Data Analysis: Kobe Bryant's Shot Selection

Danny Gomes

Dg505@kent.ac.uk

Classifying Kobe Bryant’s Shot Outcomes

Motivation and Background

- Shot success in basketball depends on shot location, timing, and shot difficulty
- Publicly available NBA shot data allows these factors to be analysed statistically
- Kobe Bryant is an interesting case due to the high difficulty and variety of his shot attempts
- Previous basketball analytics research shows that shot location and shot type explain much of shooting efficiency, while contextual factors play an additional role

Research Question

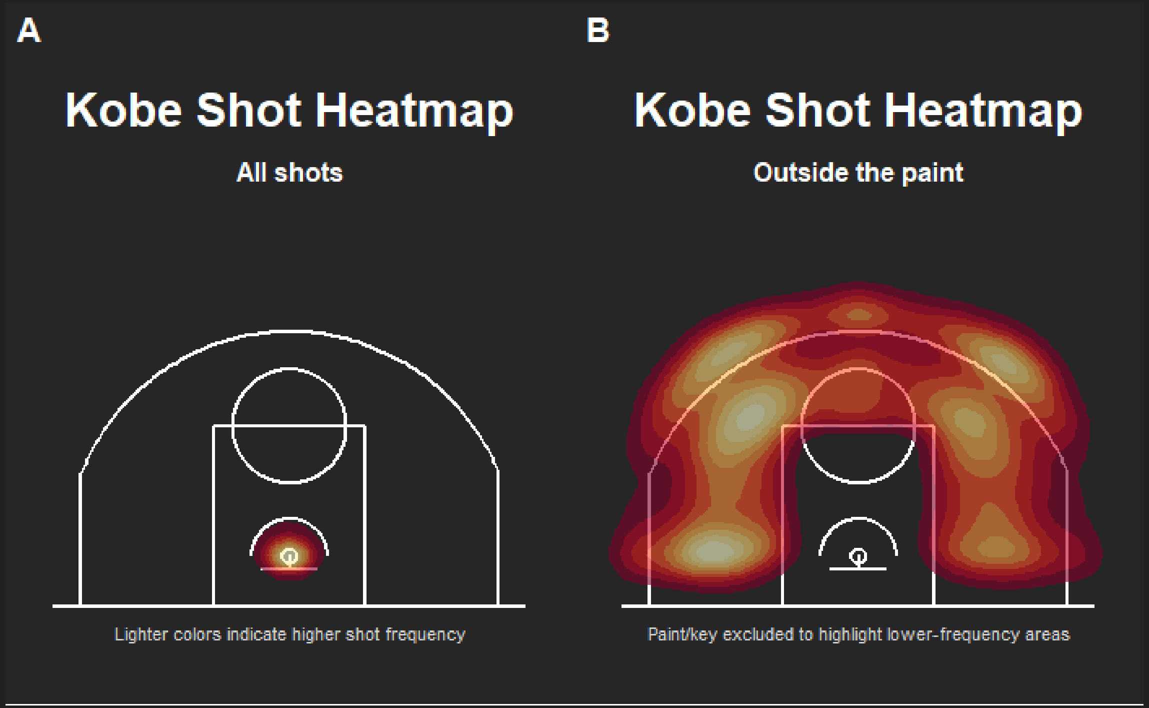
Can Kobe Bryant’s shot attempts be accurately classified as made or missed?

Data Overview

Source of the data is from Kaggle and it has 30697 observations and 25 variables
Contains data across his entire 20 seasons in the NBA from 1996-2016

- Shot-level NBA data from Kobe Bryant's career
- Binary outcome: **Made vs Missed**
- Spatial location (court coordinates, shot zones)
- Temporal context (time remaining, period)
- Shot descriptors (shot type, mechanics, clutch)

The data contains meaningful shot difficulty information, but no defensive or player-tracking variables.



Kobe Shot Heatmap

All shots

Lighter colors indicate higher shot frequency

Kobe Shot Heatmap

Outside the paint

Paint/key excluded to highlight lower-frequency areas

Methodology

- **Objective:** classify shots as *Made* or *Missed*
- **Train–test split:** 70% / 30%
- **Cross-validation:** 5-fold CV on training set
- **Preprocessing:**
  - Continuous variables standardised (training data only)
  - Same scaling applied to test set to avoid leakage

Models considered:

- Logistic Regression (baseline, interpretable)
- Random Forest (non-linear interactions)
- SVM with radial kernel (flexible decision boundary)
- Deep Learning (H2O, higher model capacity)

Model evaluation

- Probabilistic predictions retained

Primary metric: AUC and ROC

- Threshold-independent
- Robust to class imbalance

Accuracy, sensitivity, and specificity used for interpretation

Variable importance to see what metrics had the most impact

Odds ratio for Logistic Regression

Results & Model Interpretation

Model	AUC
Logistic Regression	0.620
Random Forest	0.607
SVM (Radial)	0.606
Deep Learning (H2O)	0.618

Key point:

- All models achieve similar, moderate discrimination
- Deep learning does not outperform logistic regression

Specific Model Findings:

- Logistic regression: odds ratios show large effects for shot type and zone, with many temporal variables insignificant
- Random forest & SVM: variable importance and permutation analysis confirm shot difficulty dominates prediction
- Deep learning: importance scores also prioritise shot mechanics, suggesting the network learns coarse difficulty patterns rather than subtle spatial effects

Comparing the Models:

Logistic regression, used as a classical statistical baseline, performs comparably to the deep learning model, with random forests and radial SVM also failing to provide substantial improvements. This suggests that the main predictive signal is mostly determined by shot difficulty and that increased model flexibility does not lead to meaningful performance gains given the available metrics.

Conclusions & Limitations

Conclusions

- Shot outcomes are only moderately predictable using spatial, temporal, and shot-type information
- Logistic regression performs as well as or better than more complex models, including deep learning
- Increased model complexity provides limited benefit given the available features

Limitations

- No information on defender proximity or pressure
- No data on player balance, movement, or fatigue
- Shot context beyond basic timing is not captured

Future work could incorporate player-tracking and defensive data or sequence-based models to capture richer in-play context, which is likely to be more informative than increasing model complexity alone. This could be done by extracting more advanced statistics that capture defensive pressure, player positioning, and shot difficulty more directly.

As a conclusion, data quality and model understanding matter more than model complexity when trying to predict shots