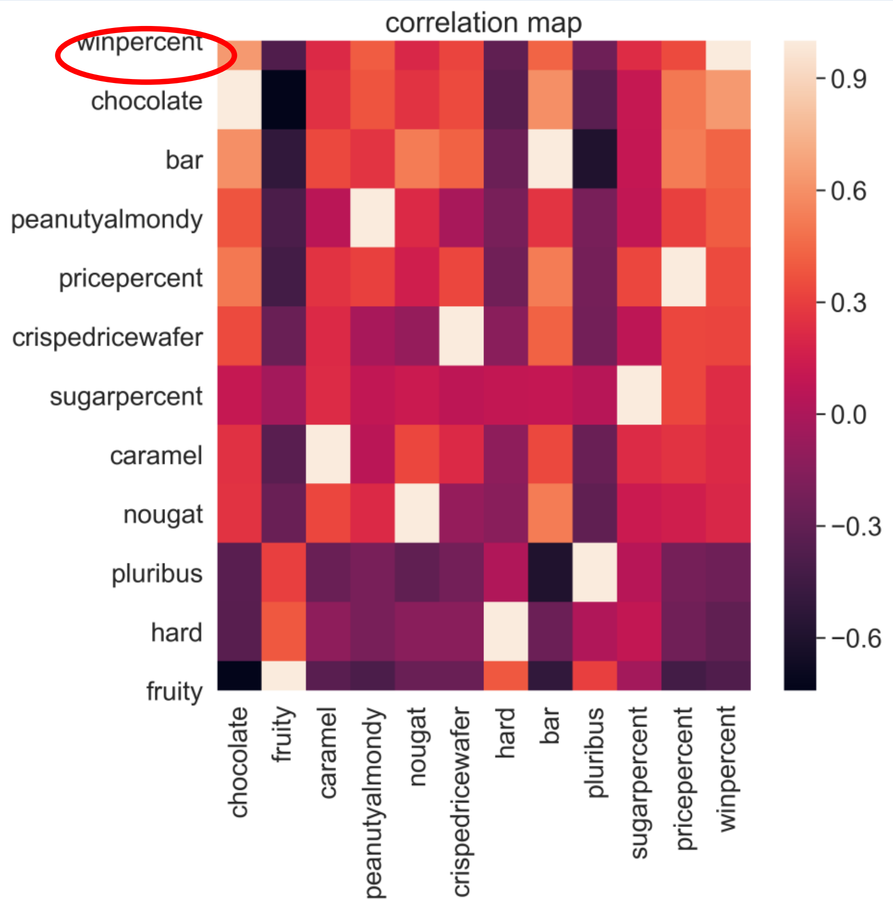
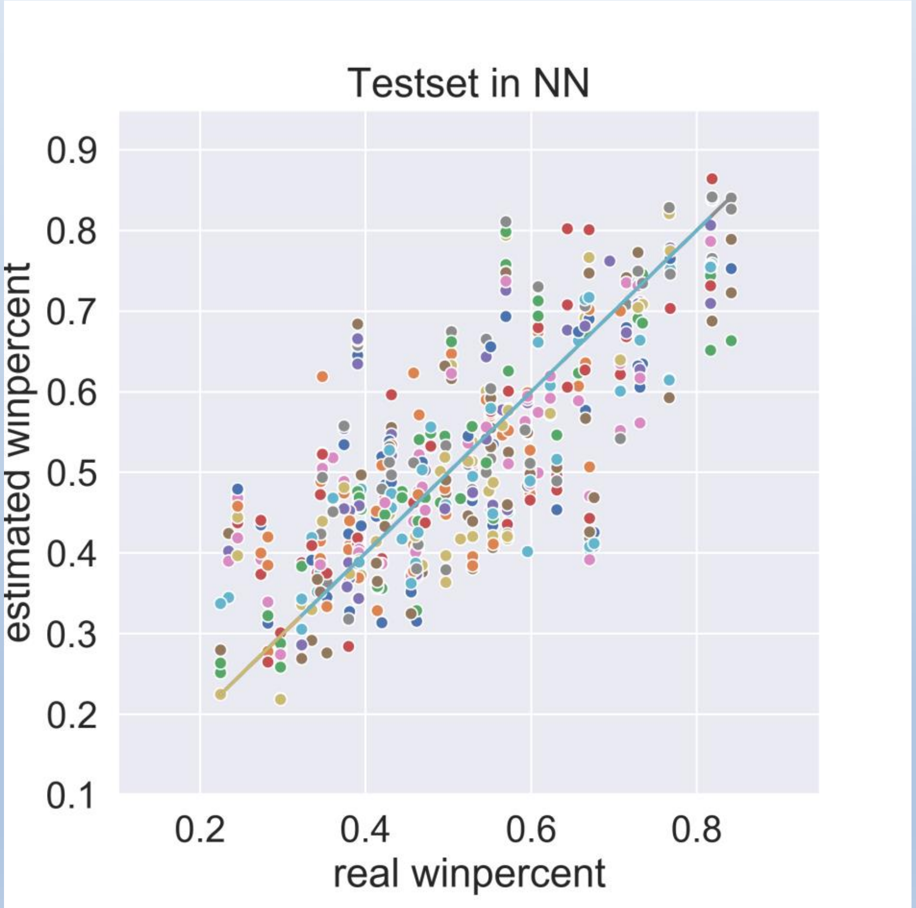
Analysis of candy-power-ranking

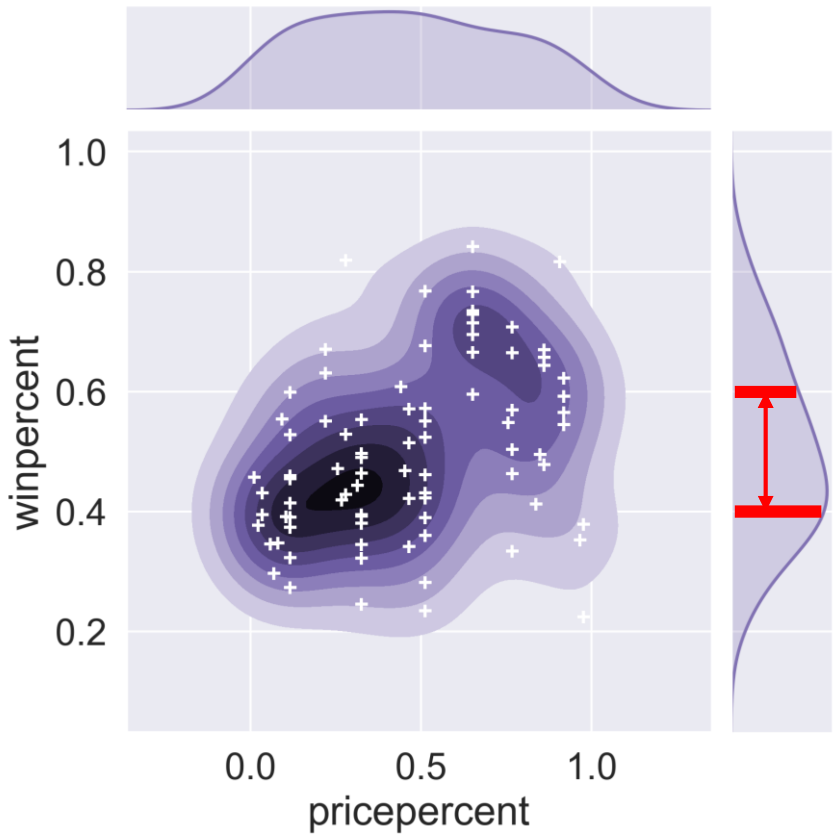
Yangbin Ma

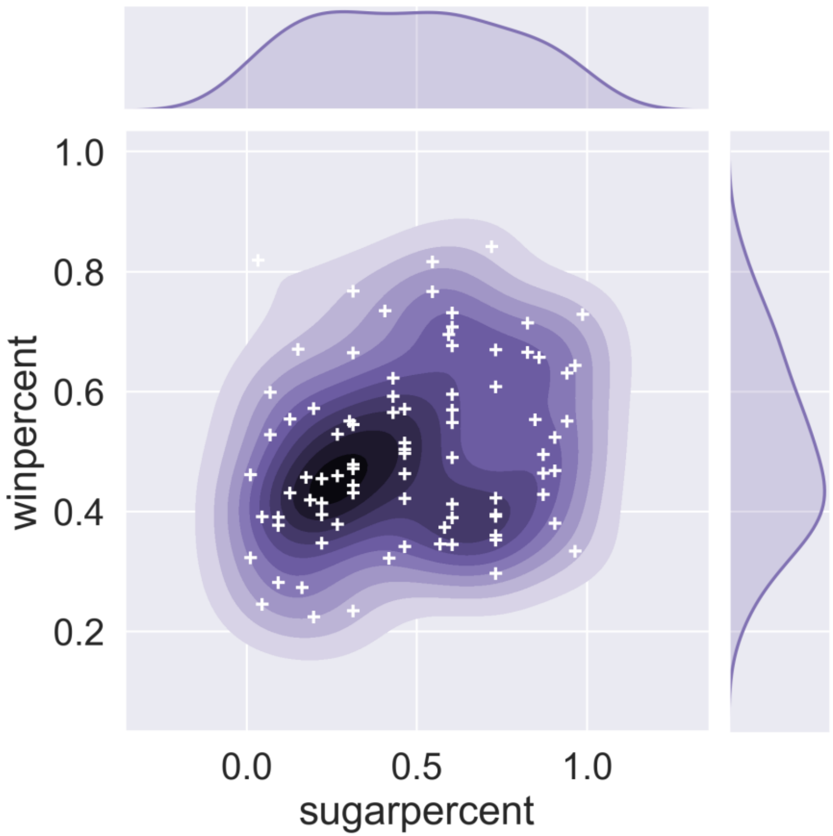
Original dataset: https://github.com/fivethirtyeight/data/blob/master/candy-power-ranking/candy-data.csv





First glance of data





Overview of winpercent
(20%~40%) occupied 25% dataset
(40%~60%) occupied 50% dataset
(60%~84%) occupied 25% dataset

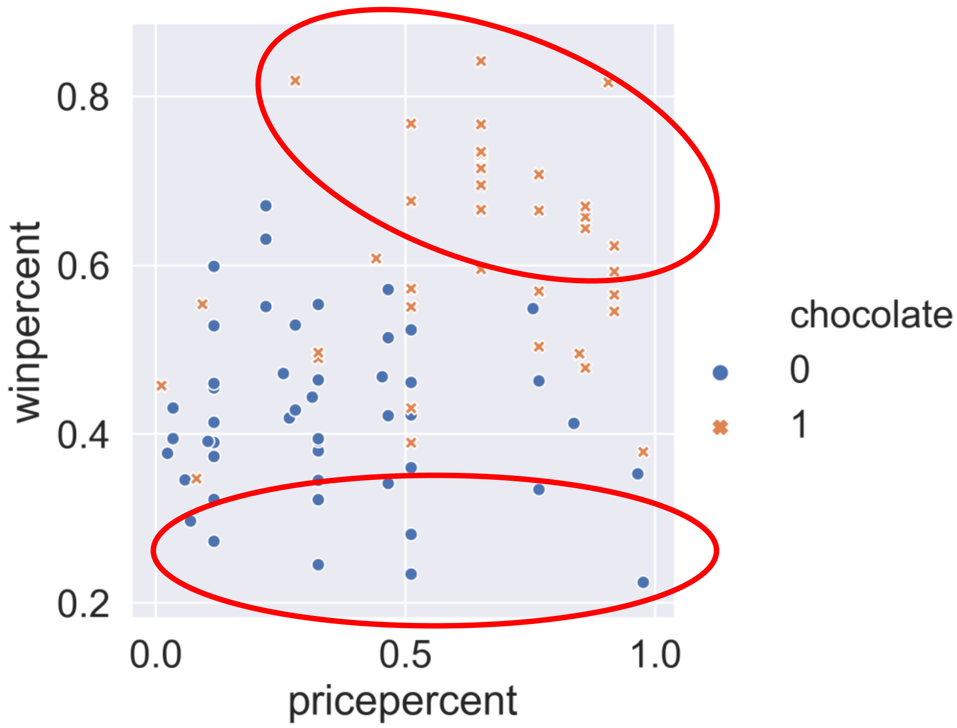
→ **Data oversampling!**

Small dataset

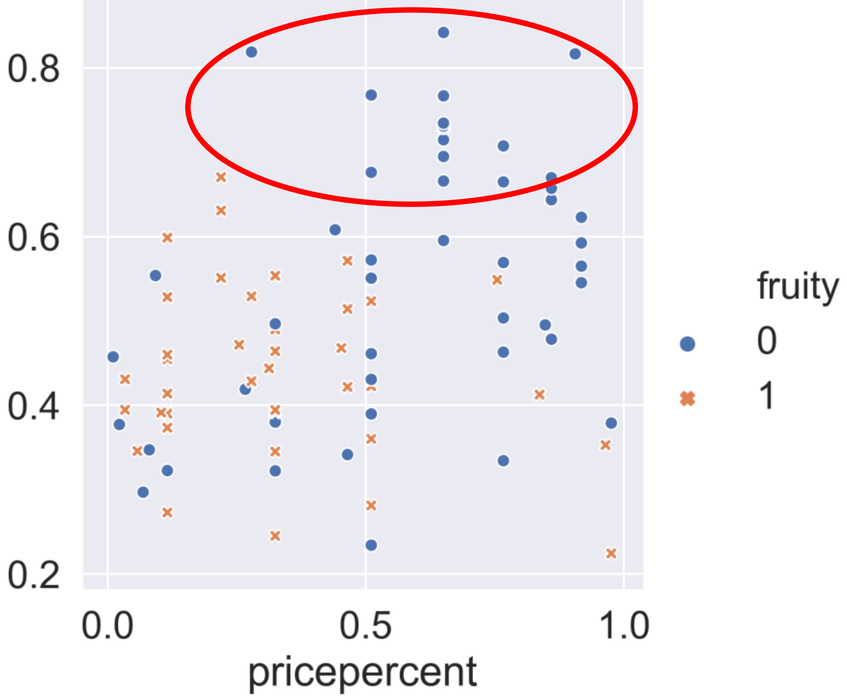
↓

Data augmentation!

First glance of data

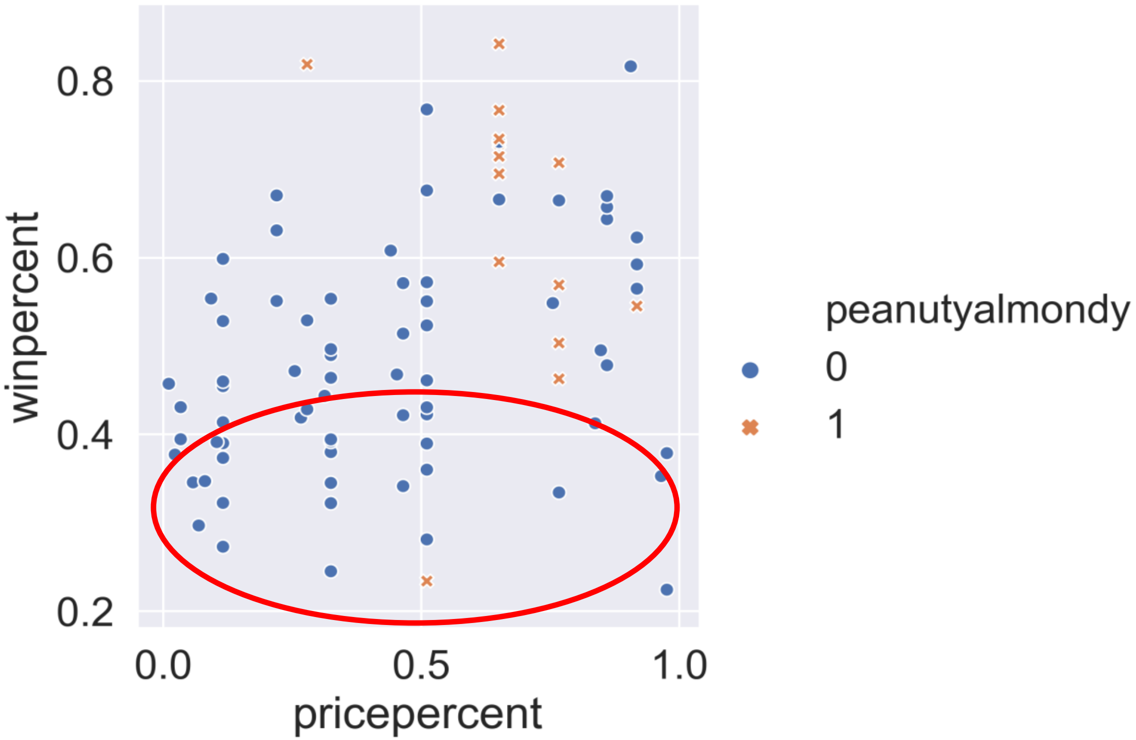


Chocolate is VIP!



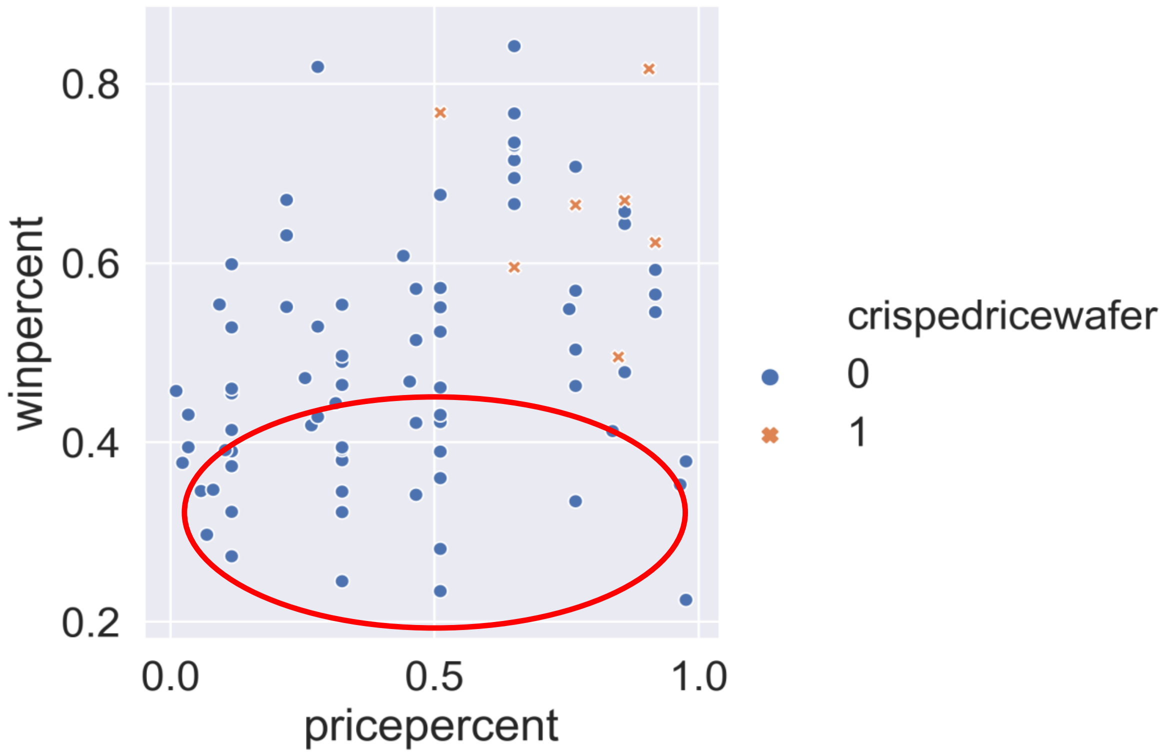
High top, no fruity

First glance of data



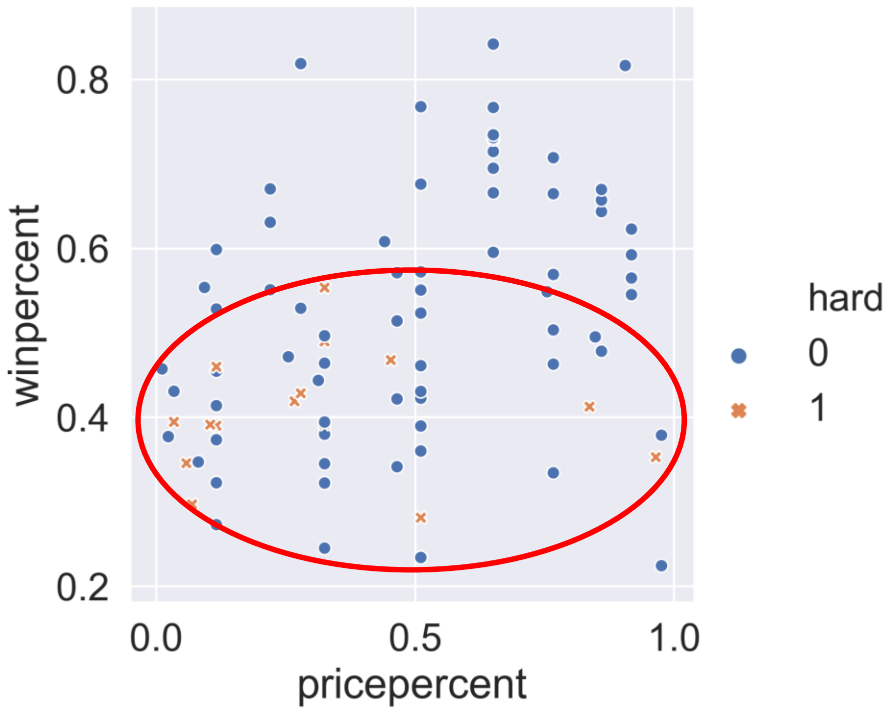
Low ranking, no peanutyalmondy

First glance of data

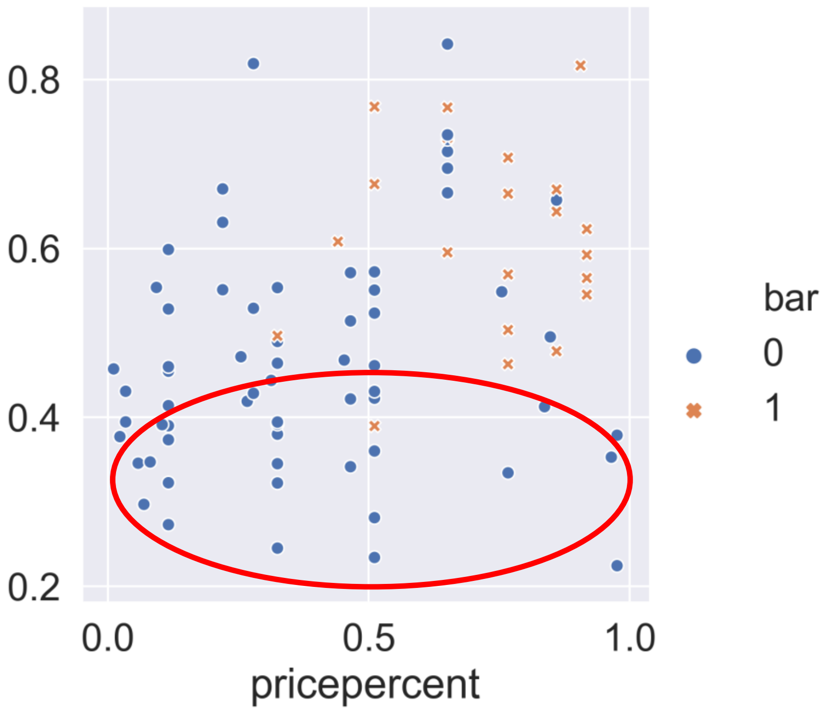


Low ranking, no crispedricewafer

First glance of data

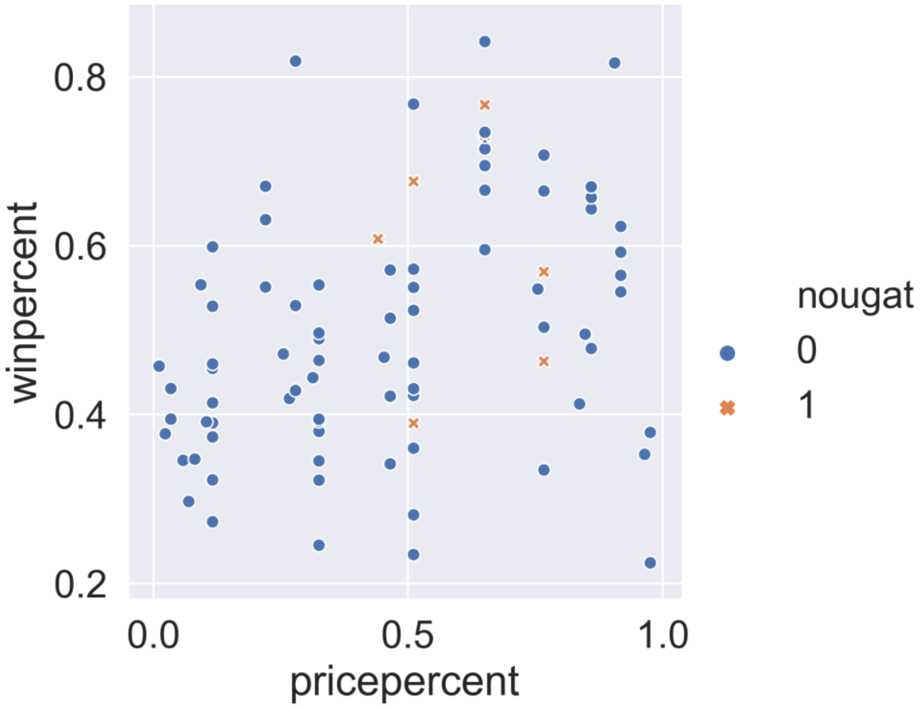


High ranking, no hard

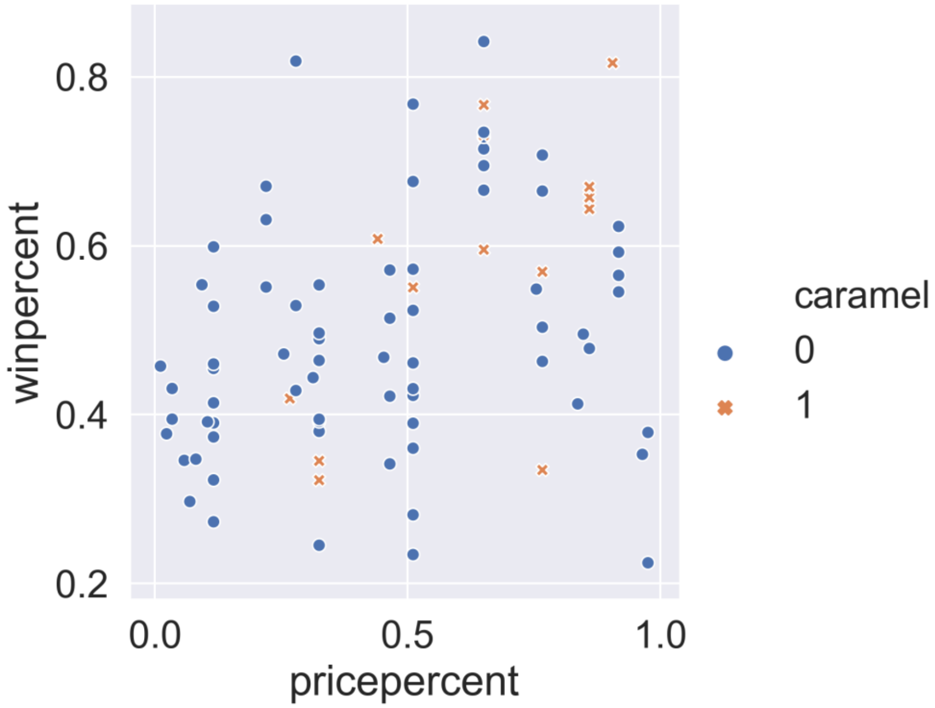


Low ranking, no bar

First glance of data

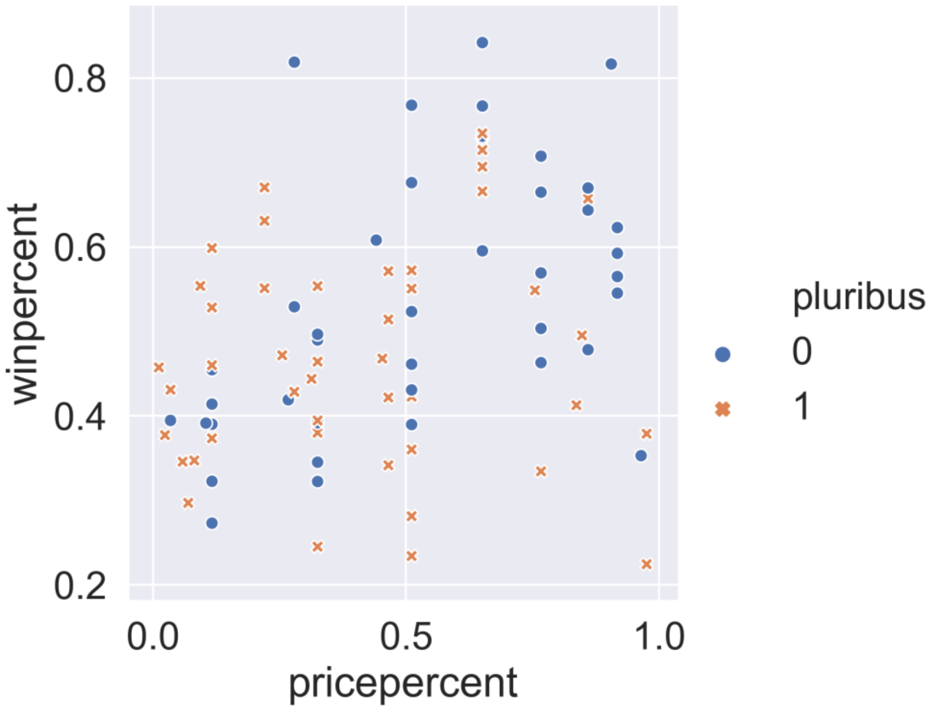


Not dominant



Not dominant

First glance of data



Not dominant

First glance of data: short summary

Overview of winpercent

- data imbalance
- 5 of 7 products with nougat rank 27 of 85 products
- 6 of 7 products with crispedricewafer rank 23 of 85 products
- top 29 products are not hard; 15 products are hard; 14 of 15 hard products has rank below 42
- 20 of 21 products with bar shape rank 47 of 85 products

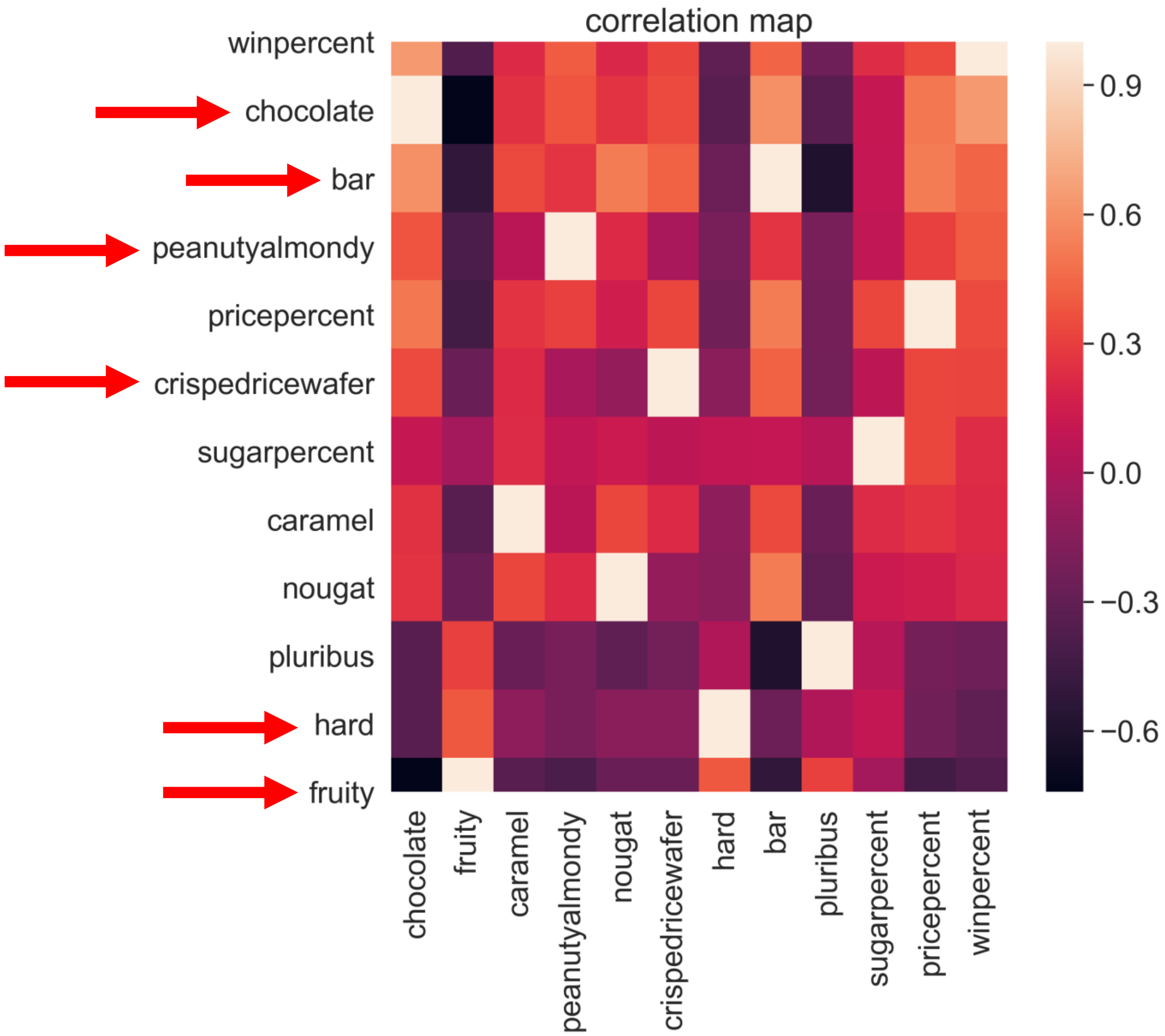
Top 12 products

- chocolate and no fruity
- either caramel or peanutyalmondy
- 8 of top 12 products have peanutyalmondy
- sugarpercent (0.57) is around 20% higher than average (0.48)
- pricepercent (0.63) is higher above the average (0.47), except reeses miniature
- Exception: reeses miniatures has low pricepercent and low sugarpercent

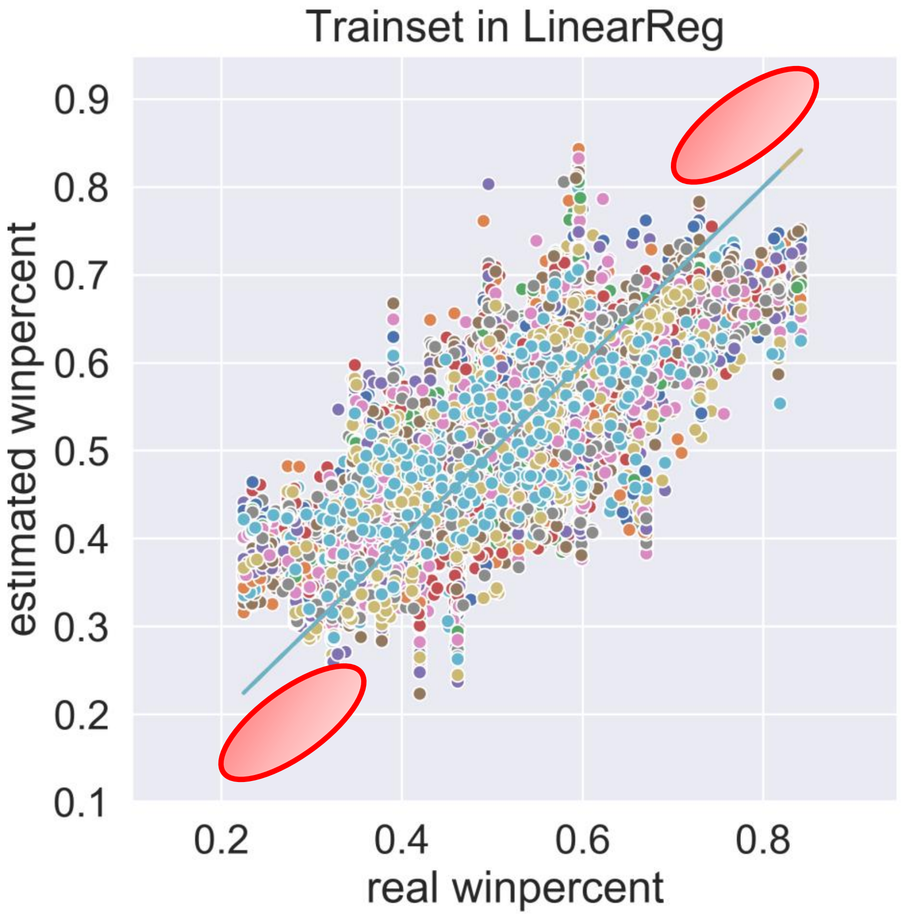
12 lowest ranking products

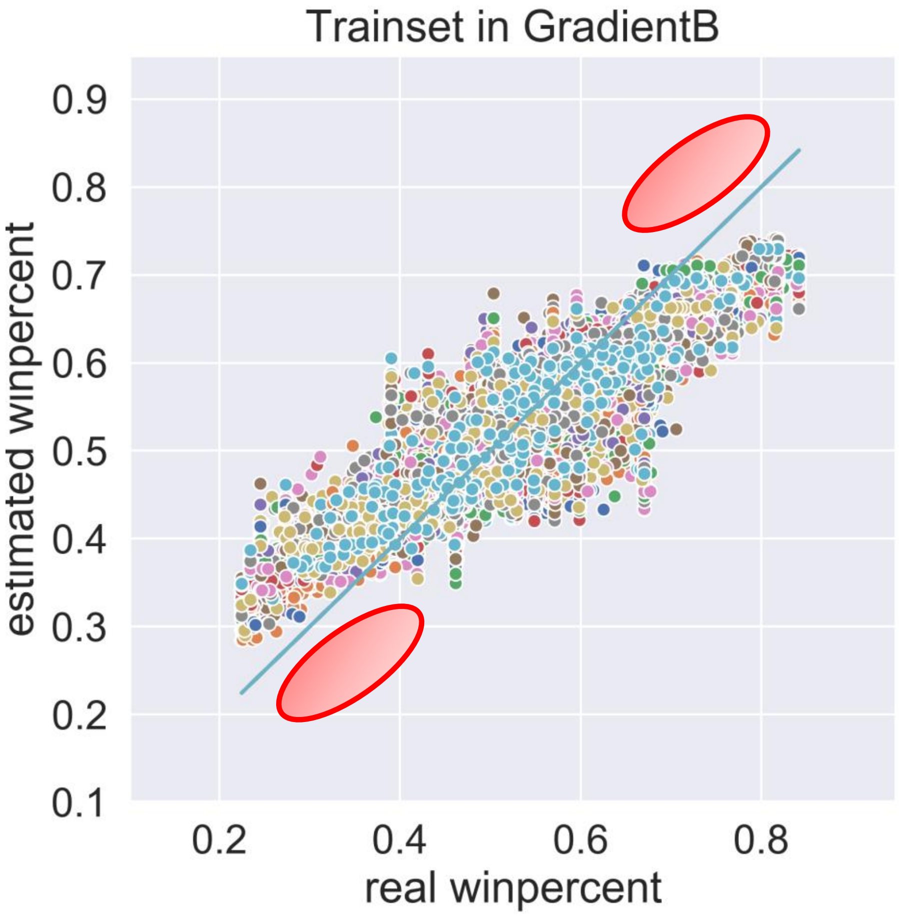
- No chocolate

Correlation map



Evaluation (train set)



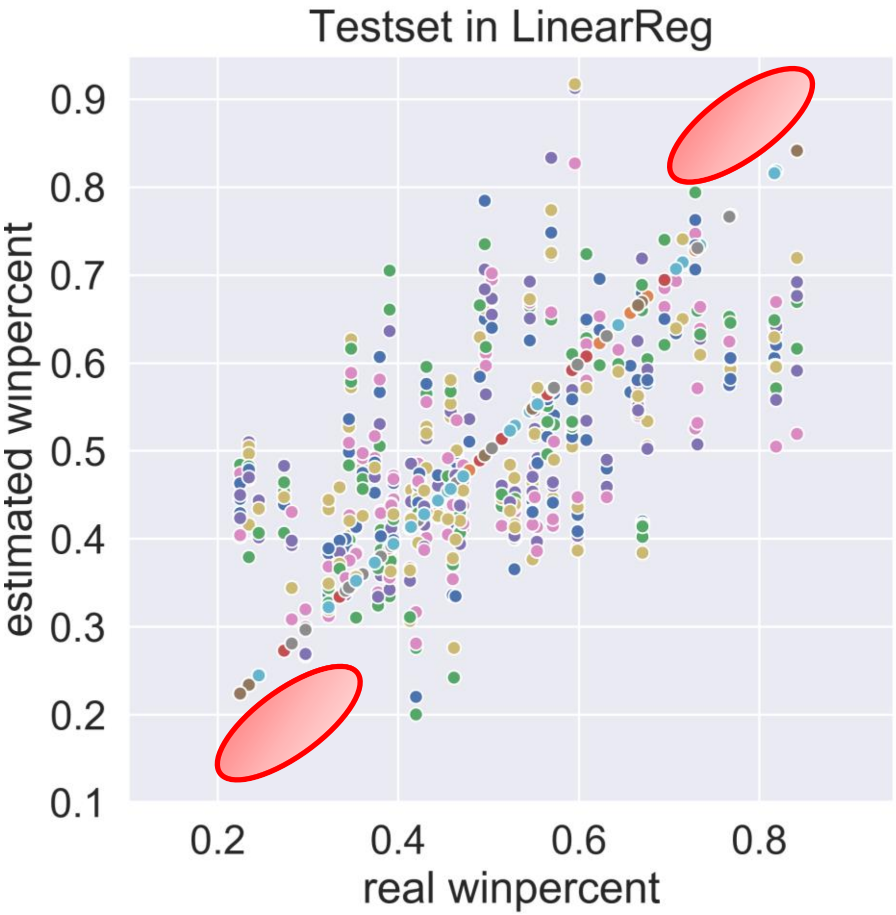


R2 score: 0.63

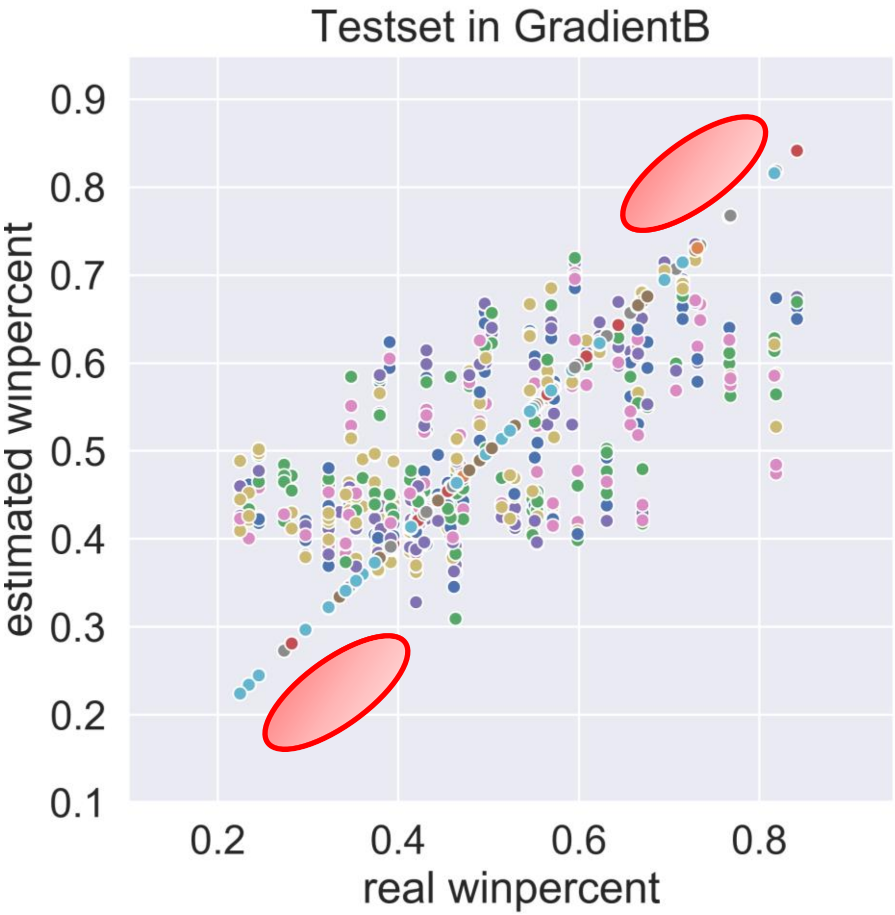
R2 score: 0.77

Boosting performs better than LinearRegessor.

Evaluation (test set)



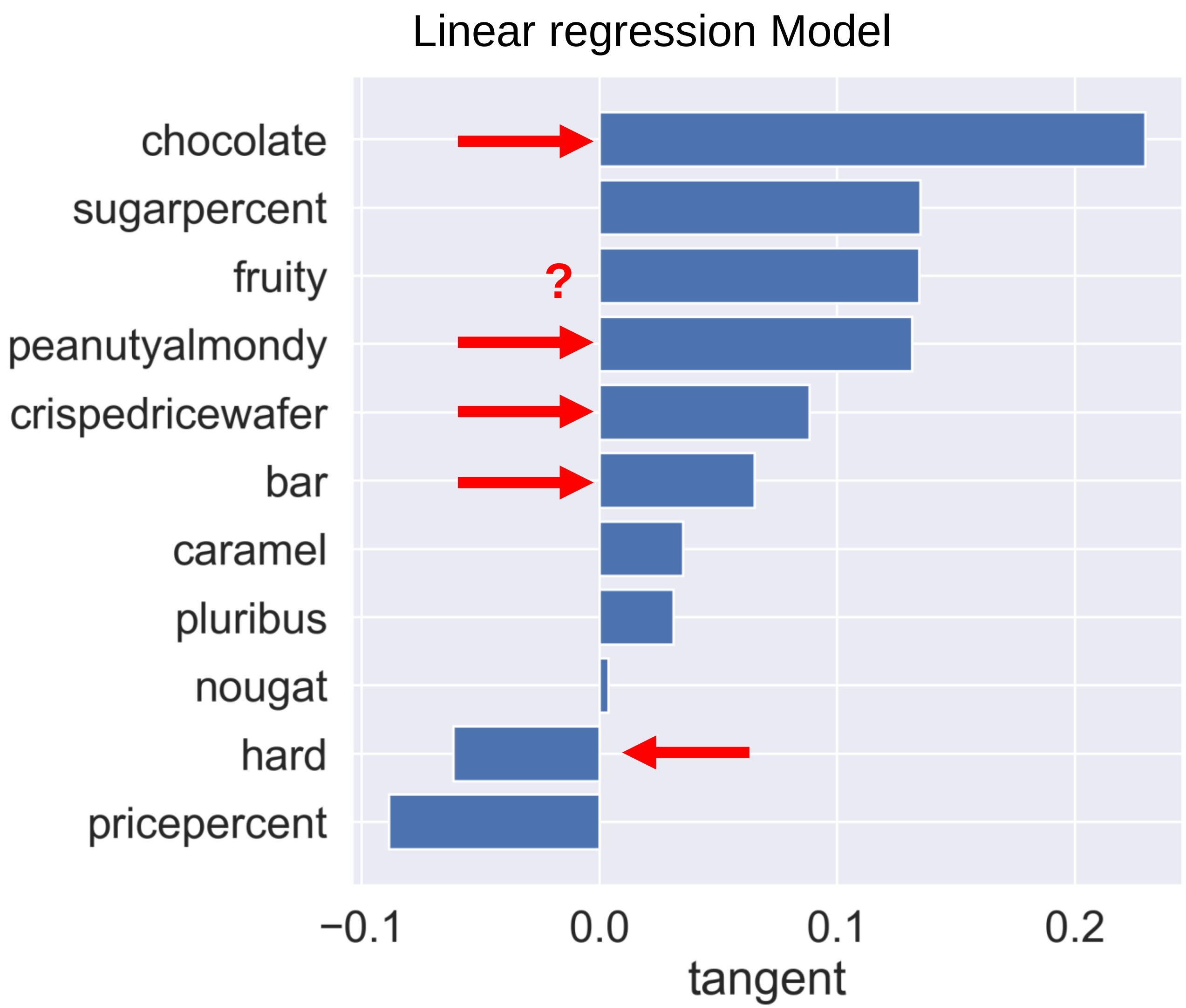
R2 score: 0.24



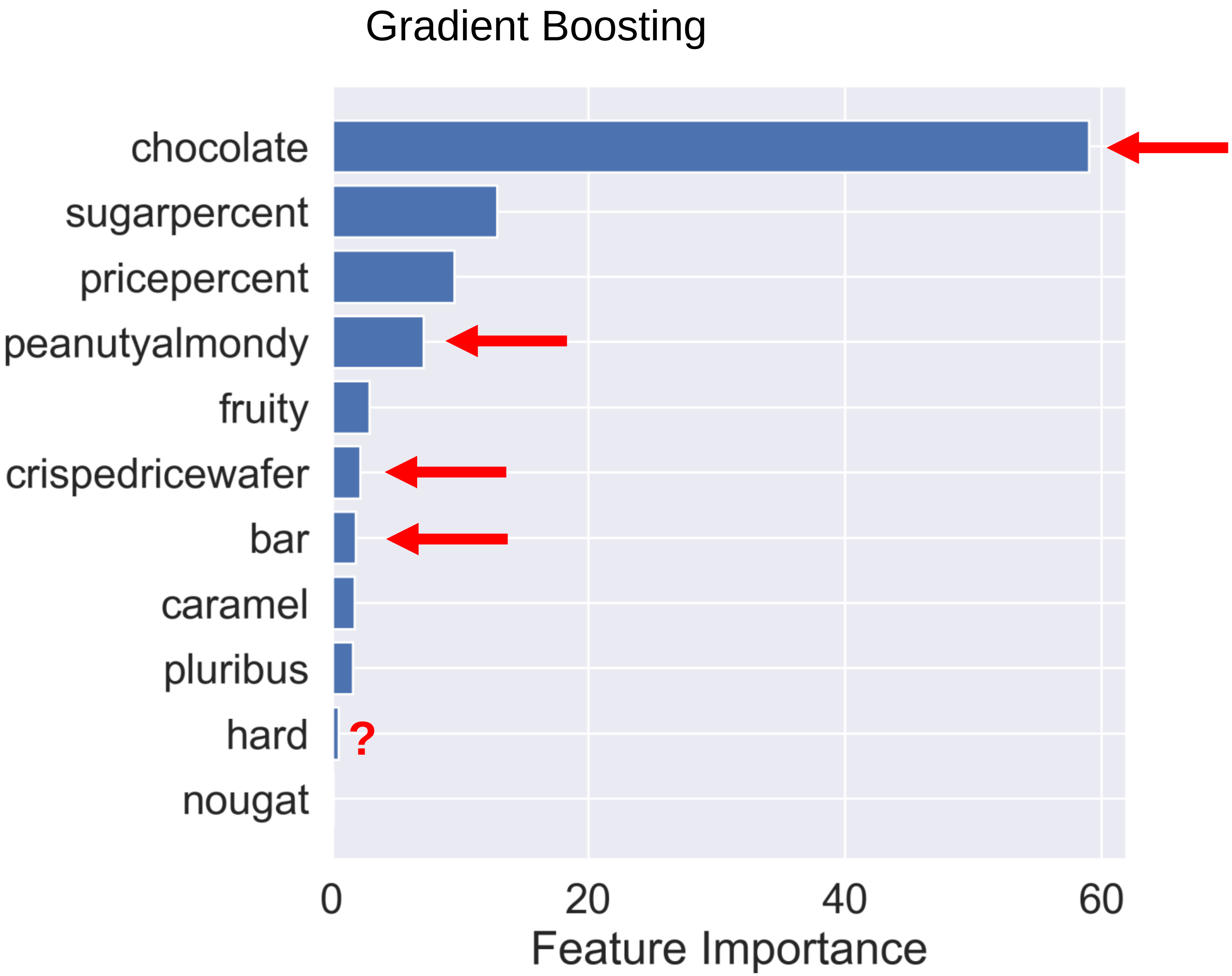
R2 score: 0.33

Boosting performs better than LinearRegessor.

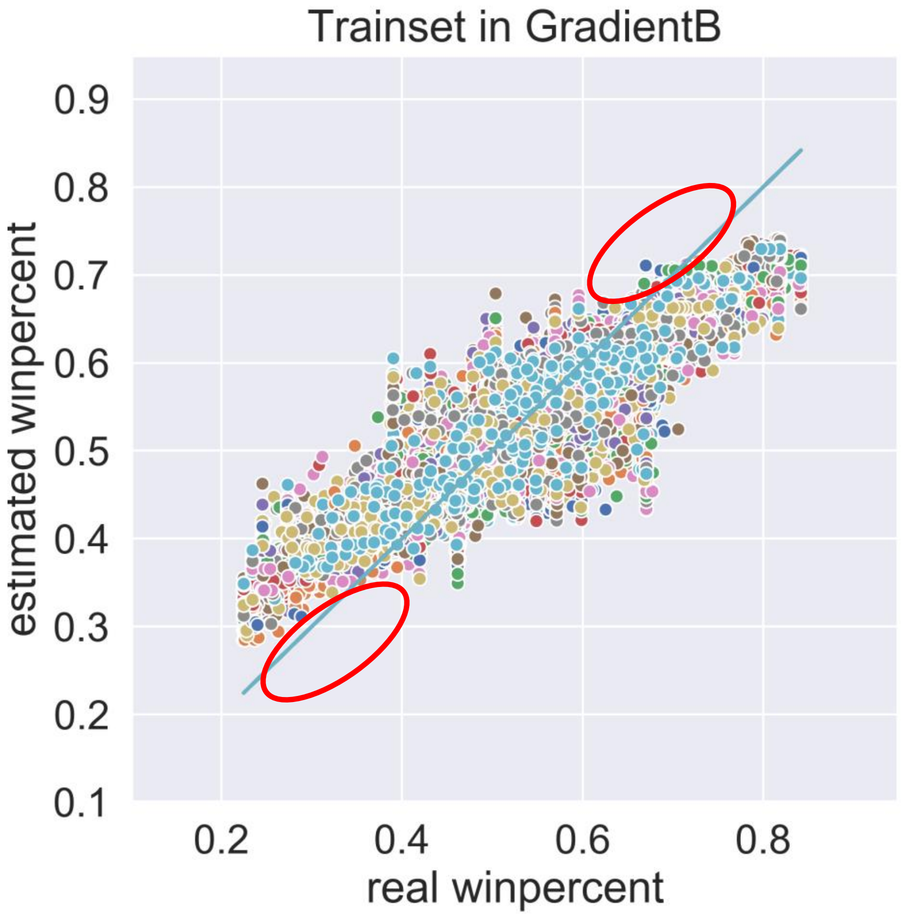
Useful information



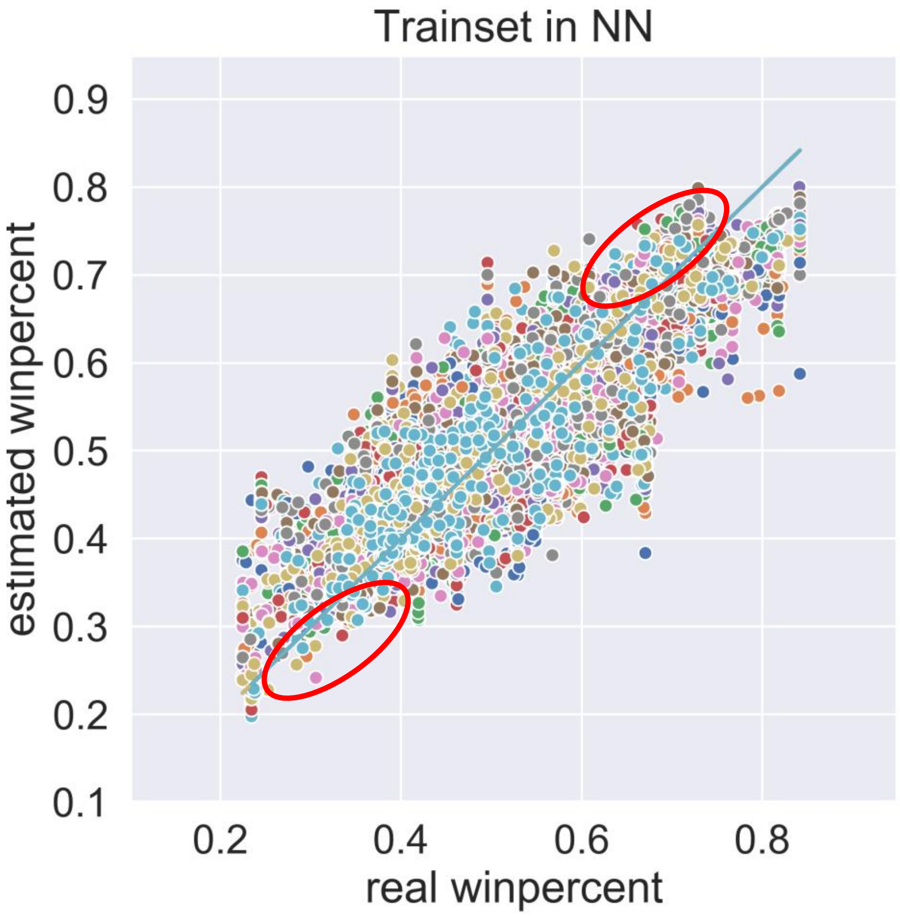
Useful information



Evaluation (train set)



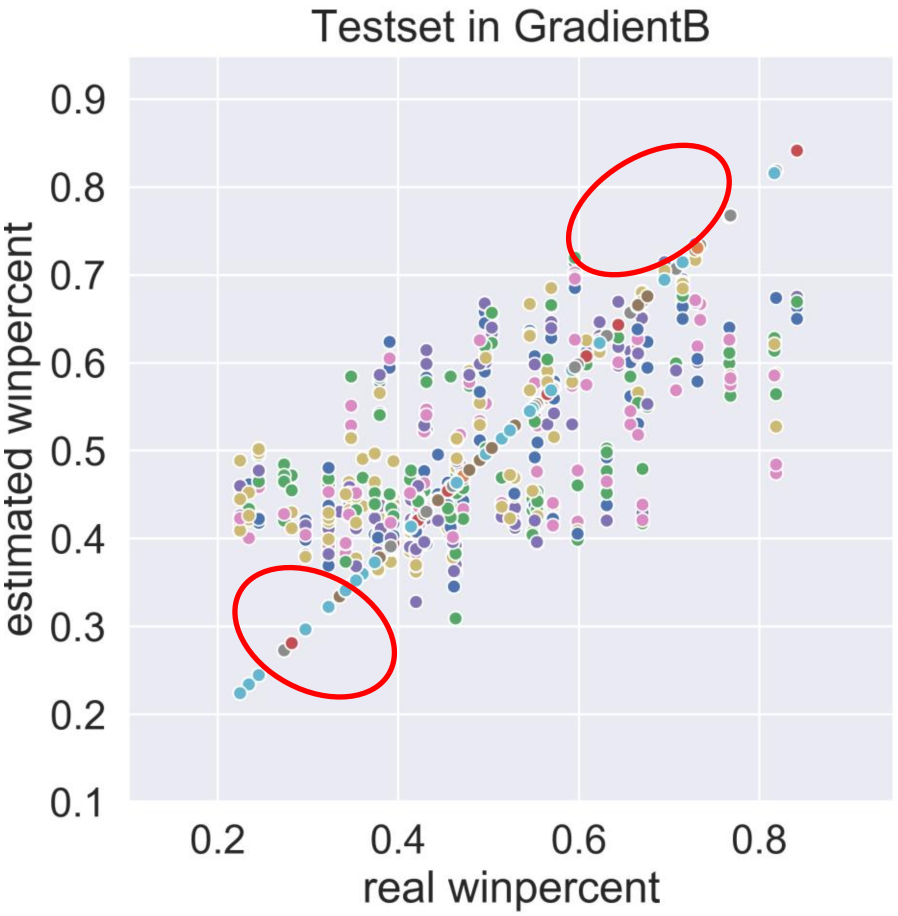
R2 score: 0.73

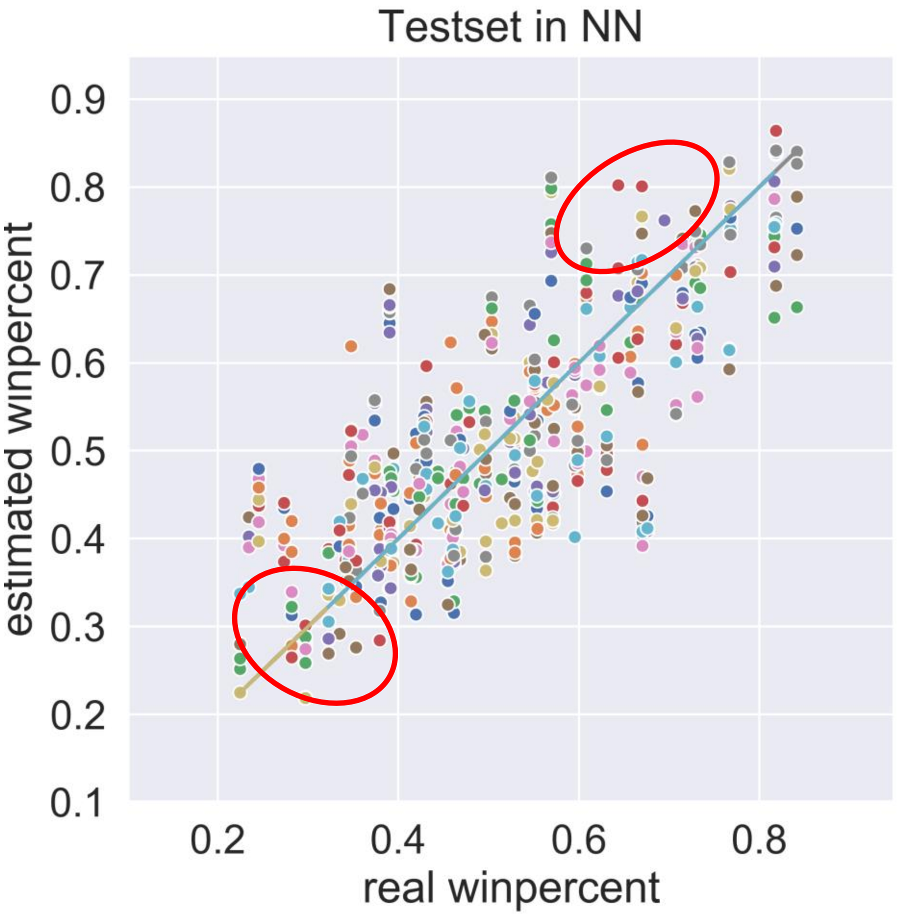


R2 score: 0.83

NN performs better than boosting

Evaluation (train set)





R2 score: 0.33

R2 score: 0.54

NN: higher prediction power

Conclusion

Preprocessing:

- Data balancing (helps a lot)
- Data augmentation (little help)

Model:

- LinearRegression, Gradientboost in sklearn (feature extraction)
- NN in pytorch (prediction)

Insights of candy data:

- Chocolate, peanut almond, crispedricewafer, bar (**+**)
- Hard (**-**)
- Price, , sugar, Pluribus, caramel (**?**)
- Many other issues, but prediction can be done in NN!

Problem:

- High and low winpercent is not well predicted (data balancing impoves performance)
- More data?