# Big Data - Lessons Learned

Tuomo Kareoja

November 4, 2019

We cannot approach this task as a normal predictive modelling exercise as it will not be useful or even possible to have the training data in use in production and stop people from getting more credit when the model tells us so. This is because we are a credit rating agency and not in command of of actual crediting and data: that is a job of the credit card company.

Creating a predictive model might still be very useful though. After the initial exploration of the data it is clear that the finding the defaulters is complex job and if we just look at graphs we might not get far with understanding the complex relationships at play. A good example of this is the fact that the bigger the debt a customer has the less likely she is to default. This makes no sense of course and it means that the correlation cannot be causation. It also means that to understand the relation of debt to defaulting we must look at trough the interactions of multiple factors. A model can automatically find these kinds of complex relations, but we must make sure that we can get this info out of the model. Unfortunately this greatly limits the kinds of models we can apply in this task.

The ultimate goal of our analysis should be to understand (and not just predict) how different factors affect defaulting. With this information we could see how modifying the credit limit would affect the predictions. We are in the end only interested in seeing how this variable affects defaulting, because it is the one that we have control over. In the end we want to find out for who and how much we should lower our credit rating and if there are customers we should not have given credit at all.