# Credit Risk Modelling

By **Gaurav Kumar**
Project Mentor: UdayaPrakash Somasundaram

# Table of contents

# Problem Statement

Based on multiple attributes or features try to determine whether loan should granted to a user or not

# Credit Risk Modelling

Credit Risk refers to the risk associated with borrower for not repaying the loan and credit risk modelling stands for developing a data driven risk models which calculates the chances of a borrower defaults on loan.

01

# Exploratory Data Analysis

Insights & Observations

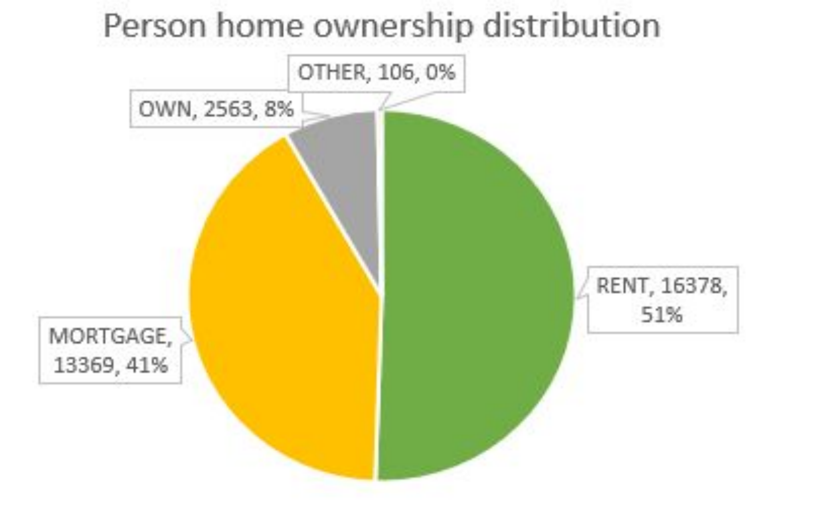
# Data Overview

- The data contains **32,416** unique records and **12** attributes
- **3982** records have some missing values
- "**loan_status**" is the target variable
- **22%** of the records belong to **positive** class and **78%** of the records belong to **negative** class

# Defaulters v/s Non Defaulters


From here we can clearly observe that the data is imbalanced and we need to take care of this during modelling by choosing a suitable metric for model evaluation

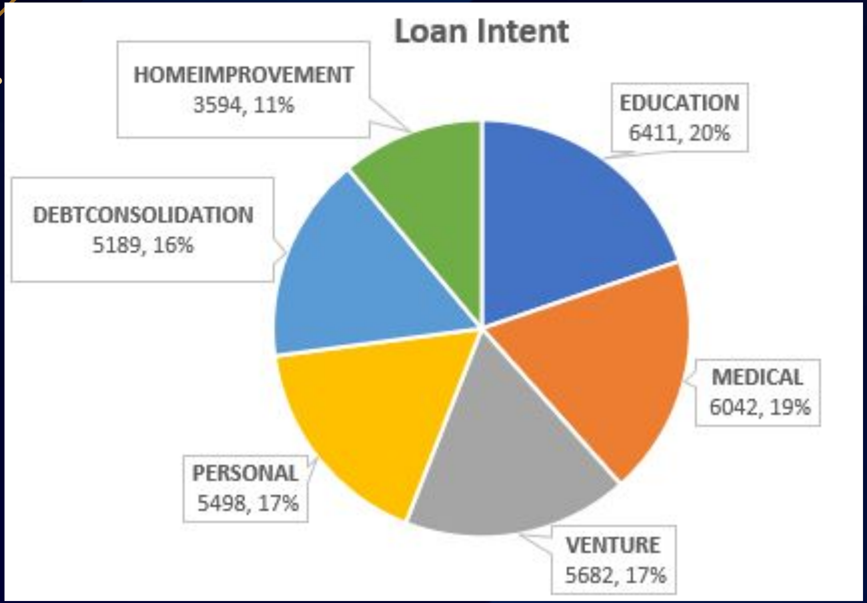
# Person Home Ownership Distribution




| person_home_ownership | loan_status | %age |
|---|---|---|
| MORTGAGE | 0 | 87.41632 |
| | 1 | 12.58368 |
| OTHER | 0 | 71.2766 |
| | 1 | 28.7234 |
| OWN | 0 | 93.33942 |
| | 1 | 6.660584 |
| RENT | 0 | 68.76503 |
| | 1 | 31.23497 |

**Observation:** Those who are living on rent have more probability of default

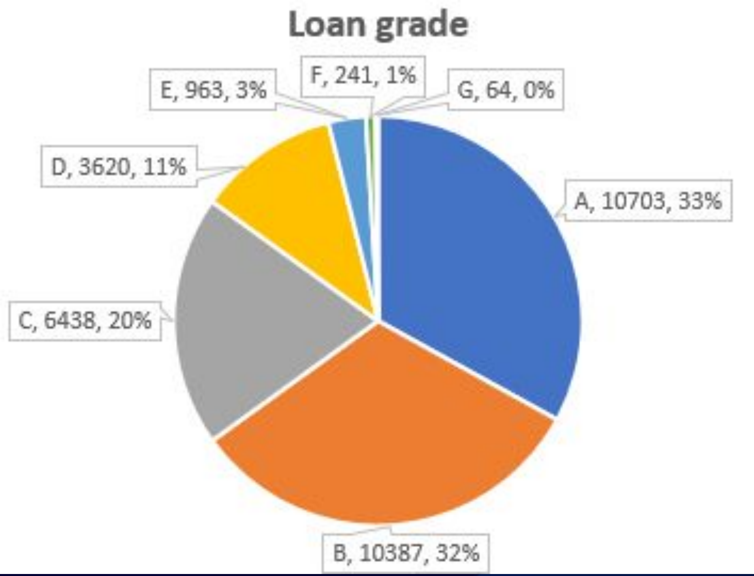
# Loan Intent Distribution




| loan_intent | loan_status | % age |
|---|---|---|
| DEBTCONSOLIDATION | 0 | 71.61 |
| | 1 | 28.39 |
| EDUCATION | 0 | 82.98 |
| | 1 | 17.02 |
| HOMEIMPROVEMENT | 0 | 74.33 |
| | 1 | 25.67 |
| MEDICAL | 0 | 73.15 |
| | 1 | 26.85 |
| PERSONAL | 0 | 80.25 |
| | 1 | 19.75 |
| VENTURE | 0 | 85.38 |
| | 1 | 14.62 |

**Observation:** From here we can understand that those who are already having debts are having more chances of default.

# Loan Grade Distribution




| loan_grade | loan_status | % age |
|---|---|---|
| A | 0 | 90.39 |
| | 1 | 9.61 |
| B | 0 | 84.12 |
| | 1 | 15.88 |
| C | 0 | 79.70 |
| | 1 | 20.30 |
| D | 0 | 40.79 |
| | 1 | 59.21 |
| E | 0 | 35.40 |
| | 1 | 64.60 |
| F | 0 | 30.14 |
| | 1 | 69.86 |
| G | 0 | 1.69 |
| | 1 | 98.31 |

**Observation:** Those who are taking a higher grade loan have more probability of default

- Also the categories E,F,G can be merged into E only

# 02.1
# Data preprocessing
# and
# Feature Engineering
# (For Logistic Regression)

# WOE and Information Gain (Overview)

The **weight of evidence** tells the predictive power of an independent variable in relation to the dependent variable. Since it evolved from credit scoring world, it is generally described as a measure of the separation of good and bad customers.

**WOE** = ln((% of Goods)/(% of Bads)) = ln((% of Non events)/(% of events))

**Information Gain** for a particular attribute is calculated using the following formula:

*IV = ∑ (% of non-events - % of events) * WOE*

It is helpful in order to rank variables on the basis of their Importance.

# WOE and Information Gain

In credit risk dataset:

- The **continuous variables** were binned using **quantile binning** method and monotonic binning was done w.r.t WOE values
- For each bin IV was calculated and summation of all these IV's gave the predictive power of each attribute
- In case of **categorical variables**, the WOE's and IV's were calculated directly for each category.
- Later each bin and  category was substituted with its respective WOE value.

**<u>Conclusion:</u> cb_person_cred_hist_length and person_age with extremely low predictive power were removed from the dataset**

# Multicollinearity Check

- Linear models are usually affected by multicollinearity since it affects the final equation:

$$\hat{Y} = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3$$

- In this case, **VIF (Variance Inflation Factor)** has been used to detect multicollinearity
- Below is the results, I got from the dataset:

| variables | VIF |
|---|---|
| person_age | 1.591779 |
| person_income | 1.803525 |
| person_emp_length | 1.084557 |
| loan_amnt | 1.856605 |
| loan_int_rate | 4.239534 |
| loan_percent_income | 1.920399 |
| cb_person_cred_hist_length | 1.549806 |
| person_home_ownership | 1.11126 |
| loan_intent | 1.018518 |
| loan_grade | 4.256187 |

**Observation:** Here we can clearly see that **loan_int_rate** and **loan_grade** are highly correlated, so one feature can be removed.

**Conclusion:** loan_int_rate was removed because of having lower predictive power and to avoid multicollinearity

# Significance Check

Logistic Regression being a statistical model considers a null hypothesis as:

$H_0$ :- There is no relationship between independent and dependent variable

Following are the P-values that I got for each variable from the LR model:

| VARIABLES | coef | std err | z | P>\|z\| |
|---|---|---|---|---|
| const | -1.3383 | 0.02 | -66.96 | 0 |
| person_income | 0.754 | 0.038 | 20.00 | 0 |
| person_emp_length | 0.4231 | 0.08 | 5.525 | 0 |
| loan_amnt | -0.2332 | 0.15 | -1.53 | 0.127 |
| loan_percent_income | 0.9966 | 0.03 | 33.185 | 0 |
| person_home_ownership | -0.8463 | 0.03 | -26.44 | 0 |
| loan_intent | -1.3368 | 0.06 | -21.409 | 0.000 |
| loan_grade | -1.1906 | 0.02 | -51.76 | 0 |
| cb_person_default_on_file | 0.0377 | 0.050 | 0.76 | 0.447 |

**Observation**: P-value of loan_amnt and cb_preson_default_on_file are more than 0.05, i.e., they are not much impacting my target variable.

**Conclusion**: Removed these two columns from the dataset.

# 02.2

# Data preprocessing and Feature Engineering (For DT and RF)

# Missing Value Imputation

| | |
|---|---|
| person_age | 0 |
| person_income | 0 |
| person_home_ownership | 0 |
| person_emp_length | 895 |
| loan_intent | 0 |
| loan_grade | 0 |
| loan_amnt | 0 |
| loan_int_rate | 3116 |
| loan_status | 0 |
| loan_percent_income | 0 |
| cb_person_default_on_file | 0 |
| cb_person_cred_hist_length | 0 |

Two columns (person_emp_length) and (loan_int_rate) were having missing values.

So, based on previous analysis, the **median imputation** was done on **loan_int_rate w.r.t loan_grade** column, since based on previous analysis, loan_int_rate and loan_grade were highly correlated.

And the remaining missing value records were dropped from the dataset.

# Next Steps

- **Feature Encoding:**
  - Performed label encoding on loan_grade column
  - Performed dummy variable encoding on rest of the categorical columns
- **Train Test Split:**
  - Splitted the data in the ratio of 80:20 for train and test
  - Used stratified sampling during the splitting'
- **Hyperparameter Tuning**:
  - Used GridSearch CV to get optimal hyperparameters, to improve the model results.

# 03

# ML Modelling And Predictions

# Performance Metrics

- **KS statistic**: It basically measures the degree of separation between the CDF's of two classes.

- **ROC AUC score**: It just gives the measurement of the area under the ROC curve made from predictions.

- **Recall Score**:- It basically tells proportion of correctly classified positives out of total positives.

  Recall = TP /( TP + FN)

- **F1 Score**:- It is just the weighted average of precision and recall.

  F1 Score = (precision * recall) / (precision + recall)

- **Capture rate**: It is the proportion of actual positives in each bin. Predicted probabilities are sorted and divided into 10 bins. A Good model shows staircase pattern from top to bottom in case of capture rate

# Results

**Logistic Regression (Using WOE)**

| Metric | Train | Test |
|---|---|---|
| KS | 60.6 | 60.8 |
| ROC AUC | 0.74 | 0.73 |
| Capture Rate (10%) | 37.07 | 36.8 |
| Capture Rate (20%) | 62.97 | 62.1 |
| Recall | 0.534 | 0.52 |
| F1 Score | 0.622 | 0.61 |

**Decision Tree**

| Metric | Train | Test |
|---|---|---|
| KS | 68.7 | 66.5 |
| ROC AUC | 0.85 | 0.84 |
| Capture Rate (10%) | 46.33 | 46.3 |
| Capture Rate (20%) | 73.9 | 72.2 |
| Recall | 0.76 | 0.75 |
| F1 Score | 0.75 | 0.74 |

**Random Forest**

| Metric | Train | Test |
|---|---|---|
| KS | 64.2 | 62.8 |
| ROC AUC | 0.788 | 0.79 |
| Capture Rate (10%) | 43.65 | 44 |
| Capture Rate (20%) | 68.78 | 68.1 |
| Recall | 0.615 | 0.62 |
| F1 Score | 0.7 | 0.7 |