Applied Data Science Project - Credit Default

Mallik Challa

December 15, 2019

Contents

Introduction

The objective of this project is to gain valuable insights about customers credit repayment ability using the "default of credit card clients Data Set" available on UCI machine learning repository via exploratory data analysis, dimension reduction techniques and Logistic regression modeling.

Primarily, the project aims to answer the following questions:

1. How accurately do the input features help in predicting customers repayment abilities?
2. Which features are the key to predict the customers repayment capability?

Data Description

The dataset is publicly available on UCI Machine Learning Repository (Yeh & Lien, 2009). The data was collected by the authors from Taiwan to research the case of customers defaulting payment and compare the predictive accuracy of probability of default among six data mining methods. The dataset contains 30000 observations and 25 features. Preliminary checks made on the data show that there is no missing data in any of the rows or columns.

The target variable is 'default.payment.next.month' which indicates whether the customer defaulted a payment following the 6-month period (values: 1=yes, 0=no).

Following is the brief description of the input features in the dataset:

ID: ID of each client.
LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit).
SEX: Gender (1=male, 2=female).
EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown).
MARRIAGE: Marital status (1=married, 2=single, 3=others).
AGE: Age in years.
PAY_0 thru PAY_6: Payment History - Repayment status in September, 2005 thru April, 2005 (-2: No consumption; -1: Paid in full; 0: The use of revolving credit; 1=payment delay for one month, 2=payment delay for two months and so on until 8=payment delay for eight months, 9=payment delay for nine months and above).
BILL_AMT1 thru BILL_AMT6: Billed amount from statements in September, 2005 thru statement in April, 2005 (NT dollar).
PAY_AMT1 thru PAY_AMT6: Amount of previous payment in September, 2005 thru previous payment in April, 2005 (NT dollar).

Note: The dependent variable 'default.payment.next.month' will be renamed as 'Target' and will be referred to as 'Target' hereafter in the paper.

Exploratory Data Analysis

Target variable counts by class (Repayment ability for next month):

First, target variable counts are examined to get some insights related to the distribution of the population with respect to the dependent variable.

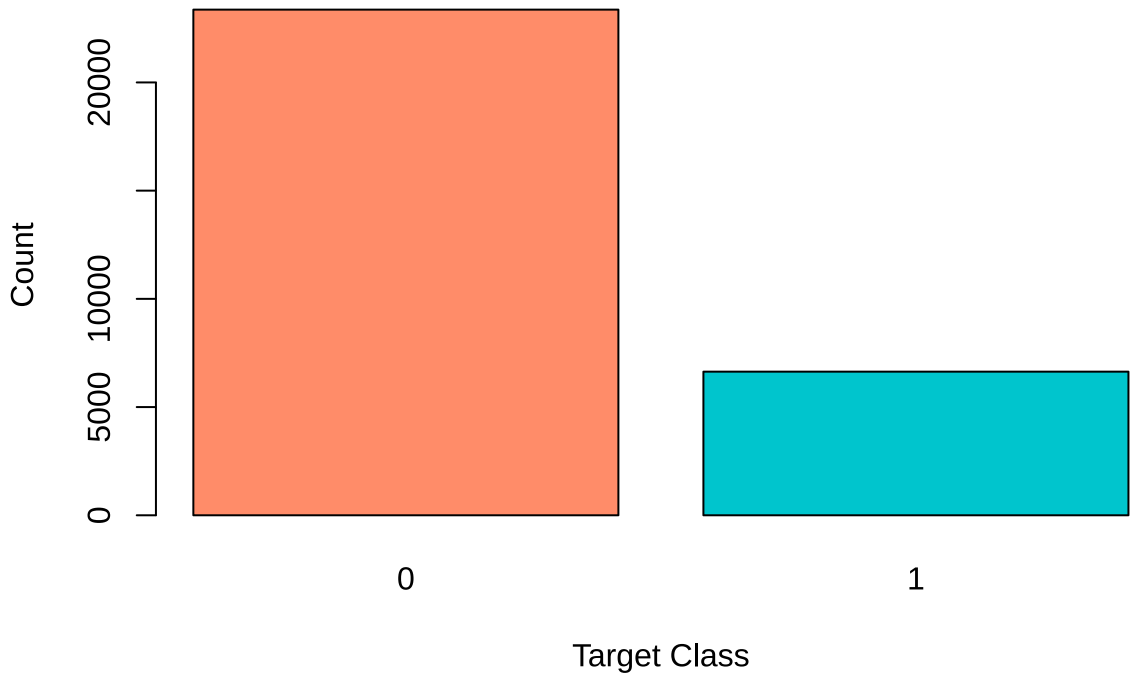

Figure 1: Counts by Target class

It is evident from the bar plot in Figure 1 that the target class is not evenly distributed. Only 22% of the sample population belong to the class of population that defaulted (i.e. class=1) on the next payment which is more than three times the other class. The imbalance in data is an important point to note since any prediction made using the given data could be more influenced by the majority class.

Visualizing Input Features

Next step is to examine the input features present in the dataset.

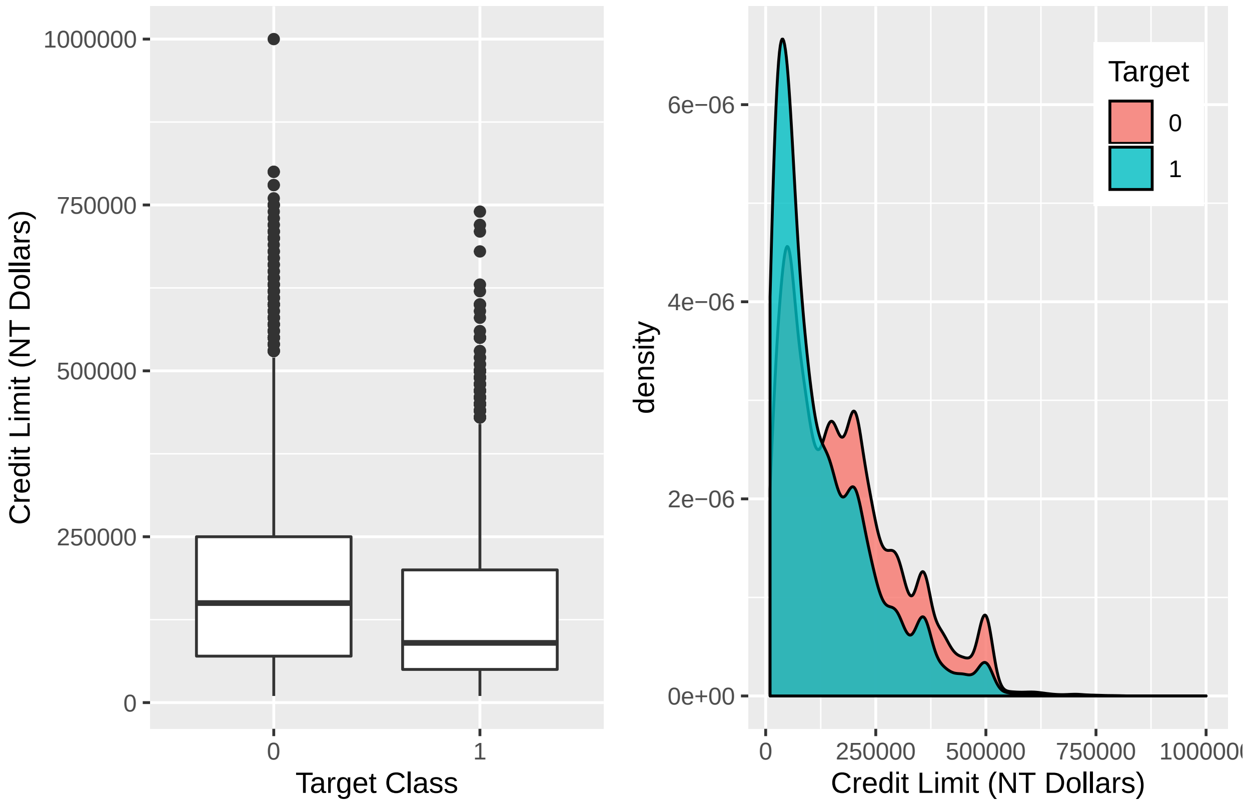

Figure 2: Credit Limit − BoxPlot and Density Plot

From the Boxplot for credit limit (LIMIT_BAL) in Figure 2, it can be noted that majority of the population has a credit limit below 250,000. The density plot also shows the peak of the distribution in both classes is below 100,000. However, the boxplot also indicates a few outliers with credit limit over 500,000 in class=0 and over 350,000 in class=1.

Further analysis with respect to credit limit indicates that the total number of outliers is 167, which is not a very high percentage compared to the total population. In order to minimize any bias due to these outliers, it would be prudent to drop the observations containing these outliers considering the smaller number.

Next set of features in the dataset include demographic variables - ‘AGE’, ‘SEX’, ‘EDUCATION’ and ‘MARRIAGE’. Following Figure-3 shows the distribution plots for each of the demographic variables by the target class.

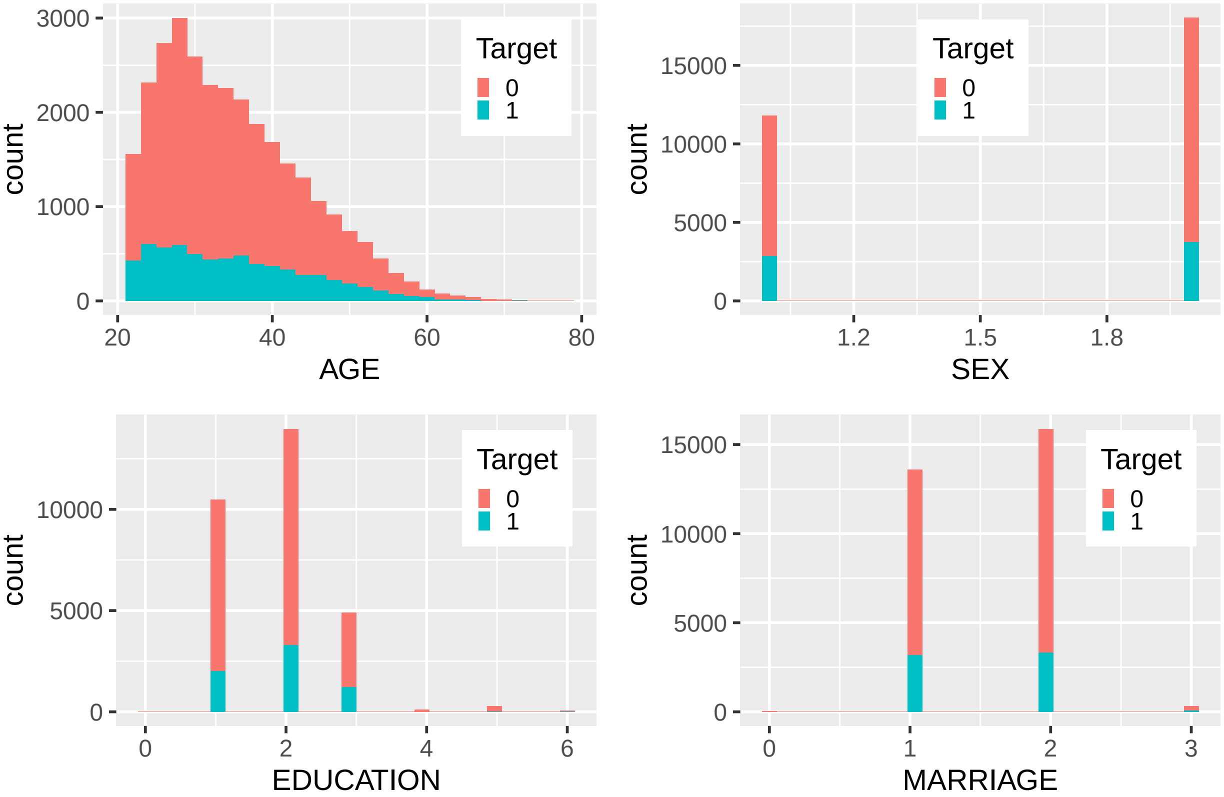

Figure 3: Distribution of Demographic variables by Target

‘AGE’ distribution (top left in Figure 3) is close to normal though with a slightly longer tail on the right (Which is expected since there cannot be customers below age 20) and the distribution peaks in between 20 and 40 years. It can also be noted that the Target class distribution is similar across all ages though the credit defaulting class appear more in between 25 to 35 years of age. Other demographic variables also show a similar pattern where in the distribution across different values appears even with respect to the target class.

There are three different categories of quantitative variables in the input each giving information for the 6-month period from April, 2005 thru September, 2005. Among these variables, payment history variables (PAY_0 thru PAy_6) could be the most important features to predict the repayment ability of the customer. Figure 4 shows the distribution for each of the 6 payment history variables by the target class.

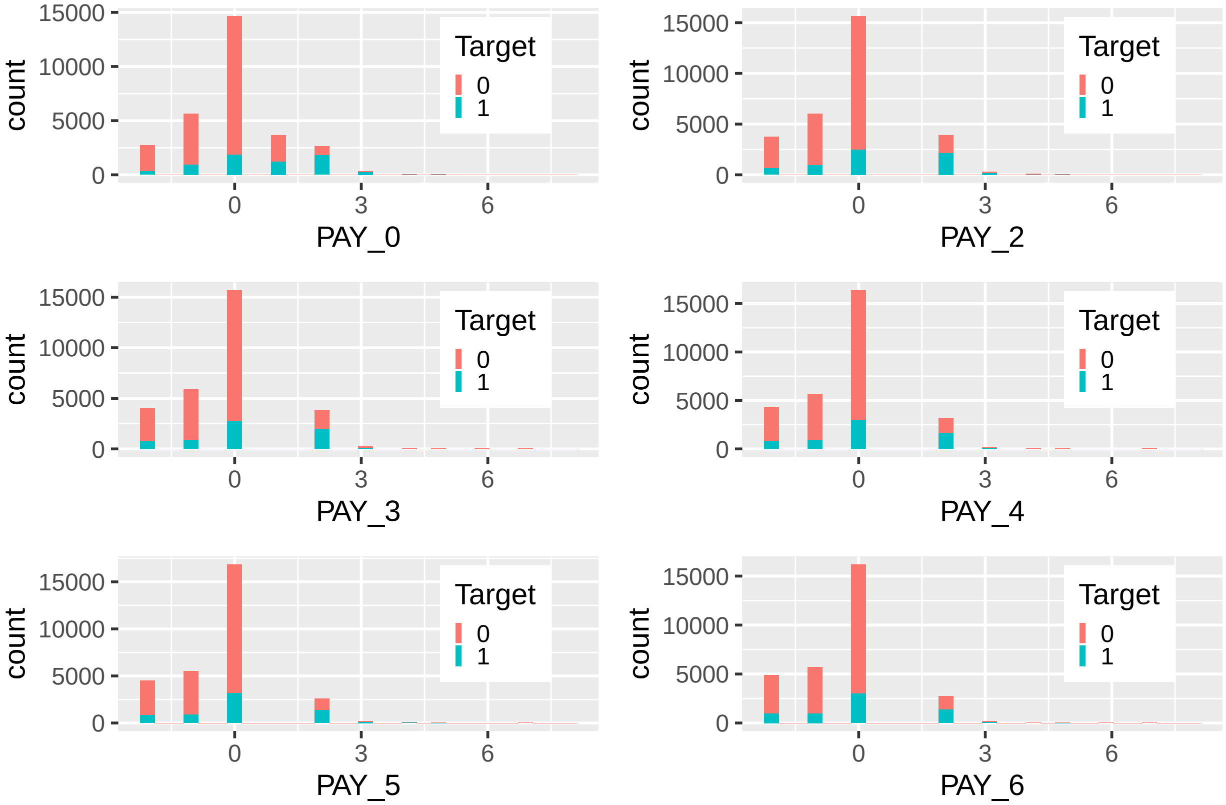

Figure 4: Distribution of Payment History by Target

The distributions of the 6 payment history variables (PAY_0 being for the most recent month) indicate that the proportion of defaulted payments i.e. Class = 1, is higher among the customers with a poor payment history i.e. those with history values greater than 0 in the histograms on the right.

In the final step of data exploration, correlation of input features can be examined to see if dimension reduction can be explored in the next step.

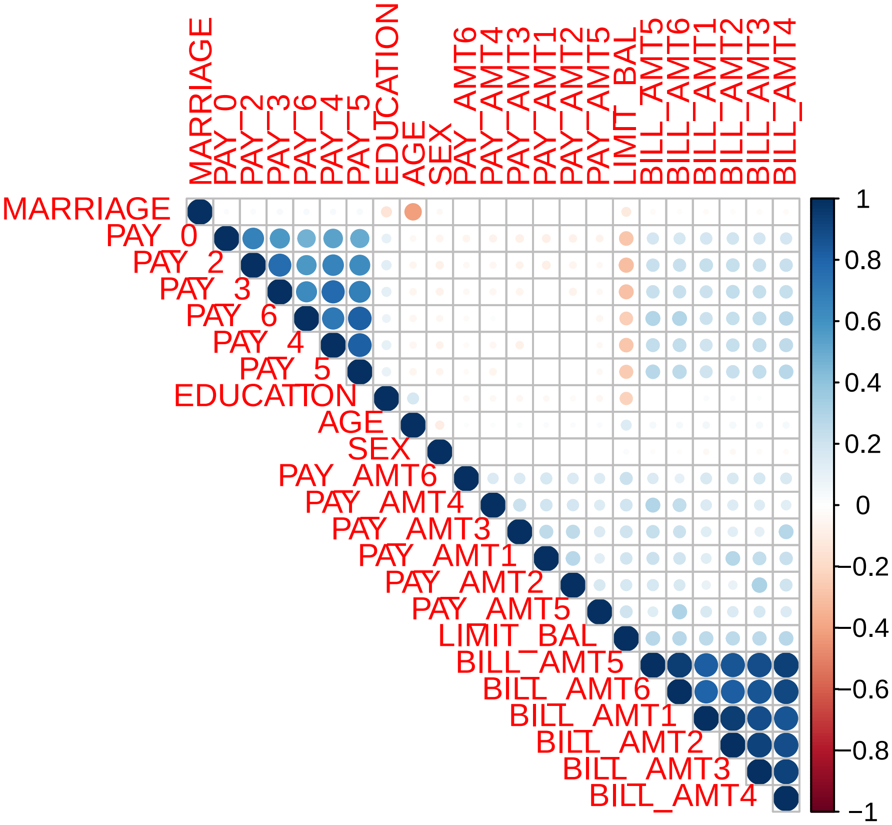

Figure 5: Correlation Plot for Input Features

Following are the observations from the correlation plot in Figure 5:

- Billed amount variables do have a strong correlation among each other which is expected since they are reflecting the cumulative amounts.
- Payment history variables also show a strong positive correlation among themselves indicating that payment history does have a pattern.
- Credit limit (LIMIT_BAL) shows some positive correlation with the billed and paid amounts while it shows some negative correlation with the payment history variables.
- Age and marital status (MARRIAGE) show some positive correlation which is quite intuitive.

Factor Analysis (EFA)

Based on the observations from correlation plot, it can be evaluated if the data is a good fit for exploratory factor analysis and thereby reduce the number of dimensions.

KMO Test for Factor Analysis

KMO test evaluation gives an **overall MSA of 0.796** which implies that factor analysis is appropriate for the given data. MSA values for most of the input variables is greater than 0.5 except for those shown in Table 1. The features with MSA much below 0.5 (like PAY_AMT2 and PAY_AMT5) could be omitted while retaining the rest of them to avoid elimination of all the payment amount variables.

Table 1: **Features with MSA values < 0.5**

	MSA
PAY_AMT1	0.47
PAY_AMT2	0.37
PAY_AMT3	0.45
PAY_AMT4	0.47
PAY_AMT5	0.38

Scree Plot

A Scree plot can be used to determine number of factors.

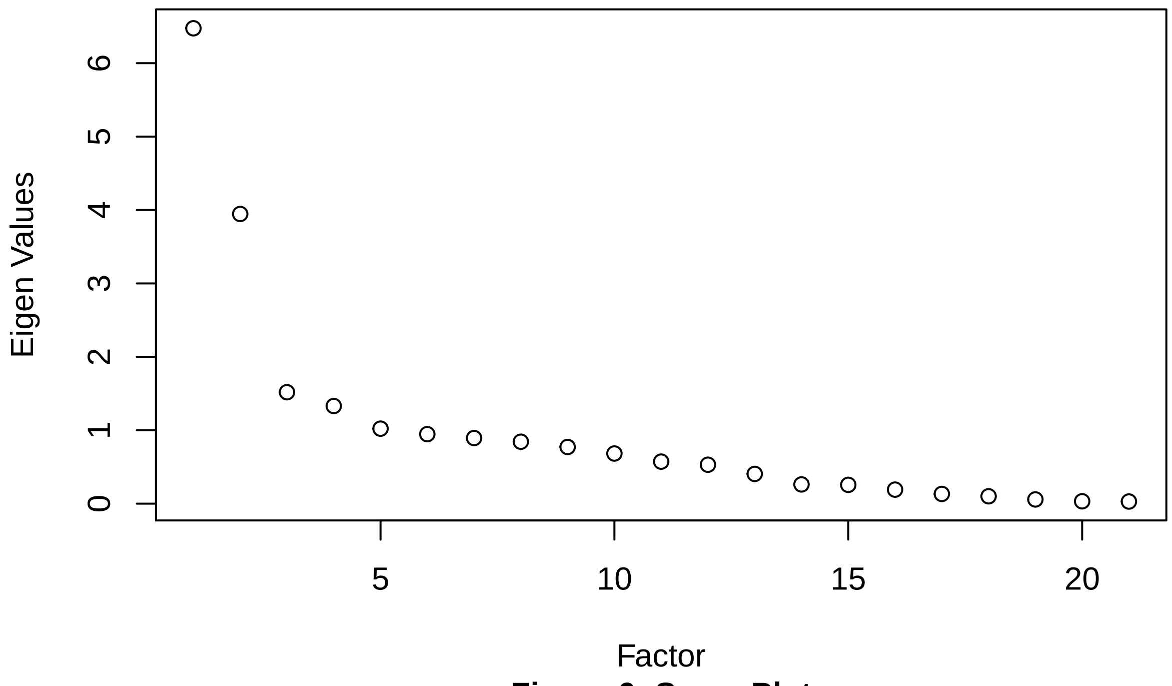

Figure 6: Scree Plot

From the Scree plot in Figure 6, we can see that the number of factors that can be considered is either 3 or 4. Choosing the factors as 3, will result in elimination of all the demographic variables. One can argue based on intuition that the demographic variables could have an influence on the customers repayment capabilities for which reason we may choose factors as **4** for subsequent analysis.

Factor Analysis with 4 Factors

Factor analysis with 4 factors will be done to get the factor scores which can be used in modeling later.

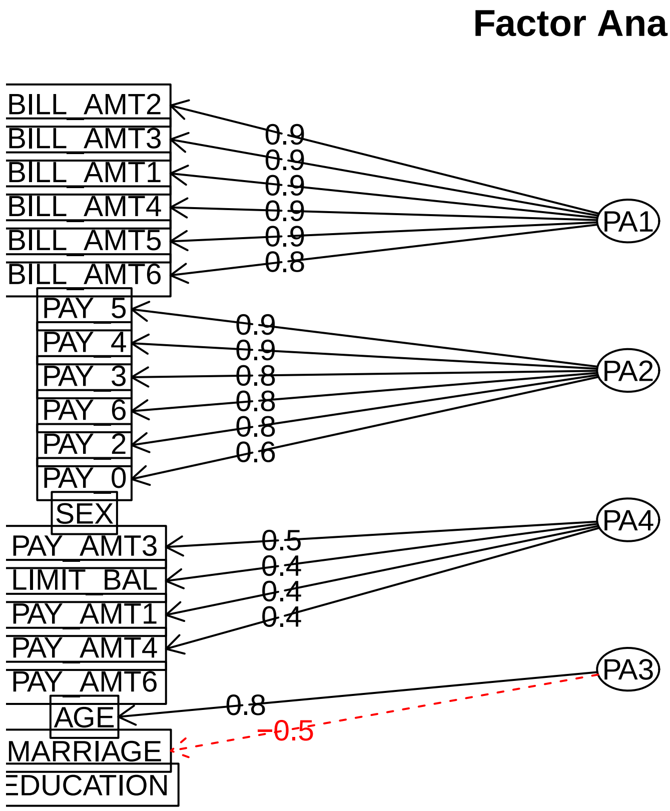

Figure 7: Factors from Factor analysis

Factor Analysis diagram in Figure 7, shows that following are the 4 factors and their corresponding variables:

- **Factor-1:** All billed amounts for preceding 6 months.
- **Factor-2:** Payment History values for preceding 6 months.
- **Factor-3:** Credit limit and all paid amounts for preceding 6 months.
- **Factor-4:** Age and Marriage status.

Logistic Regression

Given that the target data is binary, and the objective is to predict the classes of the target variable, logistic regression can be considered as an appropriate modeling algorithm.

Logistic Regression with the Factors from EFA

To perform logistic regression analysis using the factors obtained from factor analysis, first step is to combine the target variable with factor scores and name the factors appropriately. Table 2 shows sample records after creating the combined dataset.

Table 2: **Predictor and Labeled Factors samples**

Target	Bill_Amt	Pay_Hist	Pmt_Amt	Demographic
1	-0.2193560	-0.4868761	-1.4810010	-0.4429037
1	-0.7291120	0.7446294	-0.2416406	-0.8412291
0	-0.4646317	0.2542151	-0.2812157	-0.2896303

Next, the dataset is split into training set (70%) and test set (30%). Training data will be used to fit the logistic regression model and the model will be used to predict the target classes on test data.

Table 3 shows the coefficients determined by the logit model for the 4 factors. It can be noted that 3 of the 4 Factors are statistically significant while the factor including BILL_AMT variables is comparatively insignificant since the p-value is greater than 0.05.

Table 3: **Logit Model Coefficents for Factors**

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.3856532	0.0185980	-74.505651	0.0000000
Bill_Amt	-0.0209842	0.0198969	-1.054649	0.2915859
Pay_Hist	0.6523388	0.0188428	34.620010	0.0000000
Pmt_Amt	-0.5143505	0.0305724	-16.824018	0.0000000
Demographic	0.1232388	0.0208968	5.897485	0.0000000

The deviance values (residual - 2.032624×10^4 and null - 2.215207×10^4) with a difference of 4 degrees of freedom also suggests that the model as a whole fits *significantly better than an empty model.*

Model performance (with factors)

Confusion matrix can be used for deriving the accuracy which is most straightforward and intuitive metric for measuring the performance of the model. Table 4 is the confusion matrix for the logit model using the factors which indicates an **overall accuracy of around 80%**.

However, it may be noted that the prediction accuracy for minority class i.e. class 1, is **less than 15%**. This can be explained due to the imbalance in the target classes in the input dataset.

Table 4: **Confusion Matrix (using Factors)**

	0	1
0	6861	124
1	1679	286

In case of imbalanced datasets, **AUC score** (Area under the ROC curve) is a better performance metric since ROC curves are insensitive to class imbalance (Horton, 2016). Hence we can look at the AUC score as an evaluation metric for the model performance.

The logit model with factors gives an **AUC score of 0.564** which is much lower despite the an accuracy of close to 80%.

Logistic Regression with all the input features

Furthermore, logistic regression analysis can be done using all the input features and then compared with the results obtained from the previous model using factors. Before fitting the logistic model, all the independent variables need to be standardized i.e. center the values around 0 with a standard deviation of 1.

Table 5 shows the coefficients determined by the logit model for the input features. The table also gives the z-statistic which can be used to determine the most important features. As can be seen, the most recent payment history (PAY_0) feature has a significant role to play based on the z-statistic.

Table 5: **Logit Model Coefficients**

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.4765456	0.0200701	-73.5692572	0.0000000
LIMIT_BAL	-0.0893725	0.0241614	-3.6989724	0.0002165
SEX	-0.0539761	0.0180463	-2.9909808	0.0027808
EDUCATION	-0.0648000	0.0198876	-3.2583097	0.0011208
MARRIAGE	-0.0673077	0.0198424	-3.3921130	0.0006936
AGE	0.0687568	0.0197271	3.4853949	0.0004914
PAY_0	0.6500566	0.0237900	27.3247946	0.0000000
PAY_2	0.0760471	0.0290884	2.6143403	0.0089400
PAY_3	0.1067092	0.0323941	3.2940951	0.0009874
PAY_4	0.0207367	0.0351262	0.5903475	0.5549577
PAY_5	0.0504962	0.0364991	1.3834901	0.1665146
PAY_6	0.0195375	0.0302085	0.6467560	0.5177899

	Estimate	Std. Error	z value	Pr(>|z|)
BILL_AMT1	-0.5284711	0.1063839	-4.9675842	0.0000007
BILL_AMT2	0.4246054	0.1277201	3.3244986	0.0008858
BILL_AMT3	-0.0870562	0.1097776	-0.7930235	0.4277641
BILL_AMT4	0.0026282	0.1034456	0.0254066	0.9797307
BILL_AMT5	-0.0826846	0.1163539	-0.7106302	0.4773134
BILL_AMT6	0.1618360	0.0922644	1.7540465	0.0794225
PAY_AMT1	-0.2551098	0.0451275	-5.6530882	0.0000000
PAY_AMT2	-0.1496475	0.0588159	-2.5443381	0.0109485
PAY_AMT3	-0.0813347	0.0357784	-2.2732895	0.0230087
PAY_AMT4	-0.0102337	0.0285968	-0.3578611	0.7204473
PAY_AMT5	-0.1026370	0.0346419	-2.9627980	0.0030486
PAY_AMT6	-0.0575745	0.0299322	-1.9234993	0.0544174

Model performance (with all features)

Confusion matrix based on the predictions from the logistic regression on the test data using all input features in Table 6 shows a marginally increased accuracy compared to the previous model. Similarly, the new model gives a slightly higher **AUC score of 0.6**.

Table 6: **Confusion Matrix (using all features)**

	0	1
0	6732	206
1	1548	464

The increase in the metric scores is not very significant and to conclude if one model is significant than the other, additional comparision tests like t-tests over multiple runs may be required which is not in scope of this paper.

Logistic Regression using data generated by SMOTE

As observed, both the previous models have performed poorly when it comes to predicting the minority class which can be attributed to the imbalance in dataset. In order to tackle this scenario, data with minority class can be synthetically generated to have a more balanced input. This can be done using SMOTE method (Synthetic Minority Over-Sampling Technique) (Torgo, 2010). SMOTE creates new (synthetic) observations using the nearest neighbors of these cases. In the next logit model, before fitting the actual data, additional data will be generated using SMOTE to create a training set with equal distribution of both the classes.

Model performance (With SMOTE)

Table 7 shows the Confusion Matrix after running the logit model with data generated using SMOTE. The results show a similar overall accuracy level but the prediction accuracy for the minority class has increased considerably which is around **50%**. The new model gives an **AUC score of 0.686** which is also clearly higher (about 15%) compared to previous model. This shows that if the input has a more balanced data, the model performs better on the new data.

Table 7: **Confusion Matrix (After using SMOTE)**

	0	1
0	5906	1032
1	964	1048

Variable Importance

To determine the most important features which contribute towards the prediction, the absolute value of the t-statistic for each model parameter can be used in case of linear models like Logistic regression (Emerald, 2012).

Table 8 shows the top 10 features in the descending order of *score* which is the absolute z-statistic value obtained from the last logistic regression model. The table indicates the most recent payment (PAY_0) as the most important feature followed by prior month billing and payment variables. Demographic variables can be seen to exhibit almost similar importance level.

Table 8: **Variable Importance - Top 10**

	Features	Score
6	PAY_0	37.169
18	PAY_AMT1	10.290
12	BILL_AMT1	8.069
22	PAY_AMT5	6.902
1	LIMIT_BAL	5.872
4	MARRIAGE	5.773
19	PAY_AMT2	5.740
2	SEX	5.668
3	EDUCATION	5.489
5	AGE	5.338

Conclusion

How accurately do the features help in predicting customers repayment abilities?

Logistic Regression model using Factors from Factor analysis and a model with all input features have yielded a prediction accuracy of about 80% though the accuracy for predicting the minority class is much lower at 15%. Also, the AUC score, which is a better metric when evaluating imbalanced data, is only around 0.6. After addressing the issue of data imbalance by generating synthetic data for minority class, the overall accuracy remained about the same while the prediction accuracy for minority class increased significantly to 50% and the AUC score also improved to 0.69. Furthermore, different classification algorithms and neural networks may also be evaluated to increase prediction accuracy and AUC score.

Which features are the key to predict the customers repayment capability?

Logistic regression model results were used to examine the importance of features for predicting the repayment ability and it can be concluded that the most recent payment history (PAY_0) is the most important feature followed by other of payment, billing amount and demographic variables. Figure 8 shows the relative importance of all the input features. Additional modeling techniques like ensemble decision tree classifiers (Random Forest) can also be used to determine feature importance and compared for further research.

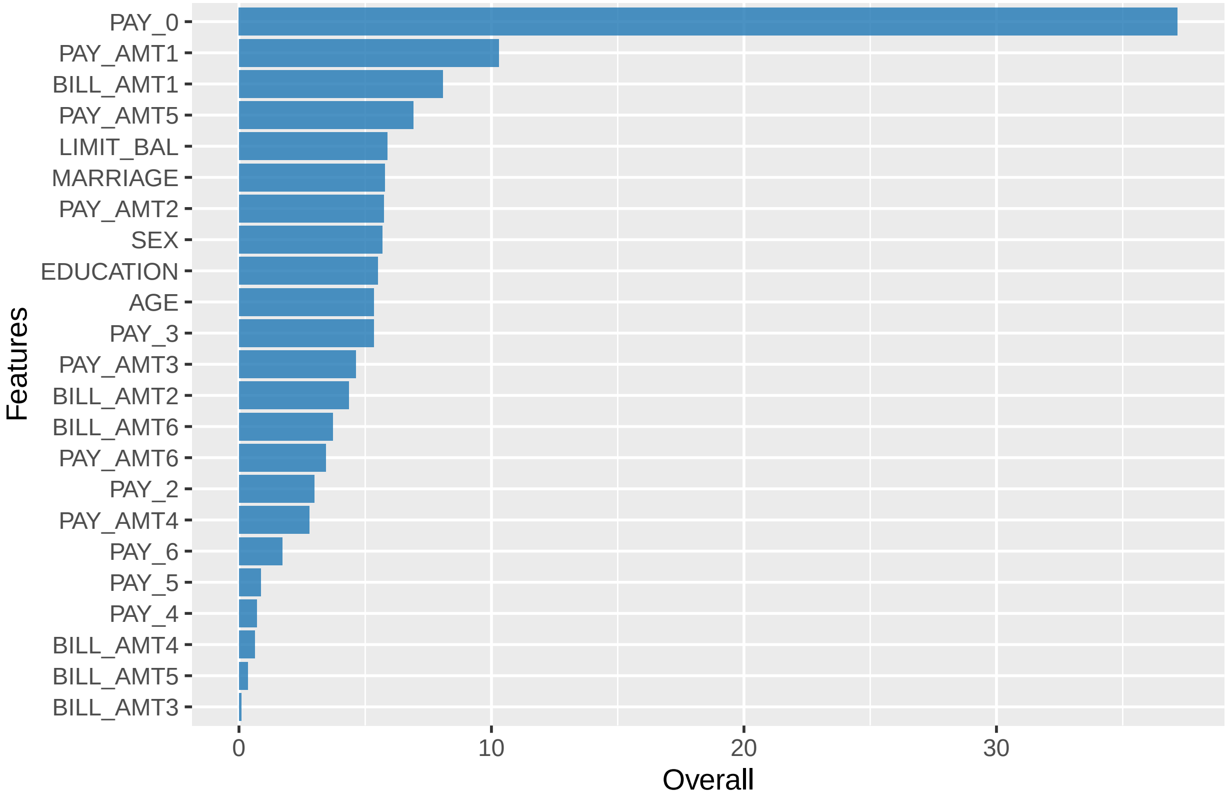

Figure 8: Variable Importance Plot

References

Code References

Byrnes, J. (2011). *Extra! Extra! Get Your gridExtra!*. Retrieved from https://www.r-bloggers.com/extra-extra-get-your-gridextra/

Holtz, Yon. (2018). *Building a nice legend with R and ggplot2*. Retrieved from https://www.r-graph-gallery.com/239-custom-layout-legend-ggplot2.html

R Core Team. (2018). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org/

Toulemonde, Emmanuel-Lin. (2019). *dataPreparation: Automated Data Preparation*. Retireved from https://CRAN.R-project.org/package=dataPreparation

Xie, Yihui. (2017). *An Introduction to the printr Package*. Retrieved from https://cran.r-project.org/web/packages/printr/vignettes/printr.html

Data and Technical References

Emerald. (2012). Variable Importance. Retrieved from https://r-forge.r-project.org/scm/viewvc.php/*checkout*/www/varimp.html?revision=894&root=caret

Horton, B. (2016). Calculating AUC the area under a ROC Curve. Retrieved from https://www.r-bloggers.com/calculating-auc-the-area-under-a-roc-curve/

Torgo, L. (2010). SMOTE. Retrieved from https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/SMOTE

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Retrieved from http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients