DSC 478 Project Proposal

Team Members: Serena Yang, Zhong Hua Xie, Jason Kates

Overview:

Nowadays, buying a car with a loan is a prevalent trend for most people; however, this will lead to a big challenge for financial institutions. Suppose the financial institutions' vehicle loans forecasting model is not good enough. In that case, it will most likely bring losses to the financial institutions since some of them may not be able to pay their loans. For this project, our team will apply some algorithms based on the users' background information, previous loan and credit history to make the best model and predict the test dataset to get prediction accuracy. We will mainly use Python for this project.

Data Schema and Size:

The dataset that we used was downloaded from the Kaggle website. Before we did data cleaning, the dataset contained 41 variables as shown below, including numeric and categorical, and 233k+ rows.

Numeric	🟦	Categorical	🟨	Date	🟩	Binary	🟥
Variable Name	Description	Type	Data clean	Variable Name	Description	Type	Data clean
UniqueID	Identifier for customers	🟨 C	Delete Column	PERFORM_CNS.SCORE.DESCRIPTION	Bureau score description	🟨	Deleterow
disbursed_amount	Amount of Loan disbursed	🟦 N	mean	PRI.NO.OF.ACCTS	count of total loans taken by the customer at the time of disbursement	🟦	mean
asset_cost	Cost of the asset	🟦	mean	PRI.ACTIVE.ACCTS	count of active loans taken by the customer at the time of disbursement	🟦	mean
ltv	Loan to Value of the asset	🟦	mean	PRI.OVERDUE.ACCTS	count of default accounts at the time of disbursement	🟦	mean
branch_id	Branch where the loan was disbursed	🟦	mean	PRI.CURRENT.BALANCE	total Principal outstanding amount of the active loans at the time of disbursement	🟦	mean
supplier_id	Vehicle Dealer where the loan was disbursed	🟦	mean	PRI.SANCTIONED.AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement	🟦	mean
manufacturer_id	Vehicle manufacturer(Hero, Honda, TVS etc.)	🟦	mean	PRI.DISBURSED.AMOUNT	total amount that was disbursed for all the loans at the time of disbursement	🟦	mean
Current_pincode	Current pincode of the customer	🟦	mean	SEC.NO.OF.ACCTS	Count of total loans taken by the customer at the time of disbursement	🟦	mean
Date.of.Birth	Date of birth of the customer	🟩 D	Delete row	SEC.ACTIVE.ACCTS	count of active loans taken by the customer at the time of disbursement	🟦	mean
Employment Type	(Salaried/Self Employed)	🟨	Delete row	SEC.OVERDUE.ACCTS	count of default accounts at the time of disbursement	🟦	mean
DisbursalDate	Date of disbursement	🟩	Delete row	SEC.CURRENT.BALANCE	total Principal outstanding amount of the active loans at the time of disbursement	🟦	mean
State_ID	State of disbursement	🟦	Delete Column	SEC.SANCTIONED.AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement	🟦	mean
Employee_code_ID	Employee of the organization who logged the disbursement	🟨		SEC.DISBURSED.AMOUNT	total amount that was disbursed for all the loans at the time of disbursement	🟦	mean
MobileNo_Avl_Flag	if Mobile no. was shared by the customer then flagged as 1	🟥 B	mean	PRIMARY.INSTAL.AMT	EMI Amount of the primary loan	🟦	mean
Aadhar_flag	If aadhar was shared by the customer then flagged as 1	🟥	mean	SEC.INSTAL.AMT	EMI Amount of the secondary loan	🟦	mean
PAN_flag	if pan was shared by the customer then flagged as 1	🟥	mean	NEW.ACCTS.IN.LAST.SIX.MONTHS	New loans taken by the customer in last 6 months before the disbursement	🟦	mean
VoterID_flag	If voter was shared by the customer then flagged as q	🟥	mean	DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS	Loans defaulted in the last 6 months	🟦	mean
Driving_flag	if DL was shared by the customer then flagged as 1	🟥	mean	AVERAGE.ACCT.AGE	Average loan tenure	🟨	Deleterow
Passport_flag	if passport was shared by the customer then flagged as 1	🟥	mean	CREDIT.HISTORY.LENGTH	Time since first loan	🟨	Delete row
Perform_CNS.Score	Bureau Score	🟦	mean	NO.OF_INQUIRIES	Enquries done by the customer for loans	🟦	mean
				Loan_Default	Whether the user got laon	🟥	Delete row

For those missing values, we will delete the rows of missing categorical and date variables; fill in the mean value of those who got loan/didn't get for missing numerical and binary variables. After we did the data cleaning, there are about 225k+ rows left. We separated this dataset into two parts: 80% as the training dataset and 20% as the test dataset.

Analysis Approach:

Our analysis approach is to conduct as many classification models as possible to predict vehicle loan default based on loaner's information(attributes). Then compare the models using the ROC curve to determine which model is ideal for vehicle loan default prediction.

The list of classifiers we will be used for this project:

- Decision tree
- Logistic Regression
- KNN
- Naive Bayes
- PCA/Factor Analysis

Plan for Evaluation – Analysis of Results/Discussion:

- Compare the different models classification metrics for performance and accuracy
    - Also include confusion matrix for each model
- Graphically represent the models with Hyperparameter tuning
- Compare the models with the ROC curve

Plan Work Distribution:

For the project, West will be assigned to format, clean, and split the dataset into train and test datasets. Normalize numeric attribute into a set range, remove or fill out null entries, convert categorical attributes into dummy variables, and split the dataset by 80 and 20 for training and testing dataset.

After West did data preparation, each of the team members will do the classifiers assigned as below:

- West – Logistic Regression
- Jason – PCA/Factor Analysis
- Jason/West - Decision Tree
- Serena – Naive Bayes
- Serena – KNN