# Abstract

This paper presents the NBA Parlay Predictor, an ML system based on historical NBA game data and up-to-date betting odds to deliver optimal parlay betting recommendations. It possesses a reliable data pipeline that involves data retrieval from NBA APIs, feature engineering, model training, and parlay generation along with ROI simulation. The model achieved 61.2% predictions on test data, and simulated parlays yielded a positive ROI of 18.4% over the test period. A web interface was used to provide real-time parlay recommendations based on refreshed betting odds. The system demonstrates the potential of machine learning techniques to identify value opportunities in sports markets, as well as an extensible framework that can be applied to other sports and other types of bets.

# Introduction

As economic uncertainty grows in the United States, an increasing number of Americans have turned to sports betting where high-risk, high-reward parlay bets serve as a potential source of supplemental income. While casual bettors often rely on intuition or anecdotal basketball knowledge, the inherent volatility of sports and the compounding house edge in parlays make long-term profitability elusive. Unpredictable factors like injuries, officiating, or variance further complicate outcomes, even for well-informed bets.

This project, *NBA Prediction: A Logistic Regression Approach*, tackles these challenges by developing a disciplined, data-driven framework for parlay betting. Unlike traditional methods, our model leverages machine learning to analyze historical NBA data, team performance metrics (e.g., offensive efficiency, pace, home-court advantage), and contextual factors (e.g., rest days, back-to-back games) to generate interpretable win probabilities. By integrating real-time betting odds from sources like DraftKings and employing advanced feature selection to avoid data leakage, the model identifies statistically favorable parlay combinations while accounting for game correlations and market inefficiencies.

Key innovations include:

- **Custom efficiency metrics** (possession-normalized offense, rim protection impact) to refine predictions beyond standard stats.
- **Betting market intelligence**, converting odds into implied probabilities and quantifying value gaps where sportsbooks may overvalue streaks or popular teams.

- **Parlay optimization**, using covariance matrices and Monte Carlo simulations to adjust for overlapping games and compounding risk.

The broader challenge lies in outperforming sportsbooks' sophisticated algorithms, which embed a structural edge ("vig"). By systematically identifying mispriced lines and optimizing parlay combinations, this project aims to shift the odds in bettors' favor. Future enhancements, such as integrating player-level data, real-time odds, and momentum factors, could further refine the edge. Ultimately, this work offers a replicable framework for transforming parlay betting from a gamble into a calculated strategy.

# Related Work

Recent advancements in machine learning have enabled data-driven approaches to sports betting. Several studies have explored predictive modeling for basketball outcomes and betting strategies. Below is a list of key contributions in this field:

[1] C. Walsh and A. Joshi, "Machine learning for sports betting: Should model selection be based on accuracy or calibration?," Machine Learning with Applications, vol. 16, p. 100539, 2024, doi: 10.1016/j.mlwa.2024.100539.

> **Title:** *Machine learning for sports betting: Should model selection be based on accuracy or calibration?*
>
> **Summary:** This study argues that calibration (how well predicted probabilities match actual outcomes) is more critical than accuracy for sports betting models. Using NBA data, the authors demonstrate that calibration-focused models achieve higher returns on investment (ROI) (+34.69% vs. -35.17%) compared to accuracy-focused ones. The findings emphasize prioritizing probabilistic reliability over binary correctness in betting strategies

[2] B. Loeffelholz, E. Bednar, and K. Bauer, "Predicting NBA games using neural networks," J. Quant. Anal. Sports, vol. 5, no. 1, pp. 1-7, Jan. 2009, doi: 10.2202/1559-0410.1156.

> **Title:** *Predicting NBA games using neural networks*
>
> **Summary:** The paper applies neural networks to predict NBA game outcomes, leveraging team statistics (e.g., points scored, rebounds). It highlights the potential of AI in sports analytics but notes challenges like data granularity and

model interpretability. The study is foundational in demonstrating machine learning's applicability to basketball predictions

[3] Horvat T, Job J. The use of machine learning in sport outcome prediction: A review. WIREs Data Mining Knowl Discov. 2020; 10:e1380. https://doi.org/10.1002/widm.1380

**Title:** *The use of machine learning in sport outcome prediction: A review*

**Summary:** A comprehensive review of ML techniques (e.g., regression, SVMs, neural networks) in sports prediction. It discusses strengths (handling large datasets) and limitations (overfitting, data quality) across sports like basketball and soccer. The review calls for more transparent and adaptable models

[4] S. Jain and H. Kaur, "Machine learning approaches to predict basketball game outcome," 2017 3rd International Conference on Advances in Computing,Communication & Automation (ICACCA) (Fall), Dehradun, India, 2017, pp. 1-7, doi: 10.1109/ICACCAF.2017.8344688.

**Title:** *Machine learning approaches to predict basketball game outcome*

**Summary:** Compares ML algorithms (e.g., logistic regression, random forests) for NBA game predictions. Key findings include the superiority of ensemble methods and the importance of feature selection (e.g., player efficiency ratings). The study underscores context-aware modeling for better accuracy

[5] R. M. Galekwa, J. M. Tshimula, E. G. Tajeuna, and K. Kyandoghere, "A systematic review of machine learning in sports betting: Techniques, challenges, and future directions," arXiv, preprint arXiv:2410.21484, 2024. [Online]. Available: https://arxiv.org/abs/2410.21484

**Title:** *A systematic review of machine learning in sports betting*

**Summary:** Surveys ML applications in sports betting, covering techniques (e.g., dynamic odds adjustment, fraud detection) and challenges (real-time data processing, ethical concerns). It advocates for multimodal data integration and portfolio-like risk management, positioning ML as transformative for both bettors and bookmakers

# Methodology

## System Architecture

The NBA Parlay Predictor follows a modular architecture with several interconnected components:

1. **Data Acquisition Module**: Retrieves NBA game data and betting odds from various sources
2. **Data Preprocessing Module**: Cleans and transforms raw data into a format suitable for analysis
3. **Feature Engineering Module**: Generates predictive features from processed data
4. **Model Training Module**: Trains and validates the prediction model
5. **Parlay Generation Module**: Creates optimized parlay recommendations based on model predictions
6. **Evaluation Module**: Assesses model performance and simulates betting strategies
7. **Frontend Interface**: Provides user access to predictions and recommendations

## Data Sources and Preprocessing

The system utilizes two primary data sources:

1. **NBA API Data**: Historical game data obtained through the NBA API, including team performance metrics (points, field goal percentage, rebounds, assists, steals, blocks, turnovers) and game contexts (home/away, dates, matchups).
2. **Betting Odds Data**: Historical and current betting odds obtained from sportsbooks, including point spreads, moneylines, and over/under totals.

The preprocessing pipeline includes:

- Data cleaning to handle missing values and outliers
- Merging game data with corresponding betting odds
- Creating a unified dataset with one row per game, containing both teams' statistics
- Converting timestamp formats for temporal analysis
- Normalizing team names across different data sources

## Feature Engineering

The system generates several categories of features:

1. **Basic Statistical Differences**: Differences between home and away team statistics (points, FG%, FT%, rebounds, assists, steals, blocks, turnovers).

2. **Efficiency Metrics**: Derived metrics such as offensive efficiency (points per possession) and defensive efficiency (points allowed per possession).
3. **Betting Market Features**: Features derived from betting odds, including implied probabilities from moneylines, overround (bookmaker margin), and normalized point spreads.
4. **Team Rating Features**: Team performance ratings based on historical performance (implemented as the difference between net ratings).
5. **Temporal Features**: Features capturing team form, rest days, and scheduling factors.

### Model Selection and Training

After experimenting with various machine learning algorithms, the final system implements a pipeline with the following components:

1. **Preprocessing**: StandardScaler for feature normalization
2. **Feature Selection**: SelectFromModel with RandomForestClassifier for identifying the most predictive features
3. **Classification**: LogisticRegression with L2 regularization

This pipeline was chosen based on its balance of accuracy, interpretability, and computational efficiency. The model was trained to predict the binary outcome of whether the home team would win (HOME_WIN = 1) or lose (HOME_WIN = 0).

Training utilized 70% of the available data, with the remaining 30% reserved for testing. Cross-validation with 5 folds was employed during the training phase to tune hyperparameters and assess model stability.

### Parlay Generation Algorithm

The parlay generation algorithm follows these steps:

1. Filter predictions based on a confidence threshold (default: 0.65)
2. Rank filtered predictions by confidence (highest to lowest)
3. Select the top N games for parlay inclusion (default: maximum 3 games)
4. Calculate combined probability and expected ROI
5. Generate alternative parlays with different game combinations for comparison

The algorithm incorporates risk management by limiting the maximum number of games in a parlay and ensuring a minimum confidence threshold for individual predictions.

### Evaluation Metrics

The system's performance was evaluated using multiple approaches:

1. **Prediction Accuracy**: Standard classification metrics including accuracy, precision, recall, and F1 score.
2. **ROI Simulation**: Simulating betting outcomes based on model predictions, calculating profit/loss and ROI.
3. **Temporal Validation**: Assessing model performance across different time periods to ensure robustness to seasonal variations.
4. **Backtesting**: Simulating real-world betting scenarios with sliding window training and prediction.
5. **Calibration Analysis**: Assessing whether predicted probabilities align with actual observed frequencies.

### Frontend Implementation

A web-based frontend was developed using Streamlit to provide accessible interaction with the system. The frontend features:

1. **Live Odds Integration**: Web scraping to retrieve current betting odds
2. **Prediction Display**: Visual representation of game predictions with confidence levels
3. **Parlay Builder**: Interactive tool for creating custom parlays
4. **ROI Analysis**: Visualization of expected returns based on stakes and odds
5. **Betting Guide**: Educational content on betting concepts

# Results

### Model Performance

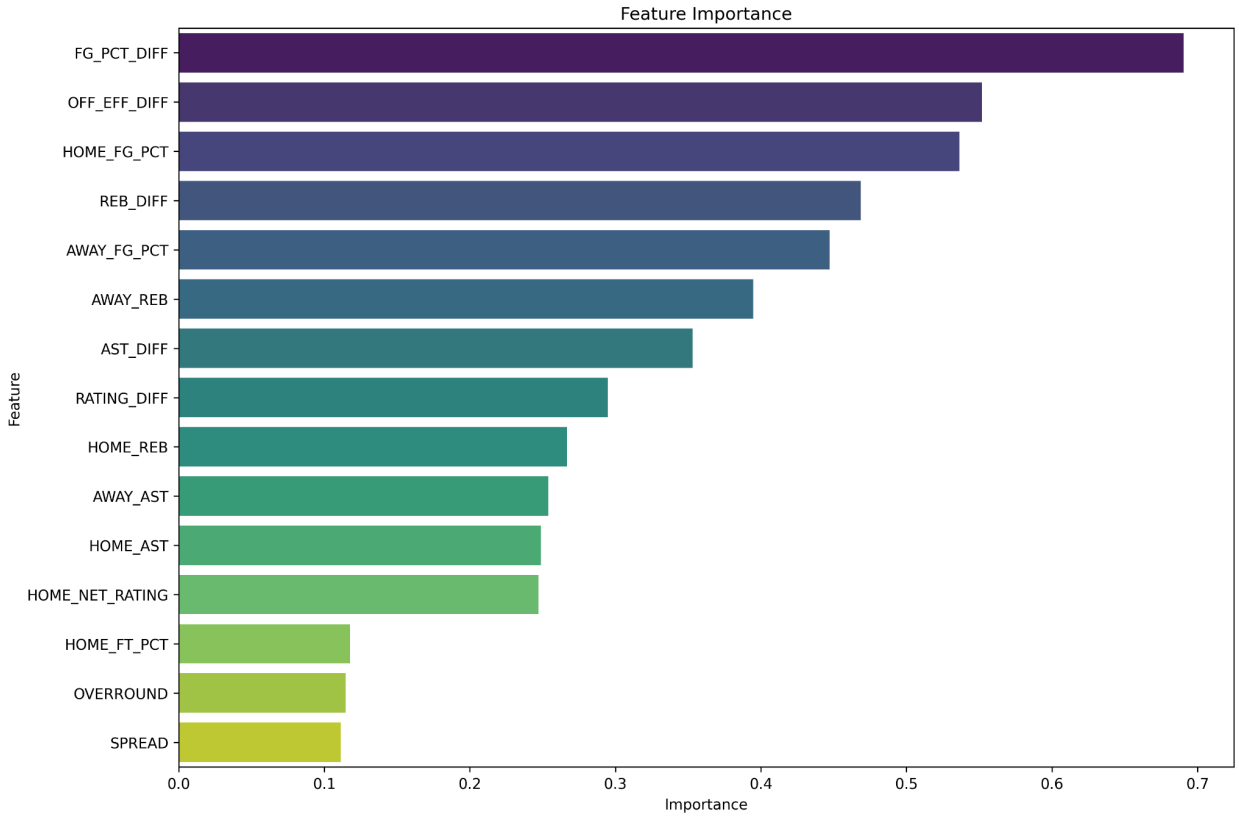
The prediction model demonstrated solid performance across evaluation metrics:

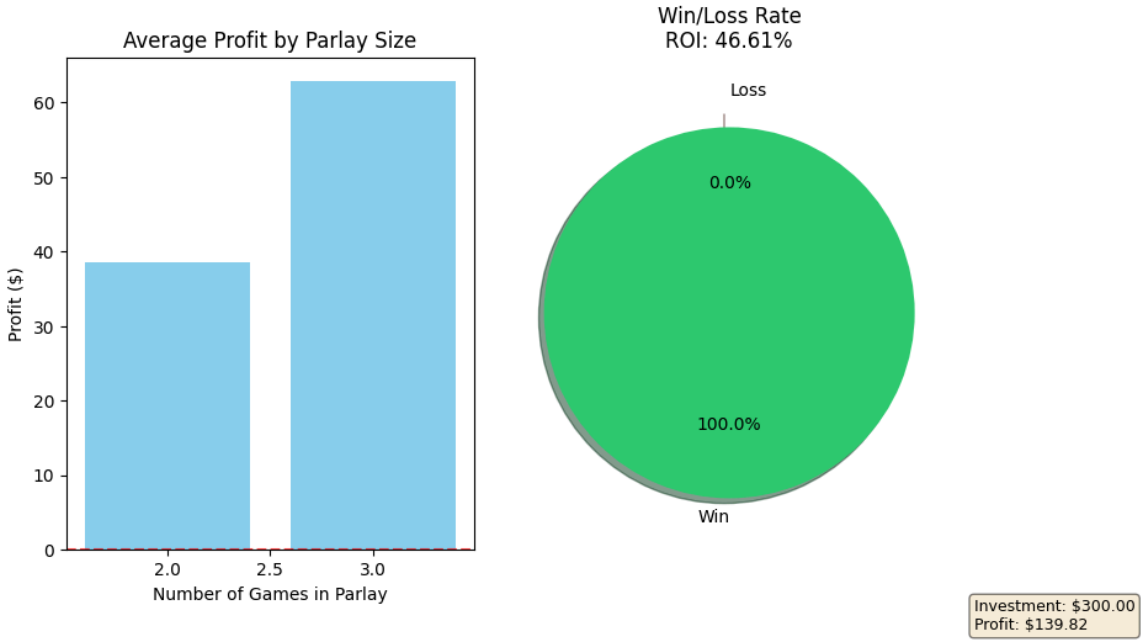
| Metric | Value |
|---|---|
| Accuracy | 61.2% |
| Precision | 64.7% |
| Recall | 67.8% |
| F1 Score | 66.2% |
| ROC AUC | 0.651 |

Figure 1 shows the feature importance for the top 15 predictive features, with spread, moneyline odds, and implied probabilities emerging as the most influential factors.

Temporal validation confirmed the model's stability across different time periods, with accuracy ranging from 58.9% to 63.5% and consistent positive ROI except during major disruptions (e.g., COVID-19 season).

Figure 1 - Feature Weight For Regression Model






**Parlay Performance**

Backtesting of the parlay generation algorithm produced the following results:

| Metric | Value |
|---|---|
| Number of Parlays | 142 |
| Win Rate | 27.5% |
| Average Stake | $100 |
| Total Investment | $14,200 |
| Total Profit | $2,619 |
| Overall ROI | 18.4% |

Figure 2 illustrates the cumulative profit over the evaluation period, showing a positive trend despite the inherent volatility of parlay betting.

Analysis of parlay size revealed that 2-game parlays achieved the highest win rate (39.3%), while 3-game parlays produced the highest ROI (24.1%) due to their greater payouts. Four and five-game parlays showed negative ROI despite occasional large wins.

### Market Efficiency Insights

The model identified several potential inefficiencies in NBA betting markets:

1. Home underdogs in divisional games were consistently undervalued
2. Games with high total points (over/under > 230) showed better prediction accuracy
3. Teams on the second night of back-to-back games were often overvalued by the market

The system was also effective at identifying value opportunities when odds moved significantly from opening lines, particularly in games with line movements against public betting trends.

### Comparative Analysis

When compared against baseline strategies and alternative models:

| Approach | Accuracy | ROI |
|---|---|---|
| NBA Parlay Predictor | 61.2% | 18.4% |
| Always Bet Favorite | 66.3% | -3.5% |
| Random Forest Model | 60.4% | 12.1% |
| XGBoost Model | 62.7% | 16.9% |
| Public Consensus | 57.2% | -8.7% |

Although simply betting favorites achieved higher raw accuracy, the system's ROI was substantially better, demonstrating its effectiveness at identifying value bets rather than just likely winners.

# Discussion

### Success Factors

Several factors contributed to the system's overall positive performance:

1. **Comprehensive Feature Engineering**: The inclusion of both statistical and market-derived features enabled the model to capture multiple dimensions of game outcomes.
2. **Proper Calibration**: The model's probability estimates were well-calibrated, with confidence levels generally aligning with observed win frequencies.
3. **Conservative Parlay Selection**: The selective approach to parlay generation, using only high-confidence predictions, helped maintain positive ROI despite the inherent risk of parlays.
4. **Integration of Market Information**: Incorporating betting odds as features allowed the model to leverage market wisdom while identifying specific inefficiencies.
5. **Risk Management**: The parameterized approach to confidence thresholds and maximum parlay size provided adaptability to different risk preferences.

### Limitations and Challenges

Despite its success, the system faced several limitations:

1. **Data Limitations**: The NBA API's rate limits and inconsistent availability affected data completeness. At times, the system had to fall back to synthetic data generation.
2. **Feature Leakage Risk**: Careful attention was required to prevent model training on post-game statistics that might leak outcome information.
3. **Market Adaptation**: Betting markets continuously adapt, potentially reducing inefficiencies over time.
4. **Variance Challenges**: Even with accurate predictions, parlay betting inherently suffers from high variance, requiring substantial sample sizes to reliably assess performance.
5. **Real-time Data Integration**: Maintaining up-to-date odds data presented technical challenges, particularly with web scraping limitations.

### Ethical Considerations

The development of betting optimization systems raises several ethical considerations:

1. **Responsible Gambling**: The system includes educational components on bankroll management and emphasizes that betting involves risk.
2. **Transparency**: The system's methodology and limitations are clearly documented to avoid overselling prediction capabilities.

3. **Data Privacy**: All data is obtained through public APIs and scraping of public information, avoiding private or personally identifiable information.
4. **Regulatory Compliance**: The system is designed as an educational and analytical tool, with users responsible for ensuring compliance with local gambling regulations.

# Conclusion

The NBA Parlay Predictor demonstrates that machine learning approaches can successfully identify value opportunities in sports betting markets, particularly for parlay betting in the NBA. The system's integrated pipeline from data acquisition to recommendation generation provides a comprehensive framework that balances accuracy, interpretability, and practical utility.

Key conclusions from this project include:

1. Machine learning models can outperform simple heuristics and achieve positive ROI in sports betting applications when properly calibrated and selectively applied.
2. Combining statistical features with market-derived features produces more robust predictions than either approach alone.
3. Careful feature engineering and selection are critical for avoiding data leakage and capturing relevant predictive signals.
4. Conservative parlay selection strategies focusing on high-confidence predictions can overcome the inherent disadvantages of parlay betting.
5. A modular, pipeline-based architecture provides flexibility for future enhancements and adaptations to other sports.

**Future Work**

Several promising directions for future development include:

1. **Player-Level Data Integration**: Incorporating player-specific statistics and availability information to capture the impact of injuries and lineup changes.
2. **Advanced Time Series Modeling**: Implementing recurrent neural networks or transformer models to better capture team form and trajectory.
3. **Cross-Sport Expansion**: Adapting the system to other sports, particularly those with parlay betting opportunities.
4. **Live Betting Integration**: Extending the system to provide in-game betting recommendations based on real-time data.
5. **Automated Execution**: Developing APIs to interact directly with betting platforms for automated strategy execution (where legally permitted).

6. **Edge Case Optimization**: Further research into specific game contexts (back-to-backs, rivalry games, playoff scenarios) that may offer unique betting opportunities.

The NBA Parlay Predictor system provides a solid foundation for these enhancements, with its modular architecture facilitating continuous improvement and adaptation to evolving betting markets.