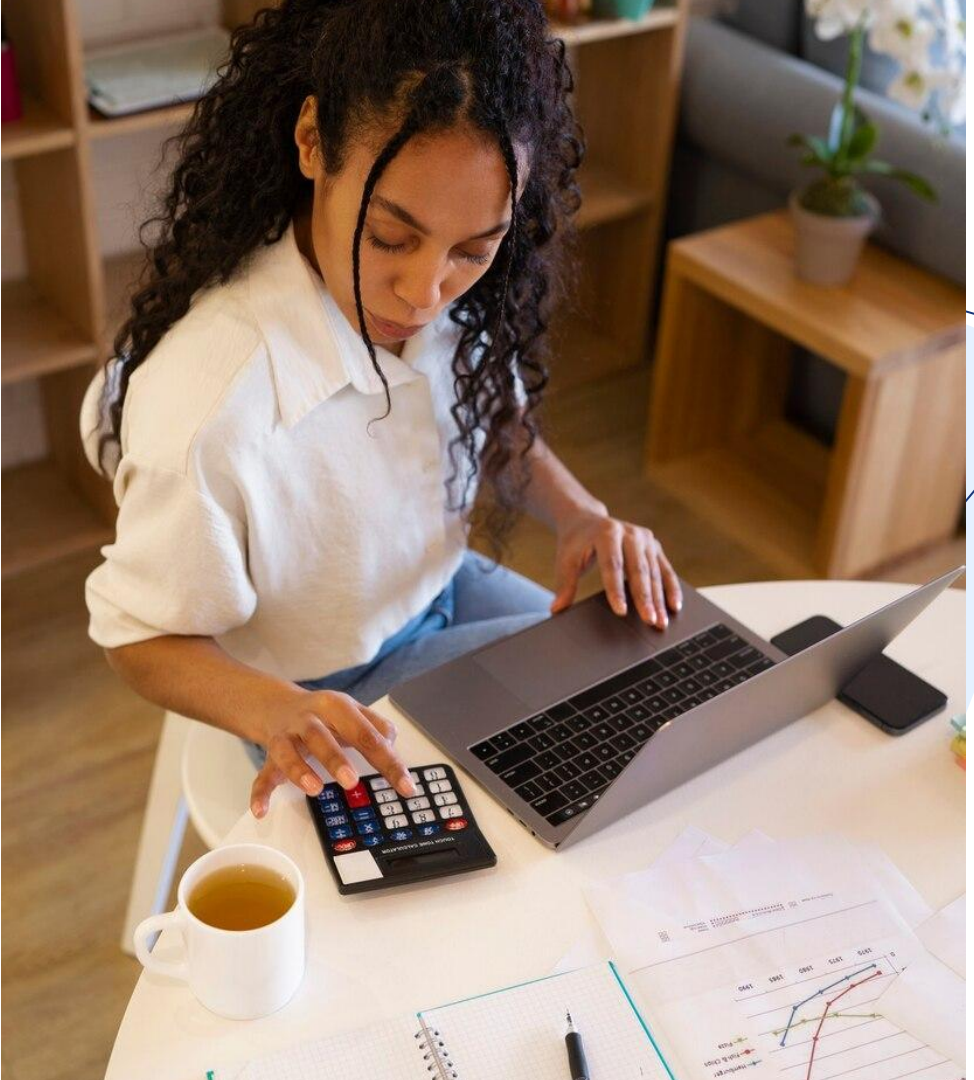
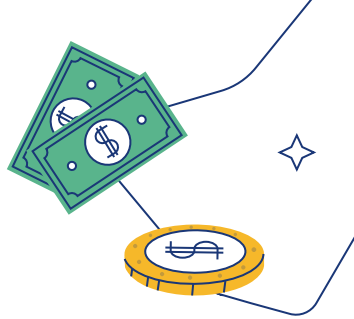
Loans Analysis - Use Case

By: Hala Sedki - Use Case Interview #2

Table of Contents





01

Business Aspects & Objectives

Business Aspects & Objectives

Aspects

When assessing a loan application, the company faces two main risks:

1. **Missed Opportunities**
2. **Financial Loss**

The dataset includes information about loan applications, categorized into two scenarios:

1. **Clients with Payment Difficulties**
2. **All Other Cases**

Decisions made during the loan process fall into four categories:

1. **Approved**
2. **Cancelled**
3. **Refused**.
4. **Unused Offer**

Objectives

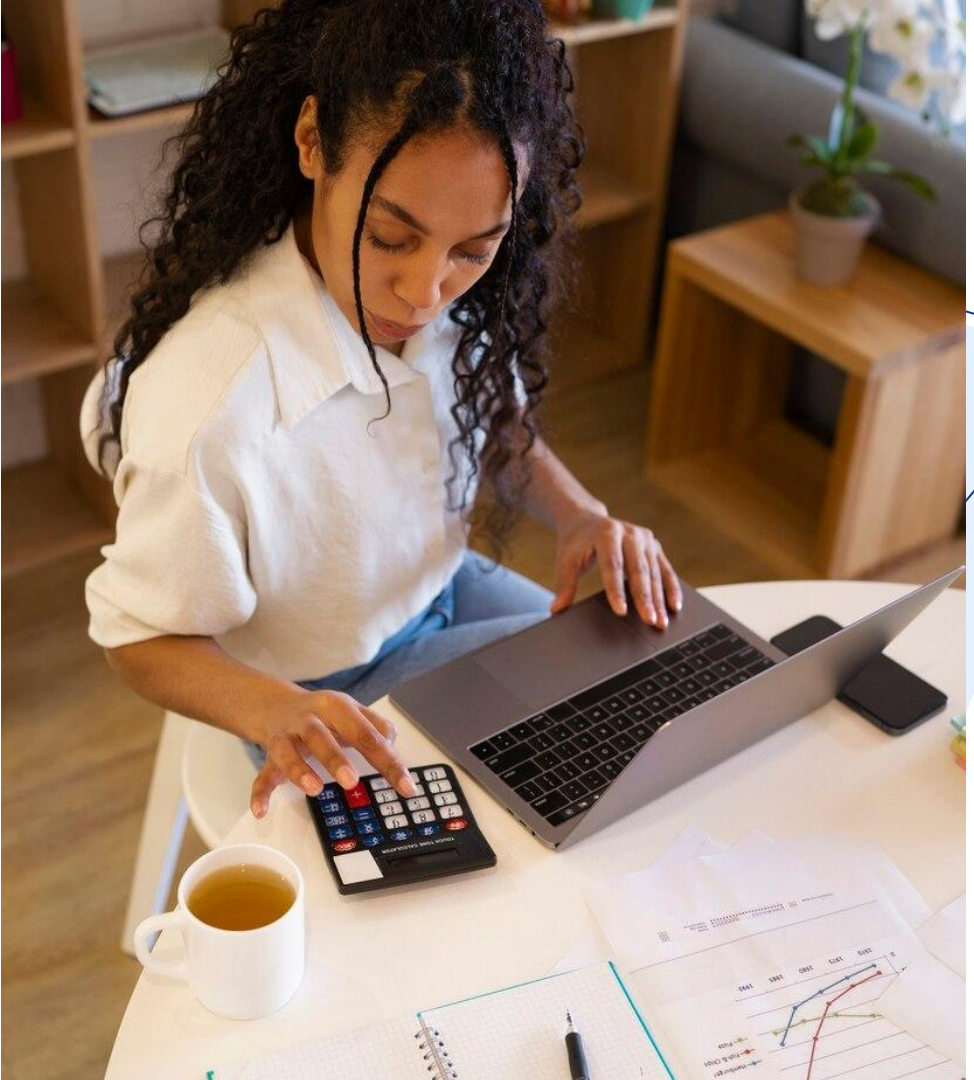
Identify patterns indicating potential difficulties in loan repayment.

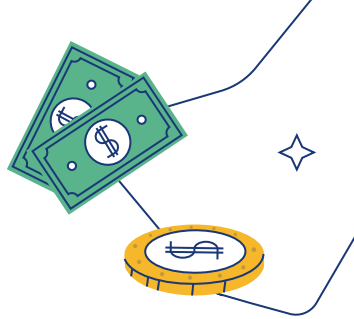
Actions Based on Findings:

- Reject high-risk loans.
- Adjust loan amounts.
- Charge higher interest rates for riskier applicants.

Goal:

- Pinpoint key factors predicting loan default.
- Improve risk assessment and management in the loan portfolio.





02

EDA Process & Insights

EDA -> Inspection & Cleaning

1 Missing Data Imputation

- Columns with more than 50% missing data can be dropped
- Columns with 13% or lower missing values, can be imputed with mode, median or mean depending on the type of the column and data distribution
- Categorical columns with > 20% but < 50% missing data can have a new type such as 'Unknown' for missing data imputation.

2 Data Type Correction

- The columns representing the number of enquires to Credit Bureau about the client are of float data type. We can change the data type to int.
- **DAYS** related columns and **CNT_FAM_MEMBERS** can be changed to int data type as they represent number of days and family member count respectively.

3 Data Standardization

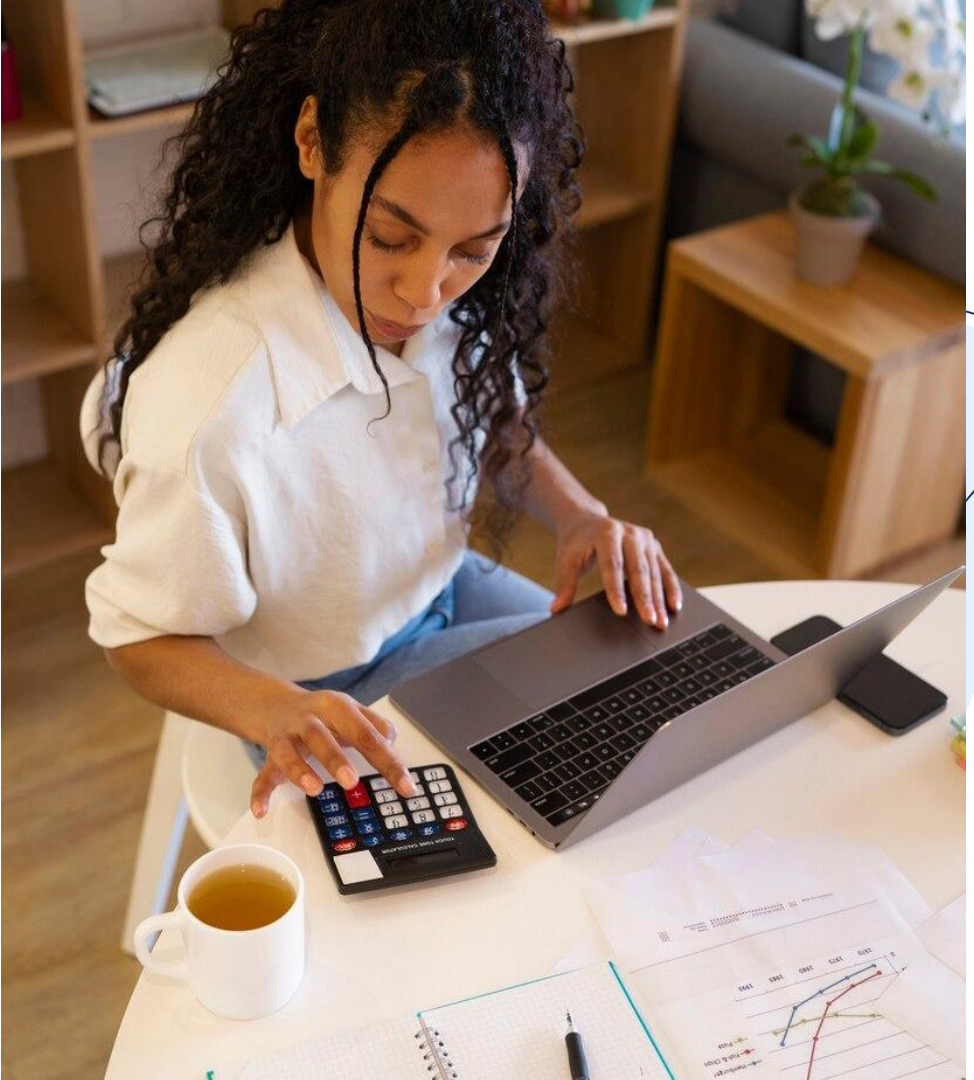
- **DAYS** related columns have some negative values. They can be replaced with their respective absolute values.
- We can create a new column based on **DAYS_BIRTH** to show the age of the applicant for better readability and then we can drop the **DAYS_BIRTH** column. Similarly we can convert the other **DAYS** columns to represent the value in years.
- **The CODE_GENDER** column has XNA values that can be replaced with nan.

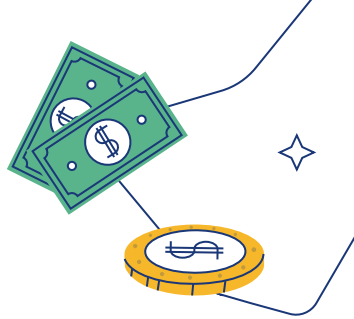
4 Outlier Analysis

- **AMT_INCOME_TOTAL** has very high valued outliers. As we know the income may vary from person to person, we can cap value here and get rid of very high incomes.
- **DAYS_EMPLOYED** has huge outliers. Some data points are showing close to 1000 years in service which is impossible. We can cap the value at a desired point after analysing the quantiles.

6 Binning

- The 'TARGET' column represents whether the client is a defaulter or not. If we segregate our dataset based on this column, and if the distribution turns out to be 50-50, then our data set would be BALANCED. In any other case, it would be considered as IMBALANCED.
- We then created 2 data sets to segregate our original data based on the TARGET column values to have defaulters in one dataframe and others in another.





03

Inferences from EDA

Most Likely to Default

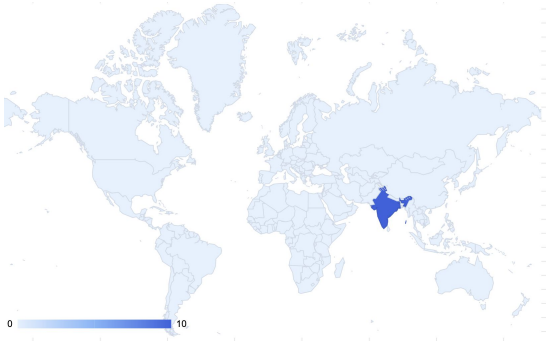
Demographic

Age	25-30 years old
Gender	Males
Occupation	Low-skill labourers, drivers
Family situation	Civil marriage, single/unmarried
Income	$5 lakhs/year
Education	Lower/Secondary

Loan Type

Income Type	On maternity leave and unemployed
Housing Type	Rented apartment or with parents
Contract Type	Cash Loan
Cash Loans Purpose	Repairs and urgent needs
Previous Loan Status	Approved

Geographic



Region	India

Follow the link in the map to modify its data and then paste the new one here. **For more info, click here**

Most Likely to Repay on Time

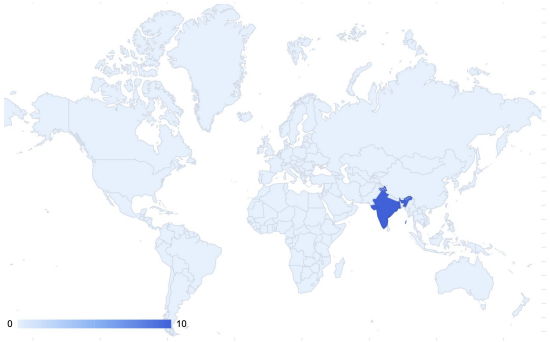
Demographic

Age	>50 years old
Gender	Females
Occupation	Managers, Students, etc.
Family situation	Married
Income	Higher Income Group
Education	Higher Education

Loan Type

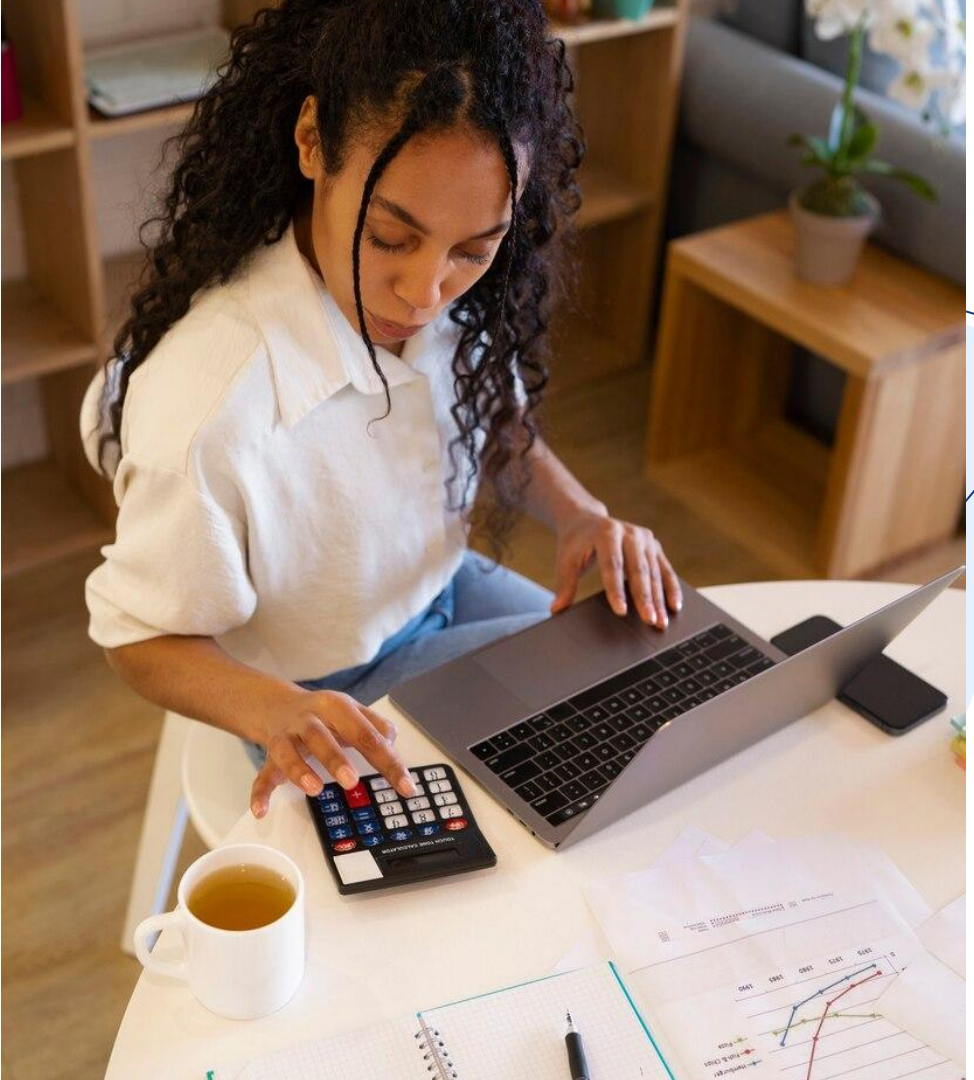
Income Type	Working Class, Students, Businessmen
Housing Type	Own Apartment/House
Contract Type	Revolving Loan
Cash Loans Purpose	Buying Garage, Homes
Previous Loan Status	Unused Offer

Geographic



Region	India

Follow the link in the map to modify its data and then paste the new one here. **For more info, click here**



04

Model

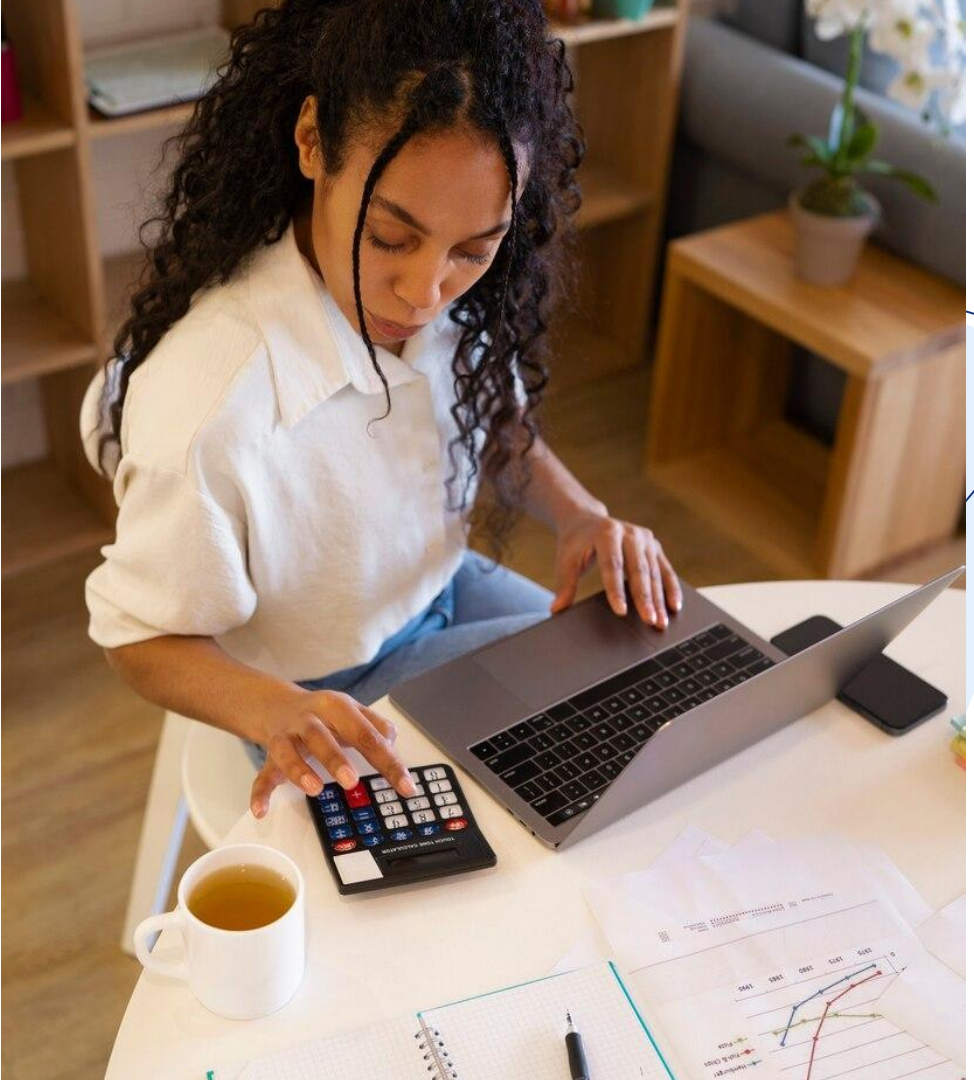
Model:

Random Forest was used for the Modeling ->

- Why I Chose RF? Random Forest is used for feature selection and variance reduction. It reduces variance by averaging the predictions of multiple decision trees, resulting in a lower variance for the overall model. Specifically, the variance of the Random Forest model decreases as the number of trees increases, following the formula: new_variance = old_variance / number_of_trees.

Evaluation Metric -> ROC Curve

- The ROC Curve is chosen as an evaluation metric because it effectively illustrates the trade-off between true positive and false positive rates, making it suitable for imbalanced datasets. However, the Precision-Recall Curve is also a strong option, as it provides a clearer view of model performance concerning the positive class.



04

Conclusion/ Recommendations

Conclusions from Data

Decisive Factor whether an applicant will be a Defaulter:

- **CODE_GENDER**: Men have a relatively higher defaulter rate than women
- **NAME_EDUCATION_TYPE**: Applicants with Lower Secondary & Secondary education has higher risk of rejecting repay loans
- **YEARS_BIRTH**: Young adults (age under 30) have the higher defaulter rate
- **YEARS_EMPLOYED**: People who have less than 5 years of employment have a relatively high defaulting rate.
- **OCCUPATION_TYPE**: Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff are the high-risk occupation since the defaulting rate is huge.
- **ORGANIZATION_TYPE**: The top 5 high-risk organization types are: Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%).
- **CNT_CHILDREN & CNT_FAM_MEMBERS**: Clients who have 4 more children (6 more family members) will have a higher defaulting rate than the other groups.
- **NAME_TYPE_SUITE**: Other_B type have the highest defaulting rate
- **NAME_FAMILY_STATUS**: People who have civil marriages or who are single default a lot.
- **NAME_HOUSING_TYPE**: Rented apartments or living with parents will potentially increase the risk of rejecting repaying loans.
- **OWN_CAR_AGE**: Applicants own cars with 10+ years has relative higher defaulting rate.
- **NAME_INCOME_TYPE**: Clients who are either on Maternity leave OR Unemployed default a lot.
- **REGION_RATING_CLIENT:** People who live in the area with Rating 3 have the highest defaults.
- **NAME_CONTRACT_TYPE**: Application with cash loans has more risk of losing loans than revolving loans.
- **AMT_GOODS_PRICE**: When the credit amount goes beyond 3M, there is an increase in defaulters.

In terms of the loan defaulter data, there are no clear boundaries to classify the loan defaulters, so I summarize the top 10 characteristics showing up most frequently in loan defaulters.

- NAME_CONTRACT_TYPE_Cash loans,
- NAME_HOUSING_TYPE_House / apartment,
- NAME_TYPE_SUITE_Unaccopanied,
- NAME_EDUCATION_TYPE_Secondary/secondary special,
- REGION_RATING_CLIENT_W_CITY_2,
- OWN_CAR_AGE_Unkown,
- FLAG_OWN_CAR_N,
- FLAG_OWN_REALITY_Y,
- CNT_CHILDREN_No_Child,
- NAME_INCOME_TYPE_Working.

To predict whether a client is a loan defaulter or not, the random forest with the best parameter setting can give a reasonably good reference according to the auc_roc_score.

Recommendations for the Bank

Young males with lower secondary education and of lower income group and staying with parents or in a rented house, applying for low-range cash contract, should be denied.

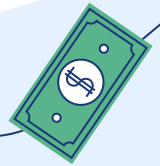
Females are likely to repay but not if they are on maternity leave. Hence, bank can reduce the loan amount for female applicants who are on maternity leave.

Since people taking cash loans for repairs and urgent needs are more likely to default, bank can refuse them.

Since the people who have unused offers are more likely to default even though they have comparatively high total income, they can be offered loan at a higher interest rate.

Banks can target businessmen, students and working class people with academic degree/ higher education as they have no difficulty in repayment.

Bank can also approve loans taken on purpose for buying home or garage as there less chances of defaulting.



Thanks!

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**