1 Pre process

In order for the model to work properly we need to make sure that:

- The data has no missing values
- Has only numerical values
- Has zero mean and unit variance

Looking at the column ”has paid” we have a boolean type representation of the data. This is replaced by the value ”1” for TRUE and ”0” for FALSE.

Column ”name_in_email” was removed due to it being deemed uninteresting.

Columns named ”merchant_category” and ”merchant_group” have a class representation of data containing strings. A solution to this could be to create a type of one-hot encoding for the different classes, but in the interest of saving time I opted to remove these columns instead (this might potentially be a bad idea if a certain merchant group is more prone to defaulting).

Some columns had a quite poor value-to-NA ratio (”worst_status_active_inv” had approximately 70 % missing values for example) which might motivate us to remove these columns. However I choose to keep them and the missing values were replaced with the mean value of respective column using excel (here median could also be used).

The data was then transformed to have zero mean and unit variance.

2 Logistic regression

The model chosen was a MLP with the following characteristics:

- 3 hidden layers with 30 nodes in each layer
- Activation function: tanh
- Output activation function: sigmoid
- Optimization method: Adam
- Cost function: Binary cross entropy
- Learning rate: 0.0005 (this value was chosen based on plots such as 1)
- Epochs: 500

I used K-fold cross validation with K=10 where I measured sensitivity, specificity, the accuracy as fraction of correctly classified cases and loss as the cross entropy loss. The mean of the training accuracy and validation accuracy is shown in table 1.

Training accuracy:	0.9858
Validation accuracy:	0.9850

Table 1: Average training and validation accuracy for an MLP with 30x3 hidden nodes

Regarding regularization we can see that the training and validation error represented in figure 1 does not show any tendencies of overtraining and hence there doesn't appear to be a need for any regularization methods.

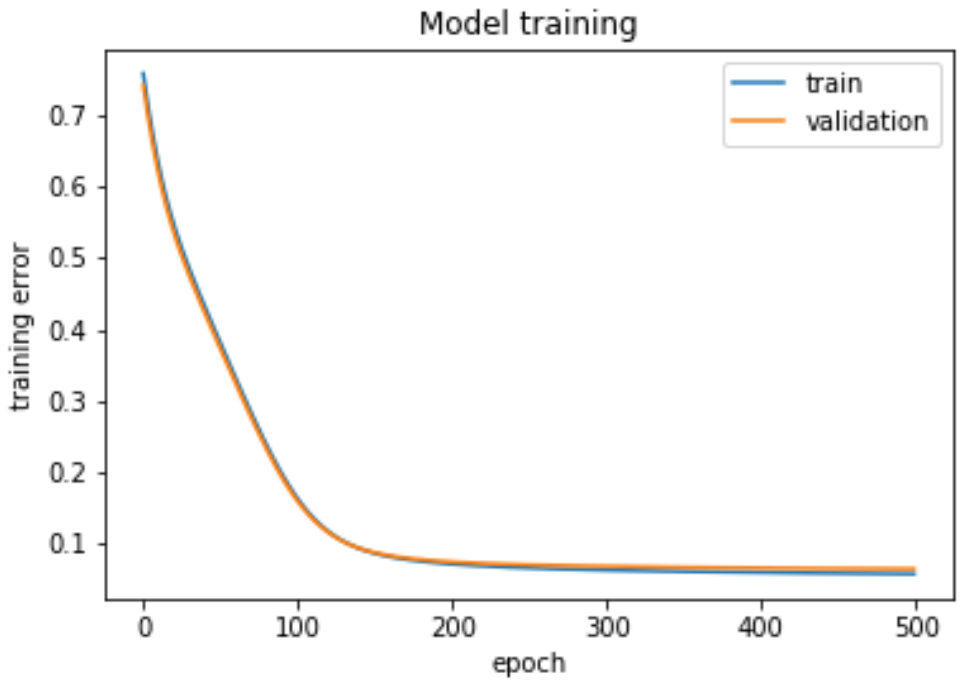

Figure 1: Training and validation error for a 30x3 MLP