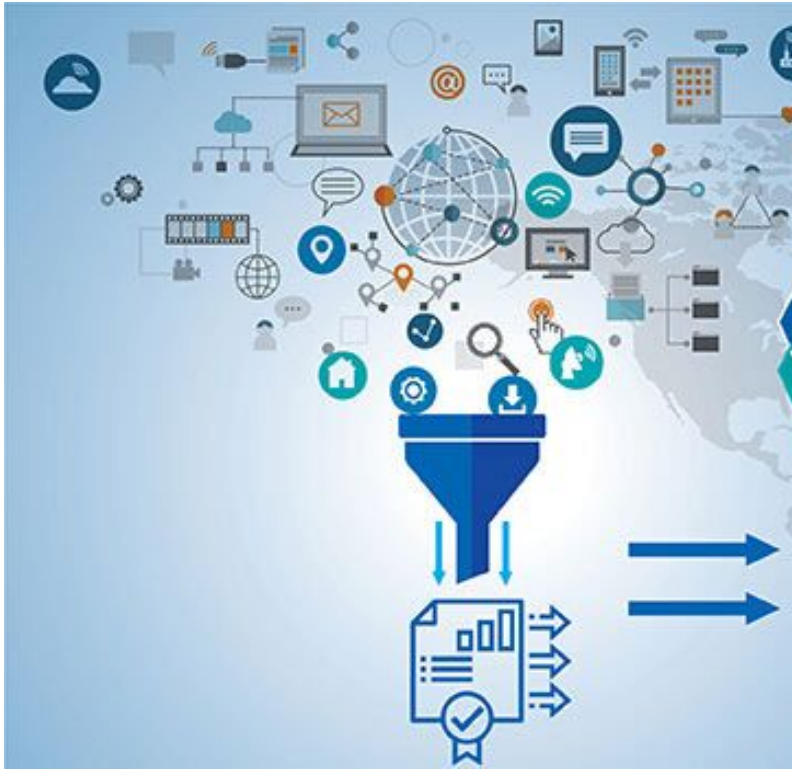
EDA Credit Risk Analysis

Authors: Anand Yati, Sreelakshmi Gosala

Dataset info



Context: Data on loan applications, with target variable as default vs not-default.

Variables provided: Loan applicant demographics, financial, education, asset, credit score, social circle etc.

Raw dataset:

- Columns: 122
- Rows: 307511

Data quality issues detected across columns:

- Missing values in categorical variables.
- Missing values: 41 columns with 50%+ missing values.
- Outliers in numerical variables.
- Wrong data-type, categorical variable detected as numerical.
- Wrong data (e.g. negative age)

Data cleansing

Data issue	Solution
Missing values	Greater than 50% for a column: Drop column
Missing values in categorical variables	● Replace with most frequently occuring class (if most frequent class occurs 50%+ times) ● Replace with a predicted value based on other columns in dataframe (if most frequent class occurs less than 50% times) ● Assign Out of Vocabulary (OOV)/”Unknown” value.
Missing values in numerical variables	● Replace by mean/median ● Replace with a predicted value based on other columns in dataframe (if most frequent class occurs less than 50% times)
Outliers	Clip the column range to remove outliers while calculations
Wrong data (e.g. negative age)	Apply absolute transformation on column to convert all values into positive.
Wrong data type	Identify data-type issues by observing unique values per column and apply data-type transformations to fix it.

Problem statement

Identify variables which are good predictor of applicants who can default on loan versus who will payback on time, so that:

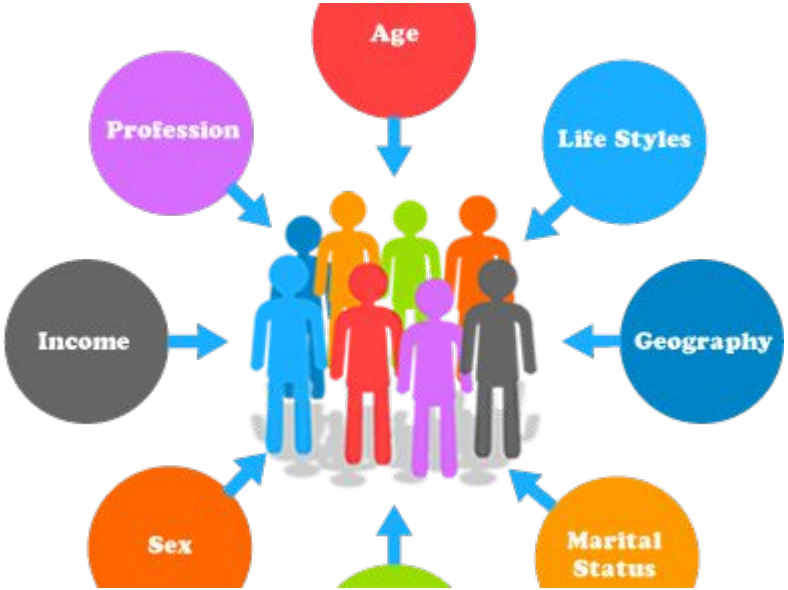
1. Client can either reject the applications with high risk or else increase interest charged to loan applicant to compensate for risk exposure.
2. Client can lower interest rate to attract high quality loan applicants and does not reject good applicants, thus maximising overall revenue and improving portfolio health.

Applicant demographic variables which indicate high likelihood of loan default

- **Gender:** Male loan applicants are at higher risk of default than female applicants.
- **Risky occupation types:** Sales staff, cleaning/cooking staff, drivers,laborers, low-skill labor, security and waiter/barmen
- **Loan type:** “Cash loans” are at a higher risk of default than “revolving loans”.
- **Education level:** Applicants with just secondary education.
- **Ownership of car:** Applicants who do not own a car.
- **Employed years:** Applicants who are employed for less than 4 years.

Applicant demographic variables which indicate low likelihood of loan default

- **Gender:** Female loan applicants are at lower risk of default than female applicants.
- **Low risk occupation types:** Private service staff, High skill tech staff, Accountants, Core staff, HR, IT, Manager, Medicine staff, secretary
- **Loan type:** “Revolving loans” have lower default rates than “cash loans”.
- **Education level:** Highly educated (completed 'HIGHER EDUCATION' or 'ACADEMIC DEGREE) applicants are less likely to default.
- **Ownership of car:** Applicants who own a car are less likely to default.
- **Income:** Applicants earning 200K+ income are at low risk of default.
- **Employed years:** Applicants who are employed for at least 6 years.



Other variables of interest for predicting likelihood of default by an applicant

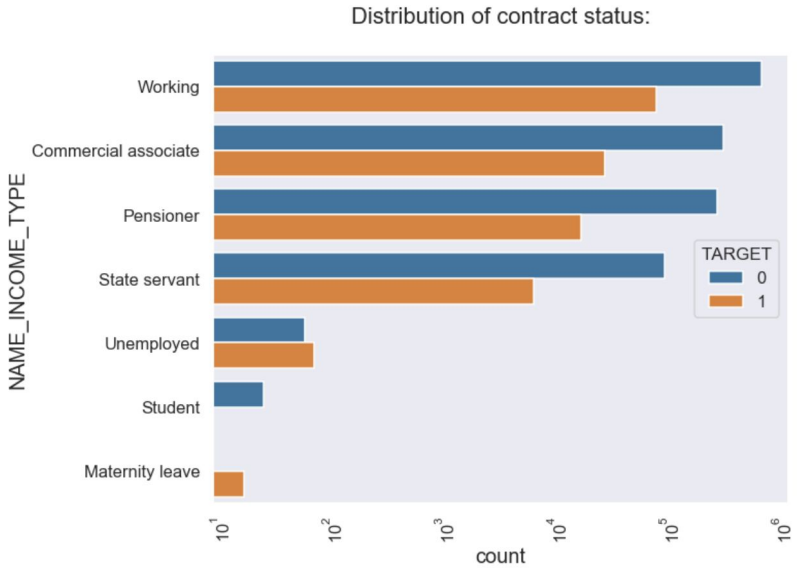
1. DAYS_REGISTRATION: How many days before the application did client change his registration
2. EXT_SOURCE_2 & EXT_SOURCE_3: External score variables
3. DAYS_ID_PUBLISH: How many days before the application did client change the identity document with which he applied for the loan
4. AGE_YEARS: Applicant age as on application day

Other meaningful insights

Who defaults on loan?

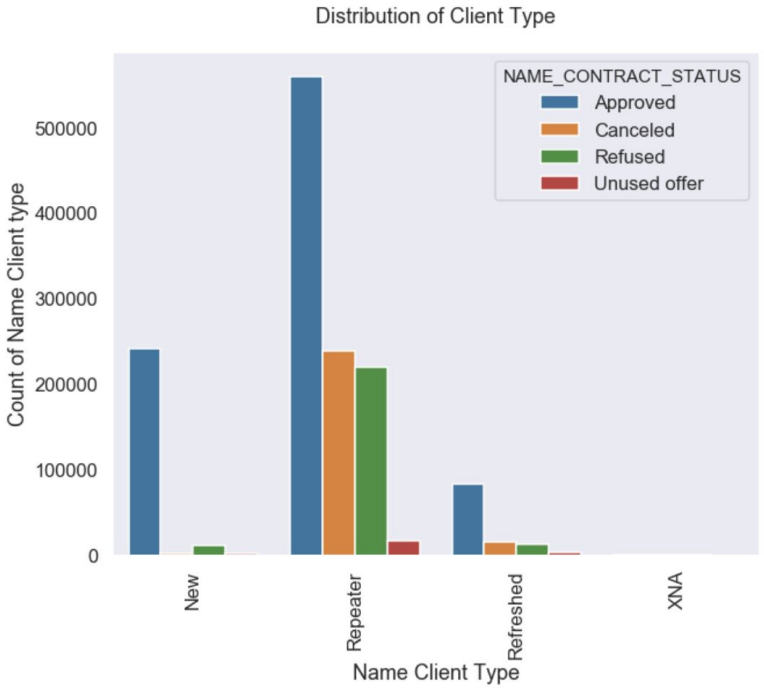
Applicants who do not have a steady income stream (unemployed and on maternity leave) have high default risk.

Students and working professional are at low risk of default.



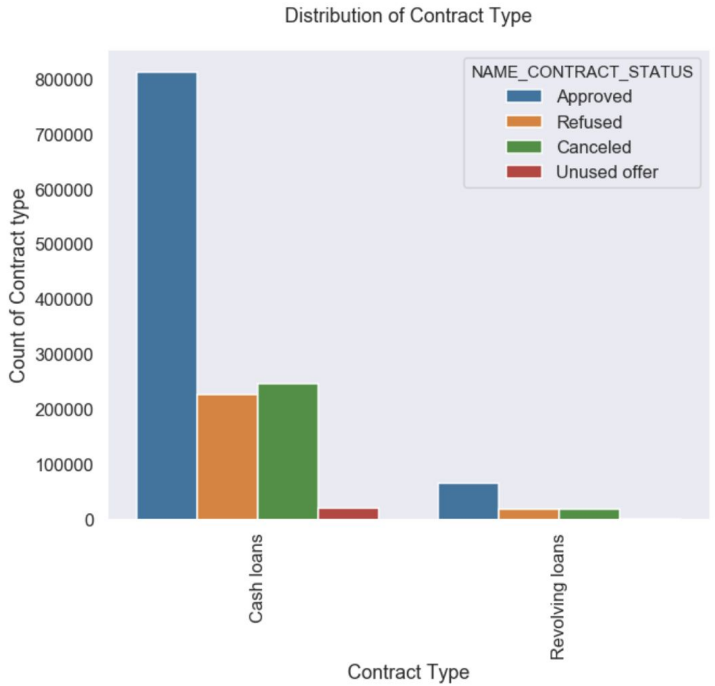
Who are our customers?

- Most of our loan applicants are repeat customers, which are safer based on their prior credit history.
- Our refusal rate for new customers is very low, which might mean we are approving riskier loans. Use variables from this presentation to better ascertain risk



Loan distribution

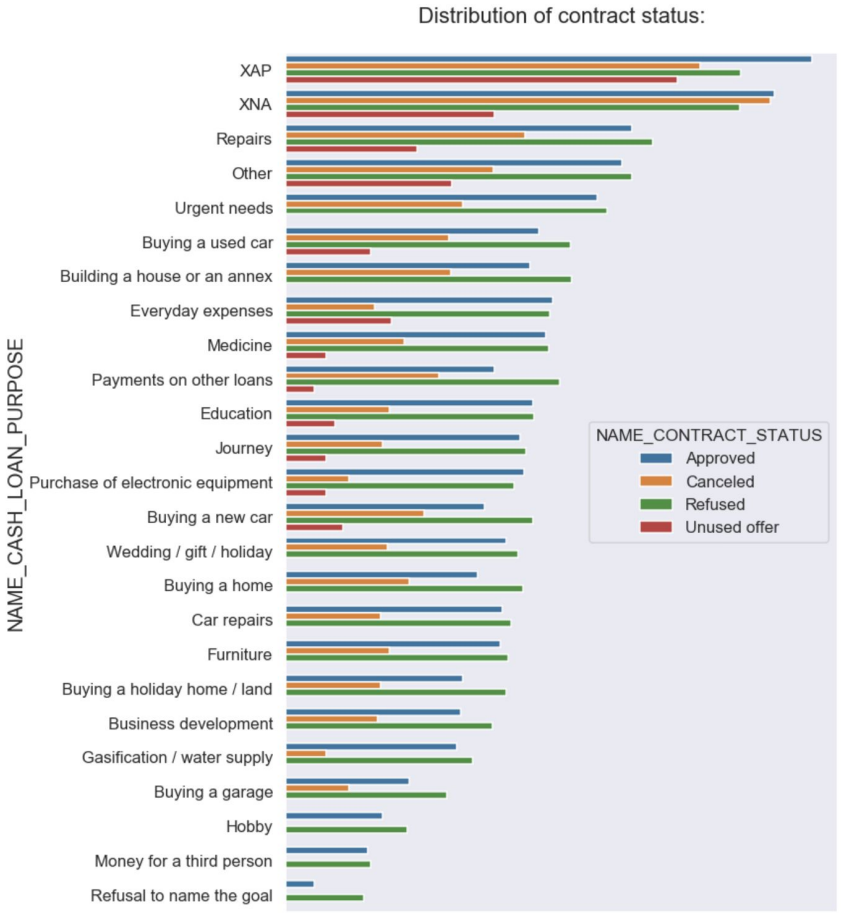
Most of our loans are “Cash loans” which are riskier, client should focus on increasing “revolving loans” in the portfolio.



Risk loan types

Loans for garage purchase, paying other loans, repairs, second hand purchase are riskier.

Education loans, small ticket loans for electronics/general-expenses are safer.



Balancing risk and reward

Not every applicant is same, applicants with higher risk should be charged higher interest rate.

Applicants with lower risk should be charged lower interest rate.

Loan portfolio should have a mix of both low and high risk loans, based on risk appetite of client and return expectations.