Loan Prediction Based on Customer Behaviour

Prepared for: Dr. Khaled Mohammed Fouad

Prepared by:

1.Mazen Mohamed 202000894

2.Youssef Ismail Mohamed 202001864

3.Ahmed Khaled Kamal 202002558

Contents

1. Introduction

Loan prediction is an essential task in the banking and financial sector. Lending institutions use a variety of information to evaluate a loan application's potential risk and decide whether to approve or reject it. Traditionally, banks have used credit scores, employment history, and income level to determine whether a borrower is likely to repay the loan. However, these factors alone may not be sufficient to assess a borrower's creditworthiness accurately. One approach to improving the accuracy of loan predictions is to analyze customer activity. By examining a borrower's transactional history, banks and Lending institutions can gain insights into their spending habits, financial stability, and repayment behavior.

The problem we aim to address is to construct a predictive model that can accurately estimate the probability of a borrower defaulting on loan payments, taking into account their transactional history and past record of loan defaults. This is a well-defined and specific problem that can assist lending institutions in making more informed and accurate lending decisions.

The importance of this problem cannot be overstated. Incorrect loan decisions can lead to financial losses for the lending institution and affect the borrower's credit score. Furthermore, it can lead to a loss of trust in the financial system, which can have long-term economic consequences.

The impact of this project could be significant in several ways. Firstly, it can help lending institutions make more informed lending decisions, leading to reduced risk and higher profitability. Secondly, it can improve access to credit for borrowers who may not have a strong credit score but have a reliable transactional history. Finally, it can contribute to the overall stability of the financial system by reducing the likelihood of loan defaults.

In the following sections, the dataset will be described, preprocessed, and explored via EDA to gain insights from the data. Finally, the chosen model will be utilized to make loan predictions, and its performance will be evaluated using various evaluation metrics.

2. Data Description

In this section, a detailed description of the dataset used for loan prediction analysis is provided. The dataset's source, size, format, data type, missing values, outliers, and data quality are described to help readers understand the nature and quality of the data.

2.1 Data source:

The dataset used in this analysis was obtained Hackathon organized by **Univ.AI**. The dataset is publicly available and was sourced from Kaggle, a popular platform for data science competitions. The dataset documentation includes information on the variables and their definitions.

2.2 Size and format:

The dataset contains 264600 rows and 13 columns. The data is provided in CSV format and includes several variables such as ID, Income, Age, Experience, Married/Single, House_Ownership, Car_Ownership, Profession, CITY STATE, CURRENT_JOB_YRS, CURRENT_HOUSE_YRS, and Risk_Flag.

```
In [5]: # shape
        df.shape

Out[5]: (264600, 13)
```

2.3 Data types:

The dataset includes both categorical and numerical data. The categorical variables include Married/Single, House_Ownership, Car_Ownership, Profession, CITY STATE, while the numerical data includes Income, Age, Experience, CURRENT_JOB_YRS, CURRENT_HOUSE_YRS, and Risk_Flag.

```
In [6]: # data types
        df.dtypes

Out[6]: Id                 int64
        Income           float64
        Age                int64
        Experience         int64
        Married/Single    object
        House_Ownership   object
        Car_Ownership     object
        Profession        object
        CITY              object
        STATE             object
        CURRENT_JOB_YRS    int64
        CURRENT_HOUSE_YRS  float64
        Risk_Flag          int64
        dtype: object
```

2.4 Missing values:

The dataset contains some missing values, with 10.52 % of the rows having at least one missing value. To handle these missing values, The mean was used to fill the missing as the data doesn't contain any outliers.

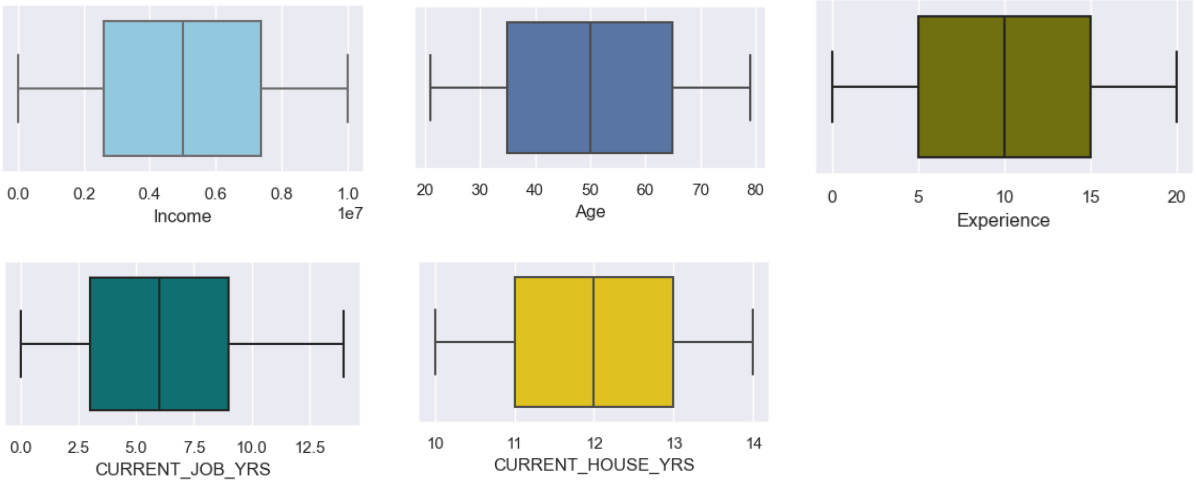
In [7]:
```
# check missing values
# percentage of missing values in each column
total_missing = df.isnull().sum().sort_values(ascending = False)
missing_percent = (total_missing / len(df)) * 100
table = pd.concat([total_missing,missing_percent], axis=1, keys=['Total','Percentage %'])
table
```

Out[7]:

	Total	Percentage %
Income	10623	4.014739
Married/Single	7692	2.907029
CURRENT_HOUSE_YRS	5326	2.012850
Car_Ownership	5314	2.008314

2.5 Outliers:

The data doesn't contain any outliers using box plots and IQR.



In [9]:
```
# check outliers
def outliers(df):
    #  first and third quartiles
    Q1 = df.quantile(0.25)
    Q3 = df.quantile(0.75)

    # IQR
    IQR = Q3 - Q1

    out = df[((df<(Q1-1.5*IQR)) | (df>(Q3+1.5*IQR)))]

    return out
```

Out[10]:

Features	Total Outliers
Income	0
Age	0
Experience	0
CURRENT_JOB_YRS	0
CURRENT_HOUSE_YRS	0

2.6 Data quality:

The dataset was found to be of high quality as the data pre-processing stage revealed no data entry errors. Although there were some duplicate rows, they were removed and the remaining data was considered suitable for analysis.

```
In [8]: # Check Duplicates
        total_dupliactes = df.duplicated().sum()
        dupliactes_percent = (total_dupliactes / len(df)) * 100
        table = pd.DataFrame({'Total':[total_dupliactes],'Percentage %':[dupliactes_percent]})
        table
```

Out[8]:

	Total	Percentage %
0	12600	4.761905

3. Data Cleaning and Pre-processing

In this section, the steps taken to clean and pre-process the loan prediction dataset before conducting any analysis were described. The data cleaning process involved the handling of missing values, duplicates, Data Validation, and the selection of a representative subset of the data for modelling purposes. The purpose of this process was to ensure the accuracy and integrity of the data and to establish the credibility of the results.

3.1 Handling Missing Values:

The dataset contained some variables, namely ‘Income’, ‘CURRENT_HOUSE_YRS’, ‘Married/Single’, and, ‘Car_Ownership’, that had missing values around 10.52% of the rows had missing data. To address this issue, certain techniques were employed to handle these missing values:

- **Income and CURRENT_HOUSE_YRS:** The missing values in previous numerical variables were imputed with the mean of the columns as they don’t have any outliers the mean seemed to be a good choice for inferring the missing values.
- **Married/Single and Car_Ownership:** The missing values in previous categorical variables were imputed with the mode of the columns.

3.2 Handling Duplicates:

The dataset contained some duplicated observations, which are rows with identical values across all columns. To handle these duplicates, a technique called deduplication or de-duping could have been used. This involves identifying and removing any duplicated rows to ensure that each observation in the dataset is unique.

3.3 Data Validation:

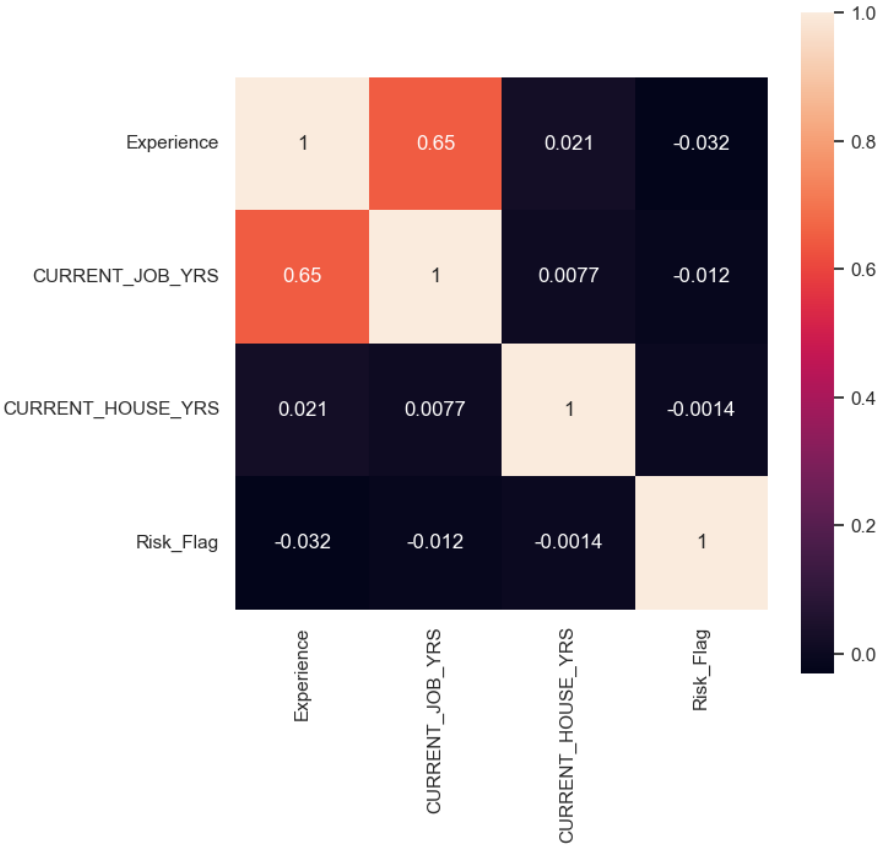
To maintain the reliability and correctness of the dataset, checks were performed to identify any errors or inconsistencies that could compromise the data quality. Some of the issues we looked out for included **inconsistent data types**, **unexpected values**, and **missing/duplicated entries**. Additionally, we ensured that the dataset was complete, meaning all the necessary variables were present and properly labelled to avoid any confusion or misunderstandings during analysis.

3.4 Data Sampling:

A stratified random sampling technique was employed to select a subset of the data for modelling purposes (55%) of the data. The subset was chosen in such a way that the proportion of loan approvals and rejections in the original dataset was maintained, thereby ensuring that the resulting model would be representative of the entire population. The sample size was selected with the aim of improving the efficiency of the model as more data means better performance.

3.5 Drop High correlated Feature:

Having highly correlated features can lead to multicollinearity, which can result in unreliable and unstable coefficient estimates in predictive models. To address this issue, one common approach is to drop one of the highly correlated features from the analysis. Since a correlation greater than 0.7 is generally considered high, the moderate correlation observed between Experience and CURRENT_JOB_YRS suggests that the two variables are positively related but not strongly so, dropping one of them might help the model for better prediction.



3.6 Feature Encoding:

Feature encoding was used to transform categorical variables such as Car_Ownership, House_Ownership, and many other into numerical values that could be used as input for the logistic regression and decision tree models label encoding was used as the categorical variables in the data have many classes.

3.7 Transformation:

In the loan prediction project, min-max scaling was used to transform numerical variables such as income and loan amount into a common range. Min-max scaling is particularly useful when there are significant differences in the scales of the numerical variables in the dataset. By scaling the variables to a common range, min-max scaling ensures that each variable contributes equally to the analysis and prevents the dominance of variables with larger values.

4. Exploratory Data Analysis (EDA)

In this section, a variety of descriptive statistics, visualizations, and correlation analyses are performed to better understand the data and identify any outliers or missing values that need to be handled as well as patterns.

4.1 Descriptive Statistics:

The loan prediction dataset includes several numerical variables, such as income, age, experience, current job years, and current house years. Descriptive statistics were calculated for these variables to provide an overview of their main characteristics.

Out[20]:

	Income	Age	Experience	CURRENT_JOB_YRS	CURRENT_HOUSE_YRS
count	1.386000e+05	138600.000000	138600.000000	138600.000000	138600.000000
mean	5.000360e+06	49.956775	10.081883	6.337330	11.996681
std	2.819461e+06	17.069516	6.001930	3.649305	1.384621
min	1.031000e+04	21.000000	0.000000	0.000000	10.000000
25%	2.596508e+06	35.000000	5.000000	3.000000	11.000000
50%	4.996003e+06	50.000000	10.000000	6.000000	12.000000
75%	7.362882e+06	65.000000	15.000000	9.000000	13.000000
max	9.999938e+06	79.000000	20.000000	14.000000	14.000000

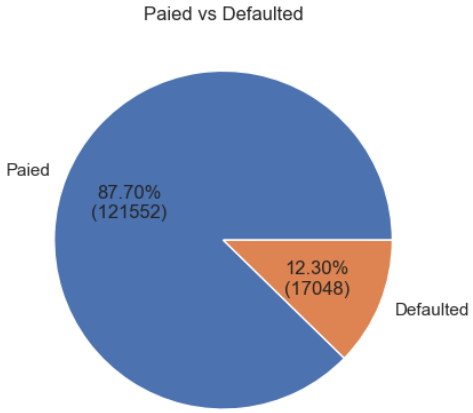
The mean **income** of loan applicants was approximately 5 million, with a standard deviation of 2.8 million, indicating a wide range of income levels. The minimum income recorded in the dataset was 10,310, while the maximum was nearly 10 million. The mean age of loan applicants was around 50 years, with a standard deviation of 17 years. The minimum age recorded in the dataset was 21, while the maximum was 79.

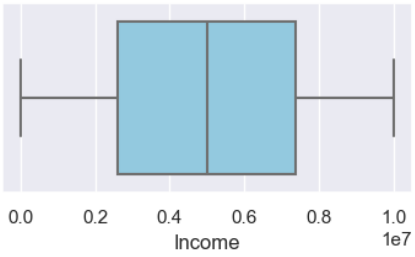
The mean **experience** of loan applicants was approximately 10 years, with a standard deviation of 6 years. The minimum experience recorded in the dataset was 0 years, while the maximum was 20 years. The mean **current job years** of loan applicants was approximately 6 years, with a standard deviation of 3.6 years. The minimum current job years recorded in the dataset was 0 years, while the maximum was 14 years.

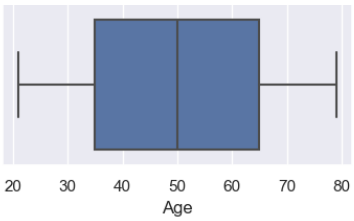
Finally, the mean **current house years** of loan applicants was nearly 12 years, with a standard deviation of 1.4 years. The minimum current house years recorded in the dataset was 10 years, while the maximum was 14 years.

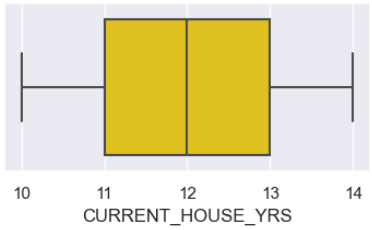
4.2 Visualization:

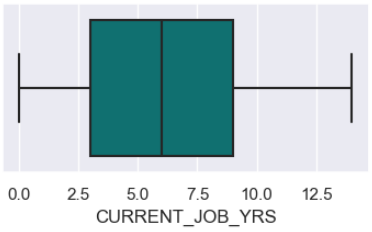
4.1 Univariate Analysis

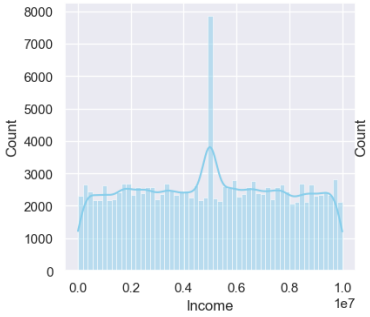


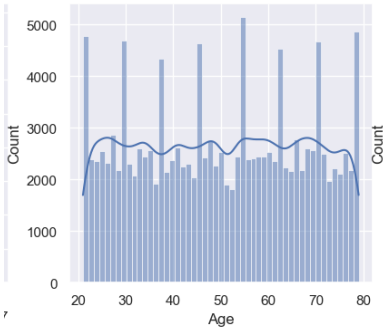


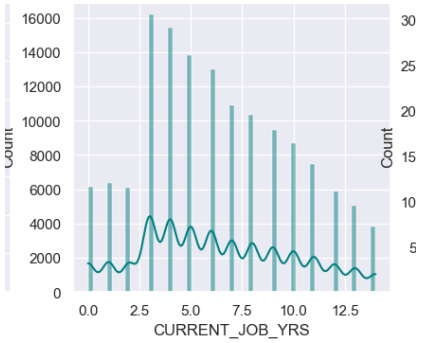


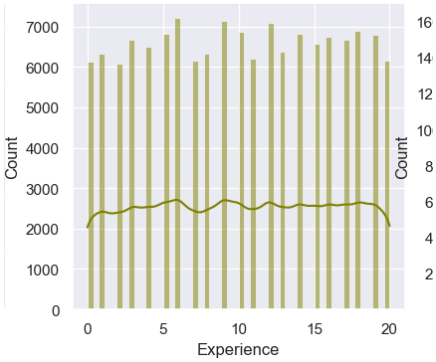


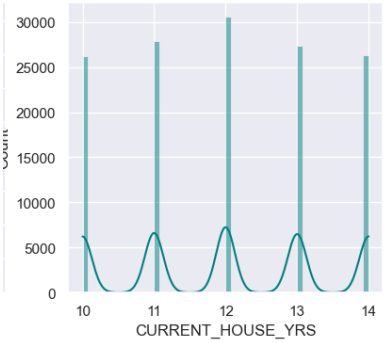








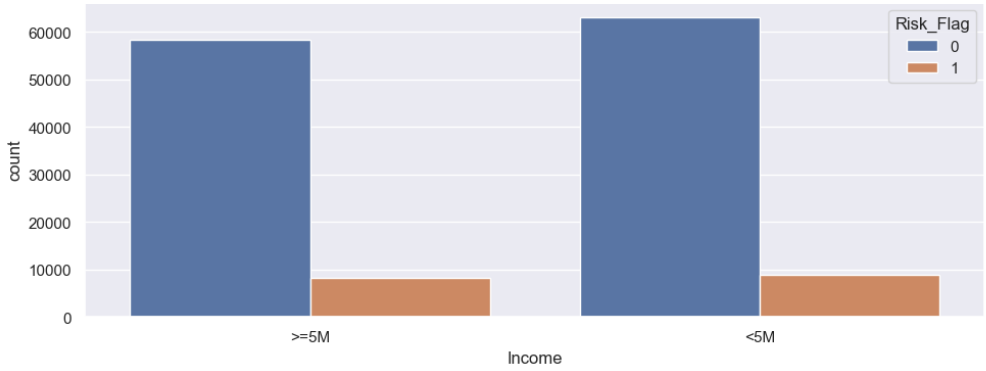


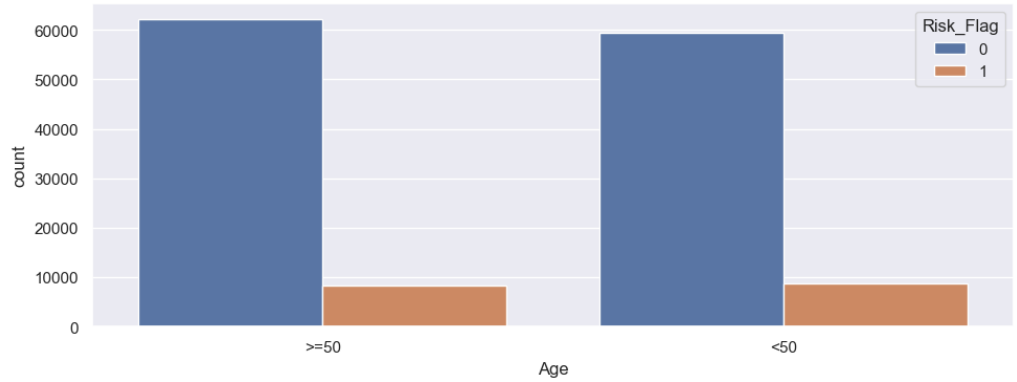


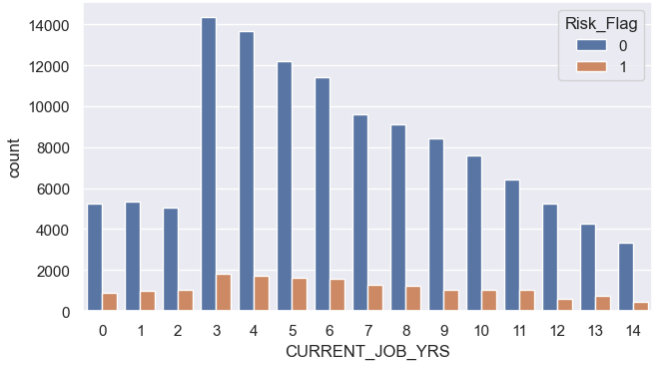
Key findings:

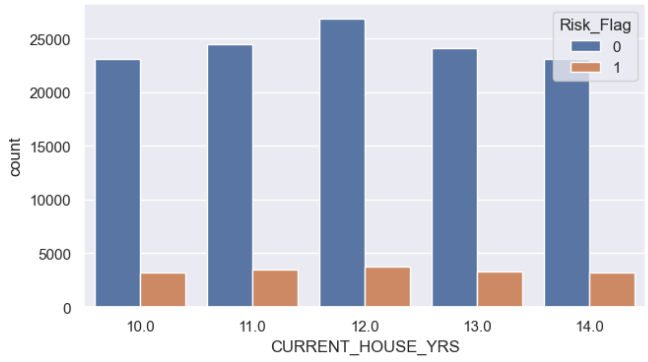
- Most of the applicates haven't been defaulted before as only 12.3% of the have been defaulted.
- There are no outliers.
- The distribution of the data is approximately symmetric there is no need for transformation.

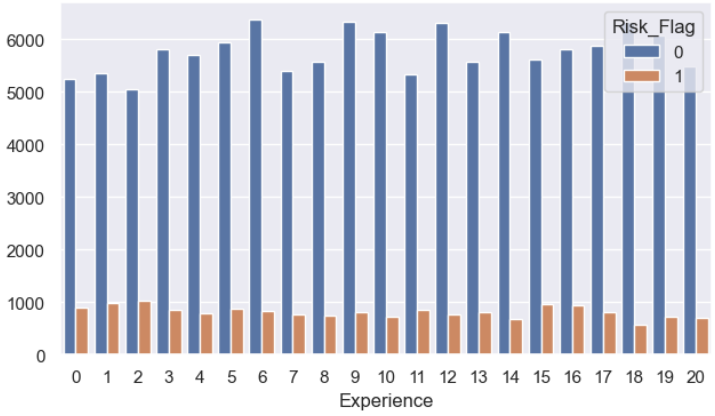
4.2 Bivariate Analysis

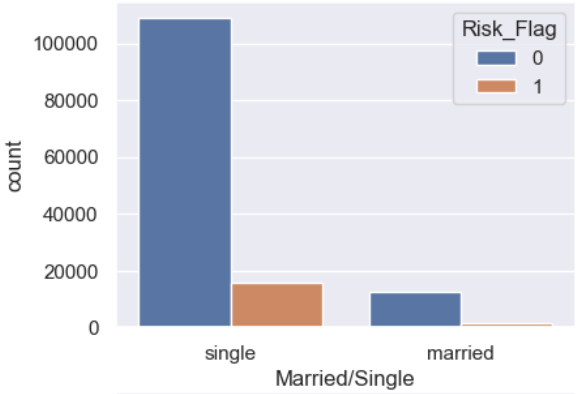


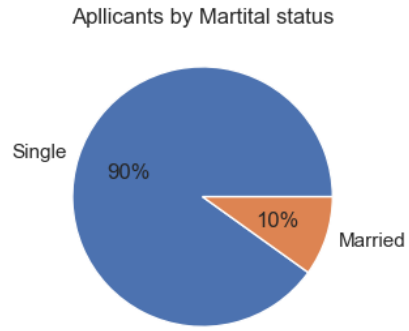


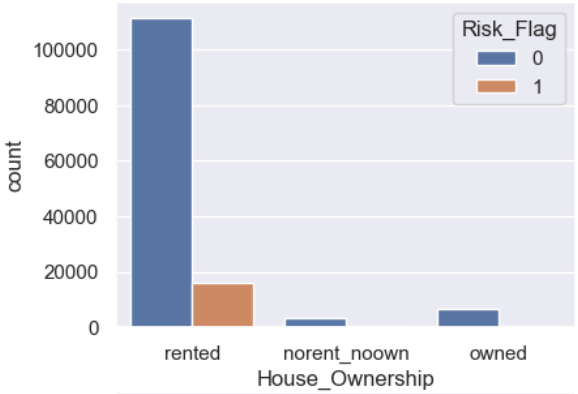


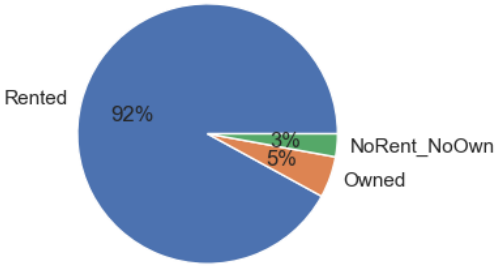


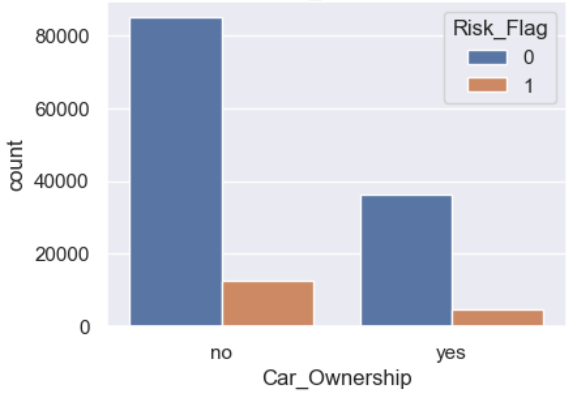


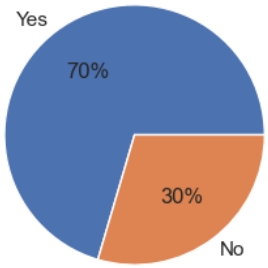












Out[31]:

Risk_Flag	0	1	All
Profession			
All	221004	30996	252000
Police_officer	4209	826	5035
Software_Developer	4303	750	5053
Air_traffic_controller	4566	715	5281
Surveyor	4000	714	4714

Out[32]:

Risk_Flag	0	1	All
Profession			
All	221004	30996	252000
Physician	5247	710	5957
Statistician	5135	671	5806
Web_designer	4808	589	5397
Drafter	4754	605	5359

Out[35]:

Risk_Flag	0	1	All
STATE			
All	221004	30996	252000
Uttar_Pradesh	25057	3343	28400
West_Bengal	20474	3009	23483
Andhra_Pradesh	22362	2935	25297
Maharashtra	22667	2895	25562

Out[36]:

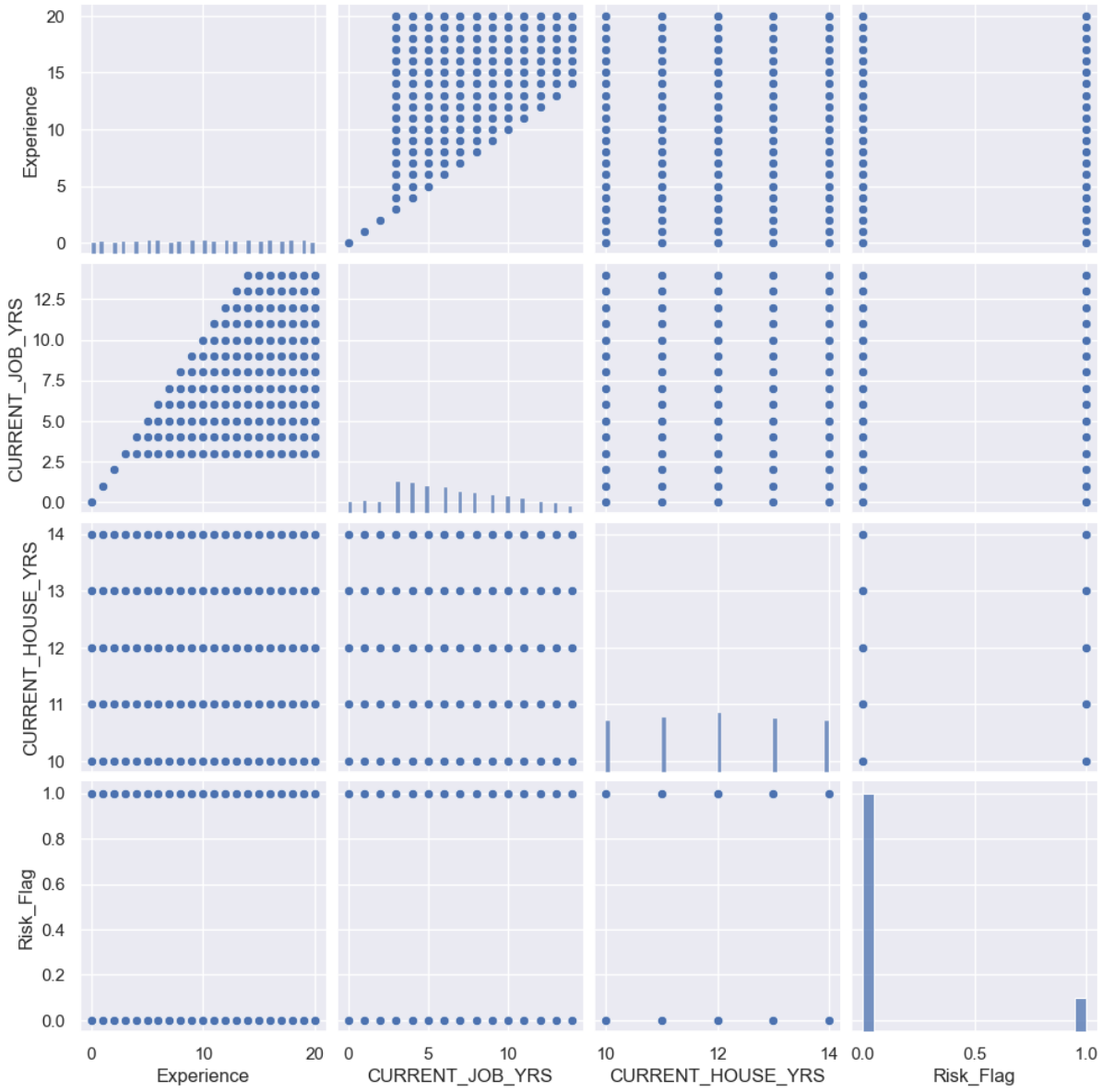
Risk_Flag	0	1	All
STATE			
All	221004	30996	252000
Uttar_Pradesh	25057	3343	28400
Maharashtra	22667	2895	25562
Andhra_Pradesh	22362	2935	25297
West_Bengal	20474	3009	23483

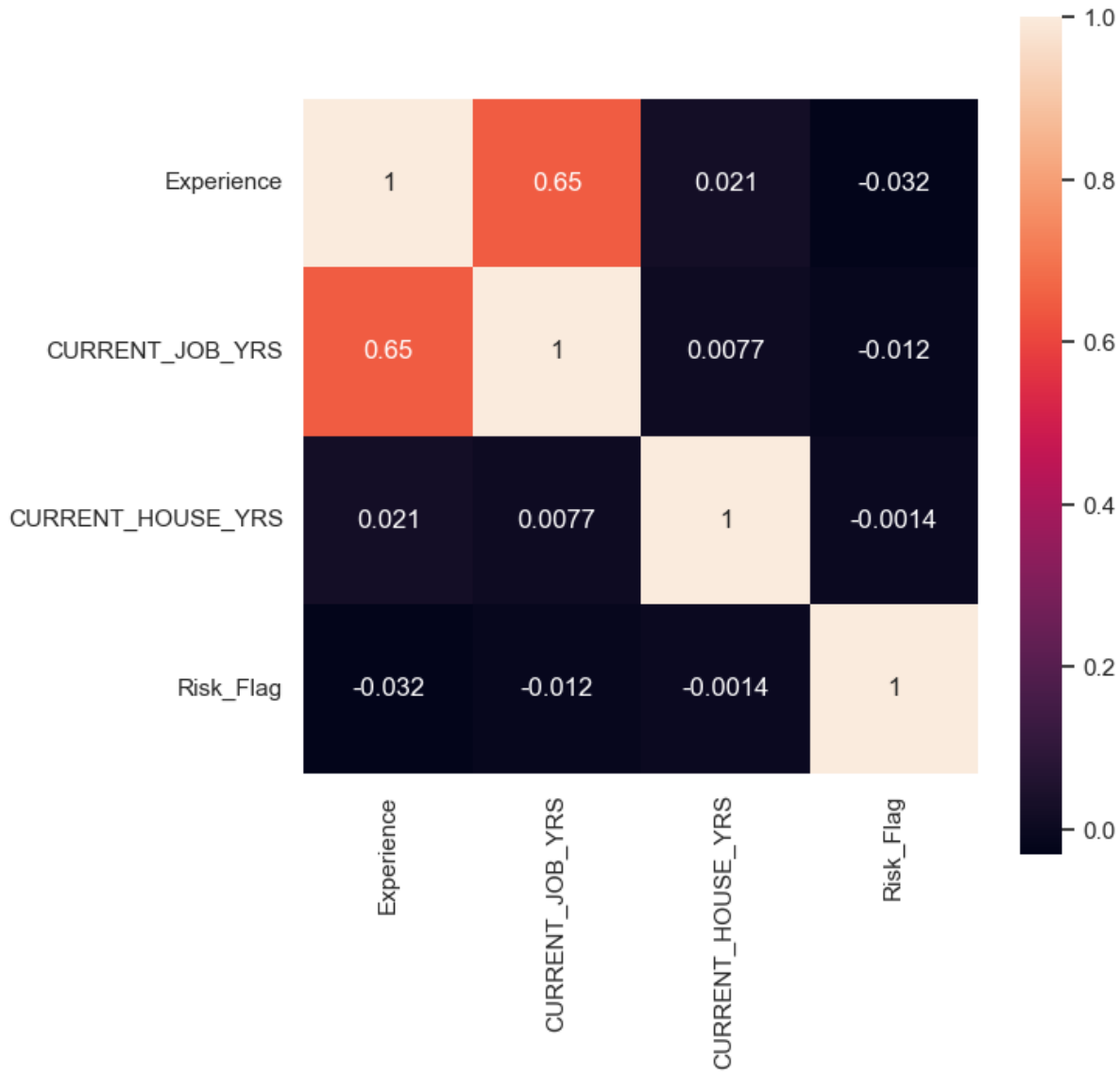
Out[38]:

Risk_Flag	0	1	All
CITY			
All	221004	30996	252000
Vijayanagaram	1110	149	1259
Saharsa[29]	1067	113	1180
Bulandshahr	1061	124	1185
Bhopal	1042	166	1208

Out[39]:

Risk_Flag	0	1	All
CITY			
All	221004	30996	252000
Kochi	718	243	961
Satna	796	232	1028
Buxar[37]	757	228	985
Srinagar	920	216	1136


Experience
CURRENT_JOB_YRS
CURRENT_HOUSE_YRS
Risk_Flag
0
5
10
15
20
0.0
2.5
5.0
7.5
10.0
12.5
10
11
12
13
14
0.2
0.4
0.6
0.8
1.0
0.5



Key findings:

- Applicants above the avg Age and Income have slightly lower risk of being defaulted compared to Applicants below the avg.

- Applicates who spend more years in their jobs are less likely to default.
- Applicates who are less experienced are more likely to default.
- Applicates w Single customers ask more for loans and at higher risk to default.
- Customers who rent ask more for loans and at higher risk to default.
- Non-car owner customers ask less for loans but at higher risk to default.
- ho live 10 yrs. or more in the same house have low level of loan default.

- There are 51 Profession.

- Physicians are the most to get loans.
- Police officers are the most to default.

- The applicants come from 29 state.
- customers of **Uttar Pradesh** are the most to take loans and default.

- There are 317 the applicants live in.
- Customers of Kochi city the most to default loans
- Customers of Vijayanagar am city the most to ask for loans
- No clear relationships between the variables.

- Experience and CURRENT_JOB_YRS suggests that the two variables are positively related but not strongly so, dropping one of them might help the model for better prediction.

Modelling:

The modelling section presents the results of the two models that were developed to analyze the loan prediction dataset: **logistic regression and decision tree**. The logistic regression model was trained using a stratified random sampling technique to maintain the proportion of loan defaulting or not in the original dataset. The model predicted the likelihood of a loan applicant defaulting on their loan repayment based on their previous history of defaulting or not, and its performance was evaluated using metrics such as accuracy. The results of the logistic regression model showed an **accuracy of 52.32 %**, indicating that the model wasn't able to accurately predict likelihood of a loan applicant defaulting on their loan repayment for the most of loan applicants.

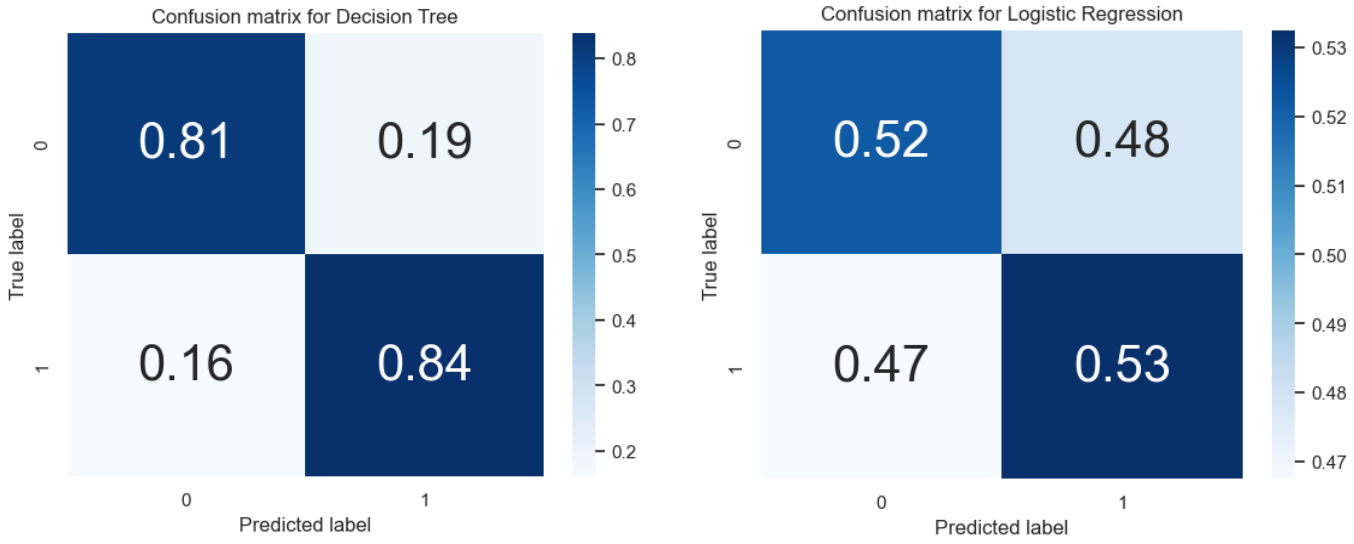
Similarly, the decision tree model was trained on the same subset of data as the logistic regression model and evaluated using the same performance metrics. The decision tree model predicted the likelihood of a loan applicant defaulting on their loan repayment based on their previous history of defaulting or not, and its performance was higher than the logistic regression model with an accuracy of 81.46%.

The potential source bias in the analysis due to an imbalance in the distribution of the target variable, where only a small percentage (12.3%) of applicants are classified as having a "Risk_Flag" (defaulted), while the majority (87.7%) are classified as not having a risk flag and the accuracy of the previous models were achieved after applying weighted loss functions and PCA for **logistic regression** which did not result in a significant improvement in the model's performance and applying PCA and Under sampling for **decision tree** which it's not be the best result for this dataset, but accepted and outperforms the Logistic Regression Model.

Results:

After applying **PCA and weighted loss functions** the results of the logistic regression model showed an **accuracy of 52.32 %**, indicating that the model wasn't able to accurately predict likelihood of a loan applicant defaulting on their loan repayment for the most of loan applicants and applying PCA and Under sampling for **decision tree** the performance was higher than the logistic regression model with an accuracy of 81.46%. The **decision tree** model was able to accurately predict likelihood of a loan applicant defaulting on their loan repayment for the most of loan applicants.

The confusion matrix for the logistic regression model revealed that the model had a relatively high false positive/negative rate, predicting 52% of non-defaulting applicants as defaulters. Additionally, the model predicted only 53% of the applicants who were actually likely to default. These findings suggest that the model had poor performance in accurately identifying loan applicants who were likely to default. On the other side, the **decision tree** model revealed that the model had a relatively low false positive/negative rate, predicting 81% of non-defaulting applicants as defaulters and 84% of the applicants who were actually likely to default.



Overall, the **decision tree models** showed promising results in predicting the likelihood of loan defaulting over **logistic regression**. However, the models have limitations and assumptions that need to be taken into consideration when interpreting the results. Further work could be done to improve the models by incorporating additional variables or using more advanced techniques such as ensemble modelling. Additionally, the model could be evaluated on a larger and more diverse dataset to ensure their generalizability.

Conclusion:

In conclusion, the loan prediction project aimed to develop and evaluate models to predict the likelihood of loan defaulting based on various factors. The logistic regression and decision tree models were developed and evaluated using techniques such as confusion matrices, and accuracy measures.

The key findings of the project were that risk flag was the most important factor in predicting loan defaulting status. The logistic regression model showed an accuracy of 52.32%, while the decision tree model showed an accuracy of 81.46%. However, both models were not able to accurately identify all loan applicants that are likely to default.

Potential areas for future work include incorporating additional variables into the models, improving data quality and completeness, and evaluating the models on larger and more diverse datasets. These improvements could help to increase the accuracy and reliability of loan prediction models, and ultimately improve the decision-making process for loan approvals.