CIS 490: Final Project Sport Science

Matt Hixon

Sport Science

- Provide concrete insights for decisions
- Predict player progressions
- Explore competitive advantages
- Estimate wins/losses

Dataset

- NBA shots from 2014-2015 season
- ~128,000 shots from ~1800 games
- Shot Distance, Defender Distance, Dribbles, Touch Time, Game Information
- Dan B from Kaggle

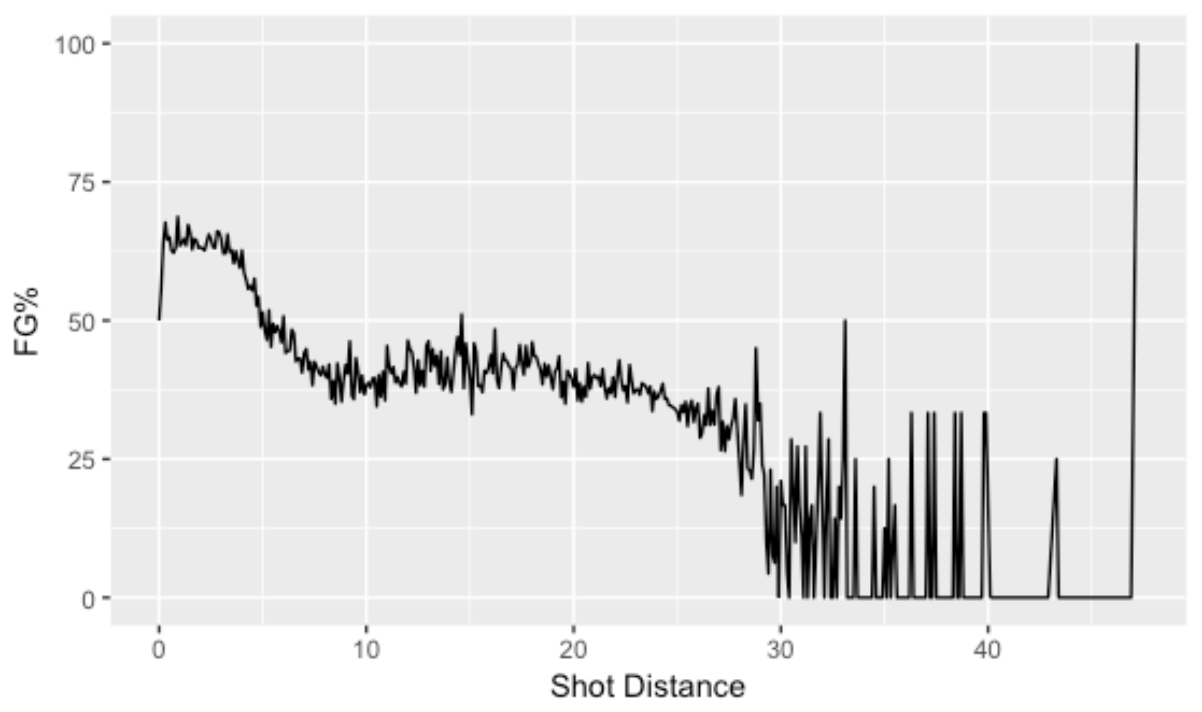
Problem Statement

- Predict a successful shot
- Visualize important factors of shots
- Model real situations with data

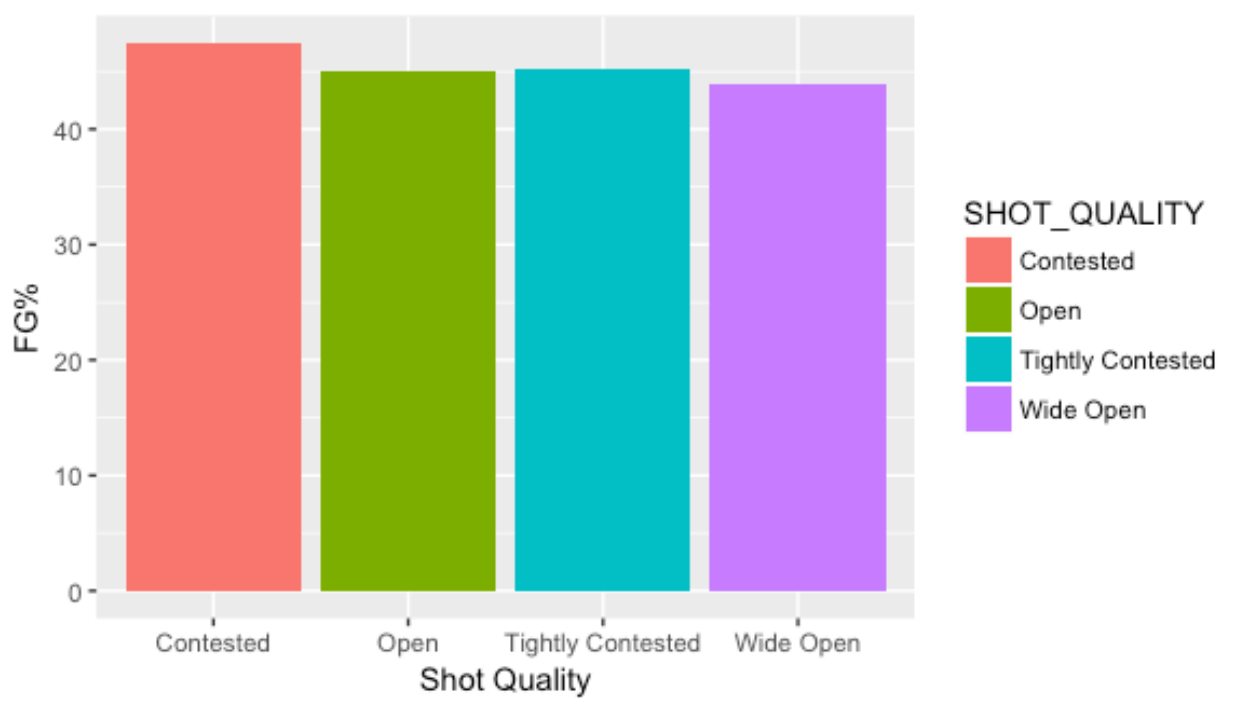
Method

- Create game situations
- Partition data into situational factors
  - Shot Type, Shot Quality, Game Clock, Shot Clock, etc.
- Use random forest model to model different combinations and find importance.

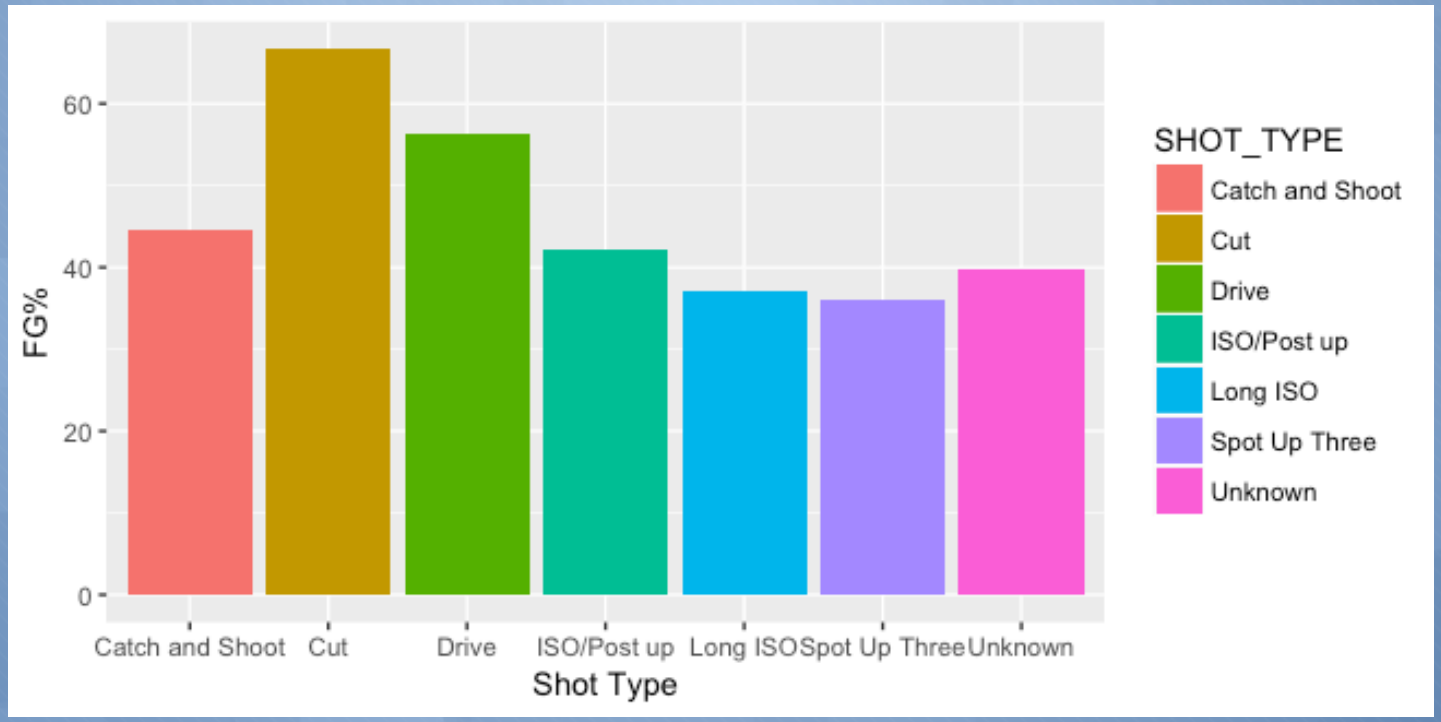
Exploration



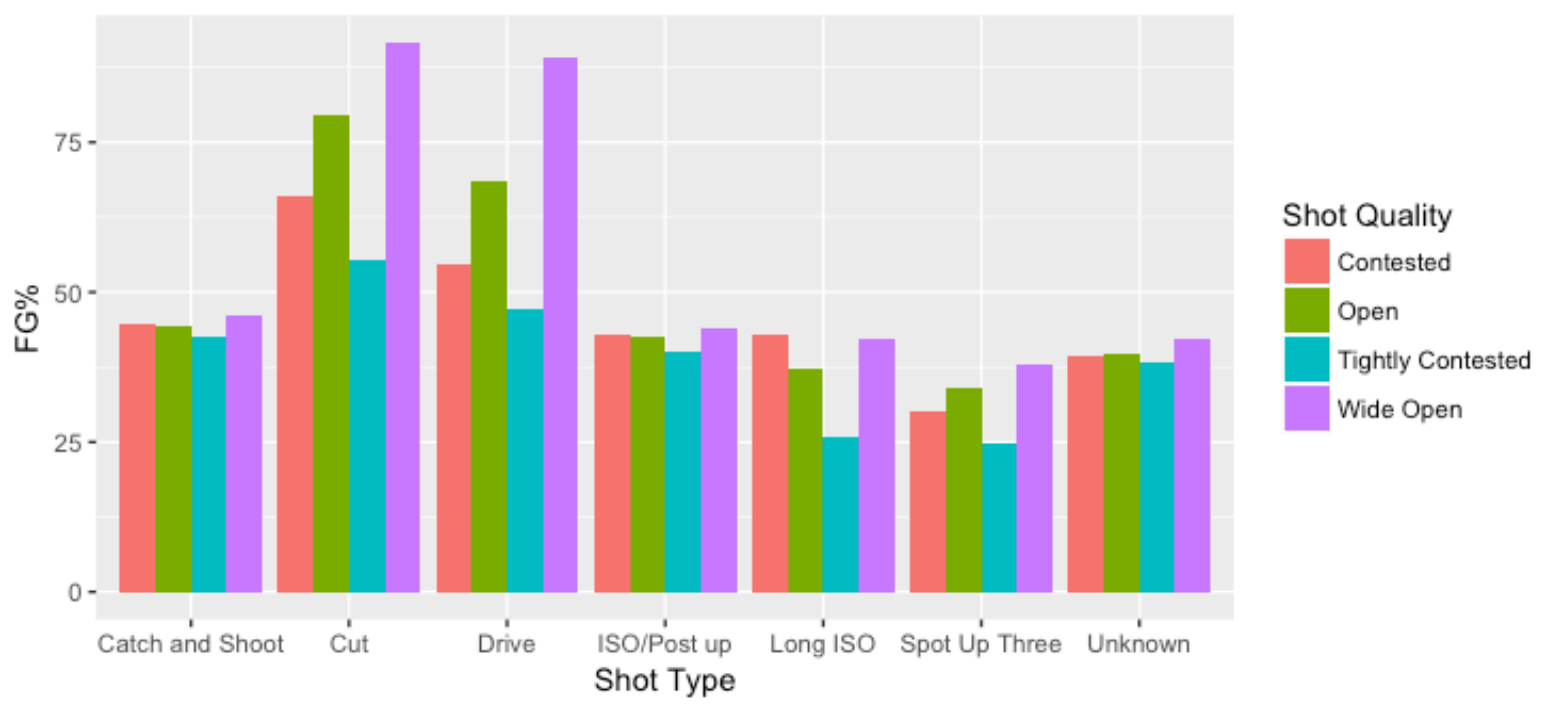
Exploration



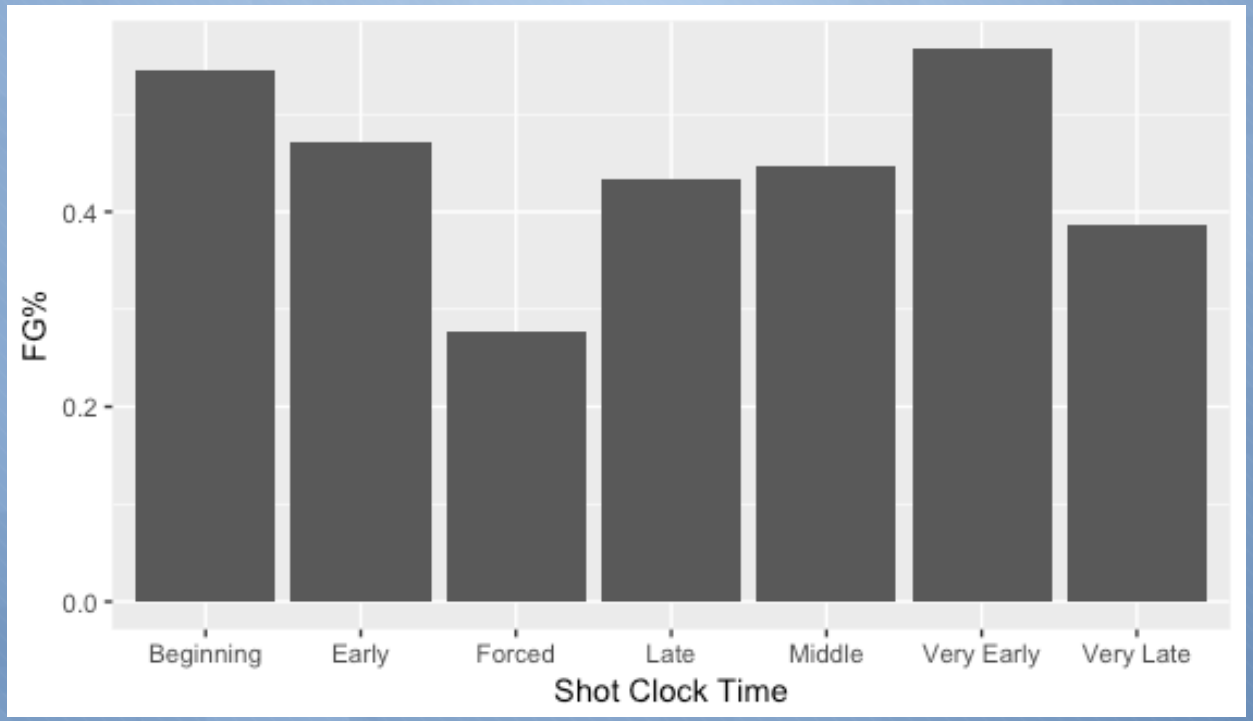
Exploration



Exploration



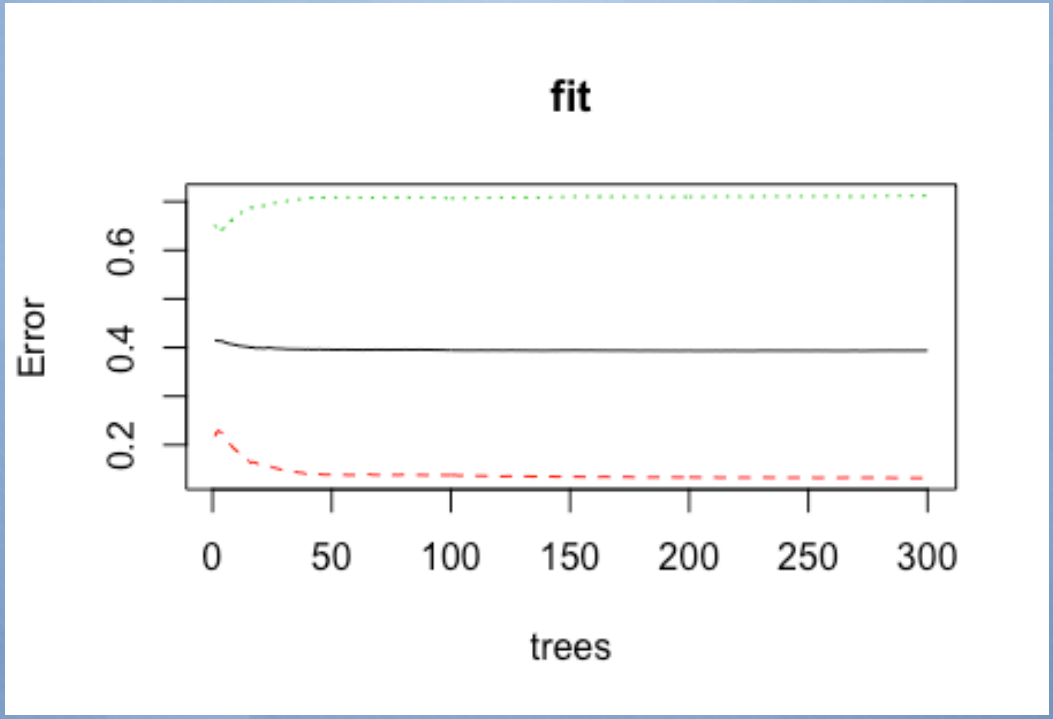
Exploration



Model

- Random forest model because of situational simulation
- Create many trees and determine importance factors

Model



Model

```
        OOB estimate of  error rate: 39.38%
Confusion matrix:
      0     1 class.error
0 45743  6954     0.131962
1 30870 12484     0.712045
                 Test set error rate: 39.17%
Confusion matrix:
      0    1 class.error
0 15197 2270   0.1299594
1 10273 4278   0.7059996
```

Conclusion

- 60.83% Accuracy Rate
- Shot situations do not universally indicate the success of a shot

Future Work

- Incorporate player data and game data
- Try different models (Naïve Bayes)
- Utilize SparkR

Most Challenging

- Determining model to use
- Preparing Data/ Feature Engineering

Things I Learned

- Manipulating data
- Feature engineering
- Basic machine learning models
- Actually predicting based off of data