

**Predicting credit card default of customers using GLRM**

**Team**: Deepika B (191046020), Swathi Patwari (191046043)

**Guide**: Sudarsan N S Acharya



**Overview**

Credit card method is one of the popular payment methods. But there are some disadvantages like credit card defaults. The main aim of this project is to predict whether the customer is going to default in the coming month or not. Along with prediction, the features are also going to be analysed to understand which are the factors leading to default. The dataset containing over 30000 samples is pre-processed, and given as the input to the Generalized low rank model. The reduced dimension dataset is classified using supervised machine learning algorithms and the accuracy, recall and precision is calculated and compared.

**Data**

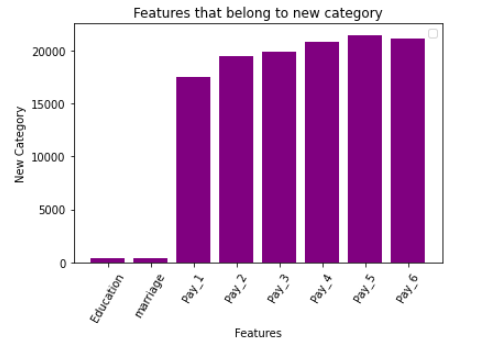
We obtained the data set from UCI machine learning repository. This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

|  |  |  |
| --- | --- | --- |
| **Variables** | **Dimension** | **Description** |
|  | ID | Identifier, redundant variable |
| X1 | LIMIT BALANCE | Amount of the given credit (NT dollar) |
| X2 | SEX | Gender (1 = male; 2 = female) |
| X3 | EDUCATION | Education (1 = graduate school; 2 = university; 3 = high school; 4 = others) |
| X4 | MARRIAGE | Marital status (1 = married; 2 = single; 3 = others) |
| X5 | AGE | Age in years |
| X6 to X11 | PAY\_0 to PAY\_6 | History of past payment.  Repayment status from September 2005 to April 2005 |
| X12 to X17 | BILL\_AMT1 to BILL\_AMT6 | amount of bill statement in September 2005 to April 2005 |
| X18 to X23 | PAY\_AMT1 to PAY\_AMT6 | amount paid in September 2005 to April 2005 |
| Y | Default payment next month | Default where 1 represent customer if defaulted and 0 represent if not defaulted(Target variable) |

Table 1: Dataset description

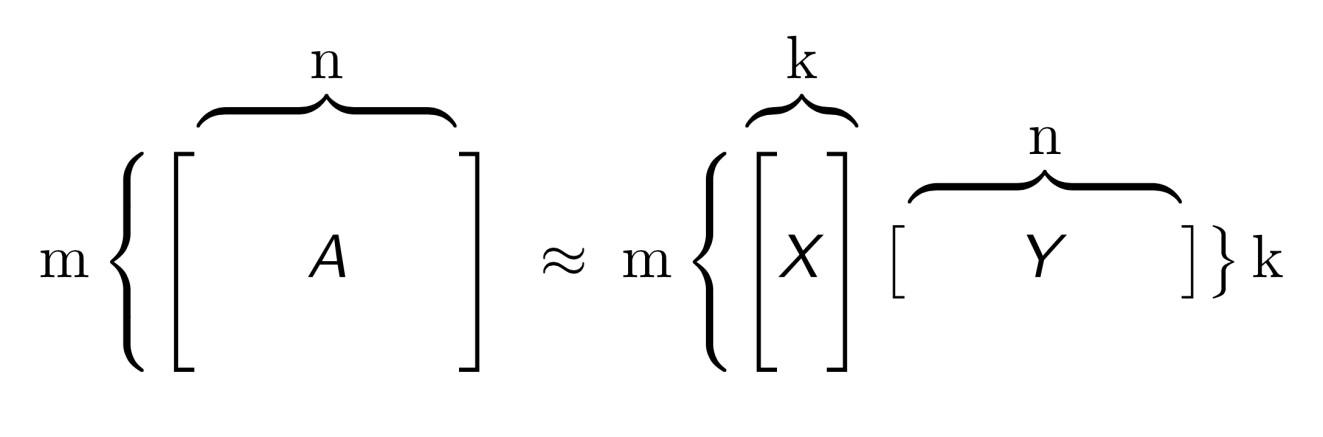
**Approach to correct the incorrect features**

Here we considered incorrect values as new category.

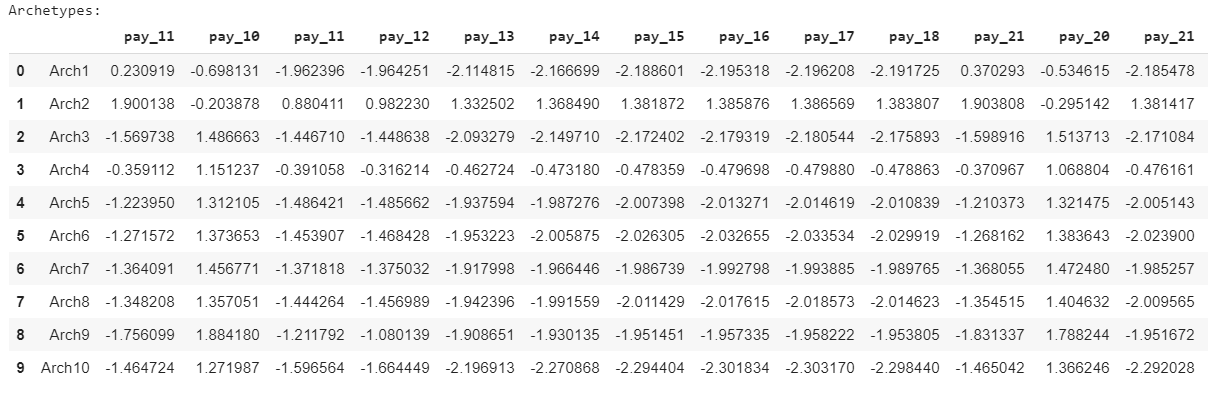


**GLRM**

GLRM is an extension of PCA, but unlike PCA, GLRM can handle Boolean, categorical, numeric data. It has the capabilities to handle mixed data types.



**Output after applying GLRM**

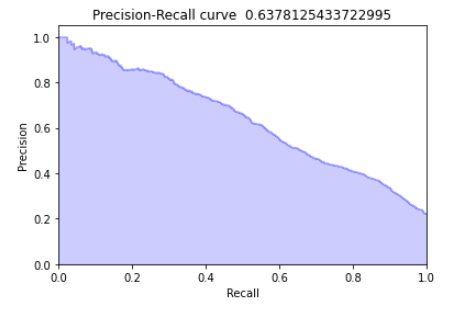


**Result**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** |  | **K=3** | **K=4** | **K=5** | **K=6** | **K=7** | **K=8** | **K=9** | **K=10** |
| **Logistic Regression** | specificity | 1.00 | 0.97 | 0.98 | 0.97 | 0.99 | 0.98 | 0.98 | 0.99 |
| sensitivity | 0.02 | 0.28 | 0.30 | 0.30 | 0.18 | 0.33 | 0.18 | 0.19 |
| precision | 1.00 | 0.79 | 0.84 | 0.78 | 0.85 | 0.83 | 0.82 | 0.89 |
| **Decision Tree** | specificity | 0.85 | 0.86 | 0.85 | 0.88 | 0.84 | 0.88 | 0.84 | 0.84 |
| sensitivity | 0.48 | 0.57 | 0.63 | 0.60 | 0.53 | 0.59 | 0.48 | 0.46 |
| precision | 0.49 | 0.55 | 0.61 | 0.60 | 0.49 | 0.59 | 0.47 | 0.45 |
| **SVM**  **(linear)** | specificity | 1.00 | 0.98 | 0.98 | 0.97 | 1.00 | 0.98 | 1.00 | 0.99 |
| sensitivity | 0.02 | 0.28 | 0.30 | 0.31 | 0.18 | 0.33 | 0.19 | 0.16 |
| precision | 1.00 | 0.79 | 0.83 | 0.78 | 0.85 | 0.83 | 0.82 | 0.89 |
| **SVM**  **(rbf)** | specificity | 0.99 | 0.98 | 0.97 | 0.97 | 0.98 | 0.97 | 0.98 | 0.98 |
| sensitivity | 0.17 | 0.26 | 0.32 | 0.36 | 0.21 | 0.37 | 0.21 | 0.22 |
| precision | 0.84 | 0.80 | 0.82 | 0.81 | 0.80 | 0.83 | 0.82 | 0.83 |
| **SVM**  **(poly)** | specificity | 0.99 | 0.97 | 0.97 | 0.97 | 0.98 | 0.96 | 0.98 | 0.98 |
| sensitivity | 0.19 | 0.28 | 0.28 | 0.33 | 0.21 | 0.40 | 0.19 | 0.20 |
| precision | 0.77 | 0.79 | 0.80 | 0.76 | 0.79 | 0.79 | 0.80 | 0.78 |

Table 2. sensitivity, specificity and precision values for different values of k for GLRM

**Precision-Recall curve**



**Summary**

Credit card Customers default was predicted using Generalized Low rank models and various algorithms like logistic regression, decision tree and Support vector machines. For different values of k, from k=3 to k=10 in GLRM, each algorithm’s sensitivity, specificity and precision was observed.

At approximately k=10 we obtained a specificity value of 99.46% and a sensitivity of 19.05%

**Future Work**

For the same dataset we can use a different approach for analysis. We can use unsupervised machine learning model like clustering to cluster the features into different groups based on correlation values. This might improve the specificity and sensitivity values further more.