REPORT

Creditworthiness Analysis

- Submitted by Dhinsha . T

Creditworthiness Analysis

Introduction

Creditworthiness is a lender's willingness to trust you to pay your debts. A borrower deemed creditworthy is one a lender considers willing, able and responsible enough to make loan payments as agreed until a loan is repaid.Lenders evaluate creditworthiness in a variety of ways, typically by reviewing your past handling of credit and debt, and, in many cases, by assessing your ability to afford the payments required to repay the debt.

This project aims to identify the factors that manage a person's credit worthiness. Whether the person is under the high risk category or not.

Problem statement

A person's creditworthiness is often associated (conversely) with the likelihood they may default on loans.

Based on the applicant details and loan history we are finding whether they were considered high risk.

0 = Low credit risk i.e high chance of paying back the loan amount

1 = High credit risk i.e low chance of paying back the loan amount

Also focusing on the following:

- How is a person's creditworthiness can be found,

Would a person with critical credit history be more creditworthy? Are young people more creditworthy? Would a person with more credit accounts be more creditworthy?

- To analyse which are the crucial factors which should be evaluated by the lenders in order to provide loans or credits.

Dataset

Basically two datasets are there, One is loan applications and other one is applicant details.

My plan of action was to analyse important variables such as high_risk_application, Months for loan taken, Purpose, amount, Loan_history, Employment_status, applicant age, Number_of_existing_loans_at_this_bank, account balance etc

Data Description :

applicant.csv: This file contains personal data about the (primary) applicant

- Unique ID: applicant_id (string)
- Other fields:
    - Primary_applicant_age_in_years (numeric)
    - Gender (string)
    - Marital_status (string)
    - Number_of_dependents (numeric)
    - Housing (string)
    - Years_at_current_residence (numeric)
    - Employment_status (string)
    - Has_been_employed_for_at_least (string)
    - Has_been_employed_for_at_most (string)
    - Telephone (string)

- Foreign_worker (numeric)
- Savings_account_balance (string)
- Balance_in_existing_bank_account_(lower_limit_of_bucket) (string)
- Balance_in_existing_bank_account_(upper_limit_of_bucket) (string)

loan.csv: This file contains data more specific to the loan application

- Target: high_risk_application (numeric)
- Other fields:
  - applicant_id (string)
  - Months_loan_taken_for (numeric)
  - Purpose (string)
  - Principal_loan_amount (numeric)
  - EMI_rate_in_percentage_of_disposable_income (numeric)
  - Property (string)
  - Has_coapplicant (numeric)
  - Has_guarantor (numeric)
  - Other_EMI_plans (string)
  - Number_of_existing_loans_at_this_bank (numeric)
  - Loan_history (string)

Methodology

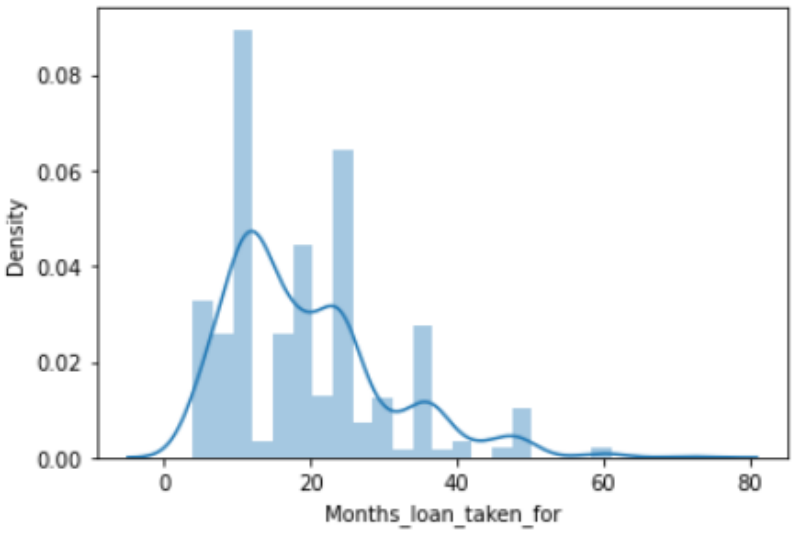
1. **Combining the Dataset**
   - We have given two datasets. One is loan applications and the other one is applicant details.
   - First of all, I merged those dataset into one data using the primary key ‘applicant_id’.
   - It contains a total of 1000 rows and 27 columns.

2. **Data preprocessing (** Handling missing values **)**
    - Data preprocessing includes cleaning and data integration, which was done in python.
    - Missing values were checked. One way of handling missing values is the deletion of the rows or columns having null values. If any columns have more than half of the values as null then you can drop the entire column. In the same way, rows can also be dropped if having one or more columns values as null.
    - its found that the columns 'Telephone', 'Other_EMI_plans', 'Balance_in_existing_bank_account_(lower_limit_of_bucket)', 'Balance_in_existing_bank_account_(upper_limit_of_bucket)', 'Has_been_employed_for_at_most' has many null values. So I dropped those columns.
    - And for the important variables which contain some null values, I dropped only those rows.
    - Checked for duplicate values.

3. **EDA**
    - Exploratory data analysis was done in Python itself to summarize the major characteristics of the dataset and to analyse the hidden trends.
    - Data visualisation was performed thus insights are drawn.
    - Few graphical representations that were plotted are as follows:

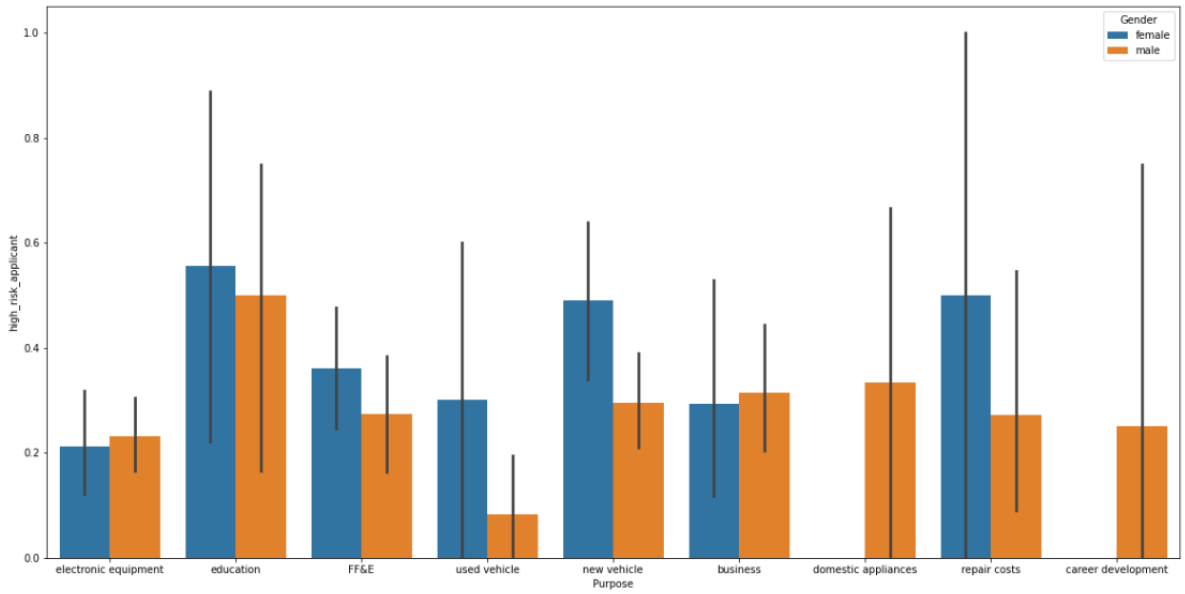
```
In [206]: sns.distplot(df1['Months_loan_taken_for'])
```

```
Out[206]: <AxesSubplot:xlabel='Months_loan_taken_for', ylabel='Density'>
```



- Majority of loans are taken for a period between 10 to 30 months

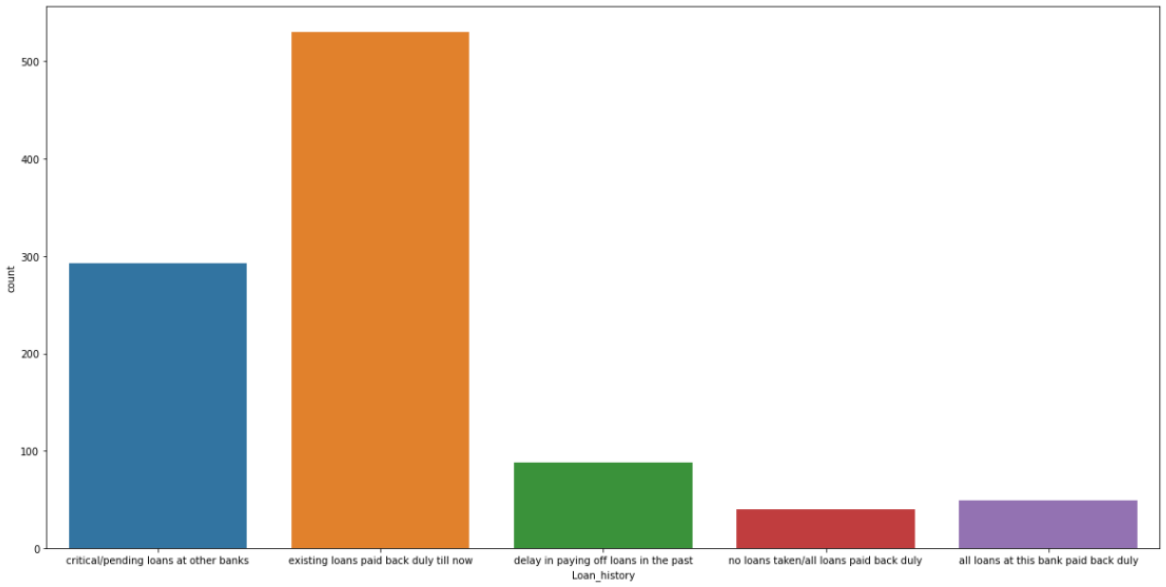
```
plt.figure(figsize=(20,10))
sns.barplot(df1['Purpose'], df1['high_risk_applicant'], hue = df1['Gender'])
plt.show()
```



- The loans taken for education, repair cost and new vehicles are high. Which are under high risk categories also.

```
fig = plt.figure(figsize =(20, 10))
sns.countplot(df1['Loan_history'])
```

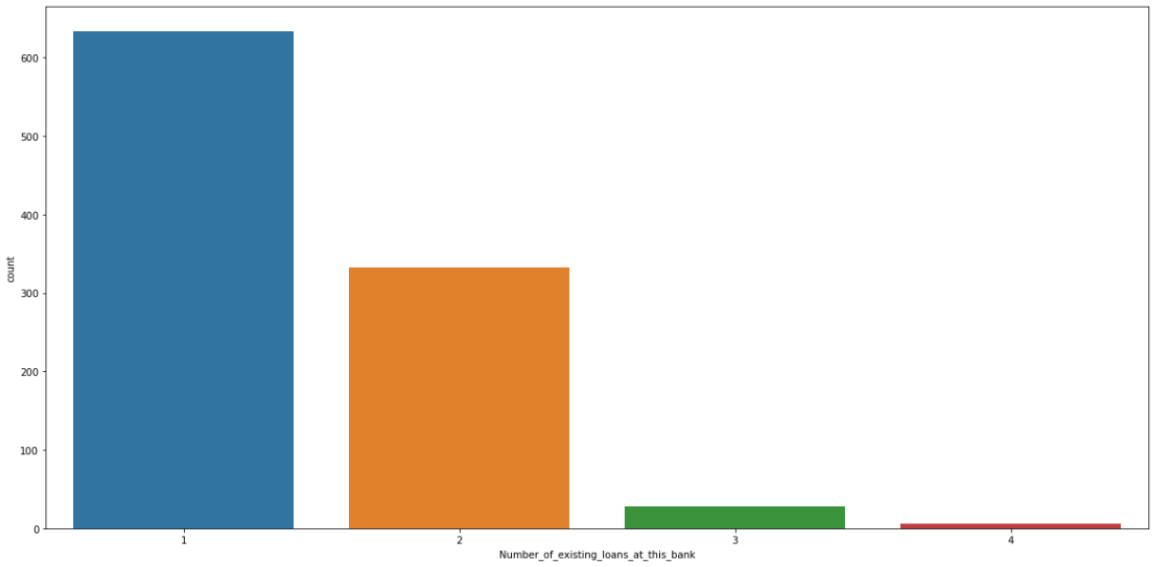
```
<AxesSubplot:xlabel='Loan_history', ylabel='count'>
```



Its found that 2nd most applicants have pending loans in other banks so is high risk.

```
fig = plt.figure(figsize =(20, 10))
sns.countplot(df1['Number_of_existing_loans_at_this_bank'])
```

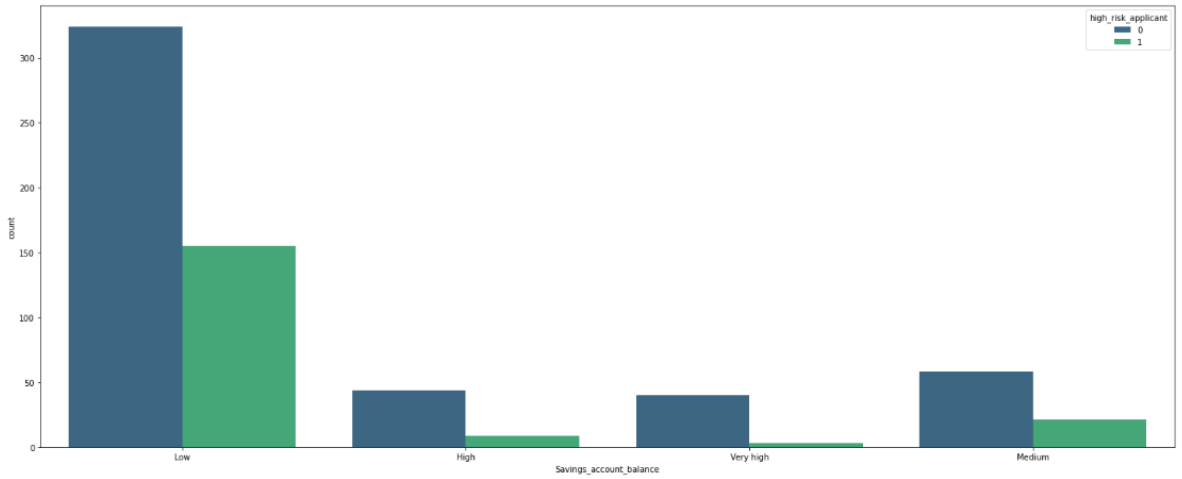
```
<AxesSubplot:xlabel='Number_of_existing_loans_at_this_bank', ylabel='count'>
```



When the number of existing loans increases risk also increases

```
plt.figure(figsize=(25,10))
sns.countplot(x=df1["Savings_account_balance"],hue=df1["high_risk_applicant"], palette="viridis")
```

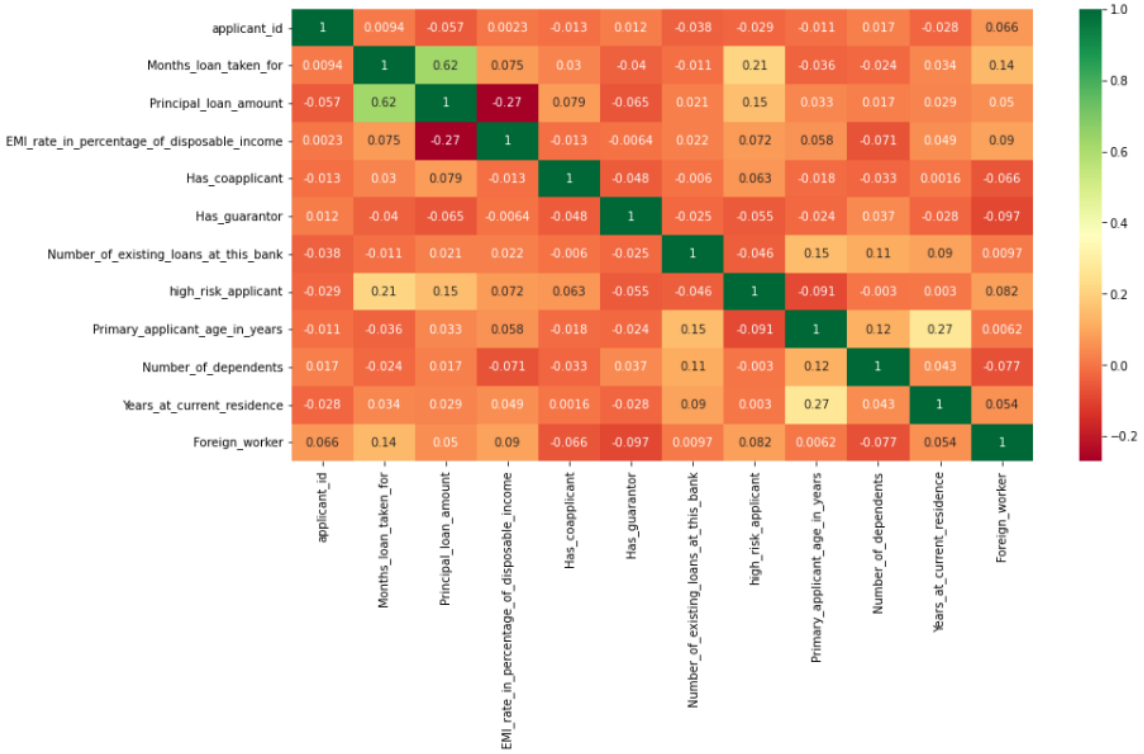
```
<AxesSubplot:xlabel='Savings_account_balance', ylabel='count'>
```



- Majority of the applicants have a low savings account balance.

```
plt.figure(figsize=(14,7))
sns.heatmap(data.corr(), annot = True, cmap = 'RdYlGn')
```

```
<AxesSubplot:>
```



- Heat map used to visualize the strength of correlation among variables. It helps find features that are best for Machine Learning model building.

Here we can see only the **principal amount** and **months of loan taken** is correlated. No other variables which have high correlation.

4. **Model building**
    - To machine learning models were built
    - First of all, the dataset was converted into integer format from object data types.
    - Splitted the data into training and testing set
    - We have used machine learning models to find the accuracy.
    - **Random forest model** and **decision tree models** were built and trained, the model and accuracy were found.

Insights

- Majority of loans are taken for a period between 10 to 30 months even though loans for more than 50 months can be considered as a high risk category.
- The loans taken for electronic equipment and new vehicles are most in number. But by checking with high risk applicants, loans taken for education, repair cost and new vehicles are high. Which seems to be a comparatively high risk category.
- Instead if it's for asset based lending then it will be under less risk category.
- It's found that 2nd most applicants have pending loans in other banks so is high risk.
- When the number of existing loans increases, risk also increases. Whether loans taken from the existing bank as well as from other banks.
- Majority applicants have less savings account balance. So it's a risk factor.

- Most applicants have no guarantor at all, its comparatively high risk factor.
- Majority applicants who are foreign workers also fall under the high risk category.
- Random Forest Classifier algorithm able to achieve the F1-score of 0.78 (78%) which is better than decision tree model (68%)