Introduction

The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

Understanding the Data

This dataset has 2 files as explained below:

1. 'application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

Data Quality checks and handling missing values:

Two data sets were provided as part of this case study

- Application Data
- Previous application data

In order to perform data quality checks and handling missing values, **we have considered Application data set and performed necessary actions as explained below**.

To find out the missing values and handle it :

- We have observed there were many missing values in this data set, so using the indexing and count of missing values in each column, identified the % of missing values for each column.

	index	Percentage of Missing Values
0	COMMONAREA_MEDI	69.872297
1	COMMONAREA_AVG	69.872297
2	COMMONAREA_MODE	69.872297
3	NONLIVINGAPARTMENTS_MODE	69.432963
4	NONLIVINGAPARTMENTS_MEDI	69.432963
5	NONLIVINGAPARTMENTS_AVG	69.432963
6	FONDKAPREMONT_MODE	68.386172
7	LIVINGAPARTMENTS_MEDI	68.354953
8	LIVINGAPARTMENTS_MODE	68.354953
9	LIVINGAPARTMENTS_AVG	68.354953
10	FLOORSMIN_MEDI	67.848630

- And dropped all columns from data frame for which missing values % is more than 50. We have also found few more columns which are having around 47% missing values. Since these are almost around 50% ,we have removed these columns as well.

- To extend the data enhancements and maintain a data frame with accurate values, we have identified columns with NULL values and even applied the same 47% logic and removed those columns from the data frame.

Columns which has less missing percentage and which can be imputed

As seen from the above missing percentage table, the following columns has around 13% missing values

1. AMT_REQ_CREDIT_BUREAU_QRT
2. AMT_REQ_CREDIT_BUREAU_YEAR
3. AMT_REQ_CREDIT_BUREAU_WEEK
4. AMT_REQ_CREDIT_BUREAU_MON
5. AMT_REQ_CREDIT_BUREAU_DAY
6. AMT_REQ_CREDIT_BUREAU_HOUR

Imputing Values Explanation

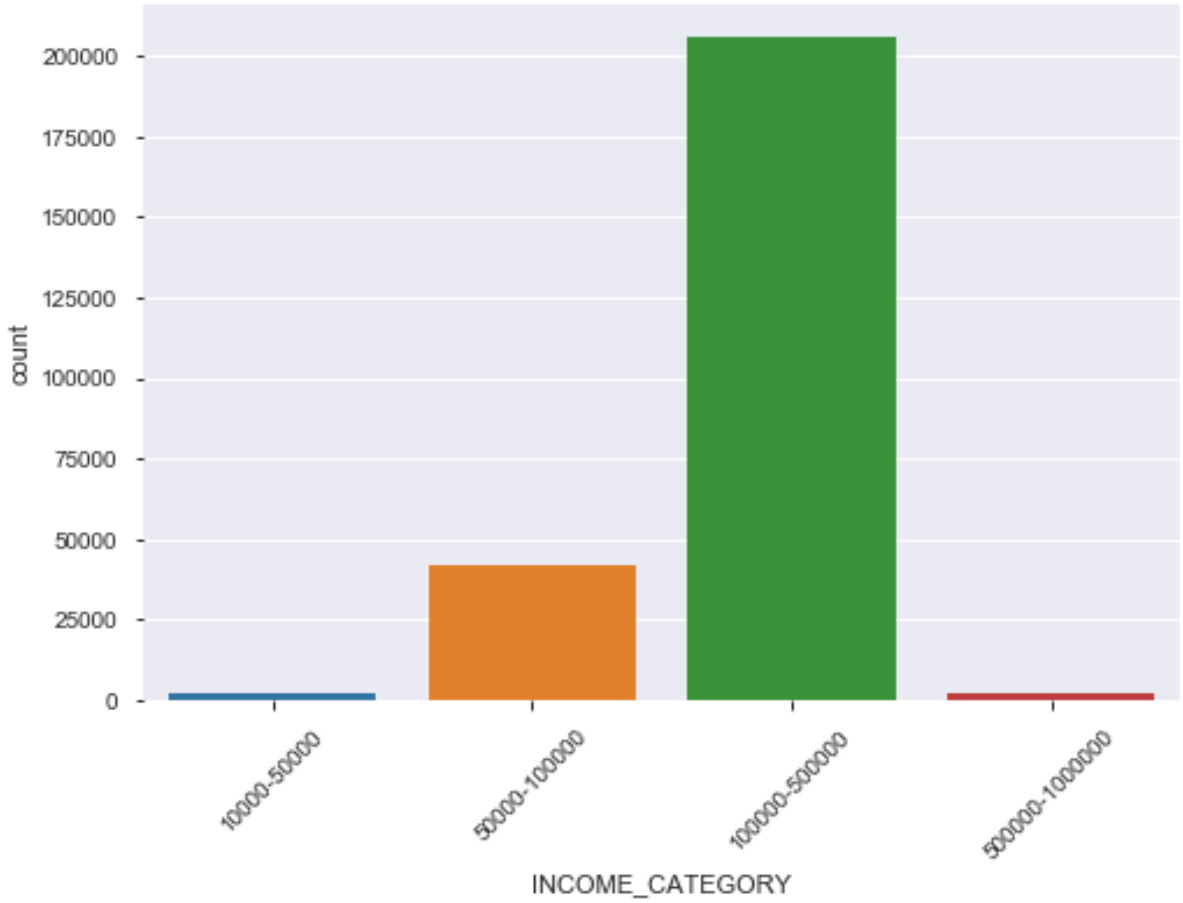
1. The afore-mentioned 6 columns are considered for imputation
2. Given that these columns describe Number of enquiries to Credit Bureau about the client for various duration of time, we can choose either mean or median to fill the missing values.
3. Some of the columns have outliers (for eg: 261 in AMT_REQ_CREDIT_BUREAU_QRT). Hence it is safe to consider median than mean for filling the missing values

Outlier Description

Amount column has outliers which needs to be removed to identify the distribution of income.

Distribution of Continuous Variables using Bucketing Technique

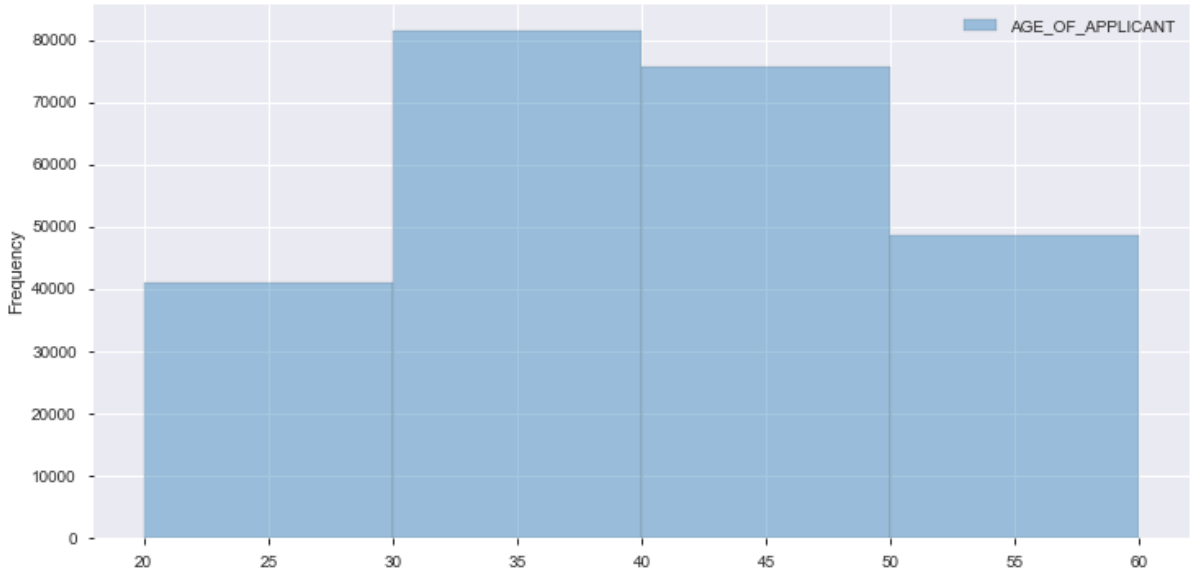
1. **Income Category**



Insights

1. As seen from above count plot, most(around 2.1 lakhs) of the applicants reside in the income category of 100000 to 500000
2. Applicants from 50000-100000 category are just below 50000
3. Applicants in other income category are very minimal for the given data

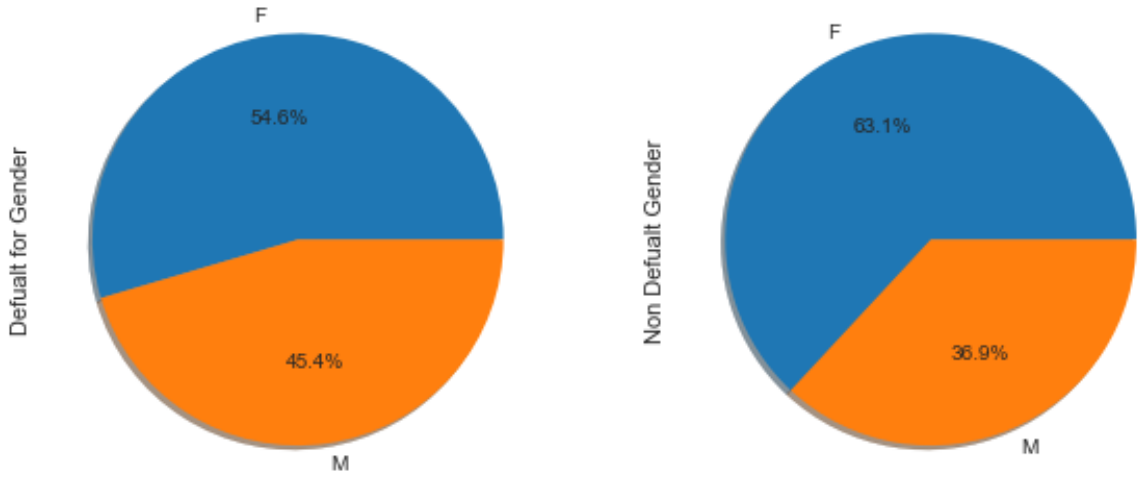
2. **Age Category**



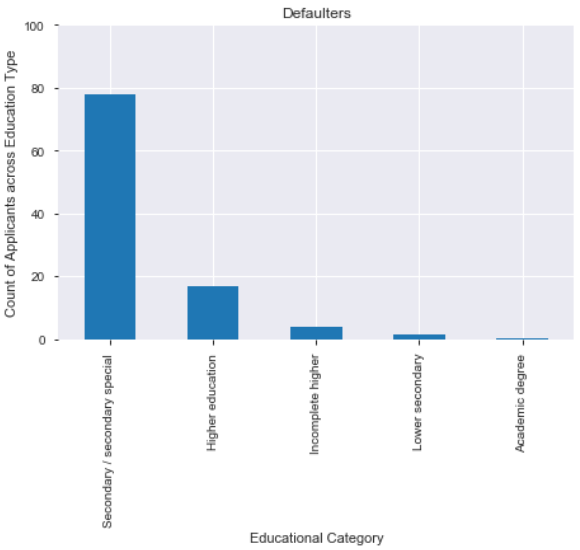
Insights

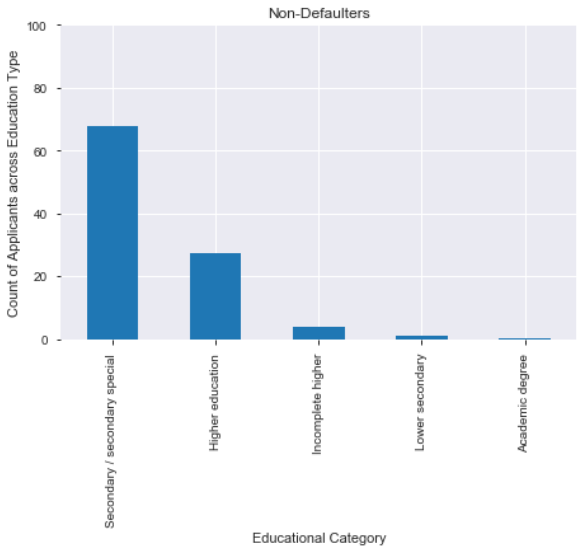
1. Maximum number of applicants resides in the age category 30-40, closely followed by 40-50

Gender Distribution

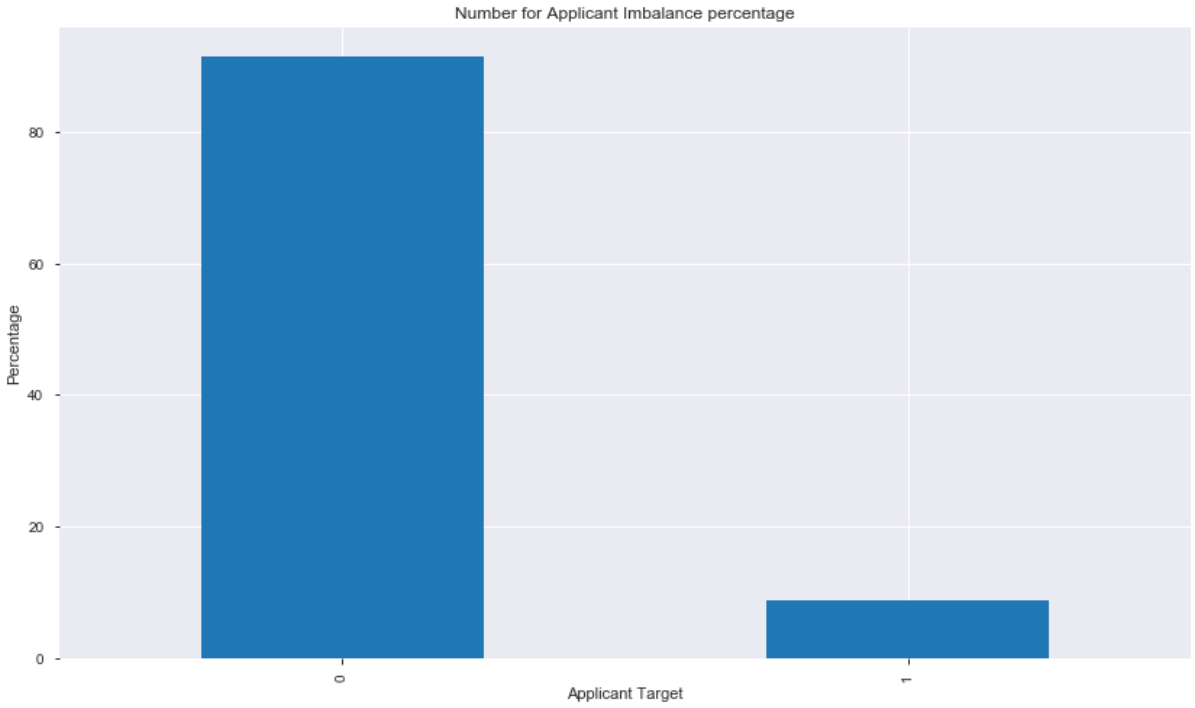


Split Up of Educational Category





Data Imbalance Percentage



As seen from the above graph there is a huge imbalance in percentage of default (around 8%) to that of non-default (around 92%)

Describing Non-Defaulter Data Set

	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_FAM_MEMBERS
count	230288.000000	230288.0	230288.000000	2.302880e+05	2.302880e+05	230276.000000	2.300510e+05	230286.000000
mean	278160.601265	0.0	0.496700	1.764979e+05	6.164767e+05	27902.517331	5.551829e+05	2.253051
std	102857.124689	0.0	0.761312	1.154978e+05	4.114329e+05	14834.207366	3.781456e+05	0.939246
min	100003.000000	0.0	0.000000	2.565000e+04	4.500000e+04	1980.000000	4.050000e+04	1.000000
25%	188934.750000	0.0	0.000000	1.125000e+05	2.762775e+05	16969.500000	2.475000e+05	2.000000
50%	278203.500000	0.0	0.000000	1.575000e+05	5.212800e+05	25843.500000	4.500000e+05	2.000000
75%	367238.250000	0.0	1.000000	2.160000e+05	8.353800e+05	35743.500000	7.020000e+05	3.000000
max	456255.000000	0.0	19.000000	1.800009e+07	4.050000e+06	258025.500000	4.050000e+06	20.000000

Describing Defaulter Data Set

	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_FAM_MEMBERS
count	21832.000000	21832.0	21832.000000	2.183200e+04	2.183200e+04	21832.000000	2.181300e+04	21832.000000
mean	277645.158117	1.0	0.517726	1.697583e+05	5.577250e+05	26859.484037	4.883938e+05	2.246977
std	102369.054202	0.0	0.782145	7.956693e+05	3.460552e+05	12475.806403	3.107725e+05	0.971970
min	100002.000000	1.0	0.000000	2.700000e+04	4.500000e+04	2844.000000	4.500000e+04	1.000000
25%	189838.250000	1.0	0.000000	1.125000e+05	2.844000e+05	17735.625000	2.385000e+05	2.000000
50%	276411.000000	1.0	0.000000	1.440000e+05	4.959855e+05	25578.000000	4.500000e+05	2.000000
75%	366373.500000	1.0	1.000000	2.025000e+05	7.290000e+05	33394.500000	6.750000e+05	3.000000
max	456254.000000	1.0	11.000000	1.170000e+08	4.027680e+06	127507.500000	3.600000e+06	13.000000

Inferences from Reports

Please note that the defaulters rate is compared against total populace to arrive at this inference

1. Comparing the three Pandas Profiling reports/Conventional charts, the following can be inferred
    - Cash loans has more defaulter rate than Revolving loans
    - Male tend to default more than Female
    - People who don't own car and people with car have not-repayment rates of around 8%
    - Significance of Own house really doesn't matter for default prediction, as the default rates for people who don't have houses and the people who have houses are almost similar
    - Below is the order of default rate based on Income Category
        - People who are in income bracket of 50k to 1 lakh default the most
        - People who are in income bracket of 10k to 50k comes second
        - People who are in income bracket of 1 lakh to 5 lakh comes third
        - People who earn above 5 lakhs and less than 10 lakh default the least
        - **So we can provide loans to persons who earns between 5 lakh to 10 lakh , who are less likely to default the loan**
    - Applicants who don't provide their home phones are likely to default than their counterparts
    - Applicants whose occupation type is Driving are ***more likely to default which is closely followed by Labourers***
    - On the other hand, Managers and Core Staff are ***less likely to default***
    - ***Elder age people (above 70) are less likely to default***

Insights and Conclusions¶

- Around 67% of Approved applicants have opted for insurance. The non-defaulter should opt for insurance for repayment of the load
- Around 73.4% customers are repeat customers, which portrays a good image about the financial institution.
- The number of female clients is almost double the number of male clients. Looking to the percent of defaulted credits, males have a higher chance of not returning their loans (around 10%), comparing with women (around 7%).
- From the looks of the above graph, irrespective of applicant owing a car or not, the default rate is almost same
- Most of the clients taking a loan have no children. The number of loans associated with the clients with one children are 4 times smaller, the number of loans associated with the clients with two children are 8 times smaller; clients with 3, 4 or more children are much more rare.
- As for clients with 9 or 11 children, the percent of loans not repaid is 100%.
- Most of applicants for loans are income from Working, followed by Commercial associate, Pensioner and State servant.
- The applicants with the type of income Maternity leave have maximum default, followed by Unemployed.
- The Lower secondary category, have the largest rate of not returning the loan (around 11%). The people with Academic degree have less than 2% not-repayment rate.
- Generally, much more people register in the city they live or work.
- The ones that register in different city than the working or living city are more frequently not-repaying the loans than the ones that register same city (work 11% or live 12%).