Machine learning

Group project report

Katarzyna Kapuscinska (S4723708)

Tristan Timpers (S4614895)

Adam Bahelka (S4887832)

Jonathan Opitz (S4804120)

Magdalena Ringle (S4555953)

Process:

Finding data: Katarzyna

Cleaning data: Magdalena, Jonathan

Linear regression: Katarzyna

Vector support machine: Tristan

Gaussian naive Bayes: Adam, Jonathan

Comparison:

In our project, we have tried to predict Total early-stage entrepreneurial activity in European countries. We based our prediction on the number of features, such as Perceived opportunities, Perceived capabilities, fear of failure rate, Entrepreneurial intentions, Established business ownership, Female/Male TEA, High job creation and expectation and a few other factors. In order to achieve this, we have used Linear regression, Support Vector Machine and Gaussian Naive Bayes.

One of the problems we have faced is the continuous values of the target. This is not a problem with linear regression, however, in order to be able to use the Gaussian naive Bayes or Support Vector machine, we needed to round the numbers to integers. Therefore, we can also generate accuracy scores and confusion matrices. Because of this difference, we will be mainly comparing the SVM with GNB, linear regression will be accessed individually.

We received the following accuracy scores:

Naive Bayes: 19%

Support Vector Machine: 22%

Therefore, we can only see a slight increase in performance in the SVM.

Linear regression:

Mean Absolute Error: 1.4528630501550506

Mean Squared Error: 3.4326069784211195

Root Mean Squared Error: 1.8527296020793536

For the Linear Regression, we can see that there is a slight error. However, it is not too big. This difficulty to accurately predict the exact number also explains the fairly low accuracy score for

SVM and GNB. It can also be seen from the confusion matrix, where most of the predictions are only one or maximum two numbers away from the actual target.

Conclusion:

Although the use of the SVM and GNB might be more efficient for different types of prediction, it can also help us in this project. The accuracy score might seem quite low at first glance, however it is still 3 times better than the random guess and therefore can help with predicting Total early-stage entrepreneurial activity.