### **Capstone Project 1: Home Loan Credibility Assessment**

### **Milestone Report**

1. **Problem Statement**

Mainstream banks and financial institutions check traditional credit score models, which include demographic characteristics, historical payment data, credit bureau data and application data, to determine repayment success. However, many unbanked individuals do not have sufficient credit scores due to their past mistakes of unavoidable circumstances. Therefore, they have to deal with unconventional means such as loan sharks when borrowing money. Moreover, most of these individuals are hard-working and should get a chance to borrow money safely. It is important to identify these individuals from the financial records to provide a positive and safe borrowing experience.

The primary object of this project is to build a model from the financial data to predict the likelihood that an applicant will experience diﬃculty in repaying their loans. The output of the proposed model is the probability which determines an applicant in terms of having at least one late payment when repaying their loan.

1. **Datasets**

The dataset is provided by Home Credit Group’s data scientists which contains personal and ﬁnancial information belonging to 356,255 individuals who had previously been recipients of loans. The data is divided into training group, which contains 307,511 records, and test group, which contains 48,744 records.

The data is acquired from seven different sources. The first dataset, application\_train/ application\_test, is the main training and testing data with information about each loan application. Each row is identified by the feature SK\_ID\_CURR. The TARGET feature in the training data represents load repaid by 0 and not repaid by 1. The second data source bureau provides the client's previous credits from other financial institutions. The third one is bureau\_balance, which provides monthly balances of previous credits in Credit Bureau. POS\_CASH\_balance provides monthly balance snapshots of the previous point of sales and cash loans that the applicant had with Home Credit. The fifth data source credit\_card\_balance presents the monthly balance snapshots of previous credit cards that the applicant has with Home Credit. All previous applications for Home Credit loans of clients who have loans are mentioned in the previous\_application. Repayment history for the previous loans is provided in the seventh data source installments\_payment.

Data Source:<https://www.kaggle.com/c/home-credit-default-risk/data>

1. **Data Cleaning/Wrangling**

I have obtained the data from Keggle and the data is in good condition. I have performed the following cleaning and wrangling on the data.

**1. Missing value treatment:** Missing data can lead to a wrong prediction or classification. There are 122 columns in the training data and 67 of them contain missing values. Most of the missing value columns contain more than 50% of missing values. Deleting the missing values is not a good option for this case.

**2. Outlier Detection and Treatment:** There are two variables that contain anomalies in our datasets. One is ‘DAYS\_EMPLOYED’ and the other one is ‘CODE\_GENDER’. The anomalies have a lower rate of default. The outliers in the ‘DAYS\_EMPLOYED’ variable are replaced with NaN value and four applications with ‘XNA’ in ‘CODE\_GENDER’ are removed.

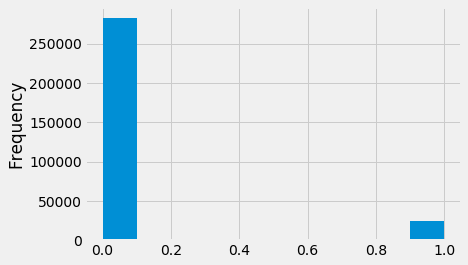
## **3. Encoding Categorical Variables:** The categorical variables need to be encoded as numbers before handing them off to the model. There are two methods to represent categorical variables, they are label encoding and one-hot encoding. The label encoding method assigns each unique category in a categorical variable with an integer without creating any new columns. However, it gives the categories an arbitrary ordering. Whereas the one-hot encoding creates a new column for each unique category in a categorical variable. Therefore, in this project, label encoding method is applied on three variables, 'CODE\_GENDER', 'FLAG\_OWN\_CAR', and 'FLAG\_OWN\_REALTY', which have only two unique values and modified the values as a 0 or a 1. The rest of the categorical variables are encoded with one-hot encoding method.

1. **Exploratory Data Analysis (EDA)**

In this section, the statistical and visualization methods are used to find trends, anomalies, patterns, or relationships within the data.

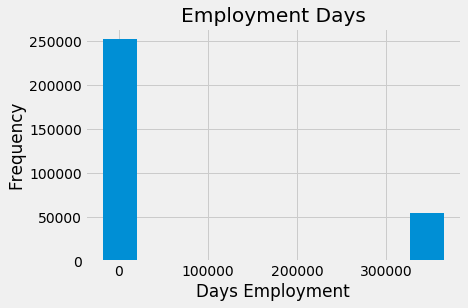
### **Examine the Distribution of the Target Column**

The target column contains two values: 0 for the loan was repaid on time and 1 indicating the client had payment difficulties. We will use this column for prediction.

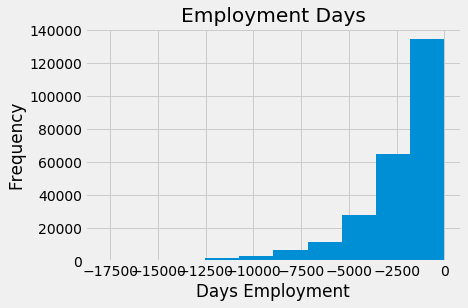


The target represents an imbalanced class problem, where far more loans that were repaid on time than loans that were not repaid. This problem can be solved by using machine learning models.

**Days Employment**

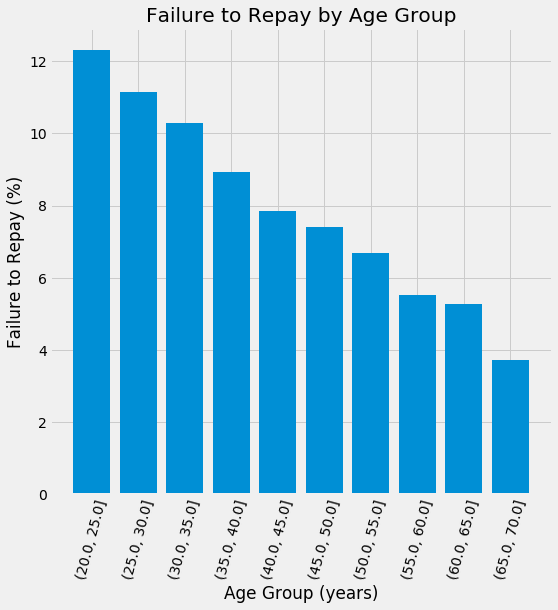
****

The days of employment column contains outliers. The maximum employment days can not be 1000 years. All the outliers in days of employment column have the exact same value. Therefore, we will fill in the anomalous values with NaN. The following figure illustrates the days of employment after outlier treatment.



**Effect of Age on Repayment**

It is one of the significant positive correlations with TARGET is the DAYS\_BIRTH.

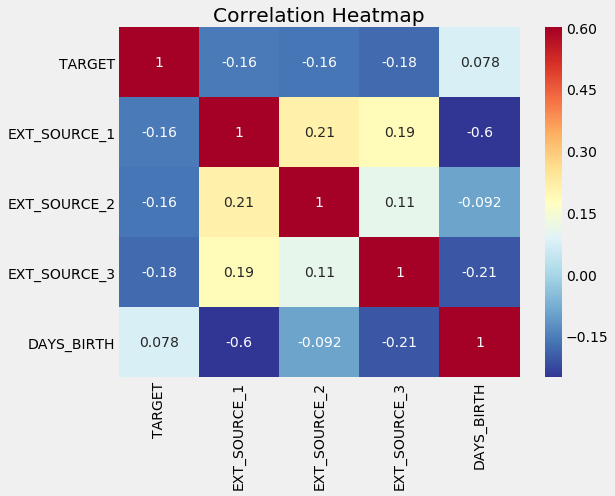


The younger applicants are more likely to not repay the loan.

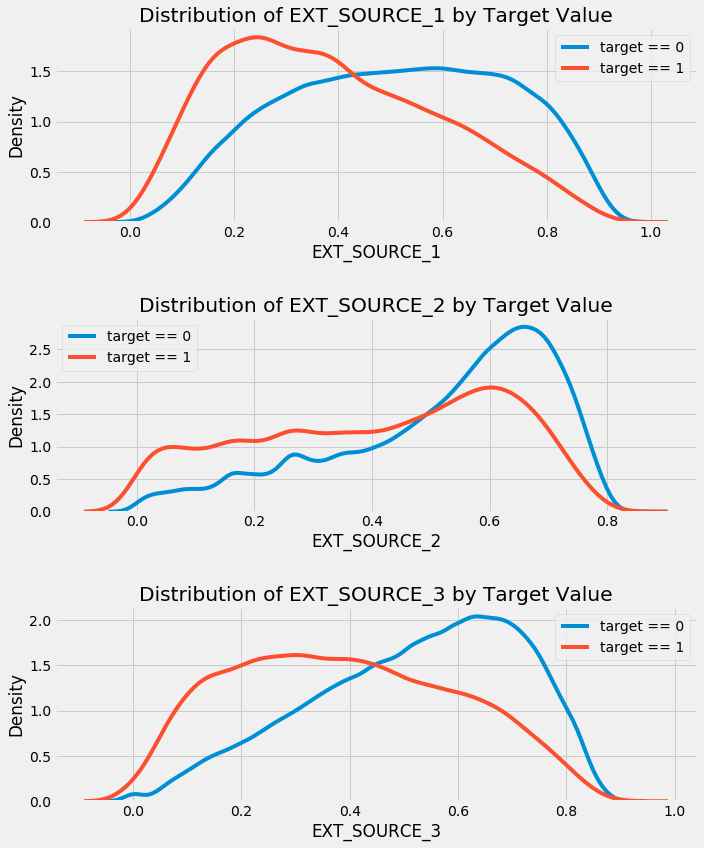
#### 

#### **Effect of Exterior Sources**

EXT\_SOURCE\_3, EXT\_SOURCE\_2, and EXT\_SOURCE\_1 are the variables which have the strongest negative correlations with the TARGET variable.



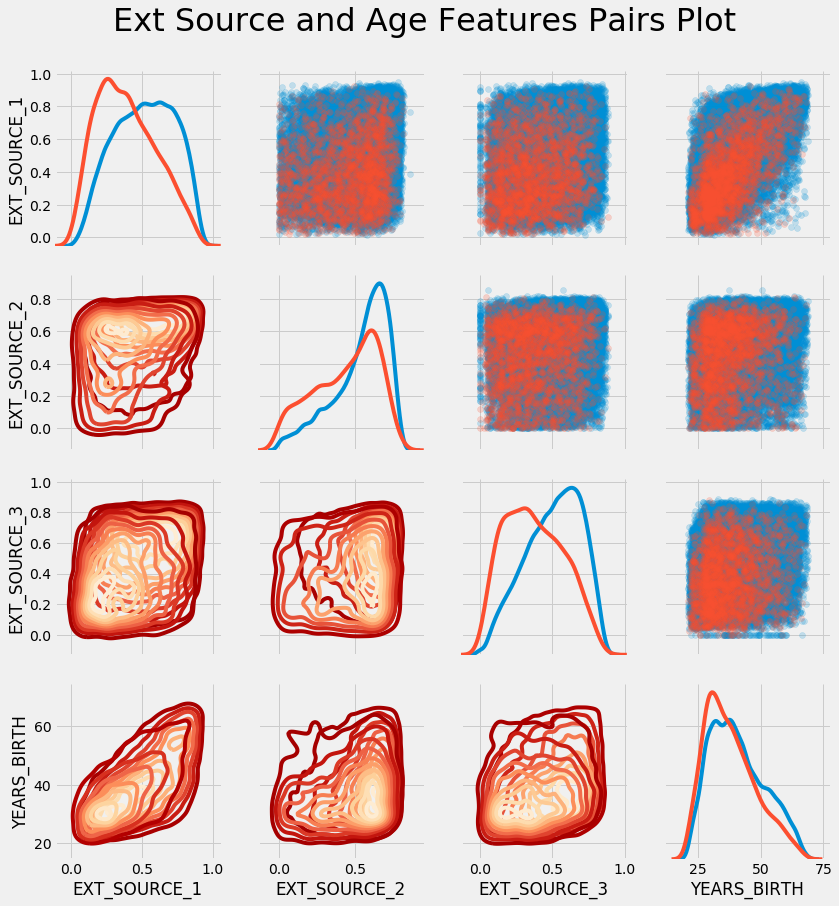
The variables EXT\_SOURCE have negative correlations with the TARGET variable. Client more likely to repay the loan when the value of the EXT\_SOURCE increases.



From the figure we can observe that the variable EXT\_SOURCE\_3 shows the greatest difference between the values of the TARGET variable. Client more likely to repay the loan if the value of the EXT\_SOURCE\_3 variable is higher.

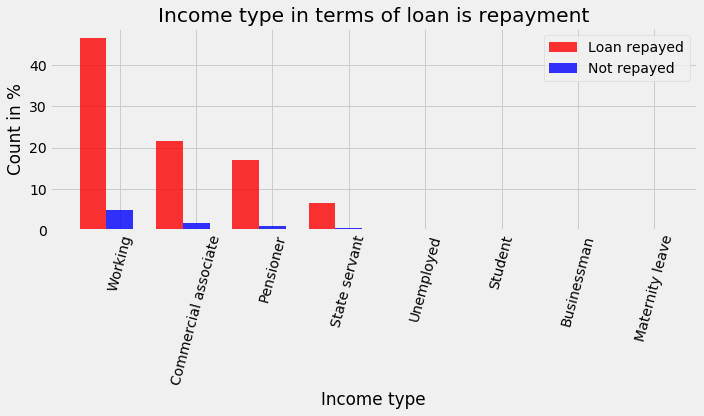
**Pairs Plot**

The Pairs Plot lets us see relationships between multiple pairs of variables as well as distributions of single variables.



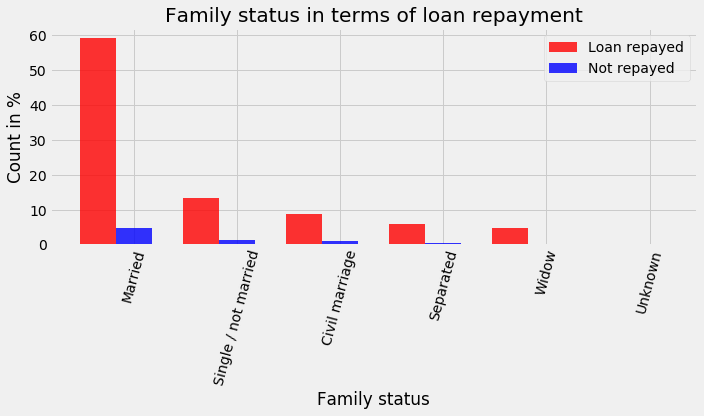
In this plot, the red indicates loans that were not repaid and the blue are loans that are paid. There is a moderate positive linear relationship between the variables EXT\_SOURCE\_1 and the DAYS\_BIRTH.

#### **Observe income type of the client's in terms of loan is repayed or not**



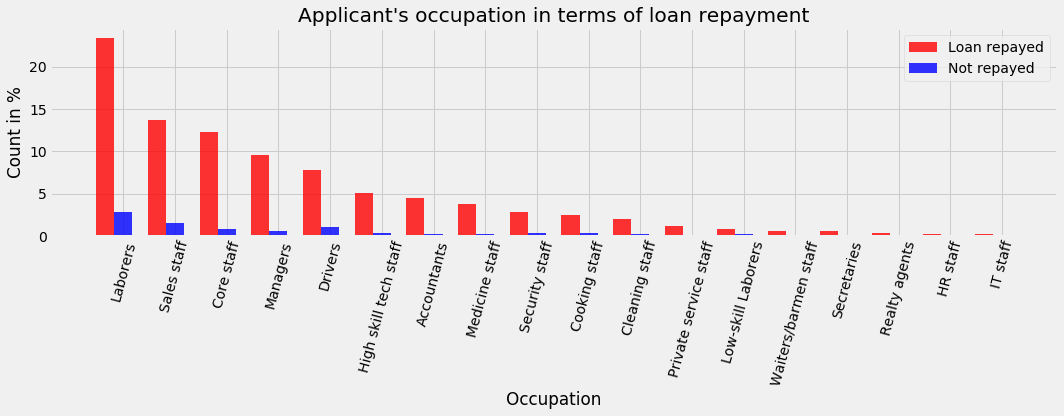
The client's with income type "working" tends to repay loans.

#### **Family status in terms of loan is repaid or not**



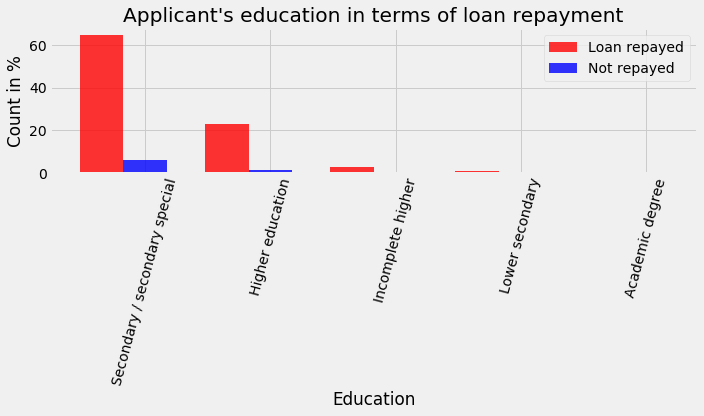
The percentage of the client with family status as "Married" tends to repay loans.

#### **Occupation of the applicant's in terms of loan is repaid or not**



Clients with occupation "Laborers" are better at loan repayment followed by "Sales staff", "core staff", "Managers", and "Drivers". However, the difference in percentage is not significant.

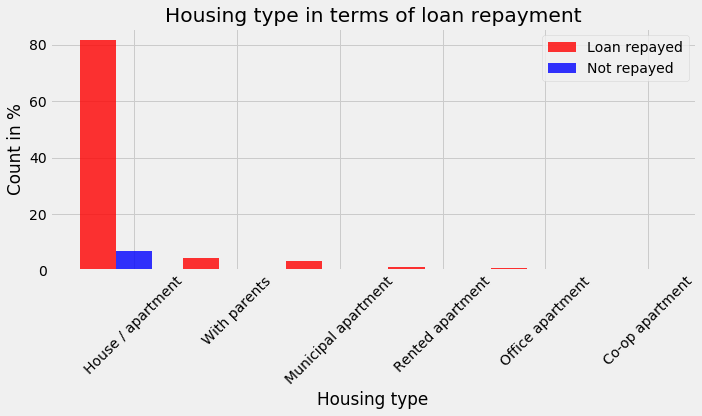
#### **Applicant's education in terms of loan is repaid or not**



The applicant's with "secondary/secondary special" education tend to repay the loan when compared to the other types of education types.

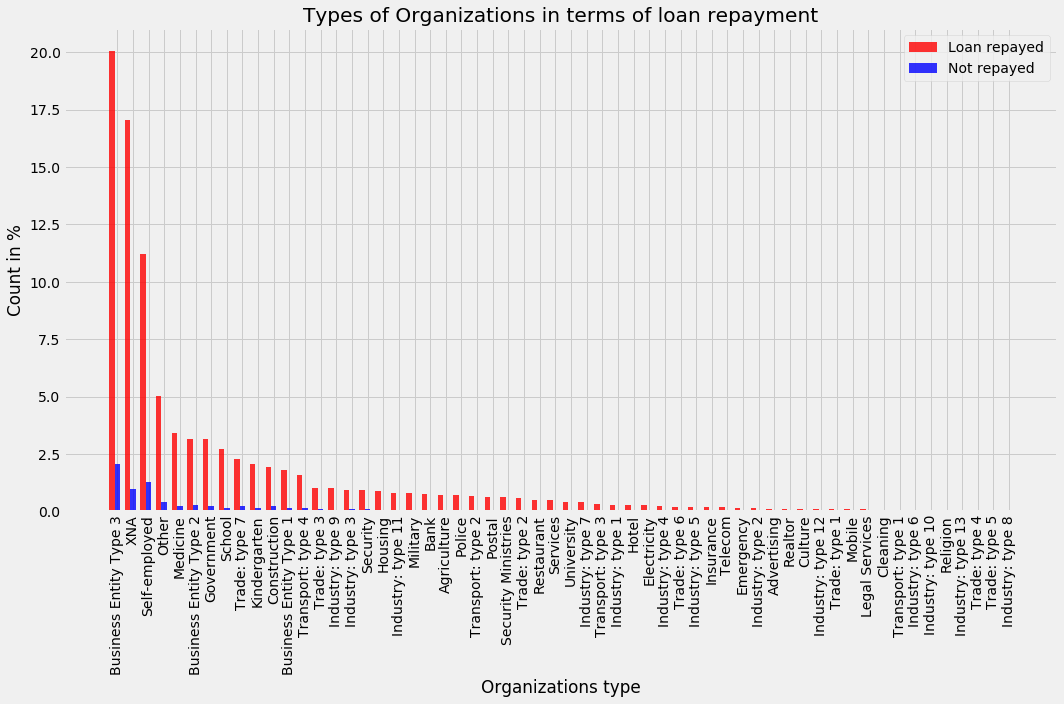
#### 

#### **Applicant's Housing type in terms of loan is repaid or not**



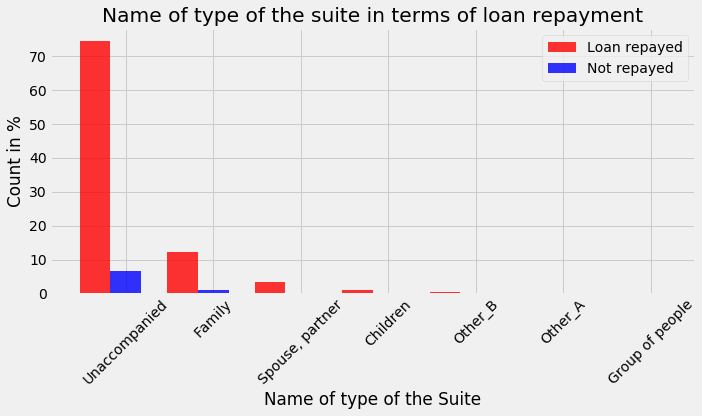
The applicant's who lives in the "House/apartment" are more likely to repay the loan.

#### **Types of Organizations of the Applicant's in terms of loan is repaid or not**



Applicant's organization type doesn't affect the probability of the loan repayment. However, among all the organization types clients' who are associated with "Business Entity Type 3" are more likely to repay the loan.

#### **Name of type of the suite of the Applicant's in terms of loan is repaid or not**

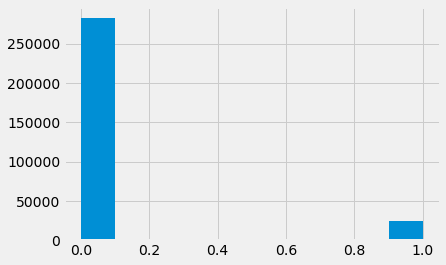


The applicant's who are "Unaccompanied" are more likely to repay the loan.

**Statistical Data Analysis**

There are 122 features in the dataset. The importance of these features will be evaluated in this section. Correlation tests will be performed with a feature against the target feature to check if they are related. I will use Kendall’s rank correlation in this study to test whether two features have a monotonic relationship. Chi-Squared test will be used to test whether categorical variables are related to the Target variable.

### **Target variable normality check:**



Using D’Agostino’s K^2 Test I got that Target variable is not look Gaussian.

### **Client's Age on loan repayment**

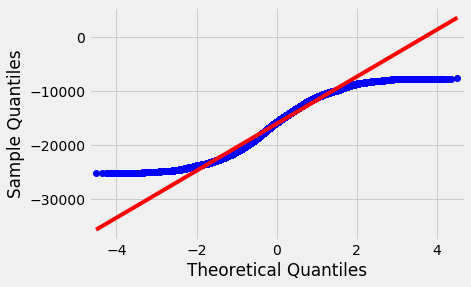
I want to observe is there a significant difference between the age groups in terms of loan repayment.

**Research questions: Is there any relation between age groups in terms of loan repayment?**

H0: There is no correlation between age groups and loan repayment.

H1: Loan repayment is related to the age groups.

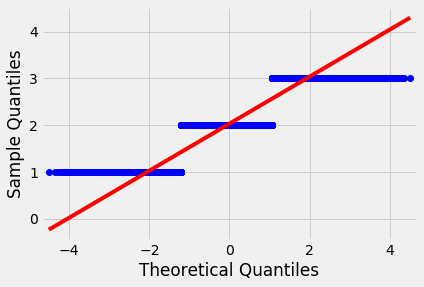
First to check whether the client's age data is Gaussian or not.



Client’s age is not Gaussian.

There is a relation between loan repayment and the age groups, which is found out using Kendall’s Rank Correlation.

**Client's Employment on loan repayment:**



Sample does not look Gaussian.

There is a relation between loan repayment and the years employed found out using Kendall’s Rank Correlation.

**Client's living region on loan repayment:** No relation found.

**External data sources on loan repayment:** There is a relation between external data sources and loan repayment.

Let us check the correlation between the categorical variables and the TARGET variables.

**Client's gender on loan repayment:** There is a dependency between the samples.

**Relation between client own car on loan repayment:** There is a dependency between the samples.

**Relation between client own house or flat on loan repayment:** There is a dependency between the samples.

**Relation between if loan is cash or revolving on loan repayment:** There is a dependency between the samples.

**Relation between accompanying person with client on loan repayment:** There is a dependency between the samples.

**Relation between Client's income type on loan repayment:** There is a dependency between the samples.

**Relation between Client's education type on loan repayment:** There is a dependency between the samples.

**Relation between Client's family status type on loan repayment:** There is a dependency between the samples.

**Relation between Client's housing situation type on loan repayment:** There is a dependency between the samples.

**Relation between Client's occupation type on loan repayment:** There is a dependency between the samples.

**Relation between on which day client's applied for loan on loan repayment:** The two samples are independent.

**Relation between Client's organization type on loan repayment:** There is a dependency between the samples.

1. **Conclusions**

Following conclusions are drawn from our analysis:

* The TARGET variable represents an imbalanced class problem.
* Most of the missing value columns contains more than 50% of missing values. Deleting the missing values is not a good option for this case.
* The days of employment column contains outliers.
* Younger applicants are more likely to not repay the loan.
* External data sources variables have the strongest negative correlations with the TARGET variable.
* Client’s gender, owning a house, income, education, family status, occupation, which day client's applied for loan, and organization have correlations with the TARGET variable.

In this report, I have inspected the data features and the insights gained from this analysis will help in the model building.