NATIONAL ECONOMICS UNIVERSITY

Faculty of Mathematical Economics

ASSIGNMENT

Course: Risk analytics

Title: Default risk prediction in L&T vehicle loans

Instructor: Ms. NGUYỄN THỊ LIÊN

Name: NGUYỄN TUẤN DUY

Student ID: 11204971

Class: DSEB 62

-

Ha Noi, March 2023

I. INTRODUCTION:

This study aims to develop a credit risk prediction model from a sample of customers from L&T Finance Holding Ltd applying for vehicle loan. By estimating the probability that a customer will default can help the institution avoid significant losses, this has led to the increasing need for a good credit risk scoring model to rejects the applications with high credit risk. With the sampled customers, this study will develop three groups of model namely Linear algorithm (Logistic regression), Probabilistic classifier (Naïve Bayes) and Tree-based models (Decision Tree, Random Forest, CatBoost) to predict the probability a loanee defaulting on a vehicle loan. Finally, the quality and performance of these models are evaluated using methods such as K-fold validation, ROC-AUC scoring, confusion matrix to identify the best one.

II. THEORETICAL BASIS:

1. The definition of risk:

1.1 Vehicle loan:

A vehicle loan is understood as a form of trade where the loanee obtains funds to purchase a vehicle for personal use. Typically, the lender loans the money (or make direct payment to the dealer on the buyer's behalf) while the buyer must pay back the full loan amount along with some interest, which is a source of profit for the financial institution. In other words, car loan is one form of personal loan designed specifically for purchasing vehicles, however, there are still some differences between the two.

Vehicle loans	**Personal loans**
Always secured.	Can be secured or unsecured.
Strictly for purchasing a vehicle.	Can be used for many different purposes.
Don't necessarily need a great credit score.	The better the credit score, the higher loanee can borrow and lower interest rate.
Borrowing amount and interest rate dictated by the vehicle's price.	Dictated by credit score or secured asset's value

Because the vehicle purchased serves as collateral for the loan (should the loanee default on the repayment, the lender can seize the vehicle), much like a mortgage, the asset remains under the lender's ownership until the final payment is made, thus, vehicle loan is a secured loan and is deemed a lower risk and interest rate on this type of loan is significantly lowered in comparison to unsecured personal loan.

1.2 Credit risk of vehicle loan:

Credit risk, or default risk emerges when the borrower fails to pay back a loan in accordance with the terms of the credit agreement. In the context of vehicle loan, credit risk can arise when the collateral has been taken back (The vehicle is revoked to offset the remaining loan as a result of borrower not being able to repay the loan). Even though the vehicle loan is secured, which makes it less risky for the lender, vehicles such as cars or motorcycles are moveable properties and depreciates rapidly over time. Moreover, the quality of the vehicle heavily depends on the borrowers, problems such as repeated maintenance can generate the incentive for the borrower to default as the maintaining cost is too high or the improper use of the vehicle can cause its values to decline to far lower than the original. In those cases of default, the common approach one financial institution can take is to

auction the collateral but even then, its value has dropped significantly and the lender will suffer losses. On the other hand, the primary source of profit for the lender is the interest accumulated from the loan, rejecting the borrower that is willing to repay the loan can also result in the loss of profit. Therefore, by accurately estimate credit risk, the loan approval task can be optimized, and profit can be maximized by avoiding approving defaulter and rejecting non-defaulter.

Credit risk can be measured by Expected Loss (EL) is the average credit loss expected from an exposure and Unexpected Loss (UL), which represents the anticipated average loss could incur when exposed:

$$EL = PD \times LGD \times EAD$$

$$UL = EAD \times \sqrt{PD \times \sigma_{LGD}^2 + LGD^2 \times \sigma_{PD}^2}$$

PD: probability of default.

LGD: Loss given default.

EAD: exposure at default.

Models utilised in this study will attempt to find the best estimation for the probability of default given the data regarding the borrower.

2. The models:

2.1. Logistic regression:

The required output for our problem is to decide whether the borrower, given the data, will default on the loan or not, thus, this problem can be classified as a binary classification problem where the target value is 1 or 0 (default or not default). Logistic regression uses sigmoid function to estimate this probability:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

One important property of the sigmoid function to estimate probability is its value is always in the range between 0 and 1, as illustrated in the graph below:

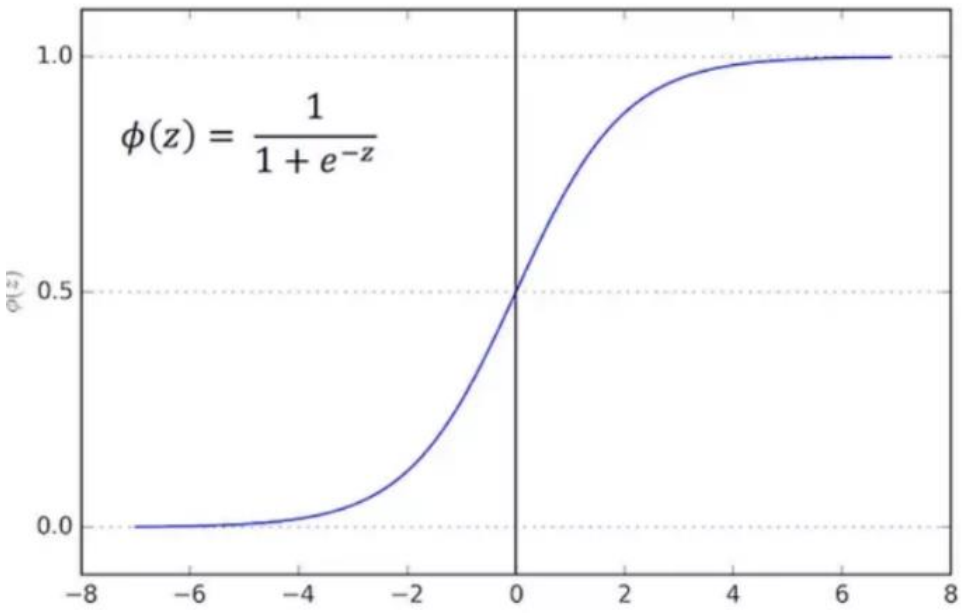


Figure 1: Distribution of sigmoid function

Given variables $(x_1, x_2, \dots, x_n)$ the predicted value for logistic regression is:

$$\hat{y}_i = \sigma(-\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_N x_N) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_N x_N)}}$$

For dataset $\phi_n = \phi(x_n)$ with $n = 1, \dots, N$, and target variable $t_n \in \{0,1\}$ The logistic regression is defined as:

$$p(C_1|\phi) = \hat{y}(\phi) = \sigma(\beta^T \phi)$$

$$p(C_2|\phi) = 1 - p(C_1|\phi)$$

The likelihood function can be written as:

$$p(t|\beta) = \prod_{n=1}^{N} y_n^{t_n} (1 - y_n)^{1-t_n}$$

By taking the negative logarithm of the likelihood, then, the loss function for logistic regression is given by:

$$L = -\frac{1}{N} \sum_{i=1}^{N} t_i \log(y_i) + (1 - t_i) \log(1 - y_i)$$

After minimizing loss function using gradient descent method, we arrive at a set of optimized parameters β, which will be used for prediction.

2.2. Naïve Bayes:

Based on the concept of conditional probability, Naïve Bayes is a probabilistic algorithm known for its simplicity and speed. Bayes' theorem states that, given class variable y and dependent feature vector x_1 through x_n:

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n|y)}{P(x_1, \dots, x_n)}$$

Bayes theorem assumes that conditional independence between every pair of features give the value of class variable, then:

$$P(x_i|y, x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x|y)$$

With this assumption, the Bayes theorem simplifies to:

$$P(y|x_1, \dots, x_n) = \frac{P(y)\ \prod_{i=1}^{n}\ P(x_i|y)}{P(x_1, \dots, x_n)}$$

Naïve Bayes algorithm will maximize this probability:

$$\hat{y} = argmax\ P(y|x_1, \dots, x_n)$$

Because $P(x_1, \dots, x_n)$ can be calculated from the data, it can be treated as a constant, the optimization problem then becomes:

$$\hat{y} = argmax\ P(y) \prod_{i=1}^{n} P(x_i|y)$$

This study implements the Bernoulli Naïve Bayes for binary target value, where:

$$P(x_i|y) = P(x_i = 1|y)x_i + (1 - P(x_i = 1|y))(1 - x_i)$$

One drawback of the Naïve Bayes model is that its underlying "naïve" assumption is that every pair of features is independent given the value of the class variable, which rarely occurs in real-life data and is often violated.

2.3. Tree-based models:

a) Decision tree:

Decision tree is a non-parametric supervised learning method used in classification and regression problems. This algorithm develops simple decision rules based on the data features. Consider one simple hypothetical sample decision tree about loan approval below:

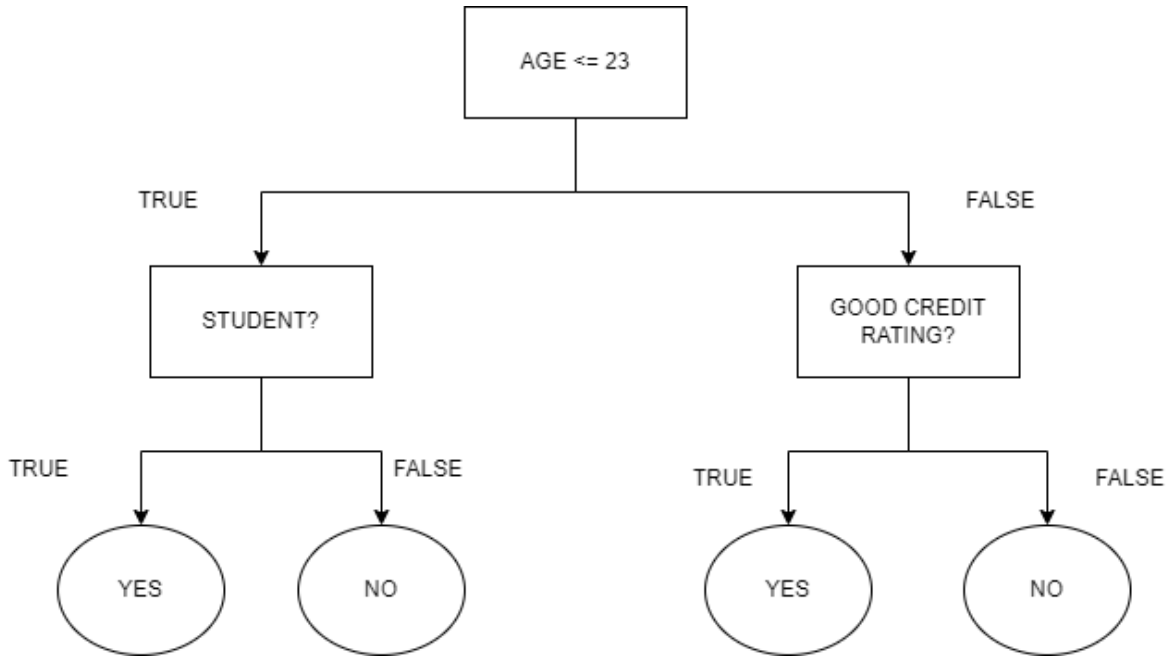


Figure 2: sample decision tree for loan approval problem

In this model, a loan approval decision is made using information about the customer. There are two kinds of nodes in the decision tree:

1. Condition nodes: These nodes have two child nodes (True or False)
2. Leaf nodes: These nodes don't have any child node containing the final decision at the end of the process.

To make a decision regarding a customer, algorithm will traverse from the root node (in this case, check if the applicant age is under 23) to its child nodes (if the applicant is above 23 years old, then check his credit rating, else check if he is a student) before eventually reach the leaf node and make the final decision.

To construct a decision tree, this study will choose condition nodes using Gini index:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

With C: number of target class.

$p_i = \frac{n_i}{N}$, n_i is the number of data points in i^{th} class and N is the total number of points in that node.

$$Gini\_index = gini(parent) - \sum_{k=1}^{K} \frac{m_k}{M} gini(child_k)$$

K: number of child nodes from parent.

M: number of datapoints in parent node.

m: number of datapoints in child node.

When splitting a node, the model tries to minimize the Gini in the child node (when Gini = 0 then all datapoints in the node belongs to one class) by maximize the Gini_index in the parent node.

b) Random forest:

In a random forest classification algorithm, multiple decision trees are created using different subsets of data by sampling with Bootstrapping method (random sampling with replacement). Using the sampled data, the model samples random features available to generate different trees. Predictions made by each tree are summarized and the final decision can be reached by taking the most popular result. Again, we consider the hypothetical example from Figure 2 but this time different subsets of data are sampled, and two more trees are generated:

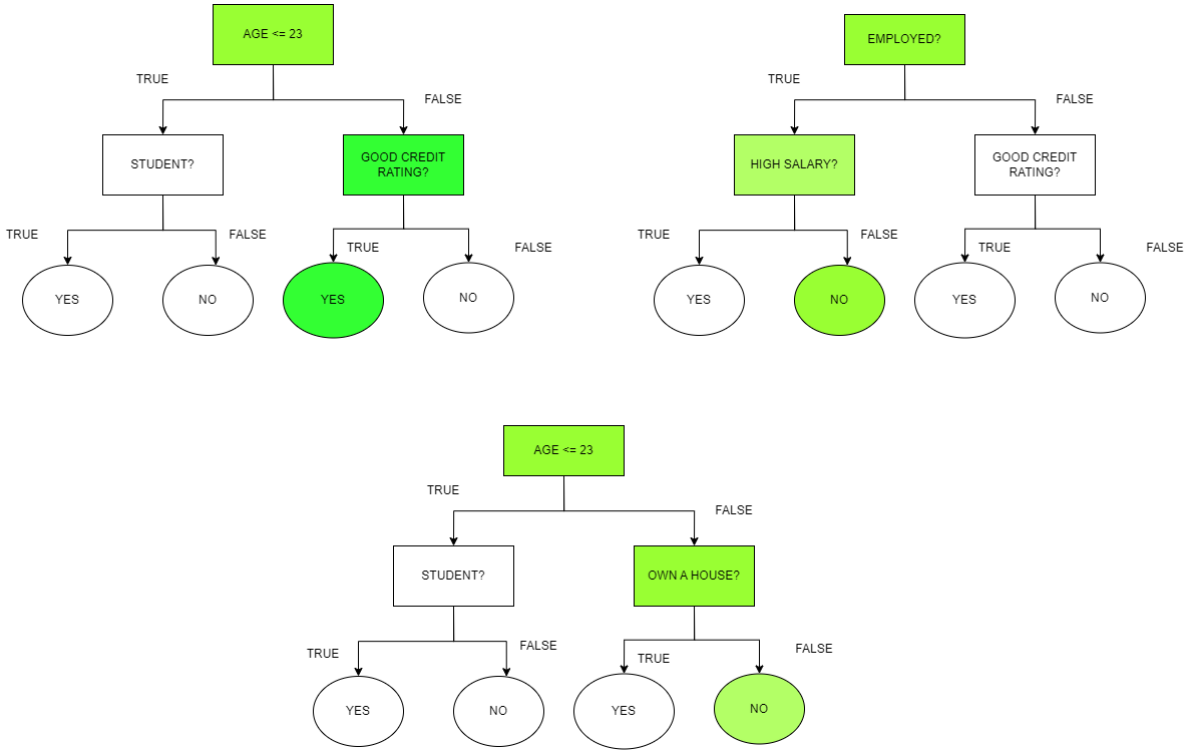


Figure 3: sample random forest for loan approval problem.

Each tree each exposed to different features and a different sample of dataset; thus, every tree can give different prediction. In this case two trees predicted the borrower will default and one says he won't, the algorithm ends with the conclusion to reject the loan request.

c) Gradient boosting:

Gradient boosting improves classification models like decision trees by providing an ensemble form of weaker prediction models (gradient boosted trees). The model tries to minimize the loss function:

$$\min_{c_n, w_n} L(y, W_{n-1} + c_n w_n)$$

Where: L: loss function

y: label.

c_n: confidence score of n^{th} weaker learner.

w_n: n^{th} weaker learner.

The gradient boosting algorithm can be summarized to the following step:

At iteration i:

1. Initialize model with equal pseudo-residuals for each data point.
2. Fit a weaker learner.
3. Compute confidence score c_i of trained model
4. Update model $W = W + c_i + w_i$
5. Recompute pseudo-residuals $-\eta \frac{\delta}{\delta w} L(W_{n-1})$

Repeat with iteration i + 1.

This study will apply CatBoost framework to implement gradient boosted tree.

3. Model evaluations:

3.1. K-fold validation:

K-fold validation is a technique to estimate the evaluation of a predictive model. The method split the training data into k consecutive folds, then fitting the data on the training sets except the selected fold, which will then be used for validation. Given a dataset S sampled from population D, the detailed steps for K-fold validation are:

1. Partition S into K equal disjoint subsets $(T_1, \dots, T_k)$
2. Perform K steps, at step k do:
    - Use $R_k = S - T_k$ as the training set.
    - Build classifier C_k using R_k
    - Use T_k as validation set for scoring (or compute error)
3. Summarize the validation scores.

The motivation for K-fold validation is that the generated model might not perform well on unseen data, by validating it on multiple subsets of unseen data, the model can be validated more accurately as under-fitting, over-fitting or well generalized. Generally, an effective model has high average score (low error) and low score variance between folds.

3.2. Confusion matrix analysis:

A confusion matrix can be used to analyse the prediction result of a classifier, for binary classifier like default prediction model, the confusion matrix can take the form:

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 4: sample confusion matrix.

Some indicators of the effectiveness of the model are:

- Sensitivity, recall or True Positive Rate:

$$TPR = \frac{TP}{TP + FN}$$

- Miss rate, or False Positive Rate:

$$FNR = \frac{FN}{TP + FN} = 1 - TPR$$

- Accuracy:

$$ACC = \frac{TP + TN}{TN + TP + FN + FP}$$

- F1 score:

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

3.3. ROC-AUC scoring:

The Receiver Operator Characteristics (ROC) curve is a probability curve that represents the true positive rate (TPR) and false positive rate (FPR) at various threshold values, and the area under the ROC curve (AUC) measures the entire two-dimensional area underneath the ROC curve. High ROC-AUC score indicates the model's high performance in distinguish between the positive and negative classes whereas low ROC-AUC the model incorrectly classifies the two classes.

III. Data:

Larsen & Toubro Ltd (L&T) is an Indian multinational conglomerate company, with business interests in engineering, construction, manufacturing, technology, information technology and financial services. To carry out this study, data about 233,154 car loan contracts of L&T finance was sampled with the need to develop an accurate credit risk prediction model and identify which borrowers are likely to default on the loan. Within the data set, the number borrowers defaulted their loans are 50,611 accounting for 22% of the total contracts recorded.

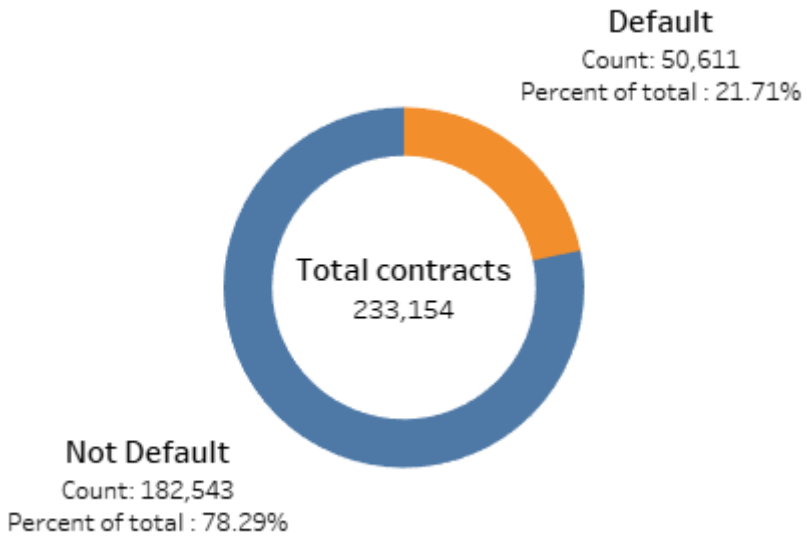


Figure 5: Distribution of default in the dataset

In the work, the sample is divided into two subsets: one containing 70% of the total number of contracts for the construction of the model (training set) and the remaining 30% to test the obtained

model. Additionally, K-fold validation technique with 5 folds is applied to validate the model before the test data is fitted.

3.1. The features:

The original data is in tabular form with 40 columns representing 39 different explanatory variables and one target columns. The variables can be divided into three groups:

1. Loanee's information (Demographic information such as age, driving license, identity proof,…)
2. Loan's information (Contract details like disbursal details, loan to value ratio,…)
3. Bureau data (Loanee's previous loans information like score, number of accounts, credit history,…)

The detailed description of the data is available in the appendix.

3.2. Data pre-processing:

Before training the models for prediction, several steps needed to be taken to the original data as it has several untidy features.

a) Handling missing value:

Using built-in function, there are 7661 missing values detected in the employment type column. In this case, these values are replaced by "Others" (the customers' employment may be unemployed or preferred not to reveal or the data is truly missing) values and treated as a separate category.

Other than employment feature, no other missing data has been found. However, some features may have missing values encoded as outliers and one of which is the Customer's birth date. It contains several values in the future with the furthest from now in the year of 2068, which leads to the problem of unrealistic negative values when calculating customer's age feature (by taking the difference of the customers' disbursal year (2018) to their birthday).

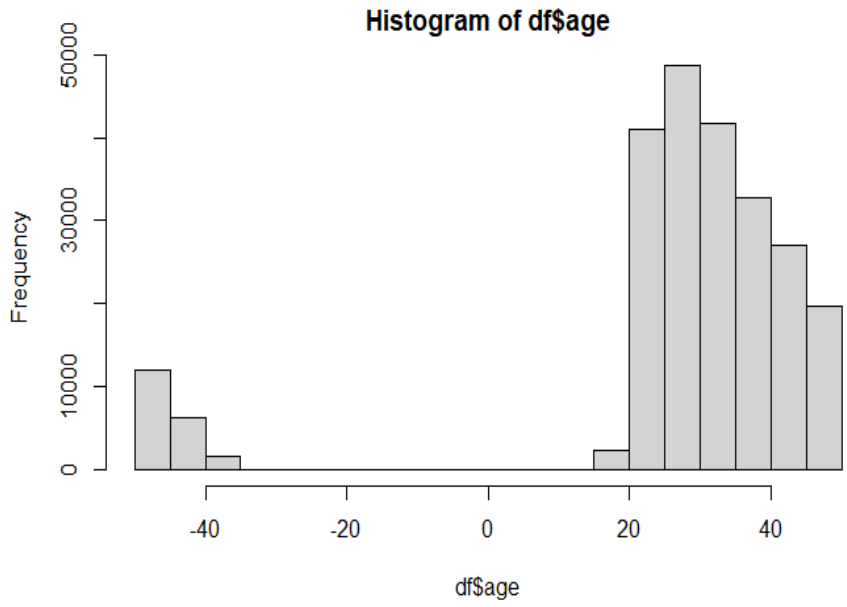


Figure 6: missing value of the age columns.

To deal with this problem, the chosen approach in this study is to drop the missing value, which led to around 8% of the original data imputed.

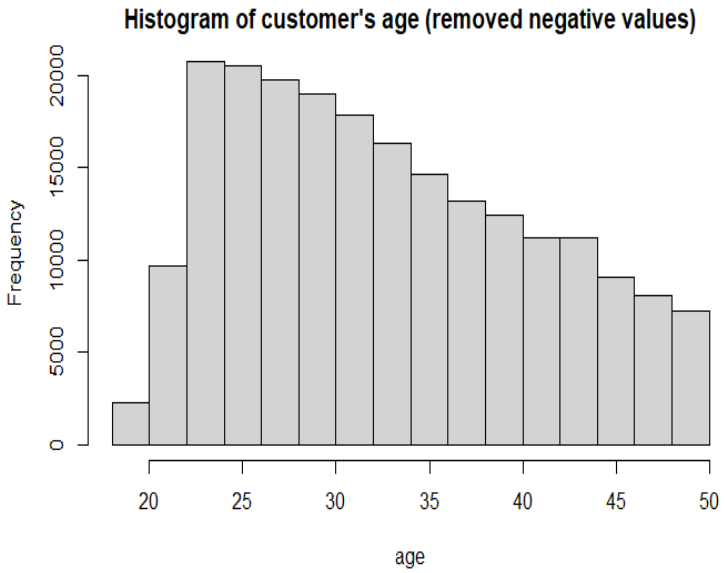


Figure 7: imputed age columns by removing missing values.

b) Treating outliers:

Some features in the datasets suffered from heavy outliers in the datasets, this study dealt with these values by implementing the Robust Scaler and rescale the data by removing the median and scales the data according to the interquartile range. The scaler's formula is as follows:

$$\frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)}$$

Using this method, we preserve the features of the datasets by making them robust to outliers.

3.3. Exploratory data analysis:

3.3.1. Loanee's information:

To decide whether to approve a loan request, it is important to investigate the applicants' personal information such as age, location and employment status as it tells us income level or the living standard of the applicants, which directly affect how likely they can repay the loan.

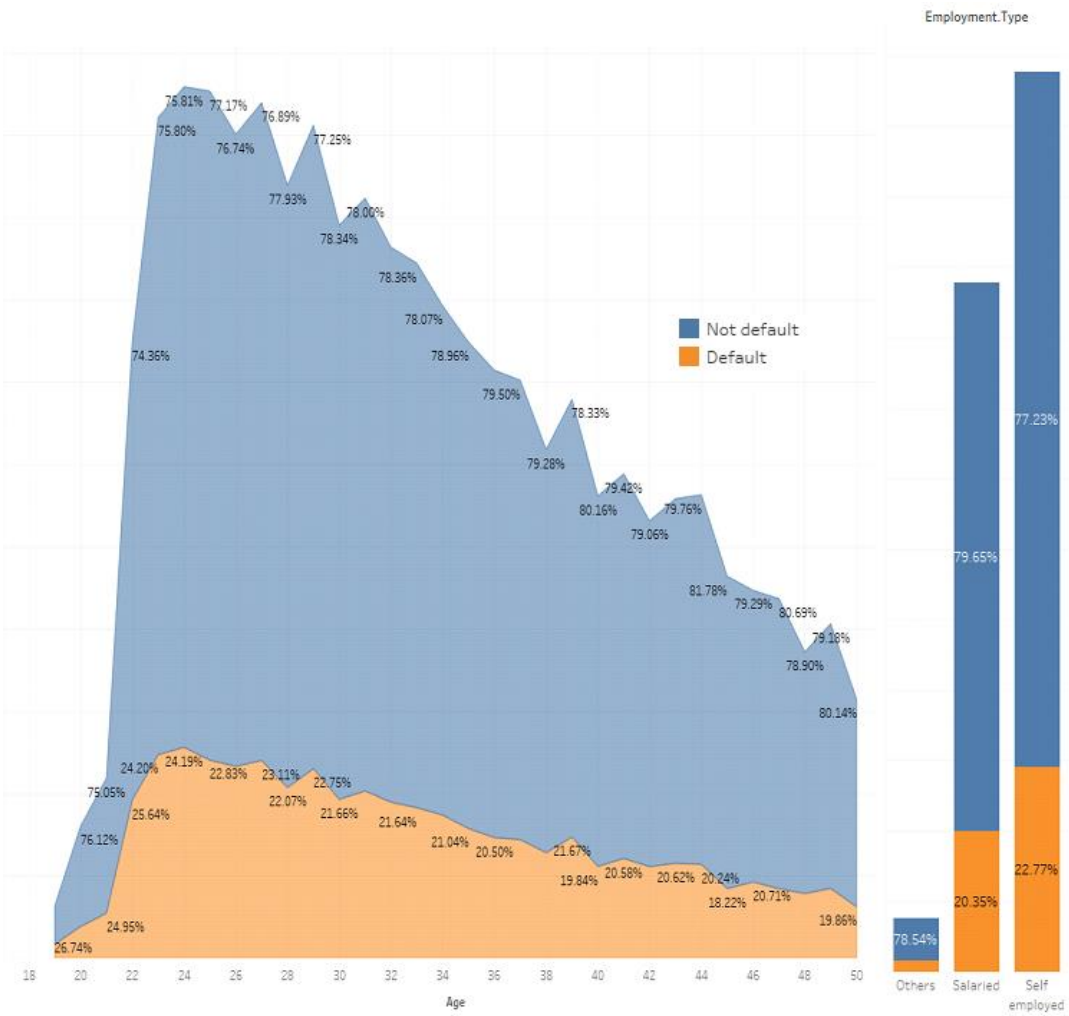


Figure 8: Default rate based on demographic information.

In this dataset, the main demographic for the vehicle loan in this institution was people from 22 to 30 years old, this age group also had high default rate as 25.64% of the 22 years old loanee defaulted. Employment can also be another factor affecting the probability of default as salaried workers was less likely to default.

3.3.2. Loan's information:

Looking at the contract details of the loans, there may be some indicators on the probability of default, one of which is the loan to value ratio (how much of the vehicle's value is covered by the loan). On average, defaulted loans had higher disbursed amount and used to cover a higher proportion of the asset's value.

	Loan Default	
	Non default	Default
Avg. Disbursed Amount	53,894	56,324
Avg. Ltv	74	77
Avg. Asset Cost	75,941	76,552

Figure 9: Average values related to the contracts.

3.3.3. Bureau data and history:

When examining the customers with previous credit history, customers having a low credit score have higher probability of default. Looking at the bar chart below, low credit risk groups (credit score from A to G) all have less than 20% of the loanee defaulted in the current loan whereas in medium and high risk groups (credit score from H to M), they are all above 24%.

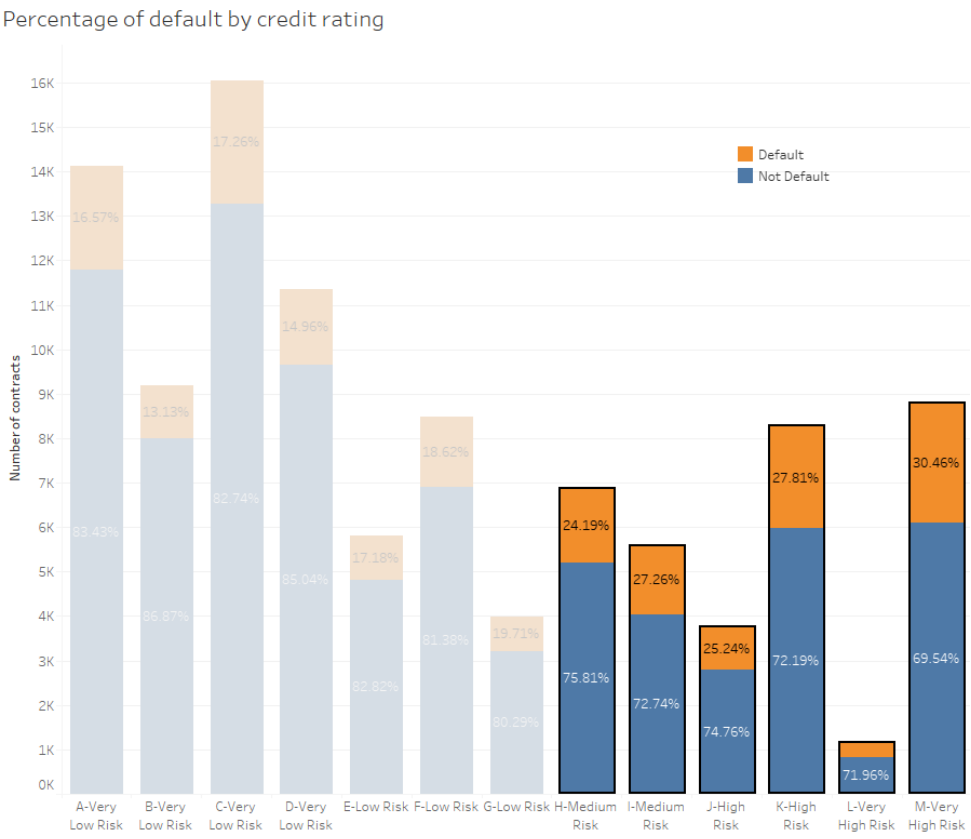


Figure 10: Percentage of default by credit rating

3.4. Features selection and features engineering:

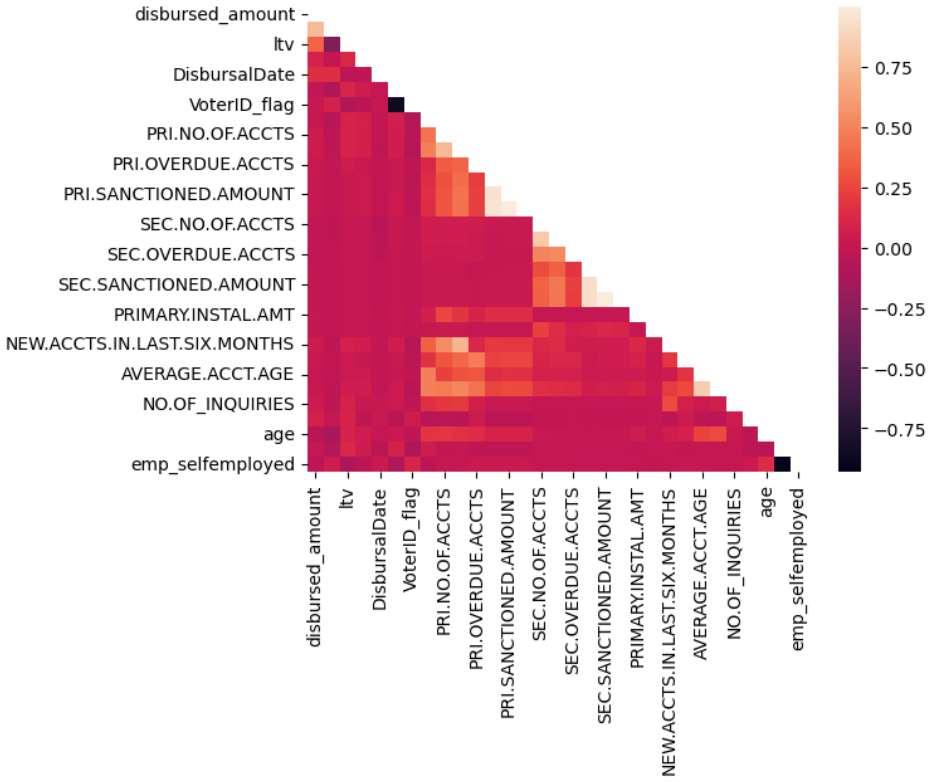


Figure 11: correlation matrix of the features.

Looking at the correlation matrix, features pairs with close meaning such as PRI.DISBURSED.AMOUNT and SEC. DISBURSED.AMOUNT was highly correlated with each other, by taking sum of the features with get the total amount for the contract.

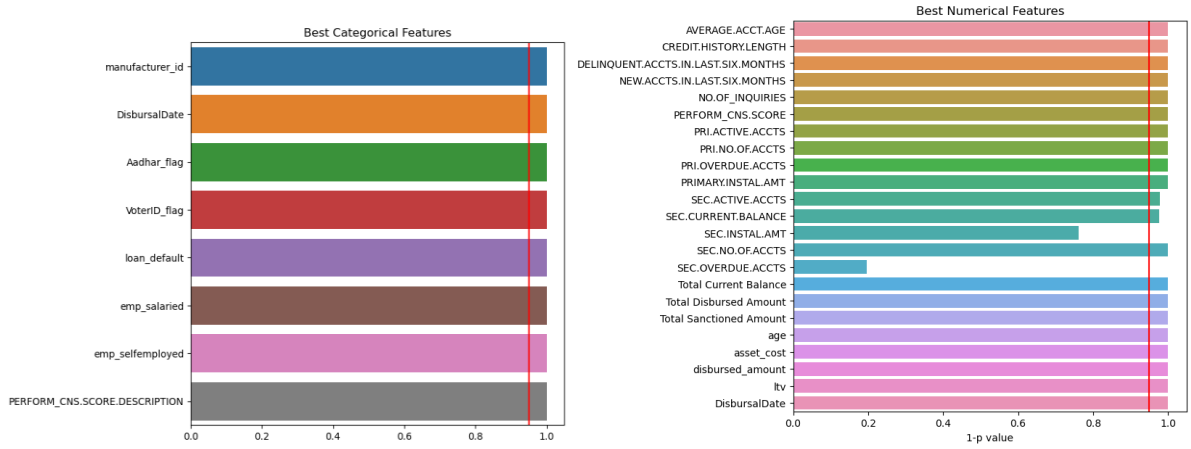


Figure 12: feature selection output

Using p-value of t-test for numerical features, and chi-square test for categorical features, two insignificant features were found in the dataset namely SEC.CURRENT.BALANCE and SEC.OVERDUE.ACCTS.

4. Findings and discussion:

4.1. In sample results:

4.1.1. Logistic regressions:

Logit Regression Results

Dep. Variable:	loan_default	No. Observations:	149301
Model:	Logit	Df Residuals:	149276
Method:	MLE	Df Model:	24
Date:	Sat, 18 Mar 2023	Pseudo R-squ.:	0.02786
Time:	21:09:36	Log-Likelihood:	-76382.
converged:	True	LL-Null:	-78571.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
disbursed_amount	4.916e-05	1.62e-06	30.259	0.000	4.6e-05	5.23e-05
asset_cost	-2.857e-05	1.04e-06	-27.434	0.000	-3.06e-05	-2.65e-05
ltv	-0.0092	0.001	-8.492	0.000	-0.011	-0.007
manufacturer_id	-0.0042	0.000	-14.448	0.000	-0.005	-0.004
DisbursalDate	-0.0131	0.007	-1.880	0.060	-0.027	0.001
Aadhar_flag	-0.2801	0.034	-8.335	0.000	-0.346	-0.214
VoterID_flag	0.0350	0.035	1.000	0.317	-0.034	0.104
PERFORM_CNS.SCORE	-0.0003	2.52e-05	-11.982	0.000	-0.000	-0.000
PRI.NO.OF.ACCTS	-0.0025	0.002	-1.175	0.240	-0.007	0.002
PRI.ACTIVE.ACCTS	-0.0464	0.008	-5.482	0.000	-0.063	-0.030
PRI.OVERDUE.ACCTS	0.2757	0.015	18.061	0.000	0.246	0.306
SEC.NO.OF.ACCTS	-0.0245	0.021	-1.146	0.252	-0.066	0.017
SEC.ACTIVE.ACCTS	0.0800	0.041	1.953	0.051	-0.000	0.160
...						
Total Current Balance	5.843e-08	4.05e-08	1.443	0.149	-2.09e-08	1.38e-07
Total Sanctioned Amount	-1.786e-07	1.27e-07	-1.407	0.159	-4.27e-07	7.02e-08
Total Disbursed Amount	1.576e-09	1.3e-07	0.012	0.990	-2.52e-07	2.55e-07

Figure 13: Logistic regression result (on unscaled train data).

For logistic regression, although this model did not perform as well as other models predictively, it is still important to analyse the output as it is interpretable in terms of multiplicative factors for the odds of a loanee's creditworthiness. In particular, the probability of default will change by a factor of e^{β_i} relative to feature i. A positive term indicates a higher odds of default, for example, when the number of primary overdue accounts (PRI.OVERDUE.ACCTS) in increases by one account (hold other factors constant will raise the probability of default by 32% ($e^{0.2757} - 1$). On the other hand, when the parameter is negative, it indicates a lower odd of default, for instance, the odds of default will decrease by 0.3% (relatively small because of the scale of the scoring system with the minimum of 0 and the maximum is 1000) when the credit score based on previous loans (`PERFORM_CNS.SCORE`) of the loanee increases by 1 point.

4.1.2. Predictions accuracy on train dataset:

Using K-fold validation using 5 folds, we can summarize the predictive capabilities of different models based on ROC-AUC score as follows:

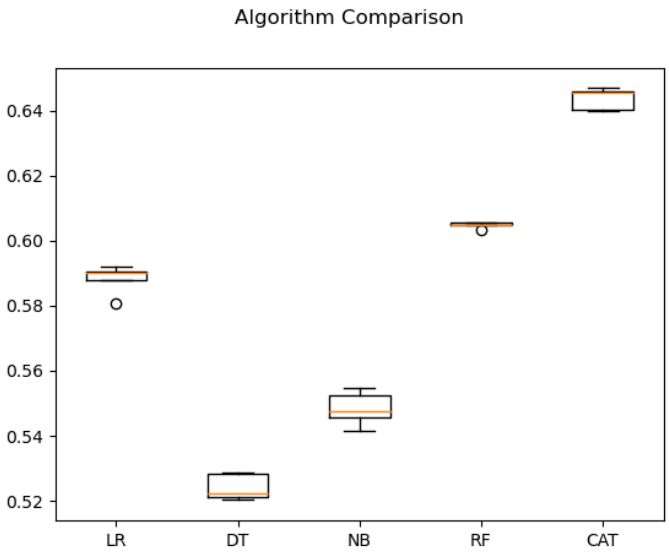
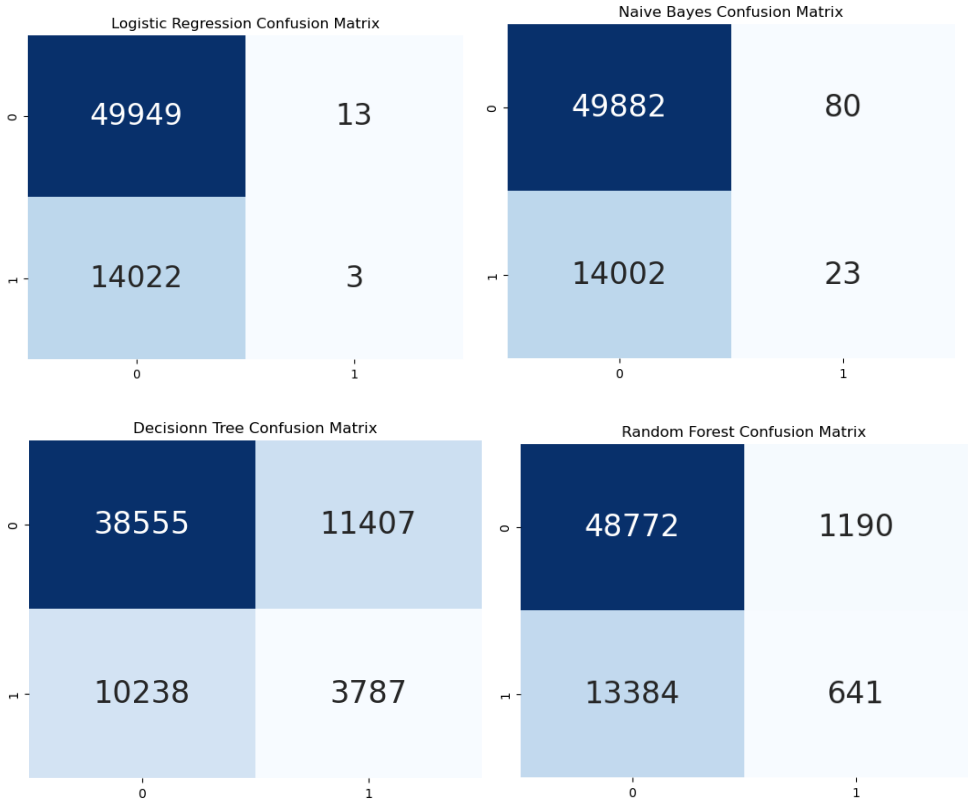


Figure 14: ROC AUC score comparison of models.

Tree-based models are the best predictors of default as random forest and catboosted forest gave the most accurate predictions on 5 folds. Even though Catboost model gives the best accuracy, random forest's accuracy did not have much variance between folds, thus it gives more stable predictions.

4.2. Out-sample results:

Training the models again on the remaining 30% of the dataset, the evaluations can be summarized and compared as follows:



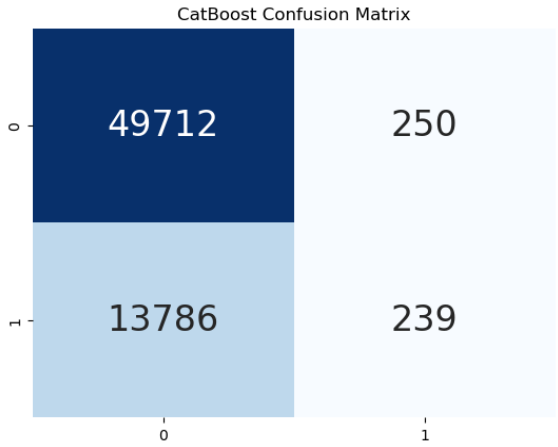


Figure 15: Confusion matrix of the models.

	Logistic	Naïve Bayes	Decision tree	Random forest	Catboost
Accuracy	78%	78%	66%	77%	78%
ROC-AUC	58%	54%	52%	60%	64%
Sensitivity (recall)	0%	0%	27%	4%	2%
F1-score	0%	0%	26%	8%	3%

Catboost and logistic regression have the greatest accuracy of 78% on the test set, when ROC-AUC score is considered Catboost is the best predictor overall. However, the model missed many of the positive case based on the low sensitivity score (2%). Additionally, 10 most important features summarized after fitting the Catboost model are:

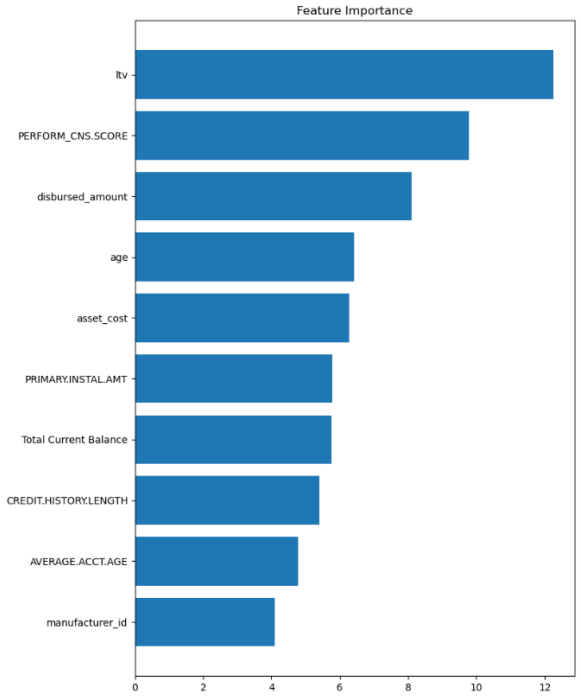


Figure 16: feature importance

4.3. Choosing the cut-off:

So far, the predictions of default is decided by determining the if the probability of default calculated by the models was higher or lower than 50%. However, the risk appetite for vehicle loan default risk is different for each financial institution, therefore, a strategy needs to be developed to determine the cut-off point that fits this appetite. One possible solution is the cumulative gain curve, for the Catboost model, the method can be applied and product the following graph:

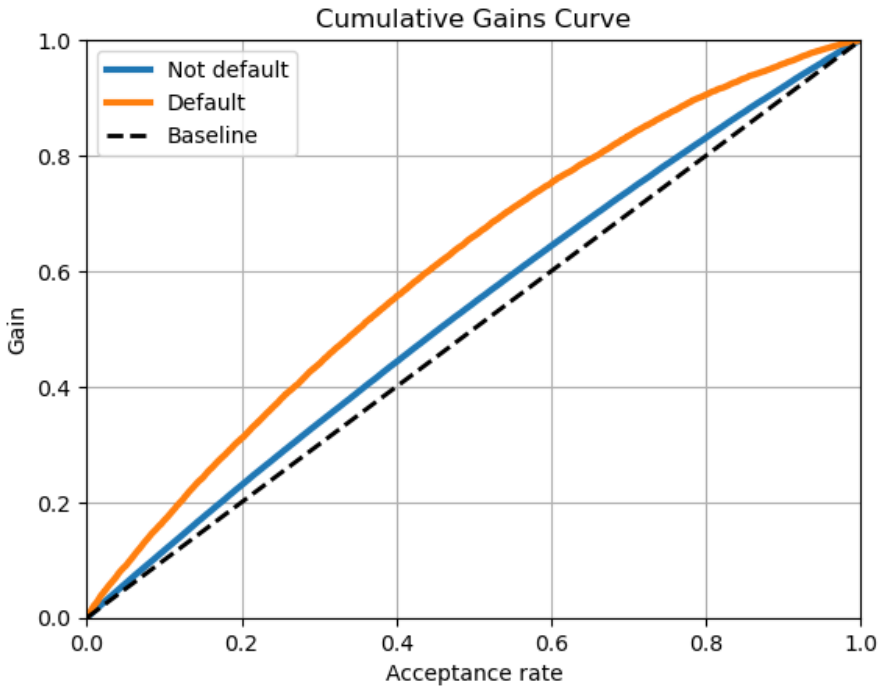


Figure 17: Cumulative gain curve for Catboost model.

By selecting the 50% acceptance rate (accepting 50% of the loan request), the models cover above 60% of the defaulters. When accepting 80% of the request is accepted (consider a request to be risky when the probability of default exceeds 20%), around 90% of the defaulter will likely be correctly predicted, which is a safer choice of cut-off as the exposure to default risk is less, but the potential profit also decreases. Another indication of a suitable cut-off point is the lift curve:

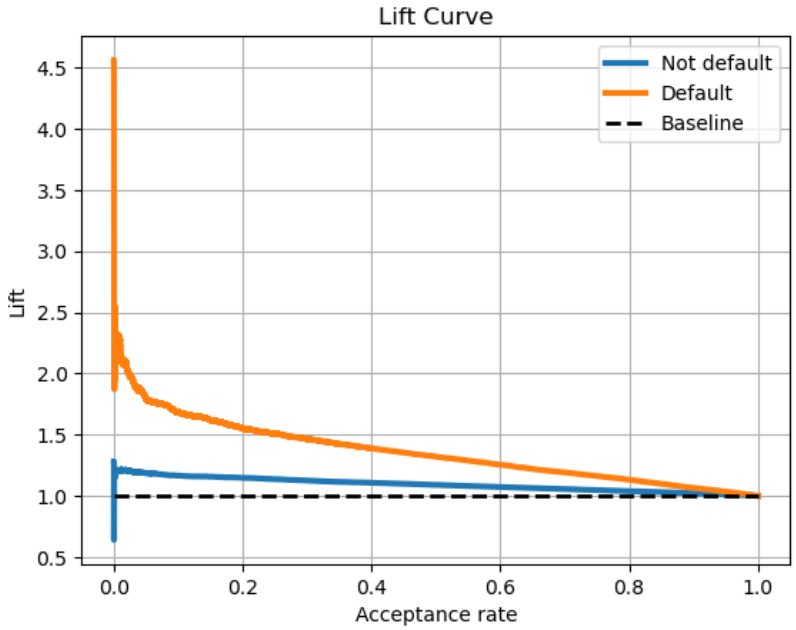


Figure 18: Lift curve for Catboost model.

Lift curve represents the gain percentage relative to the expected random result, which helps determining how effective a cut off point is. The closer the lines are to the baseline means there is little gain.

IV. CONCLUSION:

The study constructed 5 default prediction models for L&T vehicle loans using logistic regression, Naïve Bayes classifier, decision tree, random forest and gradient boosted random forest (Catboost). Through model evaluation techniques namely ROC-AUC scoring, confusion matrix analysis, Catboost model is the best method to classify the defaulters in the dataset. Although the Logistic regression did not perform as well as other tree-based models in prediction task, it is still useful to interpret the parameters to gain insight on how different features influence the probability of default. Loan to value ratio, loanee's bureau score and age are among the most important features. Furthermore, to match the model's predictions to the decision maker's risk appetite, it is important to consider indications in cumulative gain curve and lift curve to decide the suitable acceptance rate. Finally, to enhance the prediction capabilities of the models a few more steps can be taken like hyperparameter tuning (especially important for tree-based models as it tends to overfit on train data) or training different models such as Neural networks.

V. Appendix:

1. Data source: Kaggle.com
2. Code: Github
3. Data description:

	Variable Name	Description	
0	UniqueID	Identifier for customers	
1	loan_default	Payment default in the first EMI on due date	
2	disbursed_amount	Amount of Loan disbursed	
3	asset_cost	Cost of the Asset	
4	ltv	Loan to Value of the asset	
5	branch_id	Branch where the loan was disbursed	
6	supplier_id	Vehicle Dealer where the loan was disbursed	
7	manufacturer_id	Vehicle manufacturer (Hero, Honda, TVS etc.)	
8	Current_pincode	Current pin code of the customer	
9	Date.of.Birth	Date of birth of the customer	
10	Employment.Type	Employment Type of the customer (Salaried/Self Employed)	
11	DisbursalDate	Date of disbursement	
12	State_ID	State of disbursement	

13	Employee_code_ID	Employee of the organization who logged the disbursement	
14	MobileNo_Avl_Flag	if Mobile no. was shared by the customer, then flagged as 1	
15	Aadhar_flag	if Aadhar was shared by the customer then flagged as 1	
16	PAN_flag	if pan was shared by the customer, then flagged as 1	
17	VoterID_flag	if voter was shared by the customer then flagged as 1	
18	Driving_flag	if DL was shared by the customer, then flagged as 1	
19	Passport_flag	if passport was shared by the customer, then flagged as 1	
20	PERFORM_CNS.SCORE	Bureau Score	
21	PERFORM_CNS.SCORE.DESCRIPTION	Bureau score description	
22	PRI.NO.OF.ACCTS	count of total loans taken by the customer at the time of disbursement	Primary accounts are those which the customer has taken for his personal use
23	PRI.ACTIVE.ACCTS	count of active loans taken by the customer at the time of disbursement	
24	PRI.OVERDUE.ACCTS	count of default accounts at the time of disbursement	
25	PRI.CURRENT.BALANCE	total Principal outstanding amount of the active loans at the time of disbursement	
26	PRI.SANCTIONED.AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement	
27	PRI.DISBURSED.AMOUNT	total amount that was disbursed for all the loans at the time of disbursement	
28	SEC.NO.OF.ACCTS	count of total loans taken by the customer at the time of disbursement	Secondary accounts are those which the customer act as a co-applicant or guarantor
29	SEC.ACTIVE.ACCTS	count of active loans taken by the customer at the time of disbursement	
30	SEC.OVERDUE.ACCTS	count of default accounts at the time of disbursement	
31	SEC.CURRENT.BALANCE	total Principal outstanding amount of the active loans at the time of disbursement	
32	SEC.SANCTIONED.AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement	
33	SEC.DISBURSED.AMOUNT	total amount that was disbursed for all the loans at the time of disbursement	
34	PRIMARY.INSTAL.AMT	EMI Amount of the primary loan	
35	SEC.INSTAL.AMT	EMI Amount of the secondary loan	
36	NEW.ACCTS.IN.LAST.SIX.MONTHS	New loans taken by the customer in last 6 months before the disbursement	
37	DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS	Loans defaulted in the last 6 months	
38	AVERAGE.ACCT.AGE	Average loan tenure	
39	CREDIT.HISTORY.LENGTH	Time since first loan	
40	NO.OF_INQUIRIES	Enquiries done by the customer for loans	

VI. Reference:

[1] Ali Al-Aradi (2014). Credit Scoring via Logistic Regression

[2] Nazeeh Ghatasheh (2014). Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study

[3] Zhengyuan Zhang, Zhanquan Wang (2022). Research on Credit Scoring Based on Transformer-CatBoost Network Structure