Literature Review and Data Description

Submitted by : Joveria Sadain
Student number : 500979559
Supervisor : Dr. Ceni Babaoglu
Submission date : 22-02-2024

Ryerson University

Literature Review

Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis

(Mamun et al., 2022) conducted a study on a similar dataset. This research utilizes Machine Learning (ML) algorithms to analyze a standard dataset of approved loans, aiming to identify eligible loan applicants by uncovering underlying patterns. The investigation relies on customers' historical data, encompassing factors such as age, income source, loan annuity, recent credit bureau report, employment organization type, and tenure. Various ML techniques including Random Forest, XGBoost, Adaboost, Lightgbm, Decision tree, and K-Nearest Neighbor are applied to identify the most influential features that affect prediction outcomes. These algorithms are evaluated against each other using established metrics. Notably, Logistic Regression achieves the highest accuracy at 92%, emerging as the superior model with a commendable F1-Score of 96%, outperforming other ML approaches significantly.

This paper utilizes a dataset sourced from the Kaggle online platform, comprising 10,128 instances and 23 attributes. Among these attributes, one is designated as the class attribute, while the remaining 23 serve as predictive features.
In this study, they not only examined accuracy but also the AUC value, which is crucial for assessing a model's capability to differentiate between classes. This metric becomes significant in evaluating the overall performance of the model.

Developing prediction model of loan risk in banks using data mining

(Hamid and Ahmed, 2016) worked on a new model for classifying loan risk within the banking sector utilizing data mining techniques. A dataset sourced from the banking sector was chosen, formatted in the Attribute-Relation File Format (ARFF), suitable for use with Weka. This format includes tags specifying attribute names, types, values, and the data itself. Following preprocessing steps and feature selection to identify the most significant attributes, eight attributes were retained: Credit_history, Purpose, Gender, Credit_amount, Age, Housing, Job, and the Class attribute.This paper explores the utilization of three algorithms - J48, BayesNet, and NaiveBayes - to construct predictive models aimed at classifying loan applications as either good or bad based on an analysis of customer behaviors and past credit repayment history. The models were implemented using the Weka application. Following the application of classification data mining techniques, it was determined that the J48 algorithm outperformed BayesNet and NaiveBayes algorithms in loan classification. The superiority of the J48 algorithm is attributed to its high accuracy and low mean absolute error.

Loan eligibility prediction based on credit score and past history

(Senarathna et al., 2023) published a research paper that takes into account the findings of four distinct components proposed by various researchers to introduce a comprehensive and innovative strategy. The first component utilizes advanced machine learning algorithms such as Random Forest, Gradient Boosting, and Linear Regression to forecast optimal bank rate possibilities. The second component, which focuses on assessing applicants' loan eligibility, introduces a ground breaking approach that incorporates their credit histories while considering crucial factors like gender, marital status, education level, income, and credit history itself. The third element proposes a sophisticated Loan Eligibility Prediction System encompassing several vital subobjectives, including creditworthiness evaluation, income and employment verification, collateral analysis, risk profiling, fraud detection, and regulatory compliance, to ensure thorough risk assessment. To efficiently and effectively predict applicant risk levels, this component harnesses the capabilities of logistic regression. The fourth component introduces a sophisticated mortgage calculator tool designed to aid in mortgage estimation and analysis, employing Decision Tree Regression. This tool enables users to calculate and assess mortgage values by considering essential property attributes such as the number of bathrooms, bedrooms, the house's size, and its location.

This research paper focus on an in-depth 4 step process for accurate loan approval and minimizing risk associated with the process. The steps are an end-to-end procedure that aims to help not only the banking sector but also make it easier for the applicants to apply for mortgages. The processed can be summarized as:

1. **Predicting Optimal Bank Rate Options**

2. Leverage Credit History for Loan Eligibility Assessment

3. Comprehensive Risk Analysis

4. Advanced Mortgage Estimation Tools

Each step in the process involves subsequent steps to gather, analyze, and extract valuable information from data. This iterative process then applies relevant machine learning algorithms to obtain the necessary insights, which serve as the foundation for the subsequent steps in the work flow.

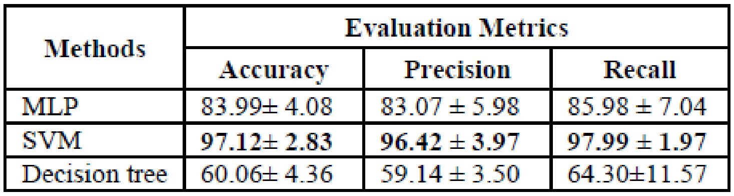
The research paper's proposal of a comprehensive system integrating multiple components has made a substantial contribution to the field of loan eligibility prediction. The study demonstrates how the evaluation of credit history, utilization of machine learning algorithms, and implementation of sophisticated prediction tools can potentially revolutionize the loan approval process.

Data Mining Approaches in Personal Loan Approval

(Pimcharee and Surinta, 2022) published a research paper that focused on applying different machine learning techniques on key factors that affect the bank loan approval decisions. The main objective of this research was to analyze the effect of feature selection on different machine learning algorithms and the accuracy achieved after them.

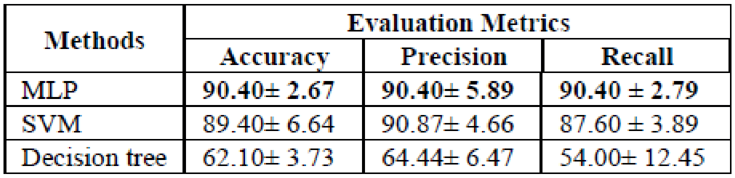
Three machine learning algorithms, Decision trees, support vector machine (SVM) and multi -layer perceptron (MLP) were deployed. The dataset used had 1000 instances and 14 variables from a banking sector. 500 applications were approved and the other half was rejected.

In the first step of the research, all the variables were used in all three machine learning algorithms and the result were as follows:



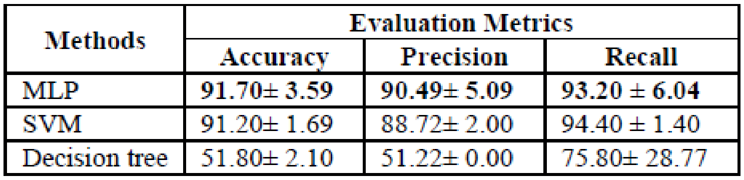
Methods	Evaluation Metrics		
	Accuracy	Precision	Recall
MLP	83.99± 4.08	83.07 ± 5.98	85.98 ± 7.04
SVM	**97.12± 2.83**	**96.42 ± 3.97**	**97.99 ± 1.97**
Decision tree	60.06± 4.36	59.14 ± 3.50	64.30±11.57

In the next step, feature selection was done on the basis of Chi-square and 10 out of 14 variables were used in all three machine learning algorithms. The results were as follows:



Methods	Evaluation Metrics		
	Accuracy	Precision	Recall
MLP	**90.40± 2.67**	**90.40± 5.89**	**90.40 ± 2.79**
SVM	89.40± 6.64	90.87± 4.66	87.60 ± 3.89
Decision tree	62.10± 3.73	64.44± 6.47	54.00± 12.45

Lastly, feature selection using Information Gain was used and only 3 out of 14 variables were selected for Machine learning algorithms. The results are shown below:



Methods	Evaluation Metrics		
	Accuracy	Precision	Recall
MLP	**91.70± 3.59**	**90.49± 5.09**	**93.20 ± 6.04**
SVM	91.20± 1.69	88.72± 2.00	94.40 ± 1.40
Decision tree	51.80± 2.10	51.22± 0.00	75.80± 28.77

The findings of this research indicate that the feature selection method is not advisable for credit approval evaluation, as it yields an accuracy rate of 91%. On the other hand, utilizing all 14 features resulted in a higher accuracy rate of 97%. Moving forward, our future research endeavors aim to explore alternative feature selection methods such as factor analysis, LightGBM, and XGBoost [25-27]. Additionally, we plan to investigate ensemble learning methods [28] to incorporate various machine learning models into the credit approval process. It is anticipated that this approach will lead to an enhanced accuracy rate for the credit approval process.

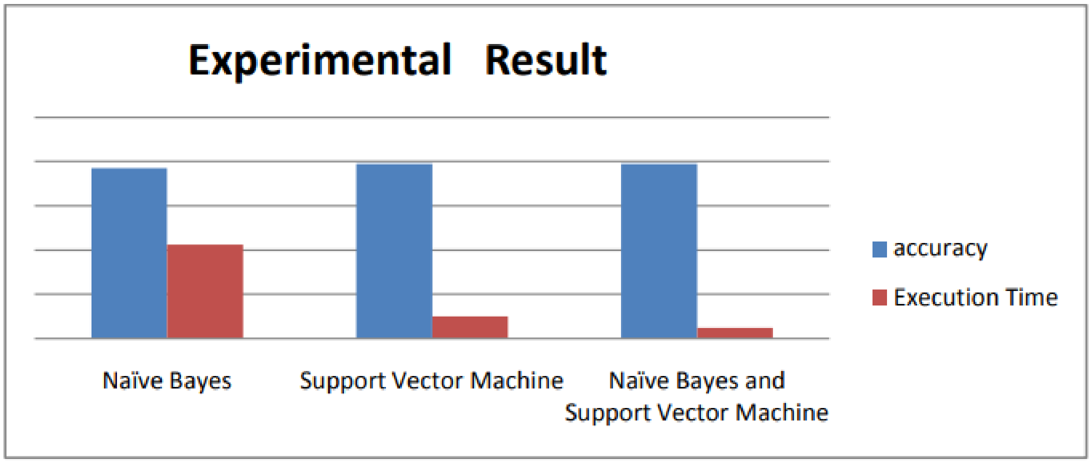
Prediction of Loan Risk using Naive Bayes and Support Vector Machine

(Vimala and Sharmili, 2018) presented their findings on a UCI provided German credit dataset having 21 attributes with their respective class. The research compared the result of using Naïve Bayes and Support vector machine (SVM) on the dataset and then evaluating their accuracy and execution time.

Naive Bayes is a straightforward technique grounded in the Bayes theorem, making predictions based on independent assumptions. Its simplicity and robustness make it a popular choice for classification tasks. Leveraging probability theory, Naive Bayes categorizes data effectively. This method aids in building models with predictive capabilities, offering a fresh perspective for interpreting data.

Support Vector Machines (SVMs) originated from statistical learning theory. SVM is a type of learning algorithm utilized to enhance classification accuracy. Widely employed in various fields including handwriting digit recognition, character recognition, text recognition, and satellite image classification, SVMs are highly versatile. At the core of SVM lies the hyperplane, also referred to as the "Decision boundary" or "Decision surface," which effectively separates positive and negative instances in the data.

The main highlight of this paper is using the Naïve Bayes and SVM algorithm together. These were the findings from the paper.



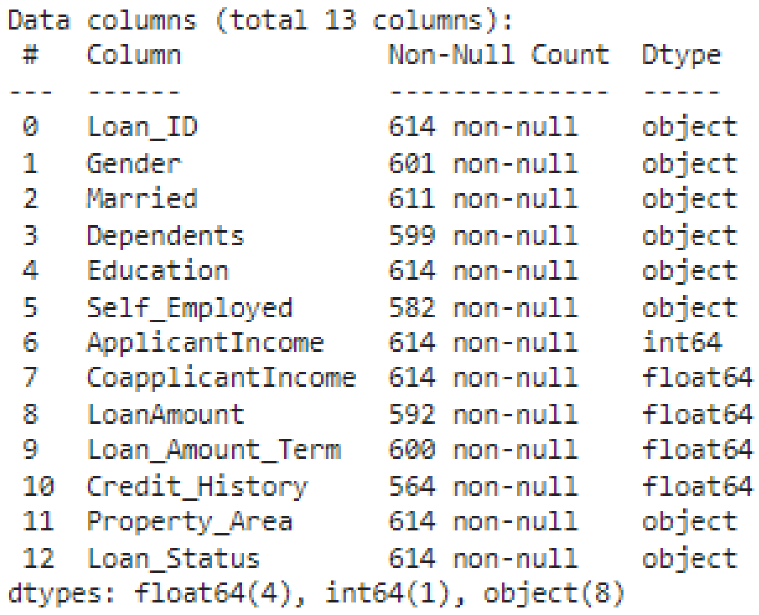
Loan approval prediction model a comparative analysis

(Khan et al., 2021) published a research paper with an objective of this project was to evaluate and compare different Loan Prediction Models to determine the most effective one with minimal error, suitable for real-world application by banks to assess loan approval decisions while considering risk factors. Following thorough comparison and analysis, the Random Forest-based prediction model emerged as the most accurate and well-suited among the options examined.

According to this research, the predictive models utilizing Logistic Regression, Decision Tree, and Random Forest achieved accuracies of 80.945%, 93.648%, and 83.388%, respectively. However, their cross-validation results are found to be 80.945%, 72.213%, and 80.130%, respectively. These findings indicate that, for the given dataset, the decision tree-based model exhibits the highest accuracy. Nonetheless, the Random Forest model demonstrates better generalization, despite its cross-validation rate not significantly surpassing that of logistic regression.

Data Description

The Dream housing Finance loan prediction dataset has been taken from kaggle’s public domain. The dataset comprises of 614 observations and 13 attributes. Here is some descriptive statistics about the data:



```
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Loan_ID            614 non-null    object 
 1   Gender             601 non-null    object 
 2   Married            611 non-null    object 
 3   Dependents         599 non-null    object 
 4   Education          614 non-null    object 
 5   Self_Employed      582 non-null    object 
 6   ApplicantIncome    614 non-null    int64  
 7   CoapplicantIncome  614 non-null    float64
 8   LoanAmount         592 non-null    float64
 9   Loan_Amount_Term   600 non-null    float64
 10  Credit_History     564 non-null    float64
 11  Property_Area      614 non-null    object 
 12  Loan_Status        614 non-null    object 
dtypes: float64(4), int64(1), object(8)
```

Univariate Analysis:

The attributes data types are as follows:

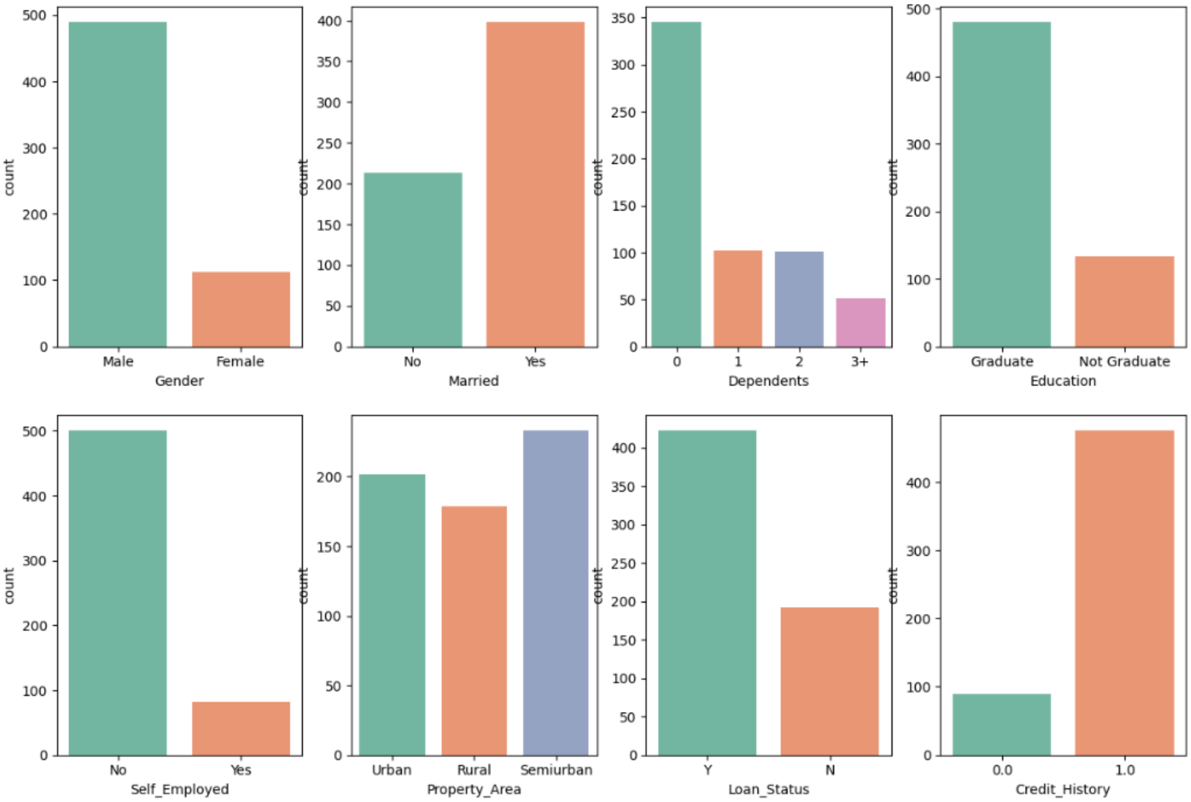
Categorical	5
Numeric	4
Boolean	3
Text	1

Listed below are the variables in the dataset and their description:

Variable	**Description**	**Data Type**
Loan_ID	Unique Loan ID	Text
Gender	Male/ Female	Categorical
Married	Applicant married (Yes/No)	Boolean
Dependents	Number of dependents on the applicant	Categorical
Education	Applicant Education (Graduate/Under Graduate)	Categorical
Self_Employed	Self-employed (Yes/No)	Boolean
ApplicantIncome	Applicant's monthly income	Numeric
CoapplicantIncome	Coapplicant's monthly income	Numeric
LoanAmount	Loan amount in thousands	Numeric
Loan_Amount_Term	Term of loan in months	Numeric

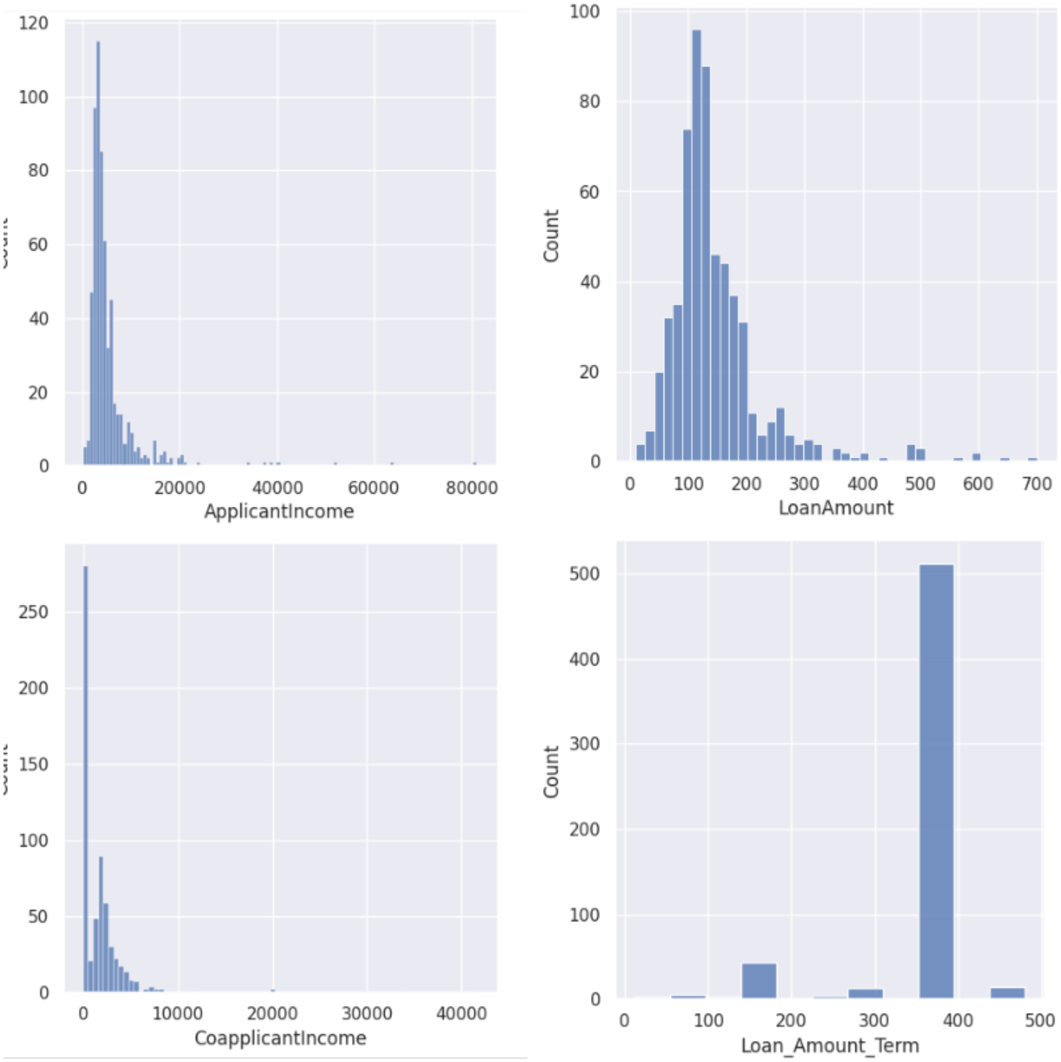
Variable	**Description**	**Data Type**
Credit_History	credit history meets guidelines (1-Yes , 0-No)	Categorical
Property_Area	Urban/ Semi Urban/ Rural	Categorical
Loan_Status	Loan approved (Y/N)	Boolean

- Categorical Variable Visualizations:

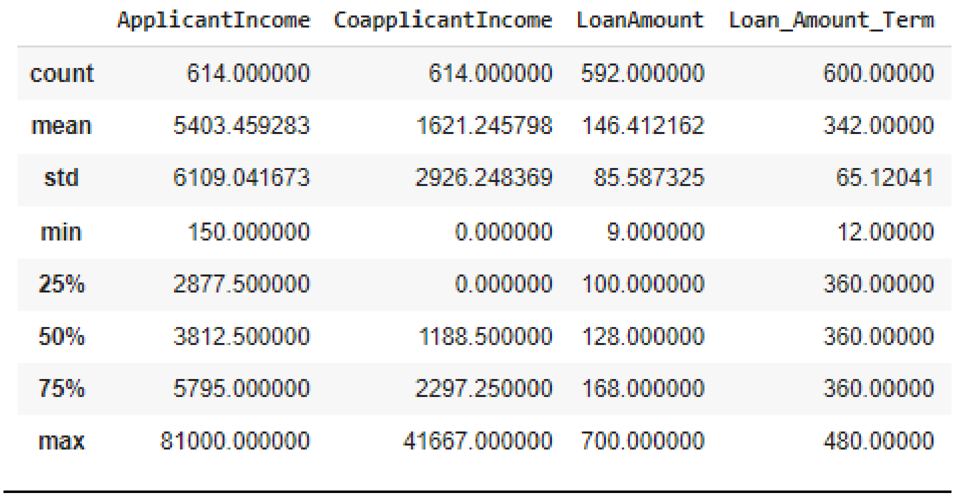


The visualizations show that there are more men than women in our dataset. Also the ratio of applicants who have applied for the loan with respect to being married is almost 2:1. The graphs also show that either the applicants have no dependents or the numbers of dependents have less variation. Most of the applicants are graduated and are working for an employer. The graph reveals that this is an imbalanced dataset with 422 cases in the positive class and 192 cases in the negative class.

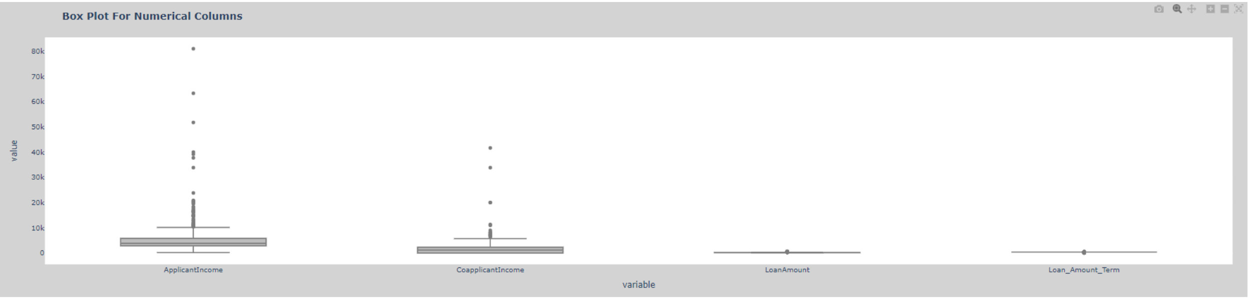
- Numerical Variable Visualizations:



From the above visualizations we can say that most of the applicants have income below 10000 monthly and also many applicants do not have a second income (co-applicants income). Also we can see that the loan amount term that is requested ranges from 0-200 in thousands whereas the term duration for which the loan is approved is mostly 360 months (30 years).



	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term
count	614.000000	614.000000	592.000000	600.00000
mean	5403.459283	1621.245798	146.412162	342.00000
std	6109.041673	2926.248369	85.587325	65.12041
min	150.000000	0.000000	9.000000	12.00000
25%	2877.500000	0.000000	100.000000	360.00000
50%	3812.500000	1188.500000	128.000000	360.00000
75%	5795.000000	2297.250000	168.000000	360.00000
max	81000.000000	41667.000000	700.000000	480.00000

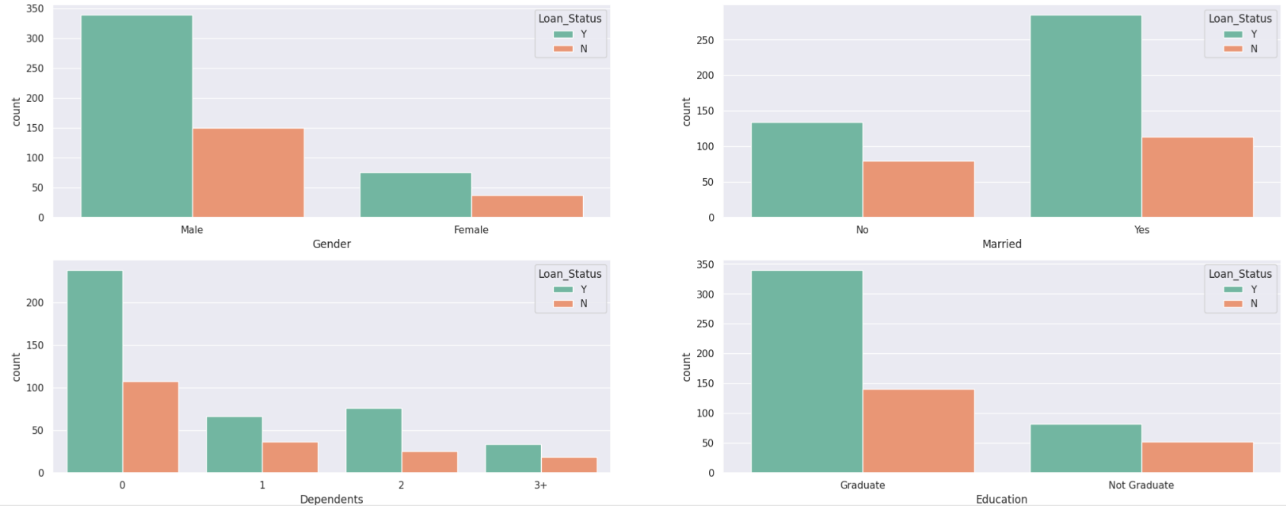


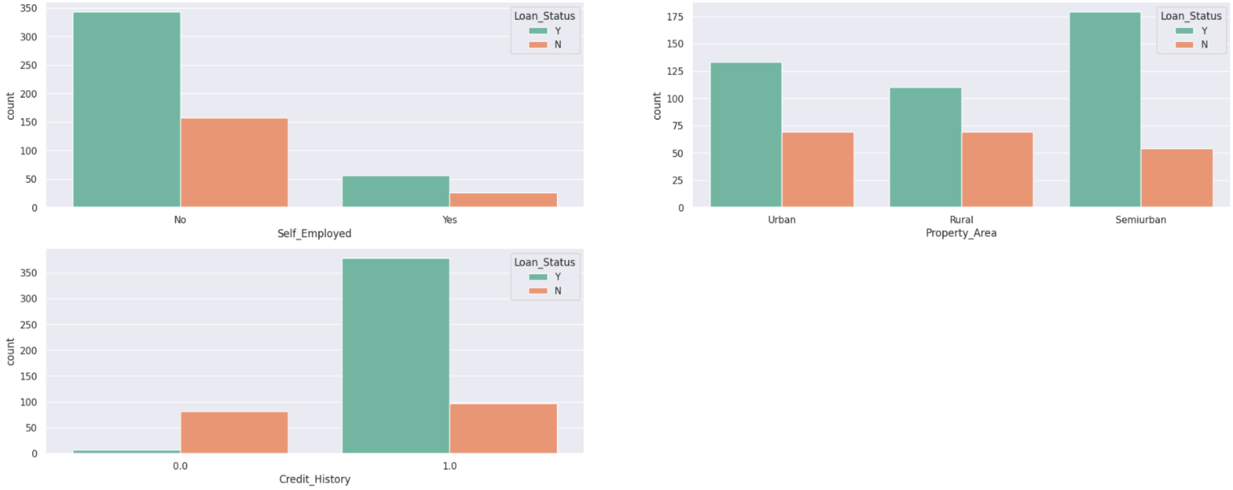
From the above statistics we can see that the data contains some outliers. We can say that because the max for ApplicantIncome is 81000 while the mean is 5403. We know that the mean is sensitive to an outlier but the fact that the mean is so small as compared to the max shows that the max is definitely an outlier which could be a possible case of human input error. Same goes for CoapplicantIncome.

There are no duplicate rows in the dataset however there can be similar data of two different applicants. There are also missing values in the dataset. Here is a table showing variables that have missing data and the percentage of missing data in it:

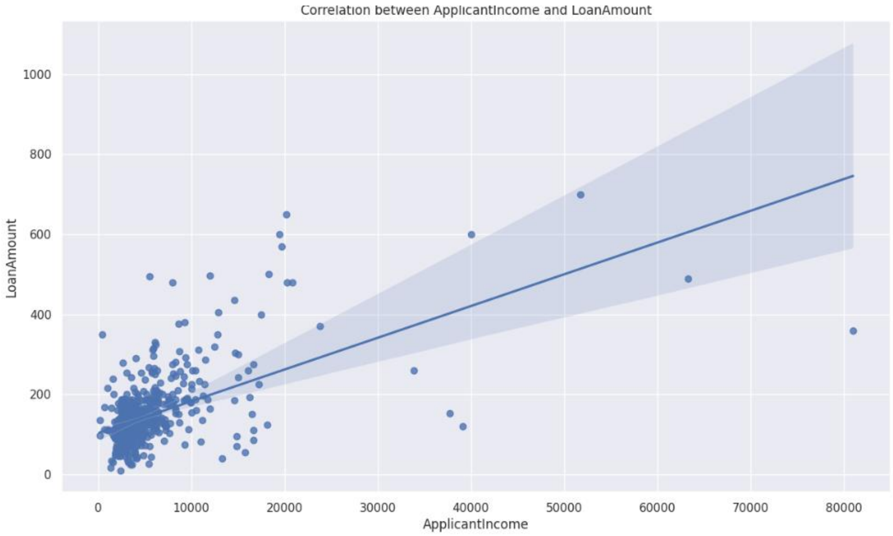
Variable	Percentage of missing data
Gender	2.1%
married	0.5%
Dependents	2.4%
Self-employed	5.2%
Loan_amount	3.6%
Loan_amount_term	2.3%
Credit_History	8.1%

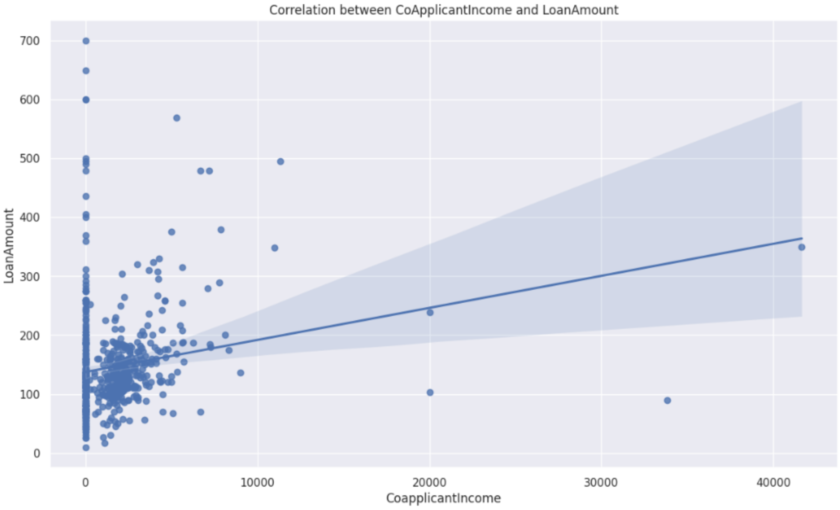
Bivariate Analysis:





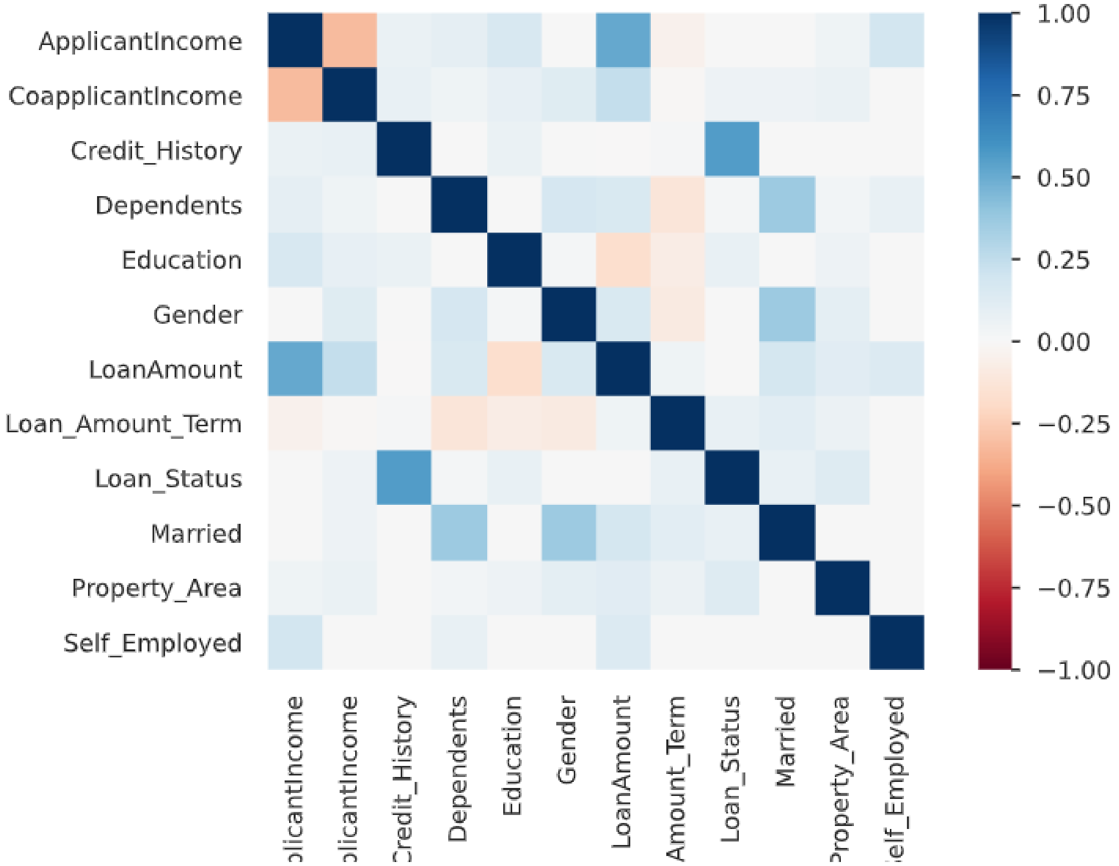
From the above visualization we can say that credit history highly impacts the loan approval status. Also most of the loan approvals granted are given to applicants residing in semi-urban areas.





From these graphs we can say the higher the applicant income, the higher the loan amount.

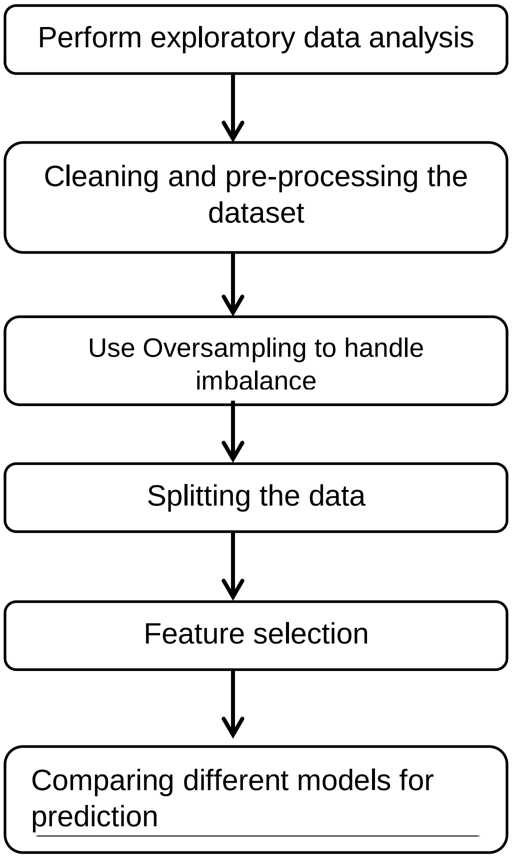
Data Correlation:



Using this initial analysis and the data correlation, these findings can be drawn:

1. Applicant Income is highly correlated to Loan Amount.
2. Credit History is highly correlated to Loan status.

Research Methodology:


Perform exploratory data analysis
Cleaning and pre-processing the dataset
Use Oversampling to handle imbalance
Splitting the data
Feature selection
Comparing different models for prediction

After performing basic exploratory data analysis we will move to cleaning the dataset and deal with missing values. After that mapping the data will be done as some data needs to be converted into binary format. Since the dataset is imbalanced, we will use oversampling to fix this issue. Moving on I will split the data into training and test set by 70:30 and then perform feature selection. After that we will use different machine learning models (logistic regression, KKN and decision trees) to make predictions and then analyze for accuracy and performance.

Github link to current coding file:

https://github.com/Jsadain/Capstone_CIND830/blob/codingfile/Copy_of_Untitled0.ipynb

Research Questions:

- Does having a strong family income increase chances for loan approval?
- Can historical credit information effectively predict loan approval outcomes?
- Identify which features have the most significant impact on loan approval decisions.
- To analyze how well do different machine learning algorithms perform in predicting loan approval outcomes and can ensemble method improve model accuracy and robustness.

References:

Al Mamun M., Farjana A. & Mamun M., “Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis” - Proceedings of the 7th North American International Conference on Industrial Engineering and Operations Management, Orlando, Florida, USA, June 12-14, 2022.
https://ieomsociety.org/proceedings/2022orlando/328.pdf

Hamid A. J. and Ahmed T. M., “Developing Prediction Model of Loan Risk in Banks using Data Mining*”, Machine Learning and Applications: An International Journal (MLAIJ), Vol.3, No.1, pp. 1-9, March 2016.*
https://www.aircconline.com/mlaij/V3N1/3116mlaij01.pdf

Senarathna B.T.N, Weerarathna K.C.M, Wickramarachchi D.S, Jayarathne S.M.P.N, Buddhima Attanyake , “Loan Eligibility Prediction Based on Credit Score and Past History” Published in *International Research Journal of Innovations in Engineering and Technology* - IRJIET, Volume 7, Issue 10, pp 532-542, October 2023.
https://doi.org/10.47001/IRJIET/2023.710070

Pimcharee K. & and Surinta O., “Data Mining Approaches in Personal Loan Approval” - *Engineering access, vol. 8, no. 1, january-june 2022*
https://www.researchgate.net/publication/353314261_Data_Mining_Approaches_in_Personal_Loan_Approval

Vimala S. & Sharmili K.C., “Prediction of Loan Risk using Naive Bayes and Support Vector Machine” - *International Conference on Advancements in Computing Technologies - ICACT 2018, Volume: 4 Issue: 2*
http://www.ijfrcsce.org/download/conferences/ICACT_2018/ICACT_2018_Track/1519367394_23-02-2018.pdf

Khan A., Bhadola E., Kumar A. & Singh N., “Loan approval prediction model a comparative” - *Advances and Applications in Mathematical Sciences Volume 20, Issue 3, January 2021, Pages 427-435*
https://www.mililink.com/upload/article/1759044670aams_vol_203_january_2020_a10_p427-435_afrah_khan_and_nidhi_singh.pdf