PROJECT 1 - CREDIT RISK FOR MORTGAGES

KAIIE, NICOLABK
IN-STK5000

1. 2.2.1

It's impossible to ensure that the policy maximizes revenue for new data. Firstly, there will always be some noise in the training data, which will cause the model to deviate from the "perfect model" to some degree. In addition, even if one was able to produce the perfect model there would be no way to confirm this. One could always find a data set where a different model outperforms the perfect model, and without infinite data one would not be able to tell that this data set was an exception set rather than a representative one. The best one can do is to produce a robust model that is likely to perform strongly on a wide variety of different datasets.

From our cursory analysis of the data we are aware that the training set is biased in certain areas; for example, the dataset is almost entirely composed of foreign workers and lack any data on single females. Bias in the data will be a problem even if we employ standard techniques like cross-validation and bootstrapping when fitting the model as we cannot take into account data that is not present in the data set. As such, the utility of the model might suffer should the model encounter a dataset largely made up of e.g. non-foreign-workers. Of course, it is still possible that the model will perform fine in these cases given that the data we are "missing" from the data set is correlated with the observed data, but there is no guarantee that this is the case.

2. 2.2.2

The existence of this (public) database raises privacy concerns assuming individuals can be uniquely identified. Although such information (names/personal ID etc) is omitted, an attacker with enough side-information could theoretically identify individuals by cross-referencing the database with this information (record linkage). Assuming instead that the database is secret and the bank wishes to publish select information about the dataset, there are still privacy concerns that should be adressed. For example, if the bank wishes to publish credit decions, i.e. the ratio of how many people who apply for loans are given loans, an attacker can still learn that an individual was granted a loan (potentially sensitive information) if the attacker knows the target exists in the data set as well as the complete information of all the other people in the data set.

We will assume that the database is private and that we have access to the real database for training the model, so that our utility will not change. We want to publish some aggregate statistics about the database, namely the ratio of loans given compared to the total requests. In order to address privacy concerns this must satisfy ϵ-differencial privacy. We know that adding random noise to the result via the Laplace mechanism meets this requirement. The full code is in the Notebook.