Data Analysis and Predictions for Personal Loans

Kunlong He, Chengfeng Xiao, Yu-Yang Chen

Simon Fraser University

CMPT353 E100

Dr. Steven Bergner

Apr 1th, 2024

1. Problem Statement

1.1 Define the Problem

The first part of the question is inspired by Vikas Ukani's Loan Eligibility post on Kaggle. A finance company specializing in home loans has a presence in urban, semi-urban, and rural areas. When a customer applies for a home loan, the company needs to validate the customer's eligibility. To streamline this process, the company seeks to automate the evaluation of loan eligibility based on customer details submitted through an online application form. These details include Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, and Credit History, among others. The company aims to use this automation to identify customer segments that are eligible for loans, enabling targeted marketing efforts.

1.2 Refine the Problem

This is a classification problem where we need to predict whether a loan will be approved. Before making predictions, it is crucial to analyze the relationships between all the independent variables and the target variable (Loan_Status). Understanding these correlations allows the company to validate customer eligibility more effectively. For instance, if a high income is strongly correlated with loan approval, this metric can be weighted more heavily in the decision-making process. Conversely, variables with low or negative correlations might be deprioritized, streamlining the assessment procedure and focusing on the most impactful factors. After analyzing the relationships and correlations among variables, we have to create a model that utilizes these features to predict the target variable. Here are the refined questions:
(1) Based on our life experiences, we can more or less guess which features might influence loan approval, although these are still subjective speculations and may not necessarily be correct. Here are some examples:
- Marital Status: Being married might be associated with greater loan approval rates, possibly indicating financial stability or dual incomes to service the loan.

- Loan Amount: Larger loan amounts might be approved less readily than smaller ones, as they entail more risk for the lender.
- Credit History: Applicants with a positive credit history are typically more likely to be approved for a loan, as this suggests a track record of managing debt responsibly.

The independent variables in the training dataset include numerical and categorical variables. Validate each speculation based on the dataset and analyze which variables are more likely to get loans approved.

(2) Typically, when we consider loan status, the first factor that comes to mind is income. However, if we trace back to what attributes might influence a person's income, it could likely be their level of education. Thus, we pose this general question: Does education influence different loan amounts by impacting income? From a statistical perspective, this question can be reformulated as:

- H0: Education does not have an impact on the customer's loan amount by impacting income
- H1: Education does have an impact on customers' loan amounts by impacting income

(3) Select an appropriate model to determine whether each customer's loan application in the loan-test dataset will be accepted or rejected.

2. Data Preprocessing

2.1 Getting Data

The loan-train and loan-test datasets were downloaded from Kaggle. The former is used for training the model and testing its accuracy, including all independent variables and the target variable. The latter contains all the independent variables, but not the target variable. We will use our model on this dataset to predict whether a customer's loan has been approved.

We will use Python to explore the data to better understand the features and the target variable. We will also employ various visualization techniques to analyze the data and summarize their main characteristics. Here are the shapes of the two datasets:

```
(train.shape, test.shape)
```

```
((614, 13), (367, 12))
```

As we can see, the size of the training dataset is almost twice that of the test dataset.

2.2 Data Cleaning

The training dataset has 614 rows and the testing dataset has 367 rows, but they contain 149 and 84 missing values respectively, which is almost a quarter of the data. Removing these entries could lead to a significant loss of information, thereby reducing the accuracy of model training and validation. Therefore, using imputation instead of dropping these rows is necessary.

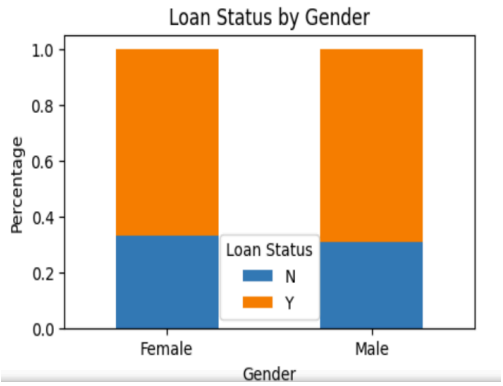
We've learned that common ways to impute missing data include using nearby values, averaging known data or mean substitution, employing linear regression of nearby values, or finding some other best fit. After using the Pandas library function `value_counts()` to count the frequency of each value in various data series, we observed that most of the data tend to cluster around a few fixed values (for example, out of 614 customers, 489 are male). Therefore, using the mode to impute missing values in these columns is justified, as such imputation preserves the overall distribution characteristics of the original data.

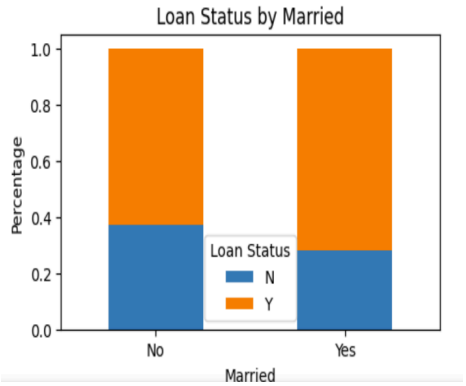
However, the situation with the LoanAmount column is different. It is numerical data with a wide distribution across 203 different values, and no single value dominates. Using the mode to impute missing values may not reflect the actual distribution of loan amounts. Therefore, for such data, we opt to use the mean for imputation, which better preserves the central tendency of the data.

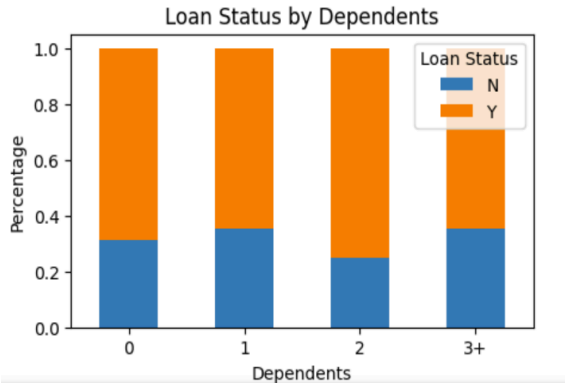
Lastly, we attempted to filter and remove any incorrect values from both datasets. Fortunately, the values in the datasets all conform to the expected range. This is great. Now, we are ready to analyze the data.

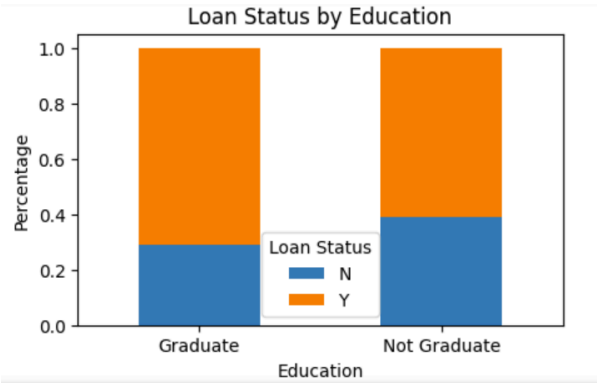
2.3 Independent Variable vs Target Variable

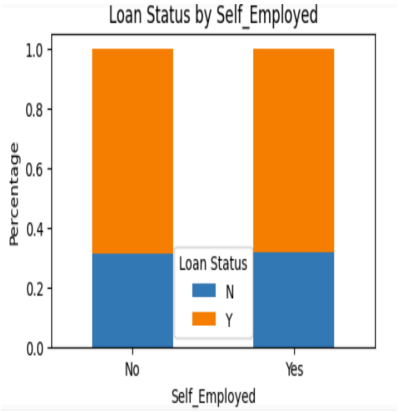
First of all, we want to identify the relation between categorical independent variables and the target variable. The `crosstab()` function from Pandas helps us visualize these relationships with stacked bar charts, demonstrating how these categorical variables are associated with loan approval status.

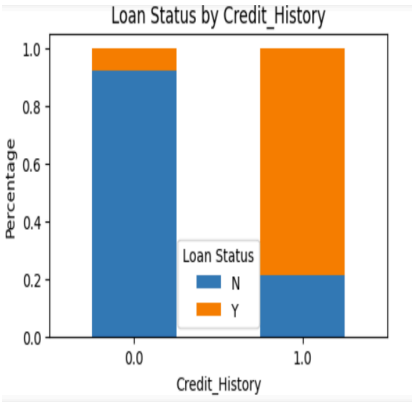


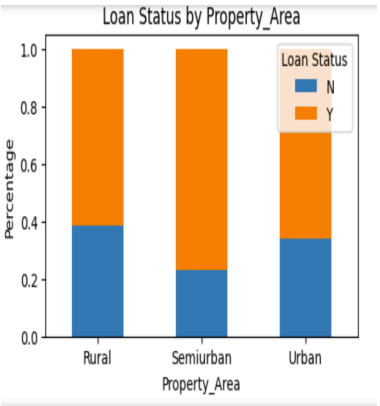








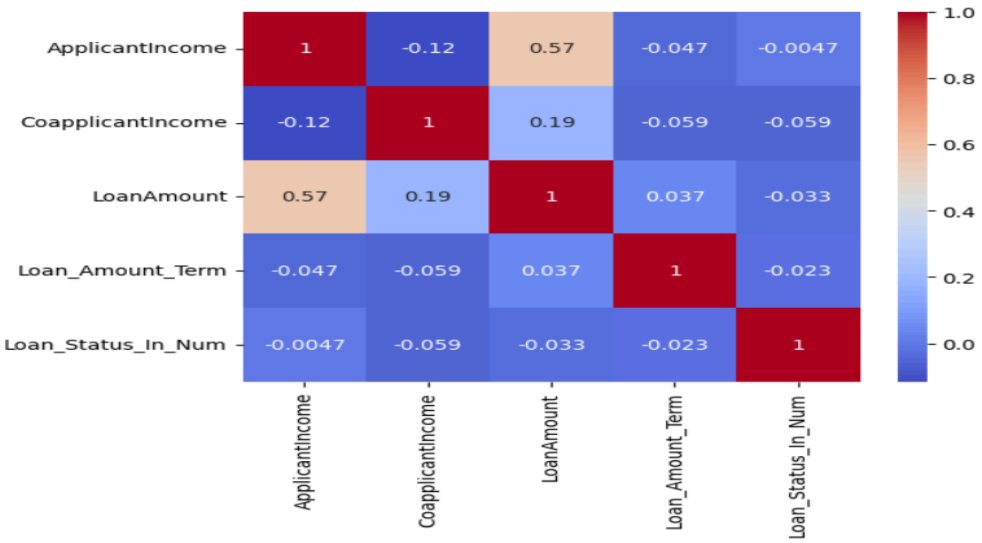




From the bar charts above, it can be inferred that the proportions of customers of different genders, marital statuses, and whether they are self-employed are roughly the same for both

approved and unapproved loans. The distribution of applicants with either one or more than three dependents is similar, and there appears to be some variation in the loan approval rates among people with different educational levels. What surprises me the most is that the percentage of loan approvals is higher in semi-urban areas compared to rural or urban areas. It is obvious that individuals with a credit history of 0 are more likely to be denied loan approval compared to those with a credit history of 1, and this contrast is the most striking. Hence, we can reasonably conclude that credit history is the most significant categorical variable affecting loan approval.

For numerical independent variables, we use the `corr()` function to compute the pairwise correlation of columns and the target variable. Then we will use a heatmap to show the correlations. A heatmap visualizes data through variations in color. Darker colors in the variables indicate higher correlations.



Applicant_Income has the highest correlation with the target variable at 0.57, indicating that among numerical variables, the customer's income is the most significant factor affecting loan approval, and the higher the income, the easier it is to get a loan approved.

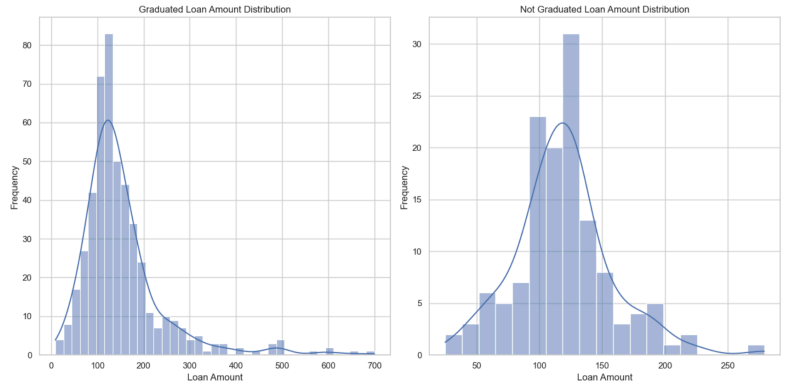
3. Statistical Analysis

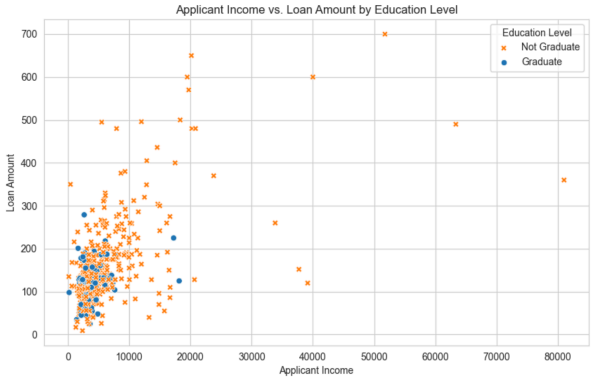
3.1 Data Transformation

From Section 2.3, We have obtained a stacked bar chart of education and loan approval status, but this is merely a visual comparison. In this section, we aim to use statistical knowledge to conclude.

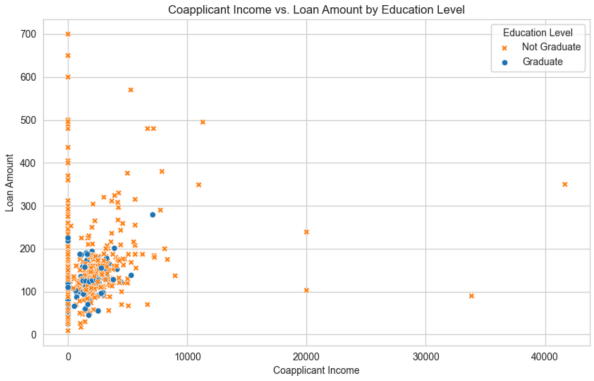
Before applying any statistical tests, A crucial preprocessing step was transforming the 'Education' variable into a binary categorical variable, allowing for a clearer comparative analysis between graduates and non-graduates. By mapping 'Graduate' to 1 and 'Not Graduate' to 0, the complexity of the analyses was significantly reduced, and we're ready to perform statistical tests.

(Figure 1)





(Figure 2)



(Figure 3)

Here are some insights into the relationship between education level and loan amount, as well as how applicants' income levels may interact with these variables.

3.2 EDA: Visualization

In Figure 1, we observe that graduates appear to receive a broader range of loan amounts, evident from the wider box, which suggests greater variability. Additionally, the median loan amount for graduates is higher than that for non-graduates, as indicated by the position of the line within the box. This initial graphical analysis supports the hypothesis that education level may have a significant impact on the loan amounts that applicants receive. Figure 2 further elaborates on this relationship, it shows a concentration of both graduates and non-graduates at the lower income and loan amount levels. There's a noticeable absence of non-graduates at higher income levels, which may suggest that individuals with higher education tend to also have higher incomes. Looking at Figure 3, a similar pattern emerges. Again, we see a cluster of points at the lower end of the income scale, with graduates spread across a wider range of loan amounts. However, we do not see co-applicants with high incomes correlating to particularly high loan amounts, which indicates that co-applicant income is not as strongly associated with the loan amount as the applicant's income or educational level.

3.3 Testing

The normal-test was used to determine the normality of loan amount distributions for graduates and non-graduates. Unfortunately, both p-values show that they’re not normally distributed, the relation of graduates combined with loan amount gives the p-value of 2.5467525185710977e-61, and ungraduates give the p-value of 0.0001097132510381849. This conclusion is significant because it has an impact on the selection of statistical tests. Since parametric tests are not suitable, non-parametric tests become the suitable analytical test.

3.4 Results

Attempts to normalize the loan amount data using a logarithmic transformation yielded distributions that did not meet the criteria for normality, so the Mann-Whitney U test was chosen to test the null hypothesis that there is no difference in median loan amounts between graduates and non-graduates. The test yielded a p-value of 4.722537698365795e-05, suggesting that the null hypothesis could be rejected. This shows

that there is a statistically significant variation in loan amounts issued to individuals based on their education level, with graduates more likely to receive larger loans than non-graduates. To further analyze the impact of education, we calculated the p-value of applicant income grouped with the loan amount. The p-value of applicant income is 3.221625380210046e-34, which shows that applicant income has more impact on loan amount than education. Another layer of interpreting this is that education has little impact on their income since education should have more impact on loan amount if applicants' income has this much.

4. Machine Learning

4.1 PCA

Within our dataset, three numerical variables, ApplicantIncome, CoapplicantIncome and Loan Amount, were analyzed to evaluate their influence on Loan Status. We employed Principal Component Analysis (PCA) on these variables to discern their relative impact. The PCA results indicate:

```
Explained variance by component for loan_df: [0.52338813 0.35616341 0.12044847]
```

So based on the result, we can tell the core influence for whether applicants can get Loan is their Income.

4.2 GaussianNB Model

Our initial foray into machine learning involved employing the GaussianNB model. The observed outcomes suggested potential deviations, leading us to conjecture that the variables might not adhere to a normal distribution. To substantiate this, we standardized the dataset. The result is shown below. This outcome lends weight to our hypothesis regarding the distribution of the variables.

```
Accuracy for GaussianNB is: 0.81
              precision    recall  f1-score   support

           0       0.89      0.47      0.61       100
           1       0.79      0.97      0.87       207

    accuracy                           0.81       307
   macro avg       0.84      0.72      0.74       307
weighted avg       0.82      0.81      0.79       307
```

4.3 RandomForest Model

Despite notable improvements through feature engineering in the GaussianNB model, two significant issues were encountered:

1. Standardization might distort the original data's features, potentially leading to inaccuracies.

2. The GaussianNB model's inability to incorporate categorical variables (e.g., Gender, Marital Status) limited its applicability.

To overcome these challenges, we pivoted to the RandomForest Model. The initial step involved normalizing the numerical data to ensure a consistent scale across variables. Subsequent results highlighted

Accuracy for RandomForest is: 0.78

	precision	recall	f1-score	support
0	0.90	0.42	0.57	43
1	0.76	0.97	0.85	80
accuracy			0.78	123
macro avg	0.83	0.70	0.71	123
weighted avg	0.81	0.78	0.75	123

Additionally, an experiment focusing solely on the three numerical variables was conducted to elucidate the categorical data's impact. This verification confirmed our earlier findings and underscored the categorical variables' significance. The result shows below.

Accuracy for three core variables is: 0.63

	precision	recall	f1-score	support
N	0.38	0.14	0.21	21
Y	0.67	0.88	0.76	41
accuracy			0.63	62
macro avg	0.52	0.51	0.48	62
weighted avg	0.57	0.63	0.57	62

This accuracy indicates that after eliminating categorical data, the accuracy decreases about 0.15, which means categorical data have a big impact on Loan Status.

By transitioning to the RandomForest model, we aimed to mitigate the GaussianNB model's limitations while capitalizing on the RandomForest model's capacity to seamlessly handle both numerical and categorical data, thereby enhancing the overall robustness and accuracy of our analysis.

4.4 Model selection

We decided to employ a RandomForest model to predict the outcomes of loan applications. Although the Gaussian Naive Bayes (GaussianNB) model exhibits higher accuracy compared to the RandomForest model, it's important to note that the data used to train the GaussianNB was standardized. This standardization assumes that the data follows a normal distribution, which may not be the case. Therefore, the validity of the GaussianNB model's superior performance could be questionable without verifying the distribution assumption of the underlying data. After training the RandomForest model, we use it to determine whether each loan application in our 'loan-test.csv' dataset will be accepted or rejected. The prediction results are then appended to the 'loan-test.csv' file, providing a comprehensive view of both the application details and their predicted outcomes.

	Gender	Married	Dependents	...	Credit_History	Property_Area	Predicted_Loan_Status
0	Male	Yes	0	...	1.0	Urban	Y
1	Male	Yes	1	...	1.0	Urban	Y
2	Male	Yes	2	...	1.0	Urban	Y
3	Male	Yes	2	...	1.0	Urban	Y
4	Male	No	0	...	1.0	Urban	N

5. Results, Findings, and Limitations

By visualizing the relationships between independent variables and the target variable, we can see that Credit_History is strongly associated with Loan_Status, with customers having a credit history of 0 being highly likely to be denied loan approval. Among the numerical variables, ApplicantIncome has the highest correlation with the target variable. Although the correlation coefficient is only 0.57, not extremely strong, it still indicates a moderate positive relationship between the variables. Generally, customers with higher incomes are more likely to get loan approvals, which aligns with the expectations of most people. Next, in order to explore whether education impacts different loan amounts, we conducted several statistical tests. Among these, the Mann-Whitney U test revealed a significant difference in median loan amounts between graduates and non-graduates, leading us to reject the null hypothesis that education does not affect loan amounts. Further, we analyzed the relationship between income and loan amounts, concluding that although education level affects loan amounts, income plays a more critical role. Lastly, we chose the RandomForest model as our final model because it effectively handles both numerical and categorical data without requiring normal distribution assumptions, providing robust and reliable predictions for loan application outcomes. In evaluating the machine learning model, we've identified two primary concerns. First, the accuracy of the model stands at 0.78, which, given the binary nature of the outcomes (Yes or No), is moderately acceptable. However, there is significant room for improvement to achieve more reliable predictions. Second, the limited size of our dataset poses another challenge, as it may not provide a sufficiently broad representation to ensure the universality of the results. The lack of actual results for the test cases limits our ability to quantitatively assess the model's accuracy on this specific dataset. There are many things that can be tried to improve the models' predictions. We can create and add more variables, and try different models with various subsets of features. If we had more time, we could also experiment with neural networks using PyTorch.

7. Contributions

Kunlong He

- Defined/refined the problem, and set clear objectives for team members
- Acquired, cleaned, and transformed data using Numpy and Pandas
- Analyzed the independent variables and the target variable to identify relationships and correlations between them
- Organized and summarized the key points of each section

Chengfeng Xiao

- Machine learning
- Performing PCA, randomforest algorithm, GaussianNB
- Limitations on machine learning

Yu-Yang Chen

- Statistical analysis
- Applying different statistical tests to preprocessed data
- Limitations on statistical analysis