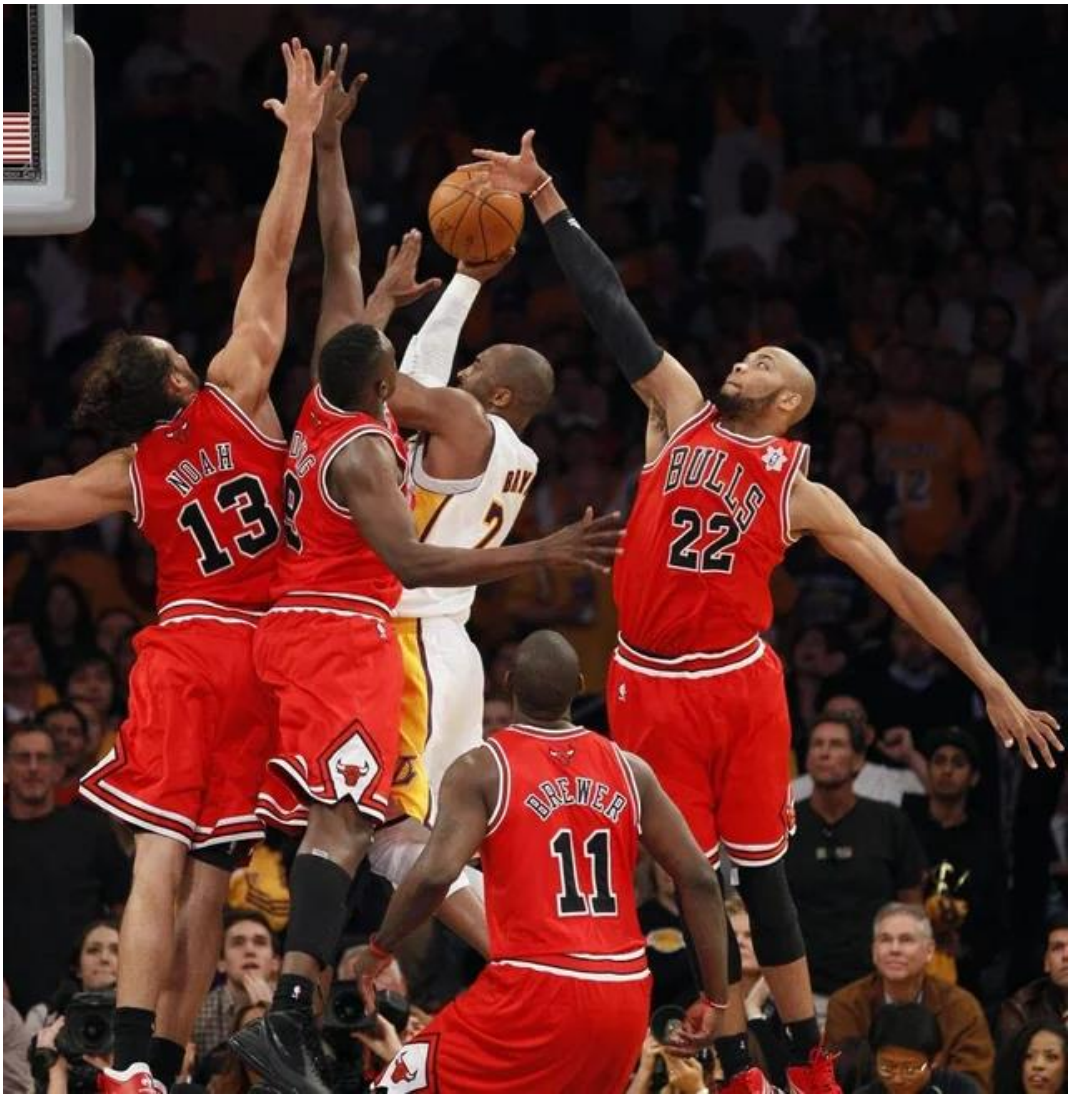
Data Analysis: Kobe Bryant's Shot Selection



Danny Gomes

dg505@kent.ac.uk

https://github.com/DxnGxm/MAST6100-Kobe-Bryant-Shots-Final-Project

Appendix

Introduction

Shot selection and shot making are central components of basketball performance analysis. Advances in data collection have made detailed shot-level information publicly available, allowing researchers to model and evaluate shooting behaviour using statistical and machine learning techniques. A natural question arising from such data is whether shot outcomes can be reliably predicted using observable characteristics such as shot location, remaining game time, and shot type.

Kobe Bryant provides an interesting case study for this task. As one of the most prolific scorers in NBA history, his career features a wide variety of shot attempts taken under diverse conditions. This diversity makes his shot data well-suited for classification analysis while also posing challenges due to the inherent difficulty of many of his attempts.

The aim of this project is to assess whether Kobe Bryant's shot attempts can be accurately classified as made or missed using spatial, temporal, and categorical shot descriptors alone. Rather than attempting to maximise predictive accuracy at all costs, the focus is on comparing classical and machine learning classifiers, understanding their limitations, and interpreting what their performance reveals about the nature of shot success in basketball.

Research Question and Objectives

The central research question of this project is:

Can Kobe Bryant's shot attempts be accurately classified as made or missed using spatial, temporal, and shot-type information?

To address this question, the following objectives are pursued:

1. To see how accurately using different factors have on predicting shot making
2. To see which factors have the most impact
3. To compare model performance using threshold-independent metrics.
4. To interpret the results in the context of basketball decision-making and data limitations.
5. To compare how modern Machine Learning Models compare against baseline statistical models

Data Description and Preprocessing

The dataset consists of shot-level observations from Kobe Bryant's NBA career. Each observation corresponds to a single shot attempt and includes information on:

- Spatial location of the shot on the court
- Remaining game time (minutes and seconds)

- Period and playoff indicator
- Shot type and combined shot mechanics
- Shot outcome (made or missed)

The response variable is binary, indicating whether the shot was made or missed.

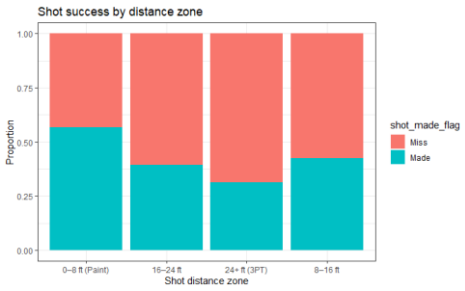
Some shot mechanics, such as bank shots and tip shots, occur far less frequently than more common actions such as jump shots or layups. As a result, estimated effects for these rare categories may be noisier due to smaller sample sizes. However, all shot-type categories contained enough observations to be retained in the analysis and excluding them risked discarding potentially informative structure.

EDA

Exploratory data analysis was conducted to examine the relationship between shot characteristics and shot success, and to inform subsequent modelling decisions. The focus was on identifying broad patterns rather than exhaustive visualisation.

Shot Success by Distance

The bar chart shows the proportion of made and missed shots across broad shot-distance zones. A clear spatial gradient is observed: shots taken within 0–8 feet (paint) exhibit the highest success rate, while three-point attempts (24+ ft) show the lowest. Mid-range shots (8–16 ft and 16–24 ft) fall between these extremes, with noticeably lower success rates than shots near the basket.

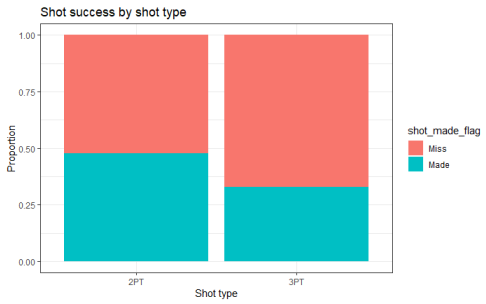


This decline in shot success as distance increases aligns with the intuition that further shots are more difficult to make and suggests that shot distance is a key determinant of shot outcome. The smooth nature of this pattern indicates that distance effects are likely to be well captured by continuous spatial variables, supporting their inclusion in both linear and non-linear models.

Shot Type: Two-Point vs Three-Point Attempts

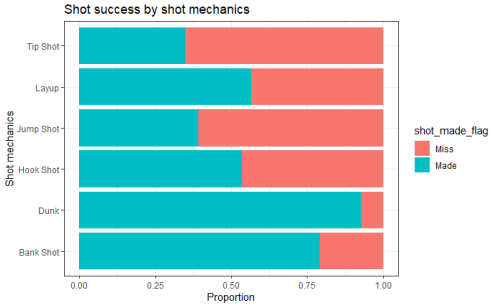
The next bar chart compares shot success rates between two-point and three-point attempts. Two-point shots are converted at a substantially higher rate than three-point shots, reinforcing the strong role of distance and shot difficulty in determining success.

However, the overlap in outcomes across shot types suggests that shot type alone is insufficient for accurate classification. This motivates the use of additional spatial and contextual variables, rather than relying solely on categorical shot-type indicators.



Shot Mechanics

This one presents shot success proportions across different shot mechanics. High-percentage actions such as dunks and layups exhibit markedly higher success rates, while jump shots and tip shots are converted less frequently. Some mechanics, such as bank shots, show high success rates but occur relatively infrequently, implying greater variability and less stable estimates.



Despite these differences, there is substantial overlap between made and missed shots across most mechanics. This suggests that while shot mechanics contain useful signal, they are unlikely to fully explain shot outcomes in isolation, particularly without information on defensive pressure or player positioning.

All these observations motivate the use of classification models that incorporate multiple predictors simultaneously and justify comparing linear and non-linear approaches. The lack of sharp univariate separation also suggests that achievable predictive performance may be inherently limited by the available features.

Train–Test Split and Cross-Validation

To evaluate model generalisation performance, the dataset was split into training (70%) and test (30%) subsets using stratified sampling. Stratification ensures that the relative frequencies of made and missed shots are preserved across both subsets, with approximately 55% missed shots and 45% made shots in each. This is particularly important given the moderate class imbalance in shot outcomes, as unstratified random sampling could lead to biased performance estimates and unstable probability calibration.

The 70/30 split represents a trade-off between providing sufficient data for model training while retaining a large, independent test set for reliable evaluation. By preserving the original class distribution, the test set more accurately reflects the true data-generating process, allowing performance metrics such as AUC, sensitivity, and specificity to be interpreted meaningfully in the context of real shot outcomes.

A 5-fold cross-validation procedure was applied within the training set to support model comparison and tuning. The training data were partitioned into five equally sized folds. In each iteration, four folds were used for training and one-fold for validation, with the process repeated until each fold had served as the validation set once.

This approach reduces variance in performance estimates and ensures that results are not driven by a particular random split.

Feature Scaling and Data Preparation

Prior to model fitting, all continuous predictors were standardised using mean centering and scaling to unit variance, with parameters estimated using the training data only. This step ensures that predictors measured on different scales contribute comparably to the model and improves numerical stability, particularly for distance-based methods such as support vector machines.

The same scaling parameters were subsequently applied to the test set to avoid information leakage.

Model performance was assessed primarily using the area under the receiver operating characteristic curve (AUC). AUC measures a classifier's ability to distinguish between classes across all possible probability thresholds.

This metric was preferred over raw classification accuracy for several reasons:

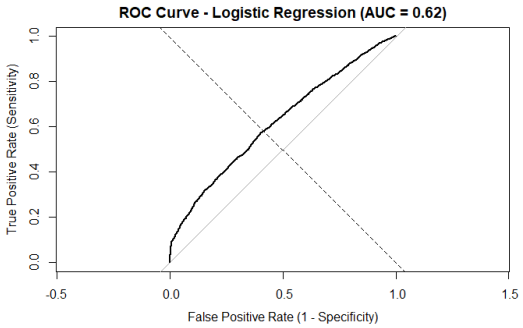
- Accuracy depends on an arbitrary threshold choice.
- AUC is threshold independent.
- AUC is less sensitive to mild class imbalance.
- The task involves probabilistic prediction rather than deterministic classification.

Class probabilities were retained for all models to enable ROC-based evaluation.

Logistic Regression

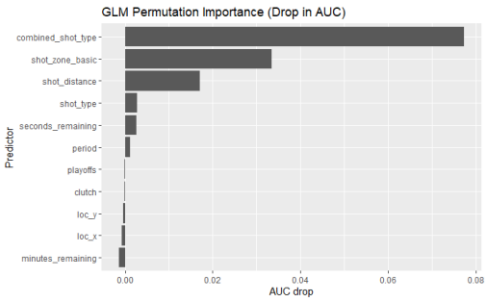
Logistic regression was used as a baseline classification model due to its interpretability and widespread use in binary outcome modelling. The model included spatial coordinates, temporal variables, and categorical shot descriptors as predictors.

Initial ROC analysis suggested poor discrimination; however, further inspection revealed that the predicted probabilities were inversely aligned with the class labels used in the ROC computation. After correcting this alignment, the logistic regression achieved an AUC of approximately 0.62, indicating moderate discriminative ability. While this performance is clearly better than random classification, it also highlights the inherent difficulty of predicting shot outcomes using location, time, and shot descriptors alone.

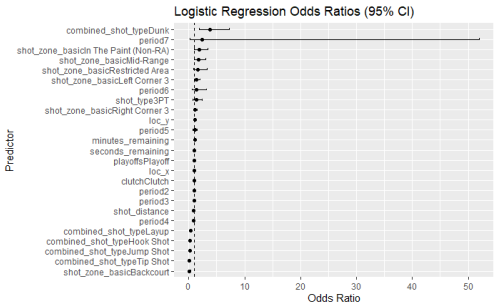


Classification accuracy at a fixed 0.5 threshold was relatively low (59%), which is expected given the threshold-dependent nature of accuracy and the low base probability of shot success. Using Youden's J statistic, an optimal classification threshold of approximately 0.58 was identified, yielding a sensitivity of 57% and a specificity of 60%. This threshold improves the balance between true positives and true negatives, although overall model evaluation is based primarily on AUC, which reflects ranking performance across all thresholds rather than any single operating point.

Permutation-based importance analysis revealed that shot mechanics and shot zone dominate the model's predictive ability. Permuting *combined shot type* resulted in the largest drop in AUC (0.077), followed by *shot zone* (0.033) and *shot distance* (0.017). In contrast, temporal variables and raw spatial coordinates produced negligible changes in AUC when permuted, suggesting that much of their information is redundant once shot type and zone are known. This indicates that discrete contextual descriptors are more informative for ranking shot difficulty than continuous timing or location variables.



Interpretation of the logistic regression coefficients supports these findings. Dunk attempts exhibited a large positive effect on shot success (odds ratio ≈ 3.76, 95% CI excluding 1), while jump shots, hook shots, tip shots, and layups were associated with significantly lower odds of conversion. Shots taken from the backcourt had extremely low odds of success (OR ≈ 0.06), reflecting their near impossibility in practice. Spatial effects were present but modest; the vertical court coordinate (*loc_y*) showed a small but statistically significant positive association with success, while horizontal location (*loc_x*) was not significant.



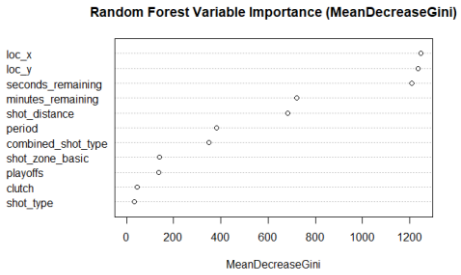
Temporal effects were generally weak. Both minutes remaining and seconds remaining showed small positive associations with shot success, suggesting slightly improved efficiency earlier in possessions. However, clutch situations and playoff games were not statistically

significant, indicating that once shot difficulty and location are accounted for, these contextual indicators do not materially affect conversion probability. Overall, the results suggest that while logistic regression captures meaningful structure in shot outcomes, much of the variability in shot success remains unexplained, reflecting the stochastic nature of shooting and the absence of defensive and situational variables.

Random Forest

Random forests extend decision trees by aggregating multiple trees built on bootstrapped samples and random feature subsets. This allows the model to capture nonlinear relationships and interactions without strong parametric assumptions.

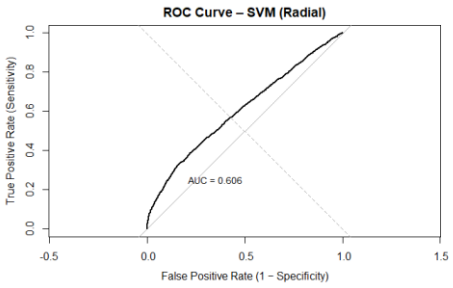
Variable importance measures indicated that shot location and temporal context were the strongest predictors of shot success. This aligns with basketball intuition, as distance from the basket and time pressure influence both shot difficulty and decision-making.



Despite its flexibility, the random forest did not dramatically outperform logistic regression in terms of AUC. This suggests that nonlinear interactions among the available features contribute only limited additional predictive power.

Support Vector Machine with Radial Kernel

A support vector machine (SVM) with a radial basis function kernel was used to allow for non-linear decision boundaries between made and missed shots. Unlike logistic regression, the RBF kernel can model interactions between predictors by mapping the data into a higher-dimensional space.



The SVM achieved an AUC of around 0.61, which is similar to the performance of the other models. However, it did not lead to a clear improvement in classification performance. This suggests that the available predictors may not contain strong non-linear patterns, or that important factors affecting shot success are not included in the dataset.

Using a 0.5 classification threshold, the SVM shows high specificity (0.85) but low sensitivity (0.32). This means the model is much better at correctly predicting missed shots than made shots. This behaviour is reasonable given that missed shots are more common and that made shots may depend on additional factors that are not captured by the available variables.

The model was fitted using moderate values for the cost parameter and a default gamma value to avoid overfitting. Overall, the results suggest that while non-linear models can capture

additional structure, the main limitation in performance comes from the lack of richer contextual information rather than the choice of model.

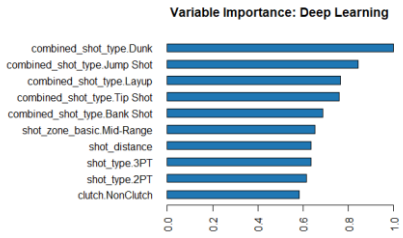
Deep Learning

To investigate whether higher-capacity models could capture nonlinear interactions beyond those identified by classical and kernel-based classifiers, a deep neural network was trained using the H2O framework. The model employed a feed-forward architecture with three hidden layers (128, 64, and 32 neurons). Dropout regularisation was applied both at the input layer (0.1) and across all hidden layers (0.3) to mitigate overfitting by preventing the network from relying excessively on any single subset of predictors.

The choice of a multi-layer architecture was motivated by the hypothesis that shot success may depend on nonlinear combinations of spatial location, shot mechanics, and contextual variables that are not easily captured by linear models. All numeric predictors were standardised before modelling. The network was trained for 50 epochs, which was a stable amount without making the time to compute too long

Model performance was evaluated on a held-out test set that was not used during training. The deep learning model achieved an AUC of approximately 0.61, indicating moderate discriminative ability. This performance is comparable to the logistic regression and other models considered, suggesting that the network can rank shot attempts better than chance but does not achieve strong separation between made and missed shots.

At a 0.5 classification threshold, the model achieved an accuracy of approximately 0.61, with a high specificity (0.85) but low sensitivity (0.32). This indicates that the model is conservative in predicting made shots: it correctly identifies a large proportion of missed shots but fails to capture many true makes. Such behaviour is consistent with a low base rate classification problem, where models that prioritise, certainty tend to favour the majority or more predictable class. As with other models, overall evaluation therefore focuses primarily on AUC rather than any single threshold-dependent metric.



Variable importance measures derived from the H2O deep learning model provide insight into which predictors the network relied on most when forming its predictions. The results indicate that shot mechanics dominate the learned representation, with dunk attempts emerging as the most influential predictor, followed by jump shots, tip shots, layups, and bank shots. Among spatial descriptors, shot zone categories such as mid-range and non-restricted-area paint shots were also assigned substantial importance.

The prominence of discrete shot-type indicators suggests that the network primarily distinguishes shot outcomes based on coarse measures of shot difficulty rather than subtle continuous effects. This finding aligns closely with the permutation importance analysis and odds-ratio results obtained from logistic regression, reinforcing the conclusion that shot mechanics and broad location categories carry the strongest signal in the available data. In contrast, several contextual indicators (e.g. clutch status, playoff indicator, and missingness flags) were assigned negligible importance, implying that these variables contribute little once shot type and zone are known.

It should be noted that variable importance in deep learning models is inherently heuristic and does not admit the same causal interpretation as coefficients in a logistic regression. Nonetheless, the consistency of the importance rankings with simpler models increases confidence that the network is learning meaningful structure rather than noise.

Comparison

Across all models considered: logistic regression, random forest, support vector machines, and deep learning. Predictive performance remained broadly similar, with no approach achieving strong discrimination. Despite its greater flexibility and capacity for hierarchical feature learning, the deep learning model did not substantially outperform simpler linear or kernel-based methods. This indicates that predictive performance is primarily constrained by the information content of the available features rather than model complexity. In particular, the absence of defensive pressure, player balance, and in-play contextual variables likely limits the ability of any model to accurately predict shot outcomes.

While non-linear models marginally improved performance relative to logistic regression, gains were modest, and the deep learning model in particular exhibited high specificity but low sensitivity, reflecting difficulty in confidently identifying made shots. Overall, these results suggest that increasing model complexity yields diminishing returns in this setting, and that meaningful performance improvements would require richer contextual data rather than more sophisticated architectures.

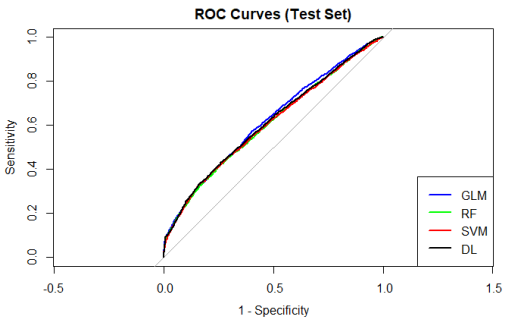


Figure compares the ROC curves for all models evaluated on the test set. All four classifiers achieve similar ROC profiles, with curves lying only modestly above the diagonal reference line, indicating moderate but limited discriminative ability across approaches. Non-linear models (random forest, SVM, and deep learning) exhibit slight improvements over logistic regression, but these gains are marginal, reinforcing the conclusion that predictive performance is constrained more by the available features than by model choice.

From an interpretability perspective, logistic regression provides the clearest insights, allowing direct assessment of coefficient signs, odds ratios, and statistical significance. These results show that shot mechanics and shot zone are the dominant determinants of shot success. This finding is consistent with variable importance analyses from the random forest and deep learning models, which also rank combined shot type (e.g. dunks, jump shots, layups) and broad shot location categories as the most influential predictors. In contrast, temporal and contextual variables such as clutch status, playoff indicator, and remaining time contribute relatively little once shot difficulty is accounted for.

Overall, the similarity of the ROC curves across models suggests that increasing model complexity yields diminishing returns in this setting. While more flexible models can capture non-linear interactions, they do not substantially outperform simpler, more interpretable methods, highlighting that meaningful performance improvements would likely require richer contextual data rather than more sophisticated classification algorithms.

Discussion

The results indicate that Kobe Bryant's shot outcomes are only weakly predictable using spatial, temporal, and shot-type information alone. While location and timing clearly influence shot success, much of the variance remains unexplained.

This reflects the inherently unpredictability of basketball shooting within games and the absence of important contextual variables such as defender proximity, player fatigue, and in-play dynamics. The findings highlight the limits of purely outcome-based modelling when key causal factors are unobserved. Especially when looking at a player like Kobe Bryant who takes difficult shots which may lead to more variance

Several limitations should be noted:

- Defensive information is not included.
- Shot context beyond basic timing is unavailable.
- Player intent and in-game strategy cannot be observed.

Future work could incorporate tracking data, defender distances, or sequence-based models to better capture shot difficulty and decision-making processes.

Conclusion

This project evaluated the feasibility of classifying Kobe Bryant's shot attempts as made or missed using commonly available shot-level features. While modest predictive performance was achieved, the results suggest that such features alone are insufficient for highly accurate classification. The findings underscore the importance of contextual information in sports analytics and demonstrate the value of careful model evaluation beyond raw accuracy.