NBA Prediction Modeling

Luke DiPerna

August 2023

Project Goal

Create a model that can predict the outcome of NBA games with 68% accuracy.

Stakeholder

Stat-Ball.com, a sports news and entertainment website.

Business Use Case

The site plans to have fantasy drafts and competitions for predicting NBA game winners, so they want an in-house model for users to compete against.

Project Tasks

Data Collection	• Web-scrape available statistics • Create database to store the datasets
Data Processing	• Determine data aggregation method • Prepare data for modeling pipeline
Modeling and Testing	• Select appropriate modeling methods • Test models and analyze results

Data Collection

Scope: Boxscore data from the past 10 regular seasons

Source: Basketball-Reference.com

Method: Web-scraper

Data Storage

SQLite Database:

- 3 tables: Game Info, Player Stats, Team Stats
- 11,979 NBA games
- 341,669 observations
- 46 features
- Kaggle link

Data Processing

Aggregation Method

Team Aggregation:

- Efficient

Player Aggregation:

- Can react to roster changes

Responsiveness

- Robust vs. Relevant

- Seasonal carryover

- 10, 20, and 30 game averages

Feature Selection

- Four Factor data

- Full dataset

- Principal Component Analysis

Model Selection

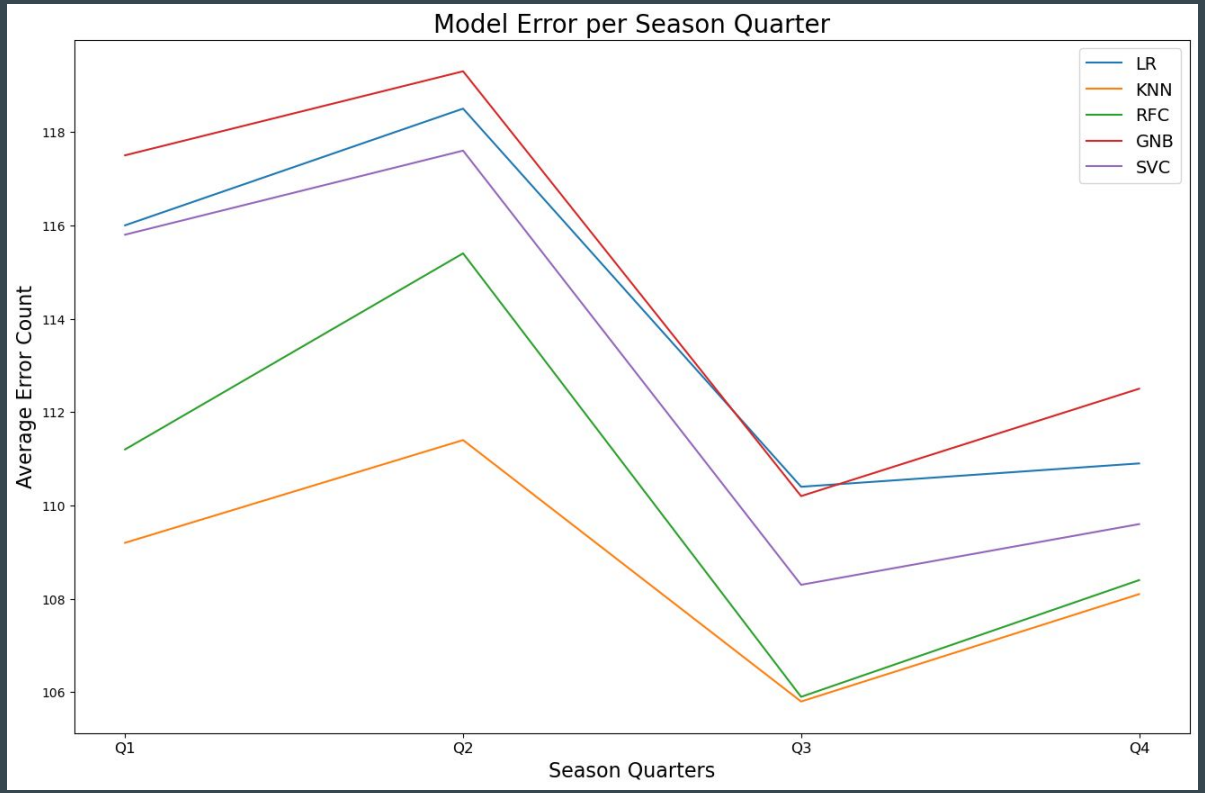
- Logistic Regression (LR)
- K-Nearest Neighbors (KNN)
- Random Forest (RF)
- Gaussian Naive-Bayes (GNB)
- Support Vector (SVC)

- Neural Network (NN)

- Elo Rating System

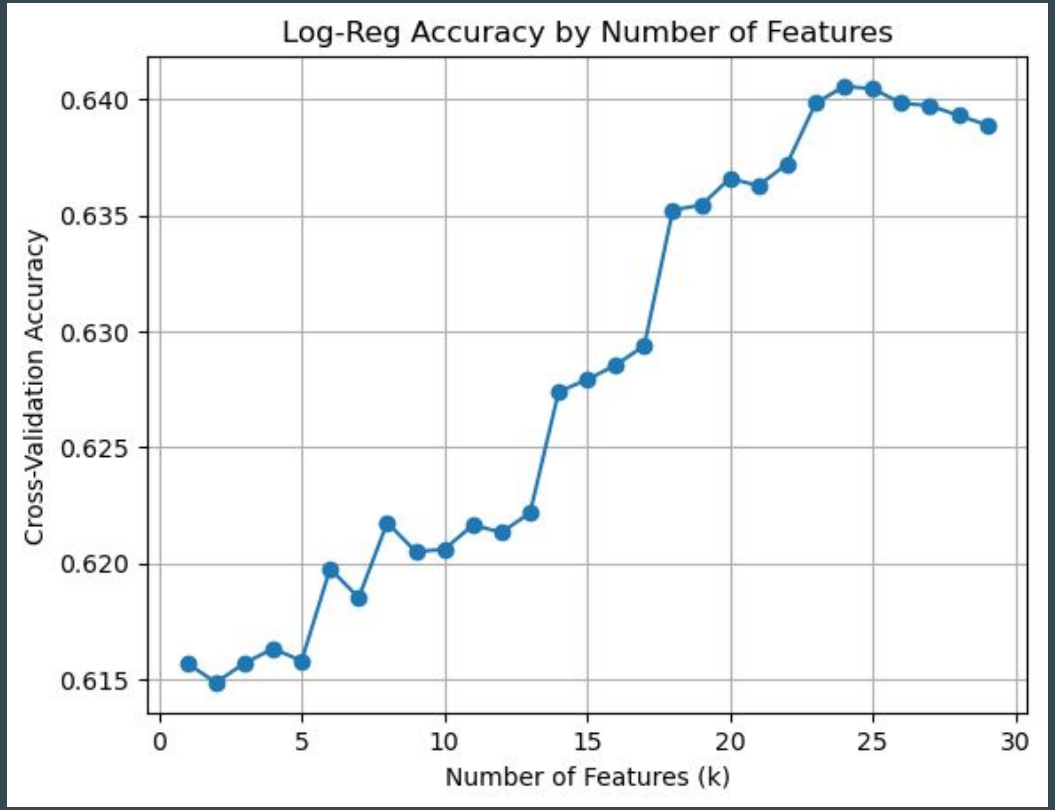
Model Comparison

- Baseline Model:
  - Accuracy: 57.2%
- Models behaved similarly:
  - Accuracy: 59-62%
  - Error Distribution
- Higher error in the first half of a season



Feature Selection

- Despite colinearity, high feature counts increased accuracy



Elo Rating System

Data Requirements:

- Team Elo ratings
- Away/Home Team
- Game Outcome

Assumptions:

- Head-to-head
- Winner: gains rating
  Loser: loses rating
- Zero-sum

Additional Adjustments:

- Margin of Victory
- Seasonal Reset

Performance:

- 65.3% accuracy

Elo Rating System

Data Requirements:

- Team Elo ratings
- Away/Home Team
- Game Outcome

Assumptions:

- Head-to-head
- Winner: gains rating
  Loser: loses rating
- Zero-sum

Additional Adjustments:

- Margin of Victory
- Seasonal Reset

Performance:

- 65.3% accuracy

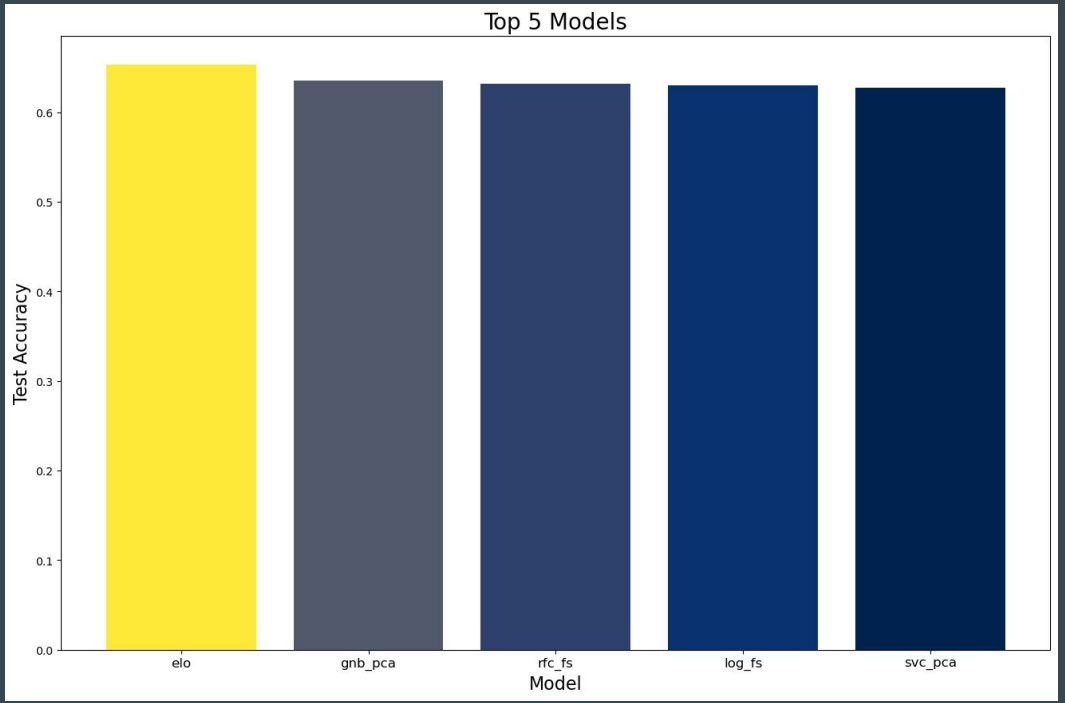
Results

Top Performing Models:

- Elo System 65.3% accuracy
- GNB 63.5% accuracy
- RFC 63.2% accuracy

Top Performing Data:

- 20-game Four Factor dataset



Recommendation:

Elo Rating System

- Highest accuracy model (65.3%)
- Lowest data requirements
- Outperforms ML models

Next Steps:

Data Collection	• Additional seasons • Playoff data
Player Aggregation	• Responsive to roster changes • Opportunity to create player-based metrics
Additional Adjustments	• Improved feature engineering/selection • Ensemble model the incorporates Elo

Questions?

Luke DiPerna

LinkedIn

GitHub