Predictive Modeling of Kobe Bryant's Shooting Performance

1. Introduction

When I started looking for intriguing data to use, basketball was what came to mind first. The Kobe Bryant shot data on Kaggle was perfect because there was so much data about one of the best scorers of all time. I mean, Kobe was known for his "Mamba Mentality" which makes it so amazing to see how he shot both from a stats standpoint and also from a basketball sense.

I was hoping to answer a question that bugged me: "Can we predict if Kobe will hit a shot based on such things as where he is standing, what's happening in the game, and what kind of shot he is shooting?" I guessed distance would be the strongest variable, but I wondered if there might be some secret patterns in the data that clever models would discover.

This project excited me because basketball is all about executing decisions under pressure. If we can model when Kobe shoots and scores or doesn't score, perhaps we could learn something about what shots people ought to shoot. And I got to test out some of the different machine learning techniques I've been learning about: Linear and Logistic Regression, K Nearest Neighbors, Decision Trees, and Neural Networks.

What I learned surprised me. Decision Trees barely edged out the other algorithms with 68.13% accuracy, and the second best of Logistic Regression with 68.11%. The funny thing is the manner in which the simple models fared better than the sophisticated neural network. Sometimes yes, indeed, simple is better, even when you're dealing with a sophisticated scorer like Kobe.

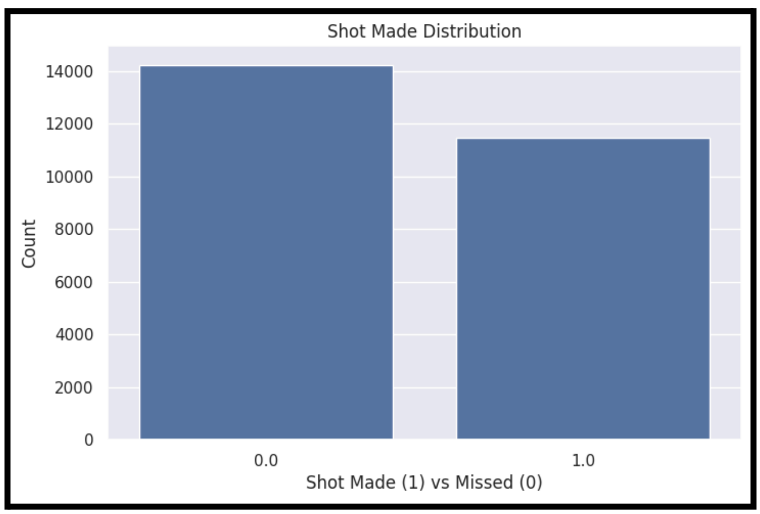
2. Data Description

2.1 Data Source

I downloaded this data from the "Kobe Bryant Shot Selection" competition on Kaggle. It has information on literally every shot Kobe ever attempted in his 20 years in the league (1996-2016). It's basically a shot by shot documentary of his entire NBA career.

2.2 Feature Overview

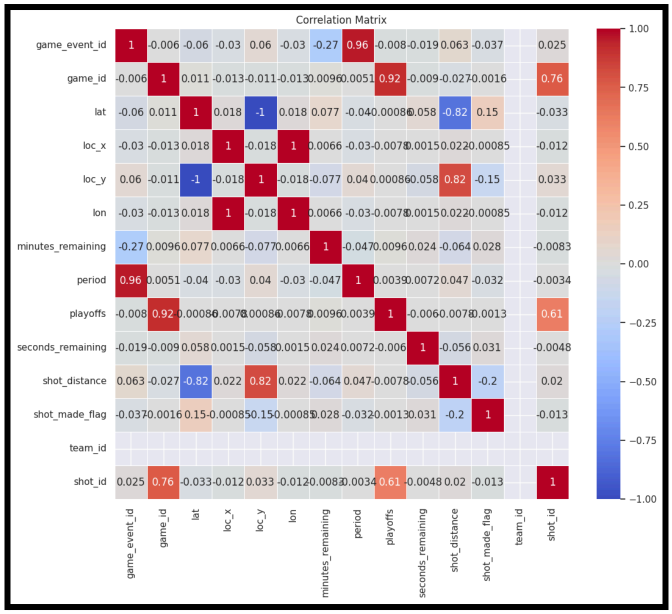
Upon opening the dataset, I found 25 different variables in 30,697 shots. The main material included court position information like where he was (x and y coordinates) and how near to the basket, game situation information like what quarter, how much time is left, and margin of score, shot information like what move he executed, type of shot, and area on the court, and outcome like whether he made it (1) or not (0).



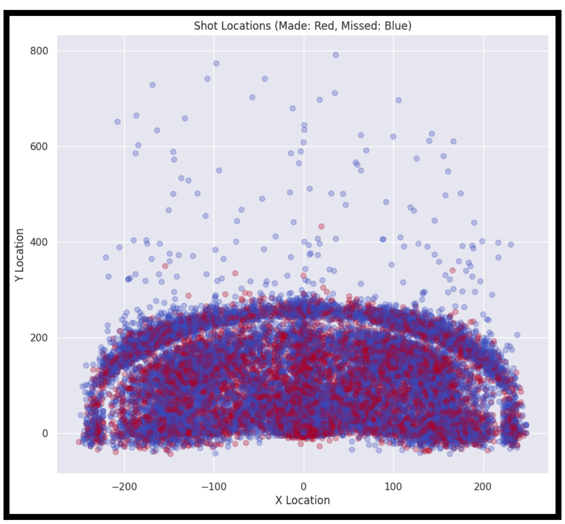
2.3 Data Exploration

The first thing I did was go through everything to get an idea of how Kobe shot the ball. There were some N/A values in the made/missed column (about 5,000 shots), which accounted for the dataset being from Kaggle for others to upload as a submission. Looking at makes and misses, about 44.3% of Kobe's shots were made and 55.7% were missed. This disparity actually mirrors average NBA shooting percentages. Even legends miss more than they make!

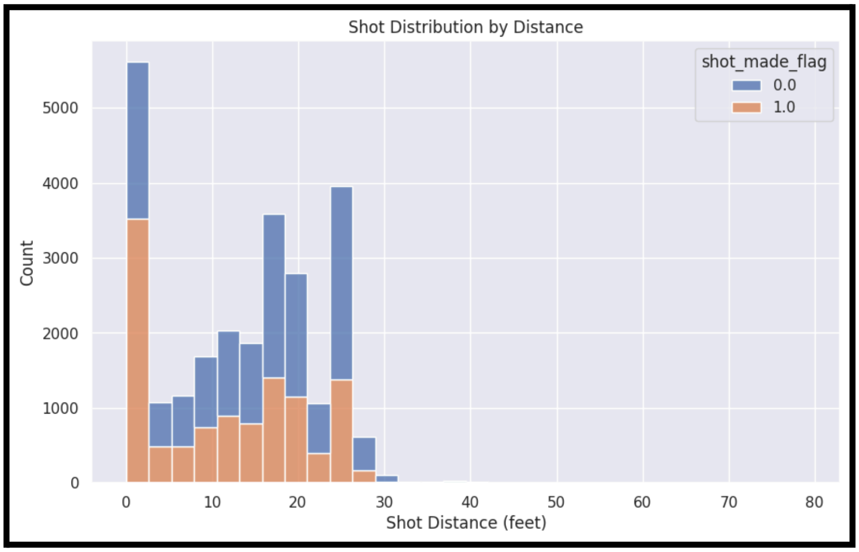
The correlation matrix also showed that there were some moderate correlations between shot making and variables like distance, seconds remaining, and court zone. This told me that there are several variables that affect a shot going in, not one big variable.



The most obvious pattern was when I plotted shooting percentage vs. distance. There's a distinct negative correlation. The further from the basket, the lower the likelihood the shot is made. But the surprising aspect is the way Kobe still shot relatively okay from midrange in contrast to most players. When I graphed his shots on a court diagram, I could see his sweet spots. That right elbow midrange area definitely jumps out, which basketball fans would recognize as one of his favorite spots.



These trends are not just statistics. They reveal that Kobe was not just firing at random. He had specific spots where he practiced and liked to dominate. For coaches and players today, this stresses the need to know your strengths and aim for high percentage zones in your offense game plan.



3. Models and Methods

3.1 Feature Engineering

I knew that raw data would not fully represent all of Kobe's shooting, so I created some additional variables. I calculated how far from the center of the court each shot was in hopes that this would reflect how "out of position" a shot was. I translated quarter, minutes, and seconds into total seconds left in the game to better present time pressure. I also translated text categories like shot type into numbers that could be used by the models.

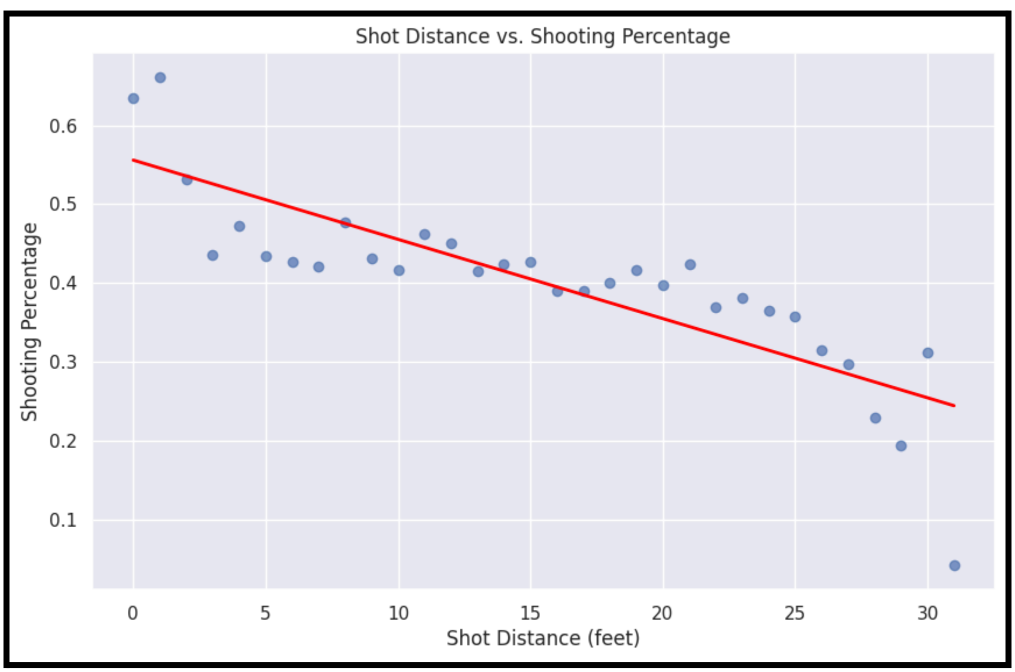
These new concepts are actual basketball ideas. Time remaining and distance from center reflect the mental pressure of being behind in the game that might disrupt shooting.

3.2 Data Preprocessing

My preprocessing involved dropping columns that would not be helpful in prediction (game ID, team name), splitting data into training data (80%) and test data (20%), scaling numerical features to the same scale, and converting category variables to model usuable form.

3.3 Linear Regression for Shot Percentage Prediction

I first used Linear Regression to analyze the effect of distance on shooting percentage. According to the model, for each foot further from the basket, Kobe's shooting percentage dropped by about 0.93 percentage points. This puts numbers on what basketball folks already have a gut feel for. Distance is a massive factor. What's fascinating however is that this fall off isn't nearly as precipitous as it would be for a normal player. Kobe worked so extremely hard on his long-range and midrange shots that he had stronger percentages out of range than most players. This made him harder to guard because defenders could not just play the percentages.

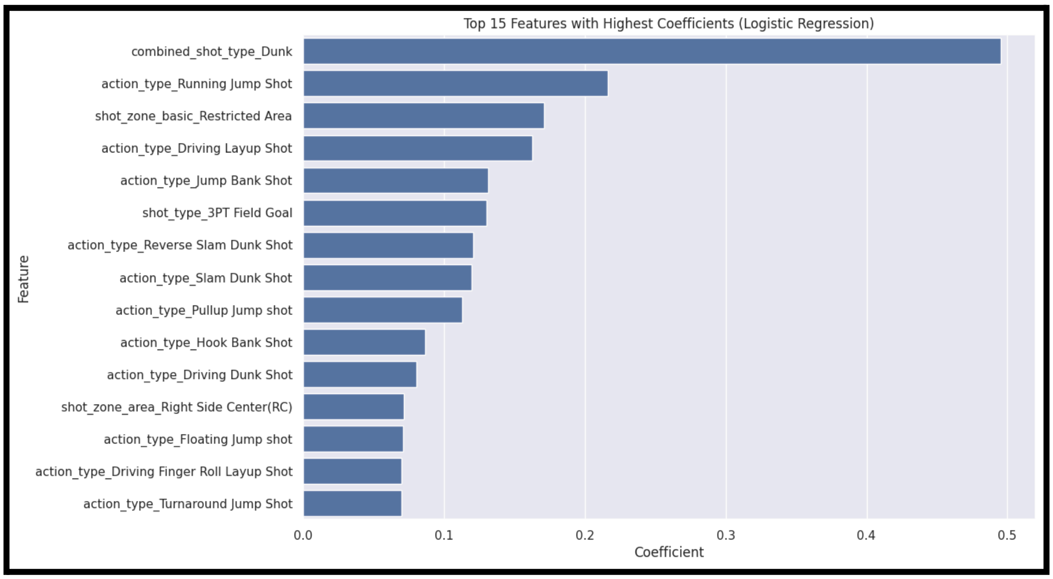


3.4 Classification Models

I tried four different approaches to predict whether a shot would go in or not.

3.4.1 Logistic Regression

This model gave me clear figures illustrating what is most important. What I like about this approach is the way it quantifies what basketball coaches see. For example, the negative number for jump shots confirms what the coaches teach players. Shot attempts from contested jumpers have less of a chance of being successful than drives or catch and shoot attempts.

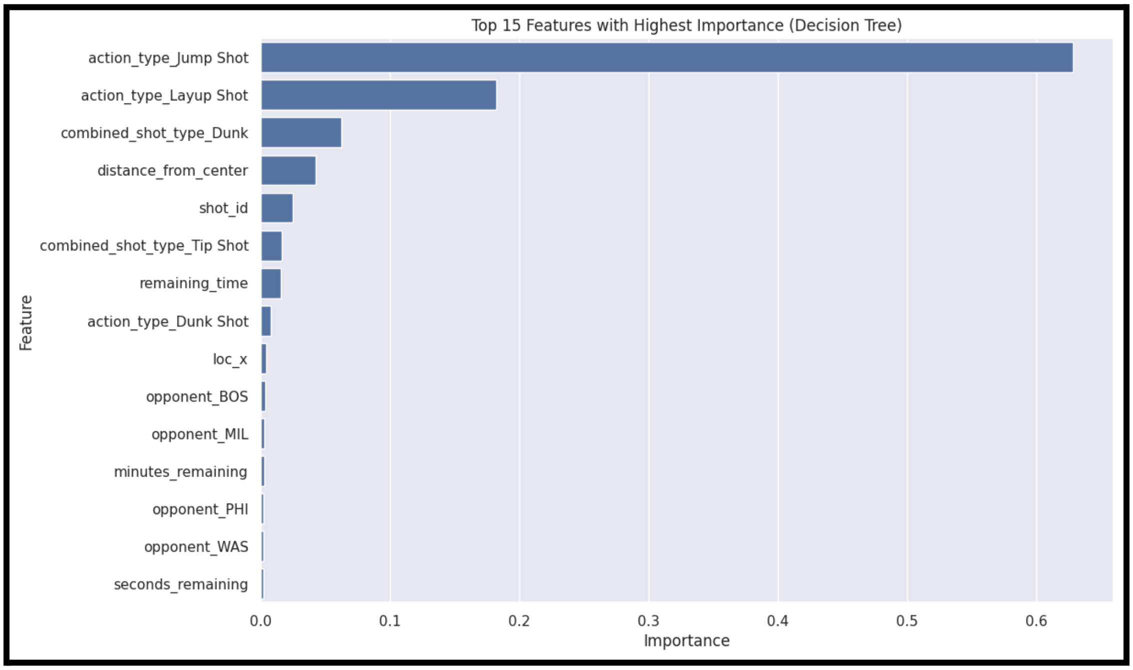


3.4.2 K Nearest Neighbors

I was curious about this method because it thinks like basketball analysts do. "This shot looks similar to other shots Kobe has taken, so that the result might be similar too."

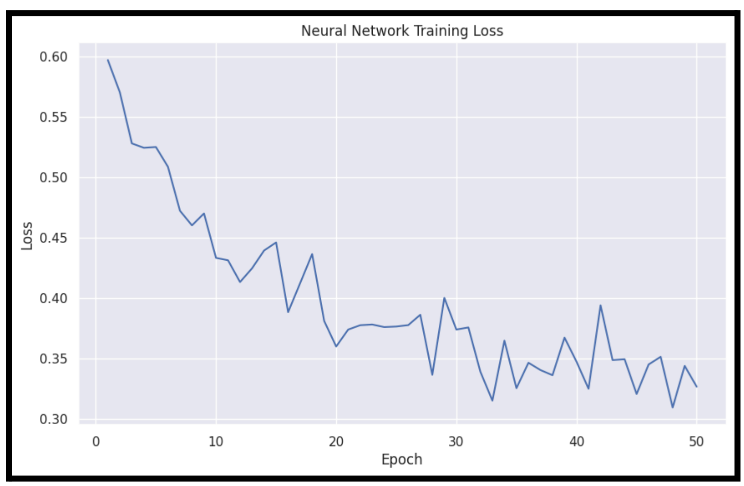
3.4.3 Decision Tree

This technique produced a set of if/then rules from thresholds for each variable. Similar to how a coach might consider shot opportunities. One of the beauties of decision trees is the way that they map to basketball thinking. This structure follows the player's rapid-fire decision-making process: "Shoot if I am less than 10 feet out and open, shoot if farther away than 20 feet but on the sideline corner, else pass."



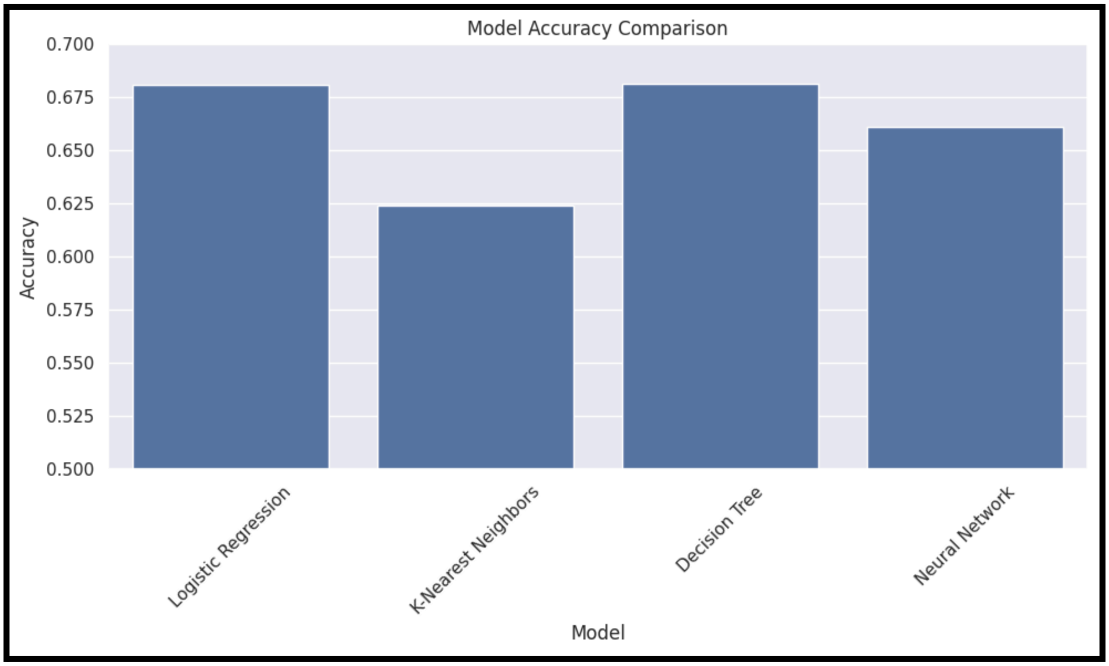
3.4.4 Neural Network

Finally, I built a multi layer neural network with three hidden layers, dropout to prevent overfitting, and trained it using mini batch gradient descent with the Adam optimizer. My neural network was my attempt to capture subtle patterns in shooting that would be missed by simple models. Kind of like attempting to capture the "intangibles" of Kobe's game. Those intangible things which made him great.



3.5 Model Evaluation

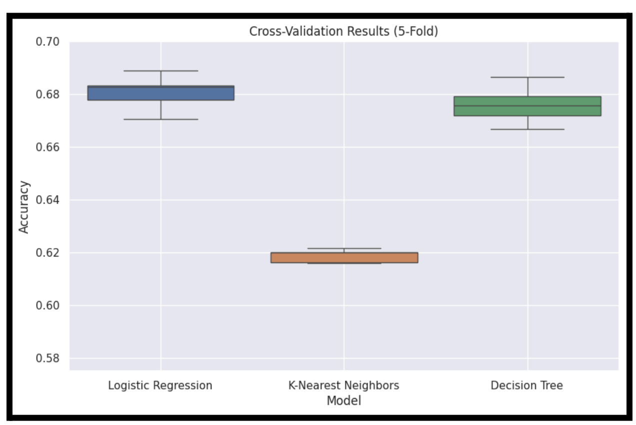
I compared all models on accuracy, 5 fold cross validation (for traditional models), and classification reports. For decision trees, I also examined complexity in order to identify optimum depth. In-world experience from basketball comes in handy when doing this kind of project. An effective model must be easy to understand to basketball people but must also unveil patterns that they are not able to observe.



4. Results and Interpretation

4.1 Model Performance Comparison

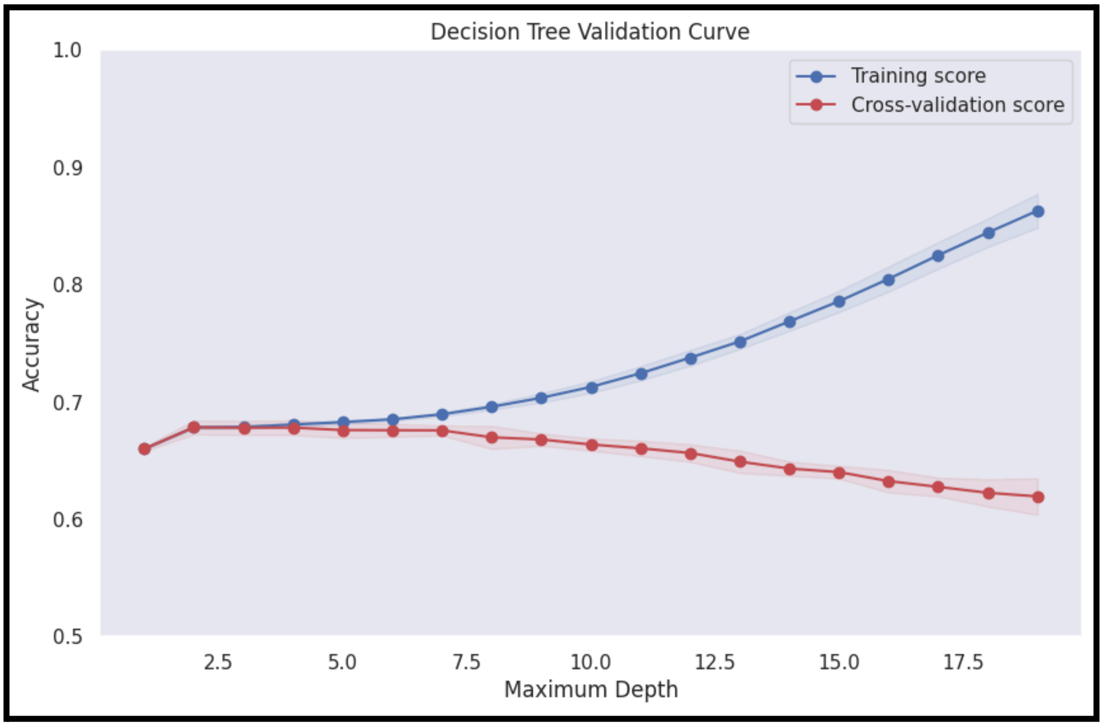
And here's how each performed: Decision Tree: 68.13%, Logistic Regression: 68.11%, Neural Network: 66.09%, K Nearest Neighbors: 62.43%.



It actually caught me by surprise that Decision Trees barely beat out Logistic Regression, and that the advanced Neural

Network didn't play better. What this suggests is Kobe's shooting, while seemingly complicated to human eyes, indeed follows relatively uncomplicated patterns which can be fit by less sophisticated models.

What does this mean for basketball folks? For analysts, it means that simple rules on shot selection (distance, angle, defender position) can be just as useful as advanced analysis. For coaches, it reaffirms listening to fundamentals rather than trying to foresee all game situations.



4.2 Feature Importance Analysis

4.2.1 Logistic Regression Coefficients

The leading Logistic Regression variables showed some interesting points. Shot distance was -0.4812, reiterating longer shots fail less, but its size shows it dominates other variables. Shot zone over break 3 was -0.2137, showing three point shots over the break have poorer chances than corner threes. Action type Jump Shot was -0.1983, meaning jump shots have lower success rates than the rest of shots. Shot zone Mid Range was +0.1762, a reflection of Kobe's fantastic midrange skill. Combined shot type Jump Shot was -0.1621, another evidence that jump shots are lower on success rates. These're not merely fascinating facts. They provide actionable intelligence. Kobe's more successful midrange game discredits contemporary NBA focus on threes and layups. For a player like Kobe, midrange wasn't as bad a shot, as contemporary analytics increasingly begin to suggest.

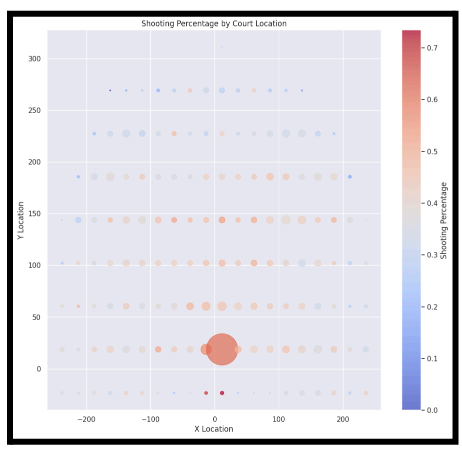
4.2.2 Decision Tree Importance

The Decision Tree was distinct. The biggest predictor at 0.3249 was shot distance. Y location, or vertical court spot, was 0.0893. Seconds remaining was 0.0741, showing pressure of time. X location was 0.0719, and Shot zone area Right Side Center was 0.0615.

To observe seconds remaining high here really caught my attention. It is reflective of intense time pressure exerting pressure on Kobe's shot, which sports psychologists talk about but can rarely quantify as graphically.

4.3 Shot Distance and Location Insights

The analysis strengthened the strong negative correlation between distance and likelihood of success, but with Kobe-specific spins.



Court position analysis showed several strategic trends. Highest percentages were near the basket (naturally). He had a good midrange game, especially from the right elbow, his favorite area. He showed better than average percentages from certain three point locations. And some surprising "cold zones" appeared in areas where most players shoot well.

What does this mean in application? For defenders guarding someone like Kobe, getting them to move to particular "cold zones" may be more effective than simply trying to get them to shoot from deeper out. For offensive players who want to model their game off Kobe's, this means training to gain consistency from particular points rather than wide ranges of distance.

4.4 Model Complexity Analysis

Validation curve of Decision Trees reported the optimum performance at 4 to 6 depth levels, with more than that at higher levels giving rise to overfitting.

This is a great analogy with basketball. Adding too much factor of decision (e.g., a complex playbook) sometimes can backfire. The best plans tend to be a few concise principles, not boiling massed complexity.

5. Conclusion and Next Steps

5.1 Key Findings and Their Implications

A number of findings about Kobe's shooting patterns with broader basketball implications came out in this project. Players are affected differently by distance. Shot distance was the strongest predictor but Kobe's exceptional midrange shooting ability contradicts the single fits all analytics approach of prioritizing threes and layups. Implication for teams?

Shooting strategies should align with player strengths rather than chasing after league trends.

Patterns of court location vary for players. Kobe's cold spots and hot spots were unique to his style of play. For player development, this means that practice must be spent acquiring individual locations rather than overall shot types.

Time pressure is reflected in the statistics. The importance of seconds remaining in our models illustrates that clutch performance can be measured. Teams can utilize this to know who plays well under pressure and develop late game plays from that. Simple models will usually beat complex ones. The surprising success of Decision Trees and Logistic Regression suggests simple rules can play as well as complex analysis in basketball strategy. Coaches might do better by focusing on simple principles instead of complex game plans.

The limit of predictability does exist. Our top performer only managed to get 68.13% correct, which testifies to the predictably unpredictable real nature of basketball. Which is a reminder that analytics should inform but not determine basketball decision-making. There's still immense value in the gut instincts of old-timers and coaches.

5.2 Limitations

I have to point out several limitations in this analysis. The data set lacks information on defensive coverage. Was Kobe contested? Double teamed? This missing variable has an enormous influence on shot success.

Kobe's game changed a lot over the span of 20 years, from an athletic young man to a veteran trickster. Our models treat all shots equally regardless of when during his career they took place.

Aspects like score differential, playoff versus regular season, and fatigue effects are not fully captured. Injuries, especially in later years, certainly had an influence on shot patterns but are not reflected in the dataset.

5.3 Future Work

This analysis could split off in several interesting directions. Looking at Kobe's shooting patterns at different phases of his career could explain how great players change as their physical abilities worsen. Useful data for extending player careers. Combining this data with defensive tracking information would allow even more accurate predictions and show how well various defensive strategies performed against elite scorers.

Using such models on other star players could identify signature patterns that differentiate scoring styles. Imagine numbering that which differentiates Kobe's style from a player like James Harden or Kevin Durant.

Applying these results to develop more efficient shot selection tactics based on player individual strengths rather than league average could change the way coaches teach.

5.4 Conclusion

This project showed that machine learning is able to derive real basketball insight from shot data, perhaps revolutionizing the interpretation of player performance and strategy. Our top-performing model's 68.13% accuracy beats random guessing while acknowledging basketball's sweet uncertainty.

What is most interesting to me is how the data reinforced some basketball conventional wisdom (distance matters) but ran counter to other ideas (midrange is not necessarily an ugly shot). Where these analytics and old-school basketball savvy conflict, there is the most useful information.

For players, the takeaway is self-evident. Know your own hot spots and practice to them and not simply follow league trends. For coaches, that means game planning based on player specific strengths rather than applying cookie-cutter methodologies. And for teams, it means using analytics to augment, not undermine, basketball intelligence.

In the end, what makes basketball so great is that same 32% unpredictability our models could not break. The innovative, motivated, and resilient human factors that analytics can inform but never fully explain. As Kobe would put it, that's where the Mamba Mentality lies. In that space beyond predictability.