account\_bal 3

payment\_status 5

purpose 10

savings\_bond\_value 5

employed\_since 5

sex\_marital 4

guarantor 3

most\_valuable\_asset 4

concurrent\_credits 3

type\_of\_housing 3

job 3

telephone 2

foreign 2

encoded categorical variables that are in set

normalized amounts for credit\_amount, duration, and age

|  | **credit\_amount** | **duration** | **age** |
| --- | --- | --- | --- |
| **count** | 1000.000000 | 1000.000000 | 1000.000000 |
| **mean** | 7.788691 | 20.903000 | 35.546000 |
| **std** | 0.776474 | 12.058814 | 11.375469 |
| **min** | 5.521461 | 4.000000 | 19.000000 |
| **25%** | 7.219276 | 12.000000 | 27.000000 |
| **50%** | 7.749107 | 18.000000 | 33.000000 |
| **75%** | 8.287088 | 24.000000 | 42.000000 |
| **max** | 9.821409 | 72.000000 | 75.000000 |

Correlation between variables

Most Positive Correlations:

sex\_marital\_A92 0.075493

type\_of\_housing\_A153 0.081556

account\_bal\_positive\_bal 0.089895

type\_of\_housing\_A151 0.092785

concurrent\_credits\_A141 0.096510

purpose\_A40 0.096900

employed\_since\_A72 0.106397

credit\_amount 0.109570

most\_valuable\_asset\_real\_estate 0.125750

payment\_status\_A31 0.134448

payment\_status\_A30 0.144767

savings\_bond\_value\_A61 0.161007

duration 0.214927

account\_bal\_neg\_bal 0.258333

target 1.000000

Name: target, dtype: float64

Most Negative Correlations:

account\_bal\_no\_acc -0.322436

payment\_status\_A34 -0.181713

type\_of\_housing\_A152 -0.134589

savings\_bond\_value\_A65 -0.129238

most\_valuable\_asset\_none -0.119300

concurrent\_credits\_A143 -0.113285

purpose\_A43 -0.106922

purpose\_A41 -0.099791

age -0.091127

savings\_bond\_value\_A64 -0.085749

foreign -0.082079

sex\_marital\_A93 -0.080677

employed\_since\_A74 -0.075980

savings\_bond\_value\_A63 -0.070954

employed\_since\_A75 -0.059733

these are the polynomial features I got using the features

['1', 'duration', 'account\_bal\_neg\_bal', 'account\_bal\_no\_acc', 'duration^2', 'duration account\_bal\_neg\_bal', 'duration account\_bal\_no\_acc', 'account\_bal\_neg\_bal^2', 'account\_bal\_neg\_bal account\_bal\_no\_acc', 'account\_bal\_no\_acc^2']

Correlation between polynomial features

account\_bal\_no\_acc -0.322436

account\_bal\_no\_acc^2 -0.322436

duration account\_bal\_no\_acc -0.232697

duration^2 0.200996

duration 0.214927

account\_bal\_neg\_bal 0.258333

account\_bal\_neg\_bal^2 0.258333

duration account\_bal\_neg\_bal 0.303343

target 1.000000

1 NaN

account\_bal\_neg\_bal account\_bal\_no\_acc NaN

list of features from the polynomial feature selection that look like that they can tell us the best about a persons account and whether they have good or bad credit

['duration^2', 'duration account\_bal\_neg\_bal', 'duration account\_bal\_no\_acc', 'account\_bal\_neg\_bal^2', 'account\_bal\_neg\_bal account\_bal\_no\_acc', 'account\_bal\_no\_acc^2']

**Evaluation criteria**

Let's have a look at the different options available.

| **Evaluation criteria** | **Description** |
| --- | --- |
| Accuracy | (true positive+ true negative) / total obs |
| Precision | true positive/ total predicted positive |
| Recall | true positive/ total actual positive |
| F1 | 2\* precision \* recall / (precision + recall) |
| AUC ROC | Area Under ROC Curve (TPR Vs. FPR for all classification thresholds) |

* Accuracy: The german dataset is an imbalanced dataset. Accuracy would give a high score by predicting the majority class but would fail to predict the minority class, which is the defaulters. Hence, this is not a suitable metric for this dataset.
* Precision: Precision is a good metric when the costs of false positive is high. Example, email spam detection.
* Recall: This metric is suitable when the costs of false negative is high. Example, predicting a defulter as not defaulter. This costs huge loss for the bank. Hence, this is a suitable metric for our case.
* F1: measure of both precision and recall.
* AUC ROC: It is the plot of TPR vs FPR. All other criteria discussed here assumes 0.5 as the decision threshold for the classification. However, it maynot be always true. The AUC helps us evaluate the performance of the model for all classification thresholds. The higher the value of the AUC metric, the better the model.
* True positive rate (TPR) = TP/ Total actual positive
* False positive rate (FPR) = FP/ Total actual negative

We will use Recall and AUC ROC as evaluation metric.

Logistic regression

LR recall\_test: 0.1 auc\_roc\_test: 0.58 LR recall\_train: 0.01 auc\_roc\_train: 0.65 precision recall f1-score support 0 0.73 1.00 0.84 141 1 1.00 0.10 0.18 59 accuracy 0.73 200 macro avg 0.86 0.55 0.51 200 weighted avg 0.81 0.73 0.65 200

| **MODEL** | **TEST ACCURACY** |
| --- | --- |
| Decision Tree | 70% |
| Logistic Regression | 71.2% |
| Random Forest | 79.2% |
| GaussianNB | 72.8% |
| KNN | 71.2% |
| SVC | 76.8% |