HOME CREDIT DEFAULT RISK MODEL

PROJECT REPORT

Group 14

Anirudha Raghava Sarma Kuchibhotla
Mobile: (617)-238-4850
Email ID: kuchibhotla.an@northeastern.edu

Kushagra Trivedi
Mobile: (617)-516-4269
Email ID: trivedi.ku@northeastern.edu

Shivie Saksenaa
Mobile: (857)-961-9549
Email ID: saksenaa.s@northeastern.edu

Percentage of Effort Contributed by Student 1: 33.33%
Percentage of Effort Contributed by Student 2: 33.33%
Percentage of Effort Contributed by Student 3: 33.33%

Signature of Student 1: Anirudha
Signature of Student 2: Kushagra
Signature of Student 3: Shivie

Submission Date: April 23, 2024

INDEX

Sno.	Description	Page Number
1	Problem Setting	3
2	Problem Definition	4
3	Data Source and Description	5
4	Exploratory Data Analysis	6
5	Model Preprocessing	13
6	Data Modeling	16
7	Results	22
8	Conclusion	24

PROBLEM SETTING

In the domain of financial services, the concept of financial inclusion has emerged as a critical imperative driven by the recognition that access to basic banking and credit facilities is a fundamental right and a key driver of economic development. Despite significant advancements in global financial systems, a substantial portion of the population remains excluded from formal banking channels, primarily due to socio-economic disparities, geographic isolation, or lack of documentation.

Traditionally, conventional lenders have relied heavily on established credit scoring models that primarily assess an individual's creditworthiness based on their credit history, income, and other traditional financial indicators. While effective for a sizable portion of the population, these models often fail to capture the creditworthiness of individuals with limited or no credit history for example, students for whom affordable credit can enable them to invest in education and entrepreneurship thereby breaking the cycle of poverty and fostering economic growth at the grassroots level.

Alternative lenders such as Home Credit have emerged with a mission to address the needs of the unbanked and underbanked population segments. These lenders recognize the inherent limitations of traditional credit scoring methods and seek innovative approaches to assess the creditworthiness of individuals who may not fit within conventional criteria.

However, extending credit to individuals with limited credit history poses unique challenges. One of the primary challenges is the risk of default, as traditional indicators may not accurately predict the repayment behavior of borrowers with sparse credit profiles. By harnessing alternative data sources and deploying innovative lending strategies, Home Credit endeavors to bridge the gap between the unbanked population and formal financial services, empowering individuals to realize their full potential and contribute to inclusive economic development.

PROBLEM DEFINITION

In the scope of our analysis, we are delving into the multifaceted factors contributing to payment difficulties among borrowers to enrich our understanding and refine our lending practices. First and foremost, we are scrutinizing demographic and socio-economic factors such as gender, age, occupation, education level, and family status to discern any correlations with payment difficulties. Additionally, we are exploring the impact of loan characteristics, including the amount of credit extended, loan type (cash loans vs. revolving loans), and the timing of loan processing (process days), to gauge their influence on repayment behavior. Moreover, we are investigating how employment status intersects with payment difficulties, identifying differences across occupational fields, and evaluating the payment behaviors of borrowers associated with specific organizations such as military or police.

Furthermore, we are analyzing repayment abilities concerning living arrangements, including suite type and housing status, to understand how these factors affect borrowers' capacity to meet their obligations. Through comparative analysis, we aim to elucidate disparities in payment difficulties between different loan types and borrower profiles. Additionally, robust data analysis is crucial for developing a comprehensive Home Credit model. It provides the necessary background information and insights to inform proper modeling techniques, ensuring that predictive models accurately capture the nuances of borrower behavior and enable proactive risk management strategies. By leveraging data analytics to its fullest extent, we can refine our lending criteria, optimize risk assessment processes, and ultimately improve the overall borrowing experience for our clients.

DATA SOURCE AND DESCRIPTION

The data for this project is sourced from the Home Credit Default Risk Kaggle competition dataset(https://www.kaggle.com/competitions/home-credit-credit-risk-model-stability/data). It encompasses a variety of tables derived from both internal and external sources, detailing the financial and personal information of clients. These tables are available in both .csv and. parquet formats, facilitating ease of access and analysis.

The dataset was about 24GB, we tried pre processing it but due to the huge volume of the data it became really difficult to work on it due to limited computing capabilities of our system. We then switched to the earlier version of the same dataset. The link to the dataset is provided below:

https://www.kaggle.com/competitions/iiitb2020-home-credit-default-risk/data

The dataset contains 122 columns and 199882 rows, providing a comprehensive array of features for developing a Home Credit model. These features encompass various aspects such as occupation type, gender, car ownership, family status, and age, among others. The target variable called ‘TARGET’ classifies individuals into two categories: Default cases (represented by 1) and Non Default cases (represented by 0). Default cases are those that have had difficulty with loan repayment while non-default did not.

This dataset offers a rich source of information for constructing predictive models aimed at assessing creditworthiness and mitigating default risks, ultimately facilitating Home Credit's mission of promoting financial inclusion and empowering individuals to achieve their financial goals.

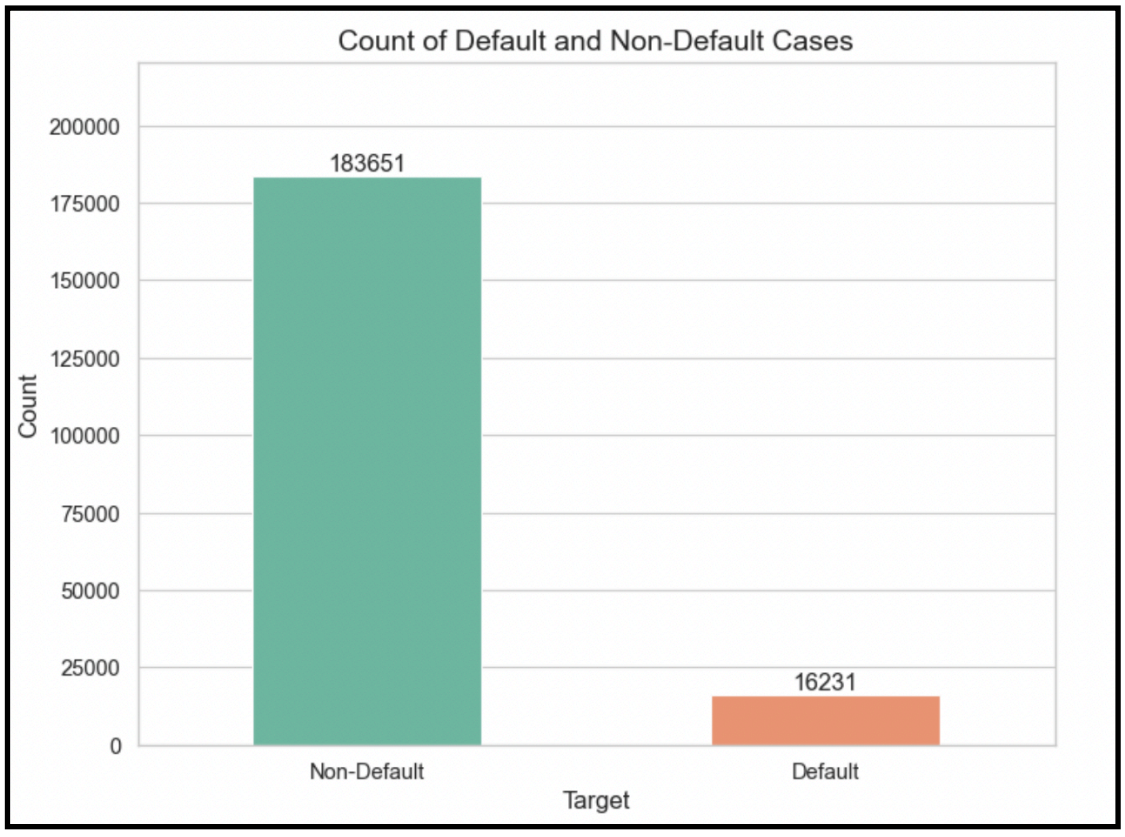
EXPLORATORY DATA ANALYSIS

We performed a univariate, bivariate and multivariate analysis through the visualizations like histograms, bar charts and box plots.

With univariate analysis we see the distribution of target variable. We see how the target is distributed with respect to some features in bivariate analysis and use a combination of features to see the distribution in multivariate analysis.

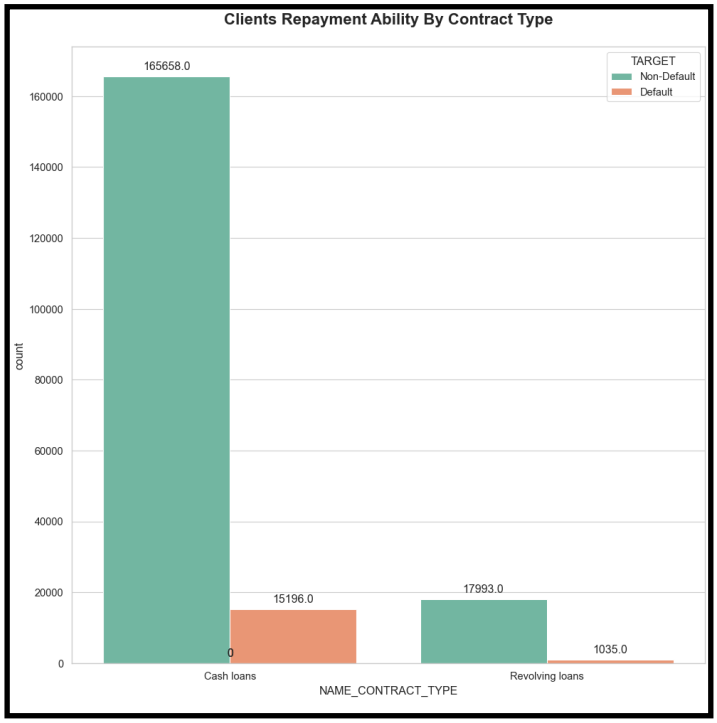
Checking For Class Imbalance

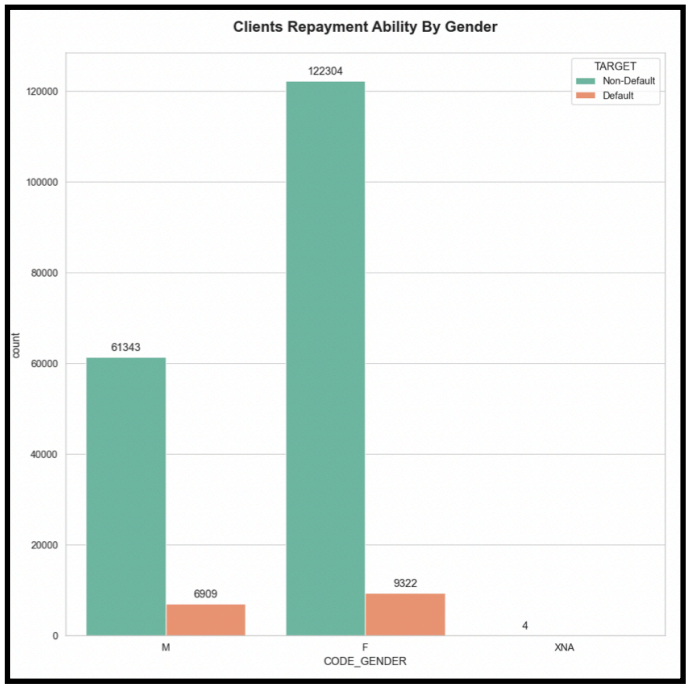
The following histogram depicts the total number of Defaulters and No defaulters.



- There is a much higher number of Non-Default cases (Class 0) compared to default cases (Class 1): 183,651 vs 16,231.
- Thus, we see a **large class imbalance**. We will need to address this before training our model by oversampling the minority class using methods like SMOTE.

Exploring The Distribution Of Target Variable For Different Categorical Features.



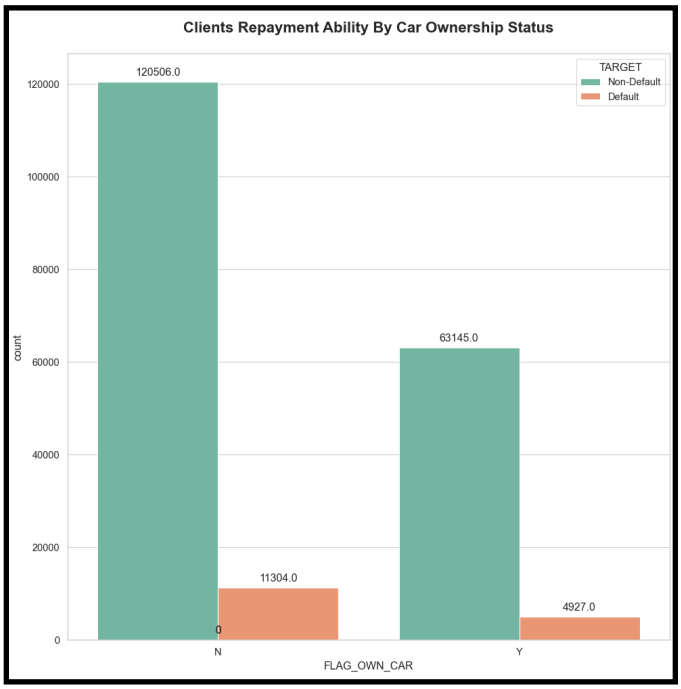


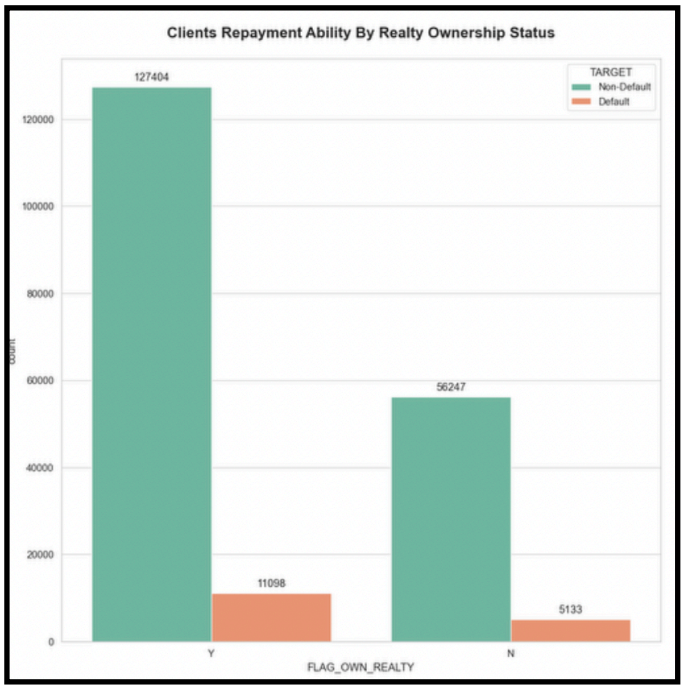
TARGET	Default	Non-Default
NAME_CONTRACT_TYPE		
Cash loans	8.40	91.60
Revolving loans	5.44	94.56

TARGET	Default	Non-Default
CODE_GENDER		
Female	7.08	92.92
Male	10.12	89.88

- We observe that there is a much larger number of clients taking out cash loans compared to revolving loans. Cash loans also have a higher default rate (8.4%) compared to revolving loans (5.4%)

- It can be seen that women have applied for a larger majority of loans. In total, there are about 202,448 loan applications applied by women, and about 105,059 applications applied by men. However, a larger percentage (about 10% of the total) of men had the problems repaying the loan compared to women clients (about 7%)



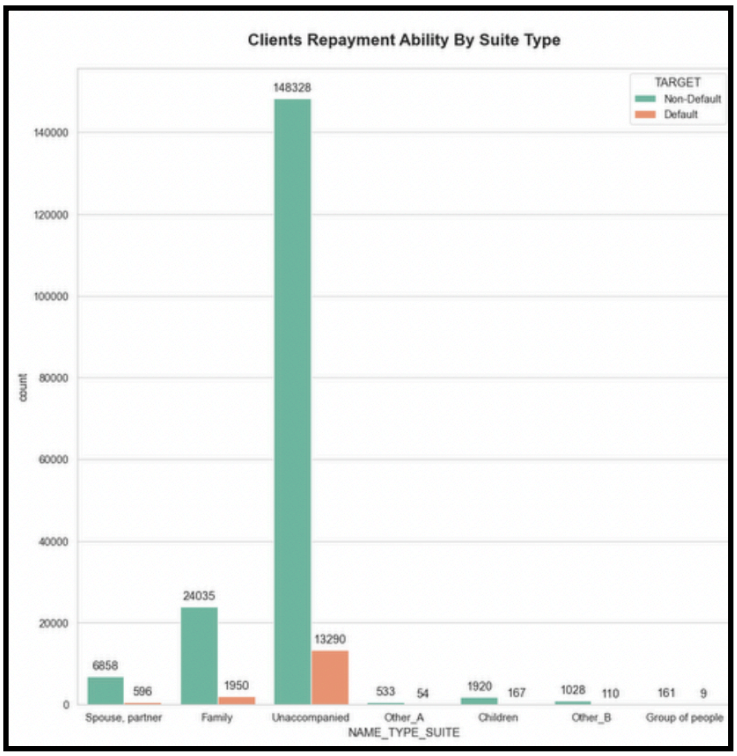


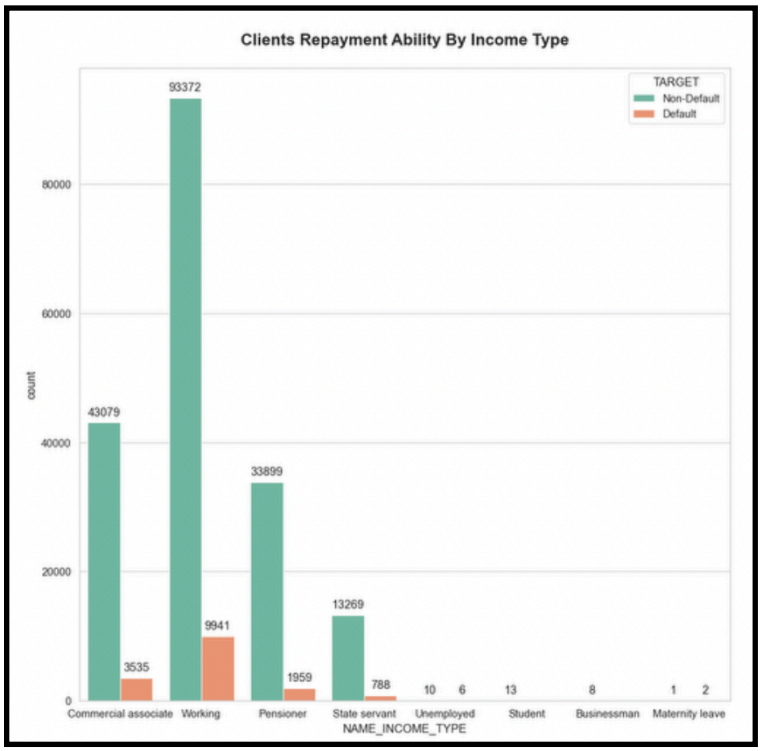
TARGET FLAG_OWN_CAR	Default	Non-Default
N	8.58	91.42
Y	7.24	92.76

TARGET FLAG_OWN_REALTY	Default	Non-Default
N	8.36	91.64
Y	8.01	91.99

- Most of the clients do not own a car, but there is a higher number of non car owners applying for loan compared to car owners. The default rate is very similar for both car owners and non car owners. Clients who own a car (about 8%) had the problems in repaying the loan compared to clients who not own a car (about 7%).

- Larger number of realty owners have applied for loans. default rate is very similar for both owners and non-owners





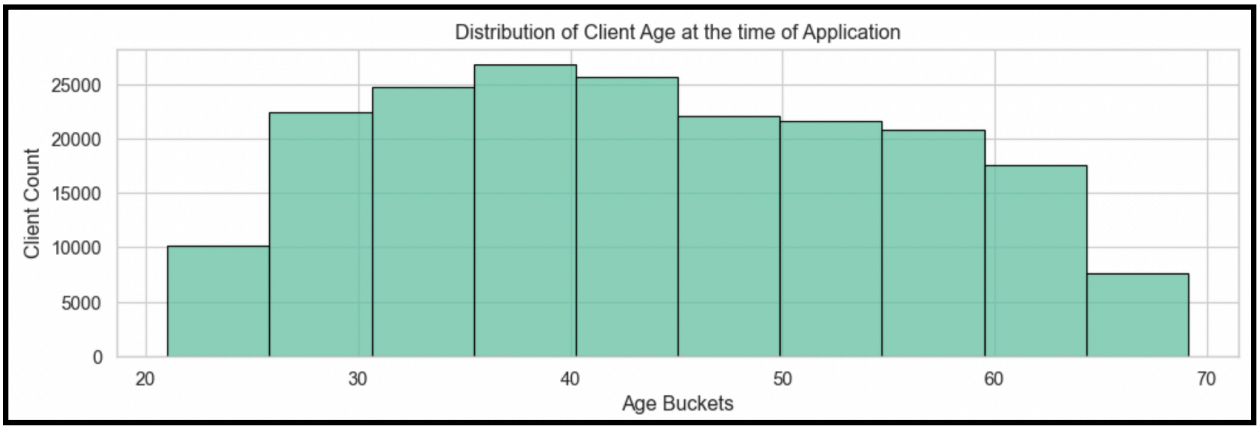
NAME_TYPE_SUITE	TARGET Default	Non-Default
Children	8.00	92.00
Family	7.50	92.50
Group of people	5.29	94.71
Other_A	9.20	90.80
Other_B	9.67	90.33
Spouse, partner	8.00	92.00
Unaccompanied	8.22	91.78

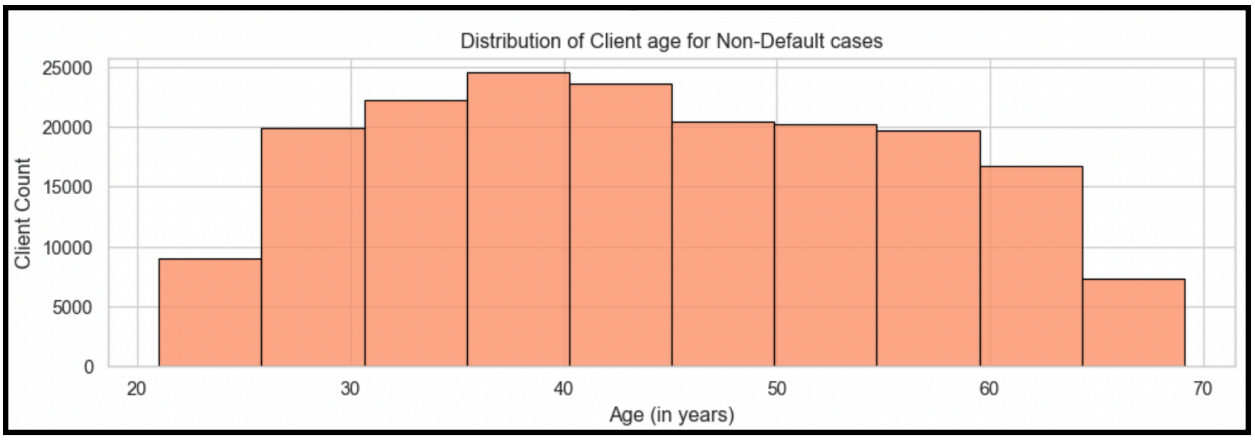
NAME_INCOME_TYPE	TARGET Default	Non-Default
Businessman	NaN	100.00
Commercial associate	7.58	92.42
Maternity leave	66.67	33.33
Pensioner	5.46	94.54
State servant	5.61	94.39
Student	NaN	100.00
Unemployed	37.50	62.50
Working	9.62	90.38

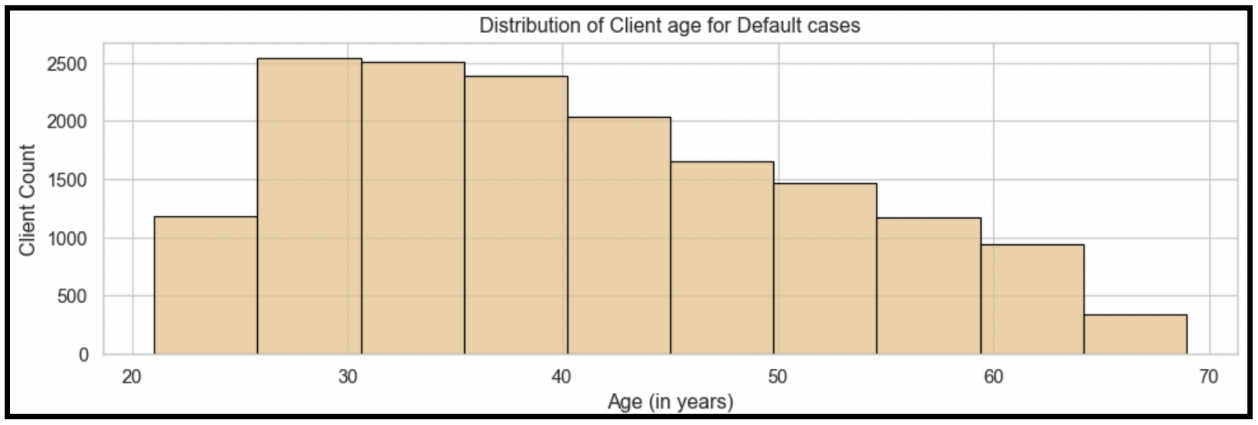
- A large majority of applicants are have Unaccompanied suite type. The default rate for group of people is much lower that other suite types at 5.29%

- Most common applicants have income type working followed by commercial associates, pensioners and state servants. We see very high default percentage for income types 'Maternity Leave' and 'Unemployed', but those are due to very low number of datapoints. Surprisingly, the most common income type (Working) has a higher than average default rate of 9.62%.

Client Age Distribution: Total vs Default Cases vs Non-Default Cases

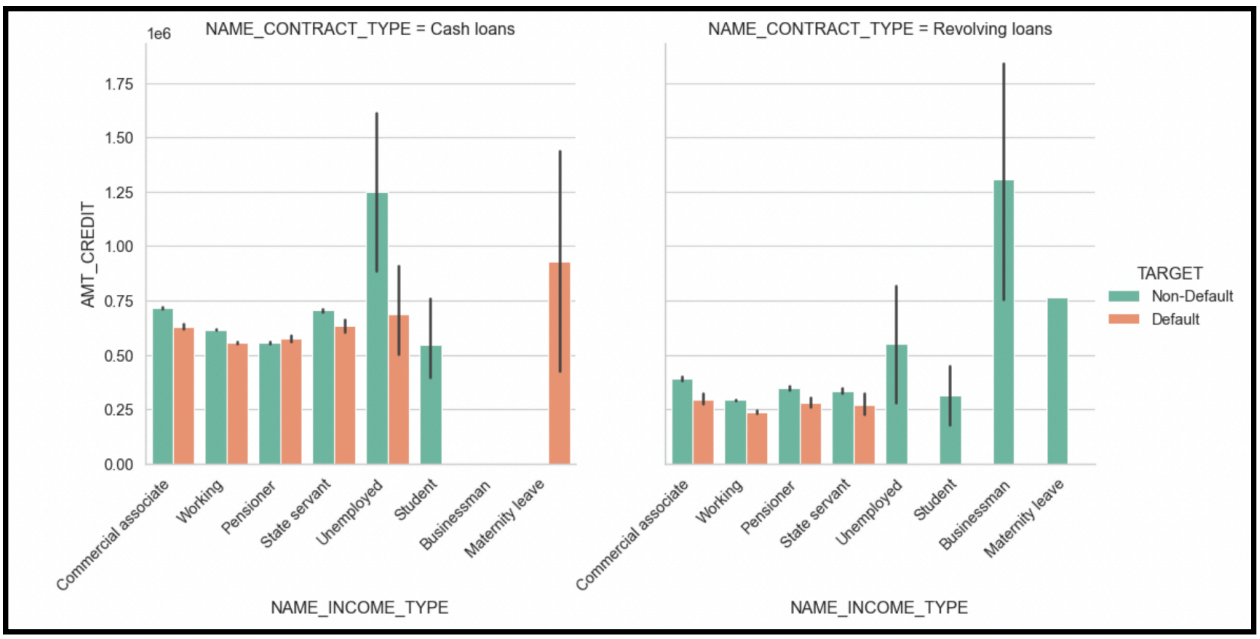






- The number of clients who apply for loans is highest for age range of 35-40 years, followed by clients in the range of 40-45 years. The number of applicants for clients at extremes (age <25 or age >65) is much lower.
- The distribution of ages for non default cases is very similar to overall age distribution with maximum clients age range of 35-40 years.
- The distribution for defaulting cases, however, is right skewed with highest number in range of 25-35 years.

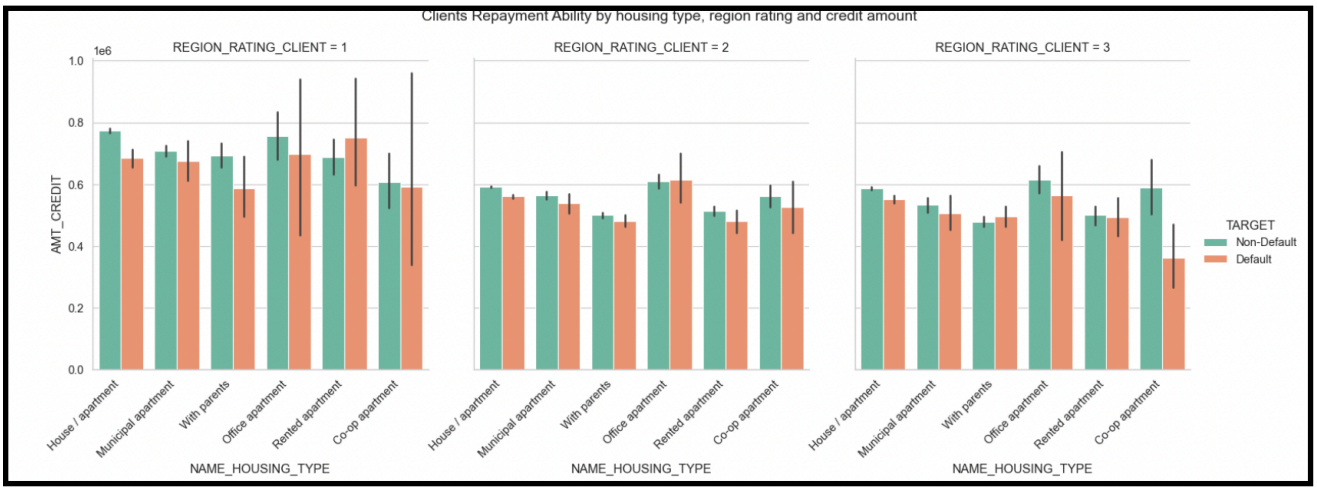
Repayment Ability Across Income, Contract, and Credit Amount



- All of the clients who have maternity leave income types and cash loans have trouble paying back the loans for a medium credit amount. All consumers who have revolving loans and maternity leaves are able to settle their debts without any trouble at all.
- More than half of unemployed clients with cash loans struggle to return loans with modest credit amounts. On the other hand, all unemployed clients who have revolving loans have no trouble paying them back.

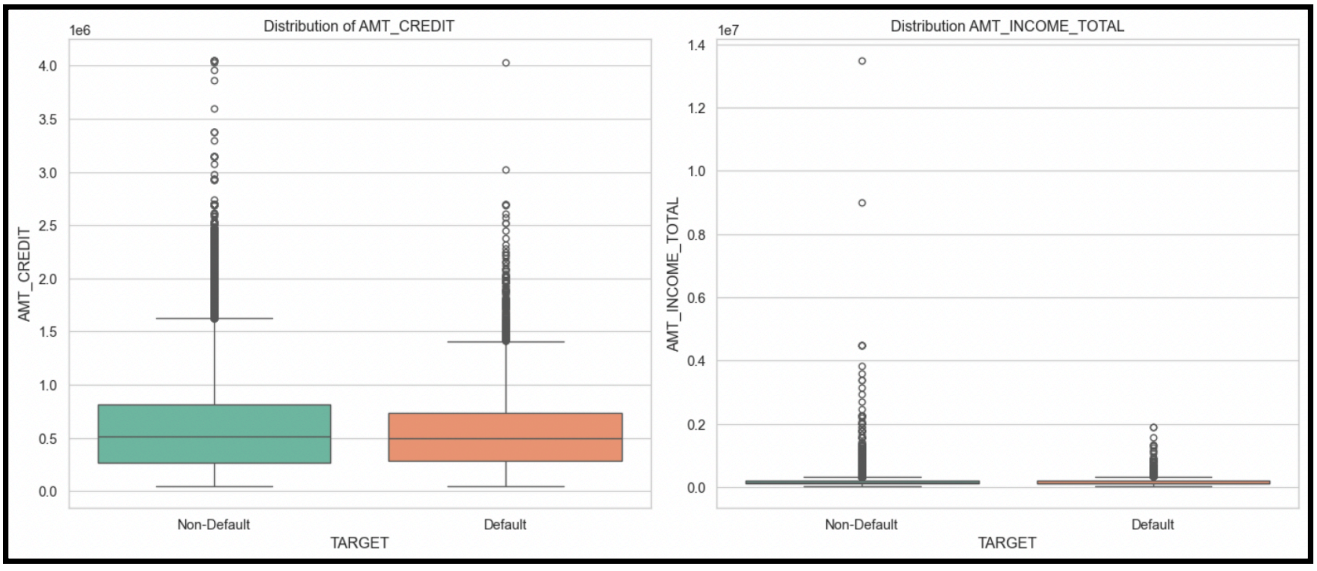
- For low to medium credit amounts, all student clients have no trouble repaying their loans, whether they are cash loans or revolving loans.

Client Repayment Proficiency: Housing, Region, and Credit Amount



- Clients who lives in rented apartment and office apartment and their region have a rating of 1, have a problem repaying the loans compared to client in region with rating of 2 for a medium credit amount of the loan.

Distribution of Amount Credit and Total Income by Default Status



- The median credit amount of the cases that did not default is slightly higher than the median value of cases that defaulted. We can conclude that people who were approved for a lighter loan amount the chances of repayment are also higher. Conversely, we see more defaults for lower credit amounts
- It can be seen that both clients who have payment difficulties as well as clients who have no payment difficulties, have mostly the same value.
- However, as the client's income increases, the probability that clients will not have difficulty repaying the loan also increases.

DATA PRE PROCESSING

Data Cleaning

- Dealt with null values
    - Removed the columns containing more than 50% null values
    - Filled the remaining columns as follows : categorical columns with mode and numerical columns with median.

```
Columns Contaning more than 50% nulls (To be removed):

OWN_CAR_AGE                 65.945908
EXT_SOURCE_1                56.434797
APARTMENTS_AVG              50.837494
BASEMENTAREA_AVG            58.590068
YEARS_BUILD_AVG             66.572278
COMMONAREA_AVG              69.949770
ELEVATORS_AVG               53.388499
ENTRANCES_AVG               50.405239
FLOORSMIN_AVG               67.931580
LANDAREA_AVG                59.439569
LIVINGAPARTMENTS_AVG        68.393852
LIVINGAREA_AVG              50.269659
NONLIVINGAPARTMENTS_AVG     69.465985
NONLIVINGAREA_AVG           55.250098
APARTMENTS_MODE             50.837494
BASEMENTAREA_MODE           58.590068
YEARS_BUILD_MODE            66.572278
COMMONAREA_MODE             69.949770
```

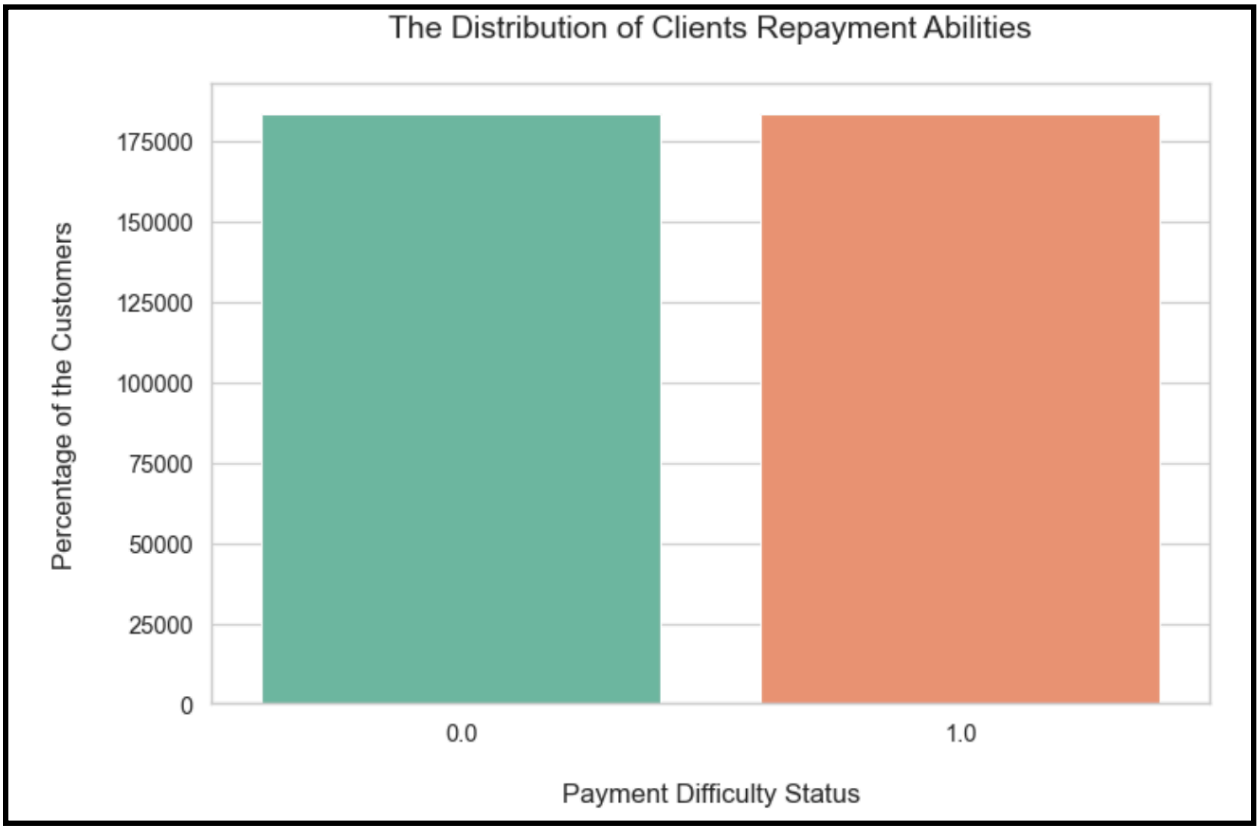
Data Encoding

- The Categorical Features That Were Remaining In The Dataset Were Non-Ordinal. For Example, Contract Type Of 'Cash' And 'Recurring' Do Not Have Any Meaningful Order. Hence We Will Go Ahead And Perform One-Hot Encoding.

```
{'NAME_CONTRACT_TYPE': 2,
 'CODE_GENDER': 3,
 'FLAG_OWN_CAR': 2,
 'FLAG_OWN_REALTY': 2,
 'NAME_TYPE_SUITE': 7,
 'NAME_INCOME_TYPE': 8,
 'NAME_EDUCATION_TYPE': 5,
 'NAME_FAMILY_STATUS': 6,
 'NAME_HOUSING_TYPE': 6,
 'OCCUPATION_TYPE': 18,
 'WEEKDAY_APPR_PROCESS_START': 7,
 'ORGANIZATION_TYPE': 58,
 'EMERGENCYSTATE_MODE': 2}
```

Handling Imbalance With Data

- To address data imbalance within the dataset, Synthetic Minority Over-sampling Technique (SMOTE) was employed, effectively augmenting the minority class (Defaulters) instances. This approach will eventually enhance the model performance by ensuring a more balanced representation across the target variable categories(Default or No Default) , thereby improving the reliability of the predictive model.



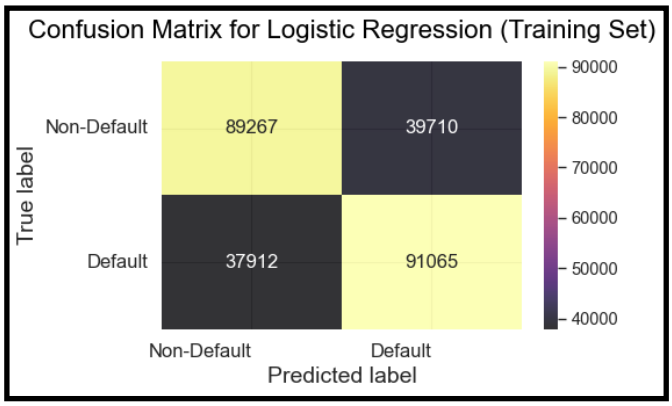
Train-Validation Split

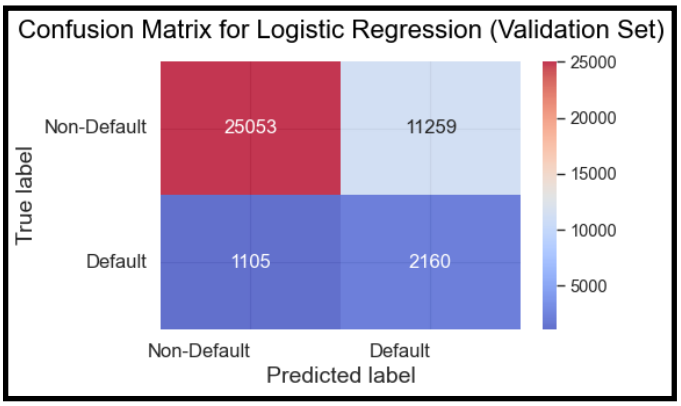
- The dataset was then partitioned into training and validation sets, with an 80-20 split, to facilitate model development and evaluation. There was no data leakage from training to testing sets as we handled them separately.
- We also normalized the features using standard scaler approach.

DATA MODELING

- We constructed baseline models using various algorithms to assess their performance on our dataset. This comparison helped us identify the most promising approach suitable for this problem.

Logistic Regression

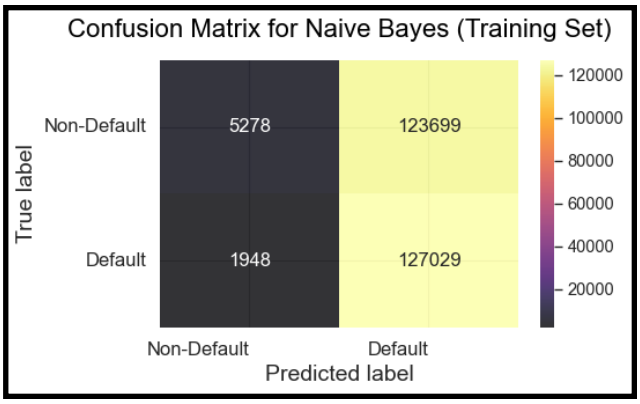


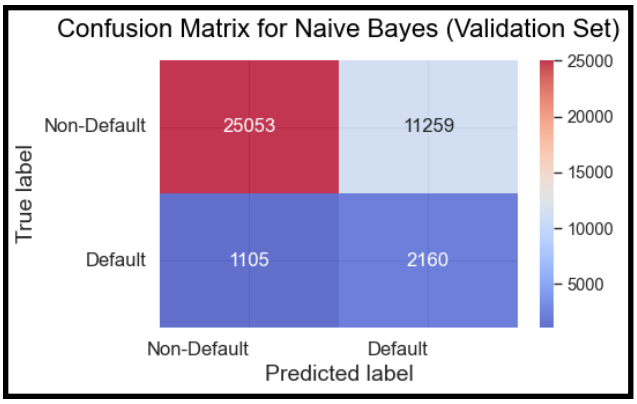



```
Logistic Regression Performance on Training Set:
              precision    recall  f1-score   support

         0.0       0.70      0.69      0.70    128977
         1.0       0.70      0.71      0.70    128977

    accuracy                           0.70    257954
   macro avg       0.70      0.70      0.70    257954
weighted avg       0.70      0.70      0.70    257954


Logistic Regression Performance on Validation Set:
              precision    recall  f1-score   support

         0.0       0.96      0.69      0.80     36312
         1.0       0.16      0.66      0.26      3265

    accuracy                           0.69     39577
   macro avg       0.56      0.68      0.53     39577
weighted avg       0.89      0.69      0.76     39577
```

- Training Accuracy: % 69.64
- Validation Accuracy: % 69.5
- ROC AUC: 0.695

Naive Bayes





Naive Bayes Performance on Training Set:

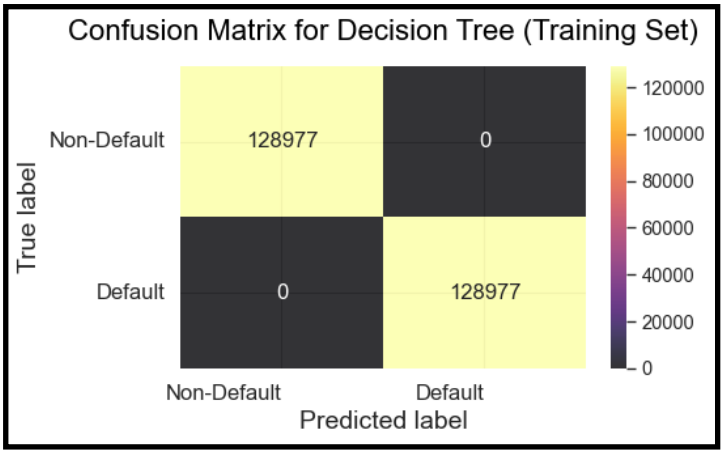
	precision	recall	f1-score	support
0.0	0.73	0.04	0.08	128977
1.0	0.51	0.98	0.67	128977
accuracy			0.51	257954
macro avg	0.62	0.51	0.37	257954
weighted avg	0.62	0.51	0.37	257954

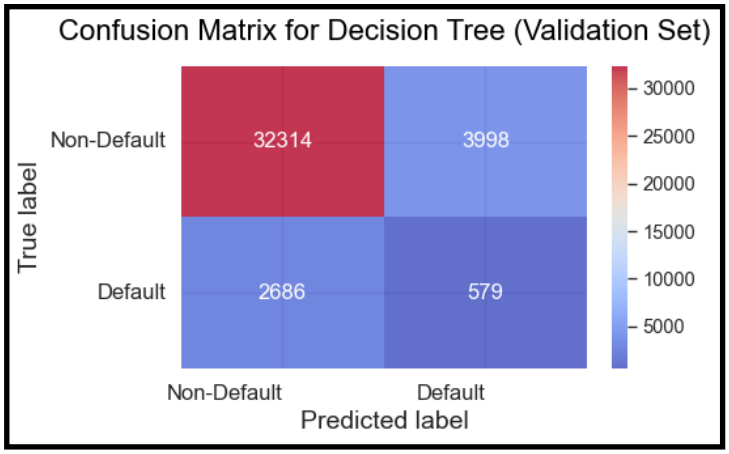
Naive Bayes Performance on Validation Set:

	precision	recall	f1-score	support
0.0	0.95	0.04	0.07	36312
1.0	0.08	0.98	0.15	3265
accuracy			0.12	39577
macro avg	0.52	0.51	0.11	39577
weighted avg	0.88	0.12	0.08	39577

- Training Accuracy: % 69.64
- Validation Accuracy: % 69.5
- ROC AUC: 0.695

Decision Tree

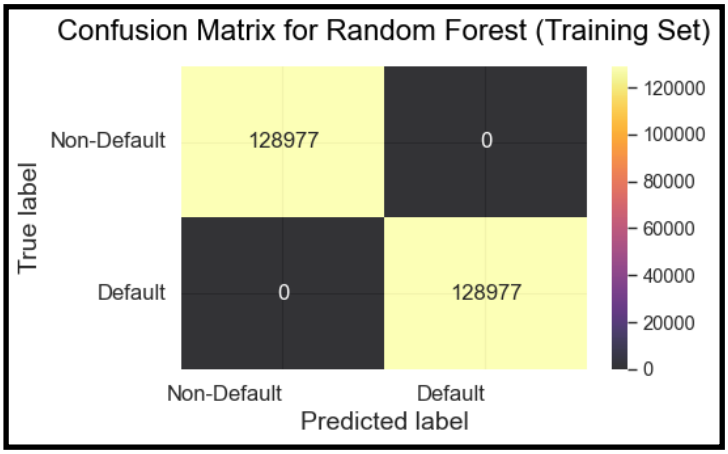


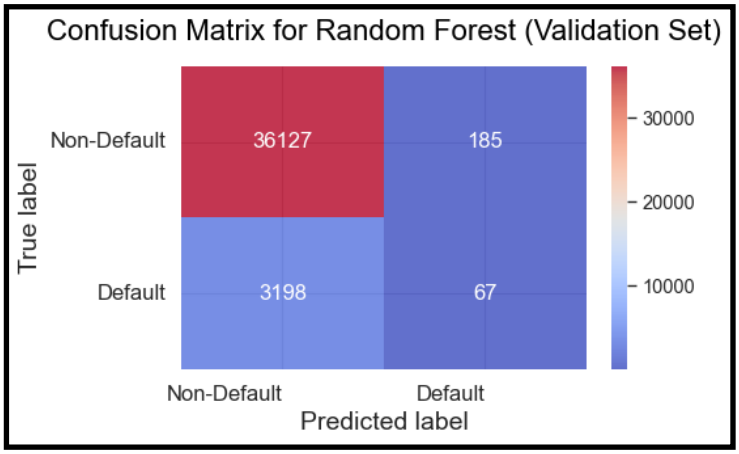



```
Decision Tree Performance on Training Set:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00    128977
         1.0       1.00      1.00      1.00    128977

    accuracy                           1.00    257954
   macro avg       1.00      1.00      1.00    257954
weighted avg       1.00      1.00      1.00    257954


Decision Tree Performance on Validation Set:
              precision    recall  f1-score   support

         0.0       0.92      0.89      0.91     36312
         1.0       0.13      0.18      0.15      3265

    accuracy                           0.83     39577
   macro avg       0.52      0.53      0.53     39577
weighted avg       0.86      0.83      0.84     39577
```

- Training Accuracy: % 69.64
- Validation Accuracy: % 69.5
- ROC AUC: 0.695

Random Forest

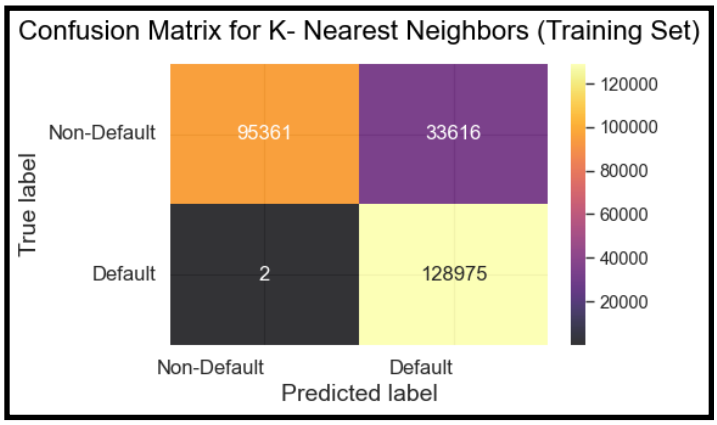


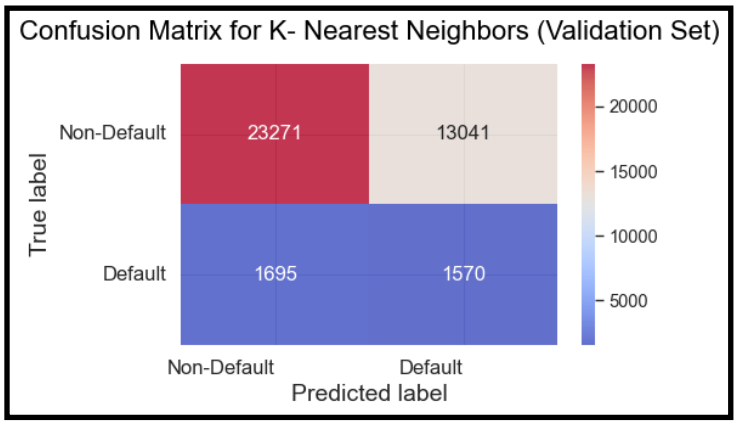



```
Random Forest Performance on Training Set:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00    128977
         1.0       1.00      1.00      1.00    128977

    accuracy                           1.00    257954
   macro avg       1.00      1.00      1.00    257954
weighted avg       1.00      1.00      1.00    257954


Random Forest Performance on Validation Set:
              precision    recall  f1-score   support

         0.0       0.92      0.99      0.96     36312
         1.0       0.27      0.02      0.04      3265

    accuracy                           0.91     39577
   macro avg       0.59      0.51      0.50     39577
weighted avg       0.86      0.91      0.88     39577
```

- Training Accuracy: 69.64%
- Validation Accuracy: 69.5%
- ROC AUC: 0.695

K-Nearest Neighbors





KNN Performance on Training Set:

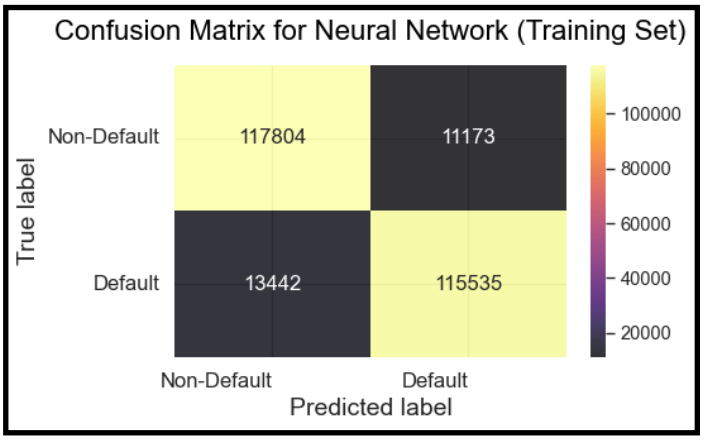
	precision	recall	f1-score	support
0.0	1.00	0.74	0.85	128977
1.0	0.79	1.00	0.88	128977
accuracy			0.87	257954
macro avg	0.90	0.87	0.87	257954
weighted avg	0.90	0.87	0.87	257954

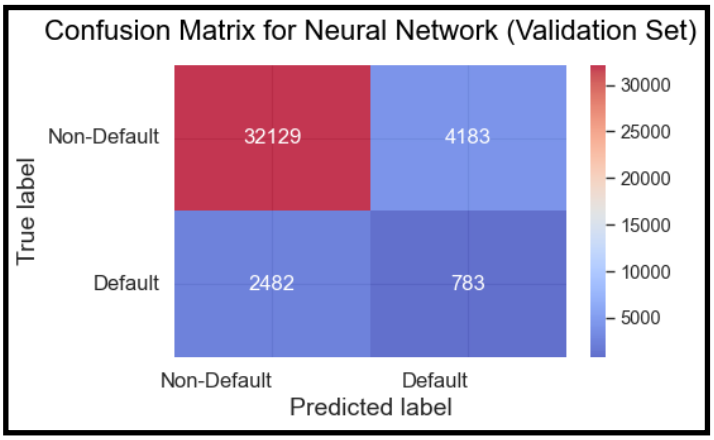
KNN Performance on Validation Set:

	precision	recall	f1-score	support
0.0	0.93	0.64	0.76	36312
1.0	0.11	0.48	0.18	3265
accuracy			0.63	39577
macro avg	0.52	0.56	0.47	39577
weighted avg	0.86	0.63	0.71	39577

- Training Accuracy: 69.64
- Validation Accuracy: % 69.5
- ROC AUC: 0.695

Neural Network





Neural Network Performance on Training Set:

	precision	recall	f1-score	support
0.0	0.90	0.91	0.91	128977
1.0	0.91	0.90	0.90	128977
accuracy			0.90	257954
macro avg	0.90	0.90	0.90	257954
weighted avg	0.90	0.90	0.90	257954

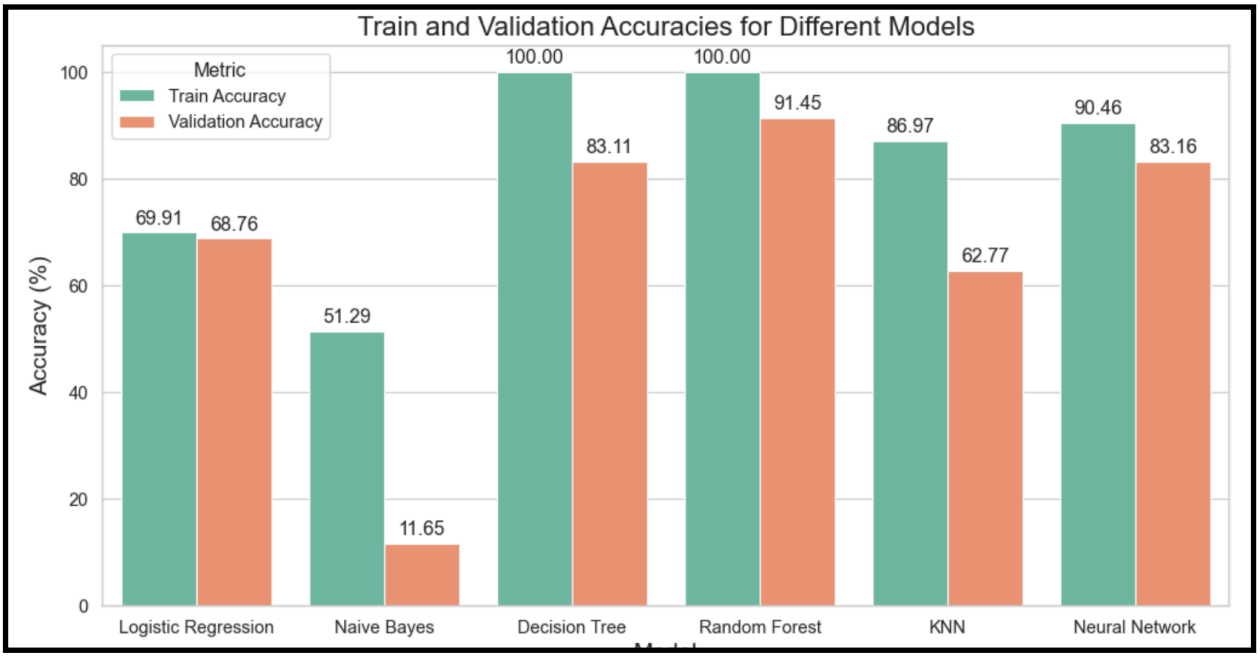
Neural Network Performance on Validation Set:

	precision	recall	f1-score	support
0.0	0.93	0.88	0.91	36312
1.0	0.16	0.24	0.19	3265
accuracy			0.83	39577
macro avg	0.54	0.56	0.55	39577
weighted avg	0.86	0.83	0.85	39577

- Training Accuracy: % 69.64
- Validation Accuracy: % 69.5
- ROC AUC: 0.695

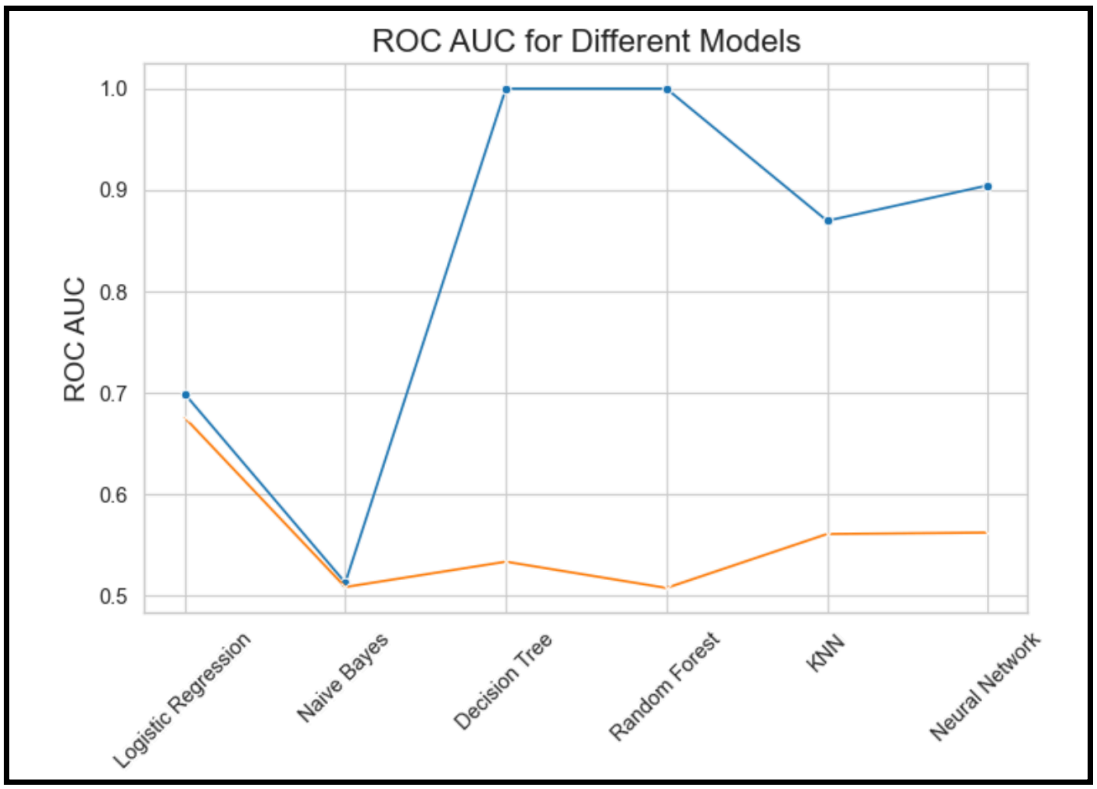
RESULTS

	Model	Train Accuracy	Validation Accuracy	Train ROC AUC	Validation ROC AUC
0	Logistic Regression	69.91	68.76	0.6991	0.6757
1	Naive Bayes	51.29	11.65	0.5129	0.5086
2	Decision Tree	100.00	83.11	1.0000	0.5336
3	Random Forest	100.00	91.45	1.0000	0.5077
4	KNN	86.97	62.77	0.8697	0.5609
5	Neural Network	90.46	83.16	0.9046	0.5623



- Of all the models, Naive Bayes performed the worst for both training and validation sets. This indicates that Naive Bayes is clearly not a good algorithm for this type of data.
- Decision Trees and Random Forests perform really well on training set which is common since these tend to overfit easily and we did not perform any pruning as a part of baseline models.
- Their performance on validation is good in terms of accuracy but average when looking at ROC AUC metric.

- KNN has good accuracy on training set but lower on validation set. However, it has a higher ROC AUC on validation set compared to decision tree and random forest.
- Neural network has higher accuracies on train and validation set but almost the same ROC AUC on validation set.
- Logistic Regression had moderately good accuracies on training and validation sets. It had the least difference between validation and training accuracy compared to all other models. This means that logistic regression is the closest to being a good fit on the data compared to all other models.
- This is solidified by the fact that logistic regression has the best ROC AUC score on the validation set outperforming all other models.



CONCLUSION

In conclusion, our analysis has uncovered key insights that inform strategic initiatives aimed at enhancing Home Credit's lending practices and improving the borrowing experience for our clients. Leveraging logistic regression as the most effective model with an accuracy of 67.57%, we are well-positioned to refine our predictive modeling efforts and prioritize proactive risk management strategies. Furthermore, our findings highlight the importance of targeting specific demographic segments, such as students, accountants, high-skill tech staff, and managers, to increase loan application interest and promote financial inclusion. Additionally, we recognize the need for further analysis, including surveys, to assess potential challenges for clients on maternity leave or unemployed individuals applying for cash loans contracts. By understanding these nuances, we can recommend the appropriate contract types to optimize approval rates for clients with diverse income sources. Moreover, our analysis underscores the significance of prioritizing clients aged 35-45, who exhibit minimal payment difficulties, as a key target demographic. By aligning our campaign strategies and lending practices with these insights, we can enhance our ability to meet the diverse needs of our clients while fostering inclusive economic development. Overall, leveraging robust data analytics enables us to refine our lending criteria, optimize risk assessment processes, and ultimately improve the overall borrowing experience for our clients, aligning with our mission of promoting financial inclusion and empowering individuals to achieve their financial goals.