Data Science for Business: Technical

Instructor: Dr. Foster Provost

PROJECT REPORT

Credit Risk Assessment for Lending Institutions

(Building a Predictive Model for Loan Decision Making)

Team:

Yogya Sharma (ys5350@nyu.edu)
Noel Nebu Panicker (nnp566@nyu.edu)
Anubhav Ghildiyal (ag8766@nyu.edu)
Viraj Parikh (vp2359@nyu.edu)

CONTENTS

1. INTRODUCTION

1.1.Objective:

The primary objective of this project is to evaluate the creditworthiness of loan applications with the intention of maximizing profitability within a specific budget the bank has which can be given as loans. This will be achieved through the development of a highly accurate predictive model capable of thoroughly assessing the credit risk associated with potential borrowers in various lending institutions. The project entails a comprehensive analysis of customer financial data, credit history, employment details, and loan repayment records, all of which will contribute to the construction of a robust credit risk assessment model. Ultimately, the aim is to make well-informed loan decisions that not only mitigate risks but also maximize profitability for the lending institution.

1.2.Plan of action:

The loan approval process involves a comprehensive evaluation of various factors to assess the suitability of a customer's loan application. These factors include the customer's credit score, age, loan type, loan amount, and other relevant information. By carefully considering these factors, we assign a credit risk value to each customer within a range of 0 to 1. Our aim is to train a robust model using a comprehensive dataset that incorporates all these features. Through this training, we seek to develop an accurate credit risk assessment model that can effectively evaluate and rate the risk associated with each customer. This model will provide lending institutions with valuable insights to make well-informed loan decisions. By leveraging customer financial data, credit history, employment information, and loan repayment history, we strive to optimize the loan approval process while safeguarding the institution's financial interests. Ultimately, our goal is to strike a balance between approving viable loan applications and ensuring the lending institution's profitability and sustainability.

Following are the steps we'll follow:

Step 1: Extract data.

Step 2: Clean the data.

Step 3: Analyze the data.

Step 4: Feature selection.

Step 5: Train and test different models.

Step 6: Implement the business requirements on the best model.

Step 7: Repeat from step 5.

2. BUSINESS ASPECT

In the subsequent step of the risk assessment process, the model will assign a range of values (confidence) to determine if they should be given a loan to each instance. This allows loans to be granted exclusively to customers whose associated risk falls below the specified threshold.

Default Risk: **Defaulting** on a loan means failing to pay it back, so each person's probability that they'll fail to pay back their loan is called their **default risk**.

While minimizing defaults can be done just with this information, it is important to note that our primary objective is to maximize profitability within a specific budget the bank has which can be given as loans. Thus, we employ various mathematical formulae to optimize profitability and achieve the best possible outcomes.

One approach involves determining an expected value from the loan portfolio by considering both the anticipated return from successful loans and the expected loss resulting from defaults. To calculate this, we utilize the following mathematical formula:

Expected Value = (1 – risk) * Interest + risk * credit amount

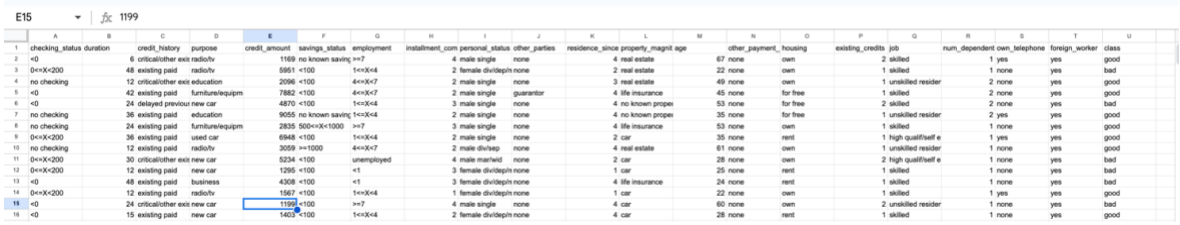
Where,

- risk: The probability that a person will default on a loan (1 – confidence).
- interest: The calculated value of applicable interest, assuming the loan is given, calculated by incorporating the associated risk.
- Credit amount: Requested loan amount.

By leveraging this formula, we can determine the best model and threshold values that lead to optimal profitability. This comprehensive approach ensures that our efforts go beyond solely minimizing defaults and focus on making well-informed decisions that maximize profitability while effectively managing credit default rates.

3. INTRODUCTION TO DATA

Here is a screenshot of the dataset (sourced from Kaggle) we are looking to use and the features we will be considering for the business proposition mentioned above:



checking_status	duration	credit_history	purpose	credit_amount	savings_status	employment	installment_com	personal_status	other_parties	residence_since	property_magnit	age	other_payment_	housing	existing_credits	job	num_dependent	own_telephone	foreign_worker	class
<0	6	critical/other exis	radio/tv	1169	no known saving	>=7	4	male single	none	4	real estate	67	none	own	2	skilled	1	yes	yes	good
0<=X<200	48	existing paid	radio/tv	5951	<100	1<=X<4	2	female div/dep/n	none	2	real estate	22	none	own	1	skilled	1	none	yes	bad
no checking	12	critical/other exis	education	2096	<100	4<=X<7	2	male single	none	3	real estate	49	none	own	1	unskilled resider	2	none	yes	good
<0	42	existing paid	furniture/equipm	7882	<100	4<=X<7	2	male single	guarantor	4	life insurance	45	none	for free	1	skilled	2	none	yes	good
<0	24	delayed previou	new car	4870	<100	1<=X<4	3	male single	none	4	no known prope	53	none	for free	2	skilled	2	none	yes	bad
no checking	36	existing paid	education	9055	no known saving	1<=X<4	2	male single	none	4	no known prope	35	none	for free	1	unskilled resider	2	yes	yes	good
no checking	24	existing paid	furniture/equipm	2835	500<=X<1000	>=7	3	male single	none	4	life insurance	53	none	own	1	skilled	1	none	yes	good
0<=X<200	36	existing paid	used car	6948	<100	1<=X<4	2	male single	none	2	car	35	none	rent	1	high qualif/self e	1	yes	yes	good
no checking	12	existing paid	radio/tv	3059	>=1000	4<=X<7	2	male div/sep	none	4	real estate	61	none	own	1	unskilled resider	1	none	yes	good
0<=X<200	30	critical/other exis	new car	5234	<100	unemployed	4	male mar/wid	none	2	car	28	none	own	2	high qualif/self e	1	none	yes	bad
0<=X<200	12	existing paid	new car	1295	<100	<1	3	female div/dep/n	none	1	car	25	none	rent	1	skilled	1	none	yes	bad
<0	48	existing paid	business	4308	<100	<1	3	female div/dep/n	none	4	life insurance	24	none	rent	1	skilled	1	none	yes	bad
0<=X<200	12	existing paid	radio/tv	1567	<100	1<=X<4	1	female div/dep/n	none	1	car	22	none	own	1	skilled	1	yes	yes	good
<0	24	critical/other exis	new car	1199	<100	>=7	4	male single	none	4	car	60	none	own	2	unskilled resider	1	none	yes	bad
<0	15	existing paid	new car	1403	<100	1<=X<4	2	female div/dep/n	none	4	car	28	none	rent	1	skilled	1	none	yes	good

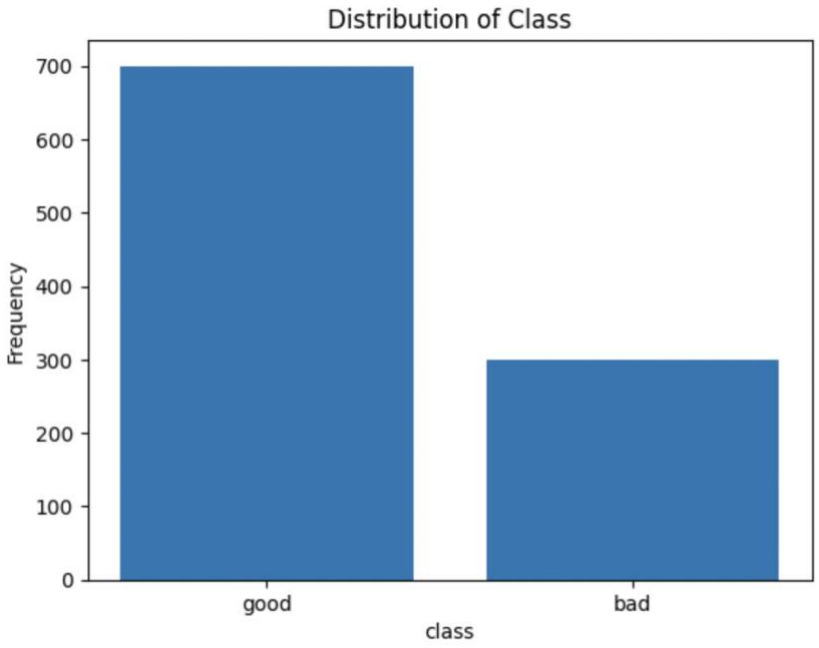
The following are the attributes taken into consideration:

Serial No.	Attribute	Description	Data Type
1.	Checking Status	Status of existing checking account	object
2.	Duration	Duration in months	float64
3.	Credit History	General representation completion payment of existing credit lines	object
4.	Purpose	Purpose of taking loan	object
5.	Credit Amount	Requested loan amount	float64
6.	Savings Status	Status of savings account/bond	object
7.	Employment	Number of years employment	object
8.	Installment	Installment rate in percentage of disposable income	float64
9.	Personal Status	Sex and marital data	object
10.	Other Parties	Other debtors/guarantors	object
11.	Residence since	Duration of residence of their current residence	float64
12.	Property magnitude	Magnitude of current assets	object
13.	Age	Age of the applicant	float64
14.	Other payment plans	Additional sources for repayment	object

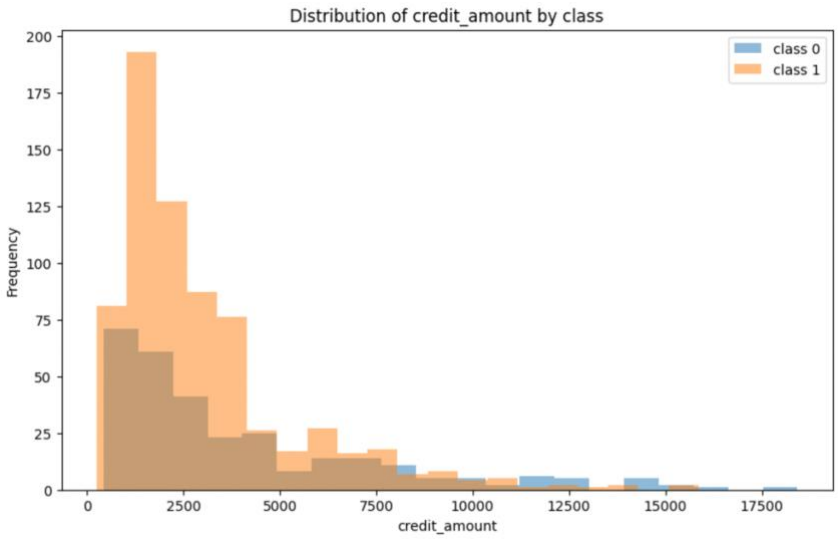
15.	Housing	Ownership/renting of current residence	object
16.	Existing credits	Any existing credit lines	float64
17.	Job	Level of skill	object
18.	Number of dependents	Dependents of the applicant	float64
19.	Own telephone	If they own a telephone	object
20.	Foreign Worker	If they are a foreign worker	object
21.	Class	If the credit request is good or bad	object

4. EXPLORATORY DATA ANALYSIS

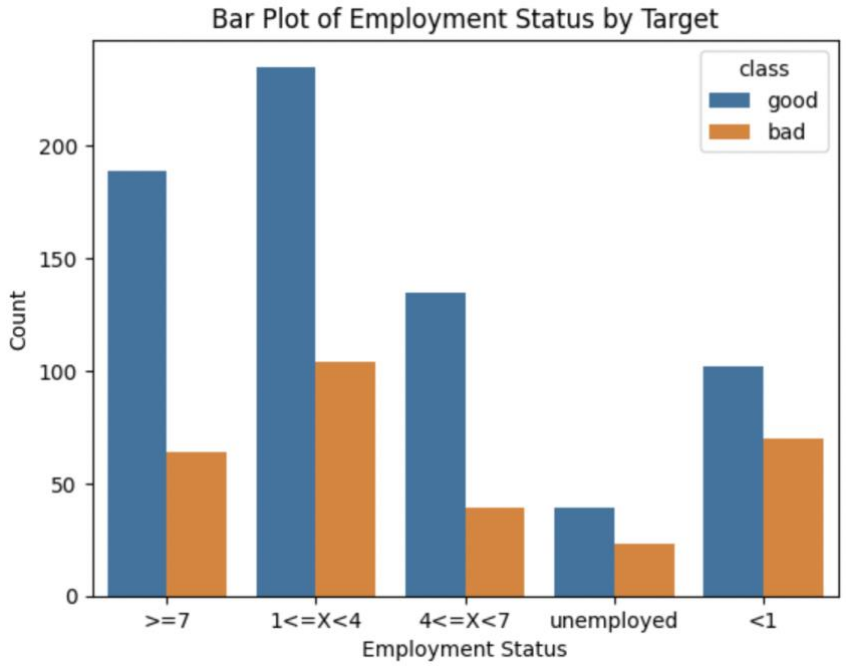
All the data was initially analyzed, wherein it was deduced that there were no null values, we had 7 continuous and 13 categorical attributes with a total of 1000 instances. We used 'class' as the target variable with many good loans, the count being 700 good and 300 bad, which implies that there is a slight disproportion in the data.

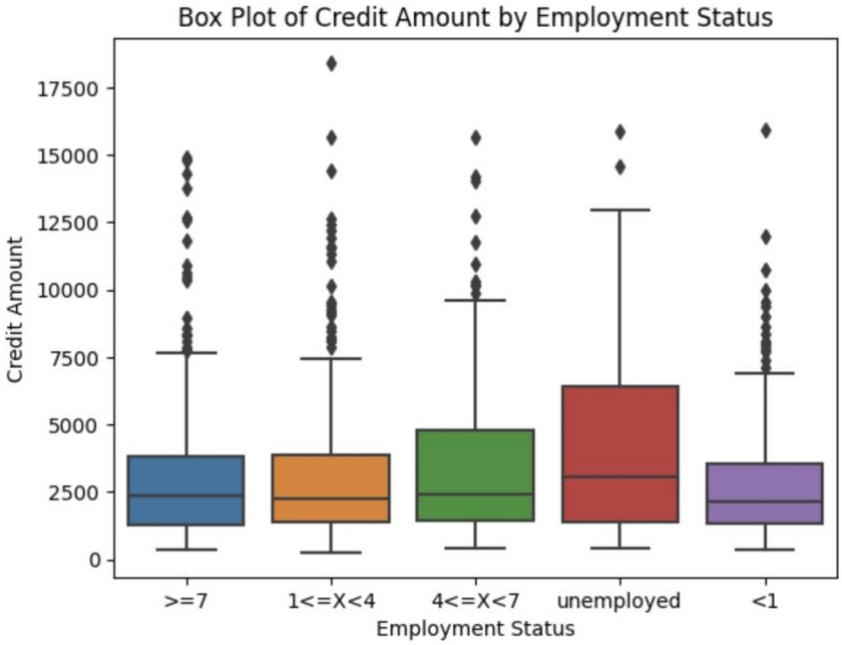


It can be observed that the maximum credit loan amounts were around $0-$5000, with highest frequency around $2500.

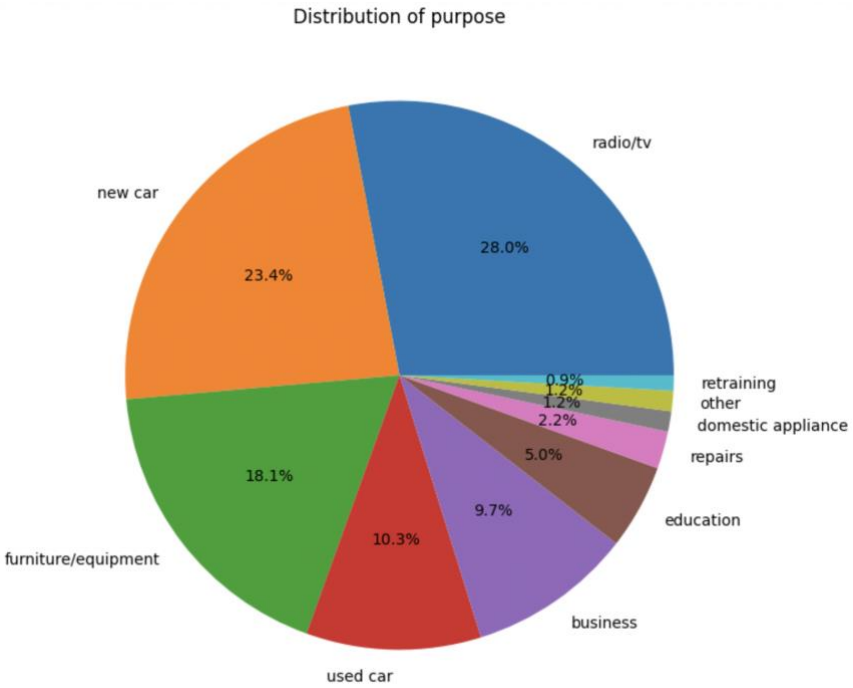


From the employment bar chart, we realized that people who were either unemployed or were employed for less than a year had a higher tendency to default. Moreover, it was the unemployed people who had requested the biggest and highest requested credit amounts. Both points suggest a better prediction mechanism to ensure that the loans are correctly given.



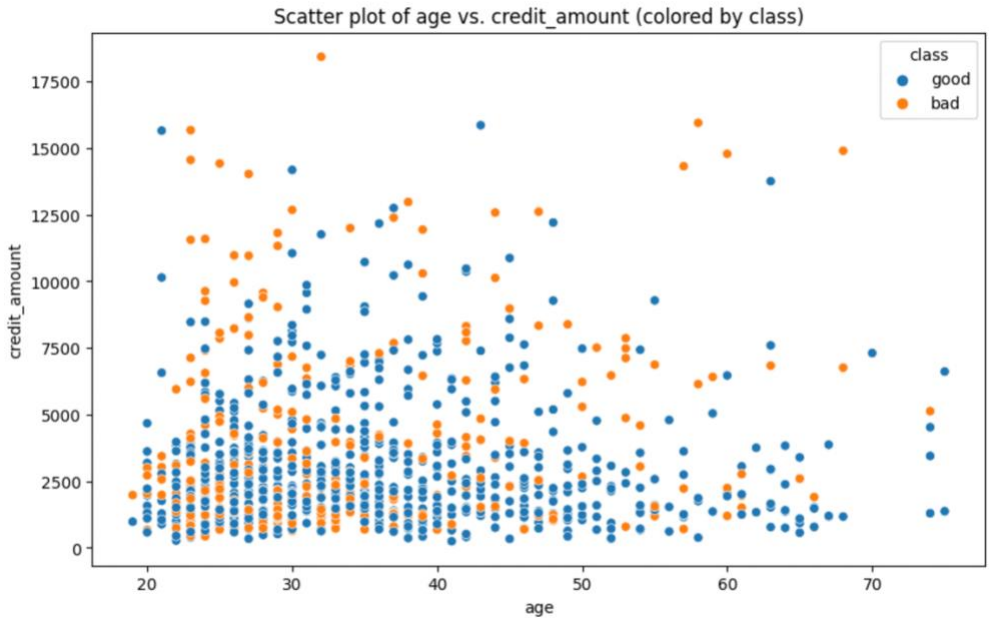


Interestingly most of the loans were taken for household appliances like Radio and TV which having relatively low costs seem to be better loans.



On the other hand, loans given out for new cars seemed to have relatively stronger tendencies to default. Moreover, a large amount of loans was given to customers with no checking accounts suggesting that having a checking amount may not always be necessary to provide loans.

We can see from the scatterplot below that most requested loans are from people between ages 20-40 with a requested credit amount between $0-$5000.

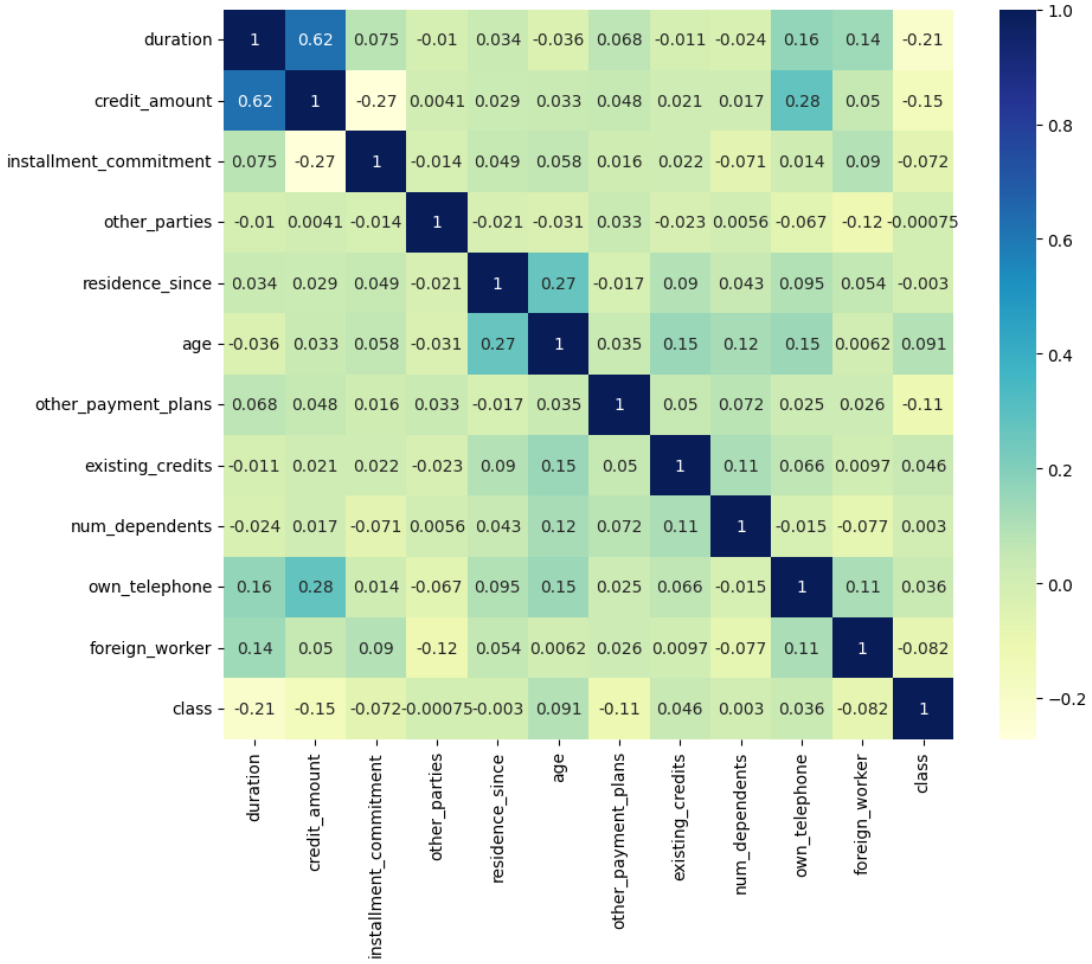


5. FEATURE ENGINEERING

As we analyzed the data, going further, we need to select and prepare our features to train the model.

5.1. Feature Selection

There were certain categorical features (other_parties, foreign_worker, own_telephone, other_payment_plans and class), which were directly converted into binary variables. Post that, we plotted the correlation matrix between the available continuous variables and the target variable (class).



From the correlation matrix we can see that there are multiple features with very low correlation value with the target class, so we eliminated any feature which has a correlation < | 0.1 | with 'class' , as they don't contribute much to it.

After, we had selected the continuous features, we had to check how much do the categorical variables contribute to the formulation of the target variable. To do this, we performed a Chi-Squared test, to get the Chi-Square metric and the corresponding p-value. By analyzing the chi-square statistic and corresponding p-values, we can identify features that exhibit strong associations with the target class. The chi-square statistic measures the difference between the observed frequencies and the expected frequencies under the assumption of independence. The corresponding p-value indicates the probability of obtaining such an extreme or more significant result if there were no association between the variables.

Choosing features based on the chi-square statistic and p-value is essential for training a predictive model. Features with high chi-square statistics suggest a substantial association with the target class. This indicates that the feature provides valuable discriminatory information that can help differentiate between different classes of the target variable. Features with low p-values indicate a high level of statistical significance, suggesting that the observed association is unlikely to occur by chance. Therefore, when selecting features for model training, it is advisable to prioritize those with both high chi-square statistics and low corresponding p-values. These features are more likely to contribute significantly to the predictive power of the model and capture important patterns related to the target class. By incorporating such features into the model, we can potentially improve its accuracy and performance in predicting the target variable.

So, we selected the features which have a chi-square statistic greater than 10 and a corresponding p-value being lesser than 0.05.

The finally selected features are: checking_status, credit_history, purpose, savings_status, employment, housing, other_payment_plans, credit_amount, property_magnitude and duration.

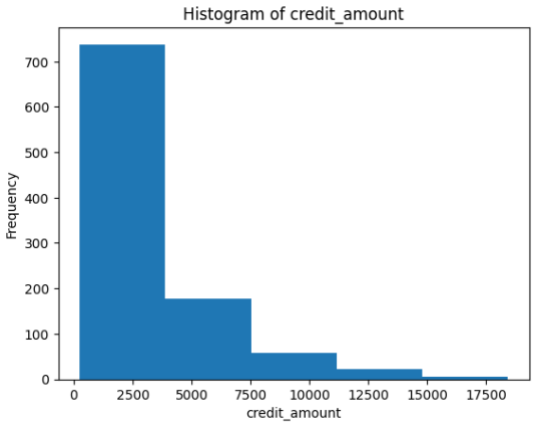
And we can see here that out of all the features, the selected ones are the most relevant to acquiring a loan.

5.2. Feature Transformation

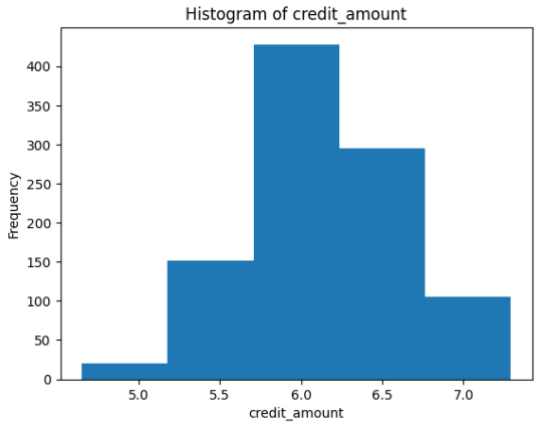
Now, that we have selected the features, we still need to transform them into being suitable as training input of the model.

We converted the categorical features into One-Hot representations. About the continuous variables, as you can see in the code, both credit_amount and duration were positively skewed and we used Box-Cox transformation to transform them.

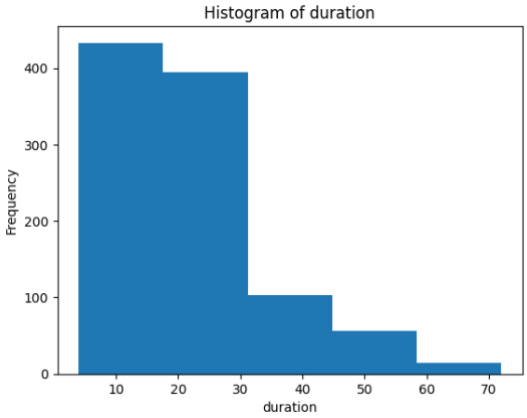
Pre-normalization:



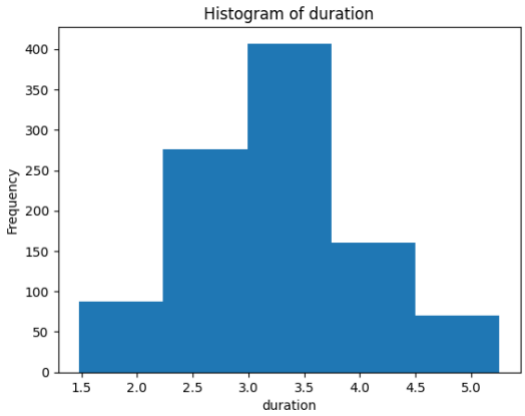
Post-normalization:



Pre-normalization:



Post-normalization:



This ensures that we filter out any outliers which would lead to a skewed prediction.

To get a better representation of the purpose of the loans we grouped all the loan types mentioned in the dataset under 4 broad loan categories:

1. Personal Loan:

- Repairs
- Retraining
- Furniture/Equipment
- Other
- Domestic Appliances

2. Auto Loan:
    - New car
    - Old car

3. Business Loan

4. Education Loan

6. MODELING AND EVALUATION

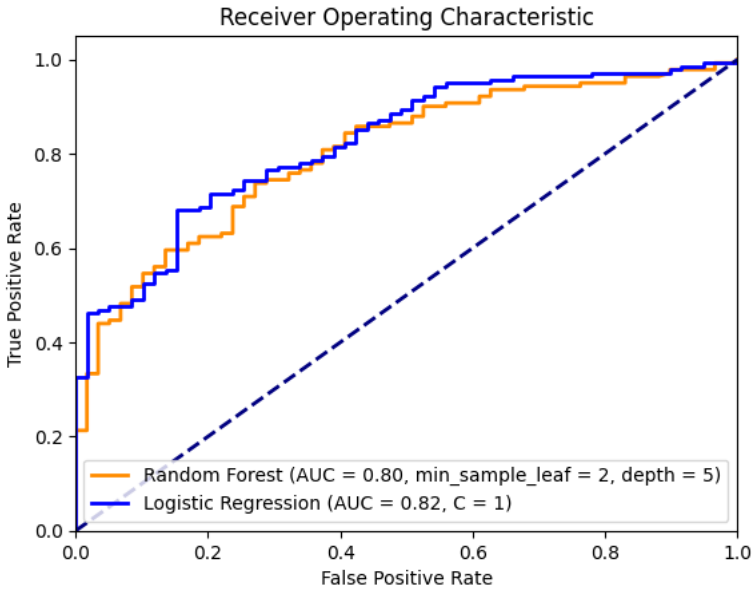
Now, that we have the data ready to be input into the model, we proceed with modeling.

Since the target variable is a yes or a no (good loan or bad loan / 1 or 0), we would be using classification models to train and test the data on.

We split the data into training and testing with a split of 80% for training and 20% for testing. We worked on two classifiers:

- Random-Forest Classifier: Tested the model for min_samples_leaf values: [1, 2, 4, 8, 16, 32, 64, 128] and tree depth ranging from 1 to 10.
- Logistic Regression Classifier: Tested the model for C values: [0.001, 0.01, 0.1, 1, 10, 100].

After finding the best set of hyperparameters for both models, we plotted the ROC curve for both of those.



Different evaluation metrics:

- Random Forest: f1 score = 0.85, accuracy = 0.76, AUC = 0.80
- Logistic Regression: f1 score = 0.85, accuracy = 0.78, AUC = 0.82

The evaluation metric we considered to choose between classifiers is AUC (Area Under the Curve). As we can see from the ROC plot that the **Logistic Regression Classifier** performs better than the Random Forest Classifier with an AUC of 0.82. We can also see that, the Logistic Regression Classifier, bows out initially being closer to the y-axis indicating a better true positive rate to false positive rate ratio.

7. DECISION LOGIC

Now that we have the best classifier to use, we calculate the confidence (probability), to see which applicants have a better credit standing to be eligible for the loan. The objective here is not only to identify the bad loans, but also identify good loans to get the best profits and the least credit default risks, which is generally a tradeoff. There may be instances where the risks associated to a loan are slightly high yet producing good profits. We calculate the risk based on the confidence levels we found previously.

Determining somebody's default risk is important since it helps calibrating their interest rates to mitigate the risk of lending money to them.

There are multiple ways we can mitigate risks associated to loans like:

1. Increasing everyone's interest rates to make up for the losses.
2. Only accepting safe borrowers
3. Charging people interest rates proportional to their default risk.

We have worked on the 3rd method as according to us it is the most promising and rewarding, giving us more flexibility to maximize profit.

We combined various purposes into 4 types of loan, and we set the base rates (set the values based on averages found on the internet):

Type of loan	**Base interest rate**
Personal Loan	7%
Auto Loan	6%
Business Loan	7.5%
Education Loan	5%

After this, we calculate risk as:

$$risk = 1 - confidence$$

and used this to evaluate the additional interest rate that needs to be added to the base rate for each applicant, so the total interest rate for each applicant becomes:

$$Interest = Base\ interest\ rate + (risk / (1 - risk)) \ [1]$$

Using this interest value, we calculated the profit that we will get if the complete loan is re-paid. Using the compound interest formula with credit_amount as principal and the duration of the loan.

We then calculated the Expected Benefit from each applicant. There are multiple cases of non-payment of dues and expected loss during the duration of the loan, but we have only considered the case whether a person either does pay the entire amount back with interest or doesn't pay anything at all, as including all other cases to evaluate the accurate Expected Benefit would require domain knowledge of the banking system.

$$\text{Expected Benefit (Future Value)} = \text{confidence} * \text{interest} - \text{risk} * \text{credit\_amount}$$

This equation doesn't however include the time value of money, so we introduced this concept to determine the 'Present Value" of the absolute value calculated from the Expected Benefit (Future Value).

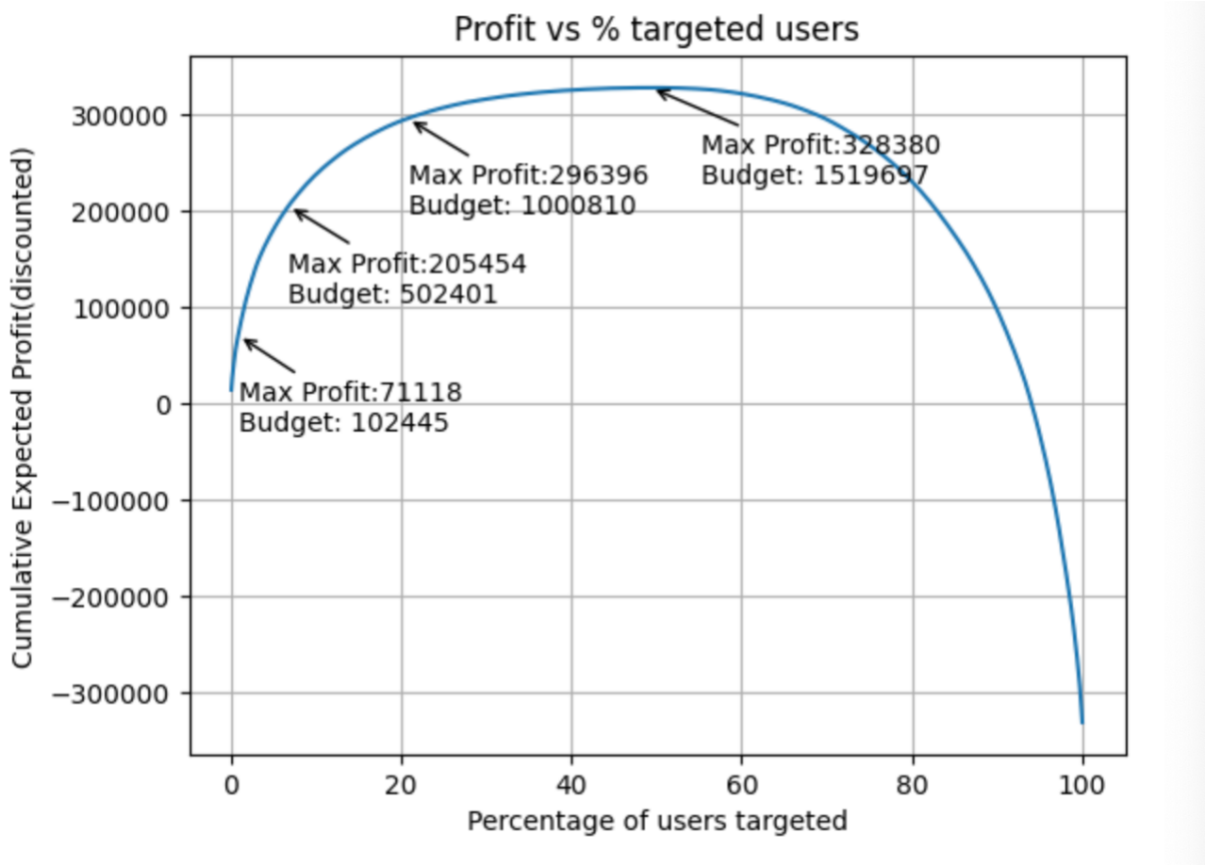
$$\text{Expected Benefit (Present Value)} = \text{Expected Benefit (Future Value)} \ / \ (1 + \text{risk})^{n} \ [2]$$

Where, n is the duration in years.

We then ranked the applicants in order of maximum expected benefit (Present Value) and plotted a profit curve which tells what percentage of users to select to maximize the profit given a set budget. Different budget-profit markers can be seen on the profit curve.

[[1] Refer to Citation 3]

[[2] Refer to Citation 4]



Profit Curve

8. CONCLUSION

We can now determine the optimal number of customers to target to achieve the desired profit while adhering to a predetermined budget, as demonstrated by the profit curve presented above. This approach has enabled us to develop an effective credit risk model along with its associated profit curves. Many organizations have successfully utilized this methodology to make informed decisions when granting loans, resulting in substantial profits. Such practices are expected to persist in the future.

At present, numerous sophisticated models exist within the industry. Therefore, the key to further enhancing profits lies in exploring alternative mathematical formulations. By delving into innovative approaches and refining existing models, we can unlock additional opportunities for maximizing profitability. Continuous research and development efforts in this realm will drive advancements in the field, enabling organizations to refine their lending strategies and achieve even greater success.

It is important to acknowledge that while mathematical models provide valuable insights and facilitate decision-making processes, they should be complemented by domain expertise and careful evaluation of real-world factors. Additionally, ongoing monitoring and validation of these models are crucial to ensure their reliability and effectiveness in various scenarios.

In conclusion, by leveraging the power of advanced models and continuously seeking improvements through innovative mathematical formulations, organizations can enhance their loan granting processes, optimize profits, and remain at the forefront of the industry. This pursuit of excellence will continue to shape the future of lending practices, benefiting both financial institutions and borrowers alike.

9. CITATIONS

1. https://www.kaggle.com/
2. https://www.google.com/
3. https://towardsdatascience.com/credit-risk-modeling-with-machine-learning-8c8a2657b4c4
4. https://www.5minutefinance.org/concepts/time-value-of-money-single-cash-flows

APPENDIX

In this appendix, we outline the individual contributions made by each member of our team to the Data Science for Business Project. The project involved various tasks related to data visualization, feature engineering, model selection, decision logic, and business understanding. We believe that acknowledging the specific efforts and expertise of each team member will provide a comprehensive overview of our collective achievements.

Yogya and Noel collaborated closely, combining their respective tasks to ensure a cohesive approach. Yogya primarily focused on data visualization and feature engineering, while Noel spearheaded the model selection process. By working together, they integrated the insights gained from visualization and feature engineering into the model selection process, resulting in a more informed and effective decision-making approach.

Anubhav and Viraj formed a collaborative partnership to tackle their respective tasks. Anubhav's primary responsibility was developing decision logic based on the models' predicted outputs, while Viraj actively collaborated with Anubhav on this aspect. Together, they analyzed the predictions generated by the models and formulated effective decision-making processes that incorporated the insights from the models.

Additionally, it is important to note that each team member made valuable contributions to the business understanding task. Collaboratively, we dedicated time and effort to thoroughly comprehend the domain, identify relevant business challenges, and define clear objectives for our data science project.

We would also like to acknowledge the valuable contributions made by Prof. Foster Provost and express our gratitude for their commitment to our project's success.