**Introduction**

The **Loan Approval Classification Dataset** from Kaggle offers a compelling glimpse into the world of Banking & Financial Services providing key insights to anyone interested in Financial Analytics and Decision making. This dataset offers a comprehensive view of Loan Applications with financial and demographical attributes that helps in assessing the **Loan Approval or Rejection**. This dataset single-handedly includes both applicants specific and loan specific data that can help our model in being accurate with its prediction. We are using Rapidminer tool to do the Data Exploration on the following dataset.

The Dataset contains 14 columns and 45000 rows. Each row represents a unique Loan Application with 14 diverse attributes associated with it. The datatypes in the dataset ranges from numerical- Integer and Real (person\_emp\_exp, loan\_amnt, person\_income) to Nominal (person\_gender, person\_edu). In this document we will be seeing Statics of Data, Data Visualization, Modelling used as well as in-depth analysis of each attribute.

This dataset serves as an invaluable resource for financial analysis whether you are building a model or an interactive Dashboard.

**Individual Attributes**

The dataset contains 14 attributes in total described below: -

* **person\_age**: - It demonstrates the age of the applicant who applying for loan. Here datatype of the column is integer. It has 60 unique values. The minimum age for the person applying for loan is 20 years and the maximum is 144. Looking at the data we can clearly say there are outliers present in the column. From [Chart-1] we can observe that the data is Positively Skewed looking at the histogram as it doesn’t follow a bell curve & most of the data is between 20 and 40 age group.
* **person\_gender**: - This column depicts the gender of the applicant for loan. It included Categorical variable with two different values in it i.e., male and female.
* **person\_education**: - It tells us the educational background of the person that applied for loan. It is a categorical/Nominal Variable. It consists of 5 distinct values those are Associate, Bachelor, Doctorate, High School, Master.
* **person\_income**: - This column demonstrates the income of the person applying. It is a Quantitate variable with the datatype integer. The values in the column range 8,000 being the minimum & 7,200,766 the maximum income. It is highly right Skewed. Most people applying for loan has salary ranging from 41000 to 76000. Three values are way to high showing the potential outliers in the data. As can be understood from [Chart-2]
* **persom\_emp\_exp**: - This column tells us about the employment experience in years of the individual. It ranges from 0 to 125 with potential outliers and the average being 5.4 years. It is an integral Column.
* **loan\_amt**: - This column depicts the amount of loan that was requested by the individual. The median here is 8,000 and maximum is 35,000Looking at the histogram created it can be observed that the smaller loans are common and larger loans are less frequent. The peaks in the histogram around specific values such as 5,000 10,000 & 15,000 suggest that these are common loan amounts.
* **person\_home\_ownership**: - it depicts home ownership status with 4 distinct values Rent, Own, Mortgage and other.
* **Loan\_intent**: - Purpose of loan. categorical variable with 6 distinct values Personal, Venture, Medical, DebtConsolidation, Education & Home-Improvement.
* **Loan\_int\_rate**: - Depicts interest rate of loan. Numerical with real values with minimum of 5.420 maximum of 20 and average being 11.007.
* **loan\_percent\_income**: - It is the ratio of persons income with loan amount which ranges from 0 to 0.6 showing loan amount is always a small percentage of individuals income.
* **cb\_person\_credit\_history\_length**: - Length of credit history in years with 2 years being the minimum and 30 the maximum.
* **credit\_score**: - It tells the analyser what the credit score of the person is.
* **Previous\_loan\_defaults\_on\_file**: - this column tells us whether the person applied has any previous loan defaults or not with just 2 distinct values YES or NO.
* **Loan\_status**: - This is Dependent column showing loan status i.e., loan Approved = 1 or loan Rejected = 0.

**Group of Attributes**

* **Age vs Loan\_amnt Distribution**: - [Chart 3] represents the relationship between persons age and amount of loan asked and the 2 distinct colour represents whether his loan application was approved (green dots) or rejected (blue dots). Young applicants aged 20 – 30 tend to apply for higher loan amounts as the concentration is dense in this range. A significant number of loan Applications from younger applicants are approved. As age increases the number of approved loan decreases. Higher loan amounts above 25,000 are less frequently approved.
* **Age vs Loan\_status Distribution with Credit Score**: - [Chart-4] represents distribution of loan status by age corresponding to average credit score for each age group. Blue bar represents loan status approved, green bar represents rejected and red line represents the avg credit score. Younger age group, in 20’s has higher counts of applicants. Count of both approved and rejected loan tends to decrease as age increases with stable avg credit score (minor Fluctuations)
* **Emp\_experience trend with Loan\_status**: - [Chart-5] represents the relationship between employment experience and loan\_status. It can be observed that as experience years increases the loan\_status both approved and rejected decreases. The applicants with 0 years of employment experience have the highest approval and rejection rates.
* **Loan\_status vs person\_education with Credit Score**: - [Chart 6] depicts how loan\_status is related to education. It can be observed that most of the loan applicants have a bachelor’s degree moreover the approval and the rejection rate is highest for them as well. And we can see that the avg credit score for each group is stable with no fluctuations.
* **Loan\_status vs House\_ownership Distribution: -** [Chart 7]demonstrates that the approval status of loan is highest with people who have a rented home. And the rejection status is highest who have house on Mortgage.
* **Approval distribution based on Loan\_amnt: -** [Chart 8] it gives us the relationship between loan\_amount and loan status which can be concluded as the approval rates are relatively high for smaller amounts. The applications drop drastically as the amount increases from 25,000 indicating fewer applicants with high amounts.
* **Applicants based on Intent: -** [Chart-9] it shows is the tree map of what’s the purpose of loan. from the tree map we can say that the applicants applying for loan with education as reason is the highest and home-improvement is the lowest.
* **Box plot for Age & Emp\_experience: - [**Chart 10**]** represents 2 box plot one for age and other for emp\_experience for age it can be depicted that there are outliers close to or above 150 years which are unrealistic & the median is around 25. WHEREAS the median of employment experience is low near to 0 suggesting lots of applicants with 0 or little to no experience.
* **Trend of age vs loan\_status: -** [Chart-11] It demonstrate how the loan status varies as age increases.
* **Loan\_status based on previous loan defaults: -** [Chart-12] the applicants who have previous loan defaults had been rejected directly for the loan which they applied with no exceptions.
* **Loan applicants by House Ownership:** - [Chart-13]It can be concluded from the pie chart that most of the applicants have rented place or are living on mortgage.
* **Loan\_status distribution Based on emp experience & gender: - [**Chart-14] Male applicants dominate loan applications as experience increases the distribution start to balance.
* **Loan distribution by Education: -** [Chart -15] it can be observed that most of the loan applicants have bachelor’s degree and out of all applicants only 435 applicants have Doctorate level degree.

**Interesting Findings**

**Demographic Insights: -**

* Younger applicants (20-30) show higher approval rates, especially for loans under 25,000. This approval rate is because of more potential return due to the applicant’s younger age. Targeting products or offering financial counselling to this demographic may increase engagement.

**Applicants profile and approval rates: -**

* **E**xperience matters less for High approval. Applicants with no work experience display high approval rates suggesting the industries focus on other factors such as income or Credit History over Employment Tenure.
* The data reveals that applicant with prior defaults face Universal rejection. Identifying ways to support or provide products for prior defaults may open a market segment.

**Approval threshold: -**

* Smaller loan amounts are easily to be approved. Stakeholders should consider Implementing policies to pre-qualify the applicants based on Requested loan size. For instance, Applicant’s requesting higher loans should require stricter loan approval criteria than those requesting smaller loan.
* Despite age variations, Credit score remains stable suggesting age might not strongly impact credit scoring.
* Most applicants hold a bachelor’s degree with highest approval and rejection rates also in the group.
* Renters show the highest loan approval rates while those with the mortgage face higher rejection rates.
* Education purpose dominate loan intent, followed by debt consolidation and home improvement. The stakeholders should pitch in tailored loan products designed for educational funding.
* Also, there is imbalance in data for gender. The count of male applicants is higher than the count of female applicants suggesting that more male applicants are applying for loan or that a larger portion of male applicants are being rejected compared to female applicants.

**APPENDIX**

Reference [1]:- <https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data>

A graph of a person's age

Description automatically generated

[Chart-1]

A graph showing a number of people

Description automatically generated

[Chart-2]

A graph with many dots

Description automatically generated with medium confidence

[Chart -3]

A graph of credit score

Description automatically generated

[Chart 4]

A graph of a graph with a number of people

Description automatically generated with medium confidence

[Chart-5]

A screenshot of a computer

Description automatically generated

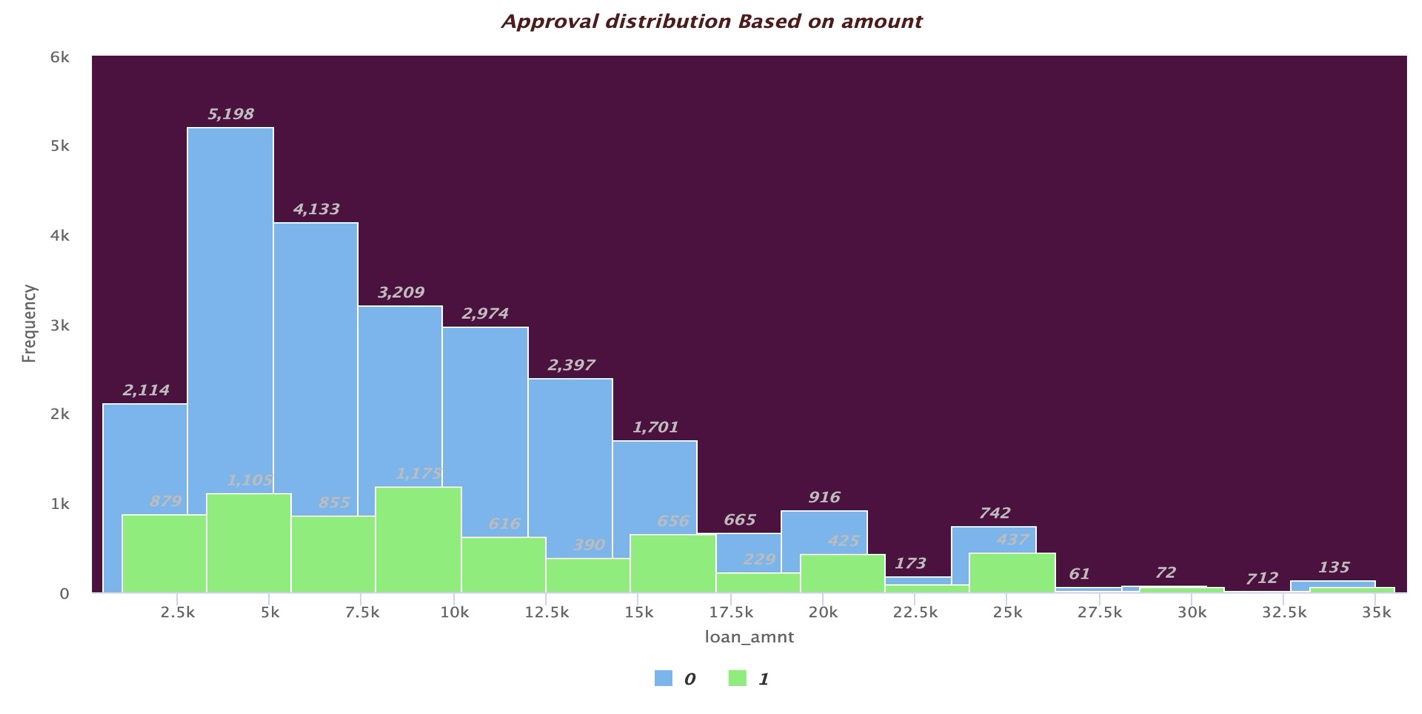
[Summary statiscs -1] ---[for reference ]

A graph of a graph with numbers and text

Description automatically generated with medium confidence A graph showing the amount of housing in the house

Description automatically generated with medium confidence

[Chart-6] [Chart-7]



[Chart -8]

A diagram of a loan distribution

Description automatically generated

[Chart-9]

A diagram of a box plot

Description automatically generated

[Chart -10]

A graph of a person with a line graph

Description automatically generated with medium confidence

[Chart-11]

A graph showing a few different colored squares

Description automatically generated with medium confidence

[Chart-12]

A pie chart with a number of different colors

Description automatically generated with medium confidence

[Chart-13]

A graph showing a number of people

Description automatically generated

[Chart-14]

A diagram of a pie chart

Description automatically generated

[Chart-15]