Christopher Dennies

Batch 1 - April 2020

Abstract

**Model Building, Tuning & Evaluating Using Performance metrics. Interpretation of the Optimum Model.**

BuSiness report

NBFC Foreclosure Prediction Notes 2

**Table of Contents**

**I. Model Building & interpretation**

**II. Model Tuning**

**III. Comparison - Optimum Model**

**IV. Business Implications**

**I. Model Building & interpretation**

**Logistics Regression**

* List of Significant Variables used for model building are below :
* The List is arrived Basis Domain Knowledge, Correlation Plots, Variation inflation factor and finally on the P-values.
* Stas Model Library was used to build a logistic model.
* The P value was obtained by the summary and insignificant predictors were removed one by one to get 13 predictors which are significant.
* One Variable NET\_LTV , though the P value is greater, retained as per domain understanding.
* Below are the significant predictors which can predict Loan Default and output from Stats model.
* **The Coeffecients of “NET\_RECEIVABLE” is positive to indicate the predictors are significant to predict the default of a loan.**

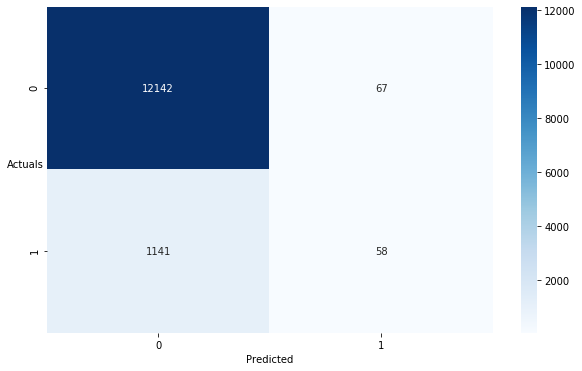
**Table 1.1**

|  |  |  |  |
| --- | --- | --- | --- |
| Logit Regression Results | | | |
| **Dep. Variable:** | FORECLOSURE | **No. Observations:** | 13408 |
| **Model:** | Logit | **Df Residuals:** | 13394 |
| **Method:** | MLE | **Df Model:** | 13 |
| **Date:** | Sun, 25 Apr 2021 | **Pseudo R-squ.:** | 0.1515 |
| **Time:** | 19:33:30 | **Log-Likelihood:** | -3426.5 |
| **converged:** | True | **LL-Null:** | -4038.5 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 1.172e-253 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | -0.2852 | 0.205 | -1.390 | 0.164 | -0.687 | 0.117 |
| **BALANCE\_TENURE** | -0.0039 | 0.001 | -5.152 | 0.000 | -0.005 | -0.002 |
| **EXCESS\_AVAILABLE** | 6.084e-05 | 1.04e-05 | 5.828 | 0.000 | 4.04e-05 | 8.13e-05 |
| **FOIR** | -0.8684 | 0.149 | -5.820 | 0.000 | -1.161 | -0.576 |
| **NET\_RECEIVABLE** | 0.0030 | 0.002 | 1.791 | 0.073 | -0.000 | 0.006 |
| **OUTSTANDING\_PRINCIPAL** | -1.167e-07 | 1.93e-08 | -6.047 | 0.000 | -1.55e-07 | -7.89e-08 |
| **PAID\_INTEREST** | 1.54e-06 | 9.8e-08 | 15.718 | 0.000 | 1.35e-06 | 1.73e-06 |
| **PAID\_PRINCIPAL** | -2.954e-06 | 2.92e-07 | -10.133 | 0.000 | -3.53e-06 | -2.38e-06 |
| **PRE\_EMI\_DUEAMT** | 1.121e-05 | 1.5e-06 | 7.462 | 0.000 | 8.26e-06 | 1.42e-05 |
| **NUM\_EMI\_CHANGES\_RANGE\_CAT** | 0.1303 | 0.050 | 2.596 | 0.009 | 0.032 | 0.229 |
| **PRODUCT** | -0.9828 | 0.045 | -22.011 | 0.000 | -1.070 | -0.895 |
| **LOAN\_AMT** | -2.548e-08 | 5.89e-09 | -4.327 | 0.000 | -3.7e-08 | -1.39e-08 |
| **NET\_LTV** | 0.0023 | 0.002 | 1.423 | 0.155 | -0.001 | 0.006 |
| **CITY\_NEW** | -0.0177 | 0.008 | -2.202 | 0.028 | -0.033 | -0.002 |

# **LOGISTIC REGRESSION - WITH DEFAULT CUTOFF 0.5**

**Figure 1.1**

****

precision recall f1-score support

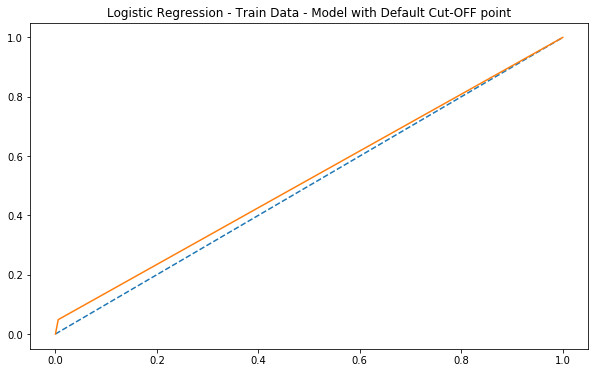
0 0.914 0.995 0.953 12209

1 0.464 0.048 0.088 1199

accuracy 0.910 13408

macro avg 0.689 0.521 0.520 13408

weighted avg 0.874 0.910 0.875 13408

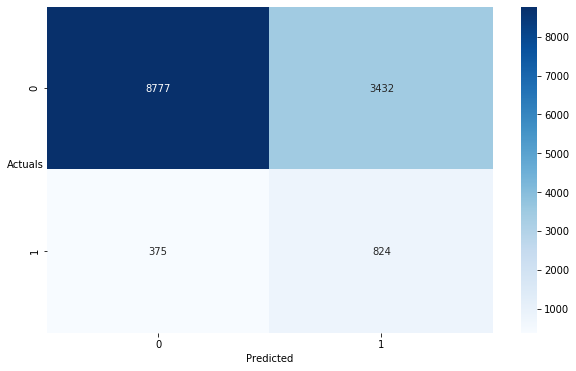


**Inference :** Recall at 4.8 percent and precision at 46.4 percent which only 4.8% defaults predicted correctly with a default cutoff 0.5. But Specificity 99 percent indicates that the most loan accounts are showing as non default.

AUC – 52

# **LOGISTIC REGRESSION – TRAIN DATA - WITH OPTIMUM CUTOFF 0.09**

**Figure 1.2**

****

precision recall f1-score support

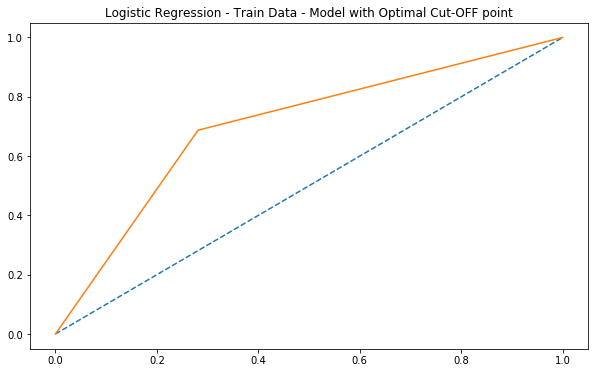
0 0.959 0.719 0.822 12209

1 0.194 0.687 0.302 1199

accuracy 0.716 13408

macro avg 0.576 0.703 0.562 13408

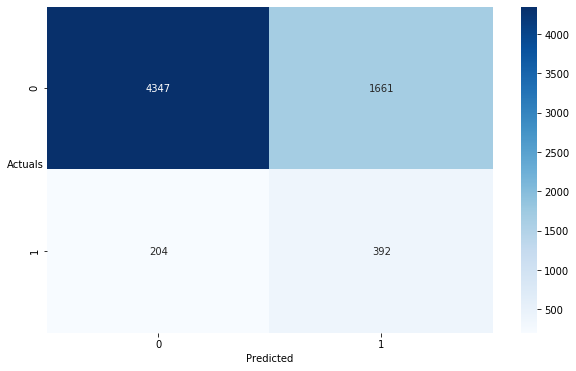
weighted avg 0.891 0.716 0.775 13408



**Inference :** Recall at 68 percent and precision at 19.4 percent is lowest, with 68% defaults is predicted correctly with a optimum cutoff 0.09. Specificity 71.9 percent. AUC – 70

# **LOGISTIC REGRESSION – TEST DATA - WITH OPTIMUM CUTOFF 0.09**

**Figure 1.3**

****

precision recall f1-score support

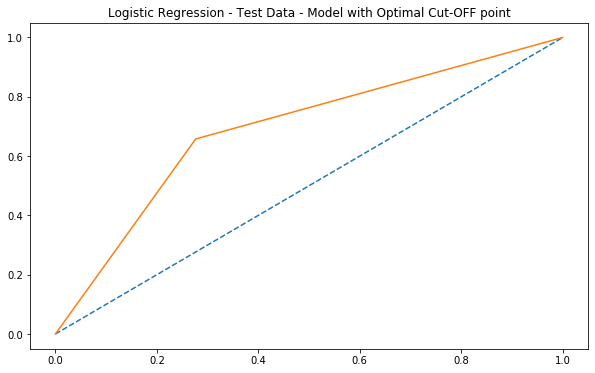
0 0.955 0.724 0.823 6008

1 0.191 0.658 0.296 596

accuracy 0.718 6604

macro avg 0.573 0.691 0.560 6604

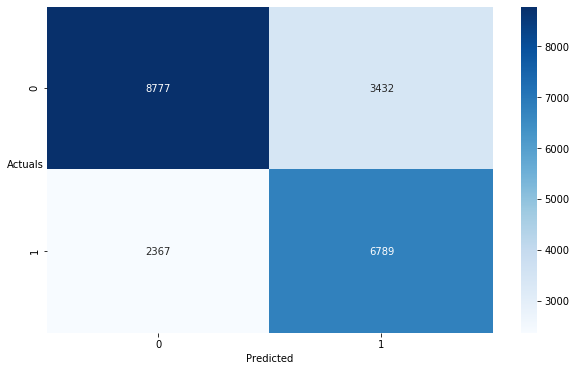
weighted avg 0.886 0.718 0.776 6604



**Inference :** Test setRecall reduced to 65.8 percent and precision at 19.1 percent is lowest, with 65.8% defaults is predicted correctly with a optimum cutoff 0.09. Specificity 72.4 percent. AUC – 69

# **LOGISTIC REGRESSION – SMOTE DATA – TRAIN DATASET – CUTOFF – 0.09**

**Figure 1.4**

****

precision recall f1-score support

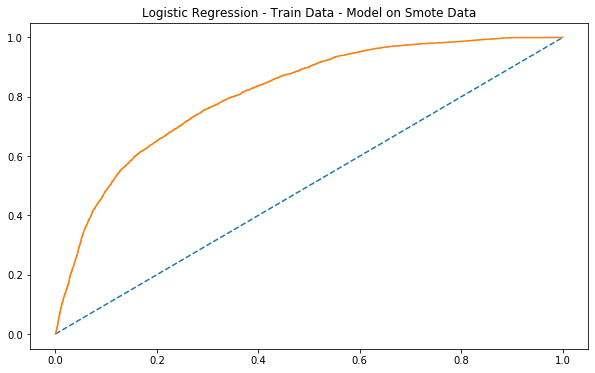
0 0.788 0.719 0.752 12209

1 0.664 0.741 0.701 9156

accuracy 0.729 21365

macro avg 0.726 0.730 0.726 21365

weighted avg 0.735 0.729 0.730 21365



**Inference :** Recall at 74 percent and precision at 66 percent which 74% of defaults predicted correctly with a optimum cutoff 0.09 is a very good model when smote is applied. Recall is at maximum compared to past 3 summary of logistic regression. Both Recall and precision are high with a regularized data.

AUC- 81.

# **LDA - LINEAR DISCRMINANT ANALYSIS**

# **LDA - LINEAR DISCRMINANT ANALYSIS**

# **With default values for both train and test datasets.**

**Table 1.2**

precision recall f1-score support

0 0.92 0.98 0.95 12209

1 0.39 0.12 0.18 1199

accuracy 0.90 13408

macro avg 0.66 0.55 0.57 13408

weighted avg 0.87 0.90 0.88 13408

**Table 1.3**

precision recall f1-score support

0 0.92 0.98 0.95 6008

1 0.37 0.10 0.16 596

accuracy 0.90 6604

macro avg 0.64 0.54 0.55 6604

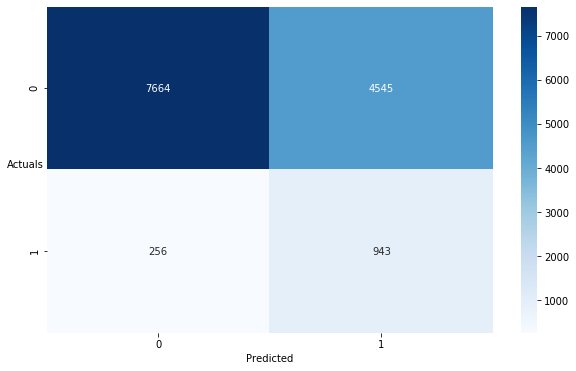
weighted avg 0.87 0.90 0.88 6604

**Inference :**

Recall for both train and test data for LDA model with default values show poor recall scores of 12 & 10 percent and having precision being lowest. Prediction of loan defaults correctly at 10 percent levels is very poor metrics.

# **LDA – TRAIN DATASET – CUTOFF – 0.06**

**Figure 1.5**

****

precision recall f1-score support

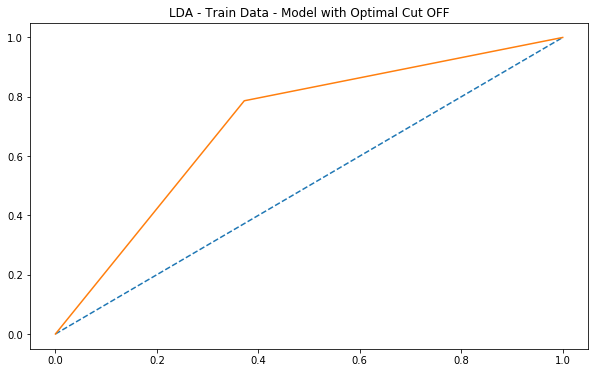
0 0.968 0.628 0.761 12209

1 0.172 0.786 0.282 1199

accuracy 0.642 13408

macro avg 0.570 0.707 0.522 13408

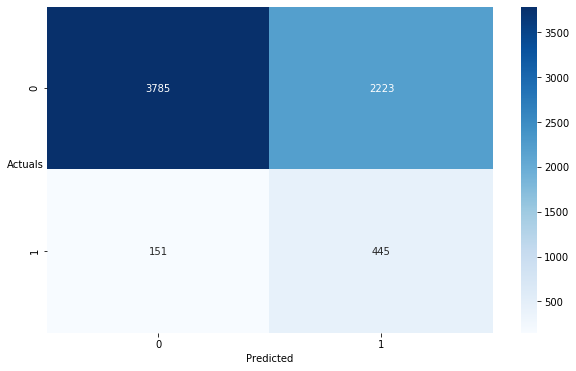
weighted avg 0.897 0.642 0.719 13408



**Inference :** Recall at 78 percent and precision at 17 percent which 78% of defaults predicted correctly with a optimum cutoff 0.06 is a very good but precision being low. AUC- 70.

# **LDA – TEST DATASET – CUTOFF – 0.06**

**Figure 1.6**

****

precision recall f1-score support

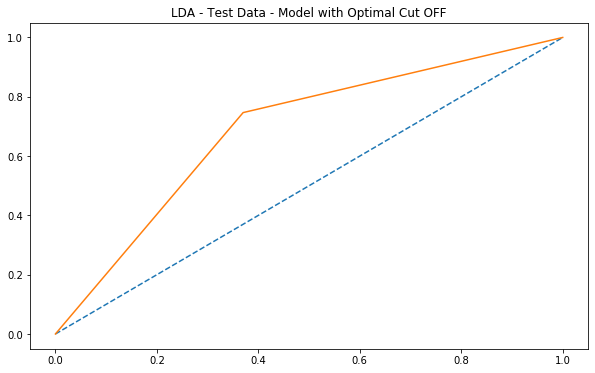
0 0.962 0.630 0.761 6008

1 0.167 0.747 0.273 596

accuracy 0.641 6604

macro avg 0.564 0.688 0.517 6604

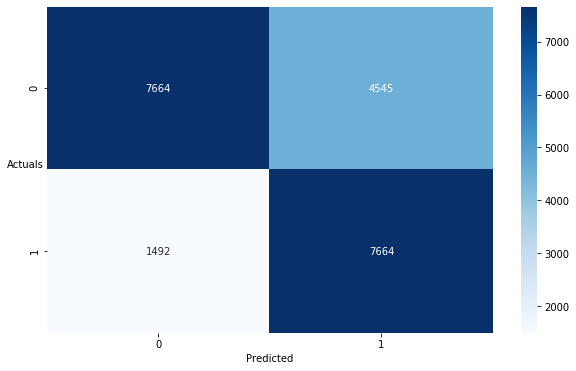
weighted avg 0.890 0.641 0.717 6604



**Inference :** Recall reduced to 74 percent on test data and precision at 16 percent which 74% of defaults predicted correctly with a optimum cutoff 0.06 is a very good but precision being low. AUC- 68.

# **LDA – SMOTE DATASET – CUTOFF – 0.06**

**Figure 1.7**

****

precision recall f1-score support

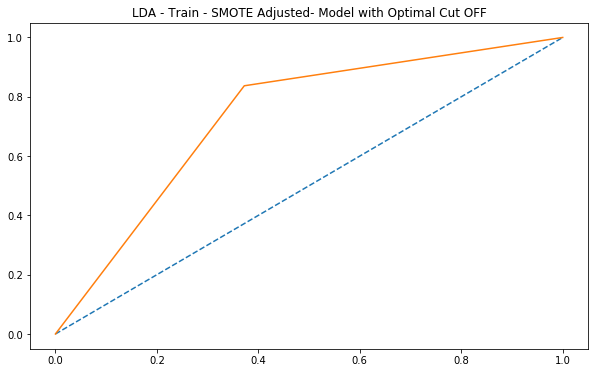
0 0.837 0.628 0.717 12209

1 0.628 0.837 0.717 9156

accuracy 0.717 21365

macro avg 0.732 0.732 0.717 21365

weighted avg 0.747 0.717 0.717 21365



**Inference :** Recall at 83 percent and precision at 62 percent which 83% of loan defaults predicted correctly with a optimum cutoff 0.06 is a very good model when smote is applied. Recall is at maximum compared to past 3 summary of LDA. Both Recall and precision are high with a regularized data.

AUC- 73.

**II. Model Tuning**

* For the purpose of model tuning, ensemble modelling Random forest was used.

# **RANDOM FOREST MODEL**

* From sklearn , imported grid search & random forest classifier ,used grid search to get the ideal features.
* Fit it on to the Train dataset.
* Got the best parameters.
* Predicted on both train, test and train with smote dataset.
* Computed confusion matrix, Summary and ROC curve AUC values.

**Train**

* Recall – 39
* Precision - 85
* Accuracy – 94
* AUC – 69

**Test**

* Recall – 31
* Precision - 77
* Accuracy – 93
* AUC - 65

**Train\_Smote**

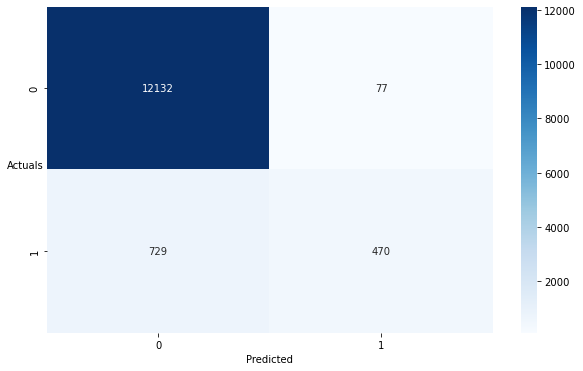
* Recall – 91
* Precision – 93
* Accuracy – 93
* AUC – 93

**Inference :**

* Both Train and test results showed low recall scores.
* By applying Smote, The Recall, Precision, and AUC has improved to a greater extend shows that the model with regularizing the data is more robust.
* Recall at 91 percent and precision at 93 percent which 91% the loan defaults are predicted correctly with a optimum grid features is a very good model when smote is applied. Recall is at maximum compared to Train and Test datasets. Both Recall and precision are high with a regularized data.
* SMOTE was used for tuning the model. Random forest achieved the maximum accuracy compared to all the models.

# **RANDOM FOREST – TRAIN DATASET**

**Figure 1.8**

****

precision recall f1-score support

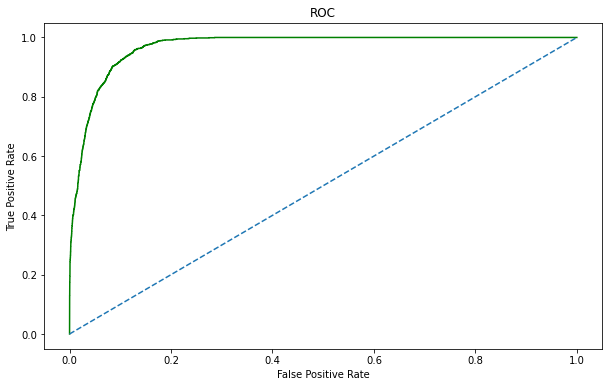
0 0.943 0.994 0.968 12209

1 0.859 0.392 0.538 1199

accuracy 0.940 13408

macro avg 0.901 0.693 0.753 13408

weighted avg 0.936 0.940 0.929 13408

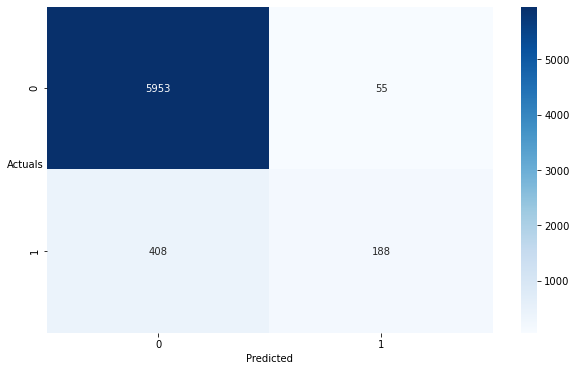


**Inference :** Recall at 39 percent and precision at 85 percent which 39% of loan defaults predicted correctly which is very low.

AUC- 69.

# **RANDOM FOREST – TEST DATASET**

**Figure 1.9**

****

precision recall f1-score support

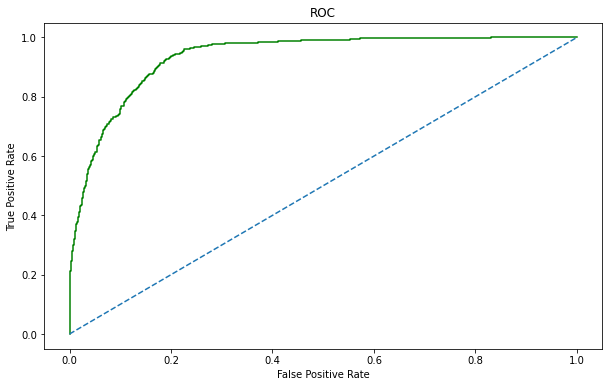
0 0.936 0.991 0.963 6008

1 0.774 0.315 0.448 596

accuracy 0.930 6604

macro avg 0.855 0.653 0.705 6604

weighted avg 0.921 0.930 0.916 6604

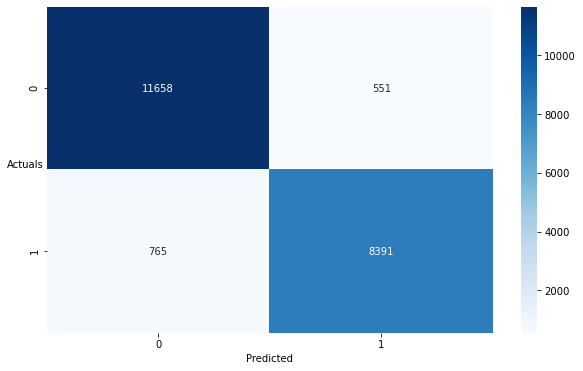


**Inference :** Recall reduced to 31 percent and precision at 77 percent which 31% of loan defaults predicted correctly which is very low.

AUC- 65.

# **RANDOM FOREST – SMOTE DATASET**

**Figure 2.0**

****

precision recall f1-score support

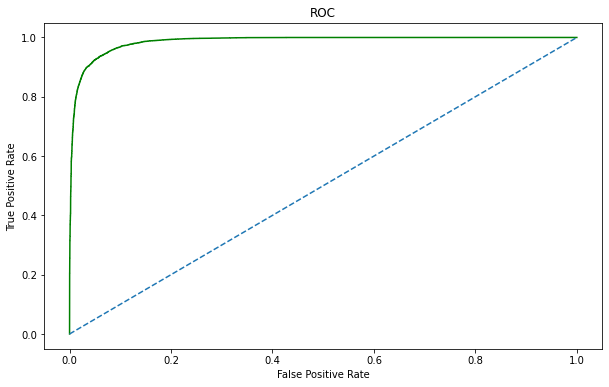
0 0.938 0.955 0.947 12209

1 0.938 0.916 0.927 9156

accuracy 0.938 21365

macro avg 0.938 0.936 0.937 21365

weighted avg 0.938 0.938 0.938 21365



**Inference :** Recall drastically increased to 91 percent and precision at 93 percent which 91% of loan defaults predicted correctly with a optimum best parameters is a very good model when smote is applied. Recall is at maximum compared to all models. Both Recall and precision are high with a regularized data. AUC- 93.

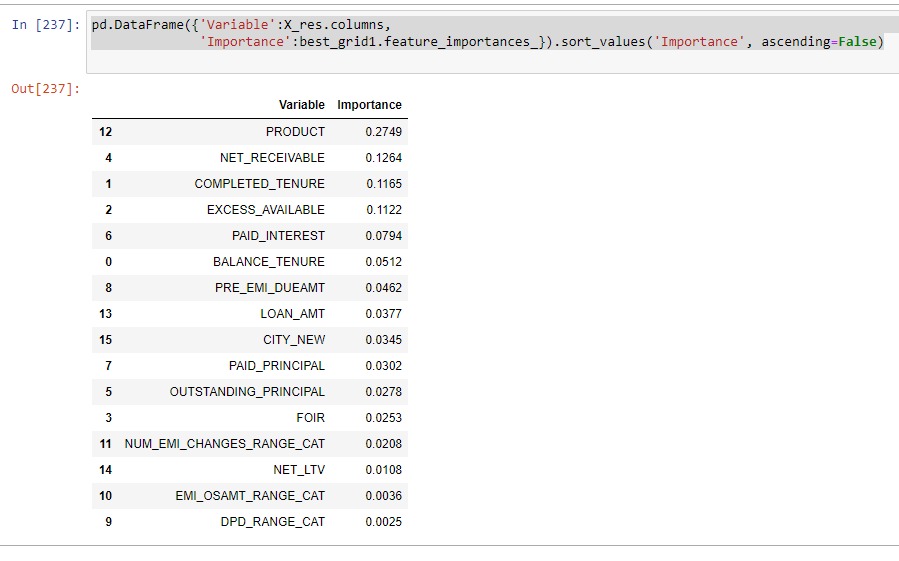
**III. Comparison - Optimum Model**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Models | Dataset | Precision | Recall | F1-Score | Accuracy | AUC |
| Logistic Regression with Default Cut-Off | Train | 0.464 | 0.048 | 0.088 | 0.91 | 0.521 |
| Logistic Regression with Optimal Cut-Off | Train | 0.194 | 0.687 | 0.302 | 0.716 | 0.703 |
| Logistic Regression with Optimal Cut-Off | Test | 0.191 | 0.658 | 0.296 | 0.718 | 0.691 |
| Logistic Regression on SMOTE | SMOTE Train | 0.664 | 0.741 | 0.701 | 0.724 | 0.812 |
| Linear Discriminant Analysis - LDA | Train | 0.39 | 0.12 | 0.18 | 0.9 | 0.785 |
| Linear Discriminant Analysis - LDA | Test | 0.37 | 0.1 | 0.16 | 0.9 | 0.772 |
| Linear Discriminant Analysis with Optimal Cut-OFF | Train | 0.172 | 0.786 | 0.282 | 0.642 | 0.707 |
| Linear Discriminant Analysis with Optimal Cut-OFF | Test | 0.167 | 0.747 | 0.273 | 0.641 | 0.688 |
| Linear Discriminant Analysis - LDA on SMOTE | SMOTE Train | 0.628 | 0.837 | 0.717 | 0.717 | 0.735 |
| Random Forest Model | Train | 0.843 | 0.38 | 0.524 | 0.938 | 0.686 |
| Random Forest Model | Test | 0.773 | 0.309 | 0.441 | 0.929 | 0.649 |
| Random Forest Model on SMOTE | SMOTE Train | 0.939 | 0.92 | 0.929 | 0.94 | 0.937 |

* **SMOTE was used to balance the data and thereby it helped to fine tune the model. By fine Tuning, Random forest model achieved the maximum accuracy compared to all the models.**
* **Random forest is an optimum model but it’s a black box model were no insights on the variables are achieved. Only magnitude of the variables is achieved.**

**IV. Business Implications**

* **Random forest is an optimum model but it’s a black box model were no insights on the variables are achieved. Only magnitude of the variables is achieved.**



* **For Business implications – Logistic Model is preferred, as it give enormous information on the variables.**

|  |  |  |
| --- | --- | --- |
| VARIABLES | COEFFICENT | Exp(Coeff) |
| NUM\_EMI\_CHANGES\_RANGE\_CAT | 0.130300000000 | 1.139170083 |
| NET\_RECEIVABLE | 0.003000000000 | 1.003004505 |
| NET\_LTV | 0.002300000000 | 1.002302647 |
| EXCESS\_AVAILABLE | 0.000060840000 | 1.000060842 |
| PRE\_EMI\_DUEAMT | 0.000011210000 | 1.00001121 |
| PAID\_INTEREST | 0.000001540000 | 1.00000154 |
| LOAN\_AMT | -0.000000025480 | 0.999999975 |
| OUTSTANDING\_PRINCIPAL | -0.000000116700 | 0.999999883 |
| PAID\_PRINCIPAL | -0.000002954000 | 0.999997046 |
| BALANCE\_TENURE | -0.003900000000 | 0.996107595 |
| CITY\_NEW | -0.017700000000 | 0.982455725 |
| Intercept | -0.285200000000 | 0.751863867 |
| FOIR | -0.868400000000 | 0.419622408 |
| PRODUCT | -0.982800000000 | 0.374261698 |

* **For every unit change in EMI – we observe 113% chance of customer defaulting the loan than not defaulting. Likewise the other variables also NET\_LTV, FOIR, Etc tend to predict well the default status of the customer.**