Rakamin Academy

id/x partners

LUTFIA HUSNA KHOIRUNNISA

CLIENT'S LOAN PREDICTION PROJECT

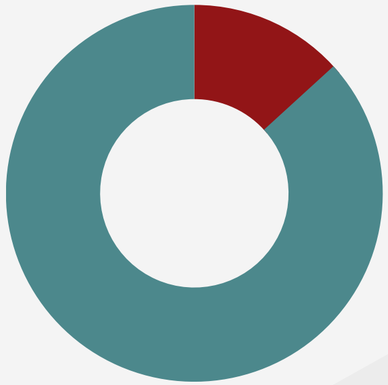
TABLE OF CONTENTS

PROBLEM STATEMENT (INTRODUCTION)

A lending company aims to reduce the number of clients with high risk profiles

HIGH RISK

11.5%



From 1990 to 2014, **there were 11.5% of clients who were classified as high-risk clients**, including clients who defaulted, were charged off, or had late payments.

The company needs to solve this issue because it has the potential to lead to financial problems due to a decrease in company revenue, disrupt the financial stability, and potentially damage the company's reputation

PROBLEM STATEMENT

Role

Data Scientist

Main Problem

Reduce the number of clients with high risk profiles

Business Metrics

Number of High Risk Client

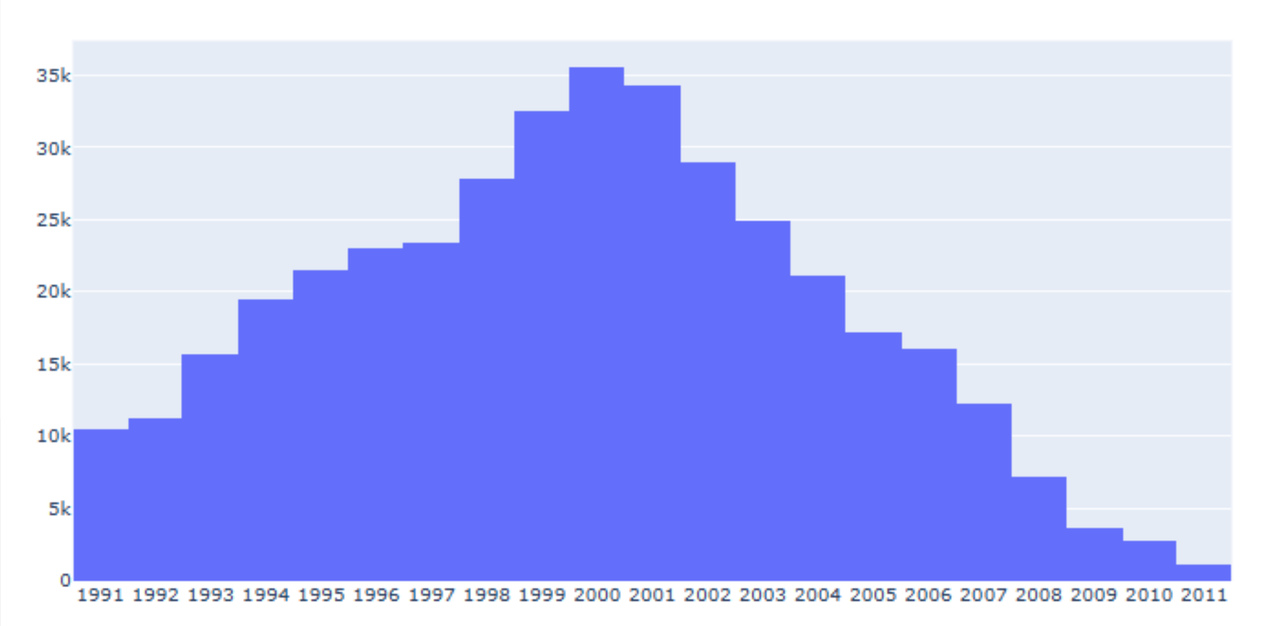
Solution

Creating a model to predict clients who are likely to have a high-risk profile in order to provide them with special treatment and mitigate potential risks.

Click here to access the project file

CLIENT OVERVIEW

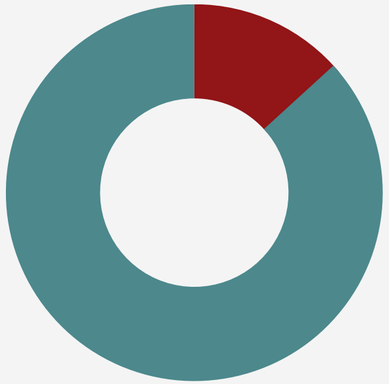
Before we proceed with further modeling, we need to examine the characteristics of the company's clients so that we can get an overall picture of the clients and gain business insights.
To maintain relevance, we will only use data from **the year 1990 to 2014.**



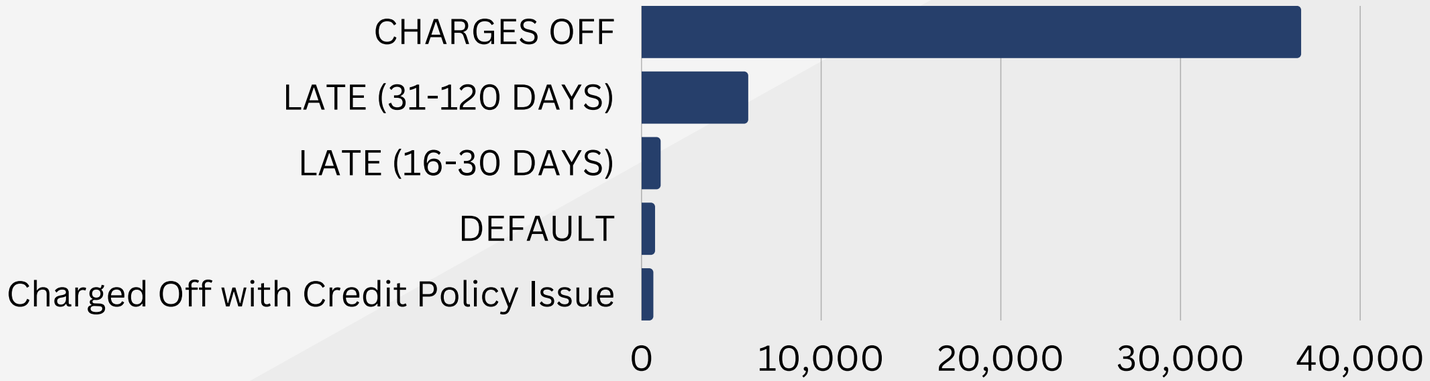
CLIENT OVERVIEW

HIGH RISK

11.5%



From **1990 to 2014,** there were 11.5% of clients who were classified as high-risk clients, including clients who defaulted, were charged off, or had late payments.



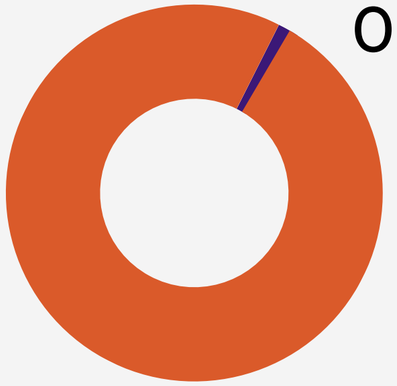
Out of all the high-risk clients, the majority of them have experienced charge-offs, which means they failed to repay their loans and the creditor considers them as losses.

In such cases, the loans are deemed unrecoverable, and the outstanding amount becomes a financial loss for the lender.

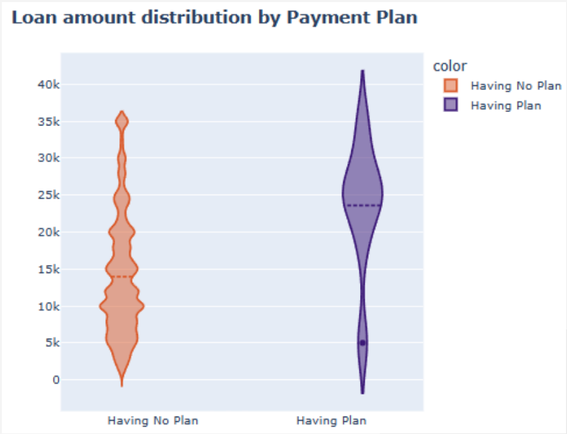
CLIENT OVERVIEW

CLIENT WITH PAYMENT PLAN

0.002%



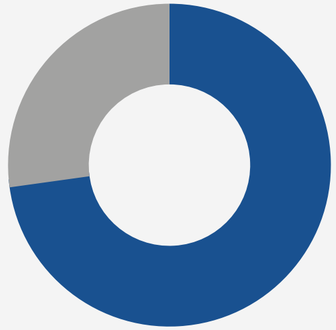
There were only 8 Clients who are having payment plan.



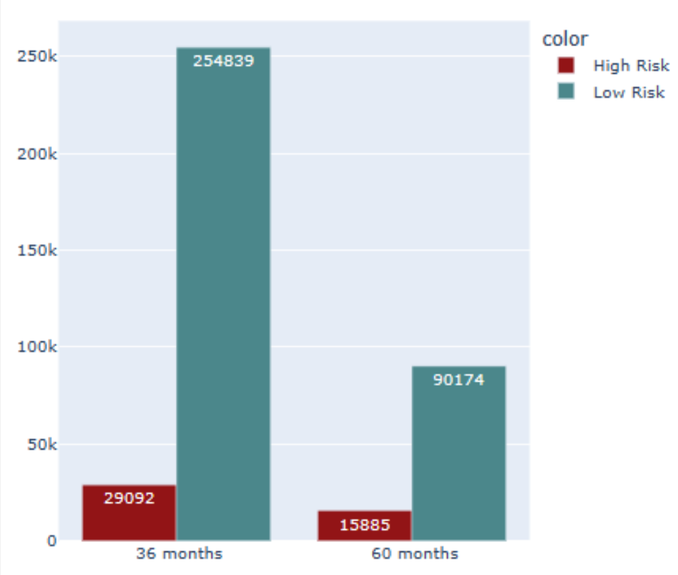
It appears that clients who have a payment plan tend to have higher loan amounts compared to clients who do not have a payment plan.
This maybe because the clients who are taking a larger loan amount typically have a clear plan in place beforehand.

And also clients with a payment plan tend to have higher interest rates compared to clients without a payment plan.
This is because clients with a payment plan often have higher loan amounts, which typically result in higher interest rates.

CLIENT OVERVIEW

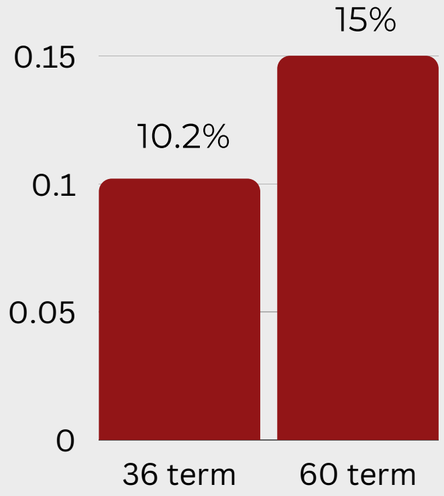


Clients tend to prefer a 36-month term, as it seems that more clients prefer for shorter terms due to lower interest burdens, resulting in lower debt loads.



The group of clients who choose a 36-month term has a higher number of clients with a high risk level.

However, when calculated as a ratio, the group of clients who choose a 60-month term has a higher high-risk rate.



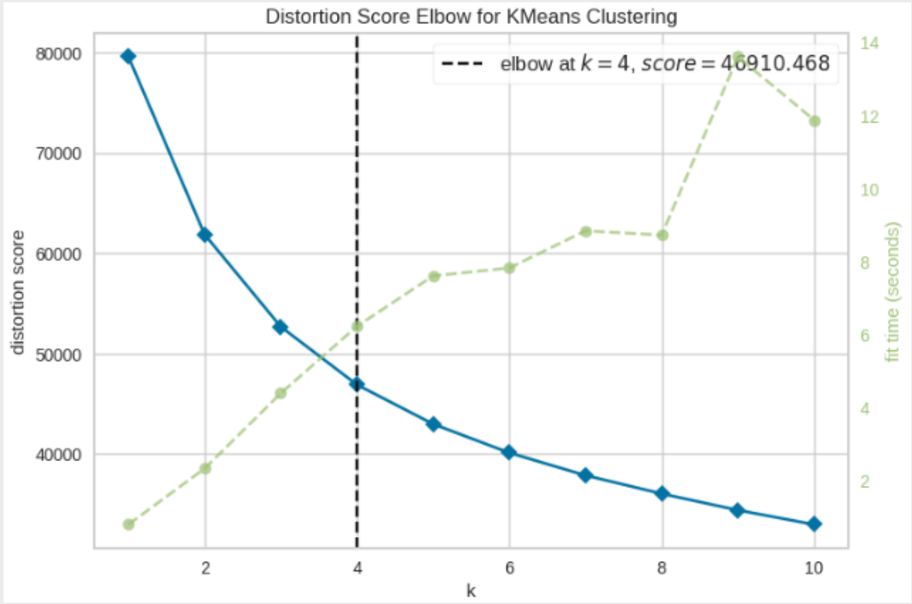
CLIENT OVERVIEW

To enhance the process of identifying clients for the company, a **client segmentation model** is developed using unsupervised learning techniques.

In grouping clients, we will consider factors such as :

- DTI (debt-to-income ratio)
- Annual income
- Interest rate
- Loan amount
- Account recency

Then an elbow test conducted to determine the optimal number of clusters or groups for data segmentation. According to the elbow test, the optimal number of clusters for data segmentation in this dataset is 4.



CLIENT OVERVIEW

Based on the obtained cluster groups, there are 4 client's characteristics:

-Cluster 0-
This cluster consists of clients with high income and low loan amounts, resulting in a low DTI. Clients in this group tend to have low interest rates because of their low loan amounts. Clients in Cluster 1 are considered low-risk clients.

-Cluster 1-
This cluster consists of clients with low annual income and low loan amount, resulting in a low debt-to-income (DTI) ratio. These clients tend to have moderate interest rates and are relatively new compared to other clusters.

-Cluster 2-
This cluster consists of clients with low income but high loan amounts, resulting in a relatively high debt-to-income ratio (DTI). Clients in this group tend to have high interest rates due to their high loan amounts and are likely long-term clients.

-Cluster 3-
This cluster comprises clients with high annual income and high loan amount, resulting in a high DTI ratio. Clients in this cluster typically have high interest rates and have been with the institution for a longer period of time..

CLIENT OVERVIEW

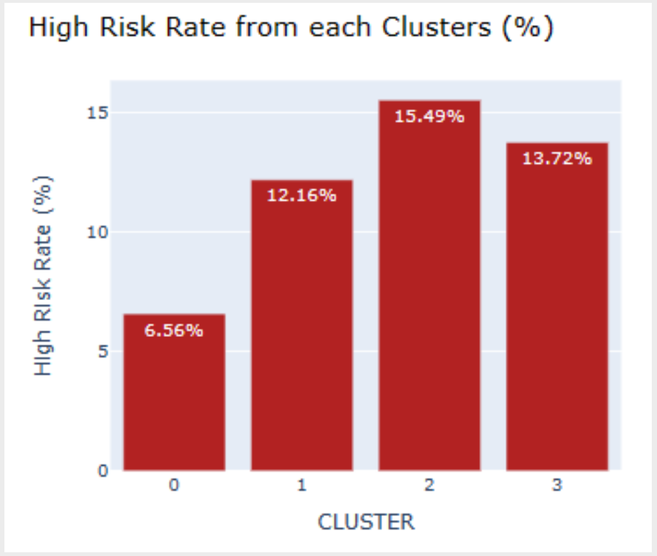
Here is a summary of the mean values for each feature in each cluster.

CLUSTER	annual_inc	dti	int_rate	loan_amnt	recency
0	74239.43	13.31	10.64	11200.0	25.84
1	47346.59	14.17	14.02	8325.0	18.72
2	48806.80	25.31	15.79	11200.0	24.01
3	87998.31	18.03	16.34	24875.0	24.66

The majority of clients are found in Cluster 0 and Cluster 1.

As expected, Cluster 2 has the highest high risk rate. Clients in Cluster 2 tend to take out high loan amounts despite having low incomes.

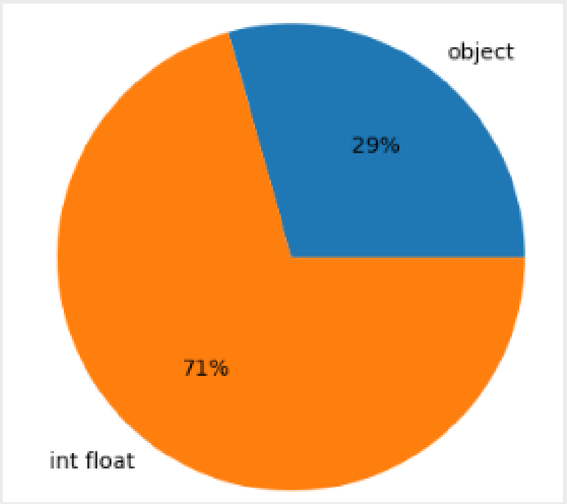
On the other hand, Cluster 0 has the lowest high risk rate. Clients in this cluster tend to take out low loan amounts despite having high incomes.

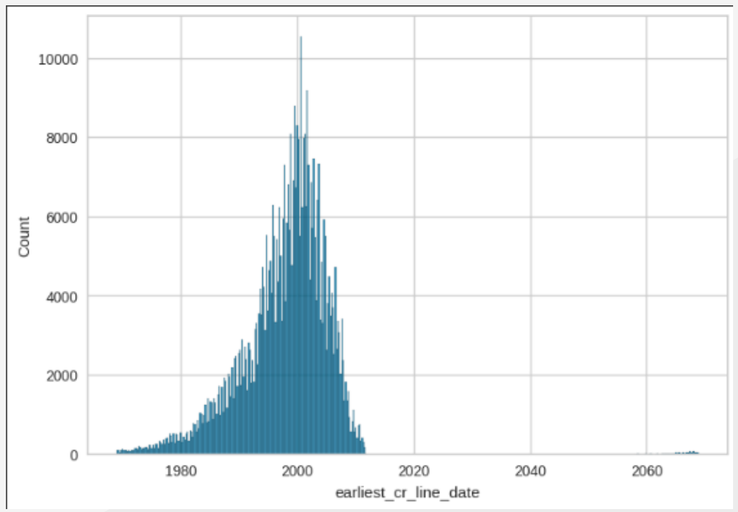


DATA OVERVIEW AND PREPARATION

This project will use a dataset that contains client's profiles and loan histories. There are 75 columns, and 466285 records of data in the dataset.

A total of 29% of the columns in the dataset have the data type "object", while the remaining columns consist of the data types "int" and "float".





The dataset used contains data from 1969, and there are incorrect dates that fall beyond 2023. We will only use data from 1990 to 2014 to maintain relevance.

Click here to access the project file

DATA PREPARATION

Missing Value Handling

Drop columns that have more than 50% missing value, and unused columns with high percentage of missing values,
Missing value handling is also done by perform imputation of missing values using the help of SKLearn's IterativeImputer.

	index	null value	percentage (%)
74	inq_last_12m	389998	100.00
65	dti_joint	389998	100.00
86	total_bal_il	389998	100.00
84	mths_since_rcnt_il	389998	100.00
83	open_il_24m	389998	100.00
82	open_il_12m	389998	100.00
81	open_il_6m	389998	100.00
80	open_acc_6m	389998	100.00
88	open_rv_24m	389998	100.00
89	max_bal_bc	389998	100.00
66	verification_status_joint	389998	100.00
64	annual_inc_joint	389998	100.00
87	open_rv_12m	389998	100.00
70	all_util	389998	100.00
72	inq_fi	389998	100.00
73	total_cu_tl	389998	100.00
85	il_util	389998	100.00
30	mths_since_last_record	339498	87.05
61	mths_since_last_major_derog	311027	79.75
20	desc	284905	73.05
29	mths_since_last_delinq	215674	55.30
48	next_pymnt_d	193567	49.63
71	total_rev_hi_lim	59461	15.25
69	tot_cur_bal	59461	15.25

Label Encoding

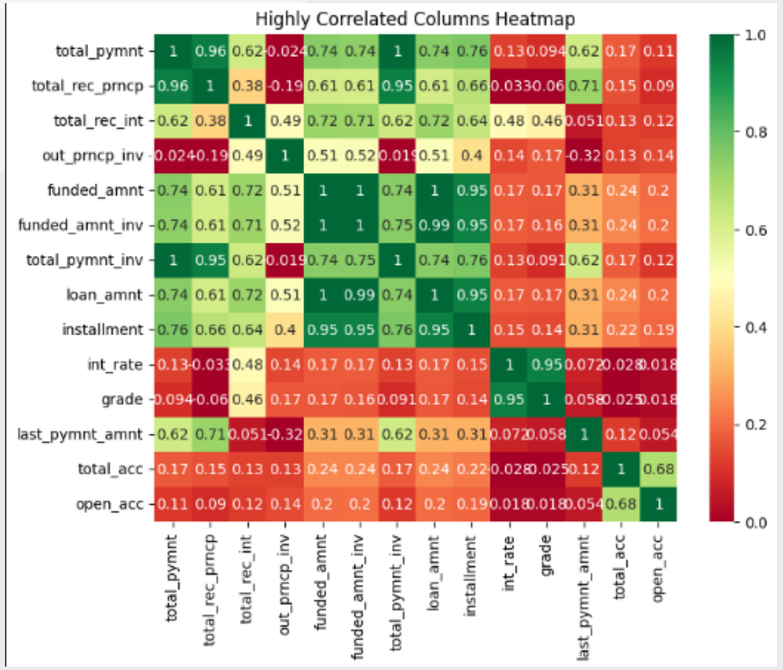
Convert categorical variables into numerical form.

```
df_clean_imputed['term'] = df_clean_imputed['term'].replace({' 36 months':0, ' 60 months':1})
df_clean_imputed['grade'] = df_clean_imputed['grade'].replace({'A':0, 'B':1, 'C':2, 'D':3, 'E':4, 'F':5, 'G':6})
df_clean_imputed['home_ownership'] = df_clean_imputed['home_ownership'].replace({'ANY':0, 'MORTGAGE':1,'OTHER':0, 'OWN':0, 'RENT':1, 'NONE':0})
```

DATA PREPARATION

Handling Multicollinearity

This is done to avoid having redundant features that are highly correlated with each other. This step is performed by checking vif score and dropping features that have high correlation with other features.



Handling Feature with Extreme Unbalanced Value

Drop features with extremely unbalanced values. This step is done to reduce bias caused by unbalanced data.

DATA PREPARATION

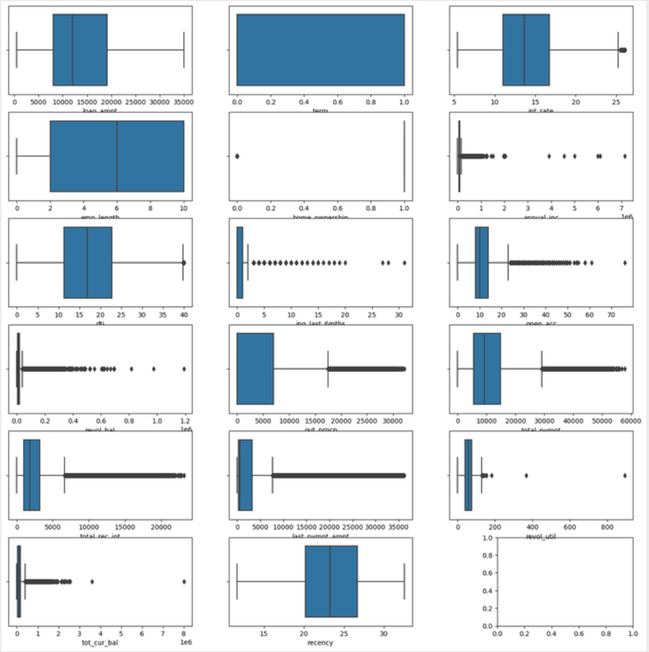
Data Splitting

Splitting the data into training and test sets is done to divide the data that will be used for training and testing. We use 80% of the data for training and 20% for testing.

Data pre train's length is 272998 records, and data test's length is 117000 records.

Handling Outliers

Outlier handling is performed by removing data records with extreme values. This outlier handling is only applied to the training data in order to improve the model's performance and reduce the impact of outliers on the training process. And the final data train's length is 272534 records of data.



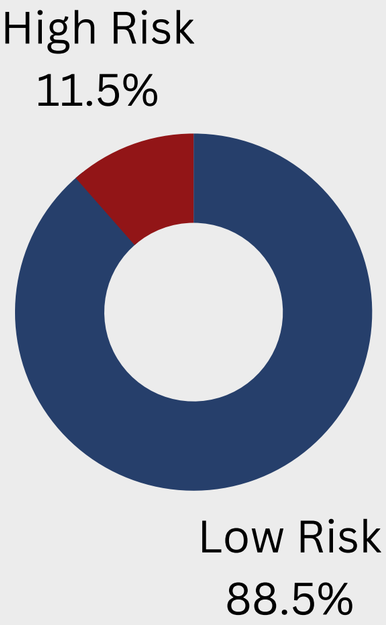
DATA PREPARATION

Feature Transformation

Feature log transformation is performed to handle features with extreme skewness, and MinMax scaling is applied to normalize the entire feature set to ensures that all features are on a similar scale.

Imbalance Class Handling

Imbalance class handling is performed to address the issue of imbalanced class distribution in the dataset. This will be done using undersampling and SMOTE methods. Furthermore, the model results will be compared using the data train without treatment and with imbalance class handling.

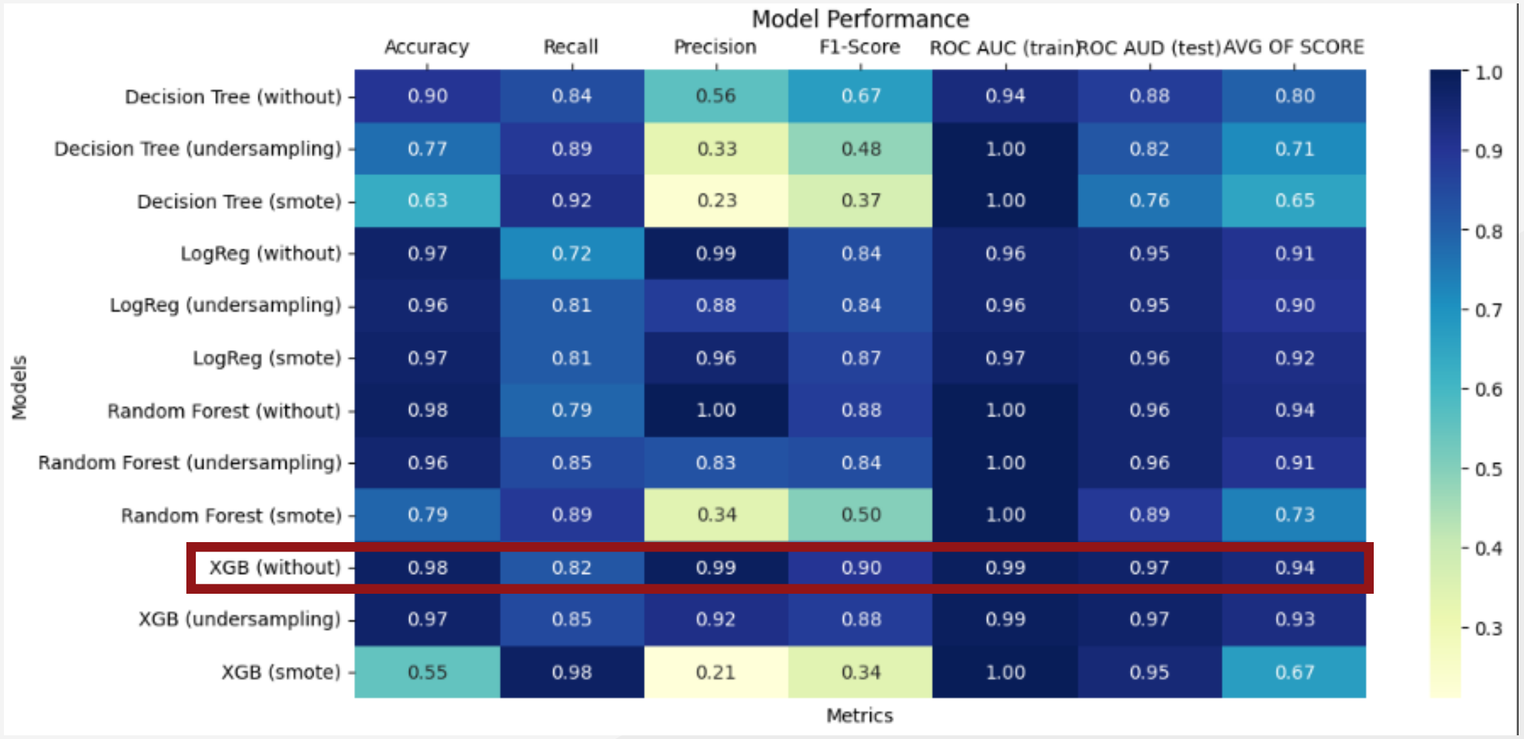


DATA MODELING

Several classification algorithms will be used and their performances will be compared.

The main metric that will be used is 'recall' because we want to minimize false negatives. Therefore, the algorithm with the highest recall value will be selected. However, it is not solely based on the recall value; other metrics will also be taken into consideration to choose the best model.

DATA MODELING



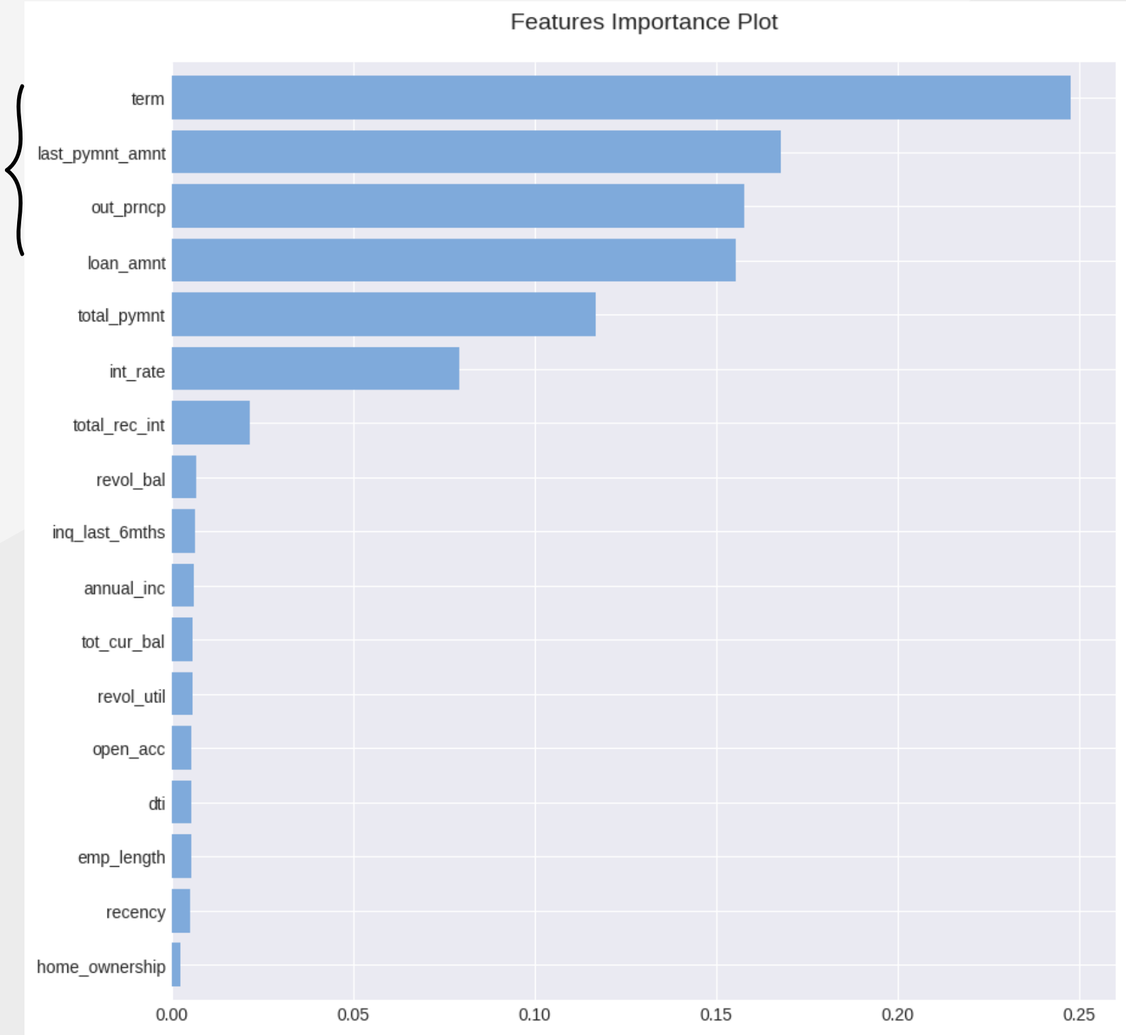
The XGBoost model with data train without imbalance handling has the highest average value across all metrics.

Although the XGBoost model with data train using SMOTE method has a higher recall value, the accuracy of the XGBoost model with data train without imbalance treatment is better. After comparing all the available metrics, we will choose the XGBoost model with data train without imbalance treatment

DATA MODELING

Feature importance

Term, last payment amount, remaining outstanding principal, and loan amount have the most significant influence on the predictions generated by the model.



DATA MODELING

	Actual Low Risk Client	Actual High Risk Client
Predicted Low Risk Client	103203	135
Predicted High Risk Client	2421	11241

Out of the 117,000 clients in the test data, there are 11,241 clients who are correctly predicted to have a high risk level.

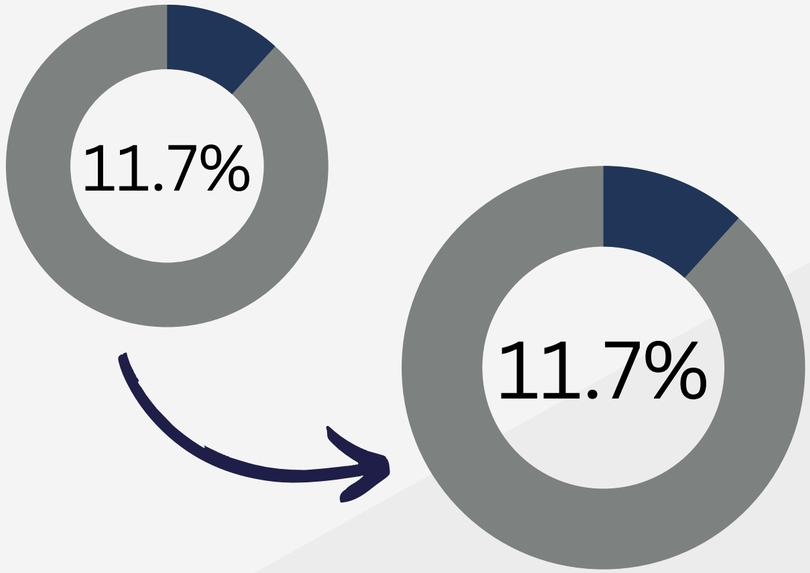
CONCLUSION

- It is found that clients in Cluster 2 have the highest ratio of high-risk clients. Therefore, it is crucial to give special focus to this particular client segment to prevent them from defaulting on their payments
- Based on the observation that the variables Term, Last Payment Amount, Remaining Outstanding Principal, and Loan Amount have the greatest influence on the prediction results, **it is recommended to focus on these factors when making business decisions.**
- It is found that clients in Cluster 2 have the highest ratio of high-risk clients. Therefore, it is crucial to give special focus to this particular client segment to prevent them from defaulting on their payments.

CONCLUSION

(Potential business impact)

To simulate the potential business impact, we will use the 117,000 clients available in the test data.



Assuming the company provides special treatment to clients predicted to be at high risk, and half of the clients correctly predicted to be at high risk are changed to low risk due to our policy, the ratio of high-risk clients can **decrease from 11.7% to 6.9%.**

CONCLUSION

(Business Recomendation)

Providing special treatment to clients predicted to have a high risk will undoubtedly incur costs, and the problem is that companies usually have limited resources.

Therefore, if we want to provide treatment to the right people, it would be better to develop **a credit scoring model** to prioritize clients who are most deserving of treatment.
This allows the company to allocate resources and provide treatment to clients with the highest priority, optimizing the cost-effectiveness of the risk management strategy.

PROFILE

Lutfia Husna K.

Junior Data Analyst

“Lutfi is a junior data analyst experienced on data analysis, business analysis, and data science with a background in mathematics. Experienced in handling and interpreting diverse data sets, extracting valuable insights, and making datadriven recommendations ”

Check out my profile and other portfolios:

Departments/Teams

Linkedin

GitHub