IMARTICUS

LEARNING

❖ Presentation on: Lending Club Data (Loan Predictions)

An Overview

Prepared by – Shubham Chaudhari

Amol Shinde

Imarticus Learning Institute, Pune

❖ Table of Content

- ✓ Introduction
- ✓ Knowing the Dataset
- ✓ Pre-processing of Dataset
- ✓ Exploratory Data Analysis
- ✓ Feature Extraction
- ✓ Selection of Model
- ✓ Conclusion

❖ Introduction

- ✓ Lending Club is a US peer-to-peer lending company, headquartered in San Francisco, California. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. Lending Club is the world's largest peer-to-peer lending platform.

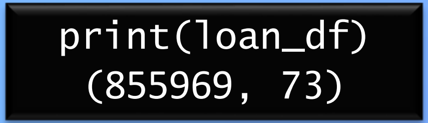
❖ Knowing the Dataset

✓ Understanding the Dataset

These files contain complete loan data for all loans issued through the June 2007- Dec 2015, including the current loan status and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter. Additional features include credit scores, number of finance inquiries, address including zip codes, and state, and collections among others. The file is a matrix of about 855969 observations and 73 variables

❖ Knowing the Dataset

✓ **Checking number of Columns and Rows**



```
print(loan_df)
(855969, 73)
```

✓ **Describe the Dataset**

	Count	Percent
emp_title	51462	5.799326
emp_length	44825	5.051393
annual_inc	4	0.000451
desc	761351	85.797726
title	152	0.017129
delinq_2yrs	29	0.003268
mths_since_last_record	750326	84.555303
open_acc	29	0.003268

✓ **Concluding X-Variables and Y-Variables**

❖ Pre-Processing of Dataset

- ✓ **Checking NA's, NULL's, ?, BLANKS**

Checked for Nulls and Blanks in the dataset

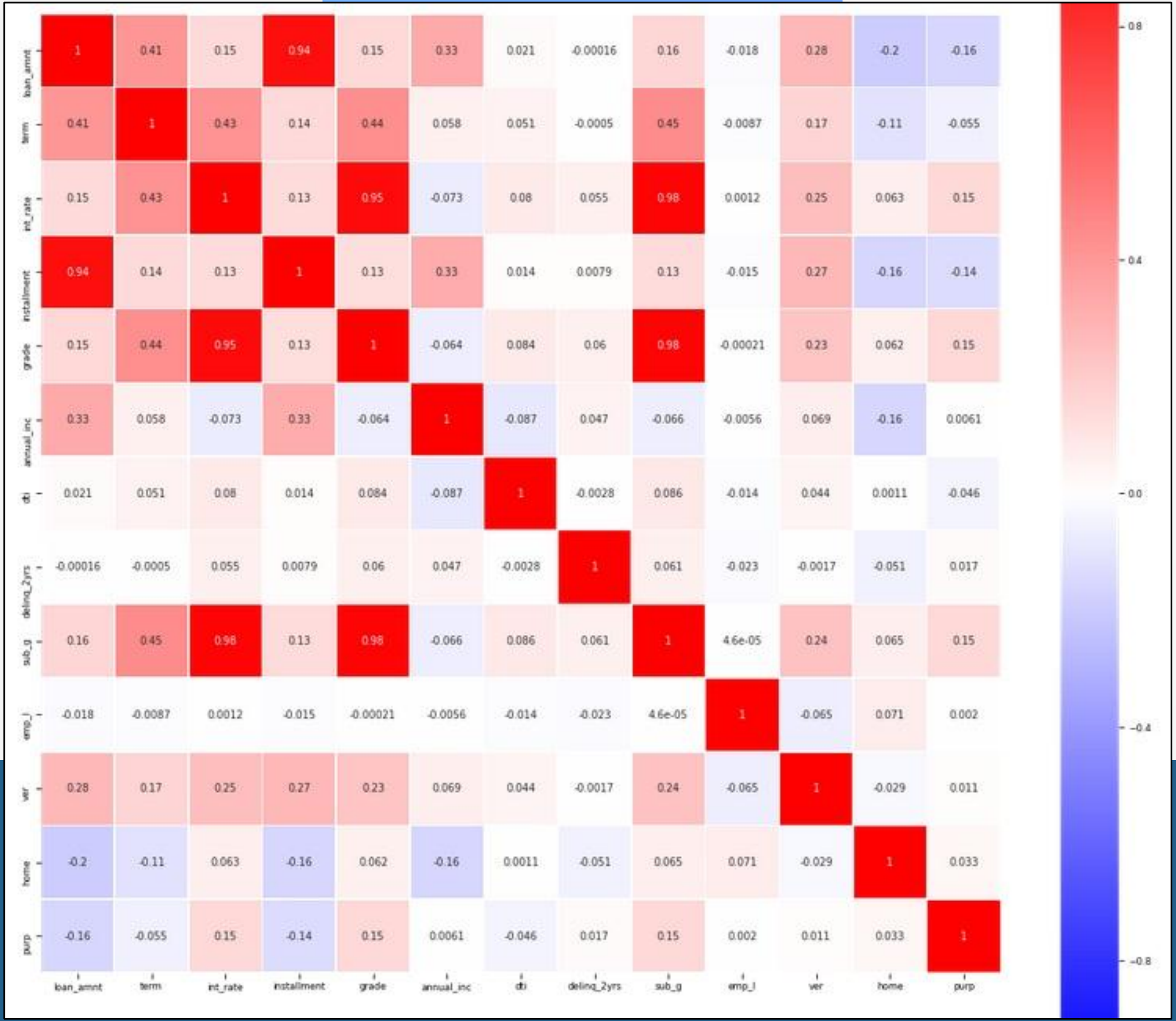
- ✓ **Dealing with missing values**

Columns having missing values greater than 80% have been dropped.

- ✓ **Dropping Unnecessary Columns**

❖ Graphical Representation

1. Correlation between the features



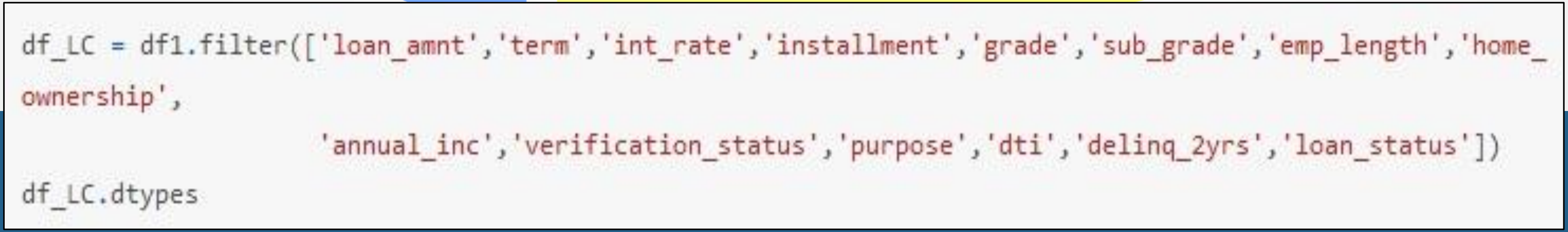
	loan_amnt	term	int_rate	installment	grade	annual_inc	dti	delinq_2yrs	sub_g	emp_l	ver	home	purp
loan_amnt	1	0.41	0.15	0.94	0.15	0.33	0.021	-0.00016	0.16	-0.018	0.28	-0.2	-0.16
term	0.41	1	0.43	0.14	0.44	0.058	0.051	-0.0005	0.45	-0.0087	0.17	-0.11	-0.055
int_rate	0.15	0.43	1	0.13	0.95	-0.073	0.08	0.055	0.98	0.0012	0.25	0.063	0.15
installment	0.94	0.14	0.13	1	0.13	0.33	0.014	0.0079	0.13	-0.015	0.27	-0.16	-0.14
grade	0.15	0.44	0.95	0.13	1	-0.064	0.084	0.06	0.98	-0.00021	0.23	0.062	0.15
annual_inc	0.33	0.058	-0.073	0.33	-0.064	1	-0.087	0.047	-0.066	-0.0056	0.069	-0.16	0.0061
dti	0.021	0.051	0.08	0.014	0.084	-0.087	1	-0.0028	0.086	-0.014	0.044	0.0011	-0.046
delinq_2yrs	-0.00016	-0.0005	0.055	0.0079	0.06	0.047	-0.0028	1	0.061	-0.023	-0.0017	-0.051	0.017
sub_g	0.16	0.45	0.98	0.13	0.98	-0.066	0.086	0.061	1	4.6e-05	0.24	0.065	0.15
emp_l	-0.018	-0.0087	0.0012	-0.015	-0.00021	-0.0056	-0.014	-0.023	4.6e-05	1	-0.065	0.071	0.002
ver	0.28	0.17	0.25	0.27	0.23	0.069	0.044	-0.0017	0.24	-0.065	1	-0.029	0.011
home	-0.2	-0.11	0.063	-0.16	0.062	-0.16	0.0011	-0.051	0.065	0.071	-0.029	1	0.033
purp	-0.16	-0.055	0.15	-0.14	0.15	0.0061	-0.046	0.017	0.15	0.002	0.011	0.033	1

Exploratory Data Analysis

✓ Finding the correlation between the variables

It can be seen from the plot above that loan amount and installment have a very high correlation amongst each other (0.94). This is intuitive since a person who takes a large sum of loan would require extra time to repay it back. Also, interest rate, sub grade and grade have a very high correlation between them. This is obvious since interest rate is decided by grades once the grades are decided, a subgrade is assigned to that loan (leading to high correlation).

✓ Selecting important variables



```
df_LC = df1.filter(['loan_amnt','term','int_rate','installment','grade','sub_grade','emp_length','home_ownership',
                    'annual_inc','verification_status','purpose','dti','delinq_2yrs','loan_status'])
df_LC.dtypes
```

❖ Model Selection

✓ Transformation

Before training the data, we would first transform the data to account for any skewness in the variable distribution. *(Float)*

Splitting Dataset into Training and Testing

We split the dataset based on 'issue_d'
Train set - (June 2007 - May 2015)
Test set – (June 2015 – Dec 2015)

✓ Model Selection

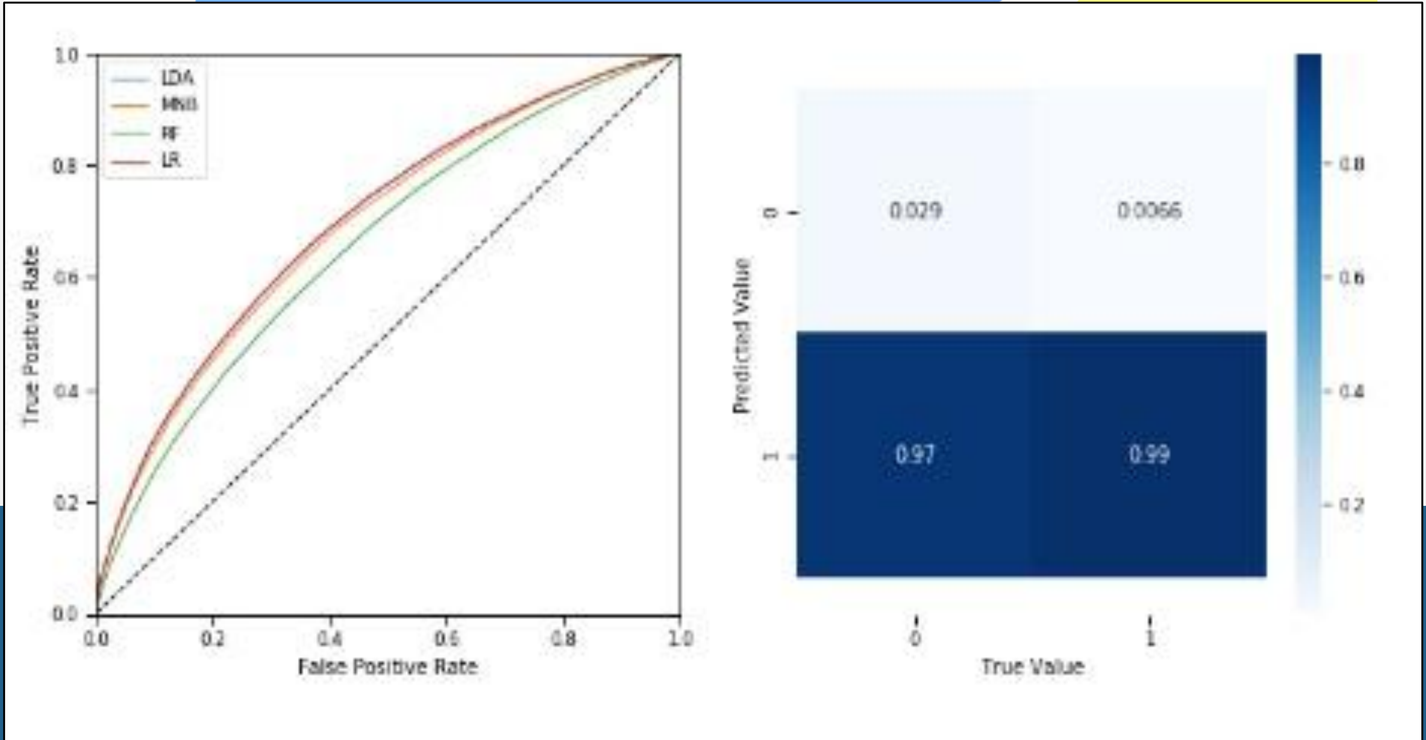
1. Logistic Regression
2. Random Forest

✓ Cross- Validation (3-Folds)

Results

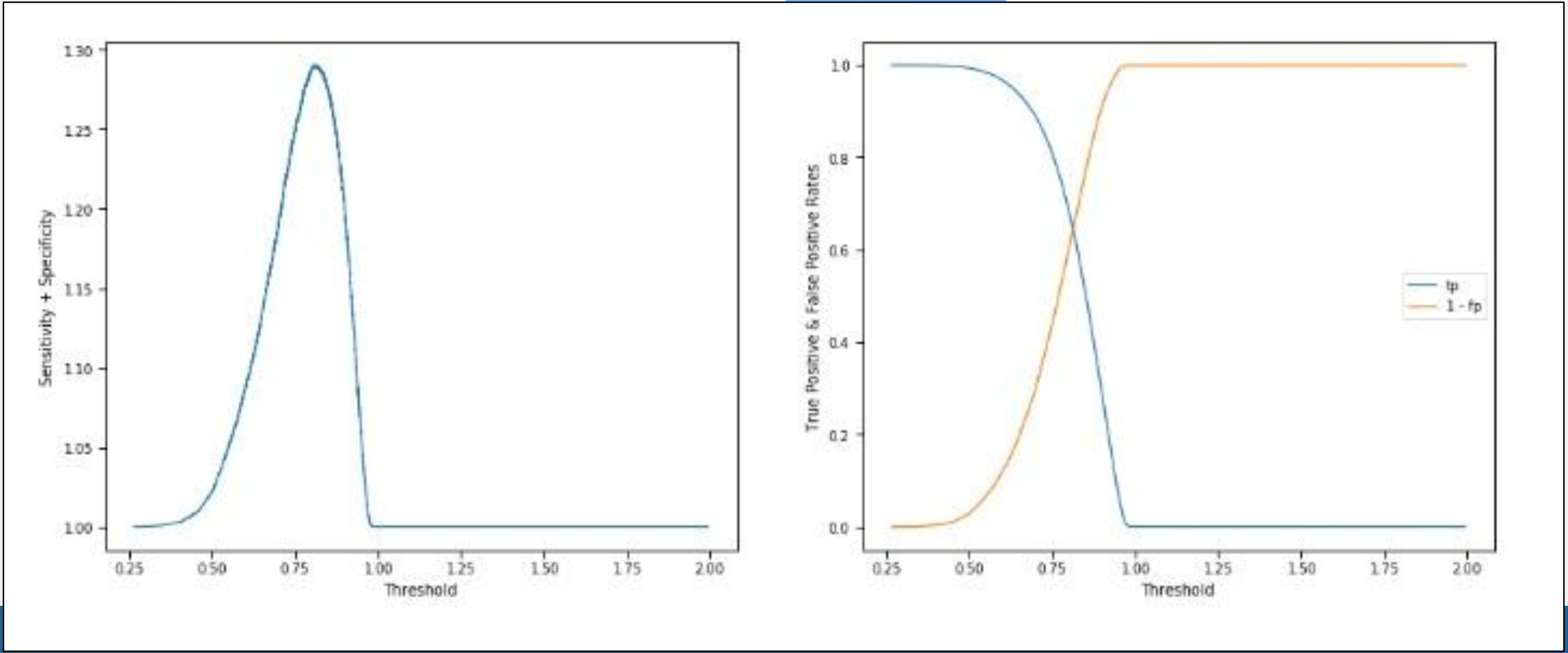
- ✓ Accuracy
- ✓ ROC Curve
- ✓ Confusion Matrix

1. LR [0.92009416 0.92294362 0.92294362] 0.9219937927023743
2. RF [0.89259241 0.90865426 0.92266815] 0.9079716039306898



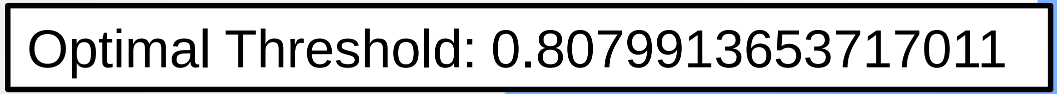
Results

✓ ROC Curve

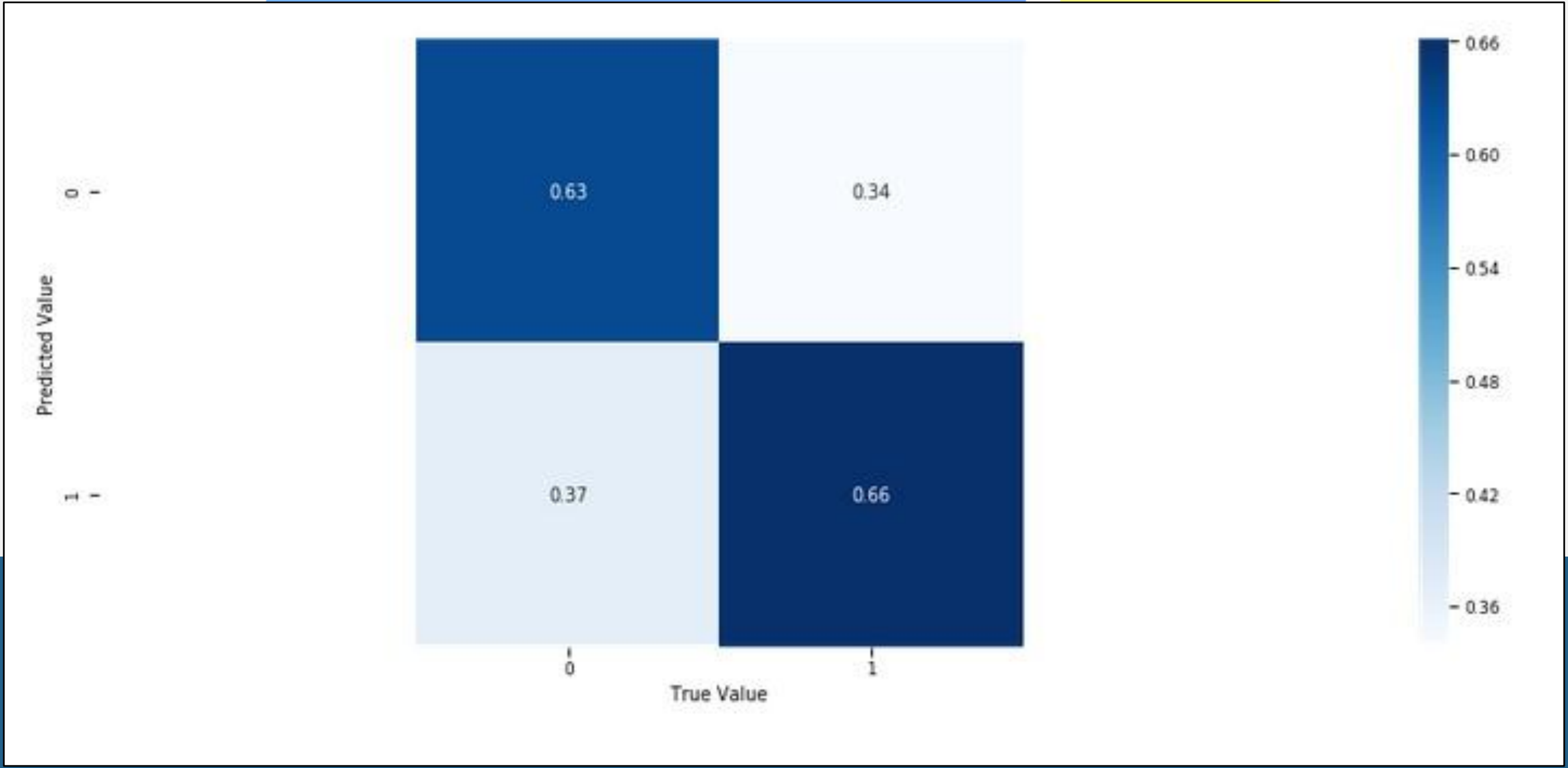


Results

✓ Optimal Threshold Values

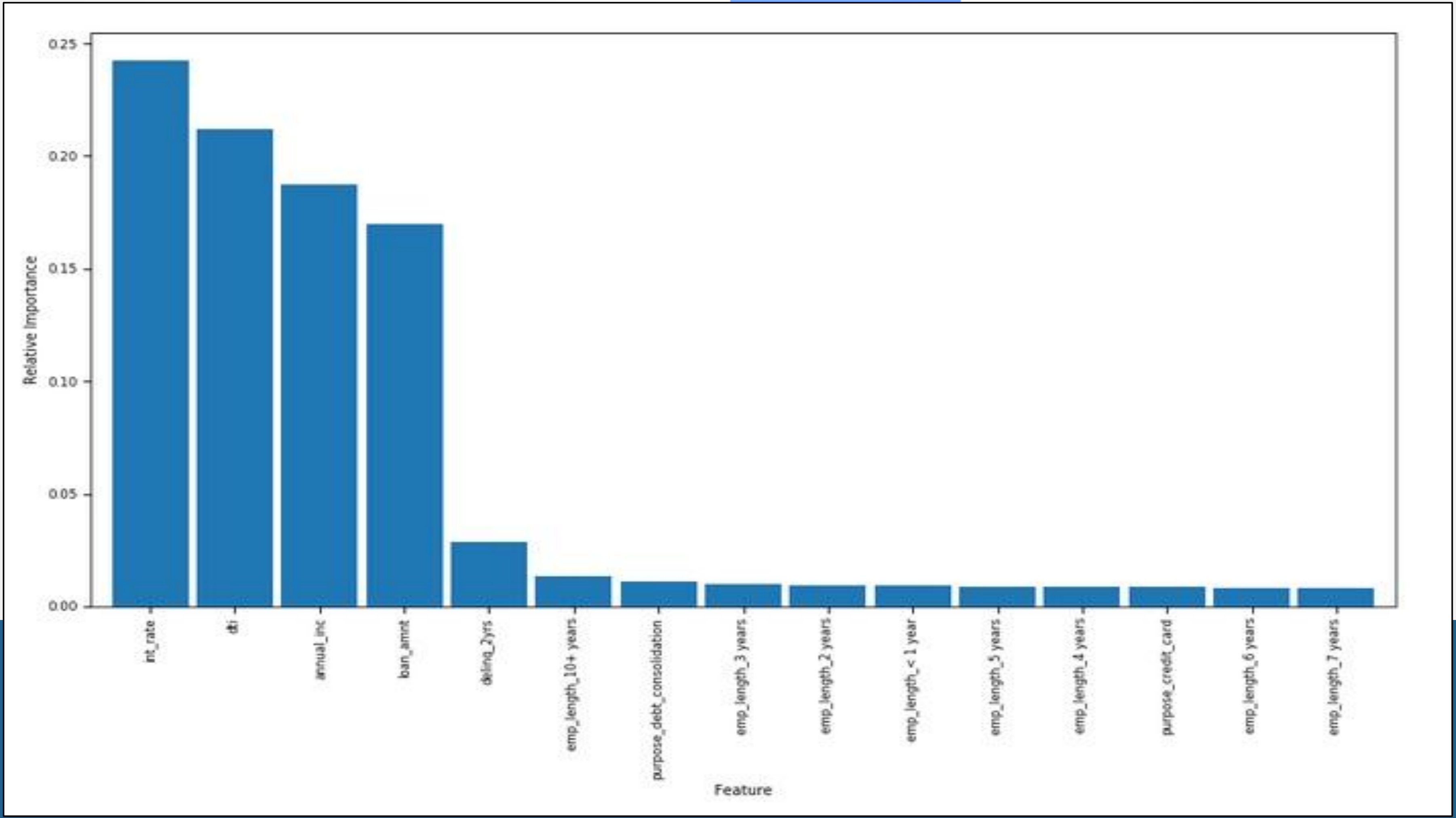


```
Optimal Threshold: 0.8079913653717011
```



Results

- ✓ **Relative Important Variables**



IMARTICUS
LEARNING

THANK YOU

Imarticus Learning Institute, Pune