Models for Qualitative Variables: LOGIT and PROBIT

by:
LI Jingxuan
YE Hongbo
ARIES Allen Jerry

Professor:
Catherine Bruneau

Empirical Application No. 1
submitted
as a requirement
in
Financial Econometrics II

7 February 2023

1 Introduction

Consider a dichotomous random variable Y that can take a value of either 0 or 1(i.e. representing two states of the world) which can be explained by a vector of X variables. A realization $y \in [0, 1]$ can be interpreted as the probability $P(Y = 1)$ of taking one state, and the complement, $P(Y = 0) = 1 - P(Y = 1)$, of taking the other state. In theory, the standard OLS regression cannot be used to predict Y since the expected value from such an OLS can take any values in $\mathbb{R}$, which may not be in $[0, 1]$. A latent variable Y^* can be introduced such that:

$$Y^* = \beta' X + \varepsilon$$

This variable is not observable and a quantitative continuous variable taking its values in $(-\infty, \infty)$. Taking a certain threshold S, the dichotomous Y can be explained by:

$$Y = 1 \text{ if } Y^* > S \text{ and}$$

$$Y = 0 \text{ if } Y^* \leq S$$

Suppose that $Y = 1$, it follows that $\beta' X + \varepsilon > S$. If $var(\varepsilon) = \sigma^2$, the standardized residual will have this relationship with the explanatory variables and the threshold:

$$\frac{\varepsilon}{\sigma} > \frac{S - \beta' X}{\sigma}$$

The model will be PROBIT if $\varepsilon \hookrightarrow \mathcal{N}(0, \sigma)$ and LOGIT if ε follows the logistic distribution whose cummulative density function is:

$$F(x; \mu, \sigma) = \frac{1}{1 + \exp(\frac{x-\mu}{\sigma})}$$

In probability theory, the *odds* of an event is easily computed as the ratio of the probability of the event happening against the probability of it not happening. The higher the probability $P(Y = 1|X = x)$, the higher will be the odds. For the case of Y:

$$odds(x) = \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)}$$

It can be deduced that $\ln(odds(x))$, defined as the logit transform of $P(Y = 1|X = x)$, is just equal to $\beta' X + \varepsilon$. The logit transform is the latent variable for the LOGIT model whose value can be compared against the the threshold S.

The performance of binary classification models can be measured in different ways. First is by looking at the the confusion matrix. Depending on the application, the decision maker can adjust the threshold S to minimize the proportions of the **False Positives** or **False Negatives**.

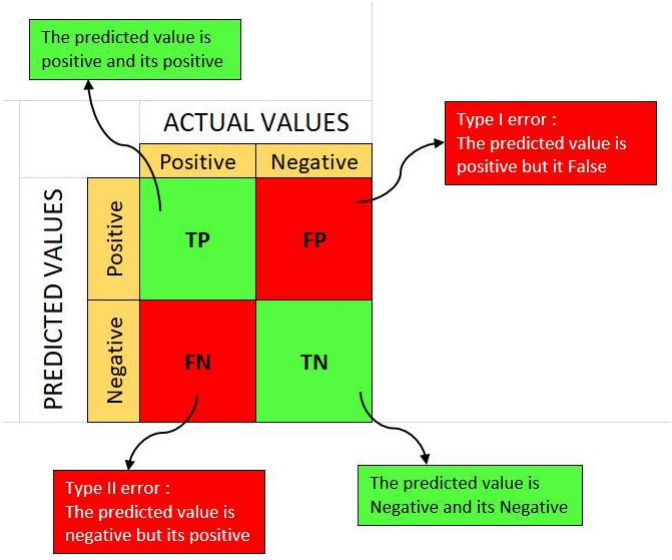

Figure 1: Confusion Matrix of Binary Classification

Another measure is the Receiving Operator Characteristic(ROC) Curve. it is the plot of the True Positive Rate (**Selectivity/ Recall / Probability of Detection**) against the False Positive Rate (**Fall-out/Probability of False Alarm**) as its discrimination threshold is varied. The Area under the ROC curve (AUC) can also be a measure of performance at various thresholds settings. It is the area between the ROC curve and the 45° line, and it represents the degree or measure of separability. The higher the separation, the better the model. Other performance measures include the log-likelihood ratio and the pseudo-R^2.

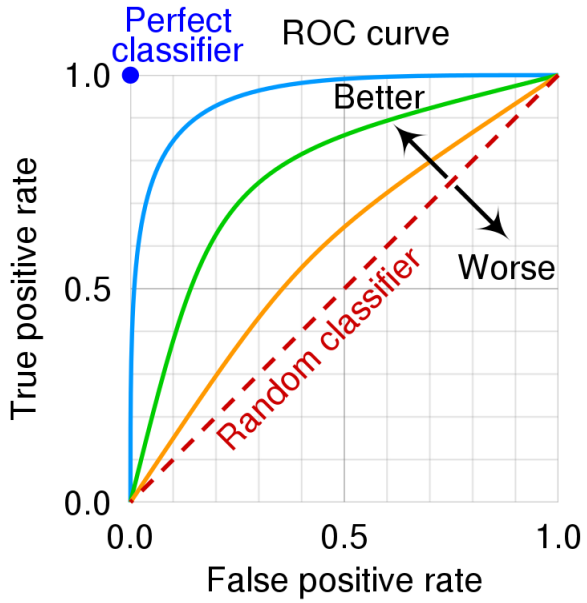

Figure 2: Receiver Operating Characteristic Curve

2 Objectives

This empirical application will apply the LOGIT and PROBIT models on **Credit Card Default Prediction**. Normally, upon application for a credit card, personal information about the applicant is collected by the bank. From this information, credit risk is assessed by the bank using its own credit scoring models. One of these is the default risk assessment where the bank decides whether to approve the credit card application or not if default is predicted based on the information provided by the applicant . In this analysis, the collected personal information will be used as explanatory variables to predict if a certain applicant will undergo default and hence the credit scoring model will be validated.

3 Database and Exploratory Analysis

The dataset was from an *anonymous* bank shared on a public domain and posted on Kaggle, a website for practicing and competitions on Data Science and Machine Learning. The dataset contains 777,715 application information data. The original dataset of the bank categorizes the number of days the client has past due. That is, 0: 1-29 days past due 1: 30-59 days past due 2: 60-89 days overdue 3: 90-119 days overdue 4: 120-149 days overdue 5: Overdue or bad debts, write-offs for more than 150 days C: paid off that month X: No loan for the month. All the clients in category five are considered to have been defaulted and are classiffied as '1' in the the regression models while those who have never defaulted as '0'.

```
#   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 1   ID                   777715 non-null  int64
 2   CODE_GENDER          777715 non-null  object
 3   FLAG_OWN_CAR         777715 non-null  object
 4   FLAG_OWN_REALTY      777715 non-null  object
 5   CNT_CHILDREN         777715 non-null  int64
 6   AMT_INCOME_TOTAL     777715 non-null  float64
 7   NAME_INCOME_TYPE     777715 non-null  object
 8   NAME_EDUCATION_TYPE  777715 non-null  object
 9   NAME_FAMILY_STATUS   777715 non-null  object
 10  NAME_HOUSING_TYPE    777715 non-null  object
 11  DAYS_BIRTH           777715 non-null  int64
 12  DAYS_EMPLOYED        777715 non-null  int64
 13  OCCUPATION_TYPE      537667 non-null  object
 14  CNT_FAM_MEMBERS      777715 non-null  int64
 15  GOT_LOAN             777715 non-null  int64
 16  DEFAULT              777715 non-null  object
```

The variable `ID` is just a client identifier and can be removed from the dataset. Columns with `int64` or `float64` as `Dtype` are continuous while the `object` type are categorical. However, the dataset contains 30% which are blank values for the variable `OCCUPATION_TYPE`. Since it is important variable in most credit scoring models, this variable is retained. Rows which has `'na'` values are also removed. Finally,the dataset has to be trimmed down to 304,354 observations. The categorical variables and its categories are summarized below:

```
Column                          Categories
DEFAULT                  [Y, N]
GOT_LOAN                 [Y, N]
CODE_GENDER              [M, F]
FLAG_OWN_CAR             [Y, N]
FLAG_OWN_REALTY          [Y, N]

NAME_INCOME_TYPE     ['Working' 'Commercial associate' 'State servant'
                        'Student' 'Pensioner']

NAME_EDUCATION_TYPE  ['Secondary / secondary special'
                        'Higher education' 'Incomplete higher'
                        'Lower secondary' 'Academic degree']

NAME_FAMILY_STATUS   ['Married' 'Single / not married'
                        'Civil marriage' 'Separated' 'Widow']

NAME_HOUSING_TYPE    ['House / apartment' 'Rented apartment'
                        'Municipal apartment' 'With parents'
                        'Co-op apartment' 'Office apartment']

OCCUPATION_TYPE      ['Security staff' 'Sales staff' 'Accountants'
                        'Laborers' 'Managers' 'Drivers' 'Core staff'
                        'High skill tech staff' 'Cleaning staff'
                        'Private service staff' 'Cooking staff' 'Low-skill Laborers'
                        'Medicine staff' 'Secretaries' 'Waiters/barmen staff'
                        'HR staff' 'Realty agents' 'IT staff']
```

For the numerical variables, Figure 3 summarizes the histograms for each. It is visible that all variables are discrete except `AMT_INCOME_TOTAL`. `CNT_CHILDREN` corresponds to the number of children in the family, `AMT_INCOME_TOTAL` is the total annual income of the client, `YEARS_BIRTH` is equivalent to age and `CNT_FAM_MEMBERS` refers to the family size. Based on the summary statistics, a typical profile of an applicant is a working 40-year old female, with at least secondary level of education, with one child, an average income of around 195K, employed for more tha 6 years, and at least 2 members in the family.

The correlation matrix (Pearson-r) of the quantitative variables in Table 1 reveals that there is a weak correlation amongst the quantitative variables except for `CNT_CHILDREN` and `CNT_FAM_MEMBERS`. The two variables cannot be used together in the model. Later in the analysis, the two will be tested if which has better explanatory power. For the categorical variables, the usual correlations analysis cannot be applied because of its bi-variate property. To address this, a measure of association called **Jaccard Similarity** will be used amongst the categorical variables. The values are summariwed in Figure 4.

Table 1: Correlation Matrix of the Quantative Variables

	(a)	(b)	(c)	(d)	(e)
CNT_CHILDREN(a)	1.000	-0.021	-0.261	-0.073	0.899
AMT_INCOME_TOTAL(b)	-0.021	1.000	0.063	0.022	-0.025
YEARS_BIRTH(c)	-0.261	0.063	1.000	0.348	-0.209
YEARS_EMPLOYED(d)	-0.073	0.022	0.348	1.000	-0.049
CNT_FAM_MEMBERS(e)	0.899	-0.025	-0.209	-0.049	1.000

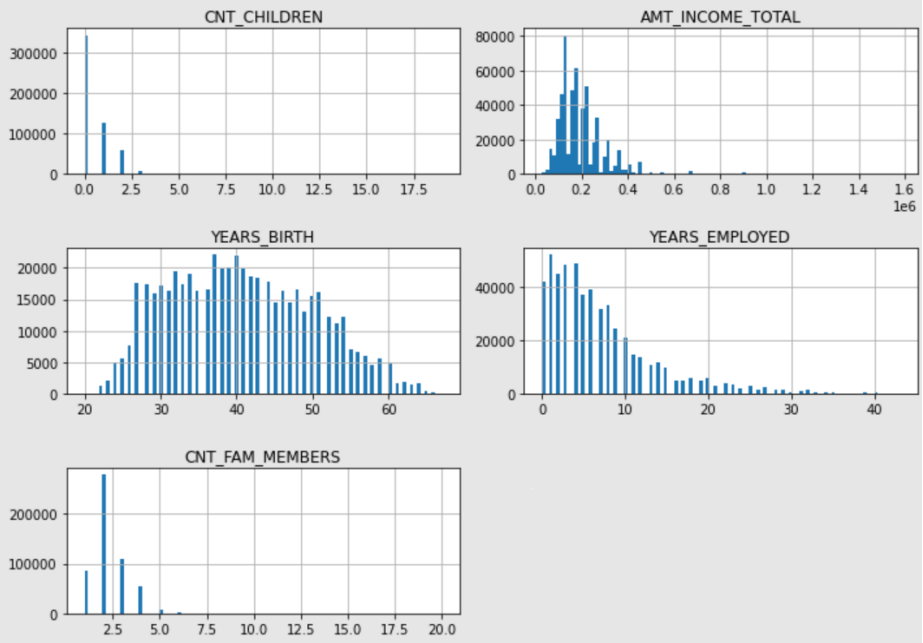

Figure 3: Histograms of the Quantitative Variables

There are a number of insights that can be interpreted from the Jaccard Similarity heatmap. One noticeable observation is the clustering of high similiarity indices between the NAME_INCOME_TYPE and NAME_HOUSING_TYPE variables. There is also high association of for the vairables CODE_GENDER_M and FLAG_OWN_CAR_Y which means, male applicants are very likely to own a car.

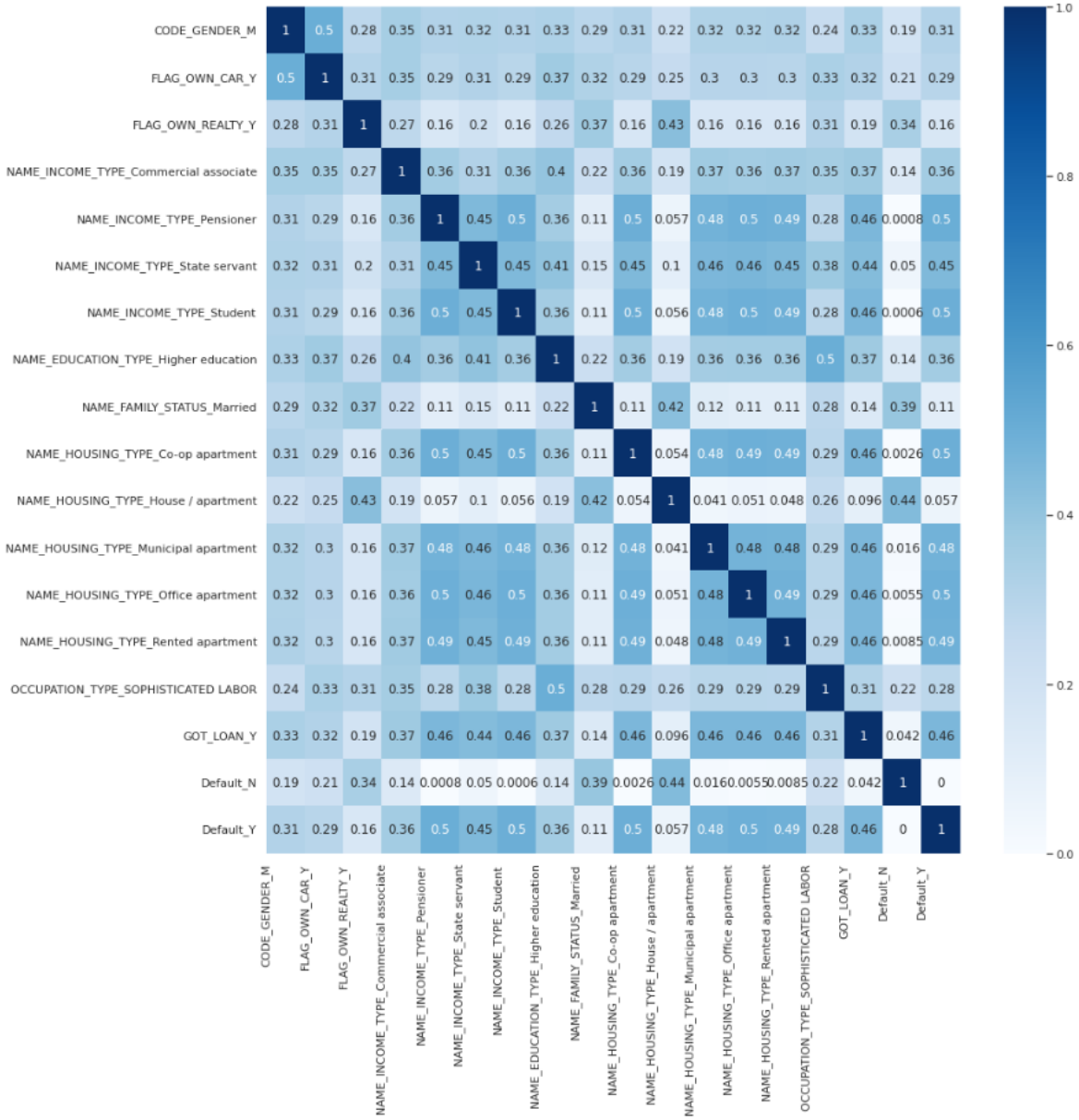

Figure 4: Jaccard Similarity Heatmap

The next step is to ensure that the independent variables do not exhibit *multicollinearity*. The **Variance Inflation Factor(VIF)** is usually utilized for this such that VIF¡5 for all the variables.

```
VIF for variable CODE_GENDER_M:                          1.65633194558255
VIF for variable FLAG_OWN_REALTY_Y:                      2.51316284531116
VIF for variable NAME_EDUCATION_TYPE_Higher education:   1.63581376004705
VIF for variable NAME_FAMILY_STATUS_Married:             3.66800816299751
VIF for variable NAME_HOUSING_TYPE_Co-op apartment:      1.00539130033695
VIF for variable NAME_HOUSING_TYPE_Municipal apartment:  1.03187217818799
VIF for variable NAME_HOUSING_TYPE_Office apartment:     1.01320332620591
VIF for variable NAME_HOUSING_TYPE_Rented apartment:     1.01971253199124
VIF for variable OCCUPATION_TYPE_SOPHISTICATED LABOR:    2.03805811809424
```

```
VIF for variable GOT_LOAN_Y:                    1.08012607280409
VIF for variable AMT_INCOME_TOTAL:              1.09348824590641
VIF for variable YEARS_BIRTH:                   1.24489859526543
VIF for variable YEARS_EMPLOYED:                1.16448724234411
VIF for variable CNT_FAM_MEMBERS:               1.30521967156679

Conclusion:
No multicollinearity
```

4 Model Selection

To address the overfitting problem, the data with L2 regularization wherein a penalty equal to the square of the magnitude of coefficients is added to the loss function. The data is divided into the training and test sets, at 90%-10%, respectively. The training set will be used to fir the data to both LOGIT and PROBIT models while the test set will be used to determine the performance of the models.

4.1 Fitting with LOGIT

The results of the Logistic Regression are as follows:

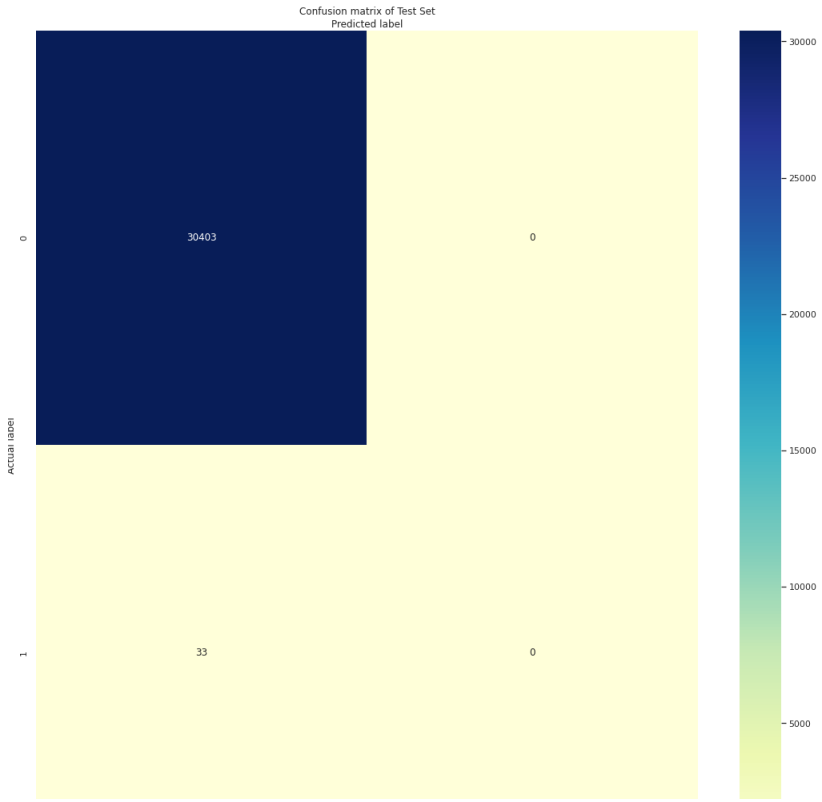

Figure 5: Test Set Confusion Matrix for Logistic Regression

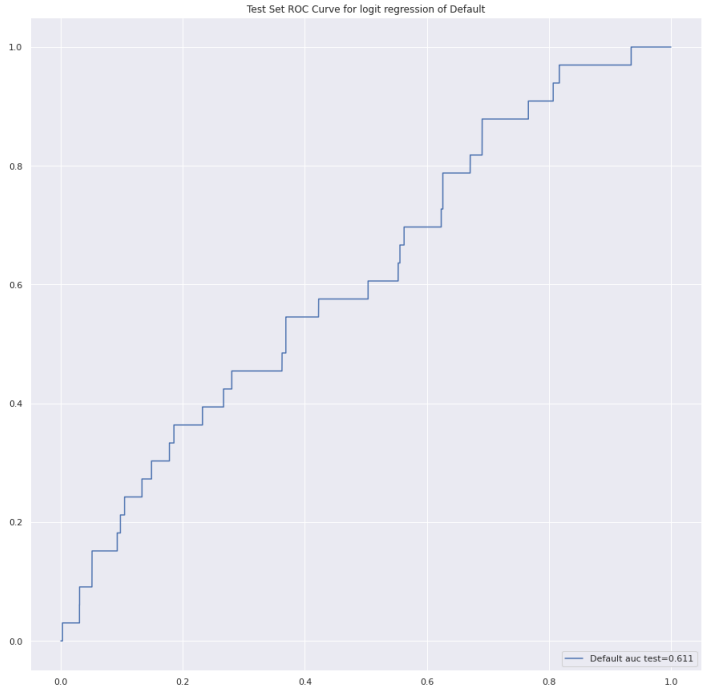

Figure 6: ROC Curve for the LOGIT Model

```
Train RMSE: 0.034179421130305594
Test RMSE: 0.03292783540702198
Train AUC 0.6132150659909794
Test AUC 0.6110596143323177
```

Logit Regression Results

Dep. Variable:	Default_Y	No. Observations:	273918
Model:	Logit	Df Residuals:	273903
Method:	MLE	Df Model:	14
Date:	Wed, 08 Feb 2023	Pseudo R-squ.:	0.01018
Time:	23:57:25	Log-Likelihood:	-2455.3
converged:	False	LL-Null:	-2480.5
Covariance Type:	nonrobust	LLR p-value:	4.982e-06

	coef	std err	z	P>|z|
const	-7.3155	0.203	-35.976	0.000
CODE_GENDER_M	0.2676	0.122	2.193	0.028
FLAG_OWN_REALTY_Y	0.4087	0.132	3.091	0.002
NAME_EDUCATION_TYPE_Higher education	0.3389	0.130	2.608	0.009
NAME_FAMILY_STATUS_Married	0.1009	0.176	0.575	0.566
NAME_HOUSING_TYPE_Co-op apartment	-11.4361	252.627	-0.045	0.964
NAME_HOUSING_TYPE_Municipal apartment	-0.5441	0.453	-1.201	0.230
NAME_HOUSING_TYPE_Office apartment	1.0314	0.341	3.025	0.002
NAME_HOUSING_TYPE_Rented apartment	-0.8737	0.711	-1.229	0.219

OCCUPATION_TYPE_SOPHISTICATED LABOR	-0.0506	0.128	-0.394	0.693
GOT_LOAN_Y	-0.2279	0.226	-1.009	0.313
AMT_INCOME_TOTAL	-0.1464	0.076	-1.932	0.053
YEARS_BIRTH	-0.1509	0.065	-2.314	0.021
YEARS_EMPLOYED	0.1729	0.057	3.016	0.003
CNT_FAM_MEMBERS	0.0084	0.069	0.122	0.903

4.2 Fitting with PROBIT Model

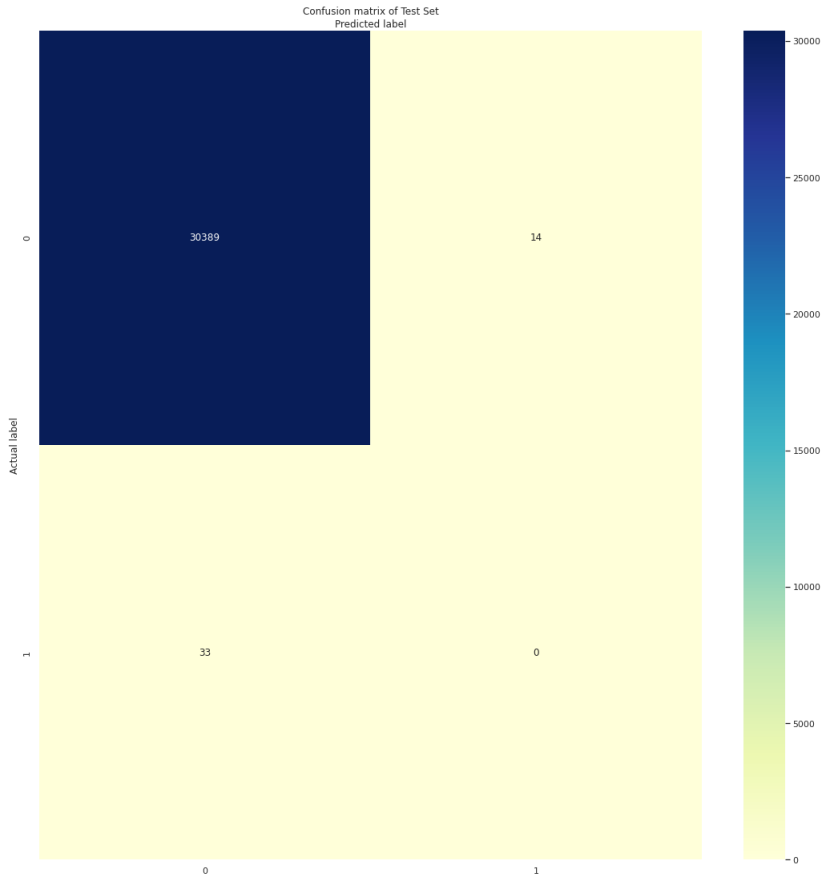

Figure 7: Confusion Matrix for PROBIT Model

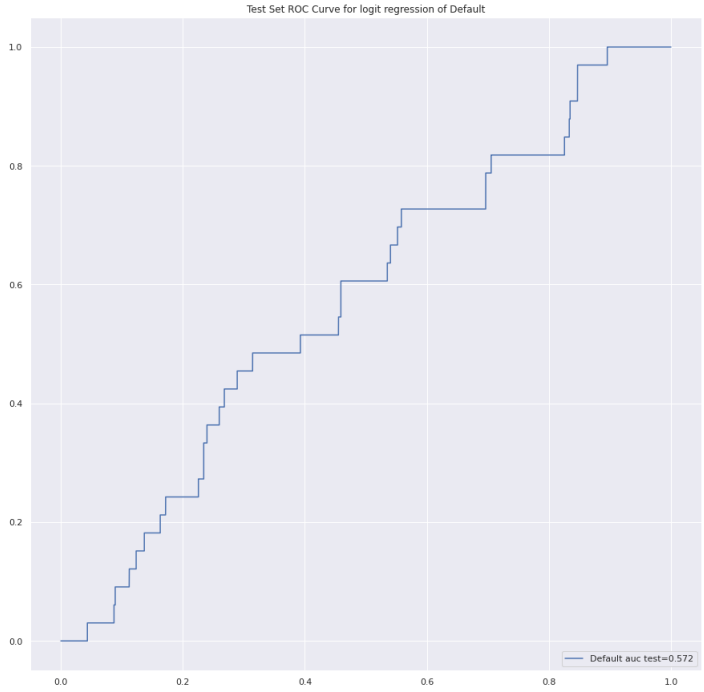

Figure 8: ROC Curve for the PROBIT Model

```
Train RMSE: 0.04120180007419628
Test RMSE: 0.039296614934723084
Train AUC 0.5275571447433827
Test AUC 0.5720209030408682
```

Probit Regression Results

Dep. Variable:	Default_Y	No. Observations:	273918
Model:	Probit	Df Residuals:	273902
Method:	MLE	Df Model:	15
Date:	Thu, 09 Feb 2023	Pseudo R-squ.:	-0.07521
Time:	00:13:21	Log-Likelihood:	-2667.1
converged:	False	LL-Null:	-2480.5
Covariance Type:	nonrobust	LLR p-value:	1.000

	coef	std err	z	P>|z|
CODE_GENDER_M	-0.0710	0.035	-2.016	0.044
FLAG_OWN_REALTY_Y	-0.0048	0.036	-0.135	0.893
NAME_EDUCATION_TYPE_Higher education	0.0607	0.039	1.551	0.121
NAME_FAMILY_STATUS_Married	-3.1848	0.057	-56.337	0.000
NAME_HOUSING_TYPE_Co-op apartment	-4.5688	5854.437	-0.001	0.999
NAME_HOUSING_TYPE_House / apartment	-0.4902	0.046	-10.723	0.000
NAME_HOUSING_TYPE_Municipal apartment	-0.7600	0.142	-5.348	0.000
NAME_HOUSING_TYPE_Office apartment	-0.1561	0.120	-1.298	0.194
NAME_HOUSING_TYPE_Rented apartment	-0.8811	0.230	-3.827	0.000
OCCUPATION_TYPE_SOPHISTICATED LABOR	-0.1398	0.037	-3.762	0.000
GOT_LOAN_Y	-0.1459	0.070	-2.095	0.036

CNT_CHILDREN	-2.4440	0.055	-44.335	0.000
AMT_INCOME_TOTAL	0.0004	0.018	0.023	0.981
YEARS_BIRTH	-0.0217	0.019	-1.112	0.266
YEARS_EMPLOYED	0.0598	0.017	3.506	0.000
CNT_FAM_MEMBERS	2.9682	0.061	48.900	0.000

5 Conclusion

In both the LOGIT and PROBIT models, the following variables are significant in predicting the credit card payment default:

```
 'CODE_GENDER_M', 'FLAG_OWN_REALTY_Y',
'NAME_EDUCATION_TYPE_Higher education',
'NAME_HOUSING_TYPE_Office apartment', 'AMT_INCOME_TOTAL',
'YEARS_BIRTH', 'YEARS_EMPLOYED', 'CNT_FAM_MEMBERS'
```

However, having an AUC score of between 0.5-0.6, both models are poor models for this application. Although, surpassing an AUC score of 0.5 makes the model better than worthless. Particular to this analysis, the LOGIT model seem to be a better predictor than the PROBIT model in terms of all the performance metrics in the test set such as the RMSE and the AUC. More recent techniques for classification using Machine Learning Models are now available for

a better prediction, wherein the rigorous methods of in the regression analysis such as correlations and multi-collinearity tests are not required anymore. Instead, a *post-facto* selection of the variables are done by *feature importance* properties of such models.

6 Python Codes

All python codes are done in Google Colaboratory and can be found in this **link**.