PROJECT DOCUMENTATION

Exploratory Data Analysis

Loan Dataset

Name : Mohan Prabhu D

Branch : Data Science

Batch Number : 1

Mode of Study : Offline

Roll Number : 150324CBR003

Table of Contents

1.Introduction

The purpose of this project is to perform an Exploratory Data Analysis (EDA) on loan Dataset to uncover underlying patterns, relationships, and insights. A loan dataset typically contains information about their features such as loan_id, no_of_dependents, education, self_employed, income_annum, loan_amount, loan_term, cibil_score, residential_assets_value, commercial_assets_value, luxury_assets_value, bank_asset_value, loan_status. Such datasets are widely used in banking analysis for machine learning applications for predictive modelling of loan approval.

2.Aim

The goal of this project is to conduct a comprehensive analysis of the dataset to derive insights and it can be utilized to address various business problems in the banking industry in loan approval.

3. Business Problem / Problem Statement

In the banking industry, loan approval is a critical process that involves assessing the risk associated with lending money to applicants. The primary business problem revolves around effectively identifying which loan applications should be approved to minimize the risk of defaults while maximizing profits.

4. Project Workflow

The project workflow includes the following steps:

1. Data Collection
2. Appending to sql
3. Data Cleaning and Preprocessing
4. Exploratory Data Analysis
5. Data Visualization
6. Deriving Insights
7. Reporting

5. Data Understanding

The dataset contains of 4269 rows and 13 columns.

The key variables include

- loan_id
- no_of_dependents
- education
- self_employed
- income_annum
- loan_amount
- loan_term
- cibil_score
- residential_assets_value
- commercial_assets_value
- luxury_assets_value
- bank_asset_value
- loan_status

Data Types

- Object
- Float
- Int

6.Data Cleaning

Data cleaning involved:

- Dropping null values:

  Since there are more null values rows present. So, dropping it.

- Identifying and treating outliers using Z-score method:

  Treating the outliers in the column's using the Z-score method.

7. Obtaining derived metrics

Extracted a columns such as " cibil_category" based on cibil_score and to get the overall insight.

income_annum	loan_amount	loan_term	cibil_score	residential_assets_value	commercial_assets_value	luxury_assets_value	bank_asset_value	loan_status	cibil_category
9600000.0	29900000	12	778	2400000	17600000	22700000	8000000	Approved	Execellent
4100000.0	12200000	8	417	2700000	2200000	8800000	3300000	Rejected	Poor
8200000.0	30700000	8	467	18200000	3300000	23300000	7900000	Rejected	Poor
9800000.0	24200000	20	382	12400000	8200000	29400000	5000000	Rejected	Poor
4800000.0	13500000	10	319	6800000	8300000	13700000	5100000	Rejected	Poor
...	...	...	...	...	...	...	...	...	...
1000000.0	2300000	12	317	2800000	500000	3300000	800000	Rejected	Poor
3300000.0	11300000	20	559	4200000	2900000	11000000	1900000	Approved	Average
6500000.0	23900000	18	457	1200000	12400000	18100000	7300000	Rejected	Poor
4100000.0	12800000	8	780	8200000	700000	14100000	5800000	Approved	Execellent

8. EDA - Univariate Analysis

Univariate Analysis revealed:

1. loan_amount:

From the above skewness value, it is confirmed that it is a symmetrical distribution. -0.5 and 0.5, the distribution of the value is almost symmetrical.

From the above we have clearly noticed that the mean value is greater than the median value, that's because the extreme values affect the mean more than the median.

A kurtosis value of -0.71 indicates that the distribution is platykurtic. Here's what it means in practical terms:

Flatter Peak: The peak of the distribution is lower and wider compared to a normal distribution.

Thinner Tails: There are fewer extreme outliers in the data than in a normal distribution. The tails of the distribution are lighter.

9. Segmented Univariate Analysis

Segmented analysis showed:

1. loan_term:

After analyzing the loan_term,

Inference:

Pattern: The distribution of loan terms appears to be somewhat uniform, with certain loan terms being more common than others. For example:

The highest frequency is observed around 5 years.

Other noticeable peaks are around 10 years, 12.5 years, and 15 years.

Line Plot Overlay:

There is a line plot overlaid on the histogram which shows the trend or density estimate of the loan term distribution. This line helps in visualizing the general trend of the data:

The line appears to fluctuate slightly but does not show any drastic peaks or valleys, suggesting that the loan terms are relatively evenly distributed with some variations.

KDE Line:

The plot also includes a Kernel Density Estimate (KDE) line, which smooths out the distribution and provides an estimate of the probability density function of the loan terms. The KDE line suggests that the probability density is relatively uniform, with slight peaks corresponding to the most common loan terms.

2. loan_amount:

After analyzing the loan_amount,

Inference:

In this box plot, there do not appear to be any points plotted beyond the whiskers, suggesting there are no significant outliers in the loan amounts.

Overall, the box plot indicates that the majority of loan amounts in the dataset are between 10,000,000 and 20,000,000, with a median loan amount of around 15,000,000. The spread of the data is relatively symmetric without any apparent outliers.

3. income_annum

After analyzing the income_annum,

The distribution of annual income appears to be roughly uniform between 0 and 10 million (1e7).

There are slight variations and peaks within this range, but no strong skewness or significant outliers are evident.

The KDE curve shows that the density is relatively flat across the income range, indicating a fairly even spread of income values within the dataset.

In summary, the plot indicates that the annual income of individuals in the dataset is quite evenly distributed across the range from 0 to 10 million, with no significant concentration of income at any specific value within this range.

10. Bivariate Analysis

Segmented Bivariate analysis showed,

1. cibil_score and loan_status:

Applicants with approved loans have higher average CIBIL scores compared to those with rejected loans.

The average CIBIL score for approved loans is around 700.
The average CIBIL score for rejected loans is around 400.

Higher CIBIL scores are strongly associated with loan approvals, while lower CIBIL scores are associated with loan rejections.

This suggests that the CIBIL score is a significant factor in the loan approval process.

Overall, the plot clearly illustrates that a higher CIBIL score increases the likelihood of loan approval, while a lower CIBIL score is associated with a higher likelihood of loan rejection.

2. loan_term and loan_status

Loans that were approved have a shorter average loan term compared to those that were rejected.

The average loan term for approved loans is around 8 years.
The average loan term for rejected loans is around 11 years.

Loans with shorter terms are more likely to be approved, while loans with longer terms are more likely to be rejected.

This suggests that the duration of the loan term is a factor that may influence the loan approval process.

Overall, the plot indicates that there is a tendency for loans with shorter terms to be approved more frequently than loans with longer terms.

3. income_annum and loan_status:

The bar for "Approved" indicates the average annual income of individuals whose loan applications were approved.

The bar for "Rejected" indicates the average annual income of individuals whose loan applications were rejected.

From the plot, it appears that the average annual income is slightly higher for rejected loan applications compared to approved ones. Both bars are very close in height, suggesting that income levels for both approved and rejected applications are quite similar, with a marginal difference favoring rejected applications.

This implies that the annual income might not be a significant factor influencing loan approval decisions, or there could be other factors at play that contribute to the loan approval process.

4. no_of_dependents and loan_status:

For all categories of the number of dependents (0 to 5), the number of approved loan applications is consistently higher than the number of rejected applications.

The count of approved loan applications appears to be quite stable across different numbers of dependents, with each category (0 to 5) showing a similar high count of approvals.

The count of rejected loan applications is relatively stable across different numbers of dependents, with each category showing a similar lower count compared to approvals.

There isn't a significant variation in the pattern of approval or rejection across different numbers of dependents, suggesting that the number of dependents might not be a major factor affecting the loan approval decision.

In summary, regardless of the number of dependents, the loan approval rate is higher than the rejection rate, indicating that having more or fewer dependents does not drastically impact the likelihood of loan approval.

11. Multivariate Analysis

Multivariate analysis revealed:

1. Correlation Heatmap for income_annum, loan_amount, loan_term:

income_annum vs. loan_amount:

The correlation coefficient is 0.92, which indicates a very strong positive correlation. This means that as the annual income (income_annum) increases, the loan amount (loan_amount) also tends to increase significantly.

income_annum vs. loan_term:

The correlation coefficient is 0.013, which indicates an almost negligible correlation. This means there is no significant linear relationship between the annual income (income_annum) and the loan term (loan_term).

loan_amount vs. loan_term:

The correlation coefficient is not shown, but the gray color suggests that the correlation is very low or close to zero, indicating little to no linear relationship between the loan amount (loan_amount) and the loan term (loan_term).

There is a very strong positive correlation between income_annum and loan_amount, suggesting that higher annual incomes are associated with higher loan amounts.

There is no significant correlation between income_annum and loan_term or between loan_amount and loan_term, indicating that these pairs of variables do not have a strong linear relationship.

2. Correlation Heatmap for residential_assets_value, loan_amount, commercial_assets_value

Residential Assets Value vs. Loan Amount: These two variables have a moderate positive correlation (0.58). As residential assets' value increases, loan amounts tend to increase as well.

Residential Assets Value vs. Commercial Assets Value: There is a weaker positive correlation (0.42) between these two variables. When residential assets' value rises, commercial assets' value also tends to increase.

Loan Amount vs. Commercial Assets Value: Similar to the previous case, there's a correlation of 0.42. As loan amounts increase, commercial assets' value tends to rise.

3. Anova test to determine the factors influencing the loan_status:

The factors determining the loan approval is ' income_annum', ' loan_term', Cibil_score'

Anova test

```
# Anova test to determine the factors influencing the loan_status
```

```
VA_C=HS[[" income_annum"," loan_amount"," loan_term"," cibil_score"," residential_assets_value"," commercial_assets_value",
       " luxury_assets_value"," bank_asset_value"]]
VA_D=HS[" loan_status"]
```

```
from sklearn.feature_selection import SelectKBest, f_classif
from sklearn.feature_selection import chi2
```

```
# Apply Anova (x^2) statistical test for feature selection using anova function

best_features=SelectKBest(score_func=f_classif,k=1)
fit=best_features.fit(VA_C,VA_D)
selected_features=VA_C.columns[fit.get_support()]
```

```
selected_features
```

```
Index([' cibil_score'], dtype='object')
```

```
# Apply Anova (x^2) statistical test for feature selection using anova function

best_features=SelectKBest(score_func=f_classif,k=2)
fit=best_features.fit(VA_C,VA_D)
selected_features=VA_C.columns[fit.get_support()]
```

```
selected_features
```

```
Index([' loan_term', ' cibil_score'], dtype='object')
```

```
# Apply Anova (x^2) statistical test for feature selection using anova function

best_features=SelectKBest(score_func=f_classif,k=3)
fit=best_features.fit(VA_C,VA_D)
selected_features=VA_C.columns[fit.get_support()]
```

```
selected_features
```

```
Index([' income_annum', ' loan_term', ' cibil_score'], dtype='object')
```

12. Overall Insights Obtained from Analysis

Based on the inferences drawn from your data exploration and visualization, here's a detailed analysis and interpretation of the findings, along with recommendations for the housing data:

Analysis and Interpretation

Key Factors Influencing Loan approval:

1. Cibil_Score:

- Higher CIBIL scores are strongly associated with loan approvals, while lower CIBIL scores are associated with loan rejections.

2. Loan_term:

- Loans with shorter terms are more likely to be approved, while loans with longer terms are more likely to be rejected.

3. Condition of houses:

- From the plot, it appears that the average annual income is slightly higher for rejected loan applications compared to approved ones. Both bars are very close in height, suggesting that

income levels for both approved and rejected applications are quite similar, with a marginal difference favoring rejected applications.

- This implies that the annual income might not be a significant factor influencing loan approval decisions, or there could be other factors at play that contribute to the loan approval process.

13. Conclusion

The loan approval process in the banking industry faces several business problems that impact risk management, operational efficiency, customer satisfaction, and regulatory compliance. By leveraging advanced analytics, automation, and customer-centric strategies, banks can address these challenges and improve their loan approval processes. Conducting comprehensive EDA on loan datasets is a crucial step in this process, providing the insights needed to make informed decisions and enhance overall performance.