Vehicle Loan Default Prediction

How Predictive Modeling reduces financial losses from defaults

The problem

Problem statement

To reduce financial loss from vehicle loan defaults, we want to accurately predict (at least 75%) borrowers who will make their first equal monthly installment (EMI) on time and those who won't. We also want to identify the top 5 most important features for prediction to assist our underwriting team.

Company

Generalized vehicle loan company with access to large database.

Needs additional models that could assist the credit and underwriting team to help identify potentially risky loan applicants

Relevant Parties

- Underwriting Managers
- R&D
- Financial Analysts
- C-Suite Leadership

Challenges deep-dive

Challenge 1

Useful Visualizations

Building visualizations of our dataset that can be useful for underwriting purposes. The interactions between features are complex.

Challenge 2

Accuracy vs Overfitting

For the prediction to be generalizable to *new* data, we tune parameters to avoid overfitting, which could mean lower accuracy score.

Exploratory Data Analysis

Our Data

40 FEATURES

- ***Borrower Information*:** loan to value ratio, loan disbursed amount, credit score, outstanding loan amounts at time of disbursement, date of birth, employment type etc.

- ***Other*:** branch, manufacturer, state

PREDICTOR VARIABLE

- ***Loan_default:*** value of 0 if borrower pays the first equal monthly installment on time (non defaulter), 1 if the borrower does not (defaulter)

Our Data

Variable Name	Description
UniqueID	Identifier for customers
loan_default	Payment default in the first EMI on due date
disbursed_amount	Amount of Loan disbursed
asset_cost	Cost of the Asset
ltv	Loan to Value of the asset
branch_id	Branch where the loan was disbursed
supplier_id	Vehicle Dealer where the loan was disbursed
manufacturer_id	Vehicle manufacturer(Hero, Honda, TVS etc.)
Current_pincode	Current pincode of the customer
Date.of.Birth	Date of birth of the customer
Employment.Type	Employment Type of the customer (Salaried/Self Employed)
DisbursalDate	Date of disbursement
State_ID	State of disbursement
Employee_code_ID	Employee of the organization who logged the disbursement
MobileNo_Avl_Flag	if Mobile no. was shared by the customer then flagged as 1
Aadhar_flag	if aadhar was shared by the customer then flagged as 1
PAN_flag	if pan was shared by the customer then flagged as 1
VoterID_flag	if voter was shared by the customer then flagged as 1
Driving_flag	if DL was shared by the customer then flagged as 1
Passport_flag	if passport was shared by the customer then flagged as 1
PERFORM_CNS.SCORE	Bureau Score
PERFORM_CNS.SCORE.DESCRIPTION	Bureau score description
PRI.NO.OF.ACCTS	count of total loans taken by the customer at the time of disbursement

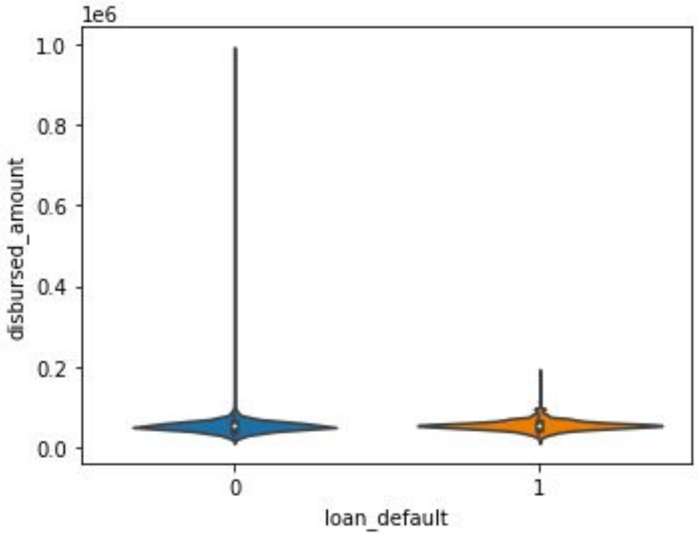
PRI.ACTIVE.ACCTS	count of active loans taken by the customer at the time of disbursement
PRI.OVERDUE.ACCTS	count of default accounts at the time of disbursement
PRI.CURRENT.BALANCE	total Principal outstanding amount of the active loans at the time of disbursement
PRI.SANCTIONED.AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement
PRI.DISBURSED.AMOUNT	total amount that was disbursed for all the loans at the time of disbursement
SEC.NO.OF.ACCTS	count of total loans taken by the customer at the time of disbursement
SEC.ACTIVE.ACCTS	count of active loans taken by the customer at the time of disbursement
SEC.OVERDUE.ACCTS	count of default accounts at the time of disbursement
SEC.CURRENT.BALANCE	total Principal outstanding amount of the active loans at the time of disbursement
SEC.SANCTIONED.AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement
SEC.DISBURSED.AMOUNT	total amount that was disbursed for all the loans at the time of disbursement
PRIMARY.INSTAL.AMT	EMI Amount of the primary loan
SEC.INSTAL.AMT	EMI Amount of the secondary loan
NEW.ACCTS.IN.LAST.SIX.MONTHS	New loans taken by the customer in last 6 months before the disbursment
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS	Loans defaulted in the last 6 months
AVERAGE.ACCT.AGE	Average loan tenure
CREDIT.HISTORY.LENGTH	Time since first loan
NO.OF_INQUIRIES	Enquries done by the customer for loans

Data Analysis Tools

Difference of Mean Independent t-test: This is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values. If the provided p-value < 0.05, we reject the null hypothesis of equal averages. ***Use this test to see if the averages of a feature between the defaulter and non-defaulter classes are different i.e. p-value < 0.05.***

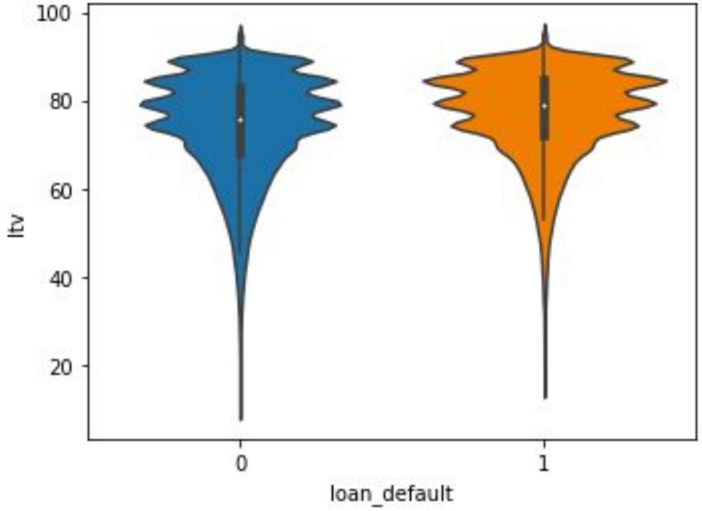
Chi-square Distribution: The chi-square test tests the null hypothesis that the categorical data has the given frequencies in our case, if the non-defaulter frequencies for a categorical feature matches the defaulter frequencies for the same feature.***If p-value > 0.05, we conclude that the frequencies do not match.***

Data Visualization for Numerical Features: *disbursed_amount*



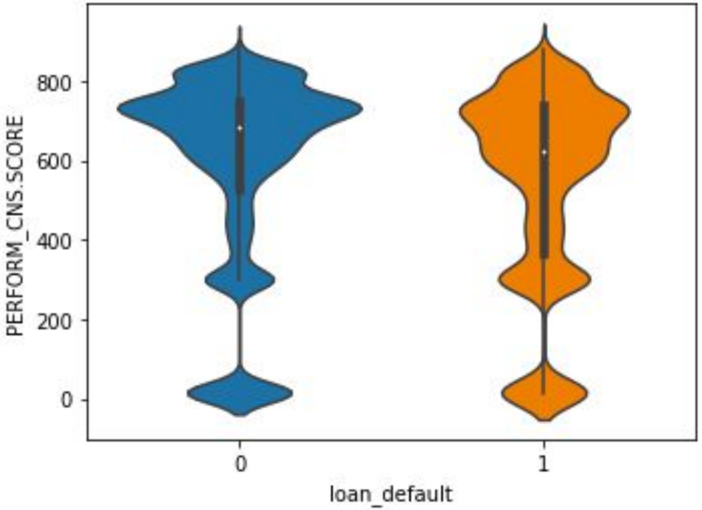

```
Nondefaulter mean for disbursed_amount : 53826.47111091633 , std: 13140.699007454747
Defaulter mean for  disbursed_amount :  56270.47386931695 , std: 12150.255527172361
Ttest_indResult(statistic=39.32291321262725, pvalue=0.0)
```

Data Visualization for Numerical Features: *ltv*



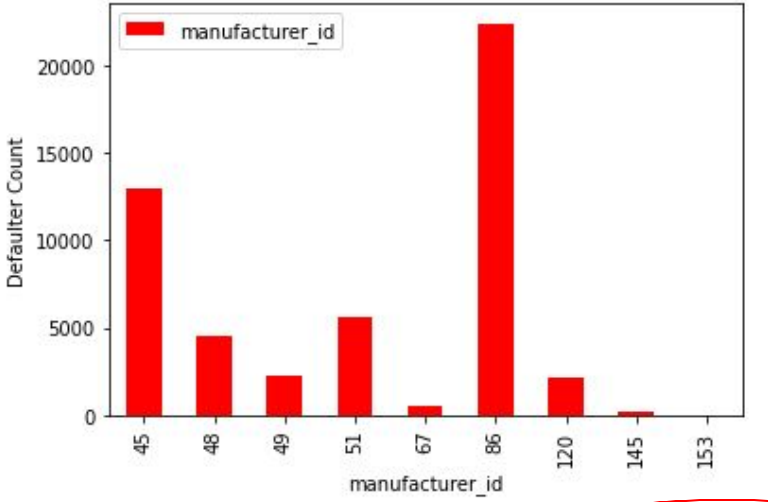

```
Nondefaulter mean for ltv : 74.15409333691578 , std: 11.681454560472389
Defaulter mean for  ltv :  76.88332180751246 , std: 10.327771446422924
Ttest_indResult(statistic=51.07804645618371, pvalue=0.0)
```

Data Visualization for Numerical Features: *PERFORM_CNS.SCORE*

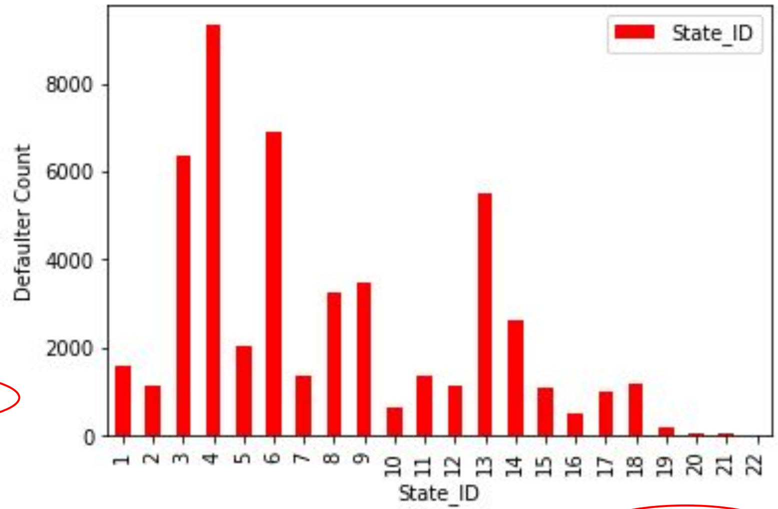



```
Nondefaulter mean for PERFORM_CNS.SCORE : 590.6796912947271 , std: 244.5933876500854
Defaulter mean for  PERFORM_CNS.SCORE :  541.8708349250817 , std: 247.84156027494095
Ttest_indResult(statistic=-27.06155806056678, pvalue=1.1001651810636842e-159)
```

Data Visualization for Categorical Features

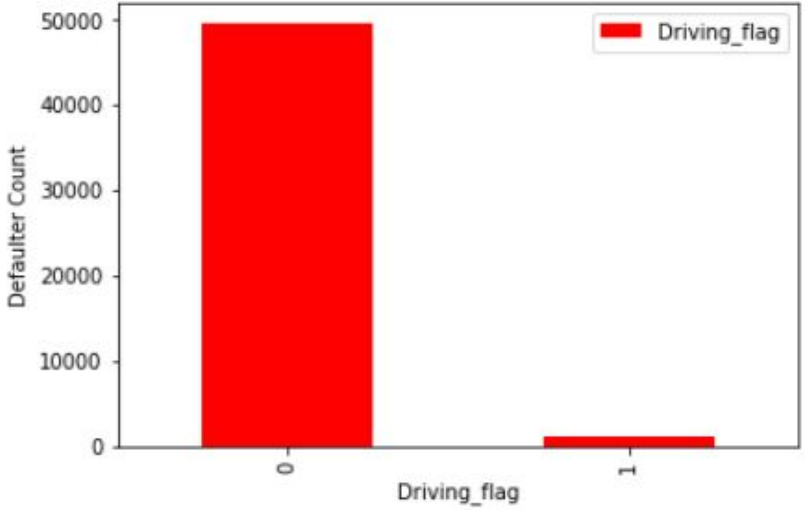


statistic=0.01227498642992129, pvalue=0.9999999999999278

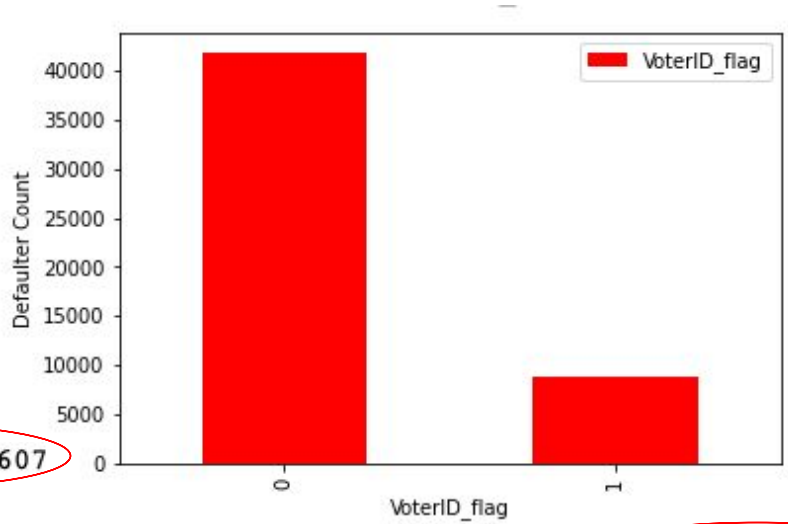


statistic=0.04404341953698229, pvalue=1.0

Data Visualization for Categorical Features



statistic=0.0001956158430401283, pvalue=0.9888409322237607



statistic=0.0118161806370888, pvalue=0.9134386517838328

Models

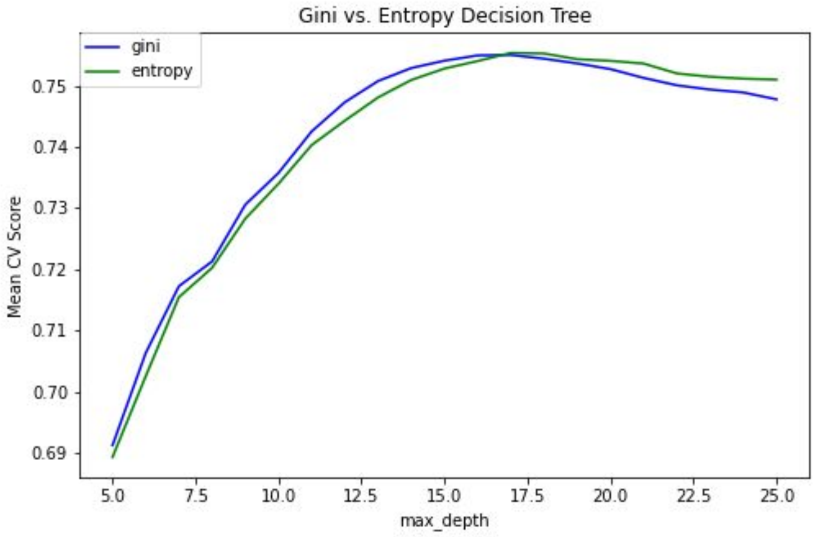
Model 1: Decision Tree Classifier

Model 2: Random Forest Classifier

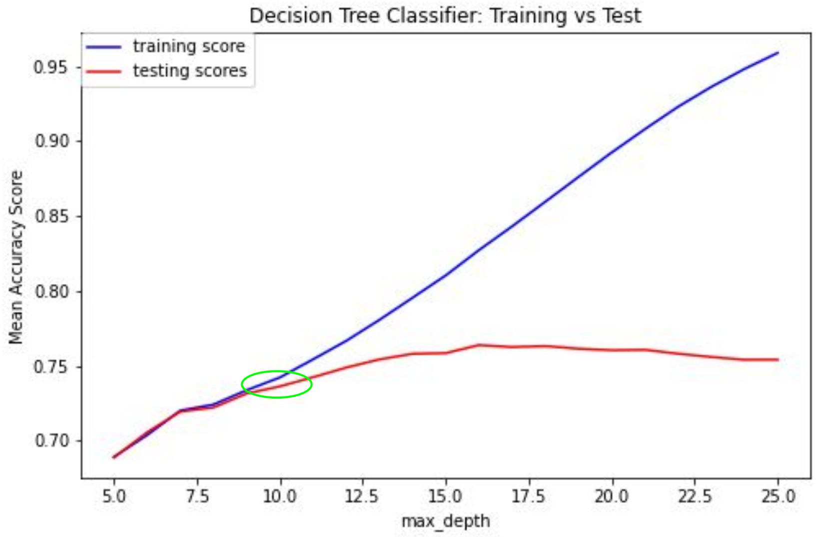
Model 3: Gradient Boost Classifier

Model 1: Decision Tree Classifier

Grid Search CV

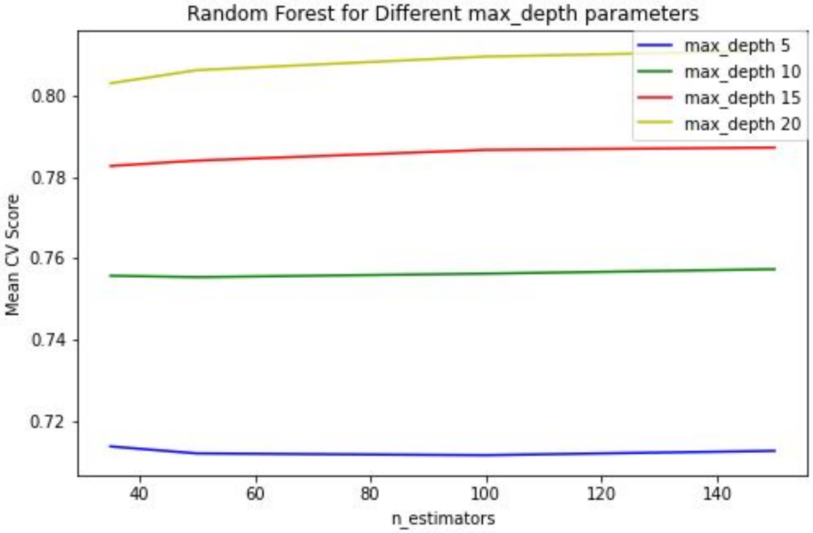


Training vs Test

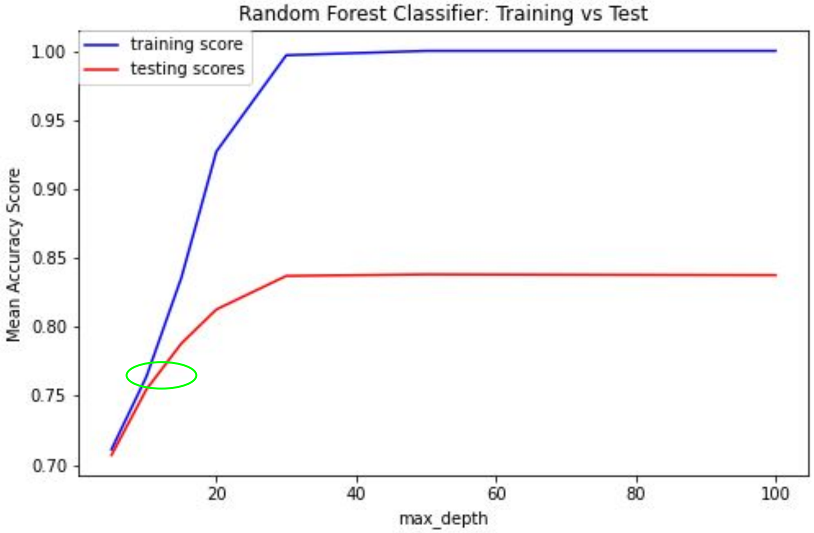


Model 2: Random Forest Classifier

Grid Search CV

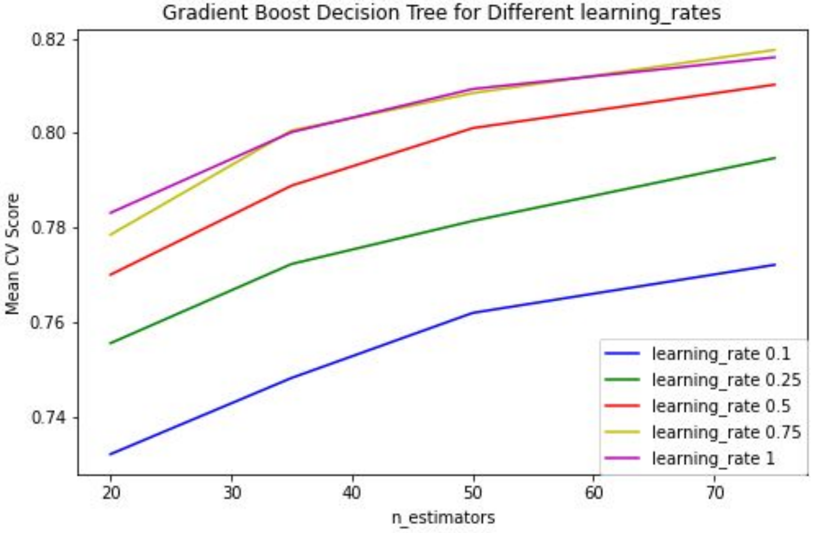


Training vs Test

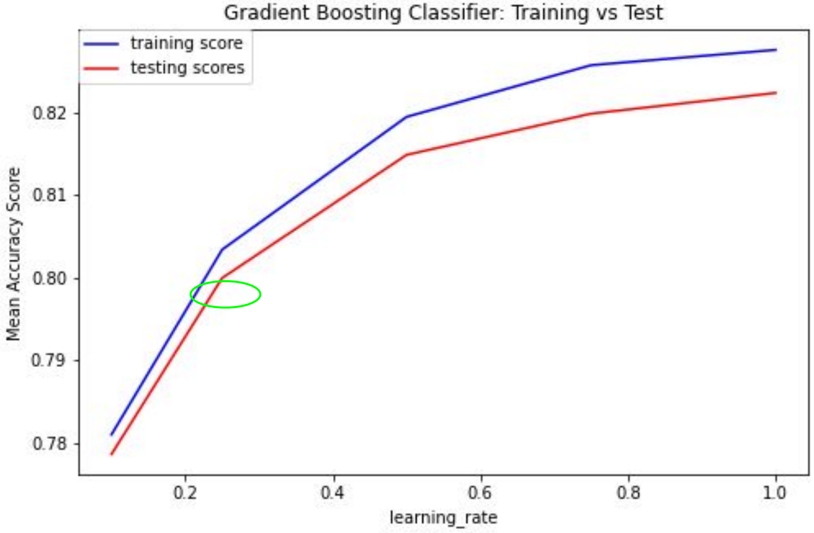


Model 3: Gradient Boost Classifier

Grid Search CV



Training vs Test



Model Results: F-1 Score

Decision Tree

F-1: 0.729

Random Forest

F-1: 0.754

Gradient Boost

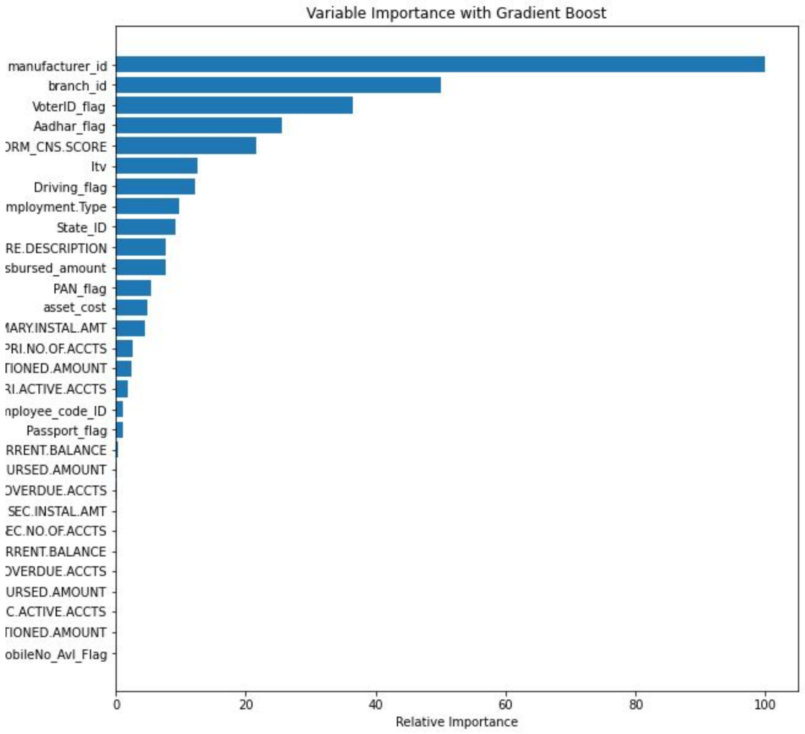
F-1: 0.809

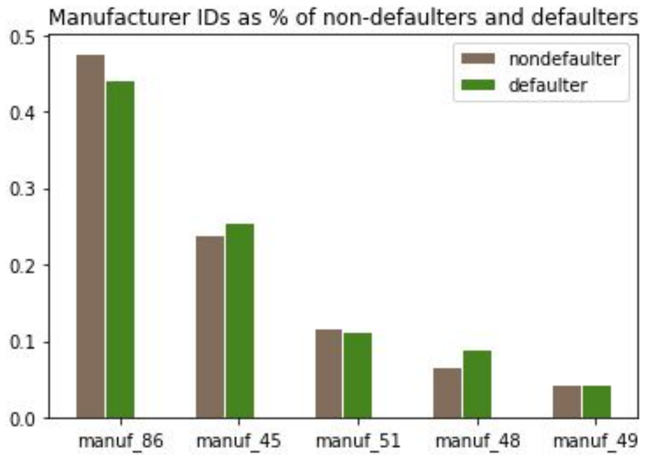
Confusion Matrix

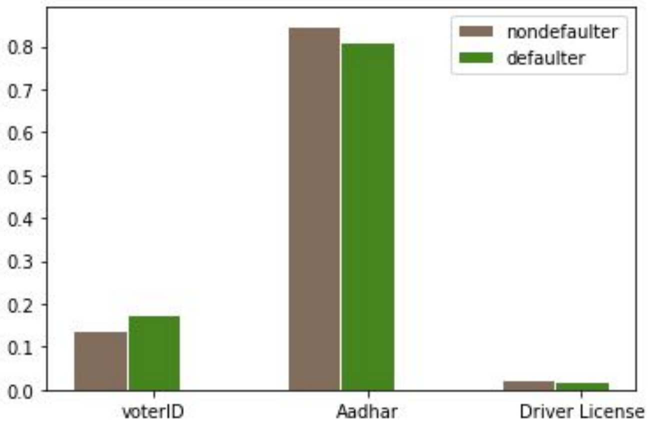
```
Gradient Boost: Accuracy=0.811
Gradient Boost: f1-score=0.809
```

	precision	recall	f1-score	support
0	0.76	0.92	0.83	36561
1	0.90	0.70	0.79	36457
accuracy			0.81	73018
macro avg	0.83	0.81	0.81	73018
weighted avg	0.83	0.81	0.81	73018

Model 3: Important Features







How the User Can Utilize our Results

Important Feature Identification

Credit Score Discrepancy

Default Prediction as Final Voice of Reason