Orange Hoops: Player Performance

Wei-Chieh Chen Peiya Qi Yash Gaikwad

2024-11-13

Abstract

This project leverages NCAA Men's Basketball data from the 2023-2024 season to identify Zach Edey as the optimal Purdue Men's Basketball player for taking game-deciding shots during clutch moments. Clutch situations are defined as the last five minutes of the second half with a score differential within five points. Using an array of machine learning models, including Logistic Regression, Ridge Regression, Support Vector Machine (SVM), and ensemble methods, SVM emerged as the top-performing model, achieving 66.7% accuracy. The final SVM model recommends players for clutch scenarios, with Edey emerging as the top choice based on his consistently high shot success probabilities across various game situations.

Introduction

In NCAA basketball, the final moments of close games are pivotal, and deciding which player should take the last shot can significantly impact the outcome. This study aims to offer a data-driven recommendation for the optimal player on the Purdue Men's Basketball team to take a winning shot in high-pressure moments. By analyzing player performance under specific game conditions—such as time remaining, score differential, and possession—the project seeks to develop a reliable model to predict shot success during clutch situations. This model will ultimately provide actionable insights for in-game decision-making.

Data Preparation

Clutch-time data from the NCAA Men's Basketball Division 1 2023-2024 season was filtered for analysis, focusing on situations in the final five minutes of the game where the score difference was within five points. Variables included score_diff, secs_remaining, win_prob, shooter, and others related to game conditions and player actions. The outcome variable, shot_success, was binary, marking each shot as "Success" or "Fail".

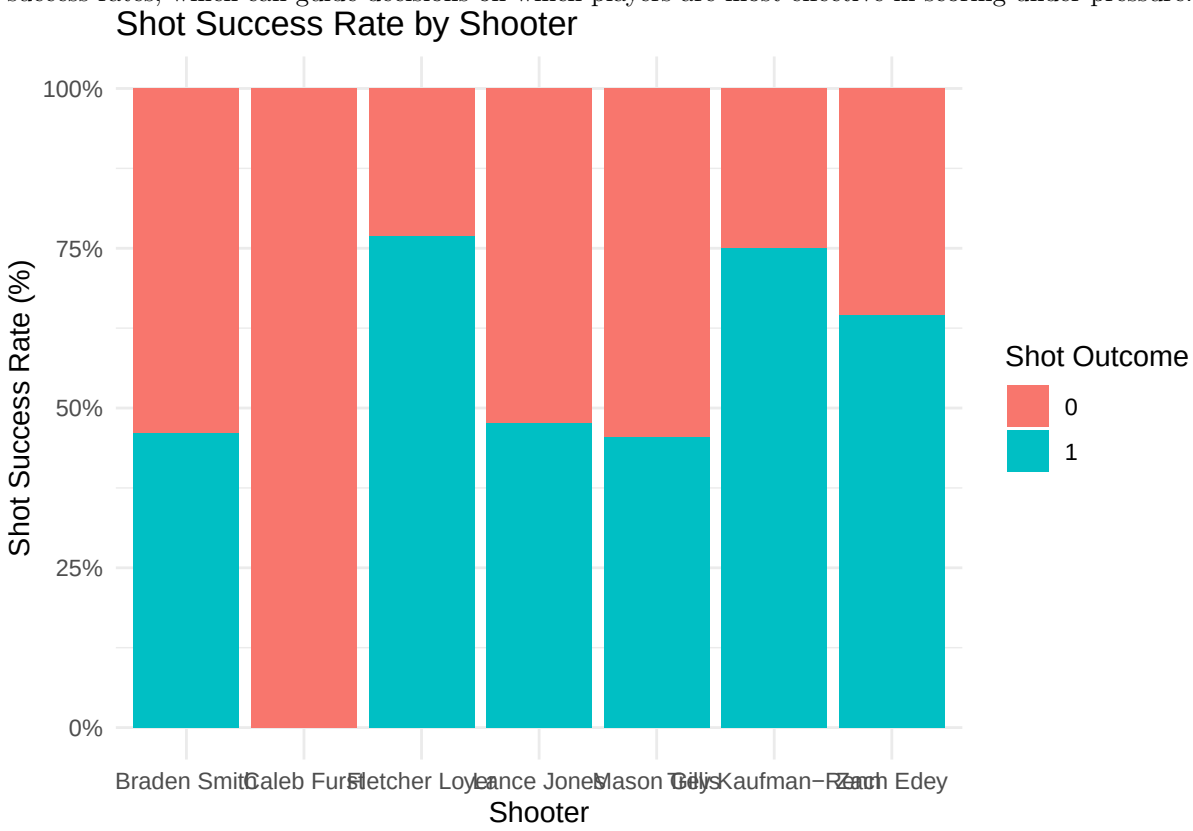
```
# Define clutch time: last 10 minutes of the second half with score difference between -5 and 5
clutch_data <- performance %>%
  filter(
    half == 2,
    secs_remaining <= 300,
    score_diff >= -5 & score_diff <= 5,
    shot_team == "Purdue"  # Filter for shots made by Purdue
  )

# Drop rows with any NA values in relevant columns for modeling
clutch_data <- clutch_data %>%
  drop_na(secs_remaining, score_diff, shot_outcome, shooter, three_pt, free_throw, possession_before, p

# Create shot success as the outcome variable
clutch_data$shot_success <- ifelse(clutch_data$shot_outcome == "made", 1, 0)

# Display data structure
# str(clutch_data)
```

Method & Result

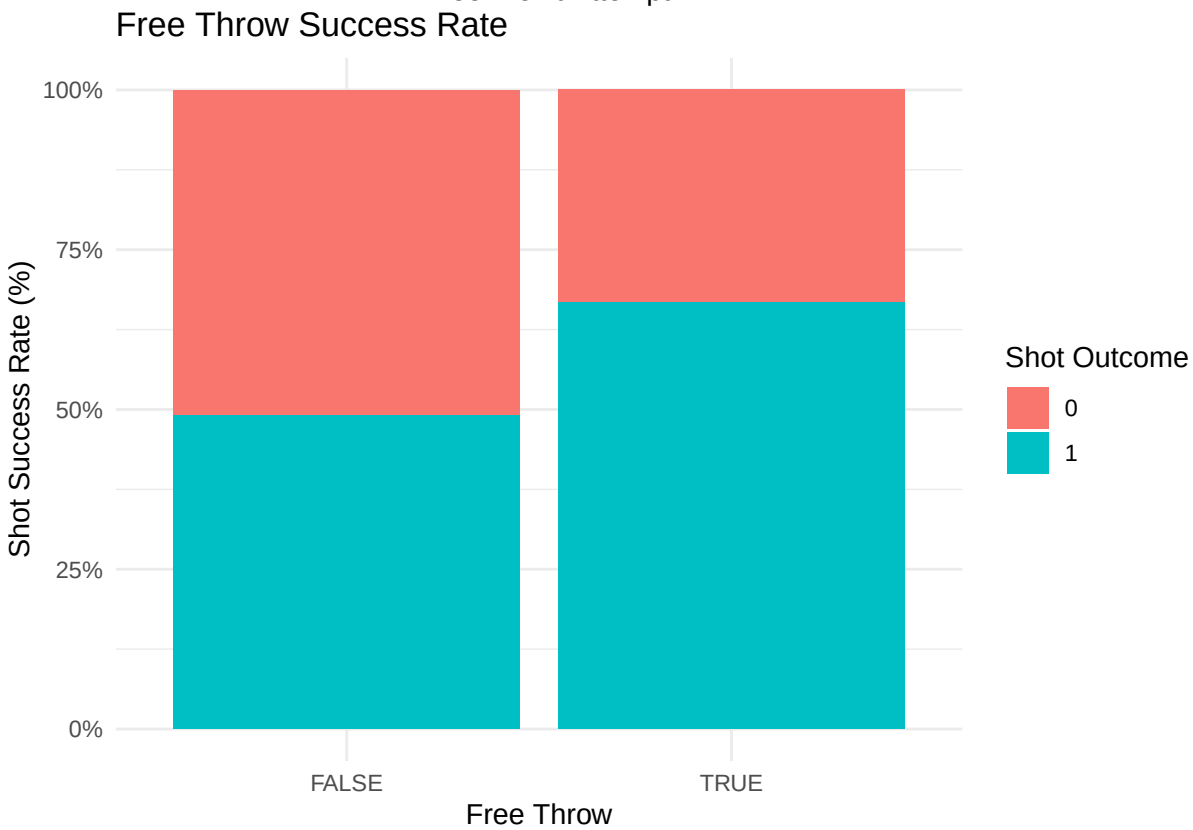
Exploratory Data Analysis

The exploratory data analysis (EDA) provides visual insights into the factors influencing shot success under clutch conditions. The following visualizations were used to understand player performance and contextual factors:

1. Distribution of Shot Success by Shooter: This bar plot shows the percentage of successful versus failed shots for each player. A higher proportion of success in any player's bar indicates a greater reliability in clutch situations. This plot helps to quickly assess individual players' shot success rates, which can guide decisions on which players are most effective in scoring under pressure.



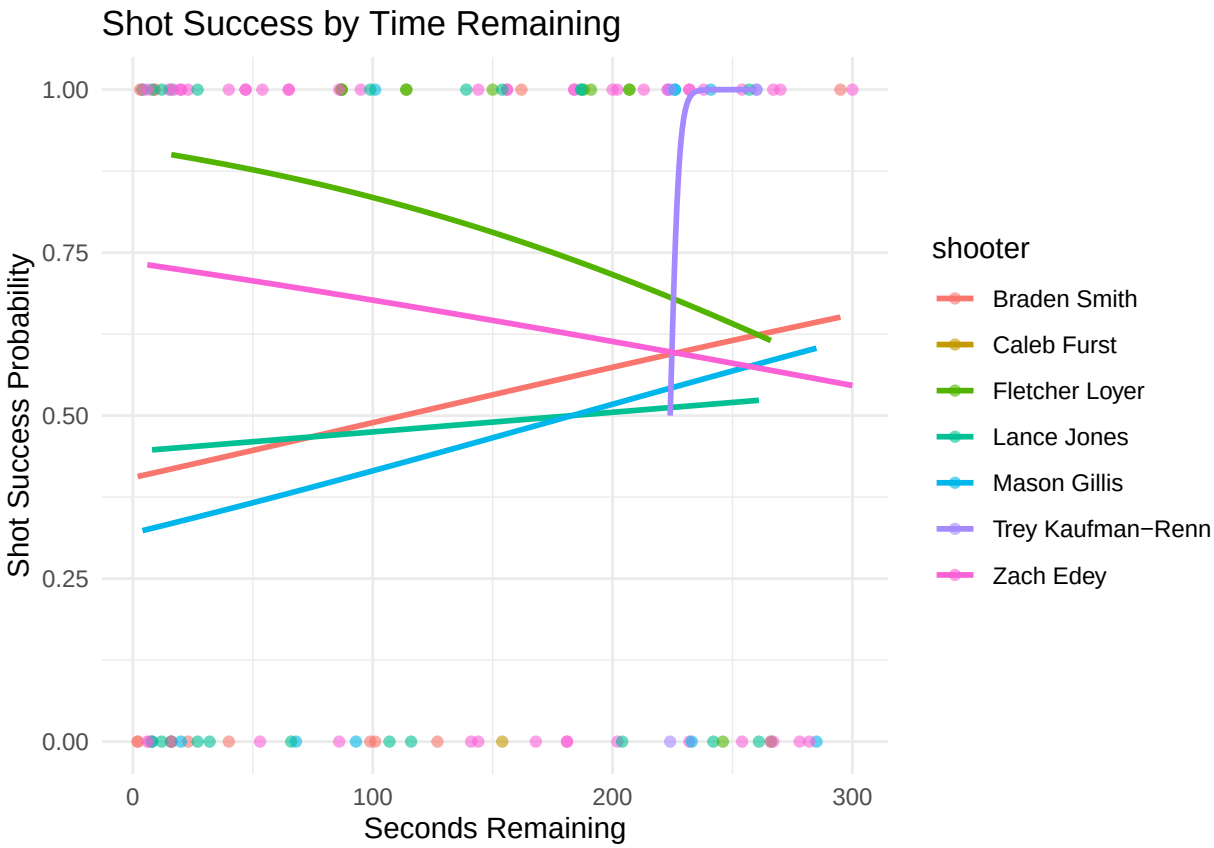
2. Distribution of Three-Point Shots and Free Throws: The plots for Three-Point Attempt Success Rate and Free Throw Success Rate provide insights into players' strengths with specific types of shots. A high success rate in the three-point plot indicates a player who is effective from beyond the arc, while a high success rate in the free throw plot highlights players who are reliable at the line. This breakdown is useful for understanding which shot types maximize each player's scoring potential.



3. Time Remaining and Shot Success: This scatter plot, with a trend line, shows how shot success probability changes as time counts down in the last five minutes of the game. If the trend line remains

steady or increases, it indicates that players maintain or improve shot success as time decreases, showcasing resilience under time pressure. Conversely, a downward trend could indicate decreasing accuracy as pressure builds, suggesting time sensitivity in shot effectiveness.

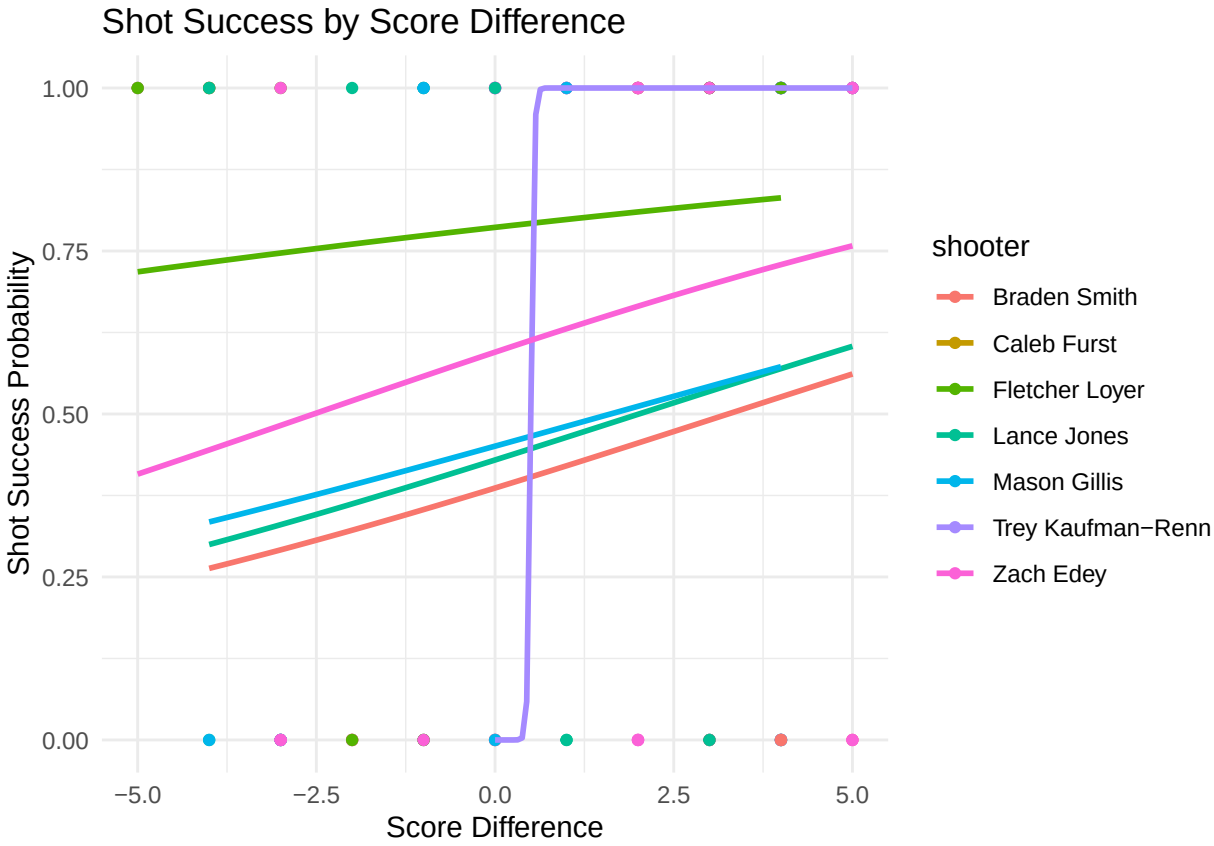
```
## `geom_smooth()` using formula = 'y ~ x'
```



4. Shot Success by Score Difference: In this plot, shot success is analyzed against the score difference between teams. A consistent trend line across different score differences suggests that players' shooting accuracy is stable regardless of the game's competitiveness. A steep slope could imply that players perform differently based on the closeness of the game, either thriving under pressure or becoming less effective as games get tighter.

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

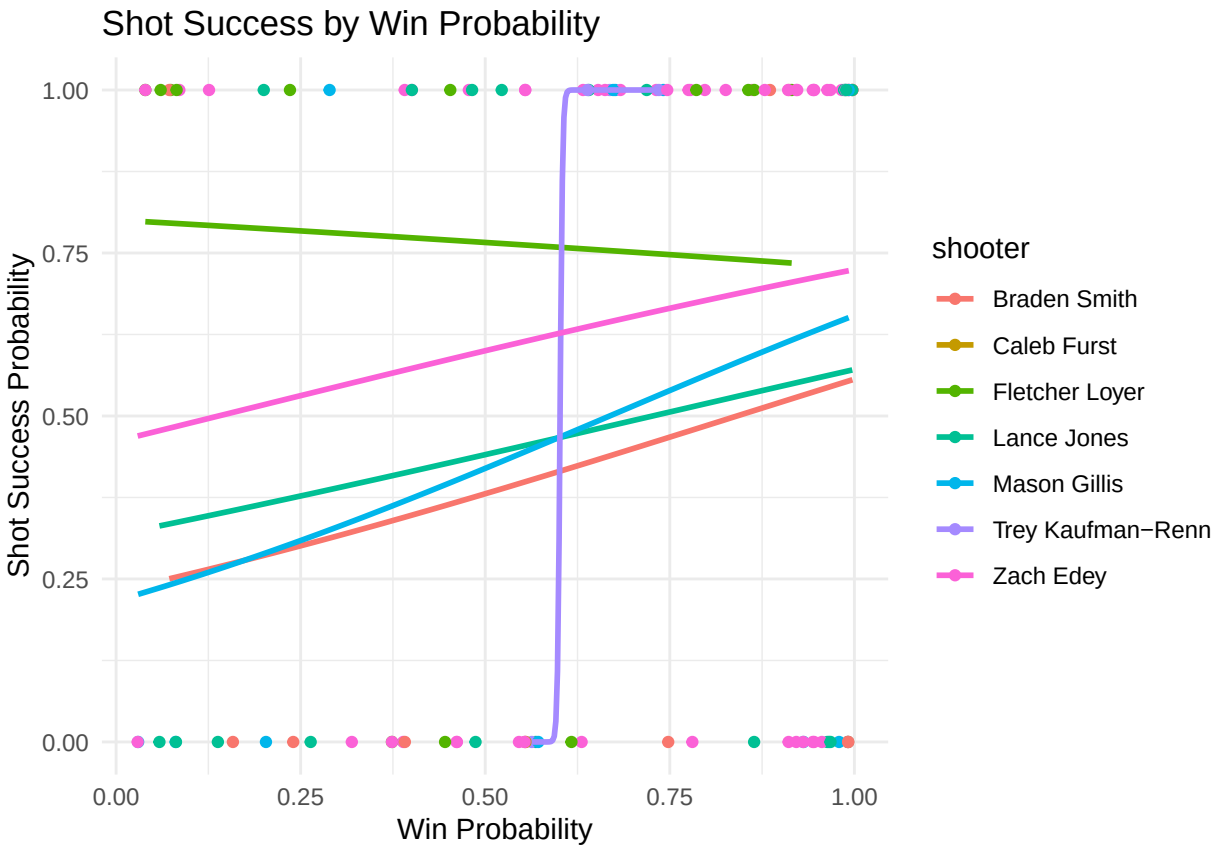


5. Shot Success Rate by Game Situation: The Shot Success by Win Probability plot reveals how players perform relative to the predicted chances of their team winning. A stable trend here suggests that players' effectiveness is independent of the game's expected outcome, while variability could indicate that certain players perform better when the odds are either in or against their favor.

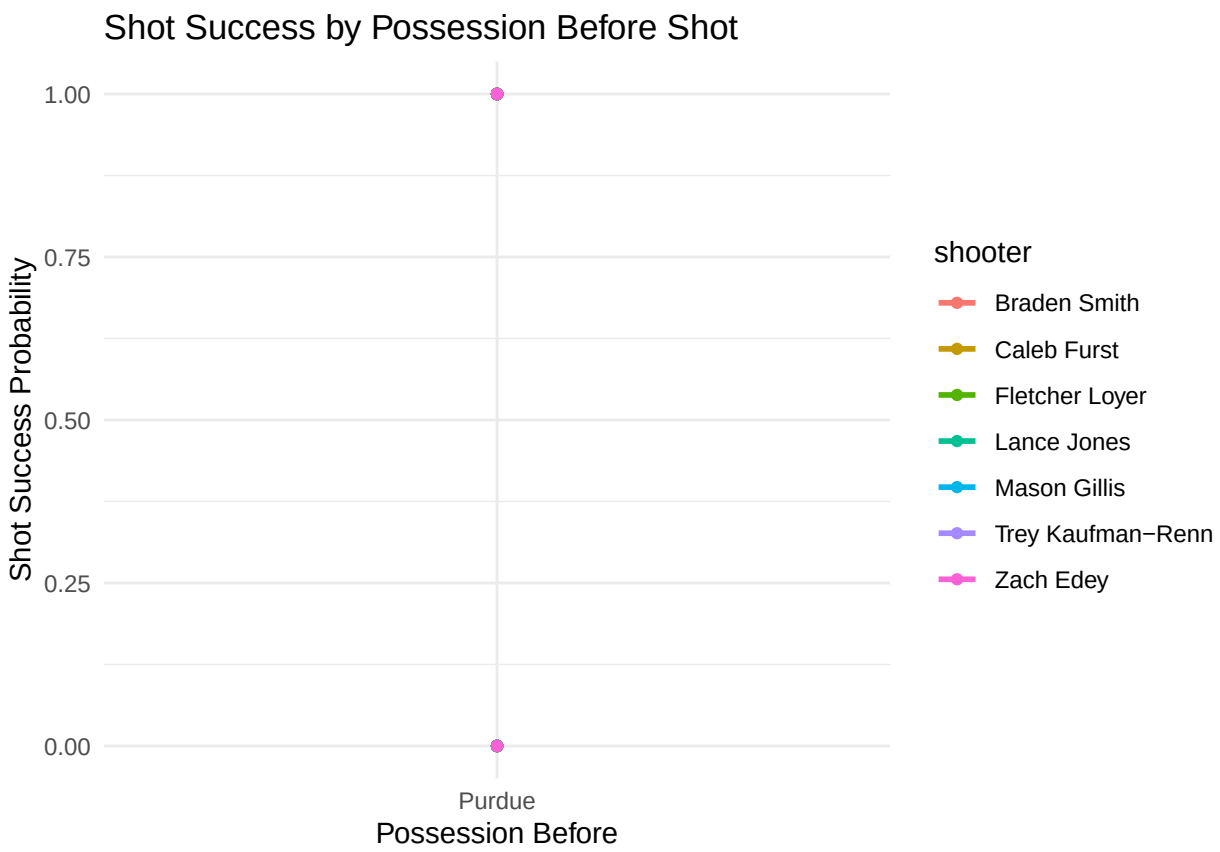
The Shot Success by Possession Before Shot plot examines how possession changes prior to the shot influence success. It may show whether a player is more likely to score following specific possession scenarios, such as a defensive rebound or steal, offering insights into the flow and rhythm that set up successful shots.

```
## `geom_smooth()` using formula = 'y ~ x'
```

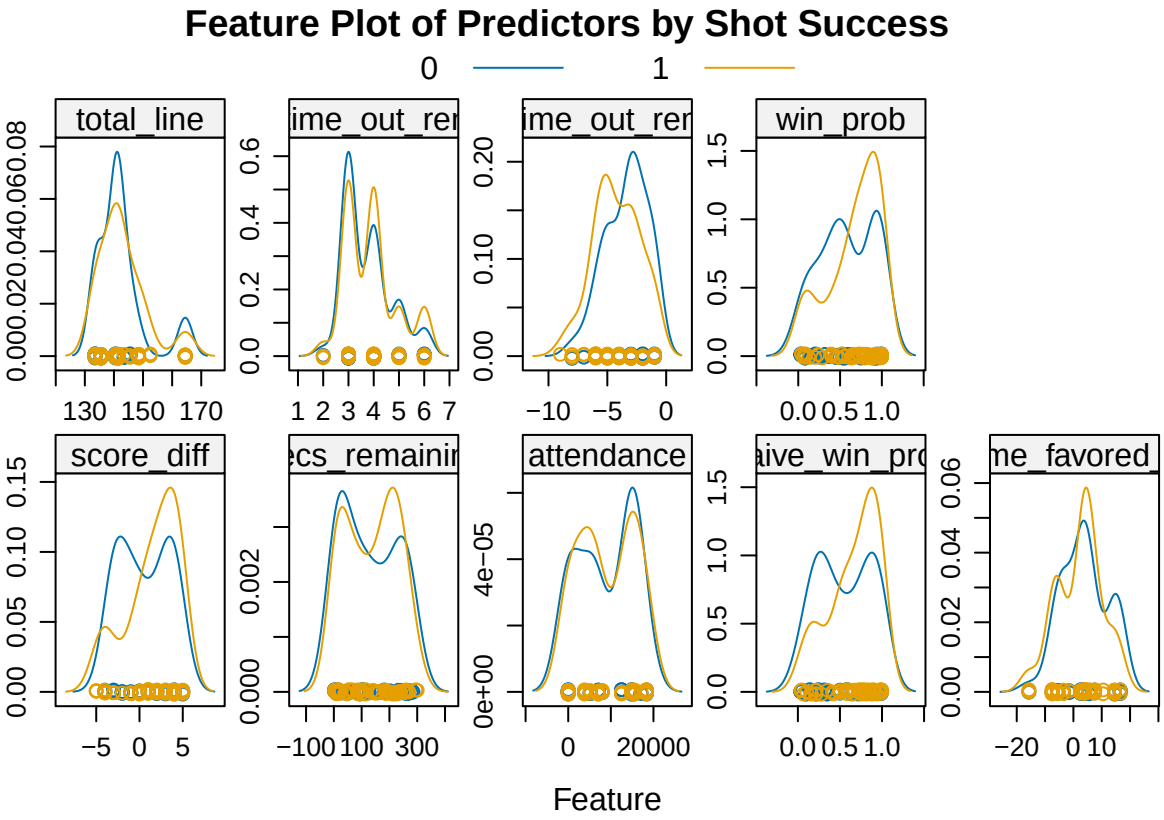
```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```




```
## `geom_smooth()` using formula = 'y ~ x'
```



6. Feature Plot: The density plots for each predictor variable allow for comparison between successful and failed shots across features like score difference, time remaining, and win probability. Peaks in the density plots for successful shots can reveal favorable conditions that correlate with higher shot success rates. This plot enables a nuanced understanding of the conditions under which players are most effective, helping to inform situational strategies based on these factors.



These EDA plots provided a foundation for selecting variables for model training, helping to capture essential aspects of player performance in clutch situations.

Model Training and Model Comparison

Multiple machine learning models were trained to predict shot success in clutch moments, specifically in the last five minutes of a close game. These models included Logistic Regression, Ridge Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting Machine (GBM), and k-Nearest Neighbors (kNN). Each model was evaluated on performance metrics such as Accuracy, Kappa, Sensitivity (true positive rate), Specificity (true negative rate), AUC (Area Under the Curve), and F1 Score to assess its ability to distinguish between successful and unsuccessful shots under high-pressure scenarios.

```
## Setting levels: control = Fail, case = Success
## Setting direction: controls < cases
## Setting levels: control = Fail, case = Success
## Setting direction: controls < cases
## Setting levels: control = Fail, case = Success
## Setting direction: controls < cases
## Setting levels: control = Fail, case = Success
## Setting direction: controls < cases
## Setting levels: control = Fail, case = Success
## Setting direction: controls < cases
## Setting levels: control = Fail, case = Success
```

```
## Setting direction: controls > cases
```

```
##                 Model  Accuracy       Kappa Sensitivity Specificity     AUC
## 1 Logistic Regression 0.5000000  0.02409639       0.625         0.4 0.56250
## 2    Ridge Regression 0.5555556  0.12195122       0.625         0.5 0.55000
## 3                 SVM 0.6666667  0.34146341       0.750         0.6 0.68750
## 4       Random Forest 0.6666667  0.34146341       0.750         0.6 0.53750
## 5                 GBM 0.6111111  0.22222222       0.625         0.6 0.52500
## 6                 KNN 0.5000000 -0.05194805       0.250         0.7 0.68125
##    F1_Score
## 1 0.5263158
## 2 0.5555556
## 3 0.6666667
## 4 0.6666667
## 5 0.5882353
## 6 0.3076923
```

Logistic Regression and Ridge Regression: These linear models offered baseline performance, with accuracies around 50-55.6%. While these models are simple and interpretable, their performance was relatively lower compared to non-linear models, suggesting limitations in capturing the complexity of clutch-time shot success.

Support Vector Machine (SVM): The SVM model, using a Radial Basis Function kernel, emerged as the top performer with an accuracy of 66.7%. It also scored high in AUC and F1 Score, indicating a strong balance between sensitivity and specificity. SVM's performance suggests it effectively handles non-linear patterns in the data, making it well-suited for this application.

Random Forest and GBM: These ensemble models achieved competitive accuracy rates (61-66.7%) and demonstrated robustness across different game situations. Although their performance was close to SVM, their slightly lower composite scores resulted in their ranking below SVM.

k-Nearest Neighbors (kNN): kNN's performance was relatively lower, indicating that its simple distance-based approach was less effective in capturing the nuanced conditions of shot success.

Selecting the Best Model

Metrics were normalized and weighted to calculate a Composite Score, providing an overall performance measure. Weights were assigned to Accuracy (30%), AUC (40%), and F1 Score (30%) to balance precision and predictive power. Based on the composite score, SVM ranked the highest, followed by Random Forest and GBM, making SVM the preferred model for player recommendations.

```
## [1] "Ranked Models by Composite Score:"
```

```
##                 Model  Accuracy       Kappa Sensitivity Specificity     AUC
## 3                 SVM 0.6666667  0.34146341       0.750         0.6 0.68750
## 4       Random Forest 0.6666667  0.34146341       0.750         0.6 0.53750
## 5                 GBM 0.6111111  0.22222222       0.625         0.6 0.52500
## 2    Ridge Regression 0.5555556  0.12195122       0.625         0.5 0.55000
## 1 Logistic Regression 0.5000000  0.02409639       0.625         0.4 0.56250
## 6                 KNN 0.5000000 -0.05194805       0.250         0.7 0.68125
##    F1_Score Accuracy_norm  AUC_norm F1_Score_norm Composite_Score
## 3 0.6666667     1.0000000 1.0000000     1.0000000       1.0000000
## 4 0.6666667     1.0000000 0.7818182     1.0000000       0.9127273
## 5 0.5882353     0.9166667 0.7636364     0.8823529       0.8451604
## 2 0.5555556     0.8333333 0.8000000     0.8333333       0.8200000
## 1 0.5263158     0.7500000 0.8181818     0.7894737       0.7891148
## 6 0.3076923     0.7500000 0.9909091     0.4615385       0.7598252
```

```
## [1] "Best Model:"
```

```
##   Model  Accuracy     Kappa Sensitivity Specificity    AUC  F1_Score
## 3   SVM 0.6666667 0.3414634        0.75         0.6 0.6875 0.6666667
##   Accuracy_norm AUC_norm F1_Score_norm Composite_Score
## 3             1        1             1               1
```

With its superior composite score, SVM was selected as the final model to drive recommendations. Its high accuracy and sensitivity to game-specific variables make it effective for identifying players with high shot success probabilities in clutch situations.

The Model Training and Comparison section shows that SVM outperformed other models, effectively predicting shot outcomes based on game conditions and player attributes. The structured comparison of models provides a clear rationale for selecting SVM, ensuring the recommendation system is backed by robust analysis and reliable metrics.

Selecting the Optimal Player

The Optimal Player section analyzes the players' clutch-time shot performance to determine the best candidate for taking the final shot in high-stakes scenarios. The Support Vector Machine (SVM) model, selected for its superior accuracy and ability to handle non-linear data patterns, was used to predict each player's likelihood of scoring under pressure.

The SVM model evaluated each player based on situational game factors such as score difference, time remaining, win probability, and other predictors. By assessing these conditions, the model calculated the probability of a successful shot for each player, ultimately identifying players with consistently high success rates in clutch scenarios.

```
## Warning in .local(x, ...): Variable(s) `' constant. Cannot scale data.

## Warning in .local(x, ...): Variable(s) `' constant. Cannot scale data.

## Warning in .local(x, ...): Variable(s) `' constant. Cannot scale data.

## Warning in .local(x, ...): Variable(s) `' constant. Cannot scale data.

## Warning in .local(x, ...): Variable(s) `' constant. Cannot scale data.

## Warning in .local(x, ...): Variable(s) `' constant. Cannot scale data.

## Warning in .local(x, ...): Variable(s) `' constant. Cannot scale data.

## Warning in .local(x, ...): Variable(s) `' constant. Cannot scale data.

## Warning in .local(x, ...): Variable(s) `' constant. Cannot scale data.

## Warning in .local(x, ...): Variable(s) `' constant. Cannot scale data.

## Adding missing grouping variables: `score_diff`

## # A tibble: 9 x 3
##   score_diff shooter        predicted_prob
##        <int> <fct>                   <dbl>
## 1         -3 Mason Gillis            0.527
## 2          3 Fletcher Loyer          0.562
## 3          4 Zach Edey               0.564
## 4          5 Zach Edey               0.548
## 5         -1 Mason Gillis            0.535
## 6         -2 Lance Jones             0.535
```

```
## 7          0 Zach Edey              0.556
## 8          1 Zach Edey              0.556
## 9          2 Zach Edey              0.555
```

Based on the SVM model's predictions, Zach Eedy consistently achieved high probabilities of successful shots across various game situations. His performance data, when assessed against key predictors, demonstrated his reliability and scoring ability under pressure. Edey's success rate in clutch-time moments set him apart from other players, making him the logical choice for critical, game-deciding shots. His consistency under varied game situations, as demonstrated by the SVM model's analysis, validated him as the optimal player for this role.

The selection of Zach Edey is grounded in objective data analysis, showing that his success rate aligns with conditions typical of clutch scenarios (e.g., close score differences, limited time remaining). This data-backed approach ensures that the decision is not based on intuition alone but on statistical evidence of his ability to perform under pressure.

Conclusion

This analysis highlights Zach Edey as the optimal player for Purdue Men's Basketball in clutch-time situations, based on a predictive model that captures game-specific and player-specific attributes. The SVM model, with its high accuracy and adaptability to non-linear patterns, proved ideal for recommending players in high-pressure moments. This project demonstrates the value of machine learning in sports analytics, showcasing how data can enhance decision-making in critical scenarios.