PROJECT BASED INTERN

Home Credit Scorecard Model

For More About The Project Click Here

PRESENTED BY

Bagus Ariobimo

Table of Content

About Project

Problem Statement

In the aftermath of the COVID-19 pandemic, the economic landscape has undergone significant changes, impacting individuals financial stability. Many borrowers are still grappling with the aftermath of economic disruptions, resulting in challenges in repaying credit loans.

Objective

Adapt credit approval post-COVID. Develop a model that avoids rejecting eligible customers, considering pandemic effects. Optimize loan terms for motivating successful repayments.

Goals

Design loan packages that not only motivate successful repayment but also adapt to the financial challenges that customers may face during economic uncertainties.

Model Metrics

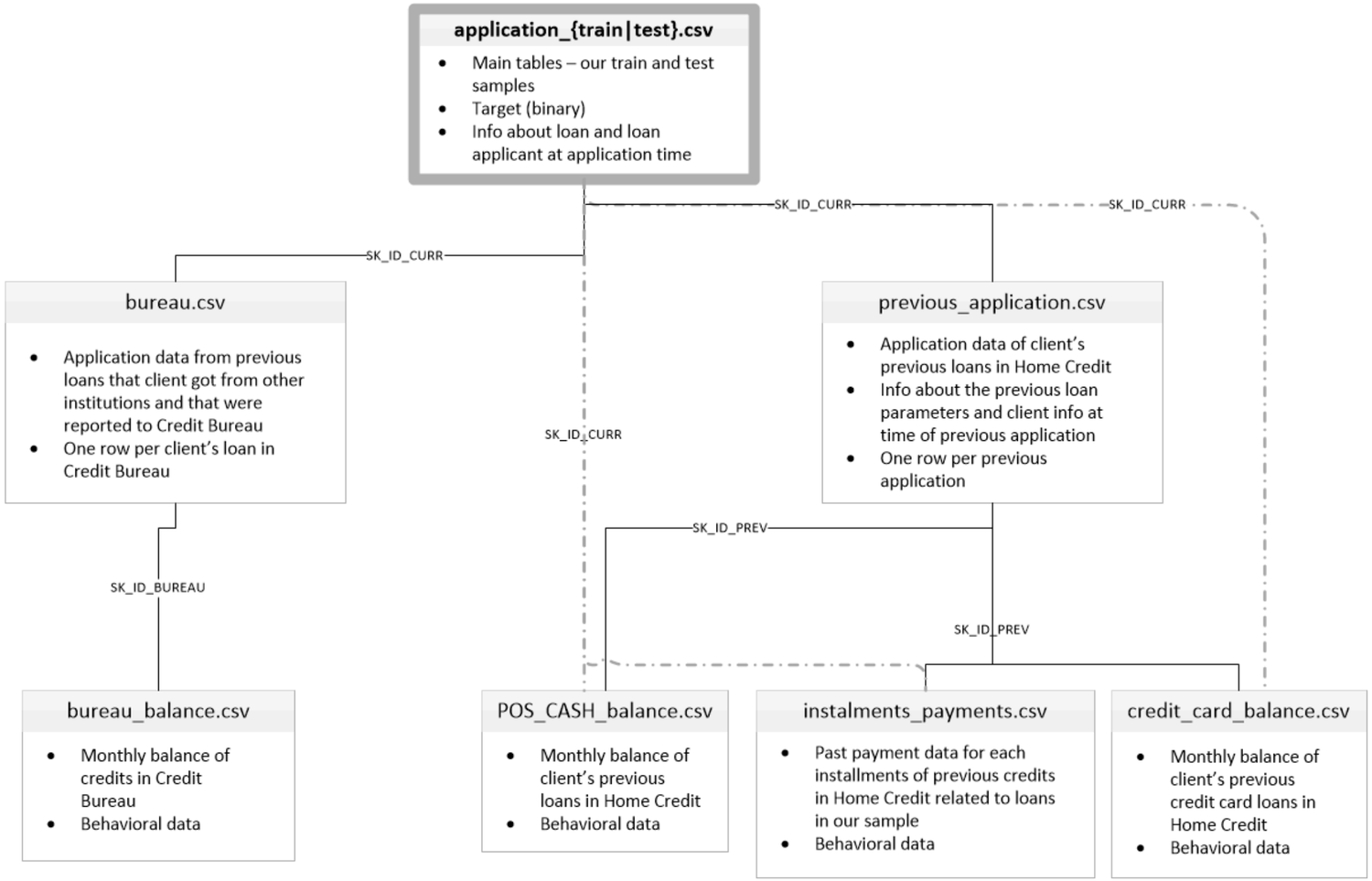
Precision: Critical for minimizing false positives and ensuring that approved applicants are genuinely creditworthy, especially in the context of economic uncertainties.

ROC AUC: Essential for evaluating the model's ability to navigate external variables, maintaining a high level of discrimination between good and bad credit risks.

Business Metrics

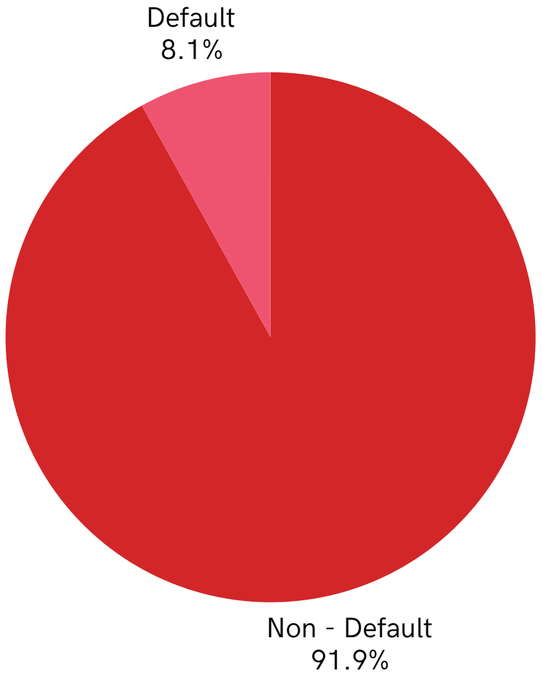
Default Rate: Assess the impact of external factors on the default rate to ensure that risk management strategies remain effective.

Dataset Understanding



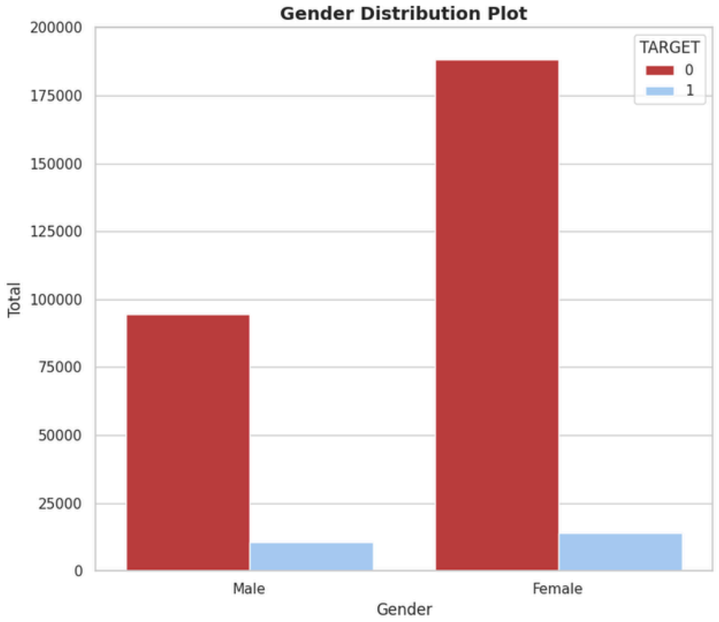
Exploratory Data Analysis

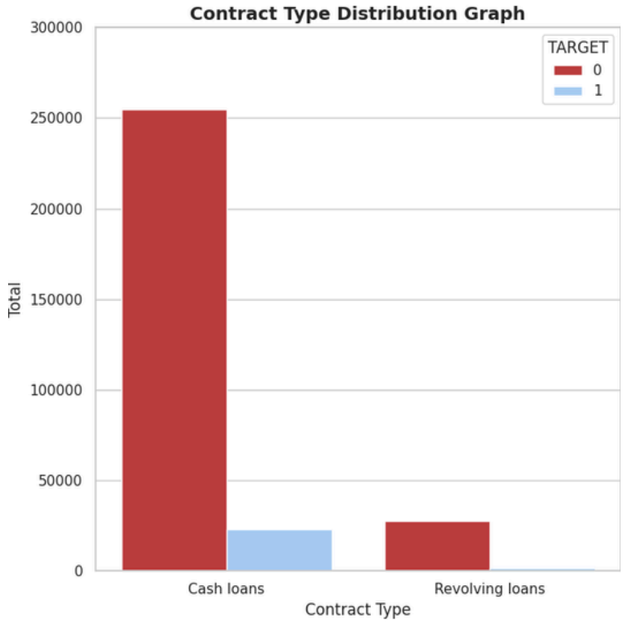
Target Percentage



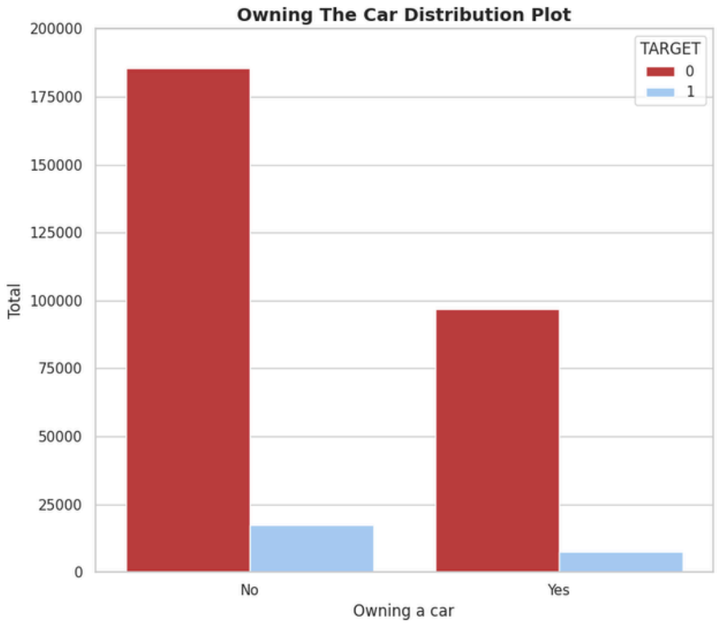
The significant disparity between the number of **Non-default instances (282,686)** and **Default instances (2,482)** highlights an extreme class imbalance, with a ratio of nearly 114:1. This imbalance risks causing the machine learning model to become biased toward predicting the majority class (Non-default) and struggle to learn patterns from the minority class (Default).

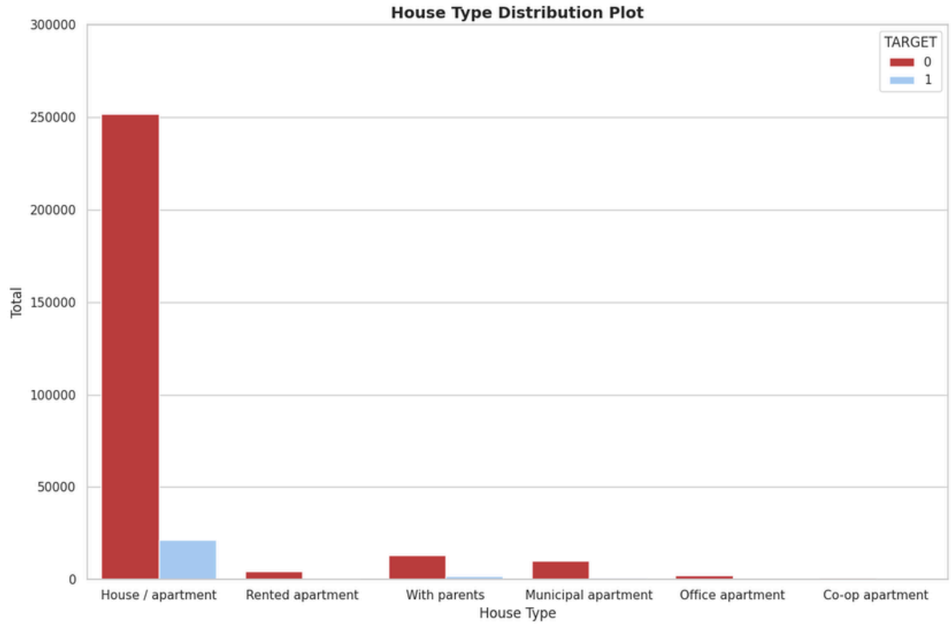
Exploratory Data Analysis Based on Personal Info





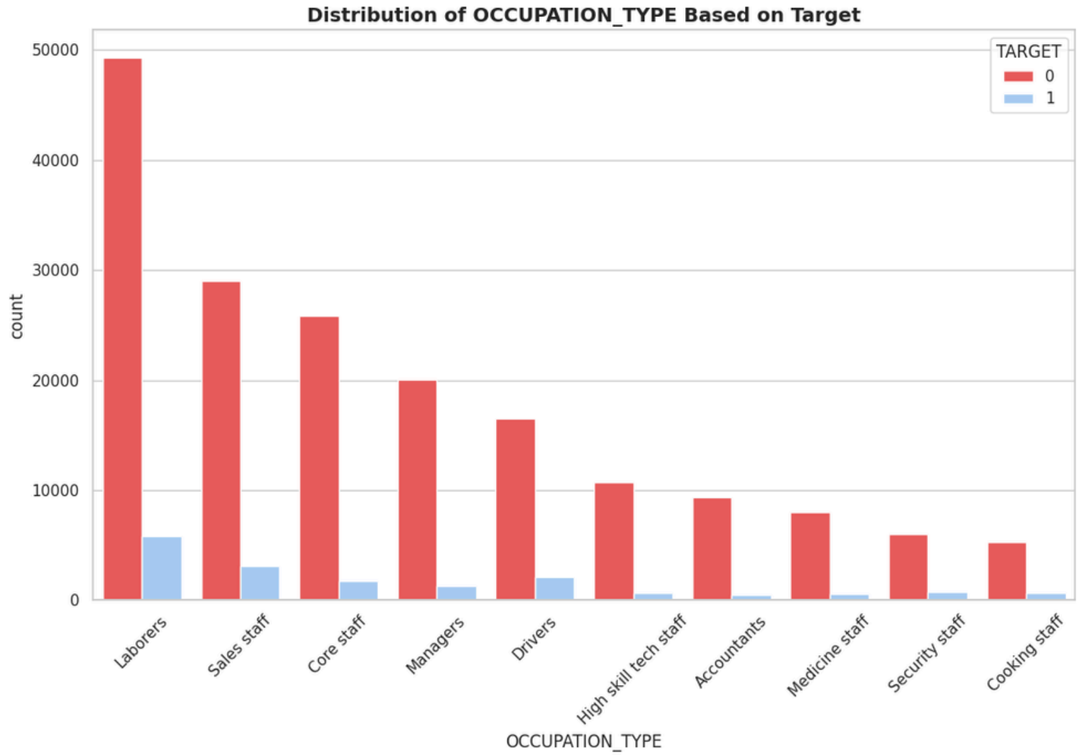
In the analysis of demographic and loan details, it's evident that a significant majority of credit users are female (202,452), surpassing the count of male users (105,059) in the 'CODE_GENDER' category. Turning to loan preferences in the 'NAME_CONTRACT_TYPE' column, the majority opt for cash loans (278,232), while a smaller proportion choose revolving loans (29,279).

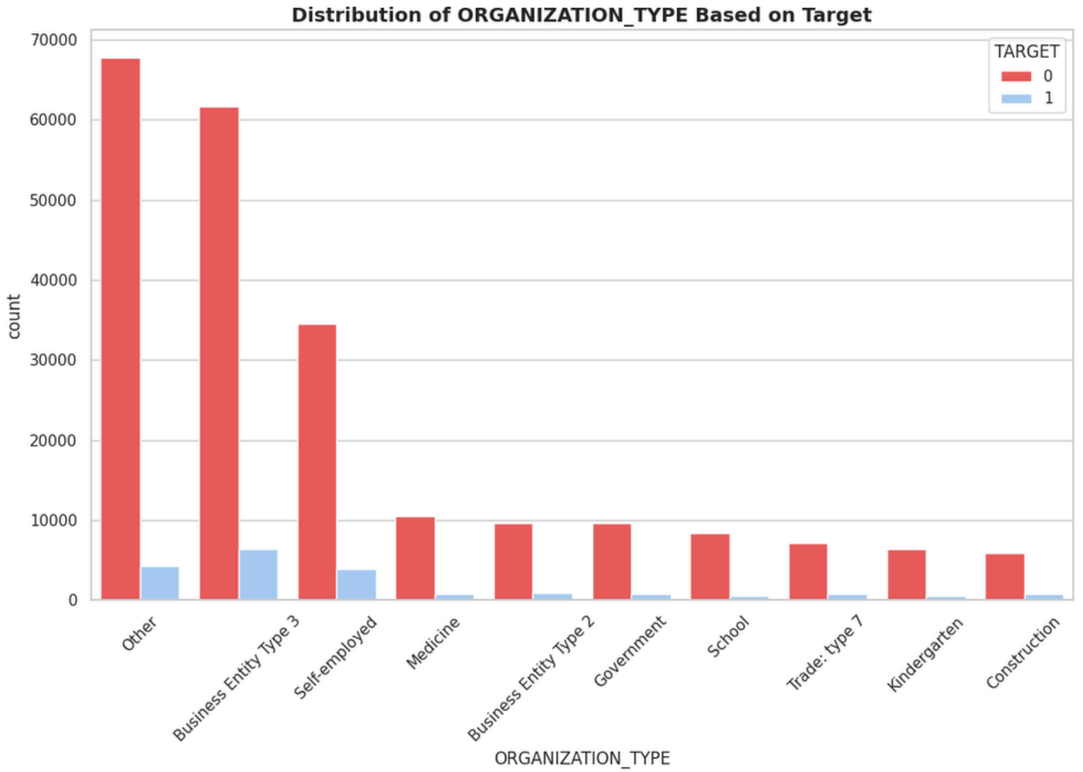




Shifting focus to housing and car ownership, a notable trend emerges. A substantial number of credit users reside in a 'House / apartment' (272,868), indicating a prevalent housing choice. Simultaneously, a considerable portion of users does not own a car (202,924). These insights offer a concise breakdown of dominant patterns in gender and loan preferences, as well as housing type and car ownership among credit users.

Exploratory Data Analysis Based on Career Info





Note: that this plot are after the data cleaning itself

The dominant occupation type among applicants is "Laborers" (55,186), followed by "Sales staff" (32,102) and "Core staff" (27,570). In terms of organization types, "Other" (72,057) holds the majority, with "Business Entity Type 3" (67,992) and "Self-employed" (38,412) also significant. These insights provide a snapshot of the workforce composition, guiding potential strategies for credit assessments and tailored loan offerings based on prevalent occupations and organizational affiliations.

Data Cleaning

Handling Missing Values

- **Removing columns that have 60% missing values**
- **Impute the columns that less than 60% with median**

Correcting Invalid Values

- **Replace the XNA and missing values using Mode**
- **the other like organization_type move the missing values to `other` Values.**

Feature Engineering

- **CREDIT_INCOME_PERCENT:** the percentage of the credit amount relative to a client's income
- **ANNUITY_INCOME_PERCENT:** the percentage of the loan annuity relative to a client's income
- **CREDIT_TERM:** the length of the payment in months (since the annuity is the monthly amount due
- **DAYS_EMPLOYED_PERCENT:** the percentage of the days employed relative to the client's age

Encoding Columns

- **Label Encoding:**

for any categorical variable (object data types) with only 2 unique categories using scikit-learn.

- **One Hot Encoding**

for any categorical variable with over than 2 unique categories

Correlation Insight

Correlations

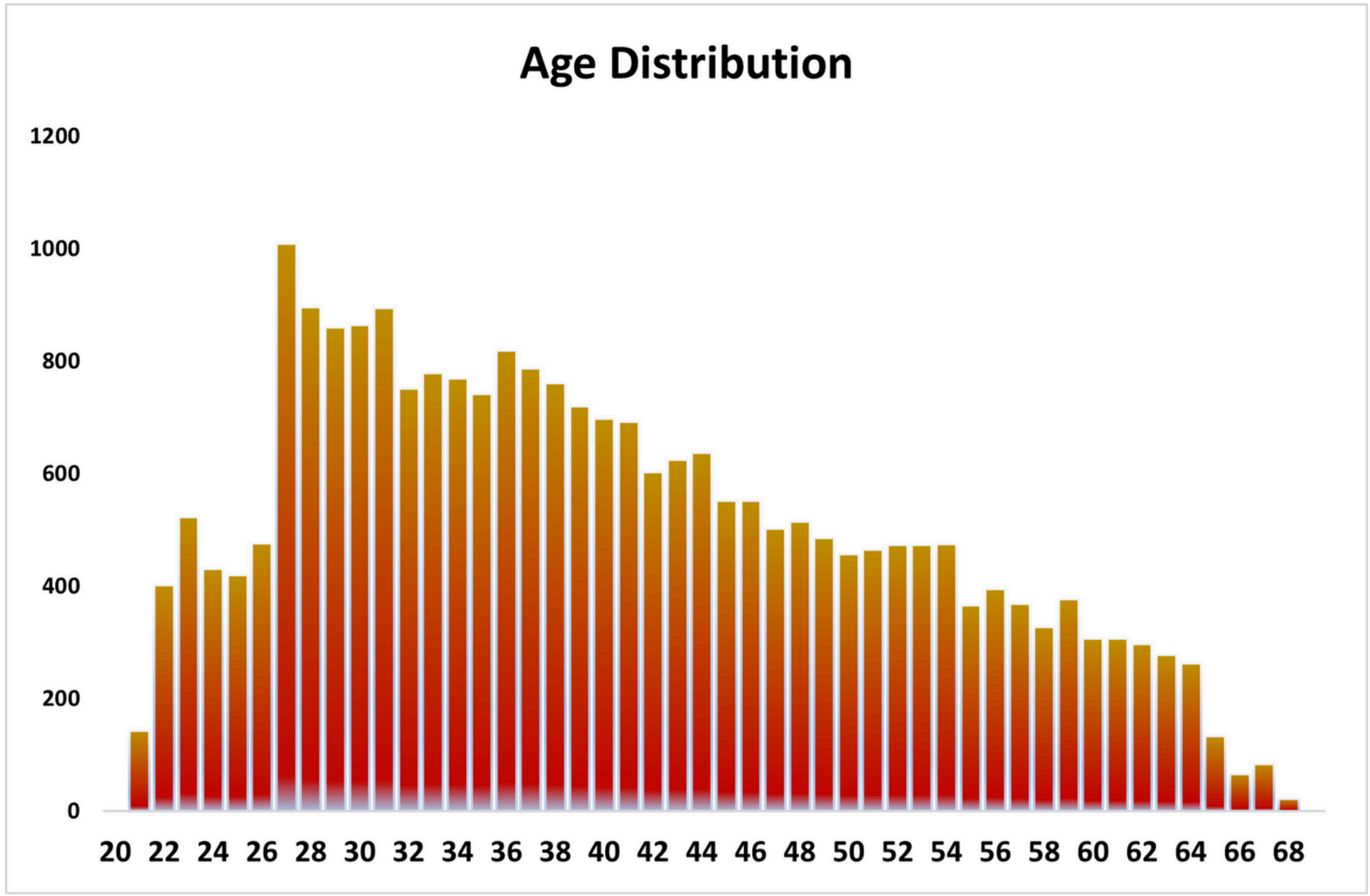
- .00-.19 “very weak”
- .20-.39 “weak”
- .40-.59 “moderate”
- .60-.79 “strong”
- .80-1.0 “very strong”

```
Most Positive Correlations:
 OCCUPATION_TYPE_Laborers                                0.043019
FLAG_DOCUMENT_3                                          0.044346
REG_CITY_NOT_LIVE_CITY                                   0.044395
FLAG_EMP_PHONE                                           0.045982
NAME_EDUCATION_TYPE_Secondary / secondary special        0.049824
REG_CITY_NOT_WORK_CITY                                   0.050994
DAYS_ID_PUBLISH                                          0.051457
CODE_GENDER_M                                            0.054713
DAYS_LAST_PHONE_CHANGE                                   0.055218
NAME_INCOME_TYPE_Working                                 0.057481
REGION_RATING_CLIENT                                     0.058899
REGION_RATING_CLIENT_W_CITY                              0.060893
DAYS_EMPLOYED                                            0.074958
DAYS_BIRTH                                               0.078239
TARGET                                                   1.000000
Name: TARGET, dtype: float64
```

Note: that this plot are after the data cleaning itself

One significant insight from the correlation analysis is the relationship between age (represented by DAYS_BIRTH). Interestingly, DAYS_BIRTH measures the client's age in negative days, which can be a bit counterintuitive. Despite the positive correlation, the negative values indicate that as clients age, they are less likely to default on their loans (where TARGET == 0). To make this relationship clearer, taking the absolute value of DAYS_BIRTH would reverse the correlation to negative, directly reflecting the trend that older clients have a lower likelihood of defaulting.

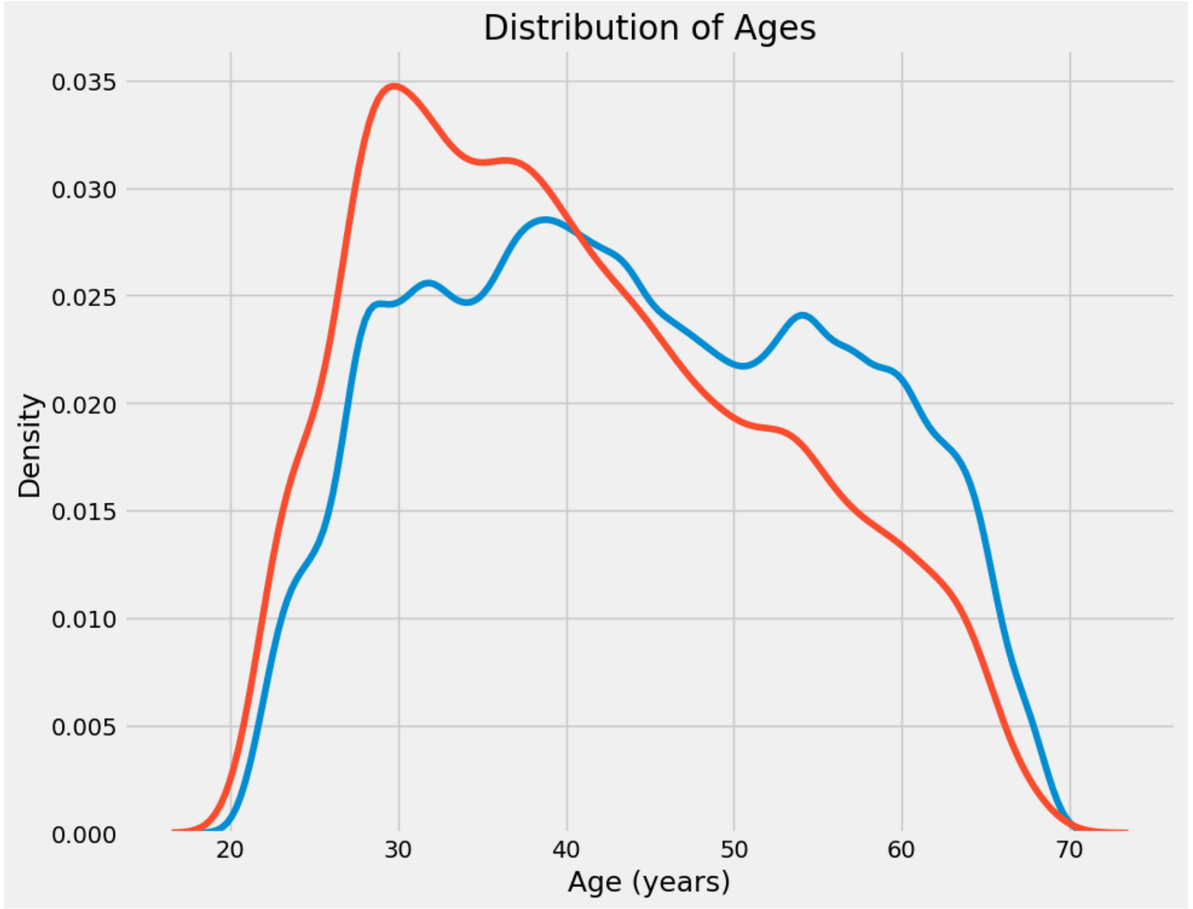
Age Clients Distribution



Note: that this plot are after the data cleaning itself

The histogram plot of ages reveals a predominant age group among users, concentrated between 27 and 46 years. This age range suggests that the majority of applicants fall within the working-age bracket, indicating a workforce-oriented demographic. Understanding this age distribution is crucial for tailoring credit offerings and repayment structures that align with the financial dynamics and life stages of individuals in this predominant age group.

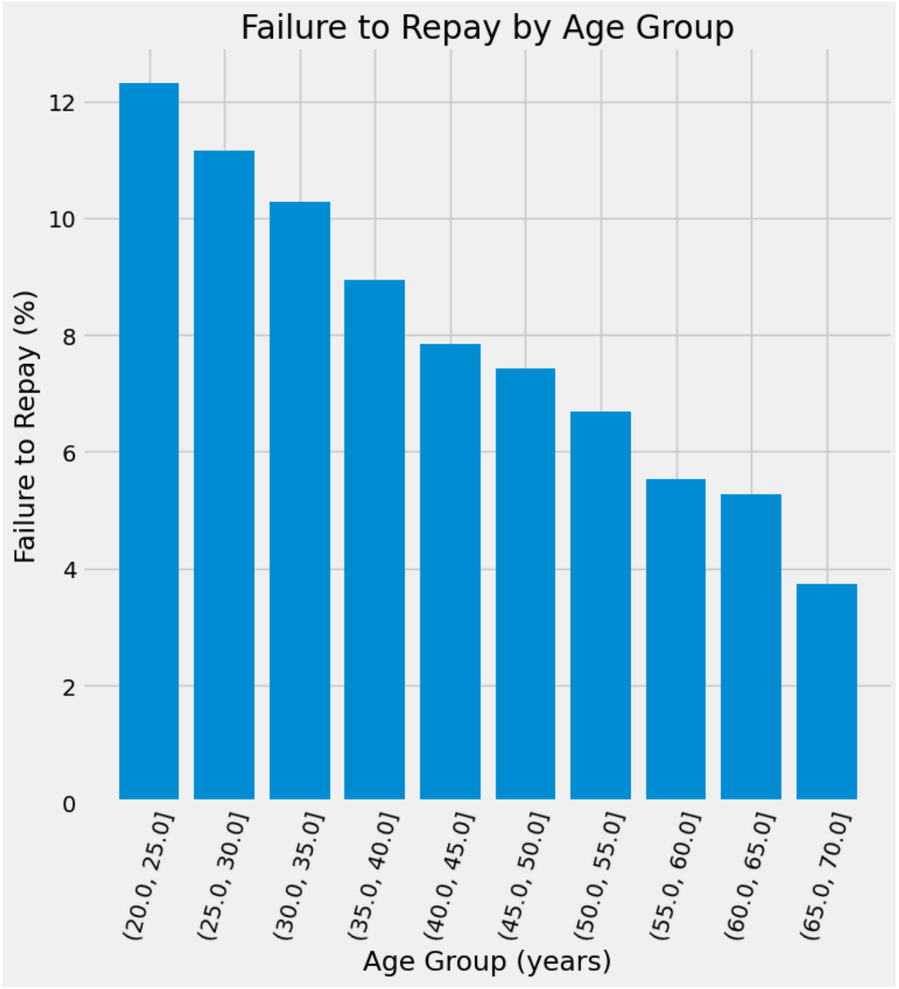
Age Clients Distribution



Note: that this plot are after the data cleaning itself

The curve for TARGET == 1 skews towards the younger end of the age range. Although this is not a significant correlation (correlation coefficient of -0.07), this variable is likely to be useful in a machine learning model as it does have an impact on the target.

Age Clients Distribution



Note: that this plot are after the data cleaning itself

There is a clear trend: younger applicants are more likely to default on their loans! The default rate exceeds 10% for the three youngest age groups, while it drops below 5% for the oldest age group. This is actionable information for banks: since younger clients are more prone to defaulting, they could benefit from additional guidance or financial planning tips. This doesn't mean banks should discriminate against younger clients, but it would be wise to take preventive measures to help them make timely payments.

Modelling

The first step in the workflow involves preparing the data for model training. Missing values are handled using median imputation, ensuring that any missing data is filled with the median of the respective feature. This step ensures the model can learn from all available data without any gaps. Additionally, the features are scaled to a range between 0 and 1 using MinMax scaling, which normalizes the data and ensures that no feature dominates the model due to its scale.

The dataset is split into training and testing sets using train_test_split, with 25% of the data reserved for testing. This allows for proper evaluation of the model's performance on unseen data.

Logistic Regression

```
[ ]  1 from sklearn.metrics import roc_auc_score, precision, recall
     2
     3 # Make predictions
     4 # Make sure to select the second column only
     5 log_reg_pred = log_reg.predict_proba(test)[:, 1]
     6
     7 print(roc_auc_score(y_test, log_reg_pred))
```

```
0.6840900298834589
```

&

```
[ ]  1 from sklearn.ensemble import RandomForestClassifier
     2
     3 # Make the random forest classifier
     4 random_forest = RandomForestClassifier(n_estimators = 100, random_state = 50, verbose = 1, n_jobs = -1)
```

```
     1 # Train on the training data
     2 random_forest.fit(train, y_train)
     3
     4 # Extract feature importances
     5 feature_importance_values = random_forest.feature_importances_
     6 feature_importances = pd.DataFrame({'feature': features, 'importance': feature_importance_values})
     7
     8 # Make predictions on the test data
     9 predictions = random_forest.predict_proba(test)[:, 1]
```

```
Show hidden output
```

```
[ ]  1 print(roc_auc_score(y_test, predictions))
```

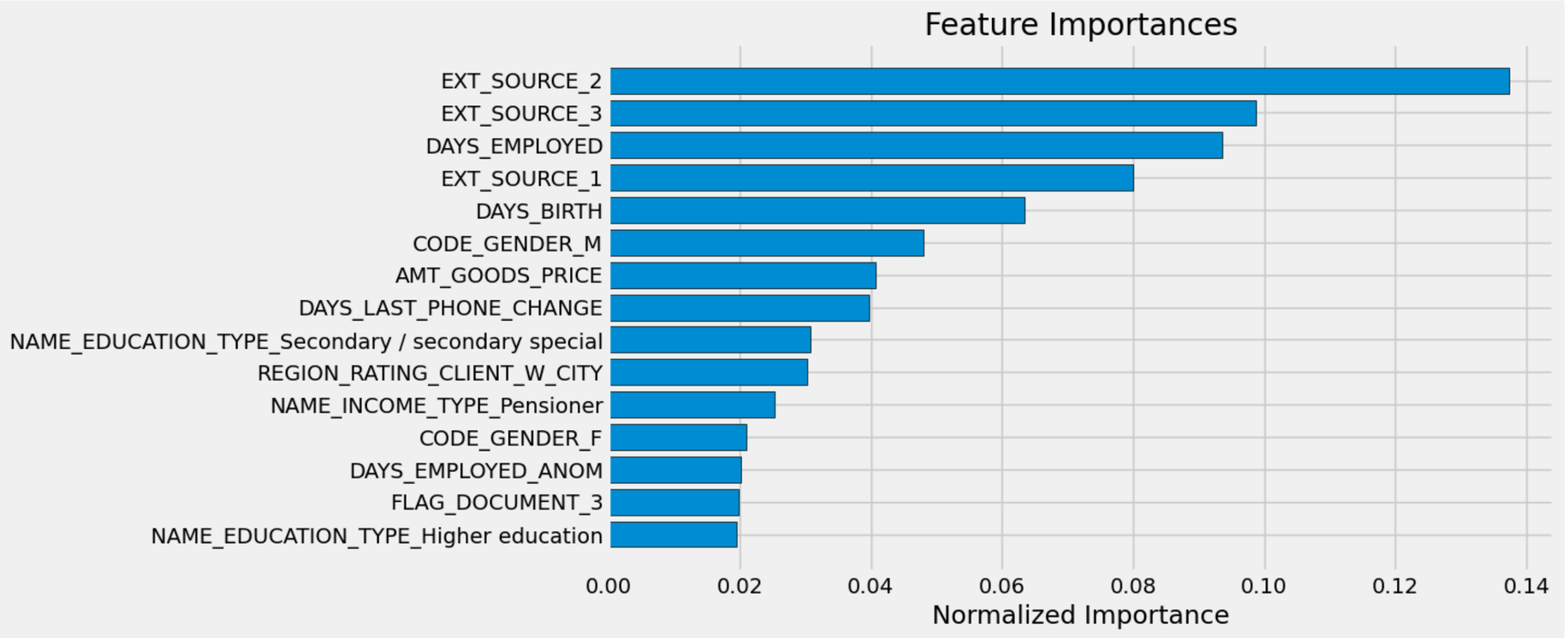
```
0.7119314418987335
```

Modelling

Apologies for any inconvenience caused. Unfortunately, due to limitations on my computing device, I wasn't able use another model. I appreciate your understanding and am happy to explore alternative approaches or provide additional information as needed.

Metric	Random Forest	Logistic Regression
ROC AUC (Test Set)	0.705	0.602
ROC AUC (Train Set)	0.703	0.614
Model Type	Ensemble (Random Forest)	Linear (Logistic Regression)
Regularization	Max Depth = 3, Class Weight = Balanced	None (Default)
Preprocessing	Median Imputation, MinMax Scaling	Median Imputation, MinMax Scaling
Performance	Better (Higher ROC AUC)	Lower (Lower ROC AUC)

Feature Selection



The feature importance analysis highlights external credit scores (EXT_SOURCE_2, EXT_SOURCE_3) and employment history (DAYS_EMPLOYED) as the most influential predictors, with normalized importance values significantly higher than other features. Client age (DAYS_BIRTH), gender (CODE_GENDER_M), and loan-related attributes (AMT_GOODS_PRICE) also play notable roles, though less pronounced. Education level, income type (e.g., pensioners), and recent customer activity (DAYS_LAST_PHONE_CHANGE) contribute moderately, while document submissions (FLAG_DOCUMENT_3) show minimal impact. The results underscore the dominance of external creditworthiness metrics and employment stability in driving predictions, with demographic and behavioral factors providing secondary insights.

Business Strategy

1. **Prioritize Creditworthiness Metrics:** External credit scores (EXT_SOURCE_2, EXT_SOURCE_3) are the strongest predictors. Strengthen partnerships with credit bureaus or invest in tools to refine credit risk models, ensuring alignment with these external metrics.
2. **Focus on Employment Stability:** DAYS_EMPLOYED (employment history) is critical. Consider targeting clients with stable employment records or introducing incentives (e.g., lower interest rates) for applicants with longer job tenure to reduce default risks.
3. **Demographic Targeting:**
    - **Age & Gender: Y**ounger clients (DAYS_BIRTH) and male applicants (CODE_GENDER_M) show higher influence. Tailor marketing campaigns to these demographics while ensuring compliance with fairness regulations.
    - **Pensioners:** The negative impact of pensioner income types suggests cautious risk assessment for retirees. Offer specialized products (e.g., smaller loans) for this group.
4. **Loan Product Optimization:** AMT_GOODS_PRICE (loan purpose) is significant. Design loan products tied to specific goods (e.g., appliances, vehicles) to align with customer needs and improve approval confidence.
5. **Behavioral Insights**: DAYS_LAST_PHONE_CHANGE (recent activity) indicates responsiveness. Use this to identify engaged customers for cross-selling or retention strategies.