Banque Haolim

Loan Application Presentation

August 3, 2025

Introduction

Welcome to Bank Haolim's loan application presentation.

The goal of this presentation is to analyze how various factors influence loan ap-
proval decisions at the bank, using a loan dataset inspired by Kaggle.

Outline

- Data understanding
- EDA (Exploration data analysis)
- Data preprocessing
- KPI
- Modeling
- Model interpretation and validation
- Deployment and prediction via API
- Conclusion

Data source: Kaggle

Data understanding

- Variable descriptions
- Missing values

Data source: Kaggle

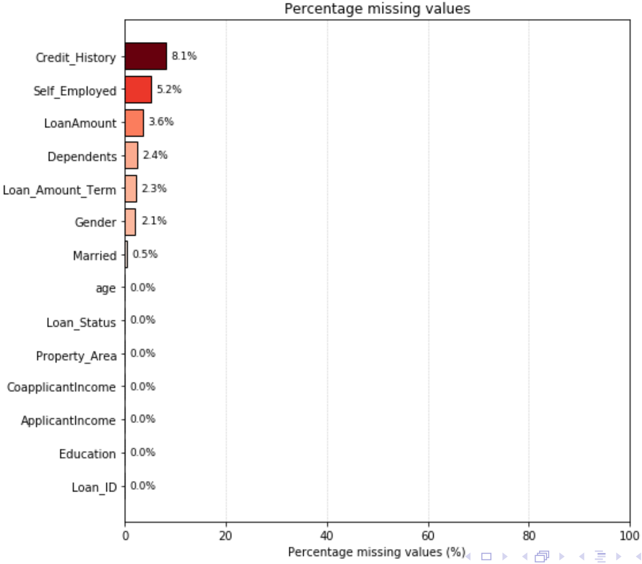
Data Understanding : Variable description

The dataset contains **14 variables** and **614 observations**. The target variable is Loan_Status, which indicates whether a loan application was approved (Y) or rejected (N). The dataset includes the following variables:

- Loan_ID: Unique identifier for each loan application
- Gender: Applicant's gender
- Married: Marital status (Yes or No)
- Dependents: Number of dependents supported by the applicant
- Education: Whether the applicant is a graduate
- Self_Employed: Whether the applicant is self-employed
- ApplicantIncome: Monthly income of the applicant
- CoapplicantIncome: Monthly income of the co-applicant
- LoanAmount: Requested loan amount
- Loan_Amount_Term: Loan duration in months
- Credit_History: Presence of credit history ($1 = \text{Yes}$, $0 = \text{No}$)
- Property_Area: Type of residential area (Urban, Semiurban, Rural)
- Age: Applicant's age (added manually)

Banque Haolim

Data understanding : Missing values



Banque Haolim

Data source: Kaggle

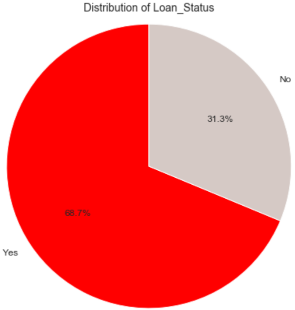
EDA : Exploratory Data Analysis

- Target variable distribution
- Categorical variable analysis
- Numerical feature exploration
- Bivariate relationships with target
- Feature correlation analysis

Data source: Kaggle

EDA : Target variable distribution

Target variable distribution: The target variable `Loan_Status` indicates whether a loan application was approved (`Y` : 68.7%) or not (`N` : 31.3 %). The distribution shows that a majority of applications were approved, revealing a slight class imbalance that may need to be considered during model training.



Banque Haolim

Data source: Kaggle

EDA : Categorical variable analysis

We examined all categorical variables, as well as numerical variables with a limited number of unique values (fewer than 10), such as `Gender`, `Married`, `Education`, `Self_Employed`, `Property_Area`, `Credit_History`, and `Dependents`.

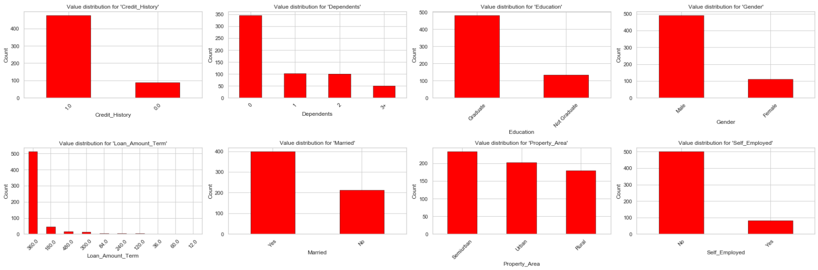


Figure: Histogramm of categorial variables

Banque Haolim

Data source: Kaggle

EDA : Numerical feature exploration

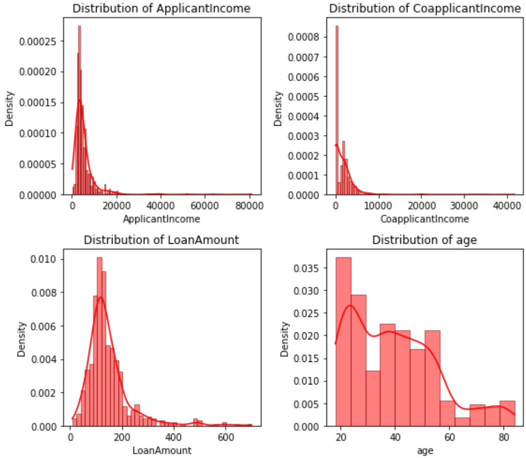


Figure: Density of numerical variables

Banque Haolim

Data source: Kaggle

EDA : Numerical feature exploration

Numerical feature exploration: We explored the distribution of key numerical variables such as `ApplicantIncome`, `CoapplicantIncome`, and `LoanAmount`. These variables showed noticeable skewness and outliers, particularly in income-related features. Understanding their distribution helps assess the need for transformations and potential feature engineering in the modeling stage.

Banque Haolim

EDA : Bivariate relationships with target (categorial variables)

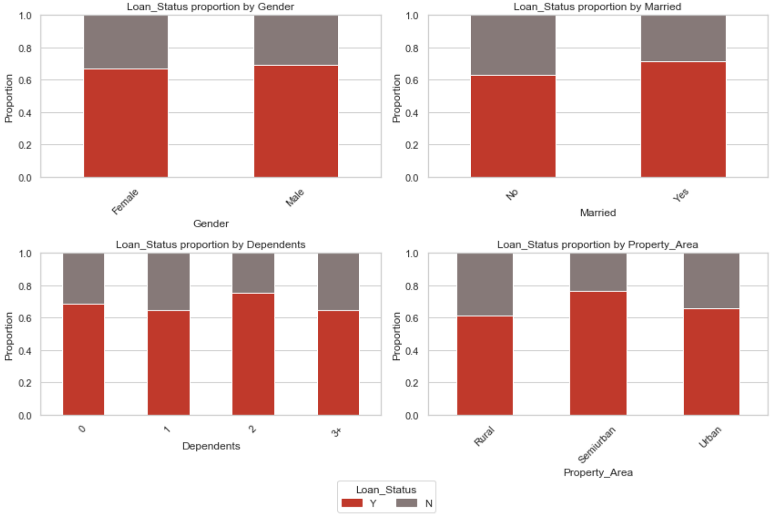


Figure: Density of numerical variables

Banque Haolim

Data source: Kaggle

EDA : Bivariate relationships with target (categorial variables)

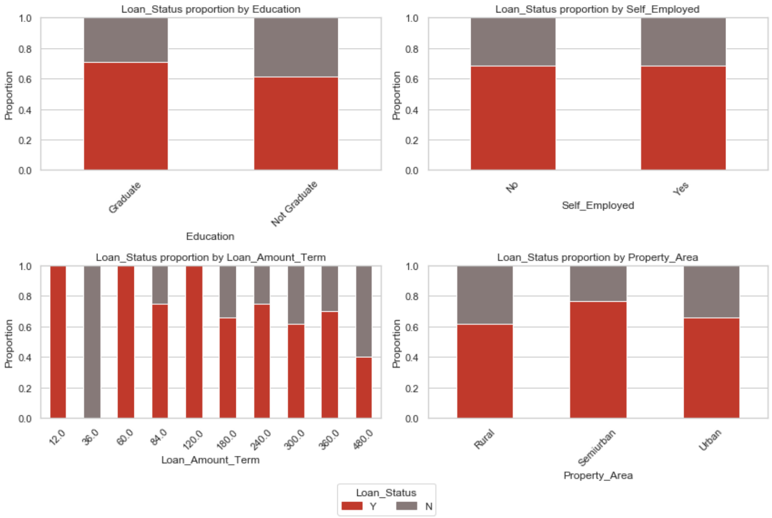


Figure: Density of numerical variables

Data source: Kaggle

EDA : Bivariate relationships with target (categorial variables)

Data source: Kaggle

EDA : Bivariate relationships with target (continuous variables)

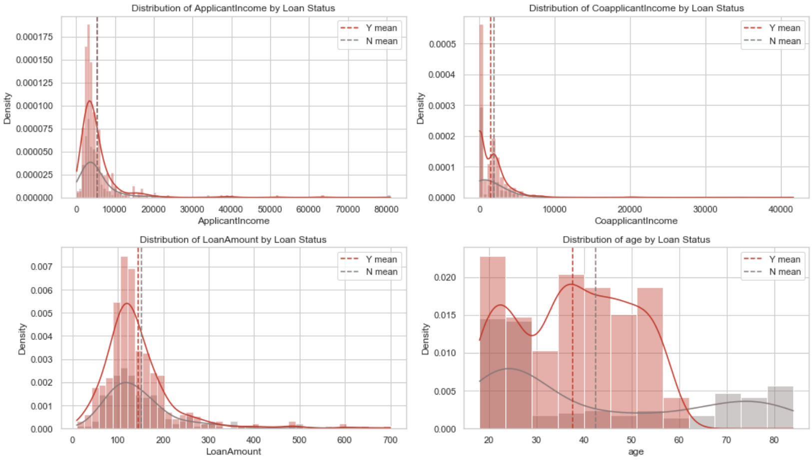


Figure: Density of numerical variables

Banque Haolim

Data source: Kaggle

EDA : Bivariate relationships with target (continuous variables)

- **LoanAmount, ApplicantIncome, and CoapplicantIncome:**
  - High dispersion, skewness, and outliers $\rightarrow$ not well described by the mean alone.
  - Refusals (N): more varied and dispersed profiles.
  - Approvals (Y): more homogeneous profiles, concentrated within a “favorable” range.
- **Age:**
  - More stable, centered around 30–40 years old.
  - Slightly more discriminant: approvals tend to cluster in this range.

Banque Haolim

Feature correlation analysis

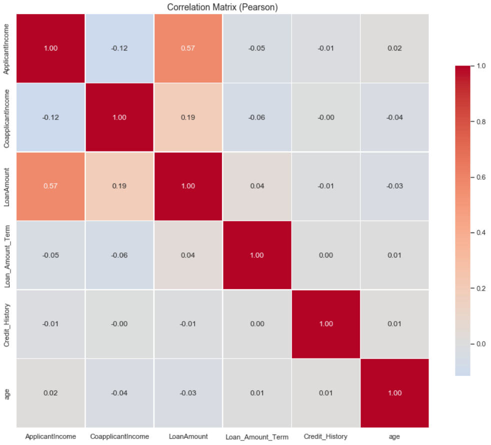


Figure: Correlation matrix

Banque Haolim

Data source: Kaggle

Data preprocessing

- Data cleaning
- Encoding Categorical Variables
- Feature Engineering

Data preprocessing : data cleaning

- Handling missing values
- Detecting and treating outliers

Data source: Kaggle

Data preprocessing : Handling missing values

- For **continuous features**, missing values were imputed using the **median**, which is more robust to outliers.
- For **categorical features**, missing **values** were filled with the **most frequent modalities** .

Banque Haolim

Data preprocessing : Detecting and treating outliers

To identify and handle outliers, we used a two-step approach:

- **Visual inspection:** Boxplots were used to visually detect potential outliers in continuous variables.
- **Statistical detection:** Z-scores were computed to quantify how far each value deviates from the mean. Observations with extreme Z-scores (typically $|Z| > 3$) were considered outliers.

Cleaning: Values identified as strongly aberrant were removed to ensure the quality of the analysis.

Banque Haolim

Data source: Kaggle

Data preprocessing : Detecting and treating outliers

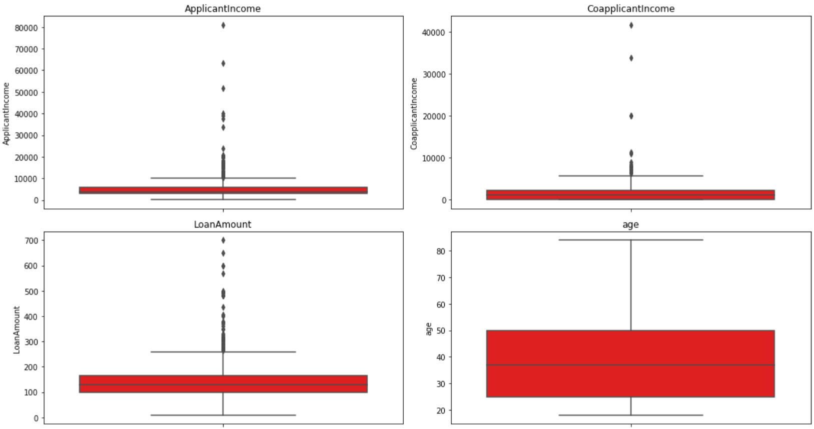


Figure: Boxplot of Numerical Variables

Banque Haolim

Data source: Kaggle

Data preprocessing : Detecting and treating outliers

The **Z-score**, or *standard score*, indicates how many standard deviations a data point is from the mean of a distribution.

Formula:

$$Z = \frac{x - \mu}{\sigma}$$

where:

- x: individual value
- μ: mean of the dataset
- σ: standard deviation

Interpretation:

- $Z = 0$: value equals the mean
- $Z > 0$: value is above the mean
- $Z < 0$: value is below the mean
- $|Z| > 3$: usually considered an outlier

Data preprocessing : Detecting and treating outliers

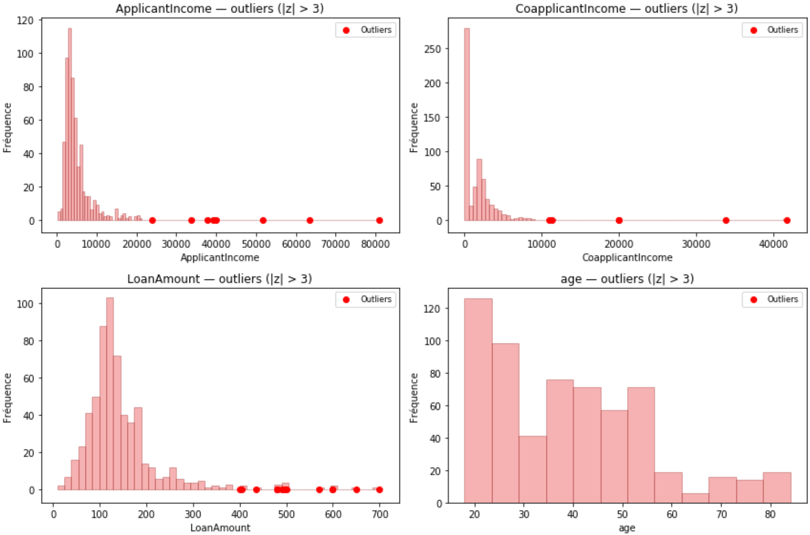


Figure: Boxplot of Numerical Var

Data source: Kaggle

Data preprocessing : Detecting and treating outliers

Based on both visual analysis using boxplots and statistical validation with the Z-test, we identified several outliers. These observations were removed to prevent them from skewing the results and affecting model accuracy.
After this step, **589 observations** remained in the dataset.

Data preprocessing : Feature Engineering

Feature engineering is the process of creating, transforming, or selecting features to improve the performance of machine learning models.

In our project, we performed the following:

Created new features:

- `TotalIncome`: Combined the applicant's and co-applicant's income to reflect total household earnings. We will discretize this continuous variable into five intervals using the `pd.cut` function. This transformation will allow for a more detailed analysis of the relationship between aggregated income levels and loan approval. The resulting variable will be named `TotalIncomeCut`.
- `MonthlyPayment`: Approximated by dividing `LoanAmount` by `Loan_Amount_Term`.
- `DebtToIncomeRatio`: Defined as `MonthlyPayment` divided by `TotalIncome`, representing the burden of the loan relative to income.

Data preprocessing : Feature engieneering

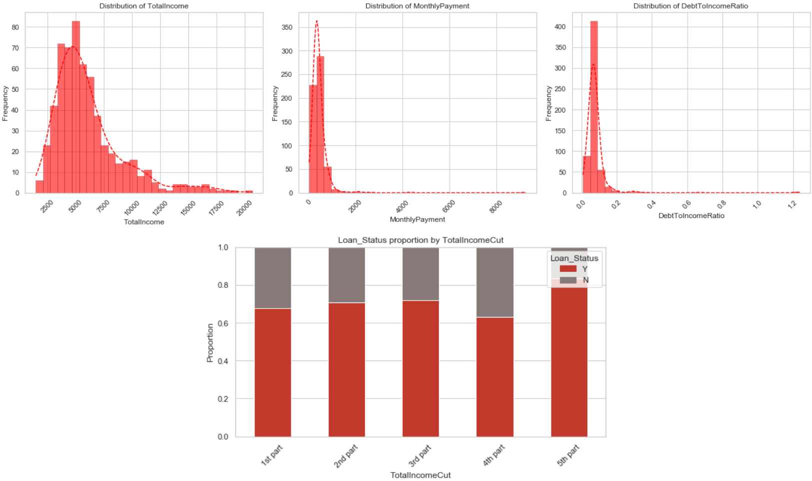


Figure: Density of newly created variables

Banque Haolim

Data source: Kaggle

Data preprocessing : Feature Engineering

- **Transformed categorical variables:** Converted categorical text fields into numerical codes to enable model processing.

These steps help the model better capture patterns and relationships within the data.

Banque Haolim

Data source: Kaggle

Data preprocessing : Feature engieneering

Encoded categorical variables as dummies: Categorical variables were transformed using one-hot encoding, creating one binary column per category. This step is essential for most machine learning models that require numerical input and cannot directly interpret text labels.

Banque Haolim

Data preprocessing : Feature engieneering

For example, the variable `Education` with categories such as `Graduate` and `Not Graduate` was converted into binary columns. One of these, `Education_Graduate`, takes the value 1 if the applicant is a graduate, and 0 otherwise. This allows the model to quantify the impact of being a graduate on the likelihood of loan approval.

This transformation also avoids introducing artificial ordinality (i.e., false numerical order) into categorical variables and ensures that the model treats each category independently.

Education_Not Graduate
0
0
0
1
0

Education
Graduate
Graduate
Graduate
Not Graduate
Graduate

Data source: Kaggle

KPI : Global Indicators

This report analyzes key performance indicators (KPIs) related to loan approval decisions using data from a Kaggle dataset. The objective is to identify which applicant characteristics and financial factors most influence whether a loan is approved.
We begin by examining general statistics related to loan approval:

Banque Haolim

KPI : Socio-Demographic Analysis

This section explores how demographic features impact loan approval:

- **Gender:** Comparison of approval rates between male and female applicants.
- **Education:** Differences between graduates and non-graduates.
- **Marital status:** Influence of family situation on loan decisions.
- **Age:** Examination of how applicant age groups impact loan approval rate.

KPI : Socio-Demographic Analysis

The **Chi-squared test of independence** is used to determine whether there is a statistically significant association between two categorical variables.

In our case, it helps answer questions such as: *Is there a relationship between Gender, Education or statut marital and Loan Approval?*

Hypotheses:

- H_0: The two variables are independent (no association).
- H_1: The two variables are dependent (there is an association).

Decision Rule:

- If the p-value is < 0.05, we reject H_0 — the variables are statistically dependent.
- If the p-value is $>= 0.05$, we fail to reject H_0 — no evidence of a relationship.

Banque Haolim

KPI : Socio-Demographic Analysis

```
=== Chi-square Test Results ===

 Variable  Chi2 Statistic  Degrees of Freedom p-value           Interpretation
   Gender           0.174                   1  0.6762          Not significant
Education           4.610                   1  0.0318  Significant (p < 0.05)
  Married           4.907                   1  0.0268  Significant (p < 0.05)
```

Based on the Chi-squared test of independence, we find that there is **no statistically significant association** between `Gender` and `Loan_Status`. In this case, $p \approx 0.68$, we **fail to reject** the null hypothesis ($\mathbf{H_0}$), meaning the two variables are considered independent.

However, both `Education` and `Marital Status` show **statistically significant associations** with `Loan_Status` (p-value < 0.05). For these variables, we **reject** the null hypothesis ($\mathbf{H_0}$), indicating a dependency with the loan approval outcome.

Banque Haolim

Data source: Kaggle

KPI : Socio-Demographic Analysis

```
T-test statistic: -2.868
P-value: 0.0045
Result: Significant difference in mean age between loan status groups.
```

For the variable `Age`, which is continuous rather than categorical, the Chi-squared test is not appropriate. Instead, since the loan status variable has only two categories (approved or not approved), a **two-sample t-test** is more suitable.

The t-test compares the means of the `Age` variable between the two loan status groups to determine if the difference in means is statistically significant.

In a t-test, the null hypothesis (H_0) states that the two group means are equal. If the p-value is less than the chosen (usually 0.05), we reject H_0, suggesting a significant difference in age between the groups. Otherwise, we fail to reject H_0, (p-value $=$ 0.0045) indicating no significant difference in mean age between approved and non-approved applicants.

KPI : Financial Situation Analysis

Here, we analyze the role of applicants' financial standing:

- **Applicant and co-applicant income:** Evaluated separately and jointly.
- **Total Income:** A combined metric to better capture repayment capacity.
- **Loan-to-Income Ratio:** Ratio between the loan amount and total income.

Data source: Kaggle

KPI : Geographical Analysis

This section looks at the impact of the applicant's location:

- **Property area:** Approval rates across Urban, Semiurban, and Rural regions.

Modeling

Before building our predictive models, we split the dataset into two distinct subsets: a training set, used to fit the models, and a test set, used to evaluate their performance on previously unseen data. This separation ensures reliable validation and helps prevent overfitting.
Then, we will test two models: logistic regression and random forest and choose the one that gives the best results.

Modeling

- Logistic regression
- Random forest voir si on en ajoute

Data source: Kaggle

Logistic regression

Logistic regression is a statistical method used for binary classification problems. It predicts the probability that a given input belongs to a certain category, typically 0 or 1.

Model Formulation

The logistic regression model estimates the probability using the sigmoid function:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}$$

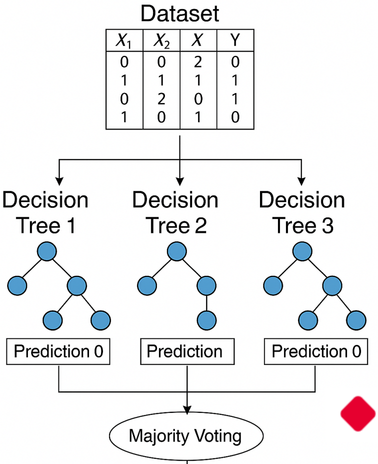
Why Logistic Regression?

- Suitable for binary outcomes (e.g., accept or reject a loan)
- Interpretable model: coefficients indicate influence on the log-odds
- Can handle multiple features

Banque Haolim

Random forest

Random forest is an ensemble learning method that builds many decision trees and combines their results to make more accurate and stable predictions. Each tree is trained on a random subset of the data and variables, which helps reduce overfitting and improve generalization.



Data source: Kaggle

Model interpretation and validation

on regarde le score de chque modele et on valide le modele dont le score est le meilleur

Deployment and prediction via API

on fait une API en récupérant le modele choisis et voir si le client peut vérifier en fonction de son profil si son pret est accepte

Conclusion

On conclue mais à mmon avis ici les bases de données sont trop petite donc on pourra parler de ca

Banque Haolim

Data source: Kaggle