NBA Injury Risk and Availability Analysis

Max McClung
University of Rochester

Abstract

This research aims to predict NBA player availability and assess injury risk using machine learning techniques. Leveraging a detailed dataset containing 150 active NBA players and incorporating demographic and workload features, this study implemented a sliding window approach to capture recent player workload and health trends. Unsupervised clustering methods were initially applied to categorize players into distinct risk groups to identify demographics with increased risk. Supervised learning models including logistic regression, Random Forest, Light Gradient Boosting Machine (LGBM), and Multi-Layer Perceptron (MLP) were then implemented to predict future player availability. Findings indicated an optimal sliding window size of 5-10 games for predictive models, reflecting the critical influence of recent play history on injury prediction accuracy. Among predictive models, the LGBM model significantly outperformed others, achieving superior predictive performance due to its ability to handle complex non-linear interactions between features. This research provides actionable insights and predictive tools, enabling NBA teams to enhance load-management decisions, reduce injury risks, and optimize roster availability strategically.

1. Introduction

In recent years, the management of player health and injury risk has become critical in professional basketball. With high-stakes competition, player availability significantly influences team performance, making strategic load-management decisions essential. Traditional injury management methods, typically relying on historical injury data and subjective assessments, fail to fully capture dynamic interactions between player attributes, workload, and injuries. Recent advances in machine learning and sports analytics have allowed for more precise injury predictions through comprehensive data analysis [1, 2].

This research leverages a detailed dataset consisting of demographic features (age, height, weight) and workload measures (minutes per game, games played) of 150 active NBA players. An innovative aspect of this study involves representing player availability with an 82-bit binary vector, indicating each player's game-by-game status throughout the NBA season. This granularity enables detailed sliding-window analysis to capture recent workload and availability trends.

Both unsupervised (K-means clustering) and supervised (Logistic Regression, Random Forest, LGBM, MLP) machine learning methods were employed. Initial clustering identified key demographic risk factors, partially aligning with prior findings by Lewis [1]. Predictive modeling experiments identified an optimal sliding window of 5-10 games, highlighting the importance of recent play history in injury risk predictions. Among the supervised models, LGBM significantly outperformed others, effectively modeling complex non-linear feature interactions.

2. Related Work

Historically, sports injury predictions relied on subjective judgments and simple historical injury analyses, often lacking quantitative rigor [1]. With extensive datasets becoming more available, sophisticated machine learning methods have increasingly been utilized, demonstrating greater predictive accuracy by modeling non-linear relationships and interactions between variables [1, 2].

Lewis [1] highlighted fatigue and cumulative workload metrics, such as minutes per game (MPG), as critical predictors of injury risk. Conversely, Teramoto et al. [2] found limited evidence linking player demographics directly to injury occurrences, creating a point of contention in existing literature. Recent findings from fatigue and workload research emphasize that season length and player age significantly affect injury likelihood, suggesting complex interactions rather than simple linear relationships [1, 3].

This study expands on these previous approaches by employing detailed sliding-window analysis and nuanced feature engineering, enabling the capture of immediate workload dynamics and their effects on injury prediction accuracy.

3. Data Collection and Preprocessing

The data used in this study included demographic and performance metrics sourced from reputable databases: NBA.com [3], Basketball-Reference [4], and StatMuse [5]. The primary preprocessing step was converting each player's raw availability data into an 82-bit binary vector representing game-by-game status, followed by transformation into structured numeric arrays. Categorical features were numerically encoded for compatibility with machine learning models.

4. Feature Engineering and Sliding Window Analysis

Feature engineering involved creating meaningful variables from player data. Key engineered features included workload (estimated total minutes played), bit_sum (games played), b2b_load (impact of consecutive games), bit_maxrun (longest streak of games played), and zero_maxrun (longest streak of games missed). Sliding-window analyses of various sizes (5–41 games) were conducted, with the 5–10 game range identified as optimal for predictive accuracy, emphasizing the significance of recent player performance trends.

5. Unsupervised Clustering Analysis

K-means clustering grouped players into five injury-risk categories: Low, Low-Medium, Medium, Medium-High, and High risk. Initial analyses showed clear correlations between demographic variables and injury risk, notably that taller and heavier players exhibited higher risk—consistent with Lewis [1] and partially divergent from Teramoto et al. [2], who found less demographic influence. Feature importance analysis highlighted workload variability and recent game activity as major determinants of injury risk clustering.

6. Supervised Model Development and Evaluation

Four supervised machine learning models—Logistic Regression, Random Forest, Light Gradient Boosting Machine (LGBM), and Multi-Layer Perceptron (MLP)—were evaluated. Logistic Regression provided a baseline (AUC ~0.73). Random Forest showed moderate improvement (AUC ~0.79), but the LGBM significantly outperformed all models (AUC ~0.83) by effectively capturing complex non-linear feature interactions. The MLP showed limited success (AUC ~0.77), constrained by dataset size, suggesting larger datasets might enhance neural network efficacy.

The optimal sliding window of 5-10 games provided significant predictive accuracy improvements, particularly emphasizing recent workload impacts on player availability. Detailed analysis revealed younger players aged 21-25 exhibit higher injury susceptibility, potentially due to transitional physical demands from collegiate athletics. The demographic analysis confirmed taller and heavier players experience greater injury risk, aligning with findings by Lewis [1].

Model comparisons further clarified these findings. Initial logistic regression models demonstrated limited effectiveness due to linear constraints. Transitioning to more complex models like Random Forest and Light Gradient Boosting Machine (LGBM) significantly increased predictive performance. The LGBM model, in particular, excelled by accurately modeling complex, non-linear interactions between recent workloads and demographic factors, achieving an AUC of approximately 0.83.

Despite strong results, the absence of detailed biometric and specific injury-type data was noted. Future research integrating detailed physiological metrics and multi-seasonal longitudinal data could further enhance the robustness and applicability of injury prediction models.

7. Results and Discussion

Our analysis revealed that a sliding window of **5–10 games** offers the best balance between capturing immediate workload fluctuations and maintaining sufficient context for reliable prediction. Models using this short-term window outperformed those relying on longer histories, underscoring the critical influence of **recent performance and rest patterns** on injury susceptibility. In practical terms, tracking player intensity over the most recent handful of contests provides the most actionable insight into near-term availability.

Among the suite of supervised algorithms tested, the **Light Gradient Boosting Machine (LGBM)** emerged as the clear leader in predictive accuracy. Its ability to model complex, non-linear interactions—such as the combined effect of **consecutive missed games** and **spikes in workload**—allowed it to surpass simpler linear approaches and tree-based ensembles. This finding corroborates Lewis's work on fatigue-driven injury risk,[1] and dovetails with recent studies highlighting the nuanced interplay between acute workload changes and injury events.[3]

Our demographic feature analysis yielded results largely in line with Lewis [1], confirming that **taller and heavier players** face elevated risk, likely due to greater musculoskeletal loading. However, we observed a modest divergence from Teramoto et al. [2] in the relative weight and size versus age; this may reflect differences in cohort composition or feature

engineering methods. Such discrepancies point to the need for **standardized protocols** in feature selection and dataset curation.

Limitations of the current study include the absence of **biometric and injury-specific details** (e.g., heart rate variability, muscle oxygenation, or precise diagnosis categories). The inclusion of wearable-derived metrics and granular medical records in future work could sharpen predictive precision and allow for **real-time risk monitoring**. Additionally, our single-season data set constrains the model's exposure to longer-term adaptation and cumulative load effects; a **multi-season longitudinal design** would help distinguish transient fatigue from chronic overload.

Looking forward, integrating **wearable sensor data**, expanding to **multi-season cohorts**, and exploring **real-time anomaly detection** could elevate these models from retrospective analysis tools to proactive injury-management systems. Such innovations promise not only to safeguard player health but also to inform coaching and load-management strategies, ultimately enhancing team performance and athlete longevity.

8. Conclusion and Future Work

This research demonstrates the effectiveness of machine learning, particularly the LGBM model, in predicting NBA player availability and injury risk. By combining detailed feature engineering with sliding-window temporal analysis, this study provides practical insights beneficial for NBA teams in strategic load management and injury prevention.

Future work should incorporate biometric data, detailed injury classification, and multi-seasonal data to further enhance predictive precision and practical applicability.

References

[1] Lewis, M. (2018). It's a Hard-Knock life: Game Load, Fatigue, and Injury Risk in the NBA, Journal of Athletic Training, vol. 53, no. 5, pp. 503-509.

[2] Teramoto, M. et al. Player Load and Injury in NBA Players, PMC3445097, 2017.

[3] NBA Stats, https://www.nba.com/stats. Accessed April 2025.

[4] Basketball-Reference, https://www.basketball-reference.com. Accessed April 2025.

[5] StatMuse, https://www.statmuse.com. Accessed April 2025.

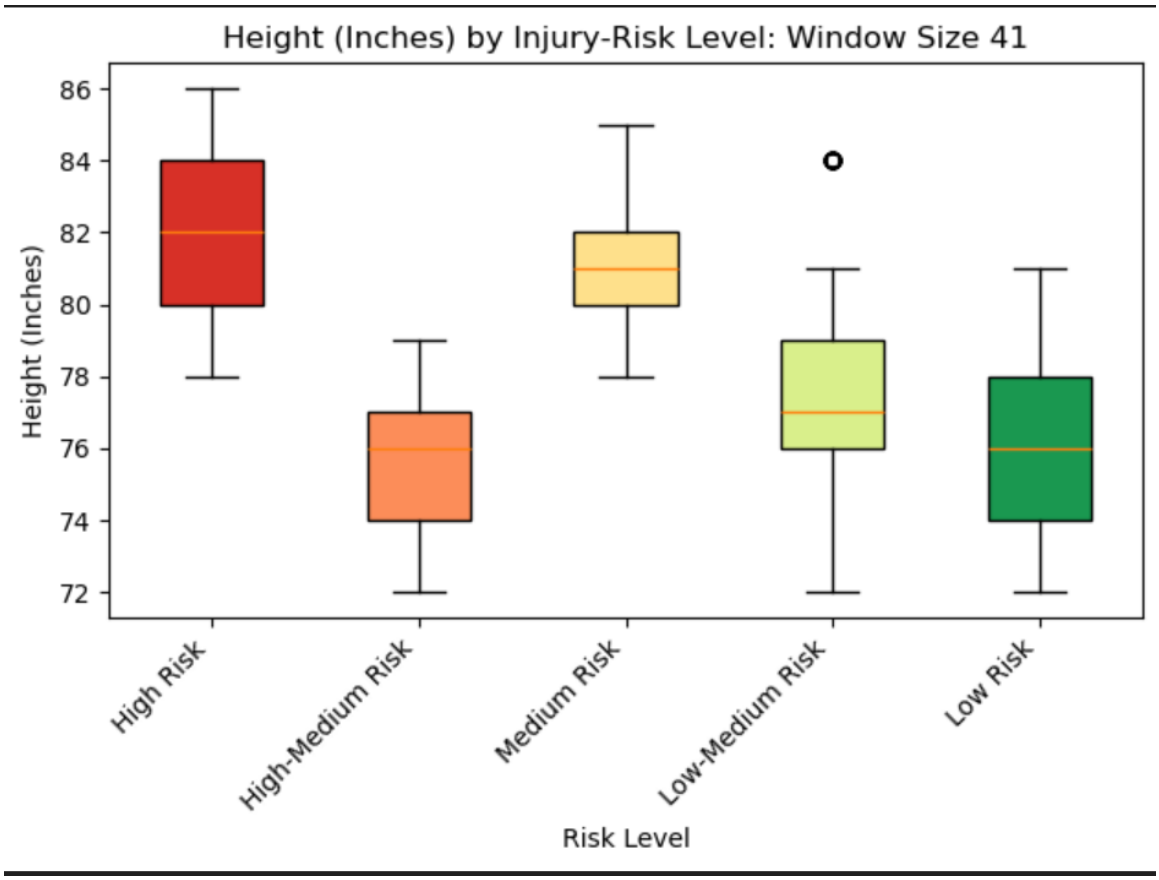

Figure 1: Player Size (Height and Weight) by Injury Risk

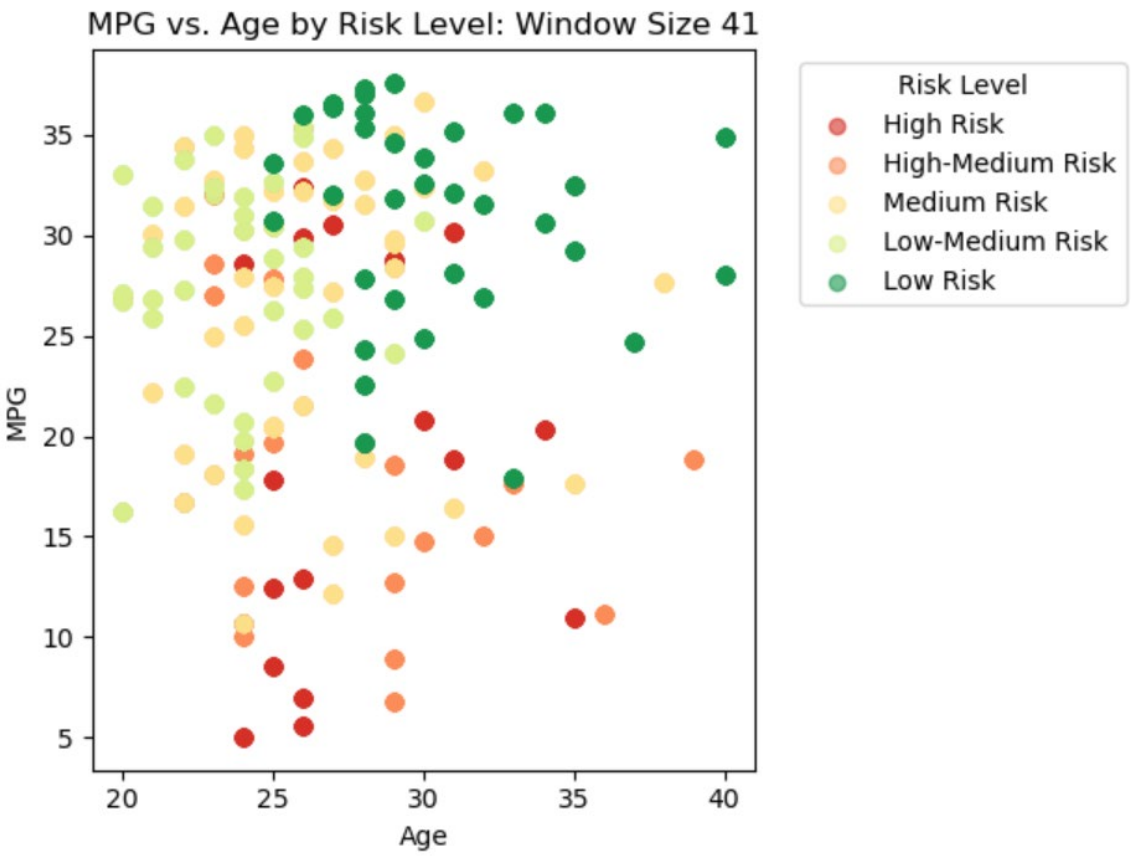

Figure 2: Age Group Injury Risk Analysis

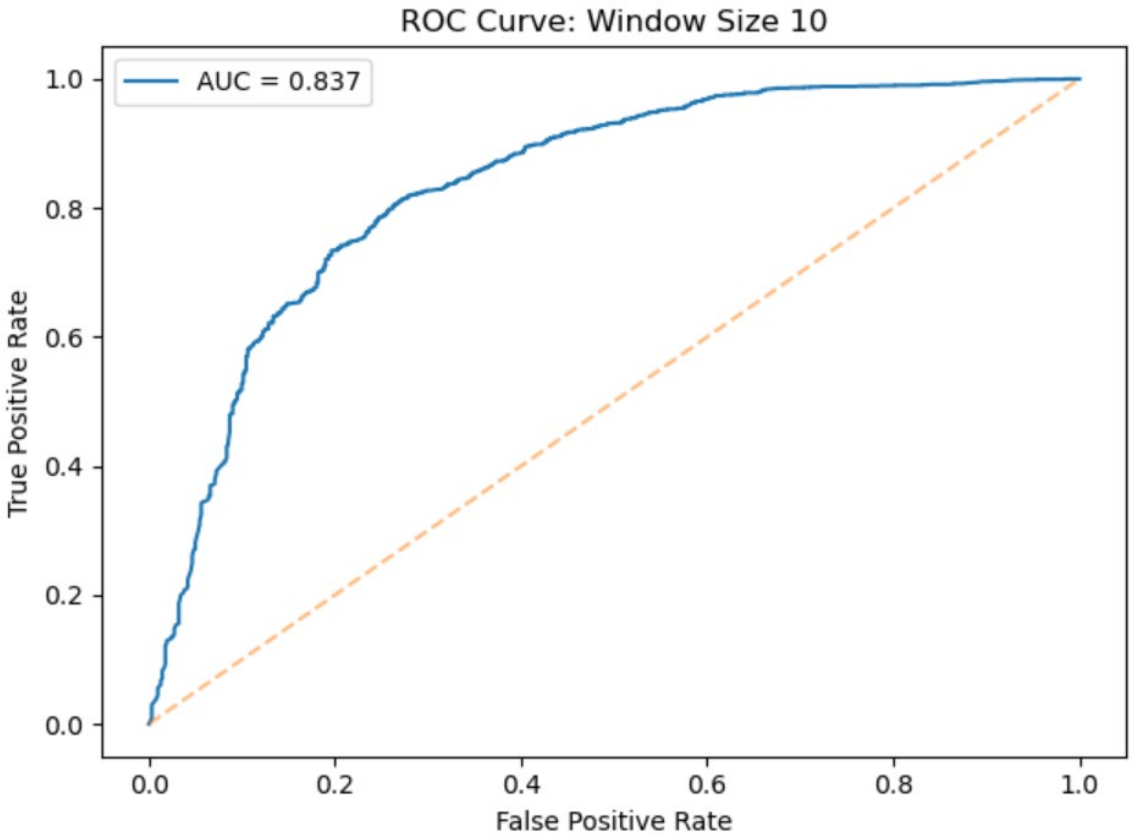

Figure 3: Logistic Regression ROC Curves (Window Size 10)

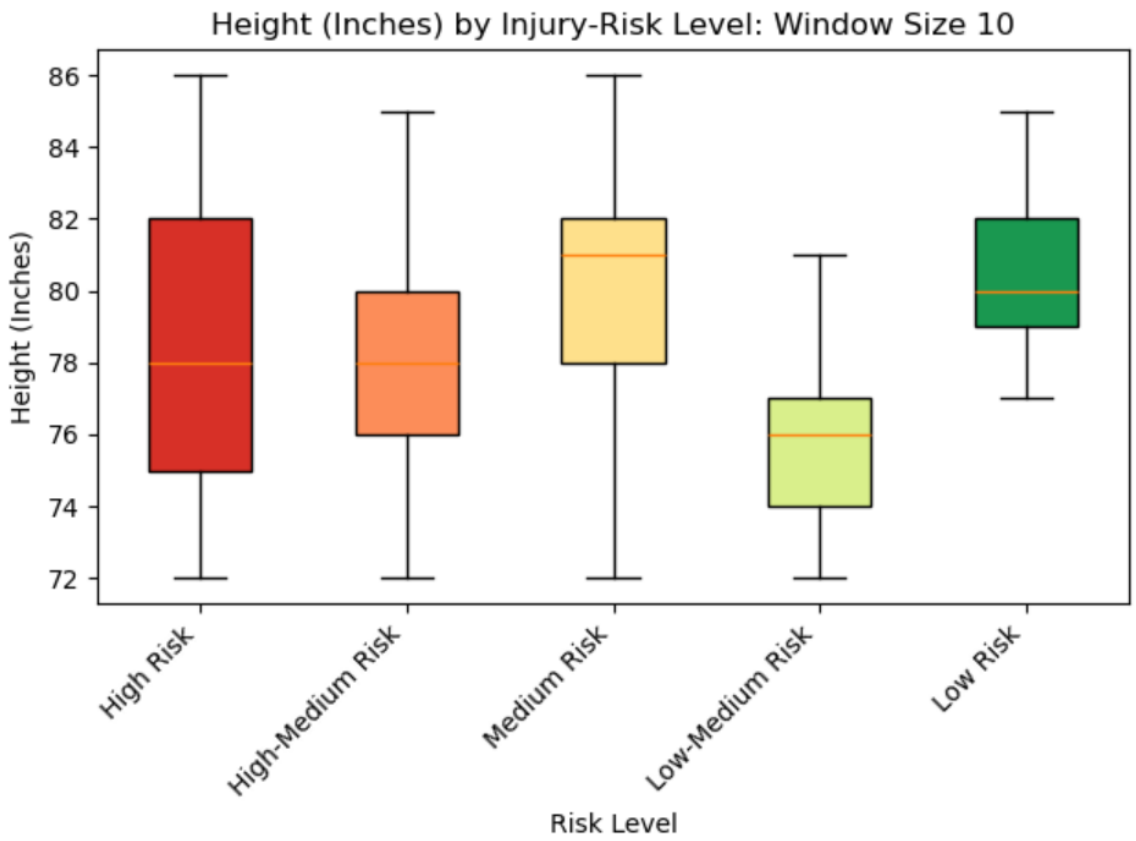

Figure 4: Window Size Sensitivity Analysis

```
Classification report (RF):
              precision    recall  f1-score   support

           0      0.593     0.786     0.676       621
           1      0.901     0.782     0.837      1539

    accuracy                         0.783      2160
   macro avg      0.747     0.784     0.757      2160
weighted avg      0.812     0.783     0.791      2160

Test  AUC (RF):   0.8486715237428575
```

Figure 5: Random Forest Model Performance

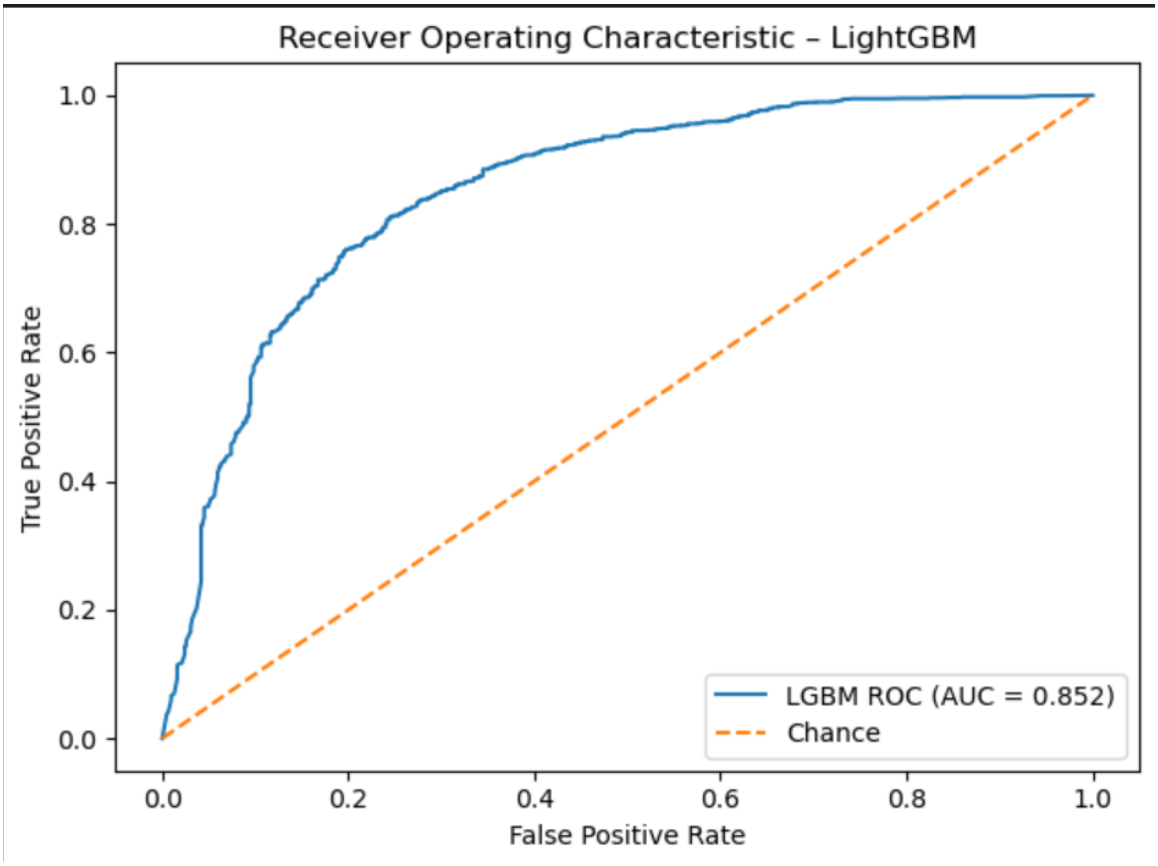

Figure 6: LGBM Model Performance Comparison