Final Project: Predicting Kobe Bryant's Shots

Belino Xhafa, Christopher Chen, Josh Felizardo, & Wenzhe (Harry) Xue

May 4, 2016

1 Research Questions

- What factors influenced whether Kobe Bryant made a given shot? Is Kobe especially talented at a particular type of shot?
- Kobe Bryant is a renowned "clutch" player. Is there a statistically significant difference between Kobe Bryant's accuracy when there is less than one minute remaining in a given game compared to his overall accuracy? What about the fourth quarter and overtime?
- Is there a statistically significant relationship between Kobe's game performance and the the ultimate result of a given game (on average)?
- Did his major injuries have a significant impact on his performance in the short/long term?
- Did Kobe's performance differ depending on whether he was playing at home or away?

2 Motivation

Given the recent retirement of NBA superstar Kobe Bryant, known to be a "clutch" player, we believe researching the factors influencing (and influenced by) Kobe's shooting accuracy will prove to be an interesting application of the statistical techniques learned in this class.

3 Hypotheses

4 Methods

4.1 Dataset Used

The dataset used as the basis for our analysis was downloaded from the Kobe Bryant Shot Selection Kaggle competition (`https://www.kaggle.com/c/kobe-bryant-shot-selection`).

4.2 Overview

Our code is divided into the following files as follows:

- **analysis.R** contains preliminary exploratory analysis of the datasets
- **cleaning.R** contains the commands we ran to clean/preprocess the data

- **hypothesis_testing.R** contains various hypothesis tests for statistical significance
- **models.R** contains the models we created based on the data

Datasets:

- **raw.csv**: Raw data downloaded from Kaggle
- **cleaned.csv**: Processed data with our modifications and dummy variables added
- **clutch_shots.csv**: Data on Kobe's shooting performance in clutch situations
- **win_loss.csv**: Win-loss data scraped from `landofbasketball.com`

4.3 Data Preprocessing

Fortunately since we were dealing with a Kaggle dataset, for the most part it was well-formatted to begin with. The main changes we made in preprocessing were to remove the rows where there was no record of whether a shot was made (these were the rows to be predicted in the Kaggle competition, but were not very useful for our purposes), add dummy variables and add in win-loss data.

4.4 Addition of Win-loss Data

Unfortunately the Kaggle dataset did not include data on whether the game during which a particular shot took place was won by the Lakers or not. Therefore we decided to gather that data ourselves. We scraped win-loss data for all Lakers games in the time period covered by the Kaggle dataset using the Chrome extension Web Scraper and merged that data with the original Kaggle dataset for use in our analysis.

4.5 New Variables Created

The following is a list of new columns that we added to the original dataset from Kaggle:

- `win`: Dummy variable for whether the Lakers won the game.
- `home`: Dummy variable for whether the game was at home or away.
- `three_pointer`: Dummy variable for whether the shot was a three-pointer.
- `jump_shot`: Dummy variable for whether the shot was a jump-shot. 0 indicates a default of layup.
- `dunk`: Dummy variable for whether the shot was a dunk. 0 indicates a default of layup.
- `tip_shot`: Dummy variable for whether the shot was a tip-shot. 0 indicates a default of layup.
- `hook_shot`: Dummy variable for whether the shot was a hook-shot. 0 indicates a default of layup.
- `bank_shot`: Dummy variable for whether the shot was a bank-shot. 0 indicates a default of layup.
- `game_date_formatted`: Reformatted date into R's native format for boolean comparisons during data processing. Pretty useless otherwise.
- `game_number`: Normalized game date. First game is 1, for game i, $game\_number[i] = i$.
- `avg`: Average shot percentage for each game.
- `shots_made`: Shots made for each game (may seem redundant, but useful for calculating averages over multiple games since we can't just average the averages)
- `shots_taken`: Shots taken per game (may seem redundant, but useful for calculating averages over multiple games since we can't just average the averages)
- `clutch_threshold`: Number of minutes remaining at which we begin counting shots as clutch shots.

- `clutch_perc`: Average shot percentage for clutch shots (shots attempted with below `clutch_threshold` minutes remaining) for each game.
- `clutch_shots_made`: number of clutch shots made for each game.
- `clutch_shots_taken`: number of clutch shots taken for each game.
- `ot`: dummy variable for whether or not the game went overtime.
- `ot_taken`: number of shots taken in OT.
- `ot_made`: number of shots made in OT.
- `ot_avg`: OT shooting percentage for each game.
- `season_norm`: represents the number of seasons Kobe has been in the NBA.

5 Assumptions

5.1 Incompleteness of Data

Since our original data is from a Kaggle competition, not all of Kobe's shots are included (5000 are hidden). However we assume that despite the incompleteness of this dataset, the data should still be an accurate representation of overall trends in Kobe Bryant's shooting performance. We believe that this is a reasonable assumption to make since it appears that Kaggle randomly assigns data into its training and testing datasets (`https://www.kaggle.com/c/DontGetKicked/forums/t/975/splitting-of-data-into-training-and-test`), so the data we had to work with should be a (very large) random sample of Kobe Bryant's shots.

6 Results

7 Limitations

7.1 Lack of Free-Throw Data

8 Challenges Faced

9 Conclusion

10 References

11 Acknowledgements

Special thanks to Phillip Huang for his kind assistance in helping us scrape win-loss data using the Web Scraper Chrome extension.

12 Theoretical Analysis