Naïve Bayes

* Ignore the feature 1 column and set creditability as the target and ID as meta data
* Split data into training and test set using data sampler, where training set included 70% of the original data and test set included the rest.
* Naïve Bayes is chosen as the predictive algorithms since it is easy to operate and the algorithms run fast to generate the prediction output.

Test and score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | AUC | CA | F1 | Precision | Recall |
| Naïve Bayes | 0.824 | 0.766 | 0.767 | 0.767 | 0.766 |

* Precision equal to 0.767, where 76.7% of the creditability is correctly identified

Confusion matrix

* Chart

  Description automatically generated number of instances
* Table

  Description automatically generated proportion of predicted
* Table

  Description automatically generated proportion of actual
* False negative equal to 67 (17.2% of actual good), and false positive is 64 (37.6% of actual bad)

ROC

Chart, line chart

Description automatically generated

Calibration plot

Graphical user interface, chart

Description automatically generated (target at bad creditability)

* The line is the cut off of good and bad creditor, on the right is bad creditor and left is the good creditor
* We can find the maximum accuracy by changing the threshold probability
* Highest F1 score occur when threshold equal to 0.43

Chart

Description automatically generated with medium confidence (target at good creditability)

* When threshold probability equal to 0.25, naïve bayes have the highest F1 score.

Cost-Sensitive learning

* In Cost-Sensitive learning, a penalty is associated with an incorrect prediction and referred to a ‘cost’ of the model
* It is worse to class a customer as good when they are bad (5), than it is to class a customer as bad when they are good (1)
* Table

  Description automatically generated
* The cost of the Naïve Bayes model is equal to 387 (using SUMPRODUCT in excel)

**Clustering (with Manifold learning)**

Diagram

Description automatically generated

* Use manifold learning in data reduction process since most data are categorical
* Choose k-means as the clustering methods and fix the number of clusters to 4
* Chart, bar chart

  Description automatically generated
* All clusters have more customers with good creditability than bad
* Timeline

  Description automatically generated
* Boxplot show the relationship of cluster and age
* The average age in Cluster 3 is the highest

**Clustering (without data reduction)**

* K-means were used with 2 cluster

Chart

Description automatically generated

* Relationship between cluster and creditability

Timeline

Description automatically generated

* Relationship between cluster and age
* People in cluster 1 are having higher age on average

Timeline

Description automatically generated

* Cluster and purpose of the loan
* In cluster 2, over half of the purpose are for new car and radio/television