NBA shot log analysis

Multivariate Analysis. UPC.

Sergio Llana and Pau Madrero.

June, 2018.

Dataset Summary

Kaggle dataset on NBA shots taken during the 2014-2015 season.

- Game actions as observations.
- Missing and wrong values in several columns.

Enriched with players' salaries and positions scrapped from ESPN.

- Joined "automatically" using `stringdist` on names.

Binary response variable: success.

Data Preprocessing

Feature Extraction

shot_difficulty based on defender distance.

"Tightly Contested", "Contested", "Open" or "Wide Open"

shot_cat based on the distance and the number of dribbles.

"Catch&Shoot", "Cut", "Drive", "ISO", "Spot up three"...

clutch based on the final result of the match and the period[1].

[1] Clutch situations are those when the player is under pressure.

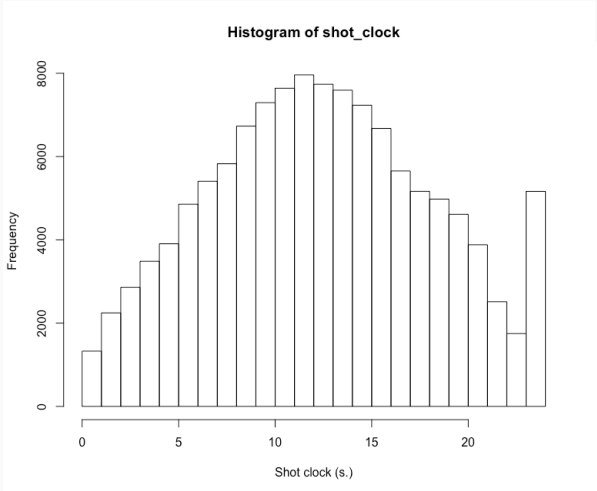
Data Cleansing

touch_time should have values in [0, 24].

- Negative values marked as NA.

Weird spike in values close to 24 in shot_clock.

- Offensive rebounds by big men close to the hoop.

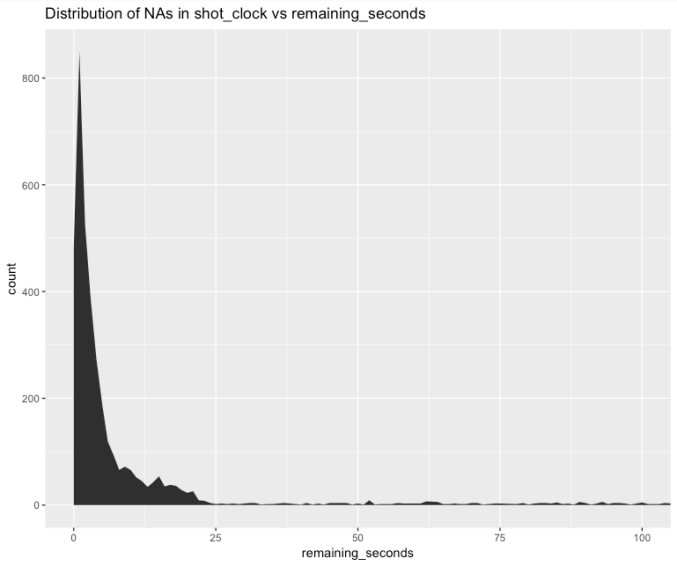


	shooter	count
	<fctr>	<int>
1	andre drummond	177
2	deandre jordan	131
3	nikola vucevic	88
4	rudy gobert	80
5	tyson chandler	79
6	jonas valanciunas	74
7	enes kanter	73
8	tristan thompson	69
9	greg monroe	66
10	joakim noah	66

... with 262 more rows

Handling Missing Values

Shot clock turned off when remaining_secs is lower than shot_clock.
Imputed randomly between [0, remaining_secs]



touch_time's NAs studied with `catdes` and imputed with 1NN.

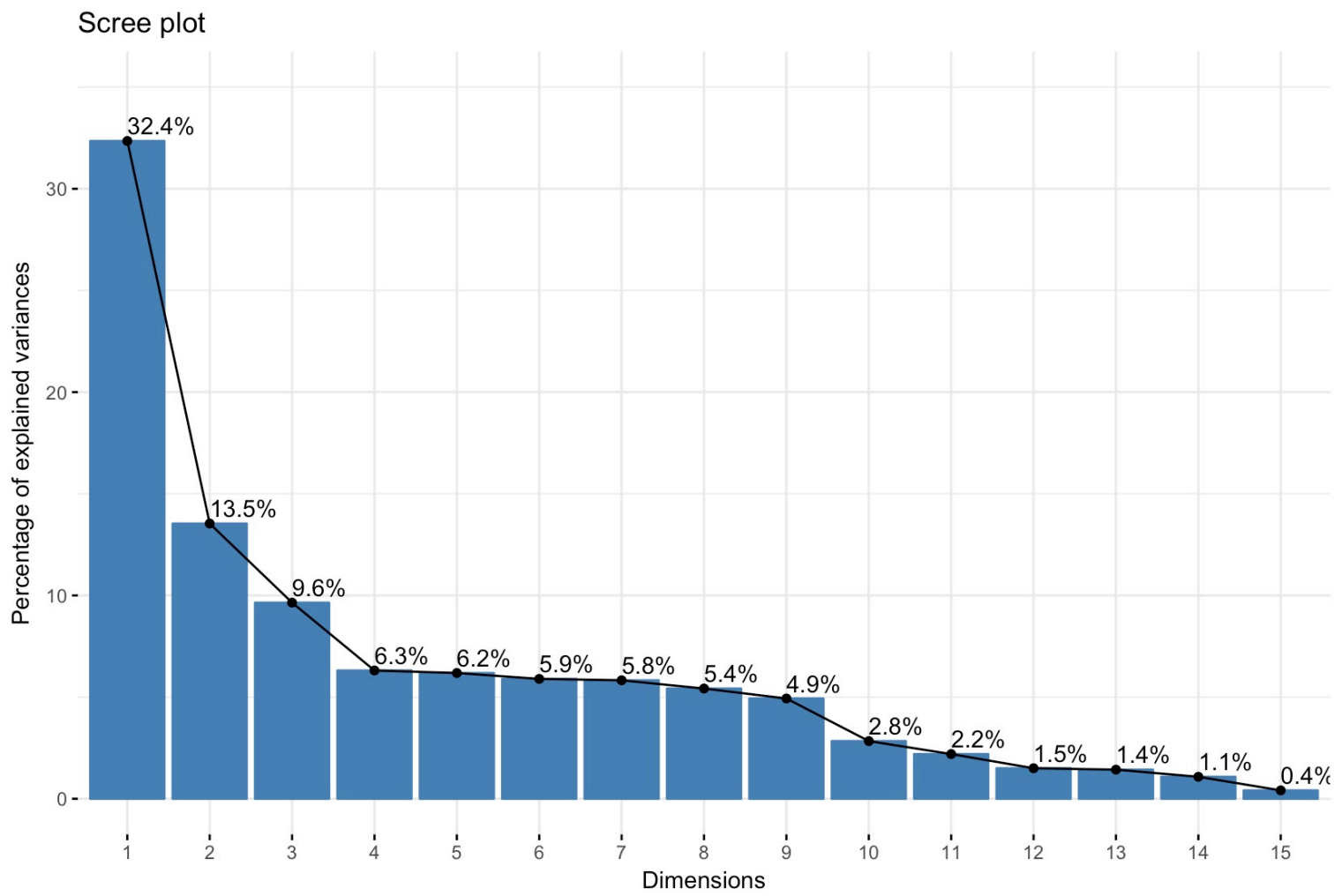
Analysis

Before PCA

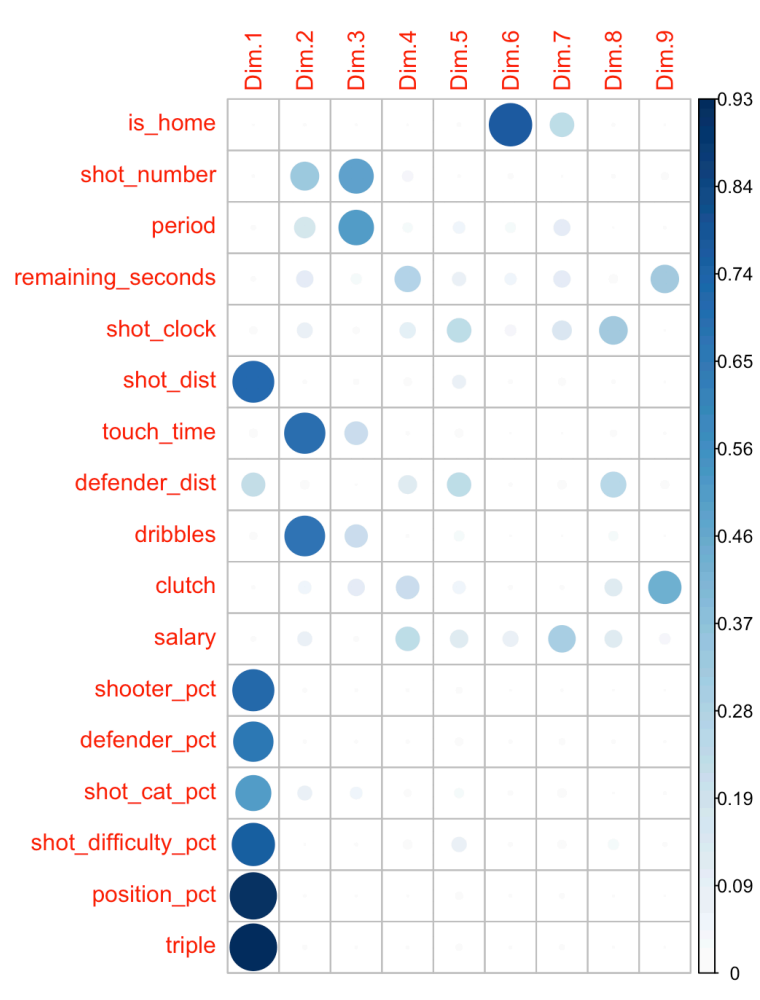
We need to transform the categorical variables into continuous.

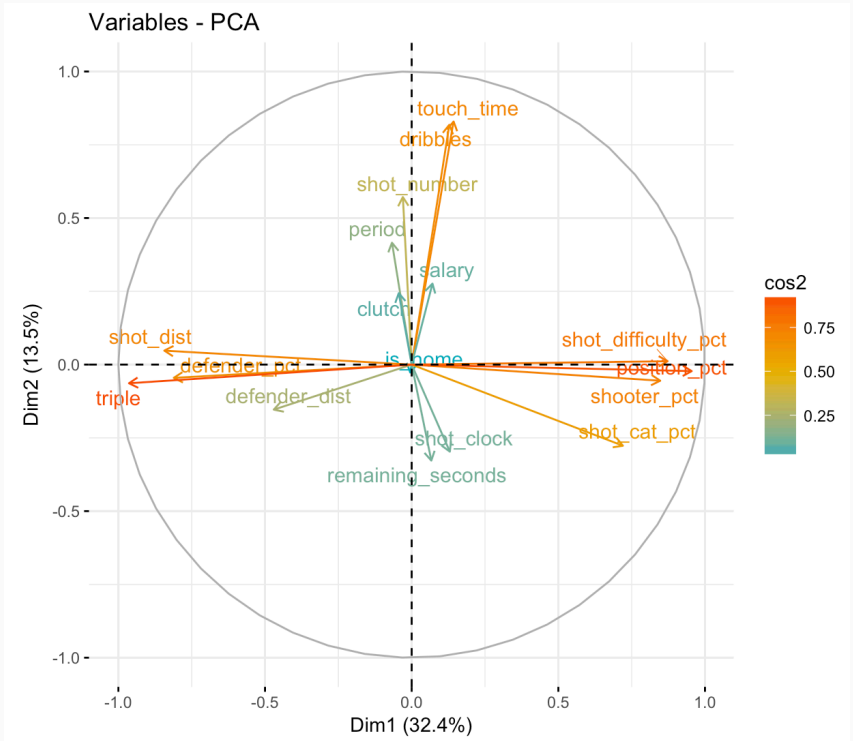
- Examples:
  - Shooter
  - Shot difficulty
  - Shooter's position
  - ...
- Based on the percentage of successful shots per modality
- Conditioned by whether the shot is a 2-pointer or a 3-pointer
- We will keep the categorical variables

PCA(I): Screeplot

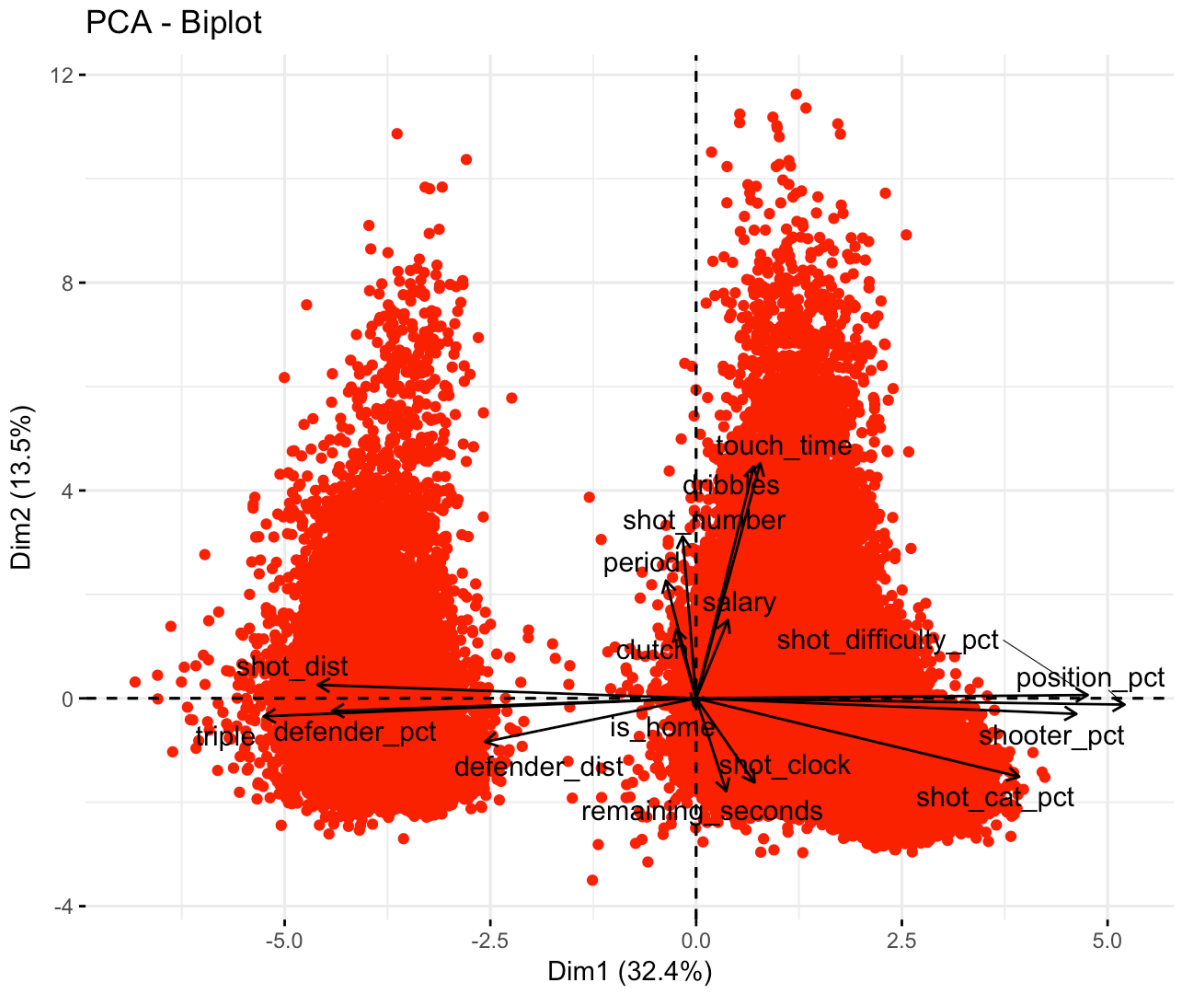


PCA (II): Variables

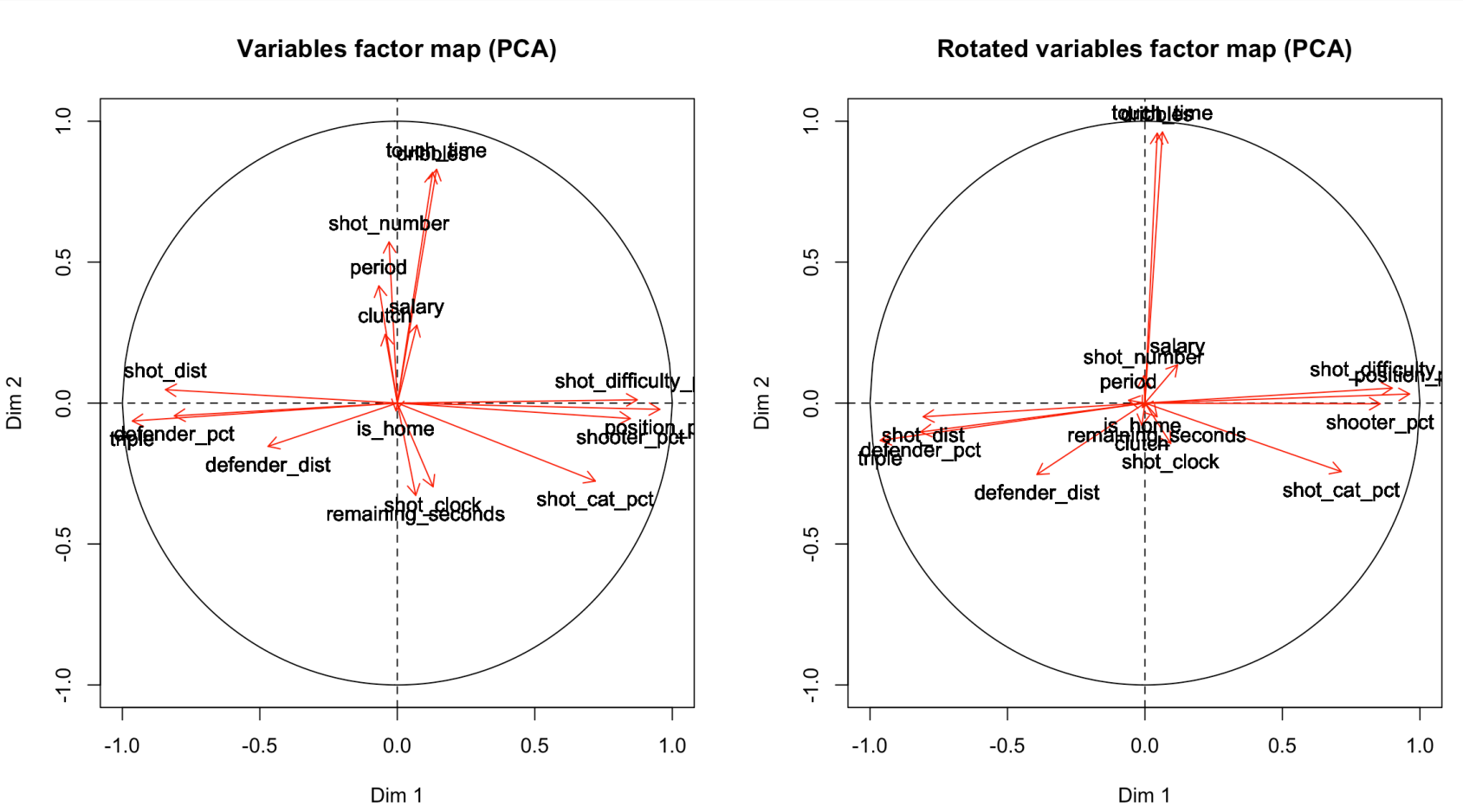




PCA (III): Biplot

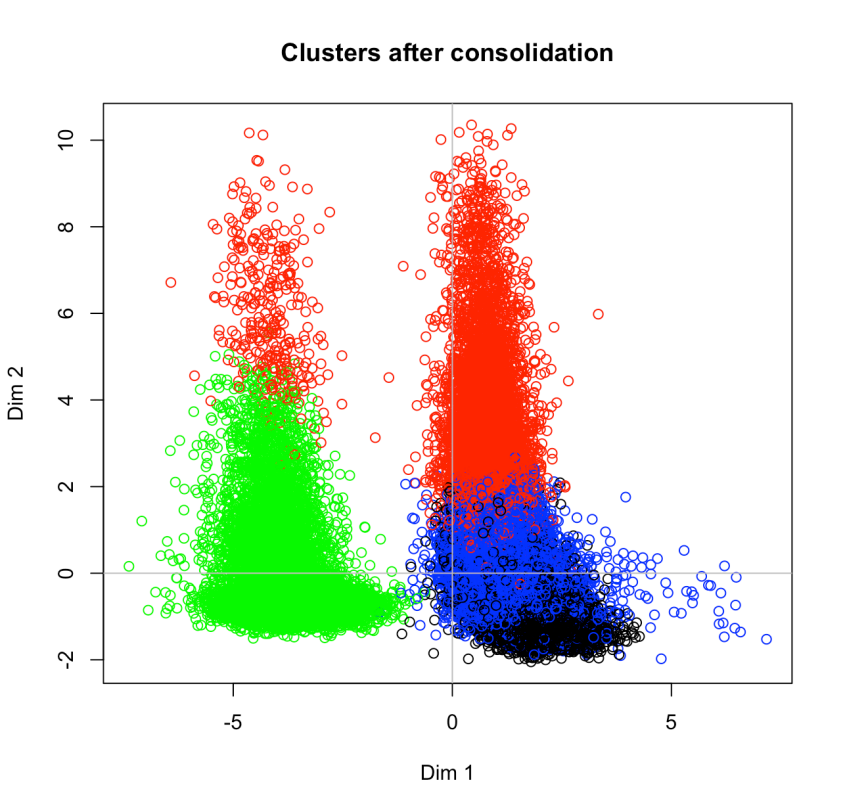


PCA (IV): Varimax



Clustering

- Clustering for large datasets.
- 4 resulting clusters:
  - Black: 2-pointer fast plays.
  - Red: Long plays with dribbles.
  - Green: 3-pointer shots.
  - Blue: 2-pointer shots.



Prediction

Validation Protocol

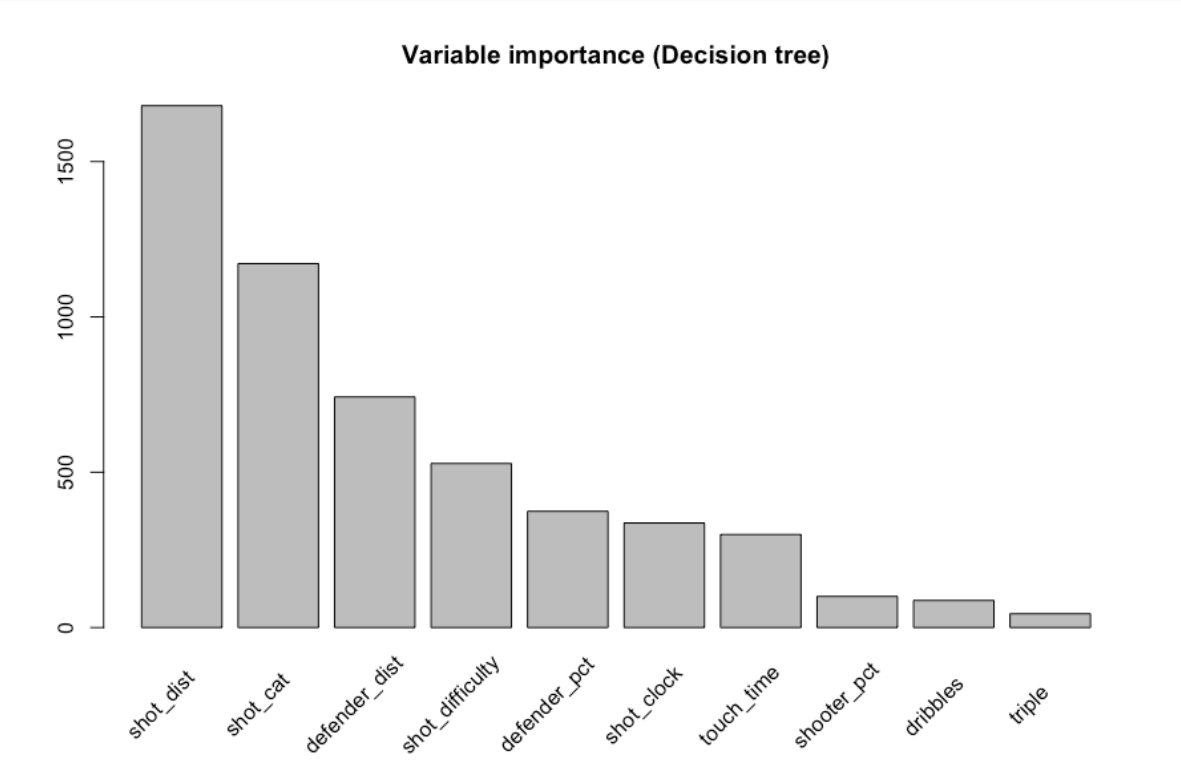
Binary classification problem: predict success of a future action.
Partition alternatives:

- Random partition: balanced response.
- Temporal partition: most recent obvs. as test set.
- Partition by players: subset of players' actions as test set.

Dataset split with 70:30 ratio of train and test data.

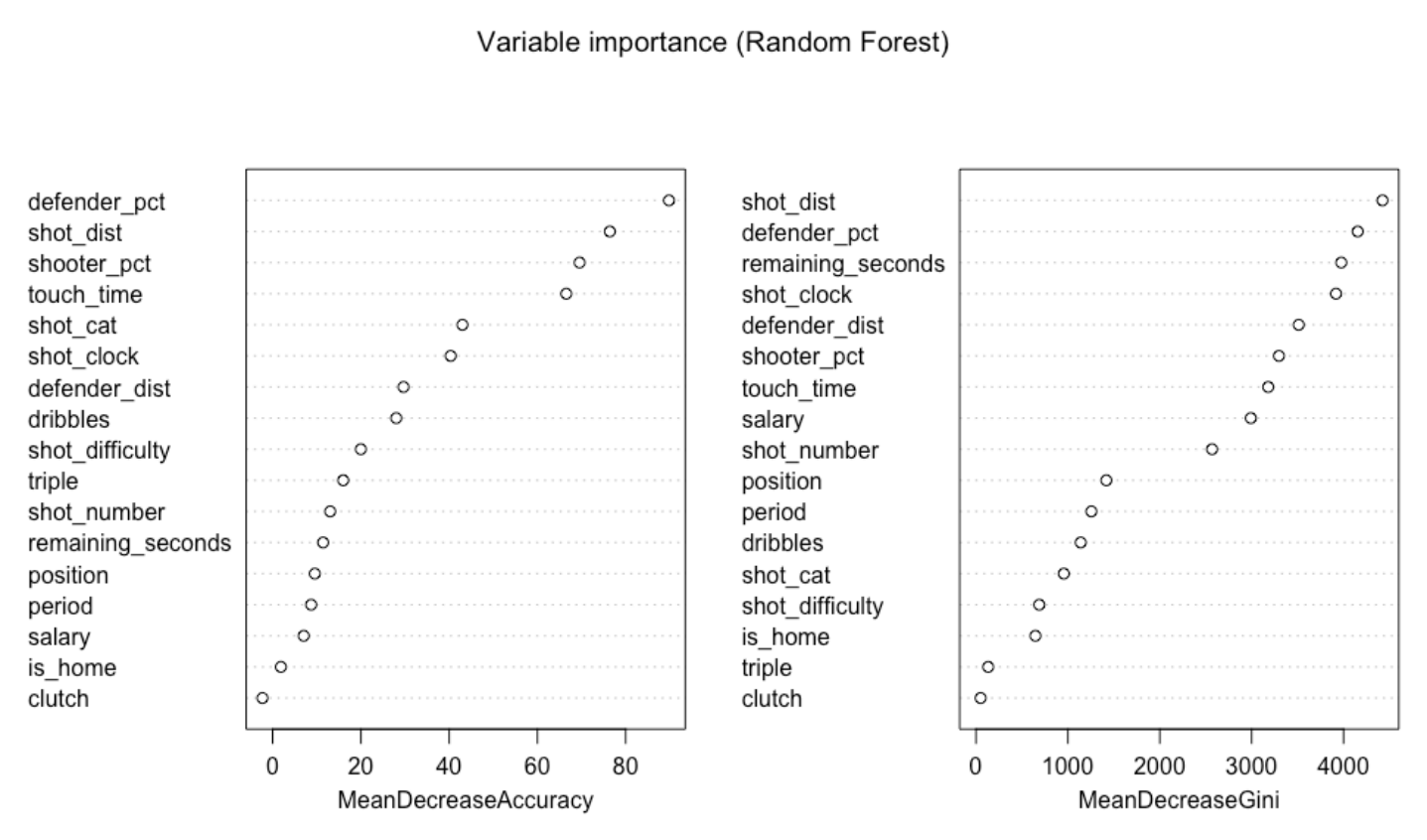
Decision Tree

10-fold Cross Validation to compare different `alphas`.
Post-pruning penalized by the size of the tree (number of leaves).

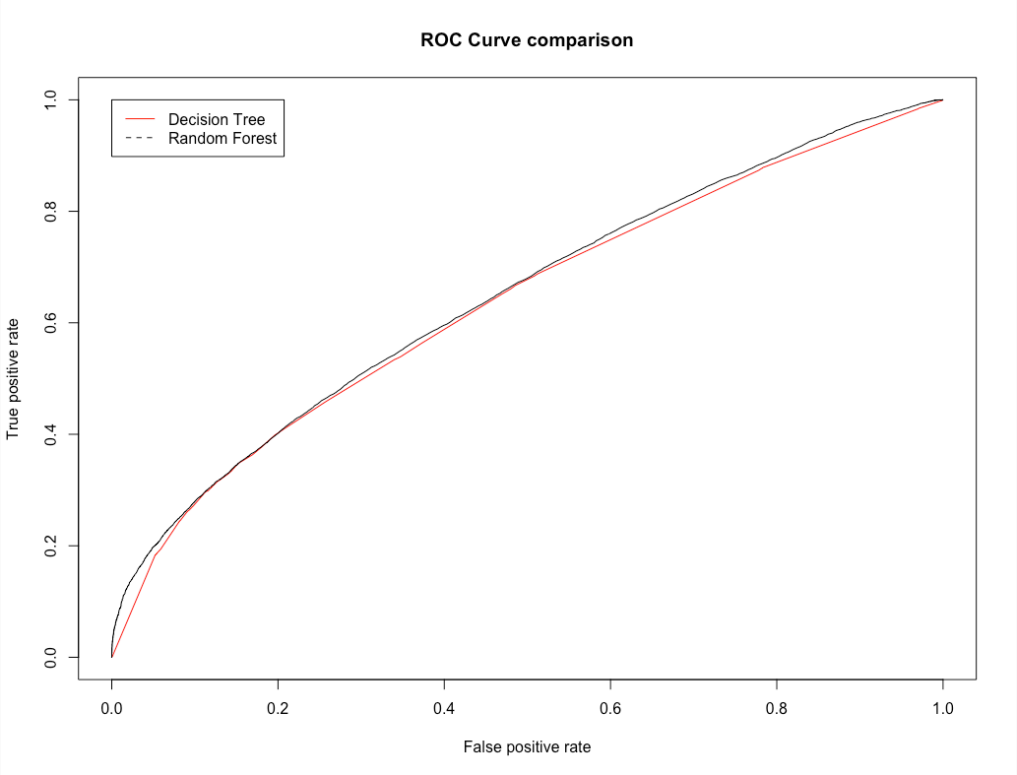


Random Forest

No need of Cross Validation thanks to the OOB error.
Parameters (`ntree` = 1000 and `mtry` = 3) optimized via grid search.



Results



	Accuracy	Precision (positive)
Decision Tree	61.91 %	37.77 %
Random Forest	61.96 %	40.07 %

Conclusions

Conclusions

- Comprehensive analysis of the dataset.
- Importance of applying domain knowledge.
- Possible extensions of the work:
  - Trying other classification models.
  - Add data from other seasons or more features.