Structure DSF Project

Prediction of clicks : Poor result (too many extreme value(boxplot), data not correlated, corrolation matrix)

Solution 1 : agregate the data per user session New goal : predict the clicks per session

Models : NN,RF, logreg…

Unfortunately, this doesn’t work/ improve the results

Solution 2 : remove extreme value and and remove clicks lower than 10

Justification : -extreme value with boxplot

-Lower bound : The company is intrested to dig in the behaviour of people which are willing to search and spend time on choosing their product. Not people who are just going through the website. Therefore 10>

Models : NN, RF, log reg, linear reg…

The results is better but still too high RMSE and MAE

Conclusion :

The prediction is not feasible because of the data.

Add data like where the clicks are located, time spending on the page, time when the click occurs