Home-Credit-Default-Risk

Can you predict how capable each applicant is of repaying a loan?

Outline

- Problem Definition
- Data
- Exploratory data analysis
- Predictive modeling
- Conclusion

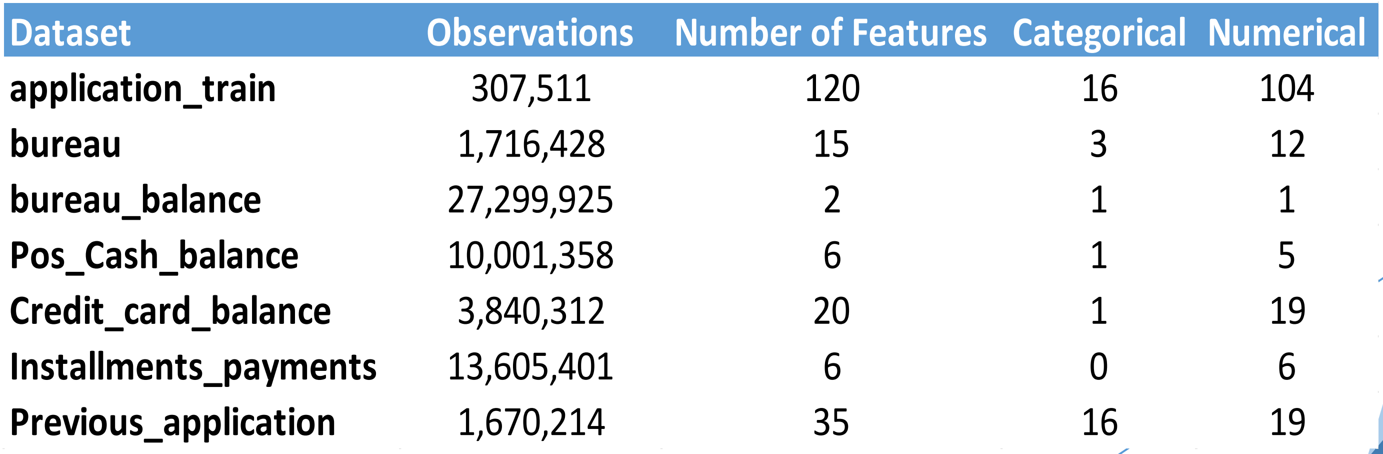
Problem DefInItIon

- Many people struggle to get loans due to insufficient or non-existent credit histories.
- This population is often taken advantage by untrustworthy lenders.
- Home credit aims to increase financial inclusion by serving the unbanked population. In order to do that it provides a dataset to analyze.
- **Goal:** identify if a new client shows a high risk for loan default.

Dataset: Kaggle’s Competition Data

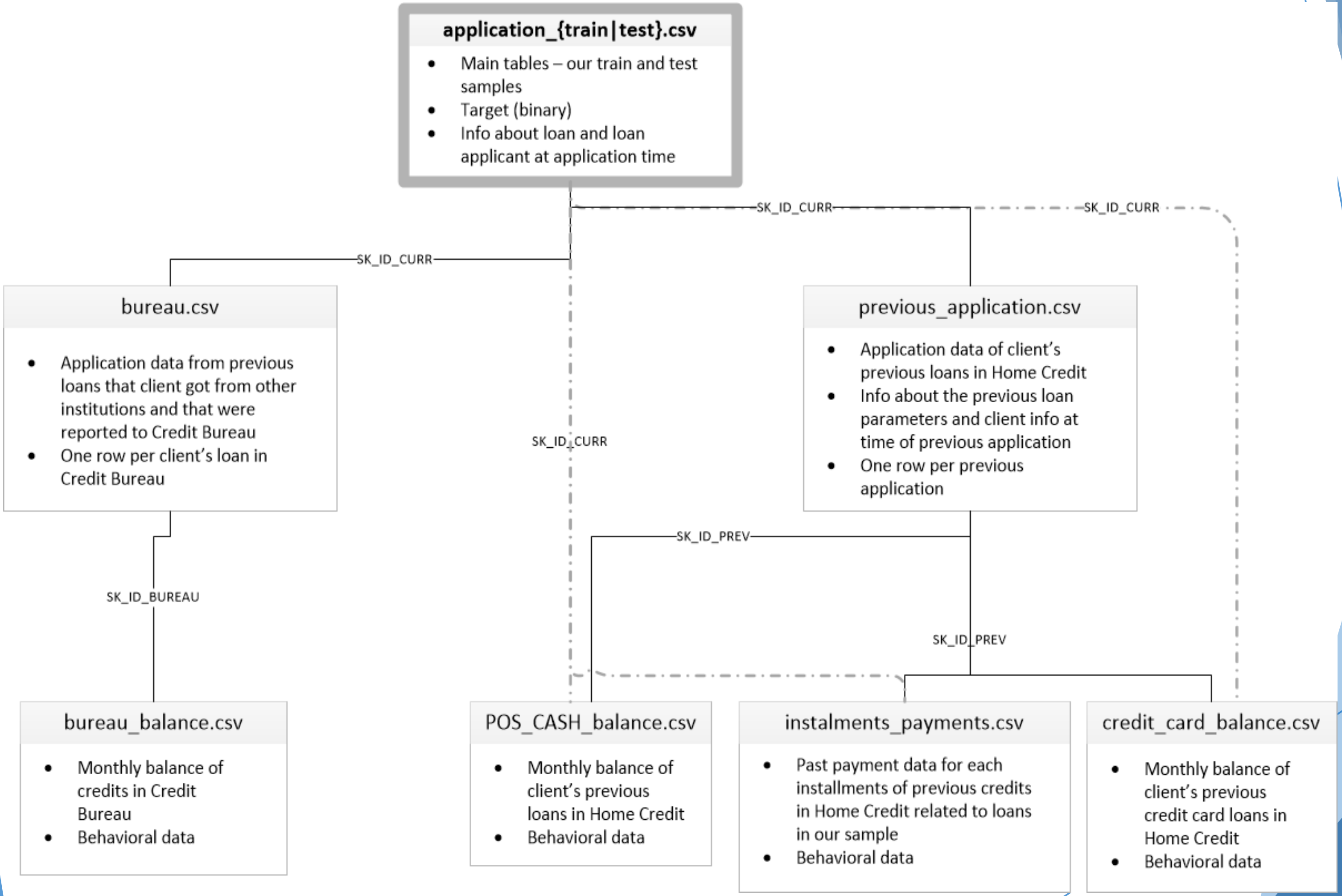
There are seven datasets

- Application_train
- Bureau
- Bureau Balance
- POS Cash Balance
- Credit Card Balance
- Previous Application
- Installments Payments



Dataset	Observations	Number of Features	Categorical	Numerical
application_train	307,511	120	16	104
bureau	1,716,428	15	3	12
bureau_balance	27,299,925	2	1	1
Pos_Cash_balance	10,001,358	6	1	5
Credit_card_balance	3,840,312	20	1	19
Installments_payments	13,605,401	6	0	6
Previous_application	1,670,214	35	16	19

Relationship Between Files



application_{train|test}.csv

- Main tables – our train and test samples
- Target (binary)
- Info about loan and loan applicant at application time

bureau.csv

- Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
- One row per client's loan in Credit Bureau

previous_application.csv

- Application data of client's previous loans in Home Credit
- Info about the previous loan parameters and client info at time of previous application
- One row per previous application

bureau_balance.csv

- Monthly balance of credits in Credit Bureau
- Behavioral data

POS_CASH_balance.csv

- Monthly balance of client's previous loans in Home Credit
- Behavioral data

instalments_payments.csv

- Past payment data for each installments of previous credits in Home Credit related to loans in our sample
- Behavioral data

credit_card_balance.csv

- Monthly balance of client's previous credit card loans in Home Credit
- Behavioral data

Methodology: Supervised Machine Learning

The objective: predict whether or not an applicant will be able to repay a loan. This is a standard supervised classification task.

Target Labels

0

- Target label of 0 indicates that there was no difficulty in repaying the loan on time.

1

- Target label of 1 indicates that there was difficulty in repaying the loan on time.

Steps taken to clean the data

- **Converting string categorical columns into numerical** – Label encoding.
- **Converting string categorical columns into numerical and adding new columns to indicate the presence of categorical variables** – One hot encoding.
- **Replacing illogical outliers with empty values** (NAN values).
- **Imputing empty cells with the median of the values**. grouping.

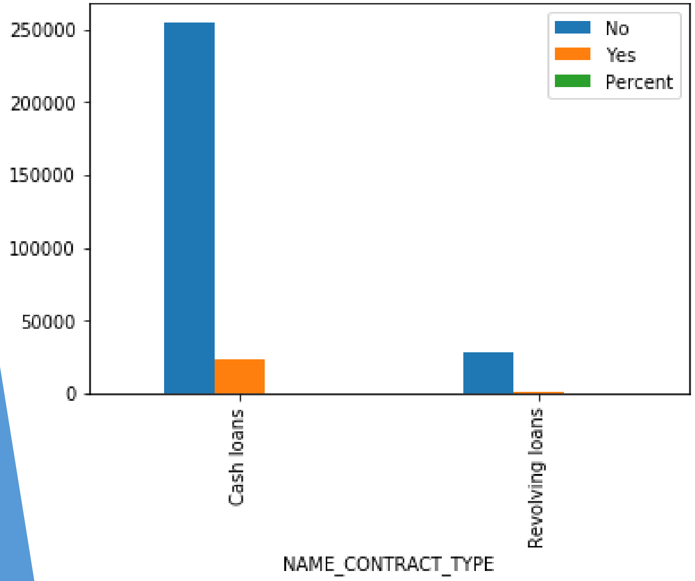
Exploratory Data Analysis

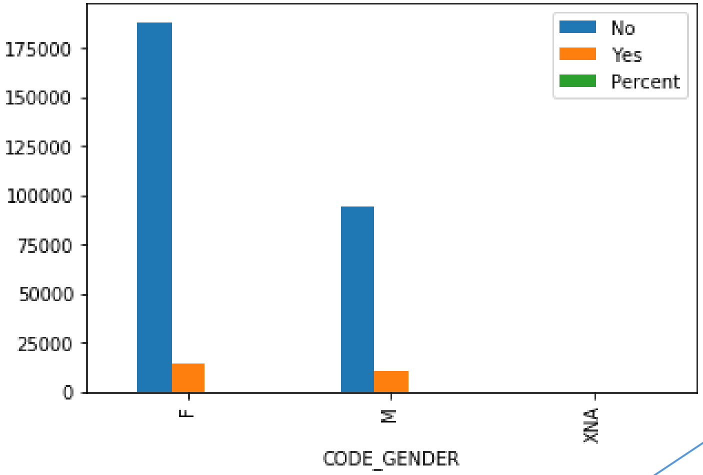
Questions

1. Is there difference in default rate depending on loan type (cash vs revolving)?
2. Is there difference in default rate depending on gender?
3. If a customer owns a car, would it affect payment of loan?
4. If a customer owns a real estate, would it affect payment of loan?
5. Does the family status of client affects payment?
6. Is having children affects payment?
7. Does income type have an affect on repayment?
8. Does the occupation have an affect on repayment?
9. Does the organization type has an affect on repayment?
10. How the education of client affects repayment rate?

Default among revolving loans is lower, males have higher default rate

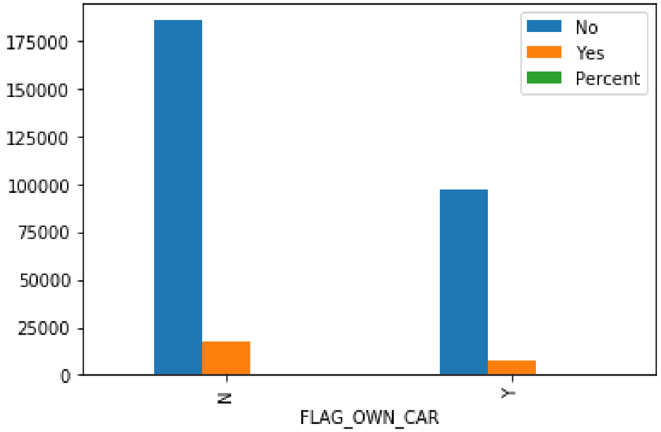
- **Default Rate: Cash loans (8.3%) vs Revolving loans (5.5%)**
- **Default Rate: Male (10.1%) vs Women (7%)**

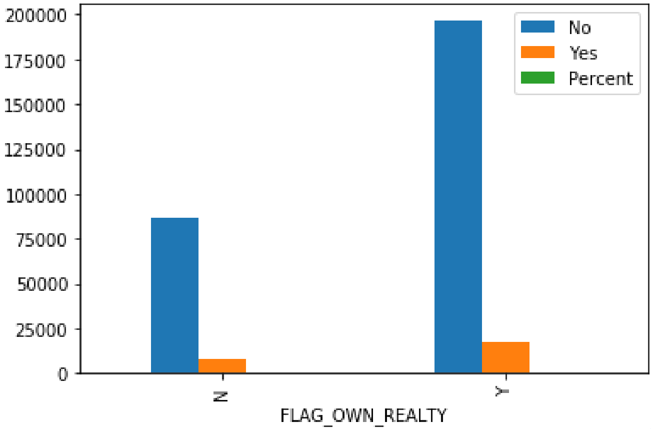




The clients with a car are less likely to default, owning a real estate doesnot effect default

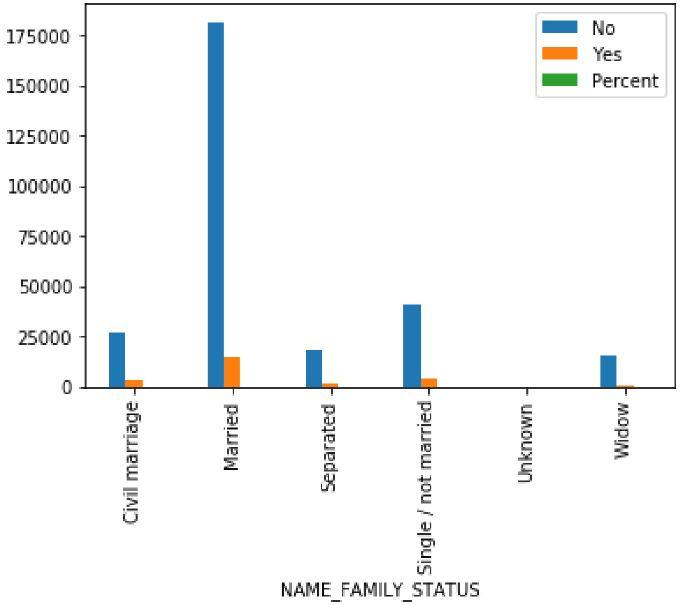
- Default Rate: no car (8.5%) vs car owner (7.2%)
- Default Rate: Both with real estate and no real estate (8%)



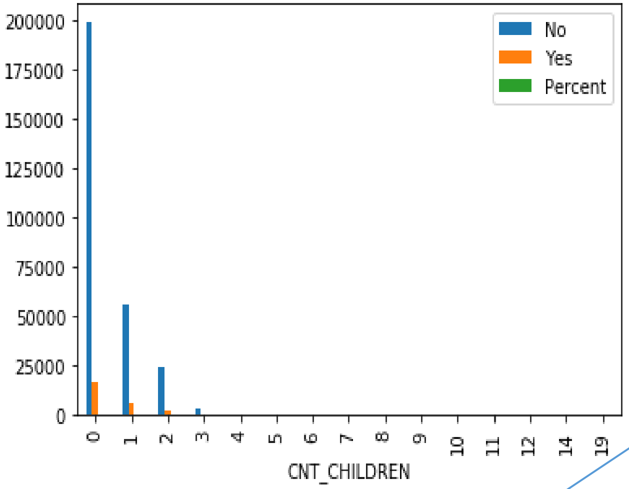


Civil marriage has the highest percent of default rate (10%)

- Civil marriage has the highest percent of default rate (10%), Widow has the lowest

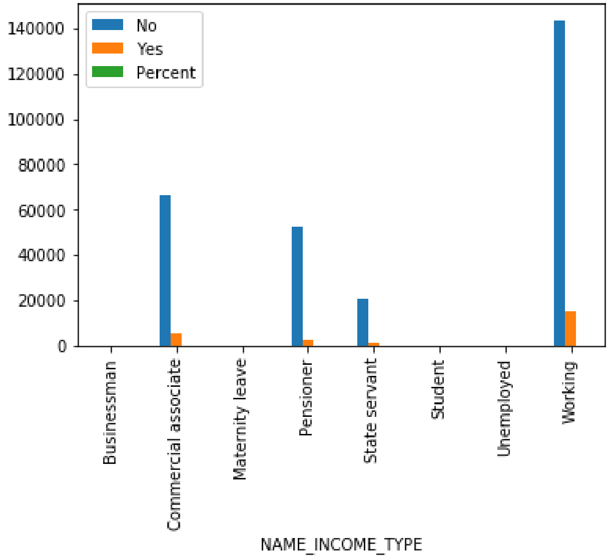


- Clients with no children, 1, 2, 3, and 5 children have percents of default rate around the average (10%).

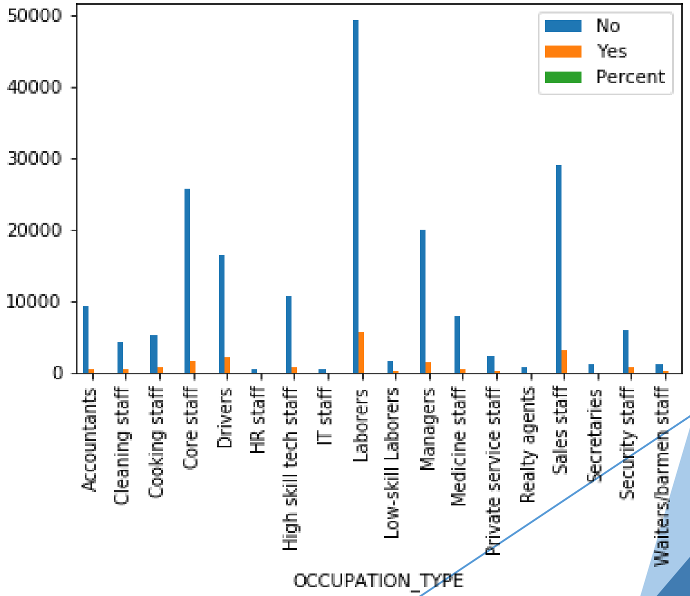


Most applicants have working job status and the loans are mostly taken by laborers and sales staff

- Applicants with the type of income Maternity and Unemployed have highest level of default.



- IT staff take the lowest amount of loans. Low-skill Laborers have the highest default rate (above 17%)

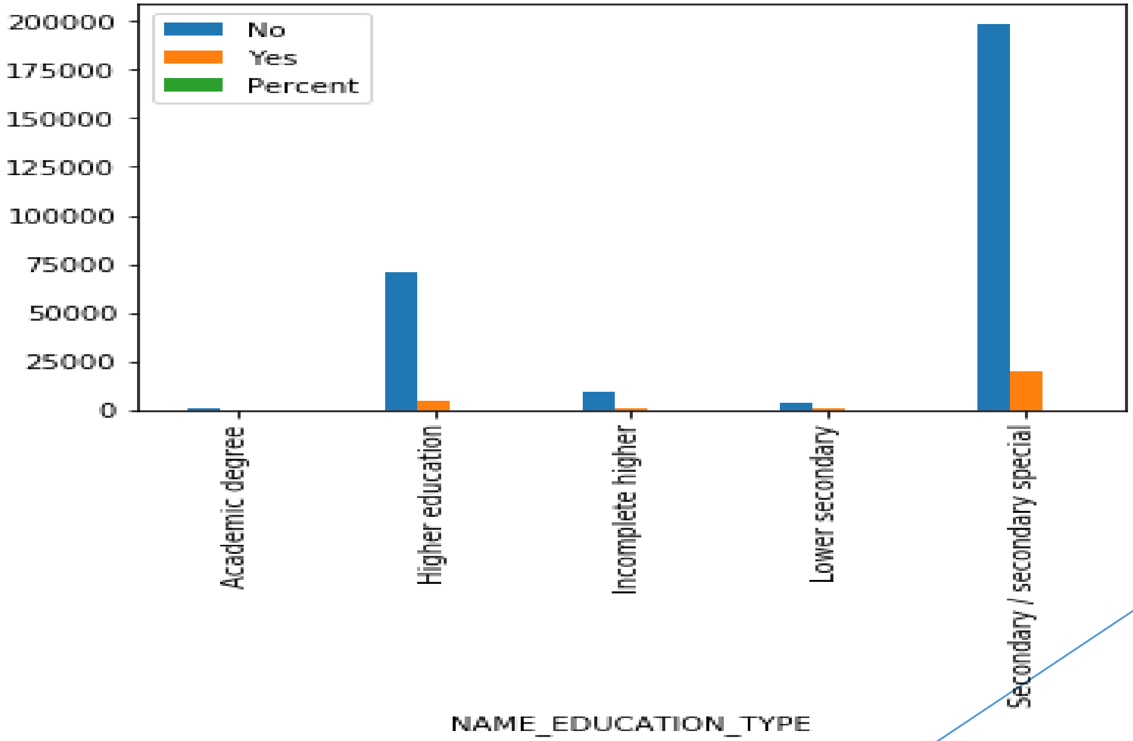


Organizations with highest default rate:

1-Transport(16%)
2-Industry(13.5%)
3-Restaurant (12%)

Education affects repayment rate…

- Majority of the clients have secondary education
- The lower secondary category have the highest default rate (11%)
- People with academic degree have less than 2% default rate.



Model: Aim is to identify whether applicant will default or not on a loan

- Expected Target Outcome: 0 or 1
  - 0 – Not default, 1 – potential default

- Perfomance Metrics used :ROC AUC

- Models currently used : Logistic regression, Random Forest

Models and Evaluation

- Training and Testing datasets used to evaluate the model
- Out of the main training dataset, a certain percentage is kept untrained to test the model’s performance.
  - Training set and validation set are split in following percentages: 66.66% : 33.33%.
- Models: Logistic Regression and Random Forest
- Metric: ROC AUC (Receiver Operating Characteristic Area Under the Curve) metric to judge the results. The ROC curve graphs the true positive rate versus the false positive rate, AUC is the area under the curve.
- ROC AUC Scores:
  - Baseline: 0.5
  - Logistic Regression Model: 0.678
  - Random Forest Model: 0.708
- F1 Score:
  - Logistic Regression Model and Random Forest Model: 0.88