

**Sri Lanka Institute of Information Technology**

Machine Learning (IT4060) – Assignment 1

Year4, Semester 2 – 2024

**TOPIC** **- Loan Rejection or Approval Status Prediction**

Group Details –

IT20648404 – B.K Tharuka

IT20658472 – D.I Weerapperuma

**Table of Contents**

1. **Introduction………………………………………………. 3**
2. **Problem……………………………………………………. 3**
3. **Dataset…………………………………………………….. 4**
4. **Methodology……………………………………………… 5**
5. **Data preprocessing………………………………5**
6. **Algorithm Selection……………………………..9**
7. **Model Architecture………………………………9**
8. **Result and discussion………………………………..14**
9. **Summary Of Evaluation Metrics…………..21**
10. **Future Work………………………………………22**
11. **Introduction**

Our topic is loan Rejection or Approval Status Prediction. With our model, we can check whether a loan is being rejected or approved.

For banks and other lenders, predicting whether a loan will be approved or rejected is essential. They want to ensure that the individuals to whom they are giving money have the ability to payback it. It can be difficult to find  out who is a safer bet.

So in there machine learning algorithm approach could be more efficient to find out if loan could be rejected or approved.

In the prediction process we have to consider various factors that affect to that long approved process. Such as Loan status Loan Amount, Property Area, Credit History and other details…

By analyzing those factors using machine learning algorithm could be able to get more accurate decision.

In here getting prediction using four supervised machine learning algorithms and build best model to predict that status.

1. **Problem**

How could be apply machine learning algorithms to get more accurate decision about loan approved or Rejected

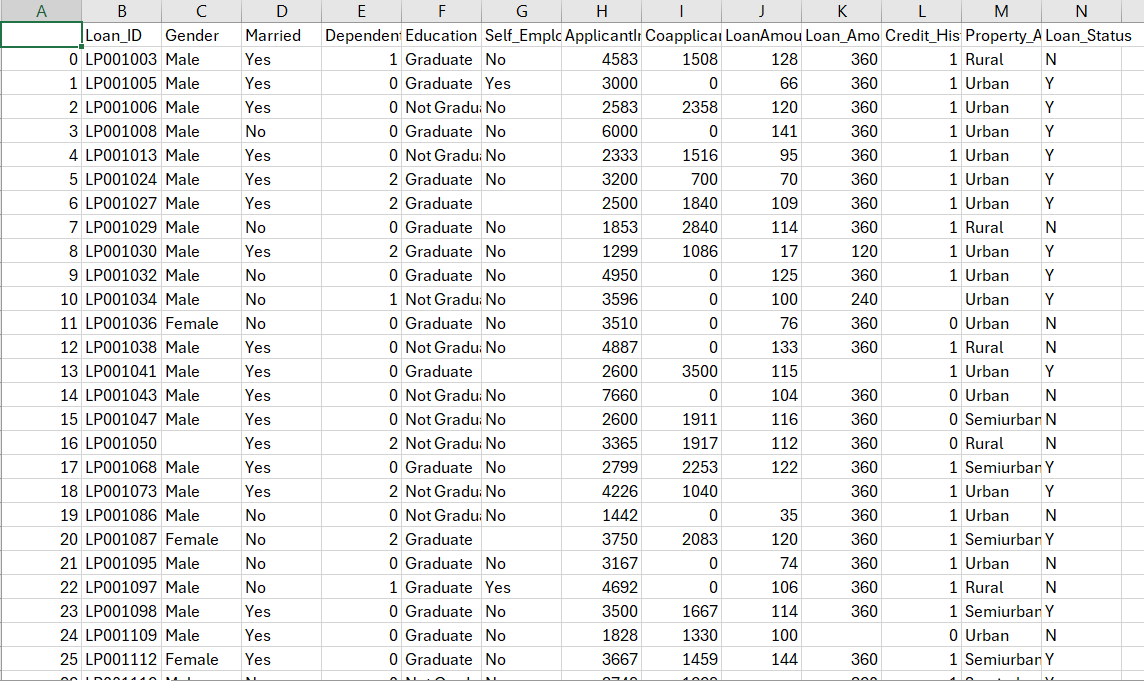
1. **Dataset**

This dataset contains information about past loan applicants, including their income, loan amount, credit history, and other factors relevant to loan approval decisions.

The goal is to build a machine learning model that can analyze this data to predict whether future loan applications should be approved or rejected.

About **loan\_data.csv** file:

* Loan\_ID: A unique loan ID.
* Gender: Either male or female.
* Married: Weather Married(yes) or Not Marttied(No).
* Dependents: Number of persons depending on the client.
* Education: Applicant Education(Graduate or Undergraduate).
* Self\_Employed: Self-employed (Yes/No).
* ApplicantIncome: Applicant income.
* CoapplicantIncome: Co-applicant income.
* LoanAmount: Loan amount in thousands.
* Loan\_Amount\_Term: Terms of the loan in months.
* Credit\_History: Credit history meets guidelines.
* Property\_Area: Applicants are living either Urban, Semi-Urban or Rural.
* Loan\_Status: Loan approved (Y/N).



A screenshot of a computer

Description automatically generated

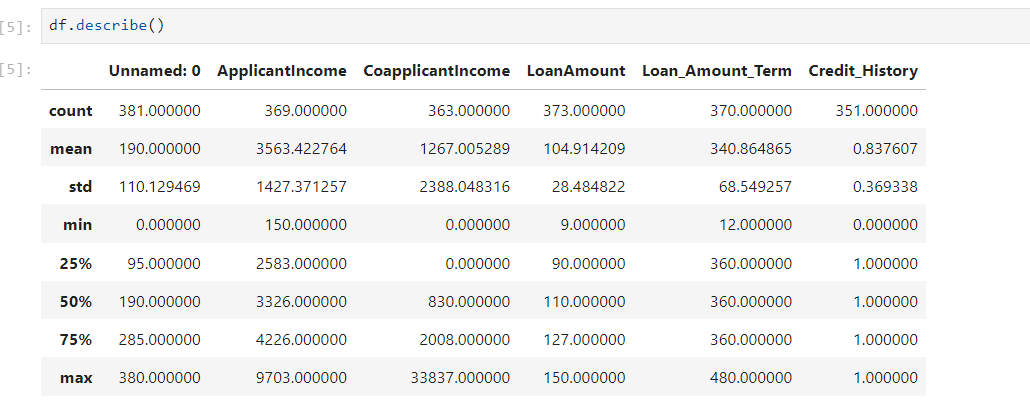
The dataset loan Rejection or Approval Status Prediction. First five records is shown in the table above

1. **Methodology**

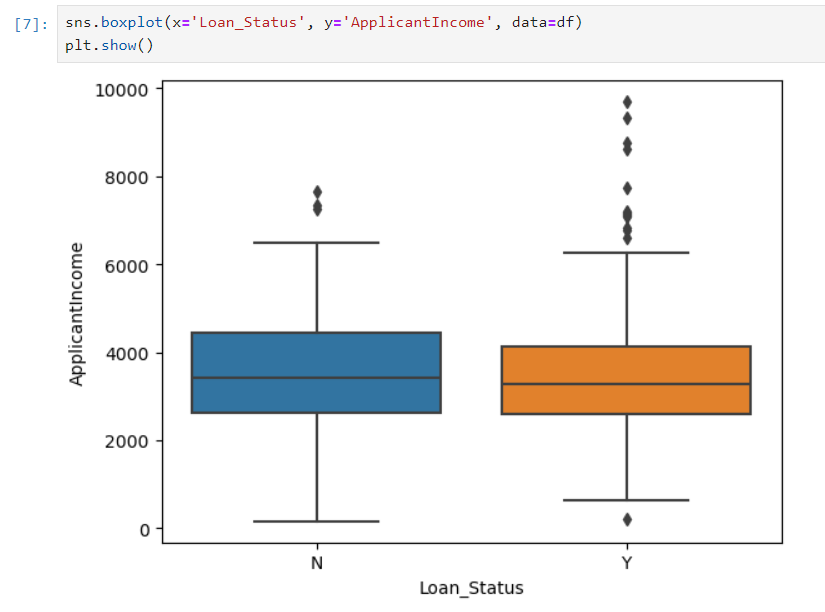
1.Data Preprocessing

A screenshot of a computer

Description automatically generatedCheck About Data Samples



In there had be tried to get some idea about how the dataset scattered around the some data points.



A screenshot of a graph

Description automatically generated

A graph of a number of numbers

Description automatically generated with medium confidence

A graph of different sizes and numbers

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated Numerical Columns missing values filled by mean

Checking for missing values

A screenshot of a computer

Description automatically generated

Drop the missing values in the categorical columns.

A screenshot of a computer

Description automatically generated

Scaling the columns data.

A screenshot of a graph

Description automatically generated

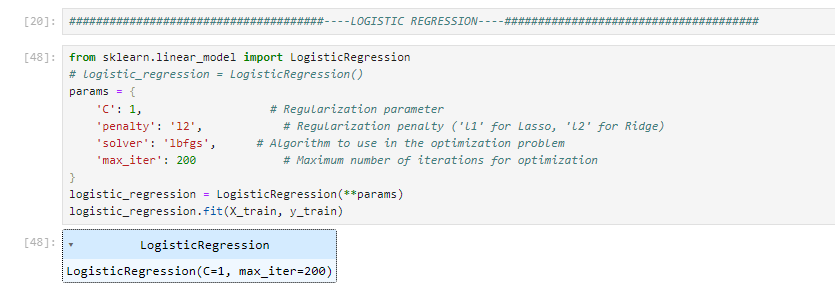
2.Algorithm Selection

In the Algorithm selection phase, we have to consider about what we want to predict through the model and what kind of dataset we have so by considering those area possible to identify in with this data set we have to do classification problem to predict that either loan “Approved” or “Rejected” and we have labeled data set. Afterwards identified, possible to do utilizing supervise learning classification approach to solve that problem. In their chosen the four algorithms to do that,

1. Logistic Regression Algorithm.
2. Decision Tree Algorithm.
3. Random Forest Algorithm.
4. Support Vector Machine Algorithm.

3.Model Architecture

**MODEL 1**



* Algorithm – Logistic Regression
* Regularization Parameter – “C = 1”
* Optimization Algorithm – “lbfgs”
* Iterations - 200

Hyperparameter Tunning

A screenshot of a computer

Description automatically generated

**MODEL 2**

**A screenshot of a computer program

Description automatically generated**

* Algorithm – Decision Tree
* Max\_depth – 5
* Min\_Samples\_Split – 2
* Min\_Samples\_Leaf – 1

Hyperparameter Tunning



**MODEL 3**

**A screenshot of a computer program

Description automatically generated**

* Algorithm – Random Forest
* N\_estimation – 200
* Max\_depth – 10
* Random\_state – 42

Hyperparameter Tunning

A screenshot of a computer program

Description automatically generated

**MODEL 4**

**A screenshot of a computer program

Description automatically generated**

* Algorithm – Support Vector machine
* Kernal - Linear
* Regularization Parameter – “C = 0.1”
* Gaussian Kernal width gamma – Scale

Hyperparameter Tunning

A screenshot of a computer program

Description automatically generated

1. **Result and Discussion**

**MODEL 1**

**A screenshot of a computer program

Description automatically generated**

* Accuracy – 80%
* Precision – 81%
* Recall – 95%
* F1-score – 87%
* The logistic regression model achieves a decent accuracy of approximately 81%.
* With a recall of 95.56%, it's quite high, meaning the model can successfully identify the majority of the positive events (loan approvals) in the dataset.
* With a precision of 81.13%, the model is likewise quite accurate in predicting loan approvals, meaning that 81.13% of the time it is right.
* The F1-score provides a balance between precision and recall, indicating overall good performance

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer program

Description automatically generatedMODEL 2**

* Accuracy – 79%
* Precision – 80%
* Recall – 93%
* F1-score – 86%
* The decision tree model achieves a similar accuracy to logistic regression but with a slightly lower recall.
* With a good recall of  93.33%, the model can successfully identify the majority of positive instances.
* The precision is also decent at 80.77%.
* When compared to logistic regression, the decision tree performs quite well overall, however it may be marginally less accurate.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer screen

Description automatically generated**

**A screenshot of a computer program

Description automatically generatedMODEL 3**

* Accuracy – 85%
* Precision – 86%
* Recall – 95%
* F1-score – 90%
* At about 85.71%, the random forest model has the highest accuracy of all the models.
* With a high recall of 95.56% and the highest precision of 86% among the models, it successfully detects the majority of positive cases while keeping a low false positive rate.
* Overall performance is good, as indicated by the F1-score.

**A screenshot of a computer screen

Description automatically generated**

**A screenshot of a computer program

Description automatically generatedMODEL 4**

* Accuracy – 80%
* Precision – 81%
* Recall – 95%
* F1-score – 87%
* The SVM model performs similarly to logistic regression in terms of F1-score, accuracy, precision, and recall.
* With a precision of 81.13% and a recall of 95.56%, it effectively identifies positive examples.
* Overall, the SVM model performs reasonably well, although it might not outperform the random forest model in terms of accuracy.

**A screen shot of a computer

Description automatically generated**

1.Summary Of Evaluation Metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 | Model 4 |
| True Positive | **43** | **42** | **43** | **43** |
| True Negative | **8** | **8** | **11** | **8** |
| False Positive | **10** | **10** | **7** | **10** |
| False Negative | **2** | **3** | **2** | **2** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 | Model 4 |
| Accuracy | **80%** | **79%** | **85%** | **80%** |
| Precision | **81%** | **80%** | **86%** | **81%** |
| Recall | **95%** | **93%** | **95%** | **95%** |
| F1-score | **87%** | **86%** | **90%** | **87%** |

* Accuracy - Random Forest has the highest accuracy (0.8571), followed by Logistic Regression and Support Vector Machine (both 0.8095), and then Decision Tree (0.7937).
* Precision and Recall - Logistic Regression, Support Vector Machine, and Random Forest have similar precision and recall values, indicating a balanced performance in correctly identifying positive instances and minimizing false positives. Decision Tree slightly lags behind in terms of recall.
* F1-score - Random Forest achieves the highest F1-score (0.9053), followed closely by Logistic Regression and Support vector machine (both 0.8776), and then Decision Tree (0.8660).
* Complexity - Decision Tree is the simplest model among the ones compared, as it doesn't require much hyperparameter tuning and is interpretable. Random Forest is more complex due to its ensemble nature, while Logistic Regression and Support Vector Machine can handle non-linear relationships with appropriate kernel functions.
* Interpretability - Decision Tree is the most interpretable model, as it can be visualized and understood easily. Random Forest is less interpretable due to its ensemble of multiple decision trees. Logistic Regression and Support Vector Machine with non-linear kernels are less interpretable compared to Decision Tree.
* Performance - Random Forest generally outperforms the other models in terms of accuracy and F1-score in this scenario. However, performance can vary depending on the dataset and hyperparameters chosen. Support Vector machine and Logistic Regression perform similarly in this case, while Decision Tree slightly lags behind.

Overall, Random Forest might be the preferred choice if maximizing accuracy is the primary goal, while Decision Tree could be favored for its simplicity and interpretability, especially when the model's transparency is essential. Logistic Regression and SVM are also good options, providing a balance between performance and interpretability.

2. Future Work

* Could be enhance the dataset. such as add data country wise.
* Could be try with different robust models
* Could be utilize efficient and accurate optimization algorithm.