**Assignment 4a**

Initially we started by exploring the available data in the training set, already here we could see that there are plenty of attributes that have a lot of missing data and in some cases over 80 percent of it. In this case we decided it is better to drop the attributes that have too much missing data and set a threshold of 20 percent (12000 cases) missing data as the maximum amount, if an attribute have more we will removed it. Attributes removed are the following: aa\_000, ab\_000, ad\_000, bk\_000, bm\_000, bn\_000, bo\_000, bp\_000, bq\_000, br\_000, cf\_000, cg\_000, co\_000, cr\_000, ct\_000, cu\_000, cv\_000, cx\_000, cz\_000, da\_000, db\_000, dc\_000.

After the missing data attributes were removed there were still some rows in the attributes that were left that had missing data and had to be addressed before we could train the model. We did an 80/20 split with the training data before we imputed the missing data that was left. The method used for this was the mean value of each attribute. This was done in turn to not make each set affect the other with its data.

At this stage we tested support vector classifiers (SVC) with different kernels (RBF, linear, Poly) and different degrees (3, 5, and 8). Overall SVC with RBF kernel with degree 8 worked best. We also tried other algorithms such as random forest (RF), k nearest neighbours (KNN) and multinomial naïve bayes (MultiNB). Already at this stage we could see that RF is outperforming the other algorithms, KNN and MultiNB are about the same and the worst one was SVC RBF.

We continued to import the test data set that was the important to try. We did the same pre-processing procedure as before although we did not split the data as we just want to test it on the already trained model. In the table below we can see the total cost of the models, as well as the confusion matrix for each model.

|  |  |
| --- | --- |
| Algorithm | Cost |
| SVC RBF | 183 070 |
| RF | 55 220 |
| MultiNB | 103 580 |
| KNN | 100 580 |

En bild som visar text

Automatiskt genererad beskrivning   
KNN confusion matrix. SVC confusion matrix.

   
MultiNB confusion matrix. RF confusion matrix.