Predictive Analysis of NFL Quarterback Performance

Introduction:

In professional football, quarterback performance is often analyzed more than any other position. Quarterbacks are central to their teams' offensive strategy, directly influencing game outcomes. This project aims to delve into quarterback performance by examining how certain measurable attributes—specifically, the team a quarterback plays for, their number of passing completions, attempts, and passing yards per game—can predict the number of touchdown passes in a season.

Dataset Overview:

Our analysis utilizes a comprehensive dataset comprising of the season's performance metrics for the NFL quarterbacks. The data includes performance indicators such as completions, attempts, passing yards, touchdowns, and the teams they played for during the season. This dataset was sourced from ESPN, which provides sports statistics that collects and compiles player performance across different seasons and games.

Significance/Relevance of the Study:

Understanding what factors contribute to a quarterback's touchdown-scoring ability not only helps teams make informed decisions regarding player drafts, trades, and training focuses but also helps coaches in game planning and strategy formulation. Additionally, this analysis helps fantasy football enthusiasts and sports analysts who rely on these predictive insights to make recommendations or create their ideal team.

Research Objectives:

Our Research Question: *Does the team a quarterback plays for, along with their number of completions, passing attempts, and passing yards per game, predict the number of touchdown passes that quarterback will have in a season?*

By using predictive modeling techniques, this project aims to:

- Identify key predictors of quarterback touchdown performance.
- Develop a model that accurately predicts touchdown passes based on pre-season and in-season metrics.
- Evaluate the effectiveness of different statistical models when analyizing quarterback performance in the NFL.

The project will use advanced statistical methods and predictive models to provide insight into the predictive aspects of sports analytics. These insights will help understand player performance in professional football, with potential uses in team management and fantasy football strategies.

Data Preparation:

For this analysis, I focused on NFL quarterback performance metrics to predict touchdown passes. This section outlines the steps to clean, prepare, and visualize the dataset for modeling.

Data Acquisition:

The dataset was sourced from ESPN's official NFL statistics page, which includes performance data for quarterbacks during the 2024 season. This dataset contains basic statistics like player names and teams as well as in-depth performance metrics such as completions, attempts, passing yards, touchdowns, and more.

Data Wrangling and Cleaning:

The dataset required several steps to ensure accuracy and usability in our models:

- **HTML Table Extraction**: The data, structured in HTML tables, was pulled using the rvest package in R and then parsed into a usable format.
- **Data Cleaning and Formatting**:
    - Player names and team affiliations were embedded in a single column, so it needed to be separated. This was achieved using string manipulation functions from the tidyverse package, in which I was able to extract the player and team.
    - The passing yards data, initially formatted as strings with commas, were converted into numeric types to help with our analysis.

A view of our dataset changes after these steps:

```
## # A tibble: 6 × 8
##   Player    Team  Completions Attempts PassingYards YardsPerGame YardsPerAttempt
##   <chr>     <chr>       <int>    <int> <chr>               <dbl>           <dbl>
## 1 Lamar Ja… BAL           254      379 3,290               253.             8.7
## 2 Jared Go… DET           276      381 3,265               251.             8.6
## 3 Sam Darn… MIN           264      386 3,299               254.             8.5
## 4 Joe Burr… CIN           302      446 3,337               278.             7.5
## 5 Tua Tago… MIA           240      325 2,456               273.             7.6
## 6 Russell … PIT           138      213 1,784               255.             8.4
```

- **Handling Missing Values**: Any missing or incomplete records were inspected. Given the completeness of the dataset directly from a reliable source like ESPN, no significant imputations or deletions were necessary.

I then summarized the data to get an understanding of the distribution of each variable:

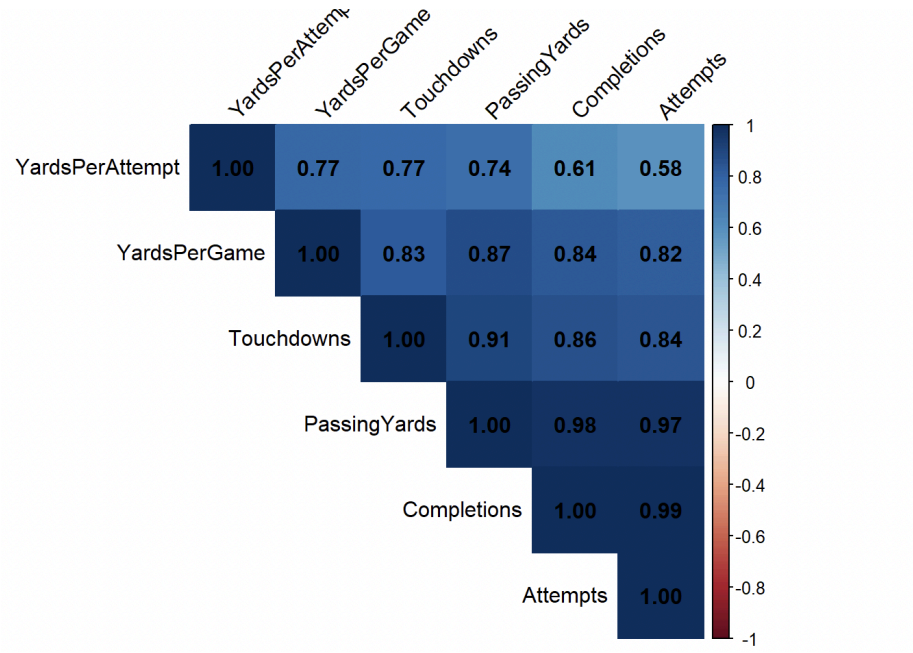
```
##     Player              Team            Completions     Attempts
##  Length:45          Length:45          Min.   : 21   Min.   : 31.0
##  Class :character   Class :character   1st Qu.:112   1st Qu.:185.0
##  Mode  :character   Mode  :character   Median :203   Median :306.0
##                                        Mean   :191   Mean   :292.2
##                                        3rd Qu.:270   3rd Qu.:393.0
##                                        Max.   :324   Max.   :466.0
##   PassingYards   YardsPerGame   YardsPerAttempt   Touchdowns
##  Min.   : 189   Min.   : 82.8   Min.   :4.600   Min.   : 0.00
##  1st Qu.:1148   1st Qu.:167.9   1st Qu.:6.300   1st Qu.: 7.00
##  Median :2070   Median :211.2   Median :6.900   Median :13.00
##  Mean   :2104   Mean   :203.0   Mean   :6.964   Mean   :12.78
##  3rd Qu.:3032   3rd Qu.:247.6   3rd Qu.:7.600   3rd Qu.:17.00
##  Max.   :3474   Max.   :278.1   Max.   :8.700   Max.   :30.00
```

TouchDown Statistics Summary:

```
## # A tibble: 1 × 5
##   mean_td median_td sd_td min_td max_td
##     <dbl>     <dbl> <dbl>  <dbl>  <dbl>
## 1    12.8        13  8.15      0     30
```

Then, I created a correlation matrix to visualize how each variable relates to each other:

Correlation Matrix of Quarterback Statistics:



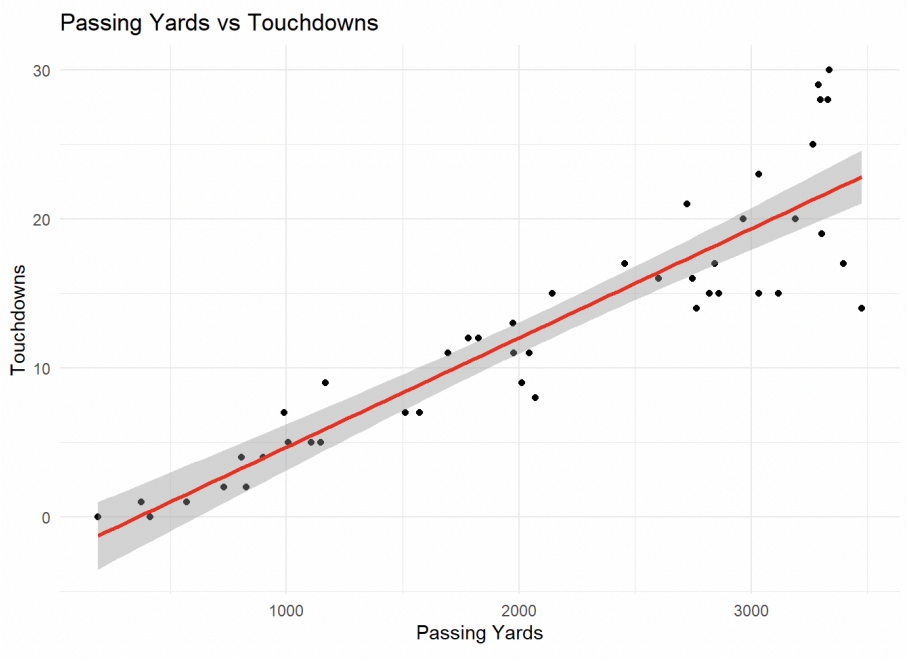
To help us prepare for the model training and validation:

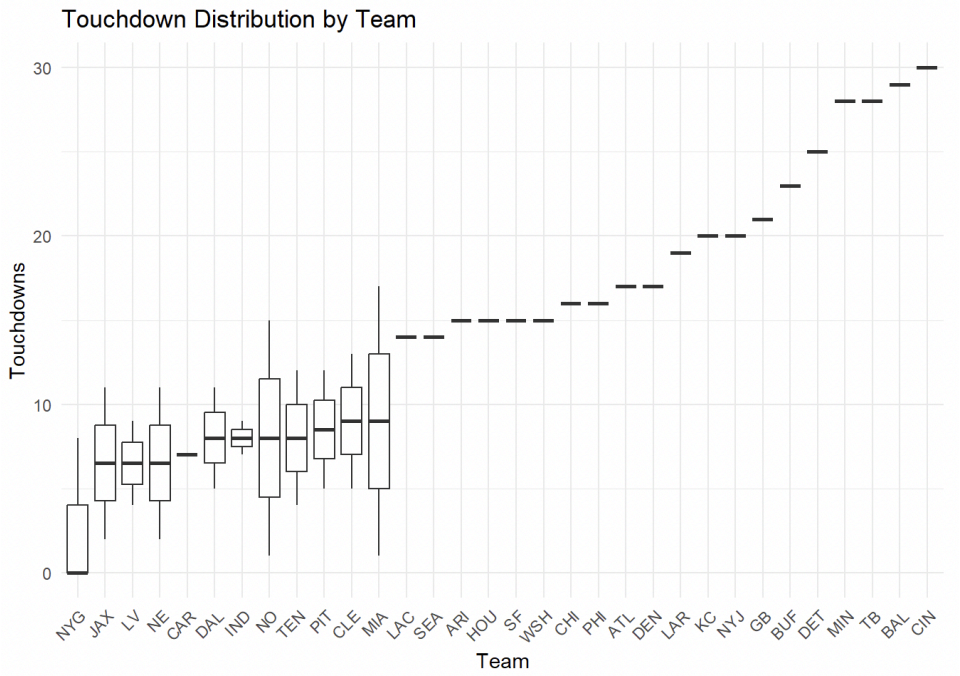
Data Splitting:

- The dataset was divided into a training set:
    - 70% of the data and a testing set 30% for the Decision Tree Model
    - 80% of the data and a testing set 20% for the Random Forest Model

- This split ensures robust model training while keeping a subset to evaluate/compare model performance.

After cleaning, an analysis was conducted to understand distributions and potential correlations among the variables. The corrplot package was used to visualize correlations, which helped in identifying multicollinearity and understanding the relationships between variables.





Model Training and Validation:

This section outlines the models considered, the training process, and the validation methods used to ensure the accuracy and robustness of our predictions.

Selection of Predictive Models

Given the continuous nature of our outcome variable (number of touchdown passes), regression models were deemed most appropriate. The two main models selected for comparison were:

- **Decision Tree Regression**: This model was chosen for its interpretability and ease of understanding. Decision trees split the data into subsets based on feature value conditions, making it straightforward to follow how predictions are made.
- **Random Forest Regression**: Random Forest model combines the results from multiple decision trees, with each tree trained on different subsets of the data with different subsets of the features. This helps with reducing the variance when compared to one decision tree, making the Random Forest predictions more reliable. Additionally, the aggregation of predicitons helps with minimizing noisy outliers, making predictions more constant.

Model Training Process

Both models were trained using the following steps:

- **Data Splitting**: Utilizing the caret package, the dataset was split into training (70% for decision tree and 80% for random forest) and testing (30% for decision tree and 20% for random forest) sets. This split was performed to validate the model on unknown data, and will help simulate how well the model would perform in real-world scenarios.
- **Feature Selection**: The models were trained using features determined to be most relevant from the data preparation phase such as: **completions, attempts, passing yards, yards per game, and team effects.**
- **Parameter Tuning**: For the Random Forest model, I tuned the number of trees (ntree) and the number of variables tried at each split (mtry) to optimize performance. This was achieved through grid search with cross-validation using the train function from the caret package.

Validation and Performance Metrics

The performance of each model was evaluated using these metrics:

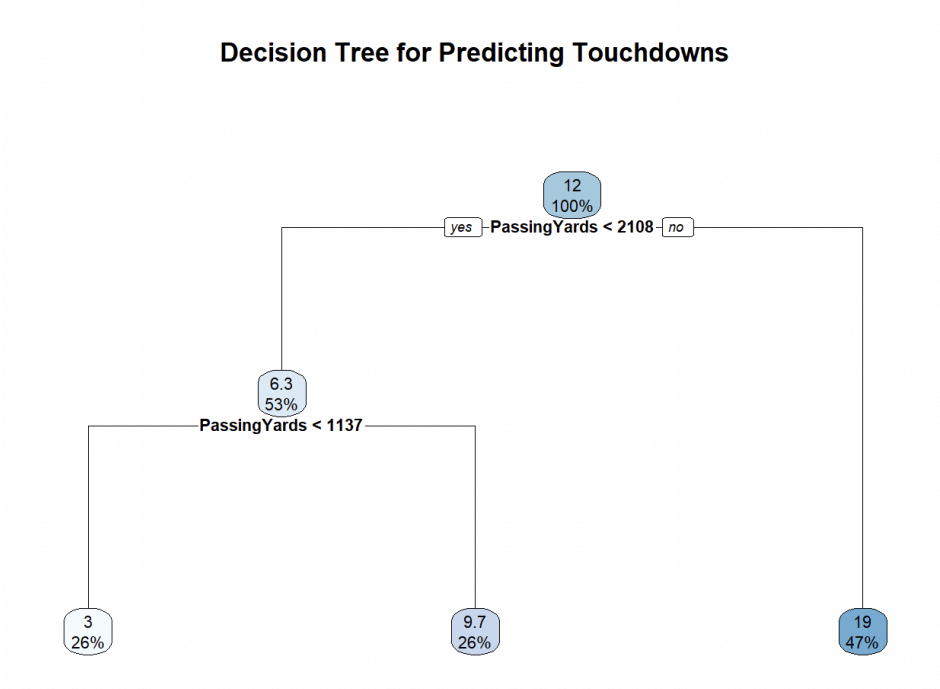
- **Mean Squared Error (MSE)** and **Root Mean Squared Error (RMSE)** were primarily used to measure the accuracy of predictions, with lower values indicating better model performance.
- **R-squared (R^2)** was used to determine how well the model's predictions matched the actual data points in terms of variance explanation.

The Results:

Model Evaluation

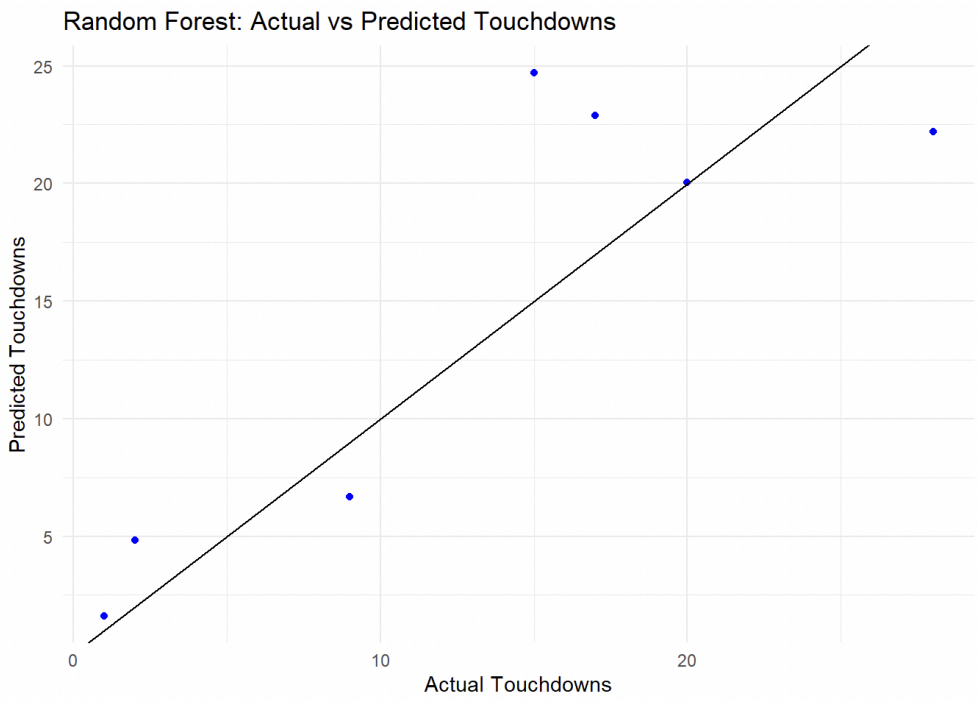
Each model was fit on the training data and then used to make predictions on the test set. The Decision Tree model was visualized to assess how decisions were made at each node, enhancing understanding of feature impacts. The Random Forest model's feature importance was analyzed to pinpoint which variables most significantly influenced touchdown outcomes.

The Decision Tree Model:



- MSE: 21.58092
- RMSE: 4.645527 (Calculated as the square root of MSE)
- R^2: 0.7180093

Random Forest Model:



MSE: 25.24918

RMSE: 5.024856 (Calculated as the square root of MSE)

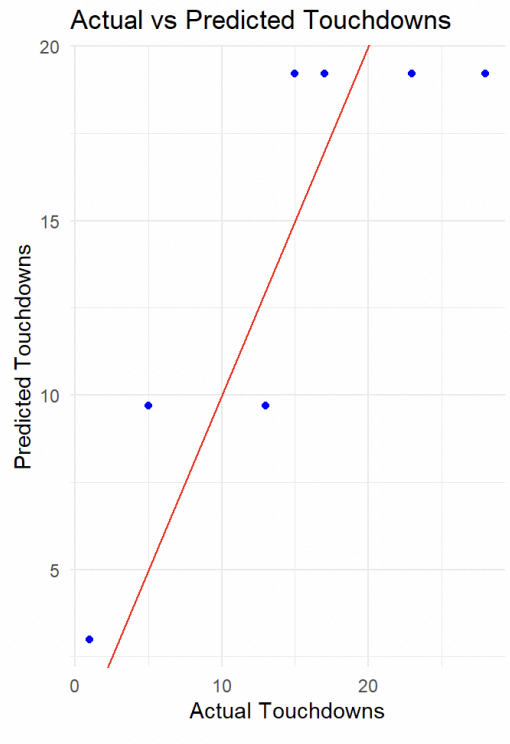
R^2: 0.6925423

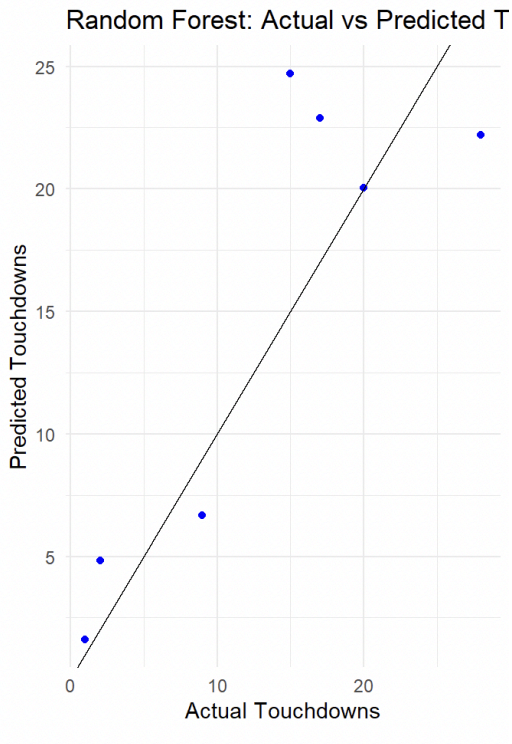
Additionally, I created a dataframe to identify the largest errors. These are the top 10 largest prediction errors.

```
##            Player Actual Predicted Difference
## 2     Brock Purdy     15      24.7        9.7
## 5    Kirk Cousins     17      22.9        5.9
## 1  Baker Mayfield     28      22.2        5.8
## 6 Jacoby Brissett      2       4.8        2.8
## 4      Joe Flacco      9       6.7        2.3
## 7   Tyler Huntley      1       1.6        0.6
## 3 Patrick Mahomes     20      20.0        0.0
```

Comparative Analysis

- The performance of each model was directly compared to assess which model provided the best balance of accuracy and complexity.
- Validation techniques, such as removal of outliers, ensured that the models were not overfitting when applied to new data.





Conclusion from Model Training:

MSE Analysis:

- **Decision Tree Model** exhibits an MSE of 21.58092, whereas the **Random Forest Model** shows an MSE of 25.24918. Contrary to the typical expectations of a Random Forest Tree being a more robust model, the Decision Tree model has a lower MSE, suggests that the Decision Tree model's predictions are generally closer to the actual touchdown scores. This lower value indicates that it might be the better model.

RMSE Analysis:

- The **Decision Tree Model** has an RMSE of 4.645527, which is lower than the **Random Forest Model's** RMSE of 5.024856, meaning that Decision Tree model is more precise, as its predictions deviate less from the observed values. The lower RMSE means the Decision Tree model has a higher accuracy to predict the number of touchdowns.

R^2 Analysis:

- When examining the R^2 values (the proportion of variance for a dependent variable that's explained by the independent variables in the model), the Decision Tree model, again, shows a higher value (0.7180093) compared to the Random Forest model (0.6925423). This higher R^2 value means the Decision Tree model can better account for the variability in quarterback touchdown passes with the variables provided.

Analysis:

Which model is better?:

The analysis of MSE, RMSE, and R^2 indicated that the **Decision Tree mode**l, in this case, outperforms the Random Forest model in predicting the number of touchdown passes. This could be due to several reasons, including the nature of the data, the interactions among the variables, or the Decision Tree model's ability to handle specific nuances in this particular dataset more effectively.

This outcome challenges the common assumption that more complex models, such as Random Forests, always produce better results, emphasizing the importance of testing various models on specific datasets.

The results from this analysis are crucial for further strategic decisions in team management, betting analysis, and fantasy football drafts, providing a robust basis for predicting quarterback performance based on observed historical data.

About this Analysis:

- The analysis conclusively shows that the team environment and quarterback performance metrics such as completions, attempts, and passing yards per game are significant predictors of the number of touchdown passes.
- For teams and sports analysts, this information is crucial. It confirms that evaluating these specific metrics can help predict quarterback performance, thereby aiding in strategic decisions regarding player selection and game strategy.
- For fantasy football players, these insights provide a statistical basis to choose quarterbacks who are likely to perform well, based on predictive factors rather than merely historical performance.

Reflection:

One of the most significant challenges I encountered was the data preparation. Ensuring data integrity and trying to determine the best way to clean this data required us to have not only an in-depth understanding of our data, but also the tools I used to prepare it. Additionally, choosing the right model to address our research question and making sure I implemented best statistical standards was something I was very careful about.

What I have learned:

- Random Forests models might not always produce better results. I know now the importance of testing various models on specific datasets.
- I have become proficient in transforming raw data into predictive models with important insights.

References:

Decision Trees: https://www.geeksforgeeks.org/decision-tree/

Comparing Decision Trees and Random Forest: https://www.geeksforgeeks.org/difference-between-random-forest-and-decision-tree/

Random Forest Trees: https://www.ibm.com/topics/random-forest

Data From ESPN: https://www.espn.com/nfl/stats/player/_/season/2024/seasontype/2/table/passing/sort/QBRating/dir/desc