LOAN DEFAULT PREDICTION

HOME EQUITY

By: < Sheidu Omuya Yusuf, Albert Osas & Jerry Tsiba - Aug'23 Batch >

Background

A major proportion of retail bank profit comes from interests in the form of home loans. These loans are borrowed by regular income/high-earning customers. Banks are most fearful of defaulters, as bad loans (NPA) usually eat up a major chunk of their profits. Therefore, it is important for banks to be judicious while approving loans for their customer base. The approval process for the loans is multifaceted. Through this process, the bank tries to check the creditworthiness of the applicant on the basis of a manual study of various aspects of the application. The entire process is not only effort-intensive but also prone to wrong judgment/approval owing to human error and biases. There have been attempts by many banks to automate this process by using heuristics. But with the advent of data science and machine learning, the focus has shifted to building machines that can learn this approval process and make it free of biases and more efficient. At the same time, one important thing to keep in mind is to make sure that the machine does not learn the biases that previously crept in because of the human approval process.

Problem Statement

A bank's consumer credit department aims to simplify the decision-making process for home equity lines of credit to be accepted. To do this, they will adopt the Equal Credit Opportunity Act's guidelines to establish an empirically derived and statistically sound model for credit scoring. The model will be based on the data obtained via the existing loan underwriting process from recent applicants who have been given credit. The model will be built from predictive modeling techniques, but the model created must be interpretable enough to provide a justification for any adverse behavior (rejections).

Objective

Build a classification model to predict clients who are likely to default on their loan and give recommendations to the bank on the important features to consider while approving a loan

Project Objective

To review many attempts by the bank that have fail including the heuristics
To replace the manual study of various aspects of creditworthiness of the applicants based on manual study of various aspect of the applications.
To find solution to the effort-intensive which could be prone to wrong judgments in the approval owing to human errors and biases.
Erase the most fearful of defaulters syndrome which eat up the chunk of bank profits. Helps them to avoid granting loans that have a high likelihood of defaulting.
One of the major points in the objectives of the project is to explore the use of machine leaning algorithms to predict loans default and improve the accuracy of default predictions.
Loan default prediction is a common problem in the financial industry, as it can help lenders or banks identify borrowers who are likely to default on their loans.
The Model created must be interpretable enough to provide a justification for any adverse behaviours (rejections).

Data Description

Column Name	Description
BAD	1 = Client defaulted on loan, 0 = loan repaid
LOAN	Amount of loan approved
MORTDUE	Amount due on the existing mortgage
VALUE	Current value of the property
REASON	Reason for the loan request (HomeImp = home improvement, DebtCon= debt consolidation which means taking out a new loan to pay off other liabilities and consumer debts)
JOB	The type of job that loan applicant has such as manager, self, etc.
YOJ	Years at present job
DEROG	Number of major derogatory reports (which indicates serious delinquency or late payments).
DELINQ	Number of delinquent credit lines (a line of credit becomes delinquent when a borrower does not make the minimum required payments 30 to 60 days past the day on which the payments were due)
CLAGE	Age of the oldest credit line in months
NINQ	Number of recent credit inquiries
CLNO	Number of existing credit lines
DEBTINC	Debt-to-income ratio (all monthly debt payments divided by gross monthly income. This number is one of the ways lenders measure a borrower's ability to manage the monthly payments to repay the money they plan to borrow)

Checking Missing Value

```
DEBTINC     1267
DEROG        708
DELINQ       580
MORTDUE      518
YOJ          515
NINQ         510
CLAGE        308
JOB          279
REASON       252
CLNO         222
VALUE        112
BAD            0
LOAN           0
dtype: int64
```

Number of Unique Values

```
BAD         2
LOAN      540
MORTDUE  5053
VALUE    5381
REASON      2
JOB         6
YOJ        99
DEROG      11
DELINQ     14
CLAGE    5314
NINQ       16
CLNO       62
DEBTINC  4693
dtype: int64
```

Data Description Cont'd

Columns	count	mean	std	min	25%	50%	75%	max
LOAN	5960	17983.17	9081.74	1100	11100	16300	23300	41600
MORTDUE	5960	70554.33	34547.64	2063	48139	65019	88200.25	149008.45
VALUE	5960	98510.55	45002.15	8000	66489.5	89235.5	119004.75	202897.4
YOJ	5960	8.59	6.83	0	3	7	12	25
DEROG	5960	0.12	0.33	0	0	0	0	1
DELINQ	5960	0.4	0.8	0	0	0	0	2
CLAGE	5960	177.43	76.15	0	117.37	173.47	227.14	385.5
NINQ	5960	1.06	1.23	0	0	1	2	5
CLNO	5960	20.9	9.04	0	15	20	26	42
DEBTINC	5960	34	5.74	20.01	30.76	34.82	37.95	48.57
BAD	5960	0.2	0.4	0	0	0	0	1

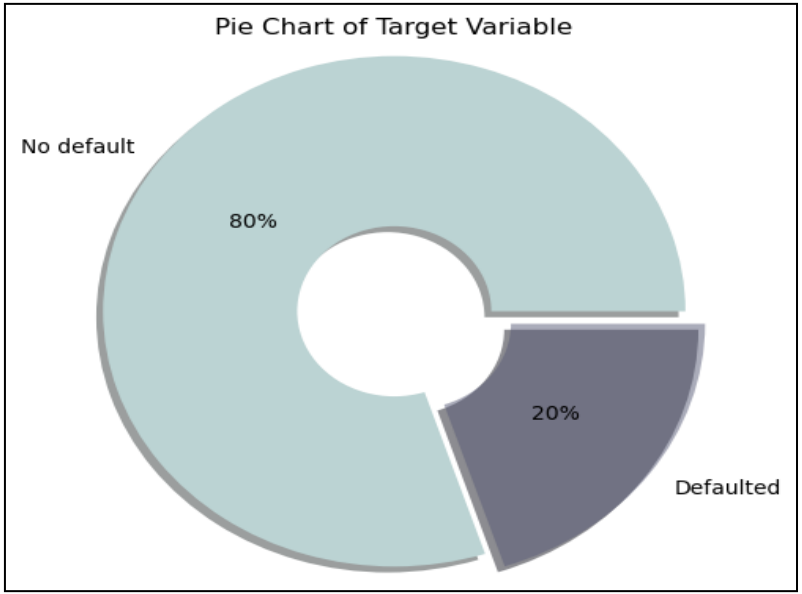
Missing data	
There are missing data	
Observations	**Features**
5960	13

Observations / Findings

- Average loan amount approved is ~ 18608 while the maximum Loan amount approved is 89900.
- The Average the due on existing Mortgage is ~ 737609
- The average Current value of property stands at 101776, while the maximum value is 855909
- Average debt-to-income ratio stands at 33.77 , which is well within favorable ratio.
- Average number of existing credit line stand at 21

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis – Univariate Analysis



Target Variable Analysis

Observations / Findings

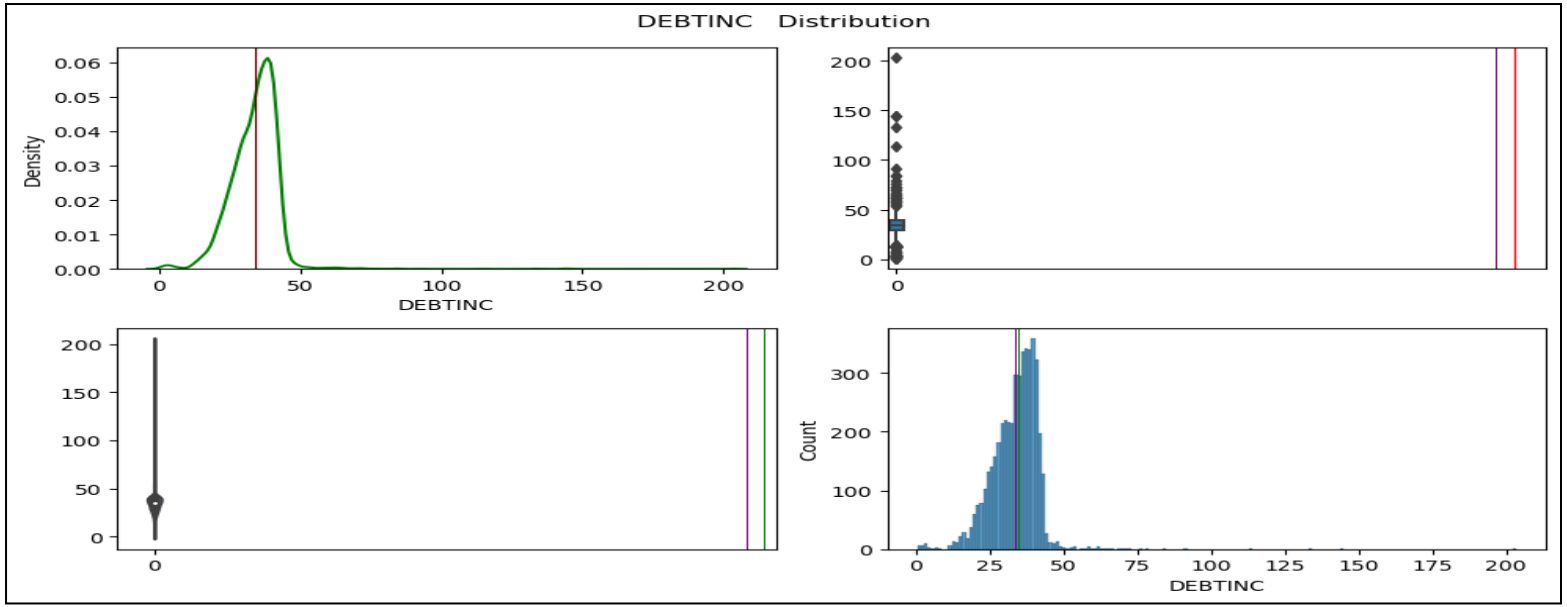
- Out of the 5960 observation 1189 loan applicant have defaulted on their obligation while 4771 repaid
- There is 20% default rate

Statistical summary of the dataset

	count	mean	std	min	25%	50%	75%	max
BAD	5960.0	0.20	0.40	0.00	0.00	0.00	0.00	1.00
LOAN	5960.0	18607.97	11207.48	1100.00	11100.00	16300.00	23300.00	89900.00
MORTDUE	5442.0	73760.82	44457.61	2063.00	46276.00	65019.00	91488.00	399550.00
VALUE	5848.0	101776.05	57385.78	8000.00	66075.50	89235.50	119824.25	855909.00
YOJ	5445.0	8.92	7.57	0.00	3.00	7.00	13.00	41.00
DEROG	5252.0	0.25	0.85	0.00	0.00	0.00	0.00	10.00
DELINQ	5380.0	0.45	1.13	0.00	0.00	0.00	0.00	15.00
CLAGE	5652.0	179.77	85.81	0.00	115.12	173.47	231.56	1168.23
NINQ	5450.0	1.19	1.73	0.00	0.00	1.00	2.00	17.00
CLNO	5738.0	21.30	10.14	0.00	15.00	20.00	26.00	71.00
DEBTINC	4693.0	33.78	8.60	0.52	29.14	34.82	39.00	203.31

- Average loan amount approved is ~ 18608 while the maximum Loan amount approved is 89900.
- The Average due on existing Mortgage is ~ 737609
- The average Current value of property stands at 101776, while the maximum value is 855909
- Average debt-to-income ratio stands at 33.77 , which is well within favourable ratio.
- Average number of existing credit line stand at 21.

Exploratory Data Analysis – Univariate Analysis

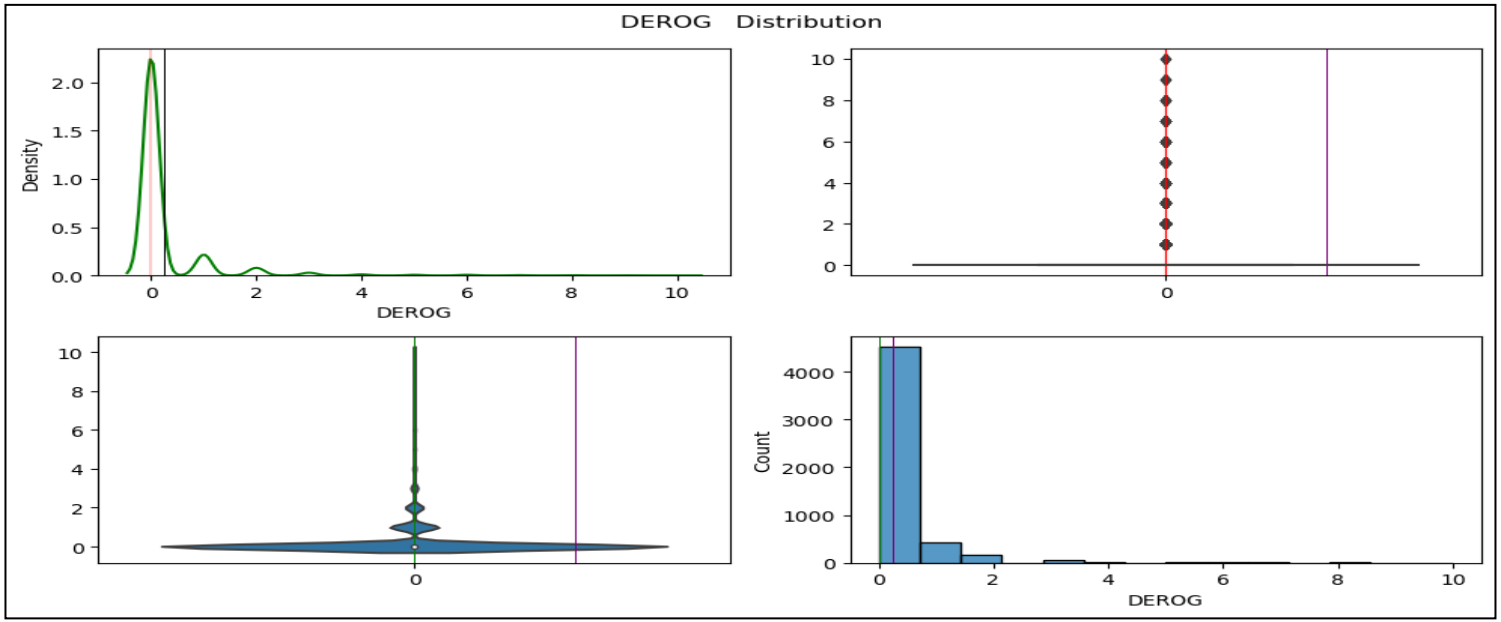


DEBTINC Distribution

Observations / Findings

- Debt –to – income rate is not normally distributed. It is skewed to the right . And has a lot of outliers

Exploratory Data Analysis – Univariate Analysis

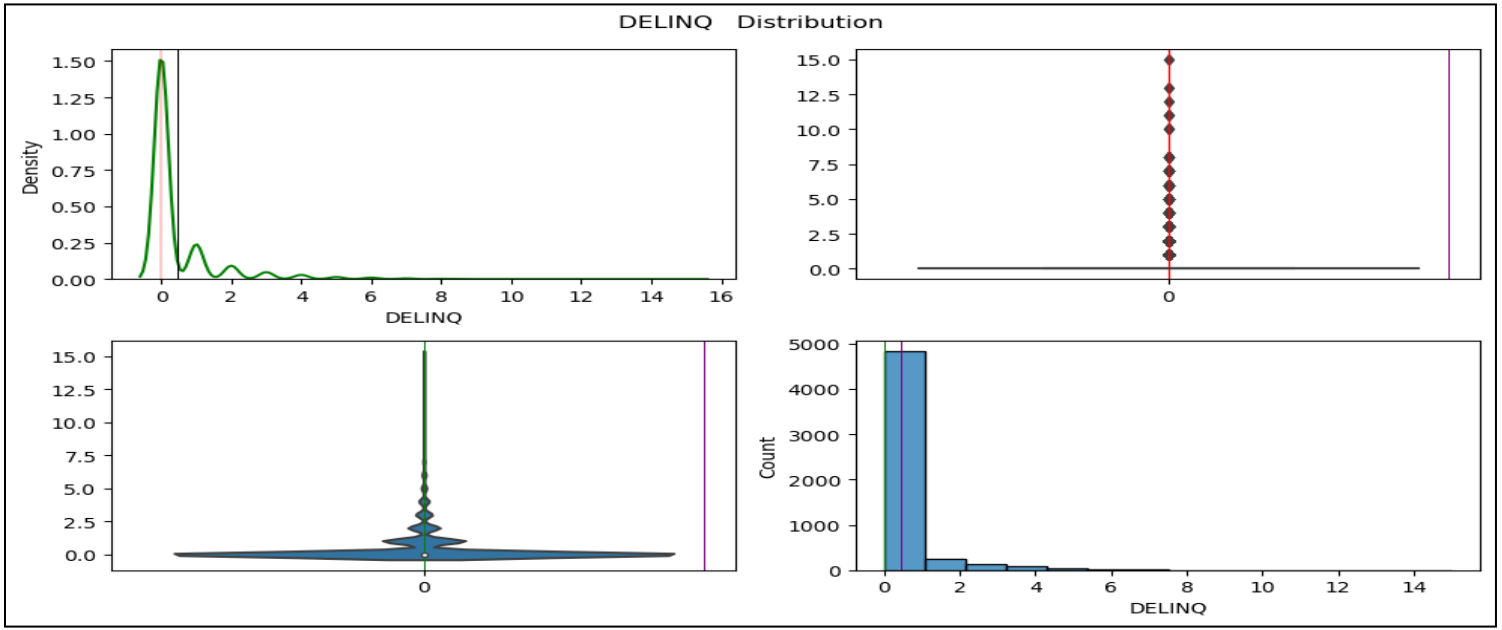


DEROG Distribution

Observations / Findings

- Debt –to – income rate is not normally distributed. It is skewed to the right. It has noticeable outliers

Exploratory Data Analysis – Univariate Analysis

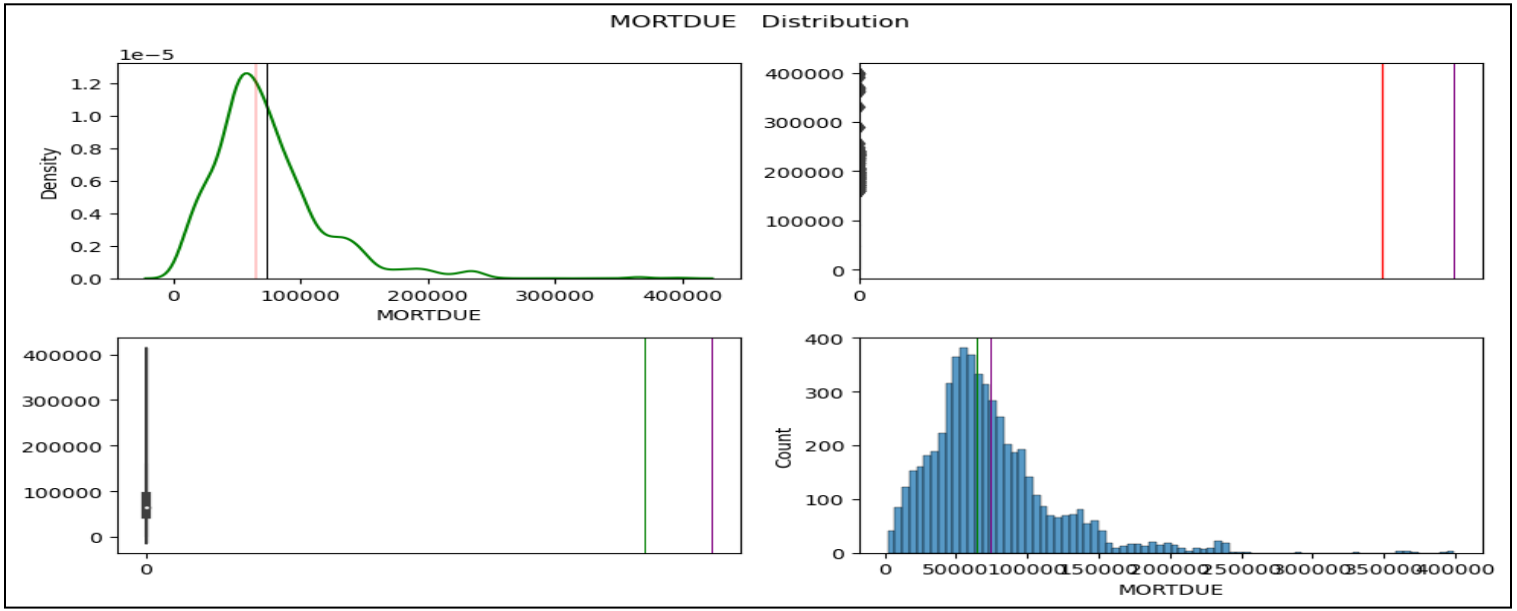


DELINQ Distribution

Observations / Findings

- Debt –to – income rate is not normally distributed. It is skewed to the right. It has outliers.

Exploratory Data Analysis – Univariate Analysis

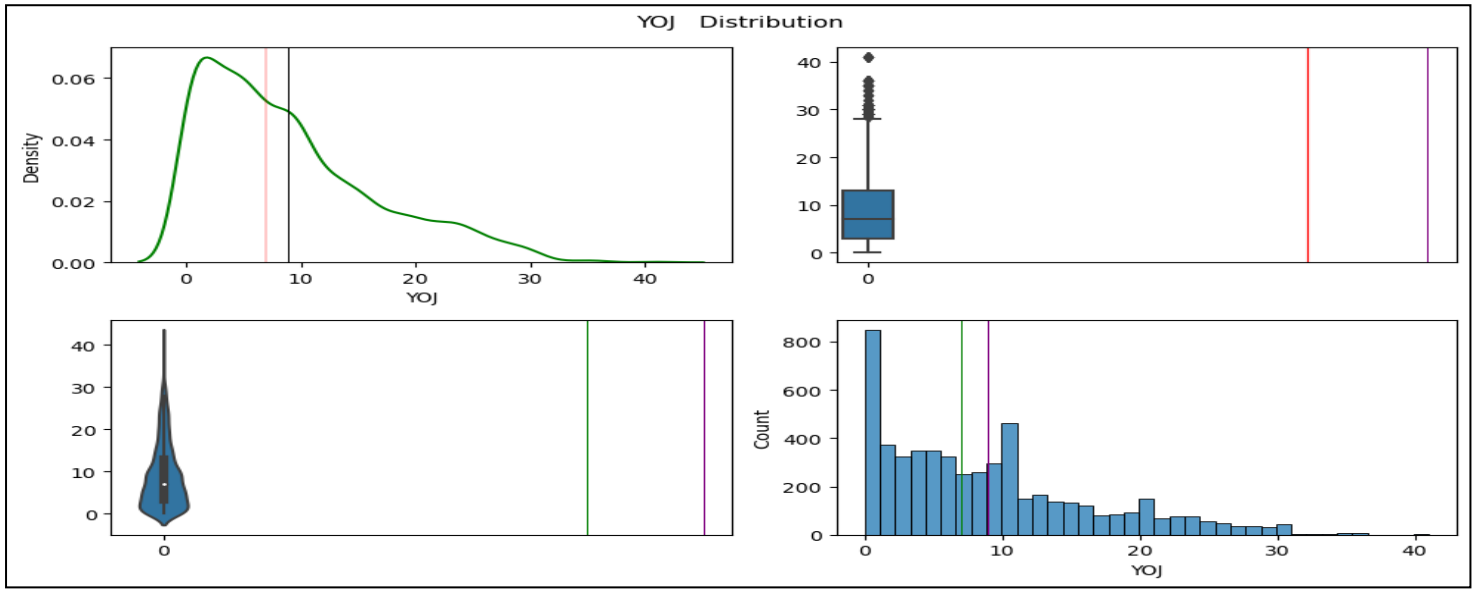


MORTDUE Distribution

Observations / Findings

- Amount due on existing mortgage is not normally distributed. It is skewed to the right signifying the presence of outliers

Exploratory Data Analysis – Univariate Analysis

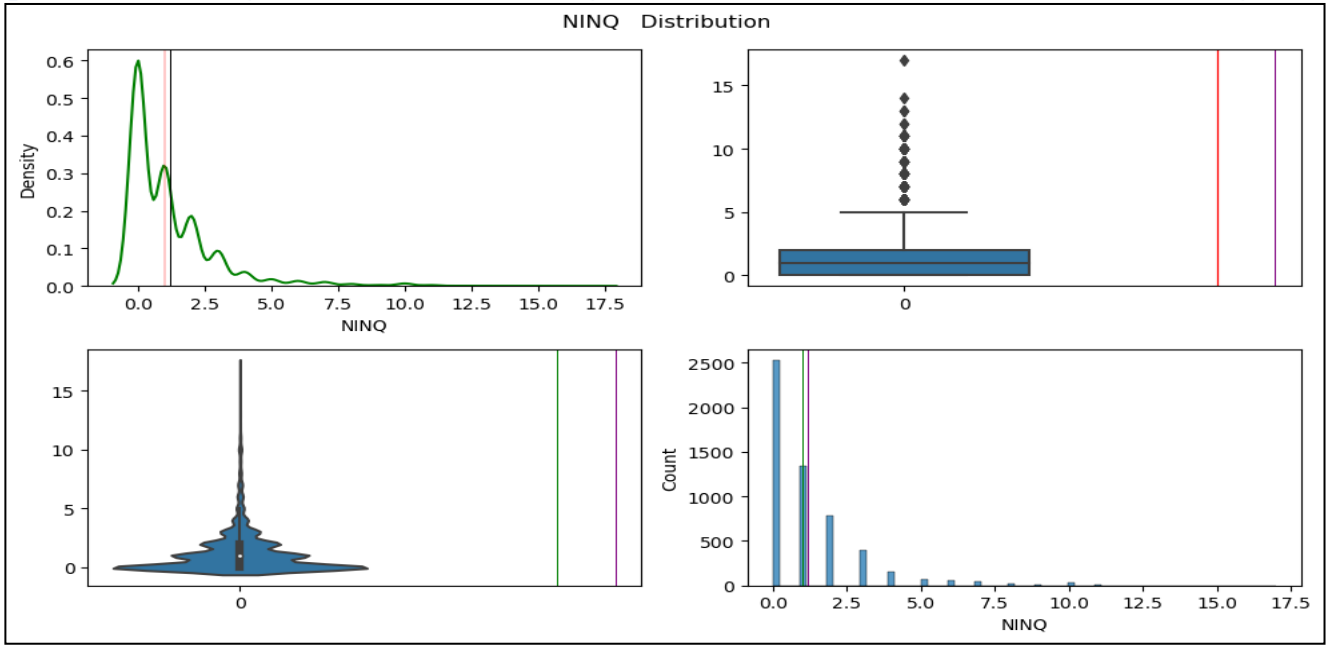


YOJ Distribution

Observations / Findings

- Year at present job is not normally distributed. It is skewed to the right signifying the presence of outliers

Exploratory Data Analysis – Univariate Analysis

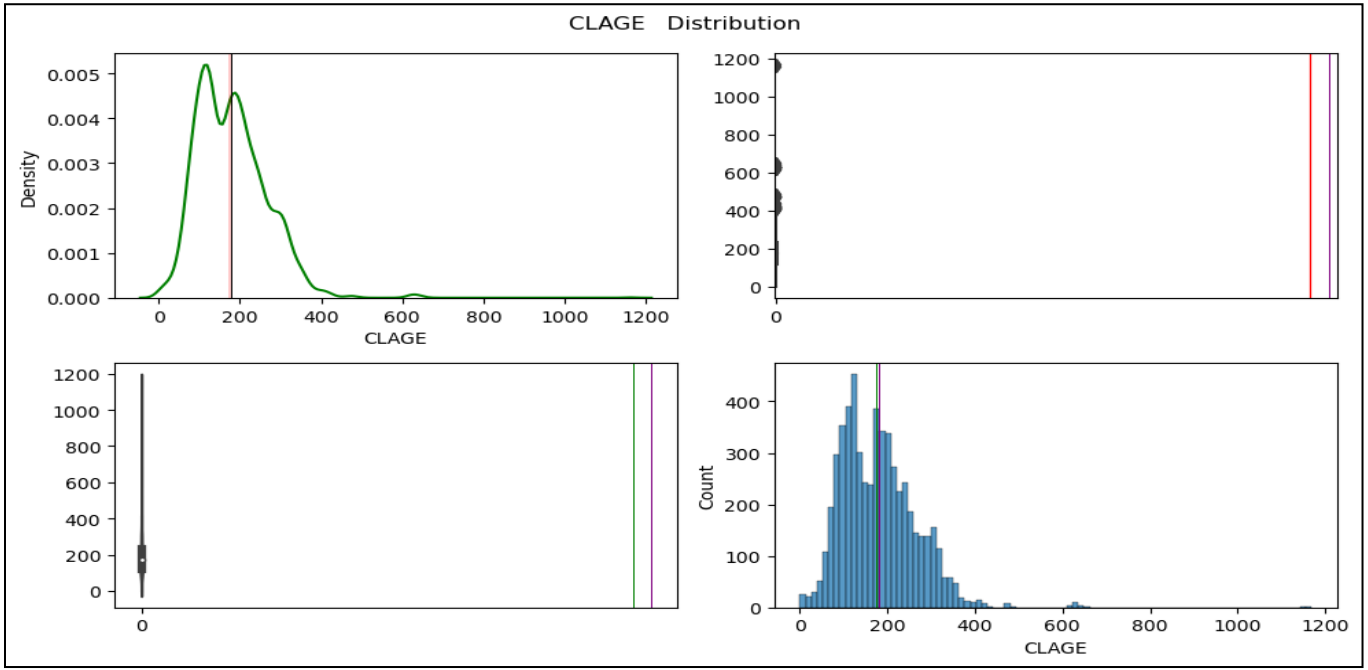


NINQ Distribution

Observations / Findings

- Number of recent credit enquiry is not normally distributed. It is skewed to the right signifying the presence of outliers

Exploratory Data Analysis – Univariate Analysis

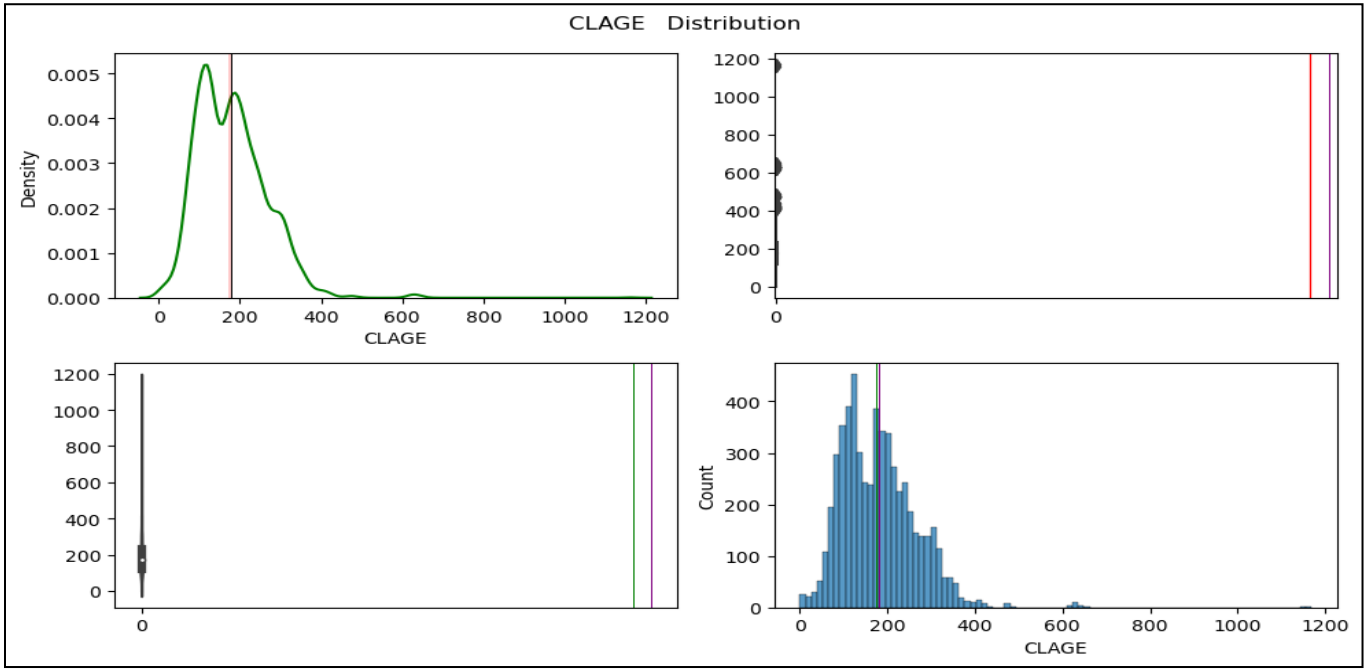


CLAGE Distribution

Observations / Findings

- Age of the oldest credit line in months is not normally distributed. It is skewed to the right. It has outliers

Exploratory Data Analysis – Univariate Analysis

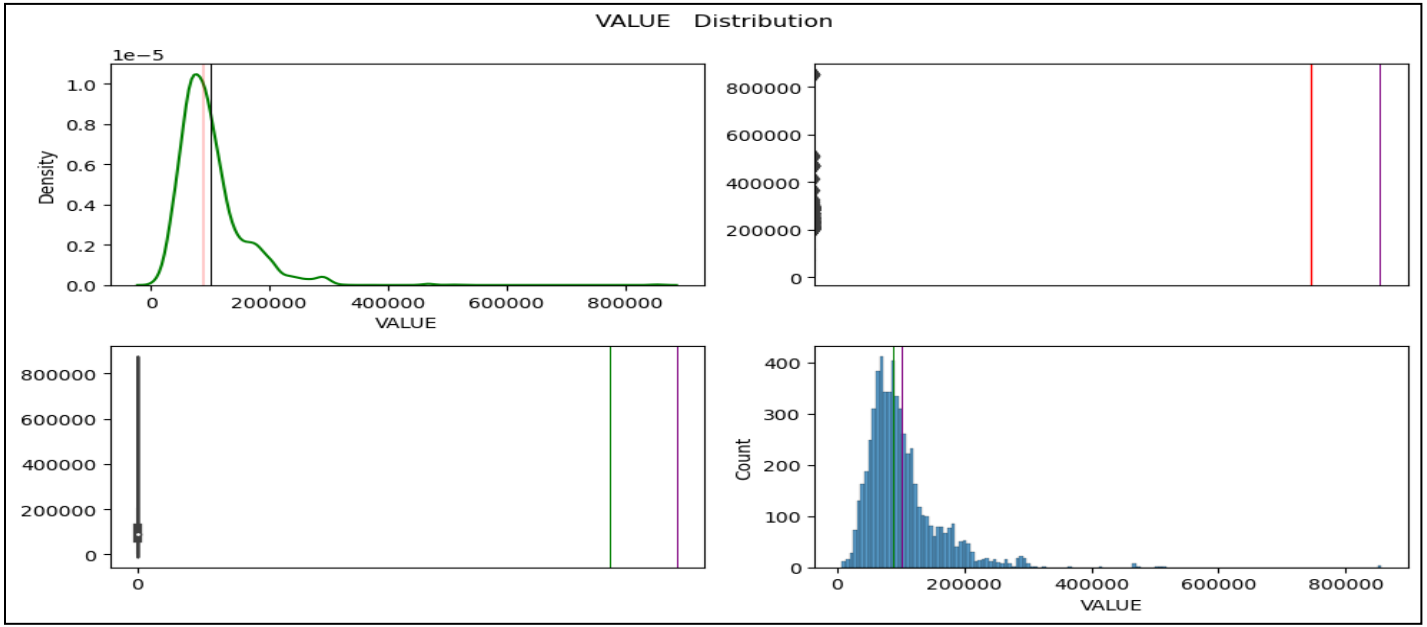


CLAGE Distribution

Observations / Findings

- Age of the oldest credit line in months is not normally distributed. It is skewed to the right. It has outliers

Exploratory Data Analysis – Univariate Analysis

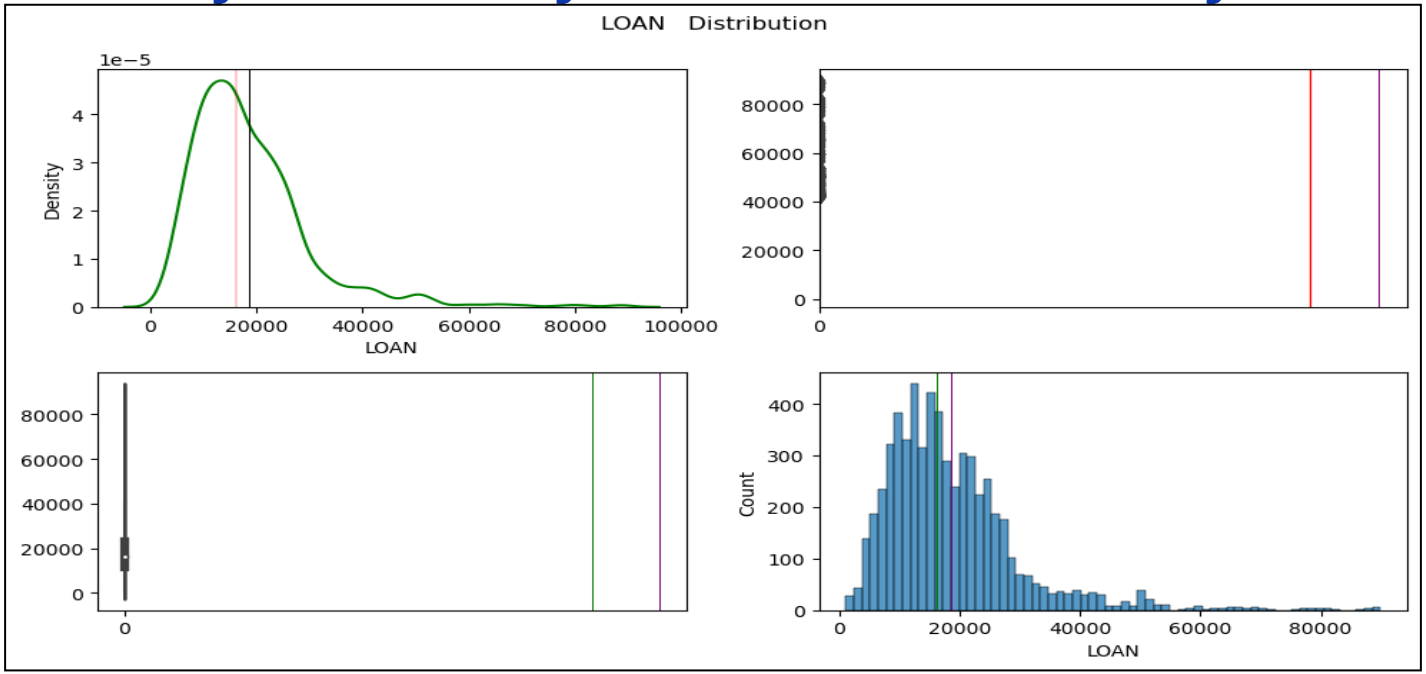


VALUE Distribution

Observations / Findings

- Current value of the property is not normally distributed. It is skewed to the right. It has outliers

Exploratory Data Analysis – Univariate Analysis

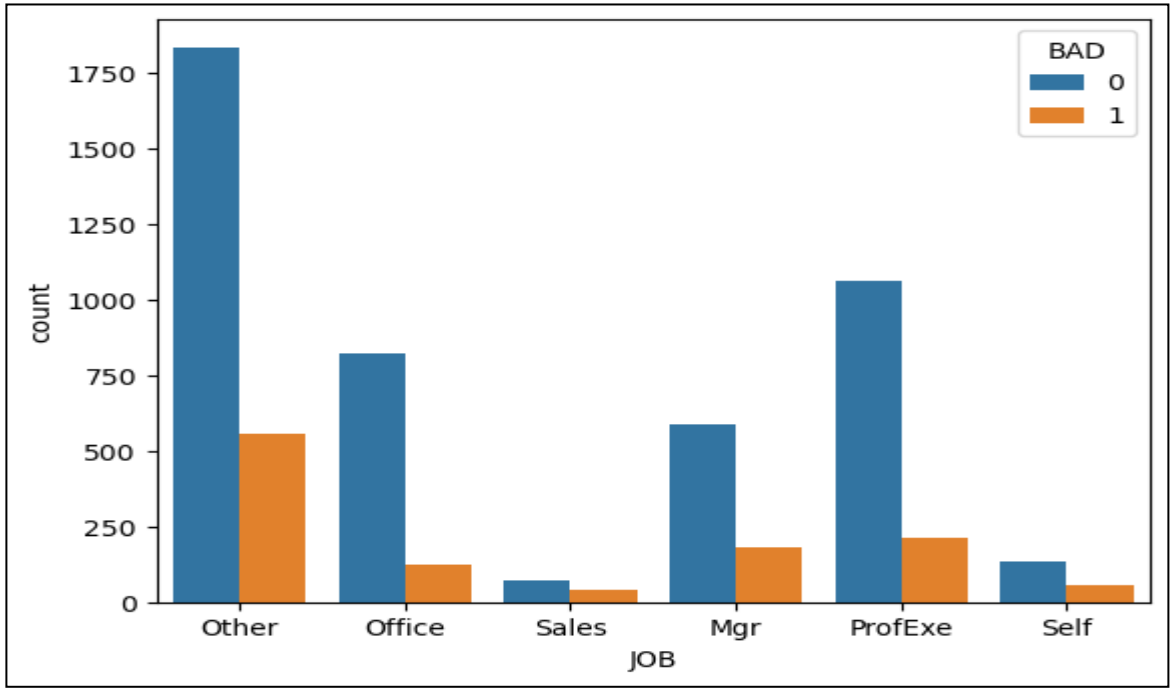


LOAN Distribution

Observations / Findings

- Amount of loan approved is not normally distributed. It is skewed to the right signifying the presence of outliers.

Exploratory Data Analysis – Bivariate Analysis

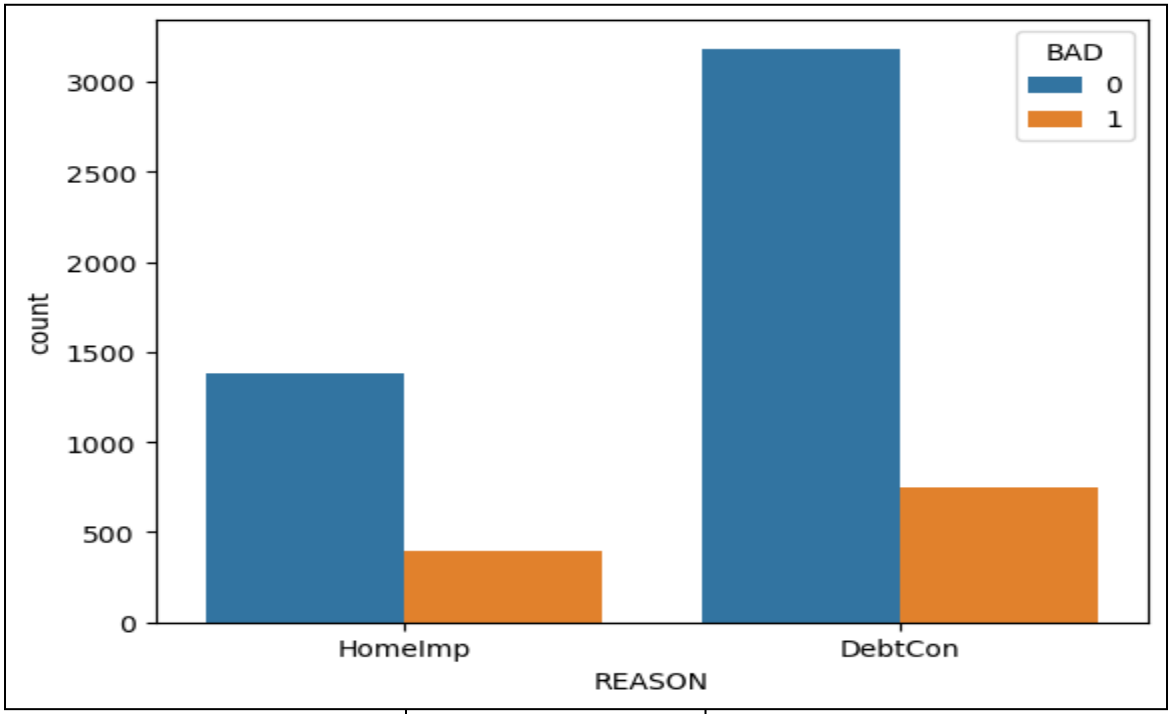


JOB vs BAD

Observations / Findings

- People who filled in others as their JOB are the highest number of clients who took loans. They have the highest defaulters
- Sales and self employed people represent the least number of clients who took loans

Exploratory Data Analysis – Bivariate Analysis

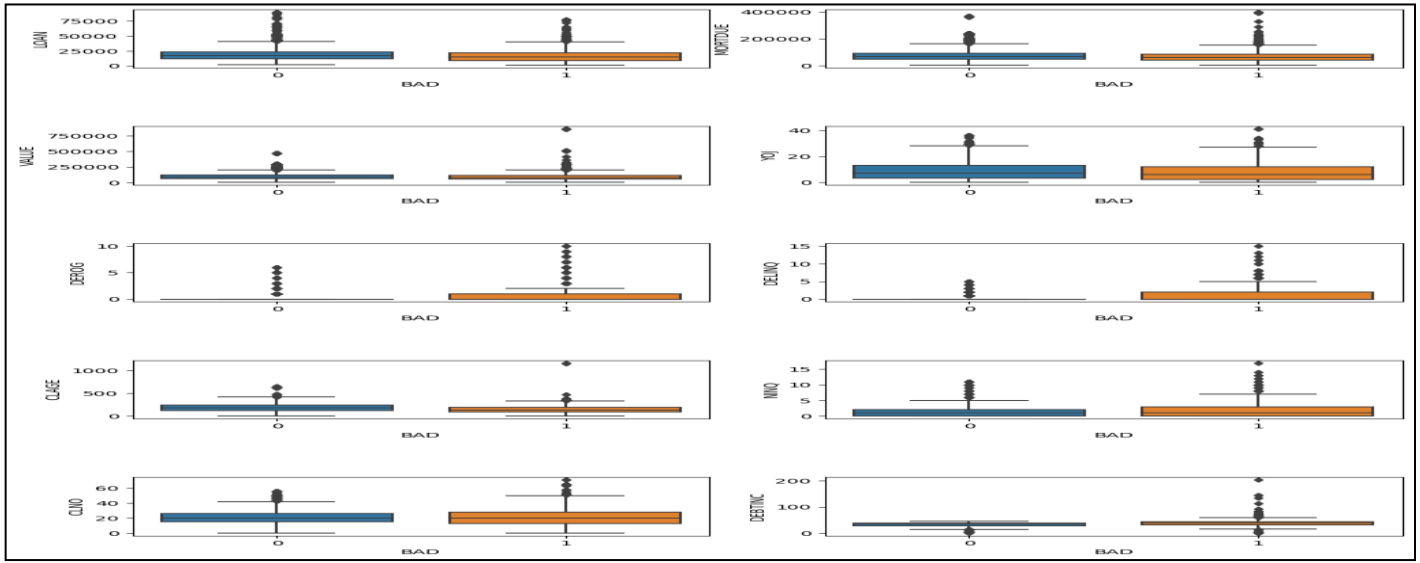


REASON vs BAD

Observations / Findings

- Client who took loan debt consolidation are more. and also have the highest number of clients who repaid their loans

Exploratory Data Analysis – Bivariate Analysis

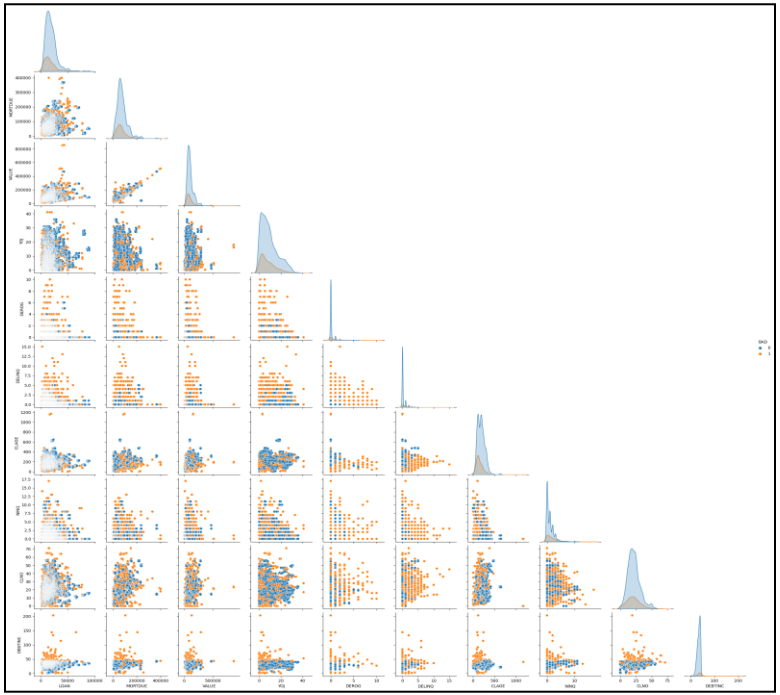


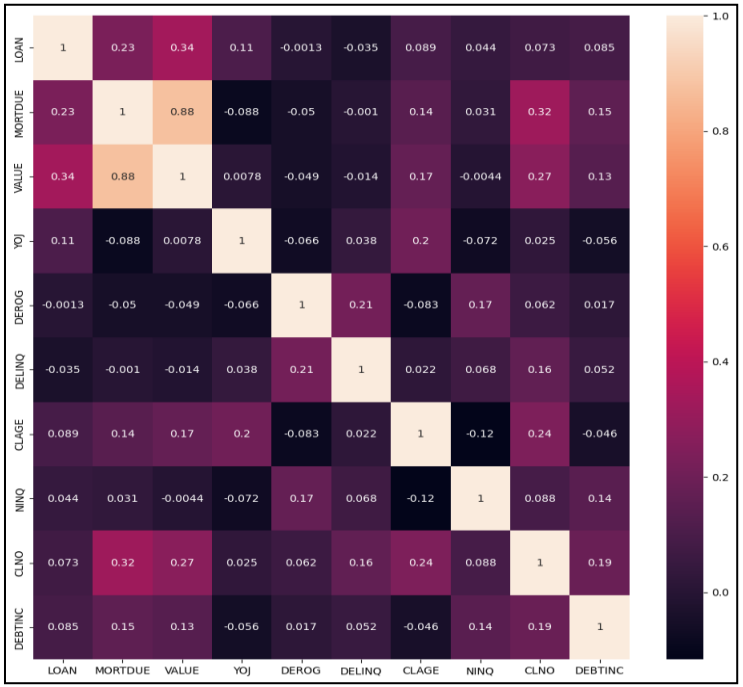
REASON vs BAD

Observations / Findings

- Clients with relatively high loan amount repaid
- more client with high amount due on existing mortgage defaulted
- More client with high current value of property seems to have defaulted
- The more the number of derogatory remarks on client the higher the possibility of default
- Also, the higher the number of delinquent credit lines the higher the chances of default
- As the number of credit enquiry increase the possibility of default also slightly increases
- The more the number of existing credit the more the chances of default
- The most obviously the higher the debt-to-income ration of a client the higher the chances of default

Exploratory Data Analysis – Bivariate Analysis



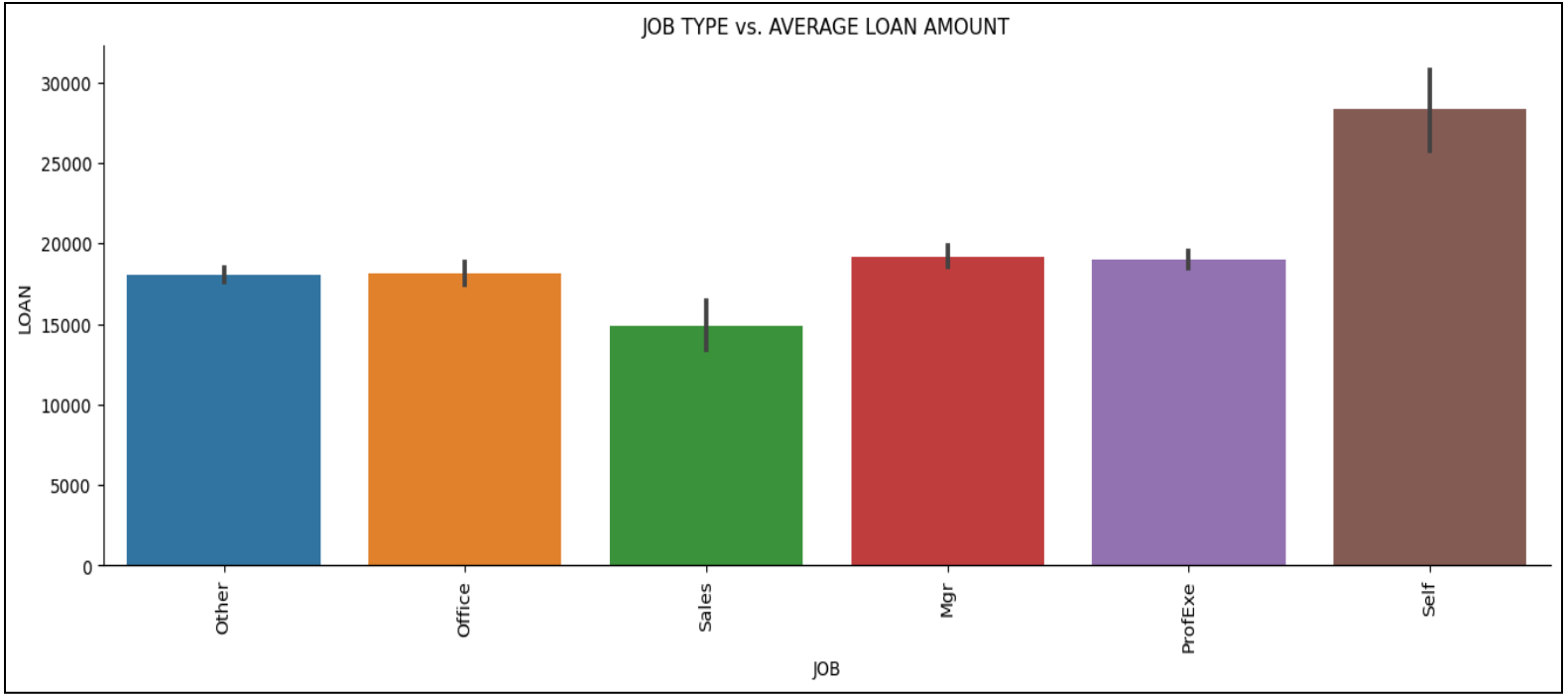


PAIR PLOT & HEAT MAP

Observations / Findings

MORTDUE is highly correlated to VALUE i.e The amount due on existing mortgage is highly correlated to Current value of the property

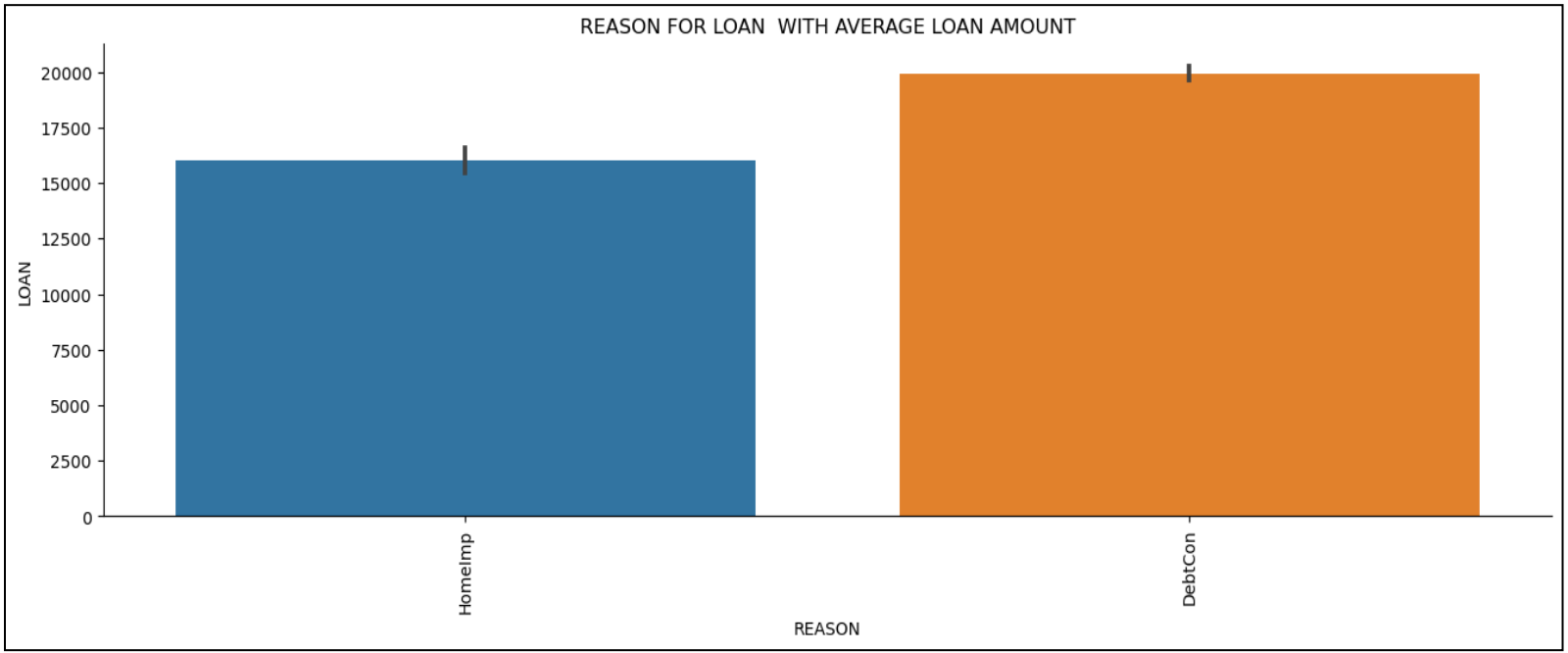
Exploratory Data Analysis – Bivariate Analysis



Observations / Findings

- Self employed client collected the highest loan amount on the average.
- Clients who are sales professionals got the least.

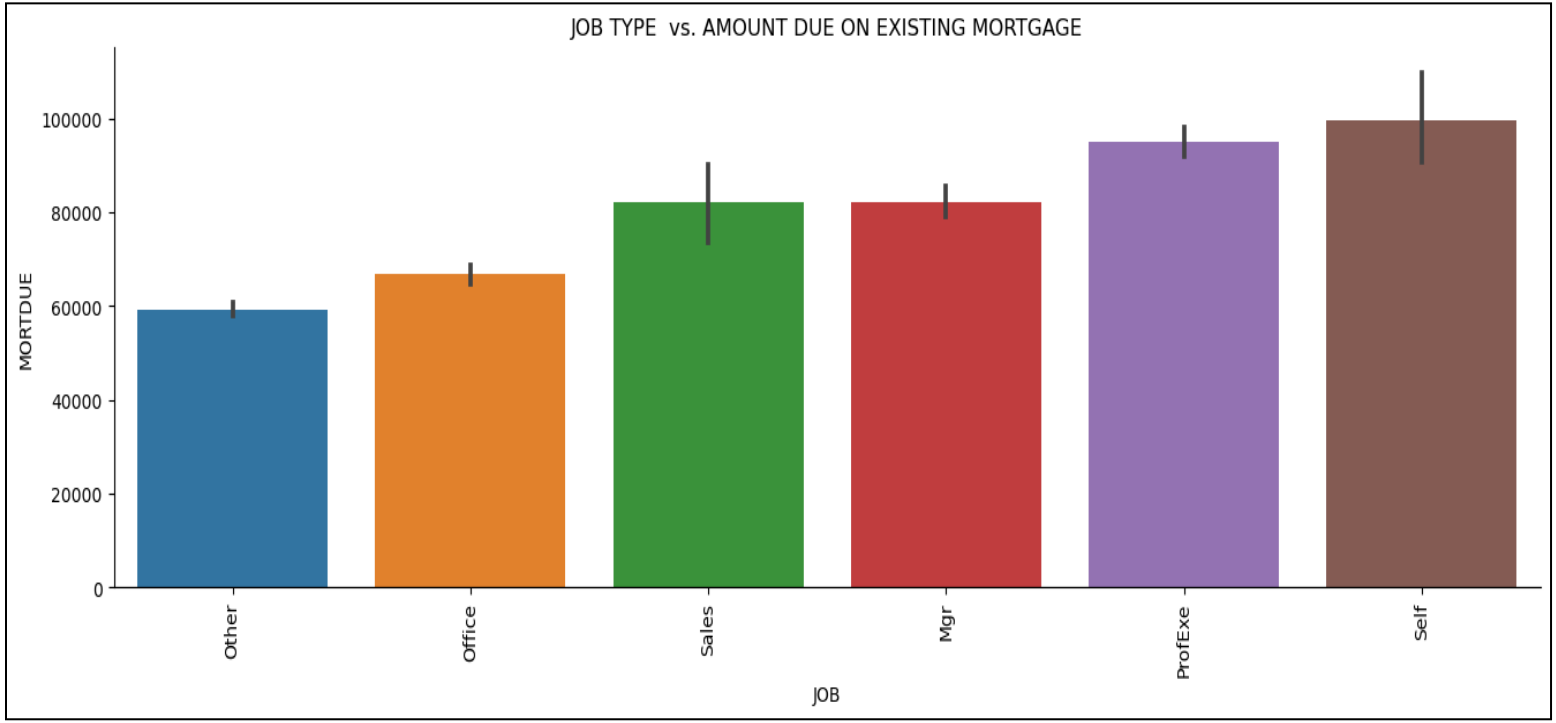
Exploratory Data Analysis – Bivariate Analysis



Observations / Findings

- Clients who took loan for debt consolidation took more on the average.

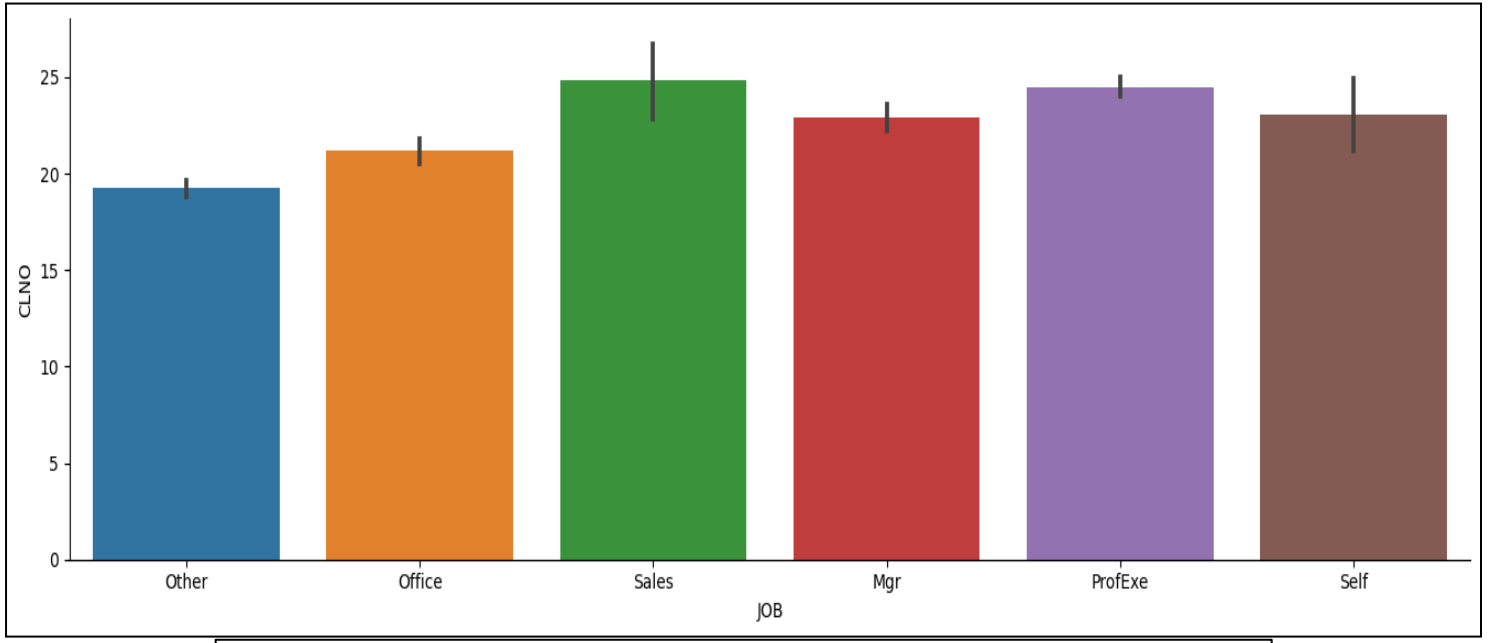
Exploratory Data Analysis – Bivariate Analysis



Observations / Findings

- Self employed client have the highest average amount due on existing mortgage, Client who chose orders as their job have the least

Exploratory Data Analysis – Bivariate Analysis



LOAN APPLICANT'S JOB FOR WITH AVERAGE NUMBER OF EXISTING CREDIT LINES

Observations / Findings

- Clients who are sales professionals have the highest average number of existing credit lines

Model Building and Evaluation

Model Building & Evaluation

Prepare dataset for modeling

```
df.dtypes
```

```
LOAN       float64
MORTDUE    float64
VALUE      float64
YOJ        float64
DEROG      float64
DELINQ     float64
CLAGE      float64
NINQ       float64
CLNO       float64
DEBTINC    float64
PROBINC    float64
BAD          int64
REASON      object
JOB         object
dtype: object
```

Model Building & Evaluation

Logistic Regression

Data preparation for Logistic Regression model

Train-Test-Split for Logistic Regression model

```
Shape of Training set :  (4172, 18)
Shape of test set :  (1788, 18)
Percentage of classes in training set:
0   0.80417
1   0.19583
Name: BAD, dtype: float64
Percentage of classes in test set:
0   0.79195
1   0.20805
Name: BAD, dtype: float64
```

Observations / Findings

We want to predict clients who are likely to default on their loan. Before we proceed to build a model, we'll have to encode categorical features. We'll split the data into train and test to be able to evaluate the model that we build on the train data

Model Building & Evaluation

Building Logistic Regression model

Logit Regression Results

Dep. Variable:	BAD	No. Observations:	4172
Model:	Logit	Df Residuals:	4154
Method:	MLE	Df Model:	17
Date:	Fri, 22 Dec 2023	Pseudo R-squ.:	0.2415
Time:	12:39:05	Log-Likelihood:	-1565.1
converged:	True	LL-Null:	-2063.3
Covariance Type:	nonrobust	LLR p-value:	5.014e-201

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.6234	1.271	2.851	0.004	1.132	6.115
LOAN	-0.1184	0.024	-4.925	0.000	-0.165	-0.071
MORTDUE	-0.8414	0.240	-3.501	0.000	-1.313	-0.370
VALUE	-0.0365	0.143	-0.256	0.798	-0.316	0.243
YOJ	-0.0777	0.136	-0.573	0.567	-0.343	0.188
DEROG	0.6250	0.059	10.656	0.000	0.510	0.740
DELINQ	0.7335	0.046	16.061	0.000	0.644	0.823
CLAGE	-0.0047	0.001	-7.182	0.000	-0.006	-0.003
NINQ	0.1667	0.025	6.597	0.000	0.117	0.216
CLNO	-0.6487	0.156	-4.167	0.000	-0.954	-0.344
DEBTINC	0.0931	0.009	10.572	0.000	0.076	0.110
PROBINC	0.0002	4.52e-05	3.610	0.000	7.46e-05	0.000
REASON_HomeImp	0.1210	0.106	1.144	0.253	-0.086	0.328
JOB_Office	-0.5862	0.182	-3.219	0.001	-0.943	-0.229
JOB_Other	-0.0562	0.141	-0.399	0.690	-0.333	0.220
JOB_ProfExe	0.0522	0.164	0.318	0.751	-0.270	0.374
JOB_Sales	0.6965	0.330	2.109	0.035	0.049	1.344
JOB_Self	0.5637	0.265	2.129	0.033	0.045	1.083

Model Building & Evaluation - Criterion

Model evaluation criterion

Model can make wrong predictions as:

Predicting a customer will not default but in reality, the customer will default to their loan obligation.
Predicting a customer will default to their obligation but in reality, the customer will not default to their obligation.

Which case is more important?

Both the cases are important as:

If we predict that a default will not be occured and the default gets occurred then the bank will lose resources and will have to bear additional costs.

If we predict that a default will get occurred and the booking doesn't get occurred then the bank might not be able to provide satisfactory services to the customer by assuming that this default will be occured. This might damage the brand equity.

How to reduce the losses?**

We will look at `F1 Score` to be maximized, greater the F1 score higher are the chances of minimizing False Negatives and False Positives.

Model Building & Evaluation

- First, let's create functions to calculate different metrics and confusion matrix so that we don't have to use the same code repeatedly for each model.
- The model_performance_classification_statsmodels function will be used to check the model performance of models.
- The confusion_matrix_statsmodels function will be used to plot the confusion matrix.

Model Building & Evaluation

Checking Logistic Regression model performance on the training set

	Accuracy	Recall	Precision	F1
0	0.85139	0.35373	0.75853	0.48247

Observations / Findings

Model Building & Evaluation

Let us check multicollinearity of variables

	feature	VIF
0	const	750.69990
1	LOAN	1.28254
2	MORTDUE	2.29975
3	VALUE	2.49333
4	YOJ	1.07623
5	DEROG	1.08064
6	DELINQ	1.07127
7	CLAGE	1.13811
8	NINQ	1.09370
9	CLNO	1.28833
10	DEBTINC	1.15834
11	PROBINC	1.24661
12	REASON_HomeImp	1.14039
13	JOB_Office	1.88896
14	JOB_Other	2.56109
15	JOB_ProfExe	2.14889
16	JOB_Sales	1.12629
17	JOB_Self	1.26024

Observations / Findings

- None of the numerical variables show moderate or high multicollinearity.
- We will ignore the VIF for the dummy variables.

Model Building & Evaluation

Let us check variables with p-value to drop them if possible

We will drop the predictor variables having a p-value greater than 0.05 as they do not significantly impact the target variable. But sometimes p-values change after dropping a variable. So, we'll not drop all variables at once.
Instead, we will do the following:

- Build a model, check the p-values of the variables, and drop the column with the highest p-value.
- Create a new model without the dropped feature, check the p-values of the variables, and drop the column with the highest p-value.
- Repeat the above two steps till there are no columns with p-value > 0.05.

The above process can also be done manually by picking one variable at a time that has a high p-value, dropping it, and building a model again. But that might be a little tedious and using a loop will be more efficient.

Observations / Findings

Model Building & Evaluation

Great Learning
POWER AHEAD

Logit Regression Results

Dep. Variable:	BAD	No. Observations:	4172
Model:	Logit	Df Residuals:	4159
Method:	MLE	Df Model:	12
Date:	Fri, 22 Dec 2023	Pseudo R-squ.:	0.2409
Time:	12:39:33	Log-Likelihood:	-1566.2
converged:	True	LL-Null:	-2063.3
Covariance Type:	nonrobust	LLR p-value:	3.256e-205

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.2490	0.813	3.998	0.000	1.656	4.842
LOAN	-0.1285	0.022	-5.851	0.000	-0.171	-0.085
MORTDUE	-0.8471	0.196	-4.313	0.000	-1.232	-0.462
DEROG	0.6265	0.058	10.736	0.000	0.512	0.741
DELINQ	0.7309	0.046	16.057	0.000	0.642	0.820
CLAGE	-0.0047	0.001	-7.246	0.000	-0.006	-0.003
NINQ	0.1645	0.025	6.573	0.000	0.115	0.214
CLNO	-0.6584	0.151	-4.348	0.000	-0.955	-0.362
DEBTINC	0.0933	0.009	10.688	0.000	0.076	0.110
PROBINC	0.0002	4.53e-05	3.694	0.000	7.86e-05	0.000
JOB_Office	-0.5569	0.144	-3.880	0.000	-0.838	-0.276
JOB_Sales	0.6921	0.307	2.257	0.024	0.091	1.293
JOB_Self	0.6223	0.236	2.636	0.008	0.160	1.085

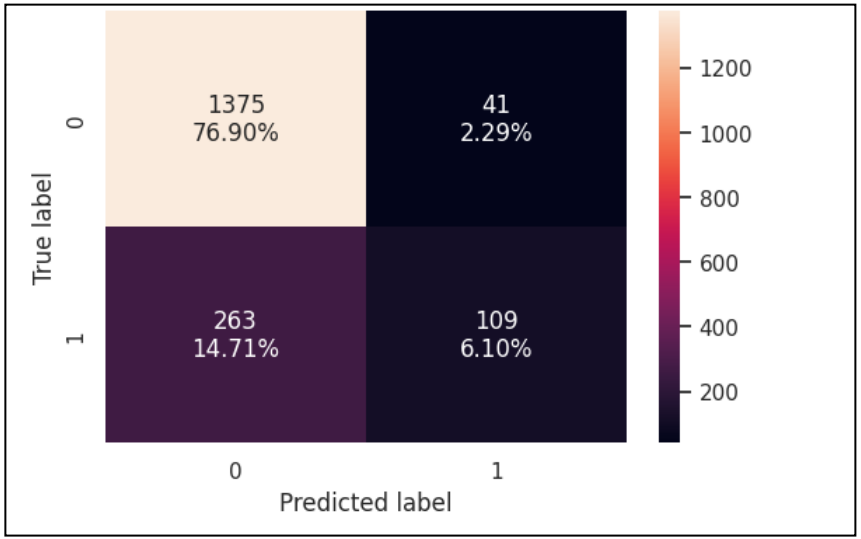
	Accuracy	Recall	Precision	F1
0	0.85163	0.35373	0.76053	0.48287

Observations / Findings

- All the variables left have p-value<0.05.
- So we can say that lg1 is the best model for making any inference.
- The performance on the training data is the same as before dropping the variables with the high p-value.

Model Building & Evaluation

Checking Logistic Regression model performance on the testing set¶



Test performance:

	Accuracy	Recall	Precision	F1
0	0.82998	0.29301	0.72667	0.41762

Observations / Findings

Model Building & Evaluation

Converting coefficients to odds

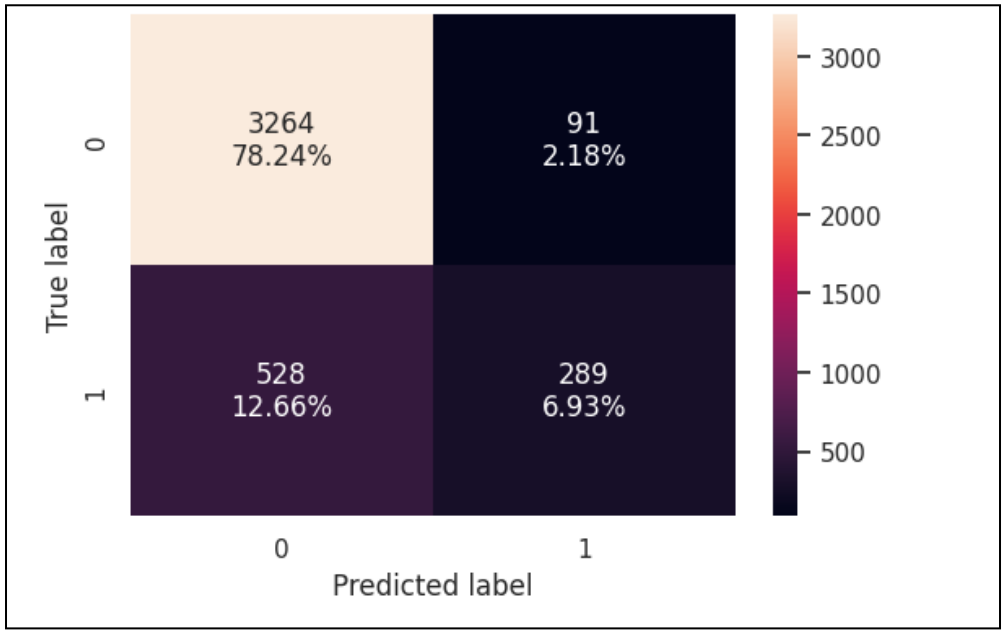
	const	LOAN	MORTDUE	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC	PROBINC	JOB_Office	JOB_Sales	JOB_Self
Odds	25.76339	0.87945	0.42864	1.87097	2.07704	0.99529	1.17879	0.51767	1.09780	1.00017	0.57299	1.99799	1.86326
Change_odd%	2476.33890	-12.05525	-57.13617	87.09705	107.70447	-0.47108	17.87866	-48.23295	9.77966	0.01673	-42.70095	99.79913	86.32564

Checking model performance on the training set

Observations / Findings

Model Building & Evaluation

Checking model performance on the training set



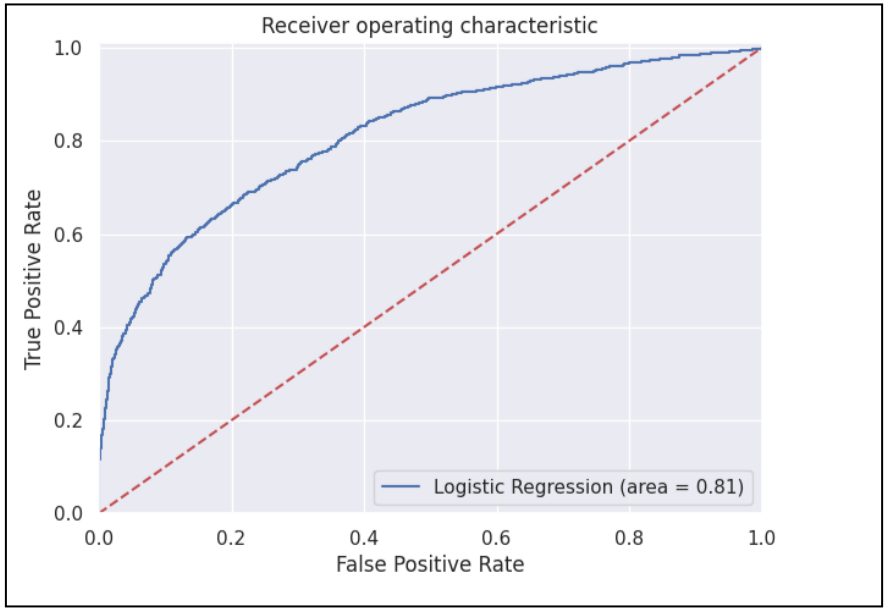
Training performance:

	Accuracy	Recall	Precision	F1
0	0.85163	0.35373	0.76053	0.48287

Model Building & Evaluation

ROC-AUC

ROC-AUC on training set

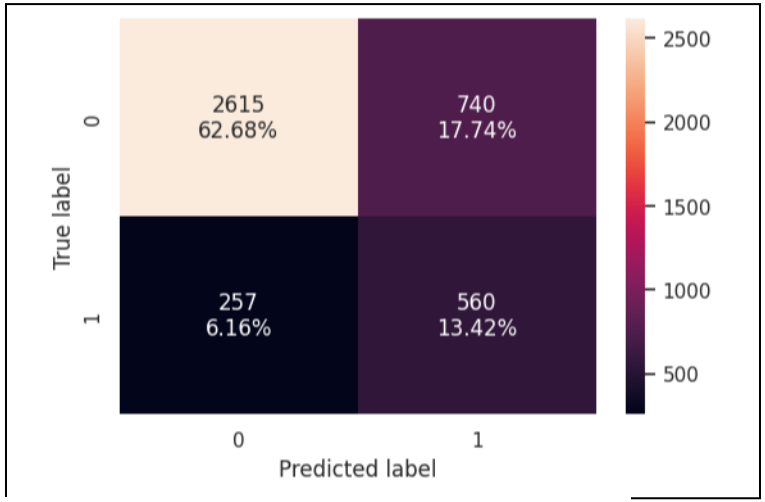


Observations / Findings

- Logistic Regression model is giving a generalized performance on training and test set.
- ROC-AUC score of 0.81 on training is quite good

Model Building & Evaluation

confusion matrix



Training performance:

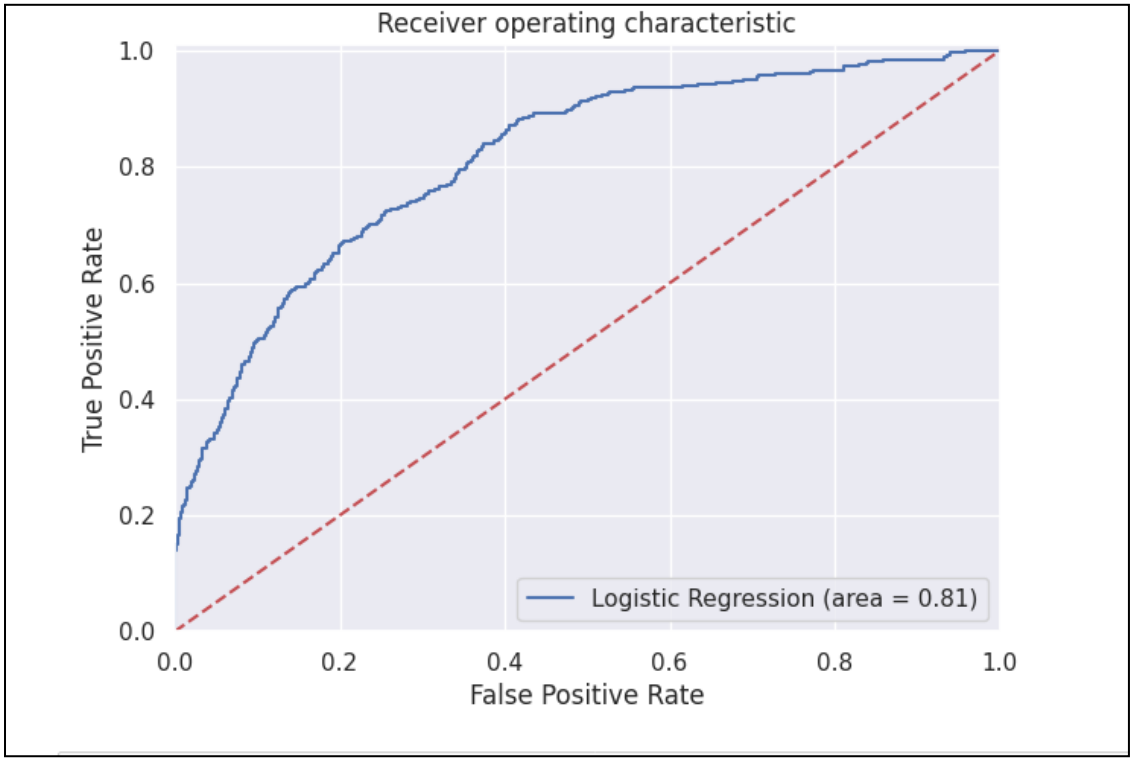
	Accuracy	Recall	Precision	F1
0	0.76103	0.68543	0.43077	0.52905

Observations / Findings

- Recall has increased significantly as compared to the previous model.
- As we will decrease the threshold value, Recall will keep on increasing and the Precision will decrease, but this is not right, we need to choose an optimal balance between recall and precision

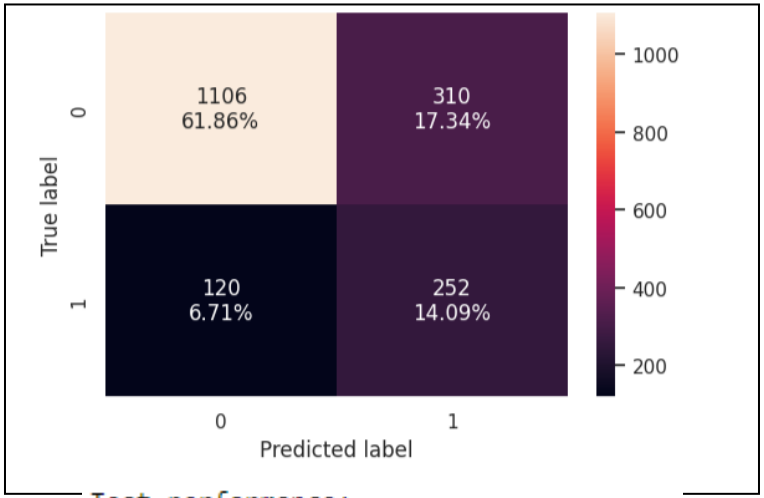
Model Building & Evaluation

Let's check the performance on the test set



Model Building & Evaluation

confusion matrix



Test performance:

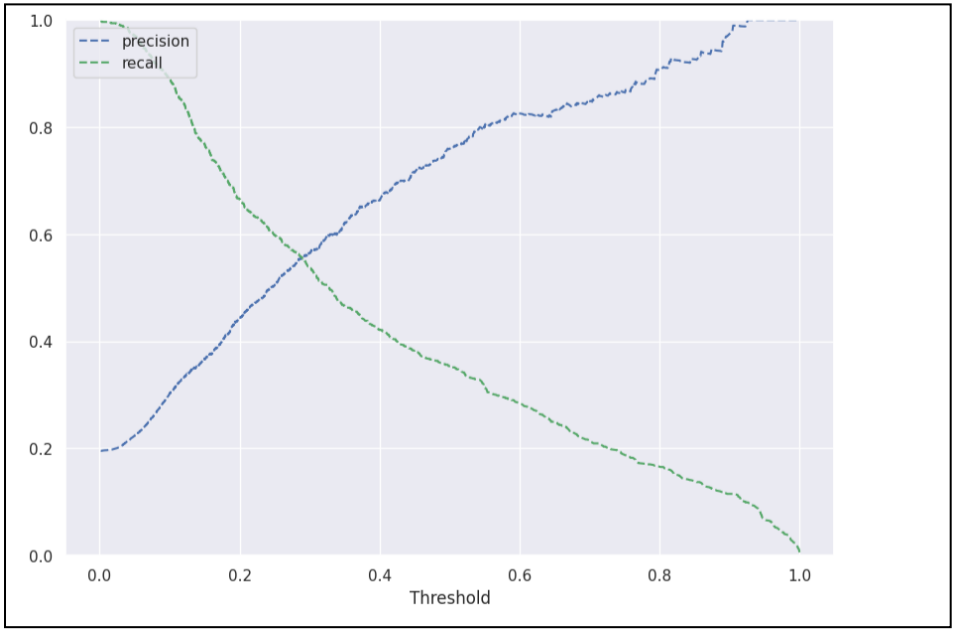
	Accuracy	Recall	Precision	F1
0	0.75951	0.67742	0.44840	0.53961

Observations / Findings

Let's use Precision-Recall curve and see if we can find a better threshold

Model Building & Evaluation

Precision-Recall curve

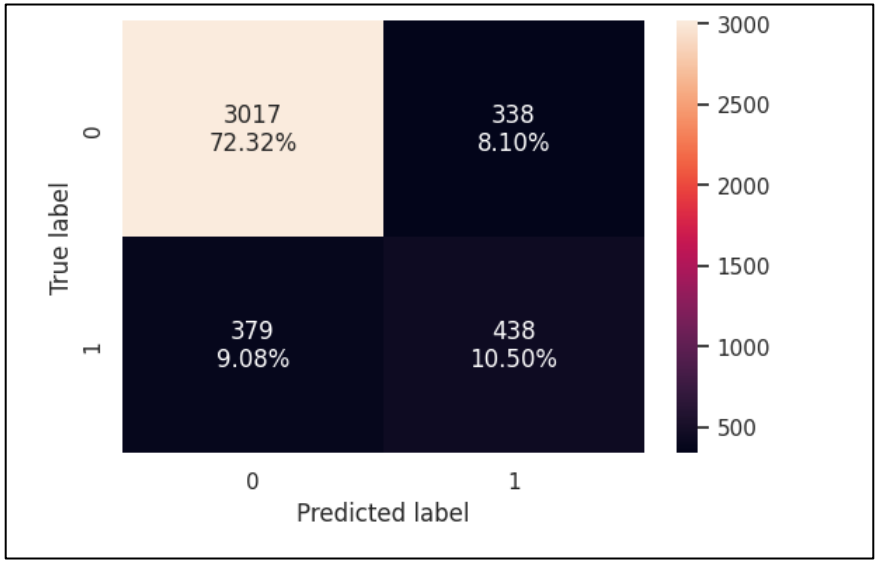


Observations / Findings

At 0.3 threshold we get a balanced precision and recall.

Model Building & Evaluation

Checking model performance on training set



Training performance:

	Accuracy	Recall	Precision	F1
0	0.82814	0.53611	0.56443	0.54991

Model Building & Evaluation

Logistic Regress Model performance summary

Training performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.85163	0.76103	0.82814
Recall	0.35373	0.68543	0.53611
Precision	0.76053	0.43077	0.56443
F1	0.48287	0.52905	0.54991

Test performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.82998	0.75951	0.81823
Recall	0.29301	0.67742	0.48656
Precision	0.72667	0.44840	0.57460
F1	0.41762	0.53961	0.52693

Observations / Findings

Model Building & Evaluation

Support Vector Machines

Data Preparation for Support vector Machines¶

Head

	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC	PROBINC	BAD	REASON	JOB
0	17.10492	3.56105	10.57221	3.02042	0.00000	0.00000	94.36667	1.00000	2.94444	34.81826	1975.70831	1	HomeImp	Other
1	17.10492	4.03347	11.13327	2.83321	0.00000	2.00000	121.83333	0.00000	3.17805	34.81826	1975.70831	1	HomeImp	Other
2	17.10492	3.28316	9.72376	2.63906	0.00000	0.00000	149.46667	1.00000	2.99573	34.81826	1975.70831	1	HomeImp	Other
3	17.10492	3.99604	11.39915	2.83321	0.00000	0.00000	173.46667	1.00000	3.40120	34.81826	1975.70831	1	DebtCon	Other
4	17.10492	4.20526	11.62634	2.56495	0.00000	0.00000	93.33333	0.00000	3.17805	34.81826	1975.70831	0	HomeImp	Office

Observations / Findings

Support Vector Machines (SVMs) are a powerful set of supervised learning methods that can be effectively used for both classification and regression tasks. We will try to use it for our classification problem and will see how it performs compared to the other models

Model Building & Evaluation

Identify categorical features and create dummies

	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC	PROBINC	BAD	REASON_DebtCon	REASON_HomeImp
0	17.10492	3.56105	10.57221	3.02042	0.00000	0.00000	94.36667	1.00000	2.94444	34.81826	1975.70831	1	0	1
1	17.10492	4.03347	11.13327	2.83321	0.00000	2.00000	121.83333	0.00000	3.17805	34.81826	1975.70831	1	0	1
2	17.10492	3.28316	9.72376	2.63906	0.00000	0.00000	149.46667	1.00000	2.99573	34.81826	1975.70831	1	0	1
3	17.10492	3.99604	11.39915	2.83321	0.00000	0.00000	173.46667	1.00000	3.40120	34.81826	1975.70831	1	1	0
4	17.10492	4.20526	11.62634	2.56495	0.00000	0.00000	93.33333	0.00000	3.17805	34.81826	1975.70831	0	0	1

Observations / Findings

Model Building & Evaluation

Train-Test-Split for SVM model¶

```
x_train.shape, x_test.shape
```

```
((4768, 19), (1192, 19))
```

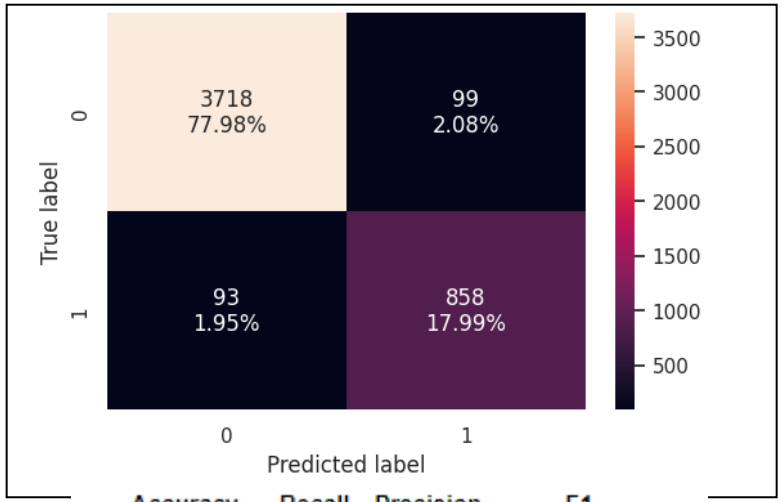
```
Shape of x_train:  (4768, 19)
Shape of y_train:  (4768,)
Shape of x_test:  (1192, 19)
Shape of y_test:  (1192,)
```

Observations / Findings

Model Building & Evaluation

Building Support Vector Machines model¶

Checking Support Vector Machines model performance on the training set

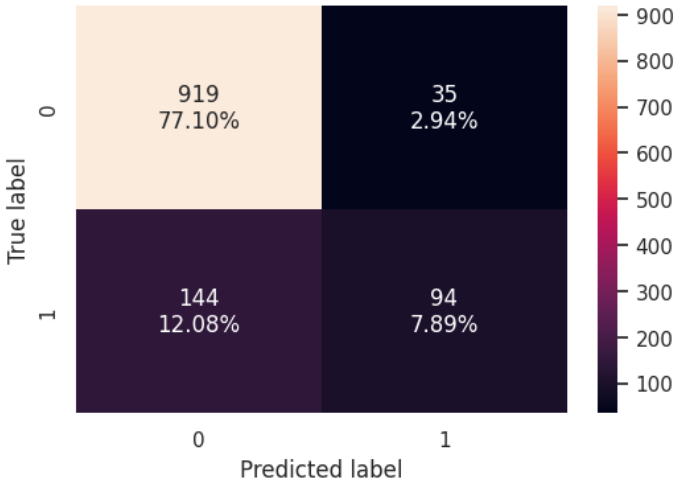


	Accuracy	Recall	Precision	F1
0	0.95973	0.90221	0.89655	0.89937

Observations / Findings

Model Building & Evaluation

Checking model performance on the testing set



	Accuracy	Recall	Precision	F1
0	0.84983	0.39496	0.72868	0.51226

Observations / Findings

Model Building & Evaluation

Support Vector Machines Models Observations

Decision Tree

Data preparation for decision tree model

```
Shape of Training set :  (4172, 19)
Shape of test set :  (1788, 19)
Percentage of classes in training set:
0   0.80417
1   0.19583
Name: BAD, dtype: float64
Percentage of classes in test set:
0   0.79195
1   0.20805
```

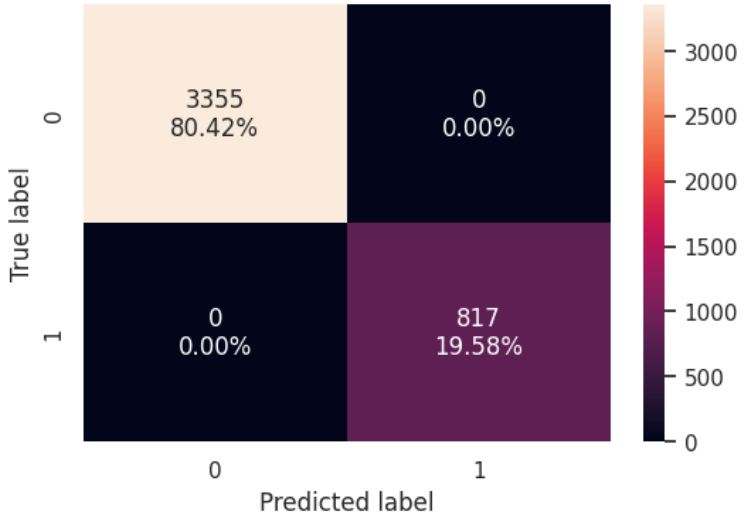
- REASON and JOB are the categorical variables
- The categorical variable REASON and JOB will be transformed using One-hot encoding

Observations / Findings

- We want to predict which bookings will be canceled.
- Before we proceed to build a model, we'll have to encode categorical features.
- We'll split the data into train and test to be able to evaluate the model that we build on the train data.

Model Building & Evaluation

Checking Decision Tree model performance on training set

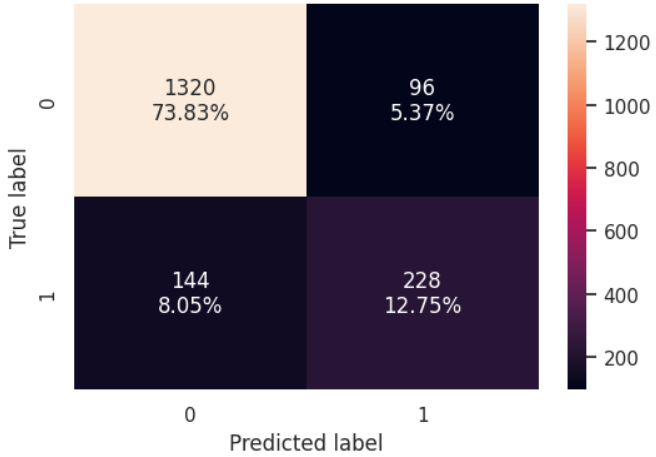


	Accuracy	Recall	Precision	F1
0	1.00000	1.00000	1.00000	1.00000

Observations / Findings

Model Building & Evaluation

Checking Decision Tree model performance on test set



	Accuracy	Recall	Precision	F1
0	0.86577	0.61290	0.70370	0.65517

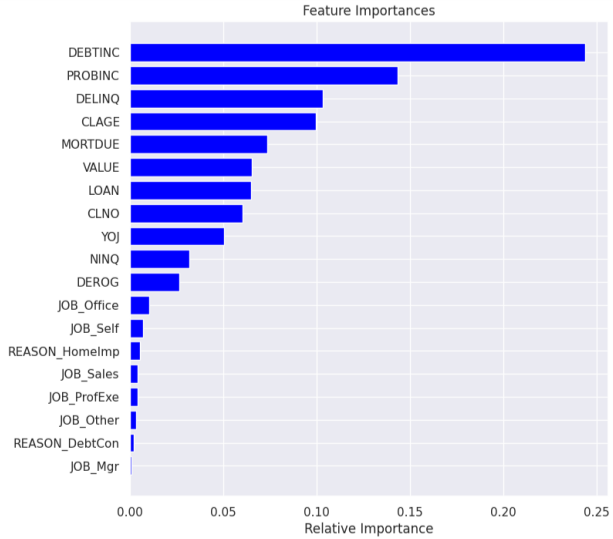
Observations / Findings

* The decision tree model is overfitting the data as expected and not able to generalize well on the test set.
* We will have to prune the decision tree.

Model Building & Evaluation

Before pruning the tree let's check the important features.

Plotting the feature importance of each variable



Observations / Findings

* DEBTINC is the most important feature followed by PROBINC.
* Now let's prune the tree to see if we can reduce the complexity.

Model Building & Evaluation

Pruning decision tree model

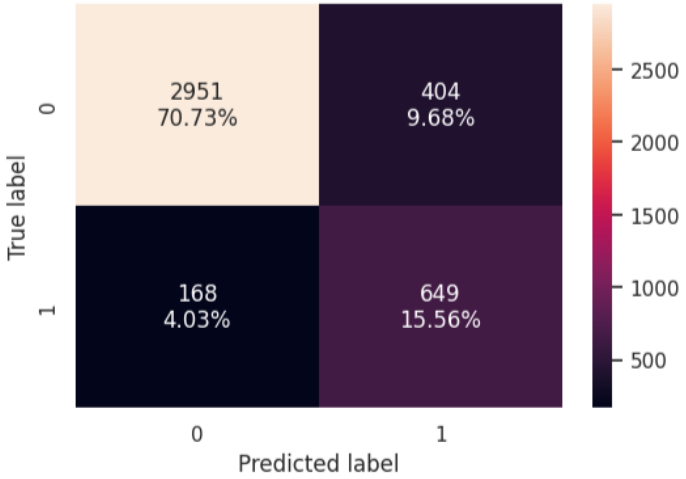
Pre-pruning

	Accuracy	Recall	Precision	F1
0	0.86290	0.79437	0.61633	0.69412

Observations / Findings

Model Building & Evaluation

Checking performance on the training set

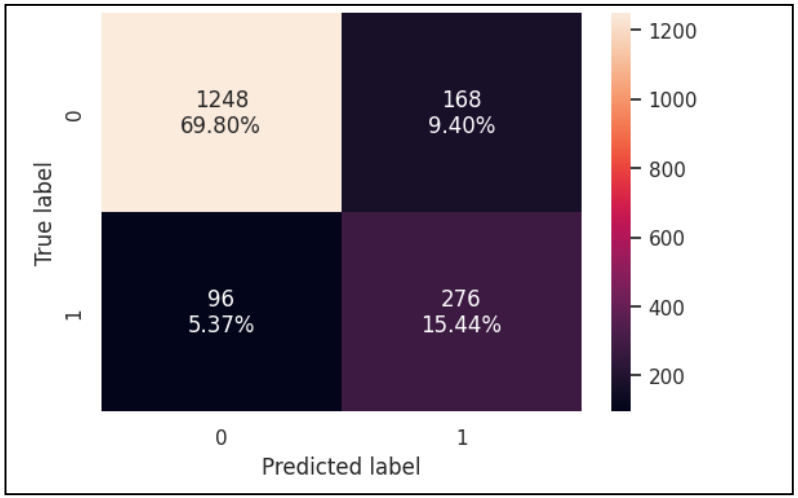


	Accuracy	Recall	Precision	F1
0	0.86290	0.79437	0.61633	0.69412

Observations / Findings

Model Building & Evaluation

Checking performance on the test set

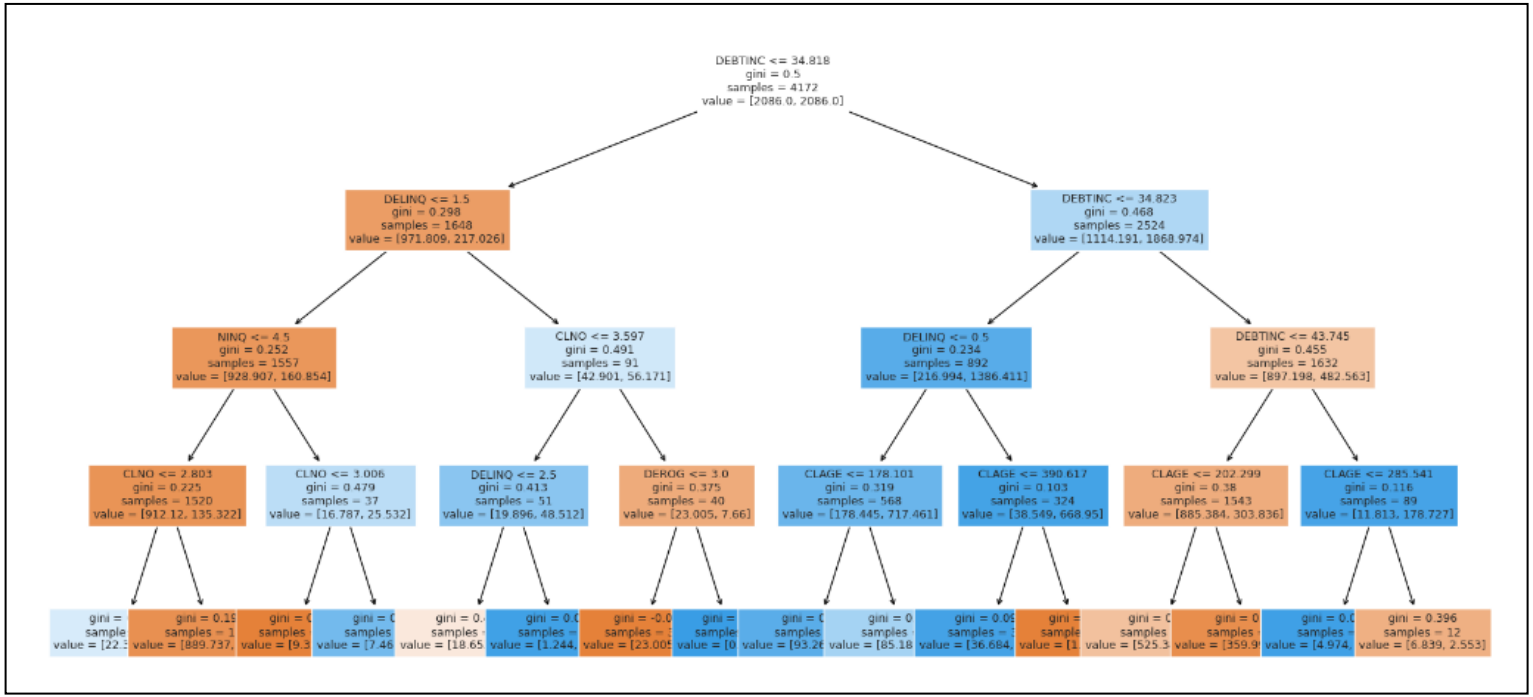


	Accuracy	Recall	Precision	F1
0	0.85235	0.74194	0.62162	0.67647

Observations / Findings

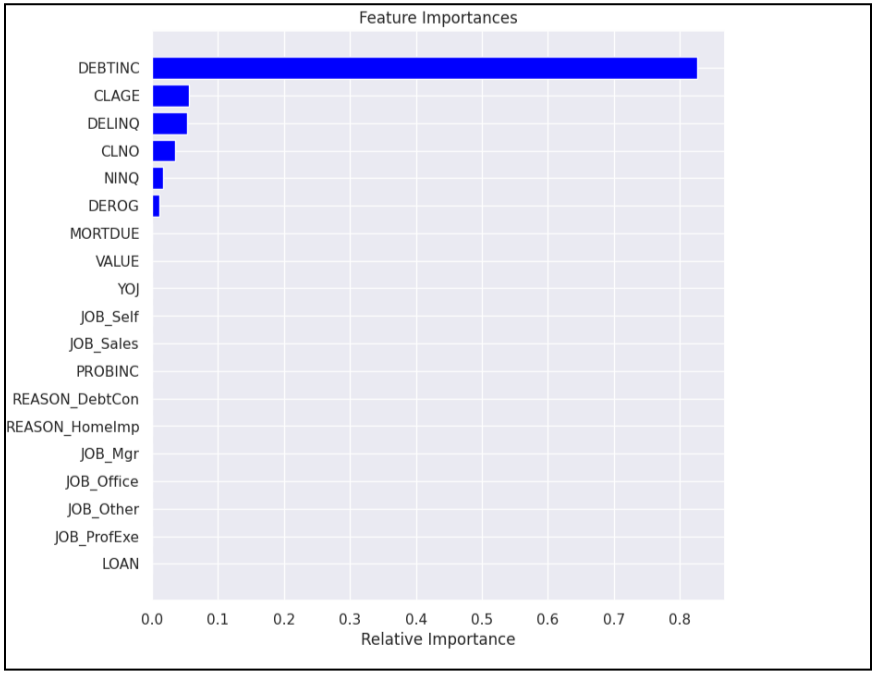
Model Building & Evaluation

Visualization of the Decision Tree



Model Building & Evaluation

Plotting the feature importance of each variable



Observations from decision tree

- We can see that the tree has become simpler and the rules of the trees are readable.
- The model performance of the model has been generalized.
- We observe that the most important features are:
  - DEBTINC
  - CLAGE
  - DELINQ
  - CLNO
  - NINQ
  - DEROG

The rules obtained from the decision tree can be interpreted as:

- The rules show that DEBTINC plays a key role in identifying if a client will default or not.

```
If we want more complex then we can go in more depth of the tree
```

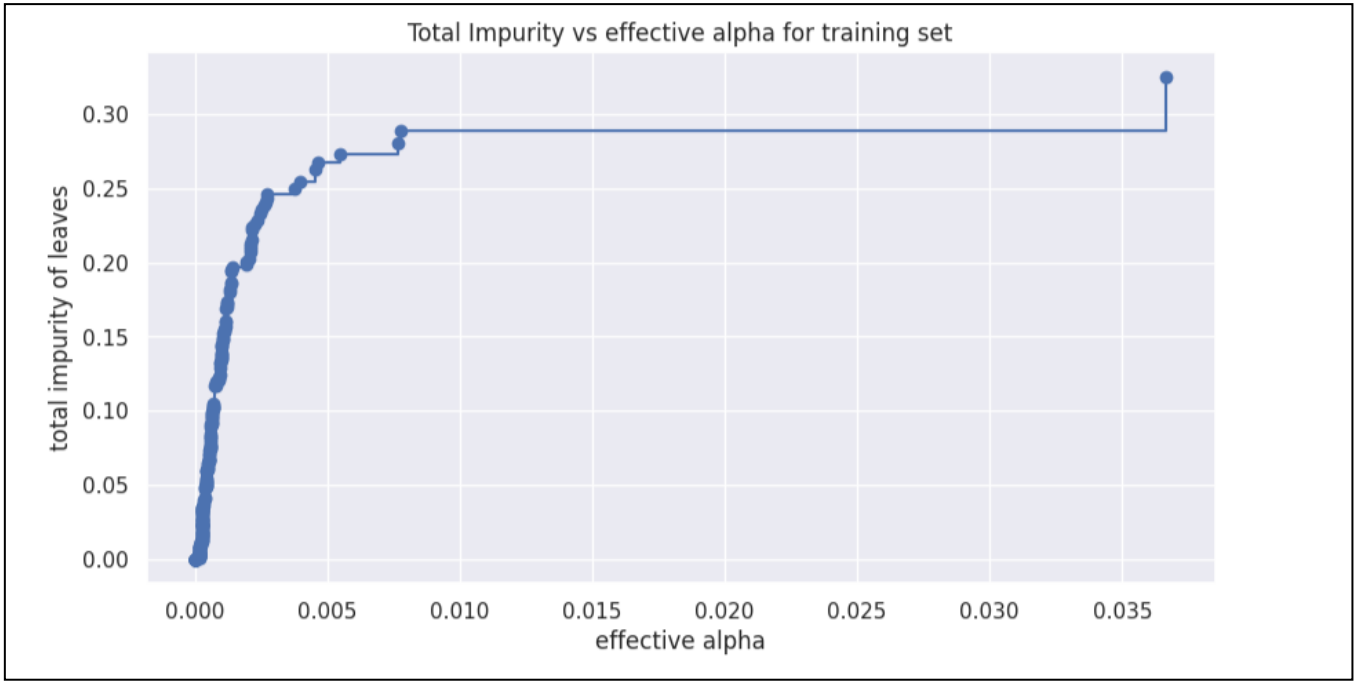
Model Building & Evaluation

Cost complexity pruning

	ccp_alphas	impurities
0	0.00000	-0.00000
1	0.00000	-0.00000
2	0.00000	-0.00000
3	0.00000	-0.00000
4	0.00000	-0.00000
...	...	...
253	0.00548	0.27333
254	0.00766	0.28099
255	0.00776	0.28875
256	0.03667	0.32542
257	0.08729	0.50000

258 rows × 2 columns

Model Building & Evaluation

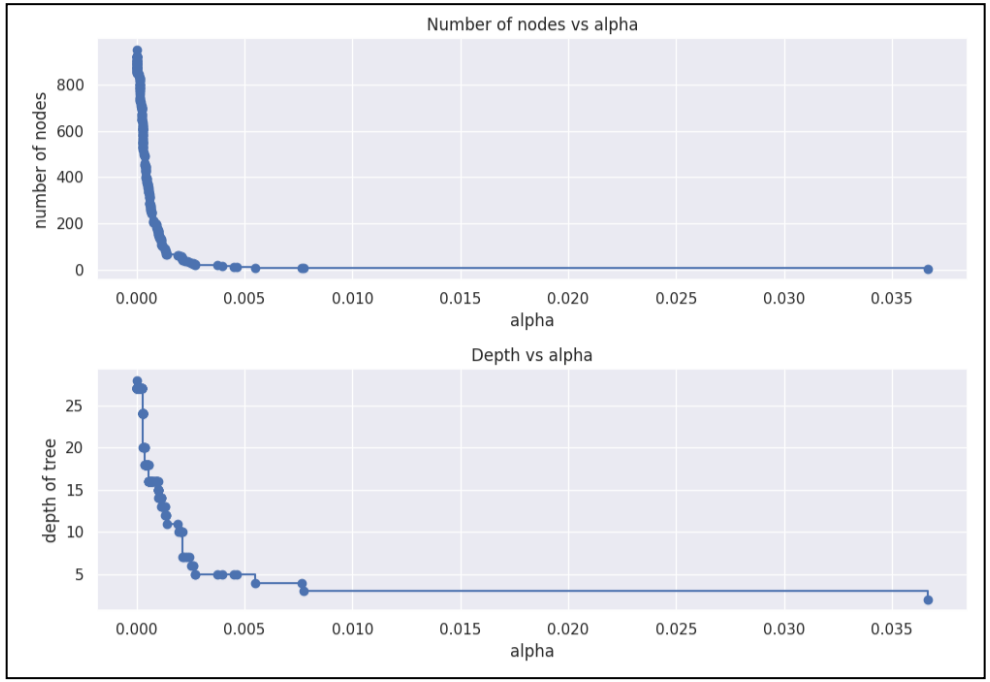


Observations / Findings

Next, we train a decision tree using effective alphas. The last value in ccp_alphas is the alpha value that prunes the whole tree, leaving the tree, clfs[-1], with one node.

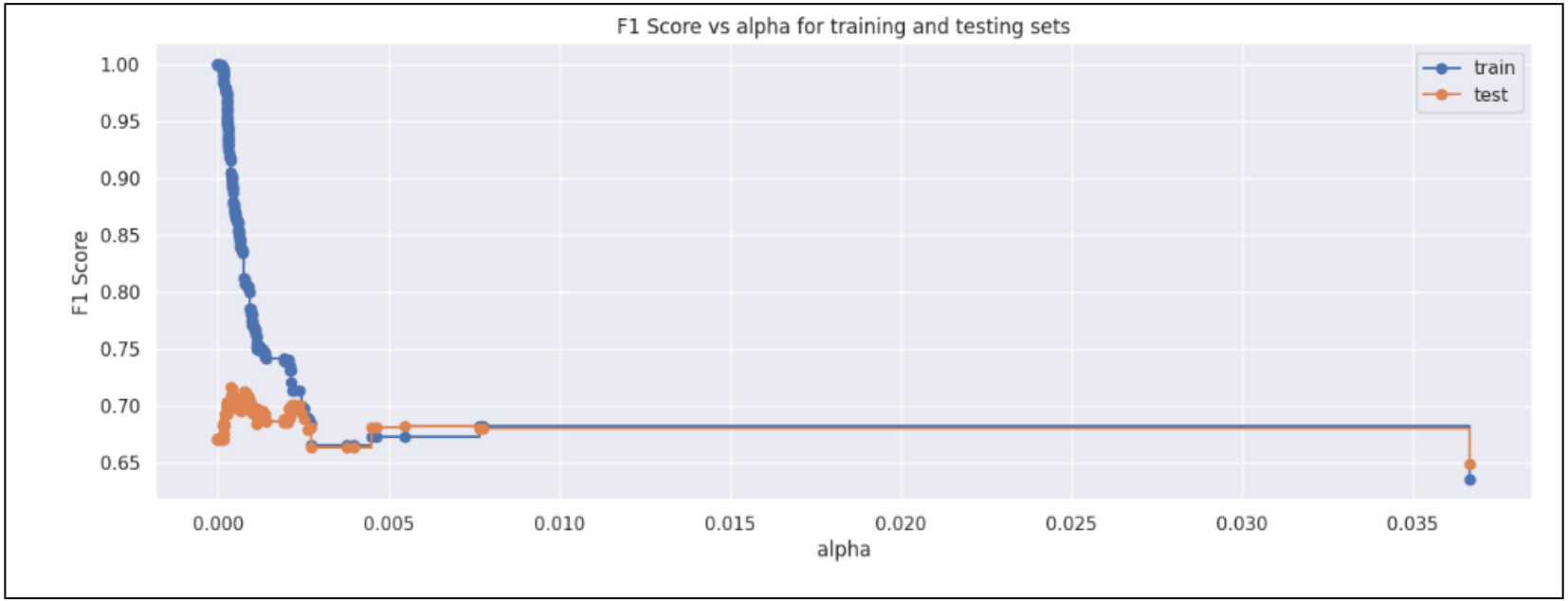
Model Building & Evaluation

```
Number of nodes in the last tree is: 1 with ccp_alpha: 0.08729057011909772
```



Model Building & Evaluation

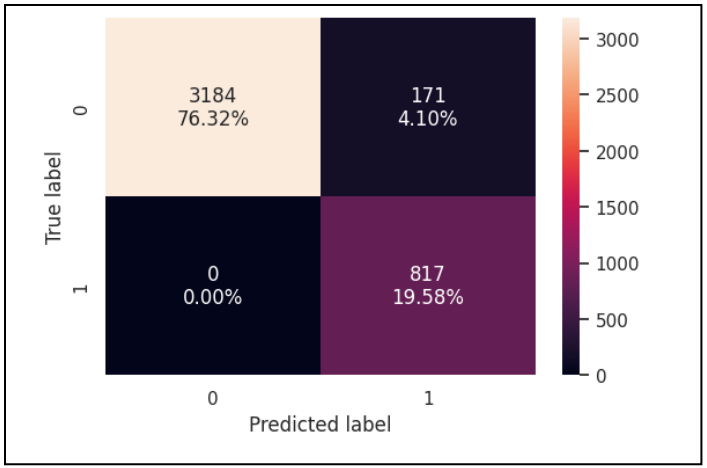
F1 Score vs alpha for training and testing sets



Model Building & Evaluation

```
DecisionTreeClassifier(ccp_alpha=0.00038111933333653467,
```

Checking performance on the training set

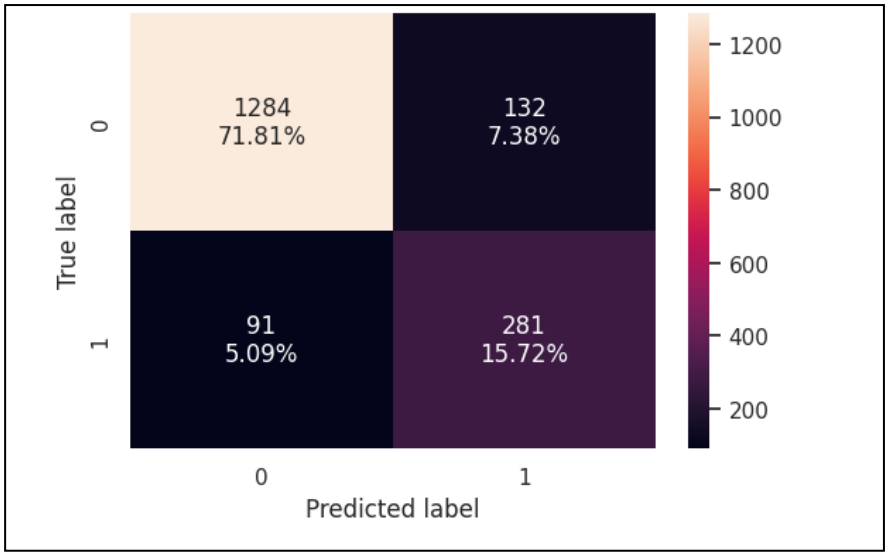


	Accuracy	Recall	Precision	F1
0	0.95901	1.00000	0.82692	0.90526

Observations / Findings

Model Building & Evaluation

Checking performance on the test set

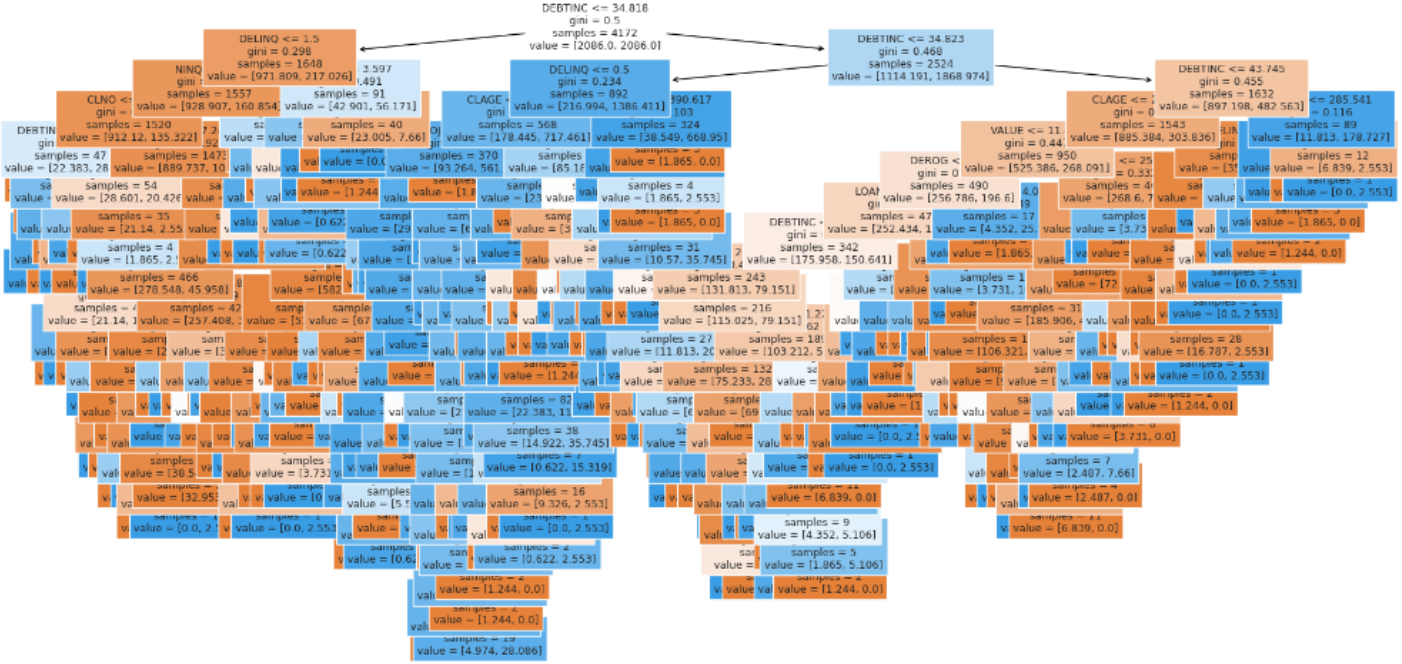


	Accuracy	Recall	Precision	F1
0	0.87528	0.75538	0.68039	0.71592

Observations / Findings

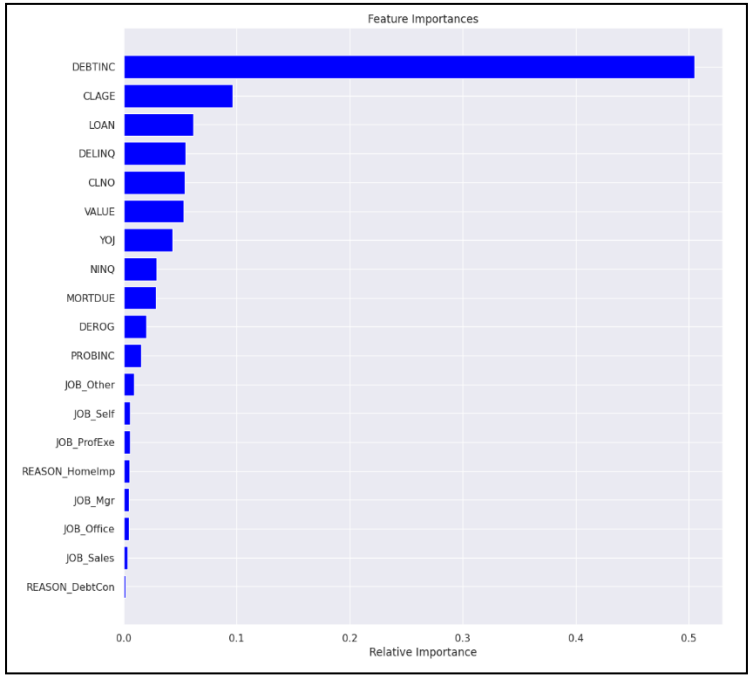
- After post pruning the decision tree the performance has generalized on training and test set.
- We are getting high recall with this model but difference between recall and precision has increased.

Model Building & Evaluation



Model Building & Evaluation

Plotting the feature importance of each variable



Observations / Findings

- The tree is quite complex as complex as compared to the pre-pruned tree.
- The feature importance is same as we got in pre-pruned tree

Model Building & Evaluation

Comparing Decision Tree Models

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	1.00000	0.86290	0.95901
Recall	1.00000	0.79437	1.00000
Precision	1.00000	0.61633	0.82692
F1	1.00000	0.69412	0.90526

Test set performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.86577	0.85235	0.87528
Recall	0.61290	0.74194	0.75538
Precision	0.70370	0.62162	0.68039
F1	0.65517	0.67647	0.71592

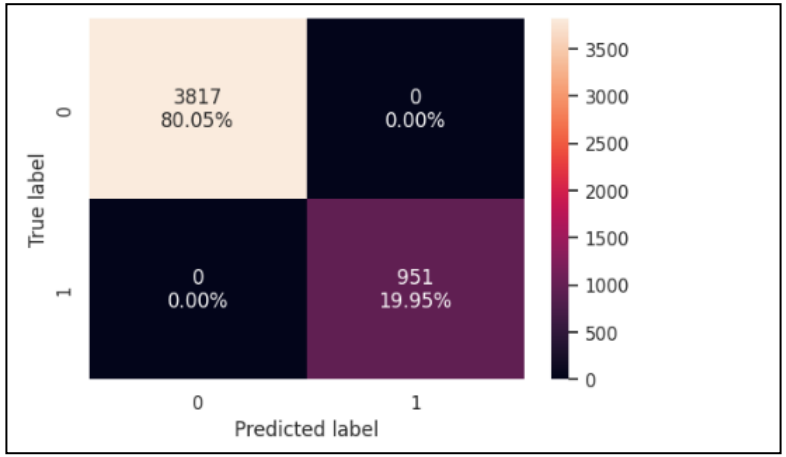
Observations / Findings

* Decision tree model with default parameters is overfitting the training data and is not able to generalize well.
* Pre-pruned tree has given a generalized performance with balanced values of precision and recall.
* Post-pruned tree is giving a high F1 score as compared to other models but the difference between precision and recall is high.
* The bank will be able to maintain a balance between resources and brand equity using the pre-pruned decision tree model.

Model Building & Evaluation

Random forest

Data preparation for random forest model



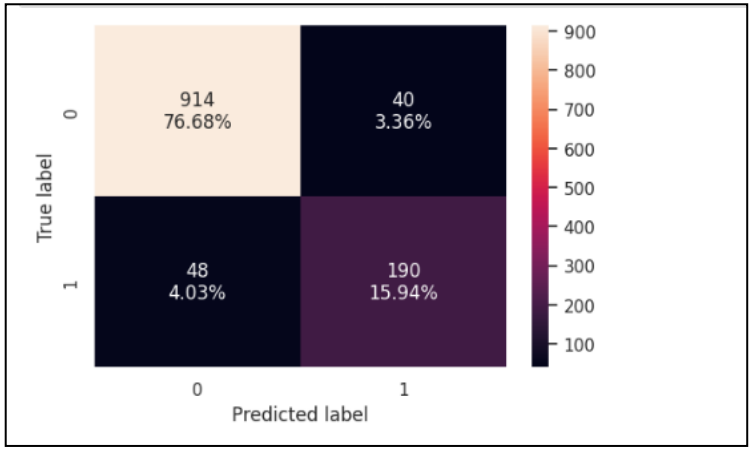
	Accuracy	Recall	Precision	F1
0	1.00000	1.00000	1.00000	1.00000

Observations / Findings

- Checking random forest model performance on the training set

Model Building & Evaluation

Checking random forest model performance on the testing set



	Accuracy	Recall	Precision	F1
0	0.92617	0.79832	0.82609	0.81197

Observations / Findings

Model Building & Evaluation

RANDOM FOREST

Building the model for bagging with Random Forest

Checking the training model performance for bagging with Random Forest

Training performance

	Accuracy	Recall	Precision	F1
0	1.00000	1.00000	1.00000	1.00000

Testing performance

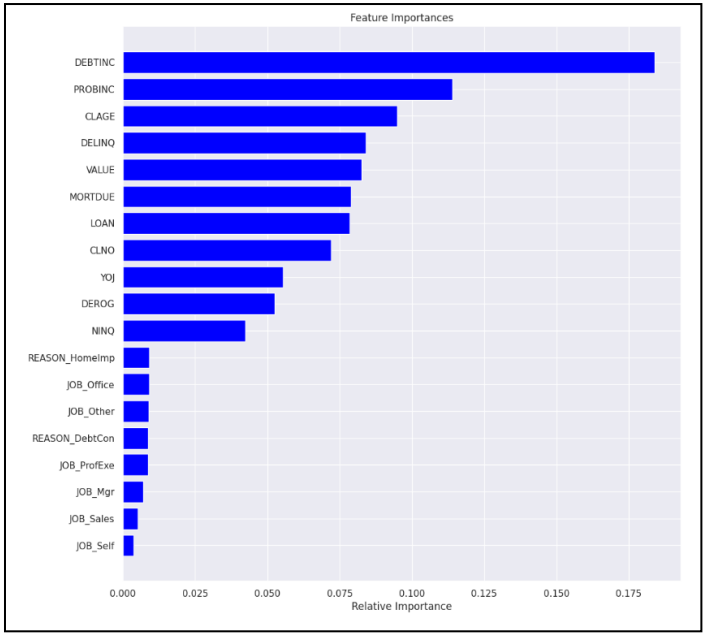
	Accuracy	Recall	Precision	F1
0	0.91163	0.68817	0.85906	0.76418

Observations / Findings

Model Building & Evaluation

Plotting the feature importance of each variable

	Imp
DEBTINC	0.18387
PROBINC	0.11402
CLAGE	0.09490
DELINQ	0.08407
VALUE	0.08254
MORTDUE	0.07885
LOAN	0.07853
CLNO	0.07212
YOJ	0.05547
DEROG	0.05260
NINQ	0.04228
REASON_HomeImp	0.00923
JOB_Office	0.00907
JOB_Other	0.00887
REASON_DebtCon	0.00868
JOB_ProfExe	0.00867
JOB_Mgr	0.00710
JOB_Sales	0.00529
JOB_Self	0.00383



Observations about outcomes of model bagging with Random Forest Classifier

Model Building & Evaluation

ADABOOST FOREST

Checking the training model performance for bagging with Adaboost

Training performance

	Accuracy	Recall	Precision	F1
0	0.89334	0.56671	0.83574	0.67542

Testing performance

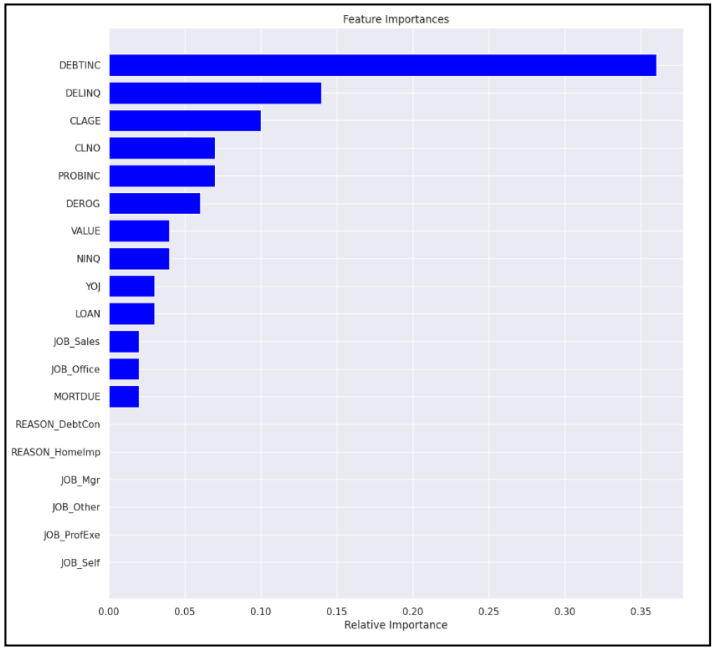
	Accuracy	Recall	Precision	F1
0	0.88647	0.54032	0.86266	0.66446

Observations / Findings

Model Building & Evaluation

Plotting the feature importance of each variable

	Imp
DEBTINC	0.36000
DELINQ	0.14000
CLAGE	0.10000
CLNO	0.07000
PROBINC	0.07000
DEROG	0.06000
VALUE	0.04000
NINQ	0.04000
LOAN	0.03000
YOJ	0.03000
MORTDUE	0.02000
JOB_Office	0.02000
JOB_Sales	0.02000
REASON_DebtCon	0.00000
REASON_HomeImp	0.00000
JOB_Mgr	0.00000
JOB_Other	0.00000
JOB_ProfExe	0.00000
JOB_Self	0.00000



Observations / Findings

Model Building & Evaluation

Gradient Boosting Classifier

Checking the training model performance for bagging with Gradient Boosting

```
Training performance
    Accuracy  Recall  Precision      F1
0   0.94535 0.78703    0.92253 0.84941
```

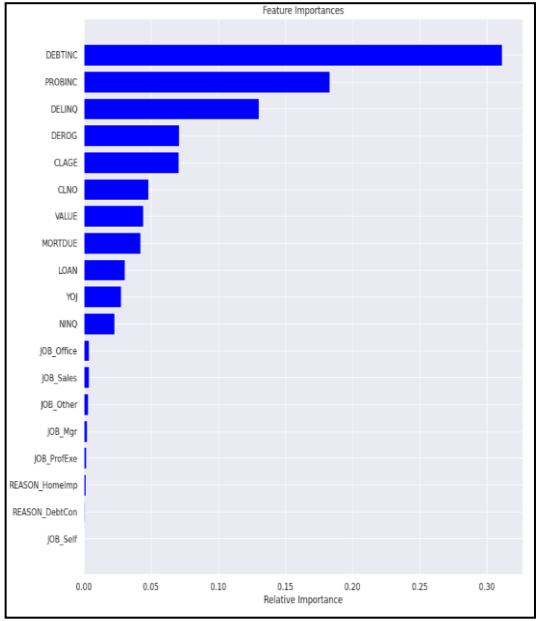
Checking the testing model performance for bagging with Gradient Boosting

```
Testing performance
    Accuracy  Recall  Precision      F1
0   0.90940 0.66667    0.86713 0.75380
```

Model Building & Evaluation

Plotting the feature importance of each variable

	Imp
DEBTINC	0.31110
PROBINC	0.18286
DELINQ	0.13044
DEROG	0.07112
CLAGE	0.07053
CLNO	0.04807
VALUE	0.04442
MORTDUE	0.04226
LOAN	0.03072
YOJ	0.02787
NINQ	0.02290
JOB_Office	0.00406
JOB_Sales	0.00405
JOB_Other	0.00308
JOB_Mgr	0.00263
JOB_ProfExe	0.00180
REASON_HomeImp	0.00152
REASON_DebtCon	0.00057
JOB_Self	0.00000



Model Building & Evaluation

XGBoost Classifier

Building the model for bagging with XG Boosting

```
Training performance
    Accuracy  Recall  Precision      F1
0   0.99832 0.99388    0.99754 0.99571
```

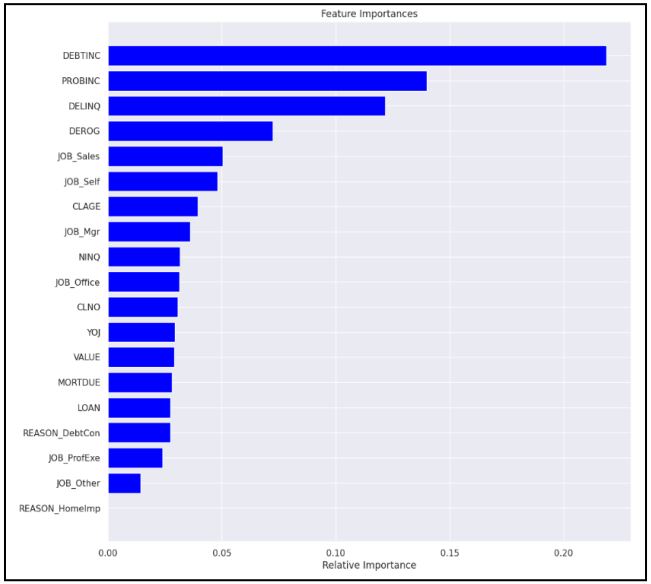
Checking the testing model performance for bagging with XG Boosting

```
 Testing performance
    Accuracy  Recall  Precision      F1
0   0.91387 0.69624    0.86333 0.77083
Testing performance
    Accuracy  Recall  Precision      F1
0   0.91387 0.69624    0.86333 0.77083
```

Model Building & Evaluation

Plotting the feature importance of each variable

	Imp
DEBTINC	0.21867
PROBINC	0.14010
DELINQ	0.12172
DEROG	0.07227
JOB_Sales	0.05035
JOB_Self	0.04820
CLAGE	0.03957
JOB_Mgr	0.03616
NINQ	0.03152
JOB_Office	0.03132
CLNO	0.03057
YOJ	0.02938
VALUE	0.02908
MORTDUE	0.02804
LOAN	0.02752
REASON_DebtCon	0.02740
JOB_ProfExe	0.02386
JOB_Other	0.01427
REASON_HomeImp	0.00000



Model Building & Evaluation

Comparing all models - Training

	Accuracy	Recall	Precision	F1
Logistic Regression-default Threshold	0.85163	0.35373	0.76053	0.48287
Logistic Regression-0.37 Threshold	0.76103	0.68543	0.43077	0.52905
Logistic Regression-0.42 Threshold	0.82814	0.53611	0.56443	0.54991
Decision Tree sklearn	1.00000	1.00000	1.00000	1.00000
Decision Tree (Pre-Pruning)	0.86290	0.79437	0.61633	0.69412
Decision Tree (Post-Pruning)	0.95901	1.00000	0.82692	0.90526
Support Vector Merchine	0.96099	0.92429	0.88520	0.90432
Random Forest (resampled)	1.00000	1.00000	1.00000	1.00000
Bagging	1.00000	1.00000	1.00000	1.00000
Ada Boost	0.89334	0.56671	0.83574	0.67542
Gradient Boost	0.94535	0.78703	0.92253	0.84941
XG Boost	0.99832	0.99388	0.99754	0.99571

Model Building & Evaluation

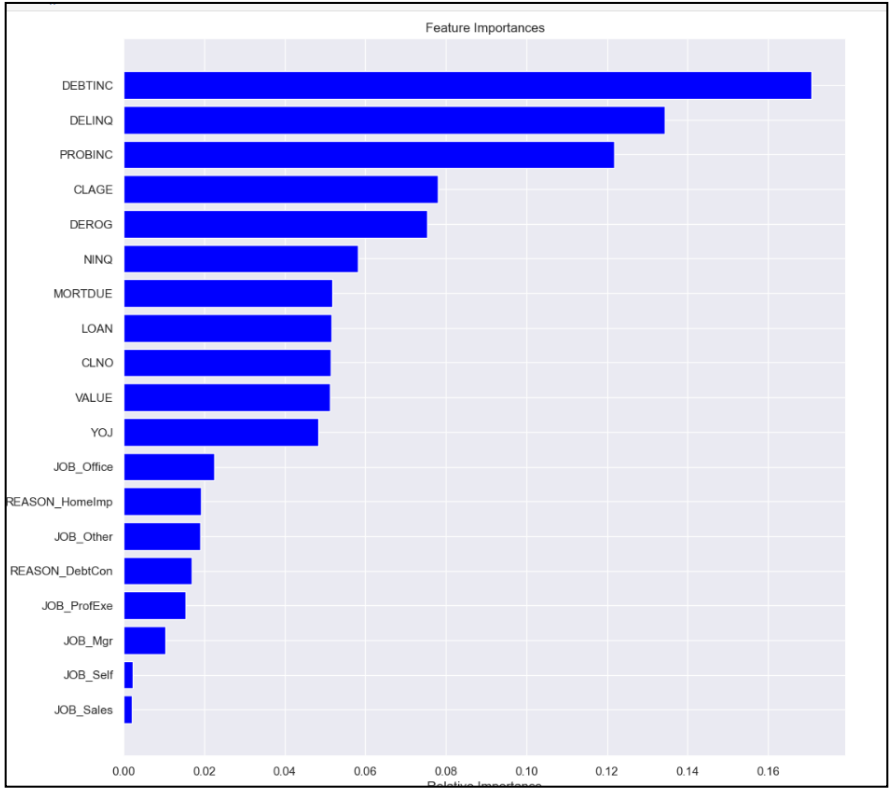
Comparing all models - Test

	Accuracy	Recall	Precision	F1
Logistic Regression-default Threshold	0.82998	0.29301	0.72667	0.41762
Logistic Regression-0.37 Threshold	0.75951	0.67742	0.44840	0.53961
Logistic Regression-0.42 Threshold	0.81823	0.48656	0.57460	0.52693
Decision Tree sklearn	0.86577	0.61290	0.70370	0.65517
Decision Tree (Pre-Pruning)	0.85235	0.74194	0.62162	0.67647
Decision Tree (Post-Pruning)	0.87528	0.75538	0.68039	0.71592
Support Vector Merchine	0.83305	0.40756	0.62581	0.49364
Random Forest (resampled)	0.91107	0.76471	0.78448	0.77447
Bagging	0.91163	0.68817	0.85906	0.76418
Ada Boost	0.88647	0.54032	0.86266	0.66446
Gradient Boost	0.90940	0.66667	0.86713	0.75380
XG Boost	0.91387	0.69624	0.86333	0.77083

Insight , Recommendation & Conclusion

Model Building & Evaluation

Feature Importance For our Best Model - Random Forest (Resampled)



Model Building & Evaluation

MODEL EVALUATION AND CONCLUSION

Model Evaluation Criterion

- The nature of predictions made by the models will translate as follows:
- True positives (TP) are defaults correctly predicted by the model.
- False negatives (FN) are defaulted clients in reality, which the model has predicted not-defaulted
- False positives (FP)are not-defaulted clients in reality, which the model has predicted, defaulted
- Model can make wrong predictions as:
- Predicting a default but in reality the client has not defaulted . Predicting a Not-defaulted but in reality the client has defaulted
- Which case is more important?
- If we predict a client will default and in actuality they do not, for banks this doesn't particularly hurt so much - no real financial loss.
- If, on the other hand, we predict the client will not default and an it does default in reality, this will be quite hurtful to the bank leading financial loss i.e loan write off, which directly impact bottom line
- To reduce this loss Recall should be maximized (Need to reduce False Negative), the greater the recall the higher the chances of identifying both the classes correctly.

Insight

1. Loan Approval and Defaults:

The dataset comprises 5,960 observations, with a 20% default rate.
The average approved loan amount is ~18,608, and the maximum is 89,900.
Clients with relatively high loan amounts seem to have repaid successfully.

2. Mortgage and Property Values:

Average due on existing mortgages is ~737,609, with a maximum of 855,909.
Higher current property values are associated with a higher default rate.

3. Debt-to-Income Ratio:
The average debt-to-income ratio is 33.77, within a favorable range.
Higher debt-to-income ratios are associated with a higher default rate.

4. Credit Lines and Enquiries:
The average number of existing credit lines is 21.
Higher numbers of derogatory remarks, delinquent credit lines, and credit inquiries are associated with a higher default rate.

5. Reasons for Loan and Job Types:
Debt consolidation is the most common reason for a loan, with the highest num Clients in the "Others" job category have the highest default rate.

Insight Cont'd

6. Distribution and Outliers:
Several variables are not normally distributed and exhibit right skewness, indicating the presence of outliers.

7. Correlations:
MORTDUE (amount due on existing mortgage) is highly correlated with VALUE (current property value).

8. Job Types and Loan Amounts: Self-employed clients have the highest average loan amount, while sales professionals have the least.

9. Debt-to-Income Ratio and Job Types:
Sales professionals have the highest average debt-to-income ratio.

10. Number of Credit Lines and Job Types:
Sales professionals have the highest average number of existing credit lines.
.

Conclusion

The Random Forest (resampled) model has demonstrated a strong recall of 76%, indicating its ability to correctly identify a significant proportion of clients who will not pay.

The overall model accuracy of 91% suggests that the model is performing well in terms of making correct predictions on the test data.

Key Features:

DELINQ (Number of Delinquent Credit Lines):

This feature is a crucial determinant, suggesting that clients with a history of delinquent credit lines are more likely to default on payments.

The company may want to pay close attention to clients with a high number of delinquent credit lines and consider additional risk assessment or targeted interventions for these individuals.

DEBTINC (Debt-to-Income Ratio):

The debt-to-income ratio is another important factor in predicting whether a client will pay or not.

Clients with a high debt-to-income ratio may be at higher risk, and the company may want to tailor its lending or credit policies for individuals with elevated ratios.

Conclusion

PROBINC (Current Debt on Mortgage/Debt-to-Income Ratio):

This feature provides insight into the relationship between current mortgage debt and the overall debt-to-income ratio.

Monitoring clients with high PROBINC values may help the company identify individuals with a potentially unsustainable level of mortgage debt relative to their income.

Business Recommendation 1

Risk Management:

Implement targeted risk management strategies for clients with a high number of delinquent credit lines (DELINQ).

Consider additional scrutiny or risk assessment for clients with elevated debt-to-income ratios (DEBTINC) to minimize the risk of default.

Policy Adjustments:

Consider adjusting lending or credit policies based on the insights from the model. This could involve setting different approval criteria for clients with varying levels of DELINQ, DEBTINC, and PROBINC.

Continuous Monitoring:

Regularly update and recalibrate the model based on new data to ensure it remains effective in predicting client payment behavior.

Continuously monitor the performance of the model and adjust business strategies accordingly.

Client Communication:

Communicate transparently with clients about the factors influencing credit decisions, especially those related to DELINQ, DEBTINC, and PROBINC.

Provide financial education and guidance to clients with higher risk profiles to help them manage their debt responsibly..

Business Recommendation 2

Focus on offering moderate to high loan amounts cautiously, considering the higher default rate.
Evaluate property values carefully, especially for clients seeking loans against higher property values.
While the average is favorable, scrutinize clients with higher debt-to-income ratios more thoroughly.
Assess clients with multiple derogatory remarks, delinquent credit lines, and credit inquiries more cautiously.
Investigate clients in the "Others" job category more rigorously. Consider offering targeted solutions for debt consolidation.
Employ robust statistical methods and outlier detection techniques during analysis to ensure accurate modeling
Consider this strong correlation when assessing a client's financial situation. It might indicate potential refinancing opportunities.
Tailor loan offerings based on the client's job type. Self-employed clients might require more customized solutions.
Provide financial counseling or advice to sales professionals to manage their debt effectively.
Monitor credit usage and provide guidance on managing multiple credit lines responsibly.

Thank You