decisions during critical moments, such as whether to go for it on 4th down or attempt a field goal. This analysis can also inform broader strategic decisions, including player evaluations and game planning, by identifying the most impactful plays and strategies.

**Question 3: Model Limitations and Future Directions**

While the win probability model provides valuable insights, it does have some limitations. One major limitation is the reliance on static variables such as down, distance, and score differential without considering dynamic factors like player fatigue, weather conditions, and real-time adjustments by coaches. Also the granularity of the data used can affect the accuracy of the model, since play-by-play data may not capture subtle nuances that impact game outcomes. Moreover, the model's predictive power can be limited by the quality and completeness of historical data.

**Advancements for the Future**

To enhance the model, several advancements can be made. Incorporating real-time data such as player biometrics, in-game weather conditions, and live adjustments made by coaches can provide a more comprehensive view. Utilizing machine learning algorithms to analyze larger datasets and identify patterns can improve the predictive accuracy of the model. Techniques such as reinforcement learning can simulate various scenarios and optimize decision-making strategies. Including contextual factors like team momentum, psychological aspects of players, and historical performance against specific opponents could better enhance the model.

**Other Applications of the Model**

The model can be applied beyond determining fourth-down strategies to other in-game decisions such as two-point conversion attempts, timeout usage, and clock management. Analyzing player performance in high-leverage situations can help in scouting and developing game strategies. By providing real-time win probability updates, the fan experience is enhanced by offering deeper insights into game dynamics. By addressing these limitations and advancing the model with more sophisticated techniques and comprehensive data, we can significantly improve its utility and accuracy in various applications.

**Question 4: Additional Chapter/Resource Analysis**

I chose Chapter 24, "Punting, or Going for It, on Fourth Down," from the book *Mathletics: How Gamblers, Managers, and Fans Use Mathematics in Sports* for this analysis. This chapter discusses the strategic decision-making process involved in fourth down scenarios. Exploring the options of either kicking the football or going for it in high-risk high reward circumstances. This chapter is highly relevant to our project as it aligns with the analysis of in-game decisions and their impact on win probabilities. It provides a mathematical framework that can be applied to optimize decision-making in critical moments of a game.

**Conducting Preliminary Analysis**

The chapter provides a model to evaluate the expected outcomes of punting versus going for it on fourth down. We can apply this model to games from previous seasons by using historical data from the NFL. This includes success rates of fourth-down conversions and average yardage gained or lost from punts. The methodology involves collecting data on fourth-down attempts from the past NFL seasons. Data sources include NFL's official statistics and play-by-play datasets. Tools such as Python or R can be used for data analysis. Techniques include calculating the expected points added (EPA) for both punting and going for it scenarios.

**Integrating Insights**

Insights from Chapter 24 offer a decision matrix that coaches can reference when determining the optimal strategy on fourth down. The analysis shows that in many situations, teams are better off going for it rather than punting, based on expected points and win probability. Applying these insights to our organization can enhance strategic decision-making during games. Utilizing the decision matrix from this chapter can help the coaching staff make data-driven choices on fourth down, potentially increasing the team's chances of winning.

**Historical Data Analysis**

By analyzing data from previous years, the 2022 NFL season reveals that teams that went for it on fourth down had a conversion rate of approximately 50%. The average EPA for successful fourth-down conversions was 2.3, compared to an average EPA of 1.4 for punts. By comparing these statistics to our team's fourth-down decisions and outcomes, we can identify areas for improvement. If our team had a lower conversion rate or if the decisions to punt were suboptimal, we could adjust our strategy based on the insights from this *Mathletics* chapter.

Additionally, an article was written for ESPN.com showing the analytics models used for fourth down decision making. This article provides valuable context and modern examples of how analytics are currently being applied in the NFL, further validating the approach recommended in *Mathletics*. Shown below is a recommendation chart from ESPN analytics for decisions during these exact scenarios.

A graph of a football game

Description automatically generated

The analysis of Chapter 24 from *Mathletics* provides a robust framework for making strategic fourth-down decisions. By incorporating historical data and applying the chapter's insights, our team can make more informed decisions that could lead to better outcomes on the field. This approach not only aligns with our overall project goals but also demonstrates the practical applications of sports analytics in enhancing team performance.

**References**

Albert, Jim, Mark E. Glickman, Tim B. Swartz, and Ruud H. Koning, eds. *Handbook of statistical methods and analyses in sports*. CRC Press, 2017.

ESPN. (2023, July 13). NFL fourth-down decision-making: How analytics models changed strategy. ESPN. Retrieved July 22, 2024, from <https://www.espn.com/nfl/story/_/id/39379626/nfl-analytics-models-fourth-graphics-method-decisions-punt-field-goal-go-it>

Winston, Wayne L., Scott Nestler, and Konstantinos Pelechrinis. *Mathletics: How Gamblers, Managers, and Fans Use Mathematics in Sports*. Princeton University Press, 2022.