# Loan Approval Prediction

**Group 18**
Kothapalli Bhuvana Chandrika (bk24542)
Spandan Pal (sp54969)
John Hwang (jh75576)
Mansi Sharma (ms89743)
Serena Song (fs9589)

# Table of contents

# 01

# Introduction

# Introduction and Objective

## Introduction

- The dataset has been obtained from **Kaggle**.
- Accurate and faster loan approval processes can
  - Reduce risk of financial institutions for granting loans
  - Streamline operation in the loan evaluation process

## Objectives

- Our goal here is to **predict the loan approval** of an applicant.
- Here, we will try to determine the factors that influence the loan approval status .
- We will also different models to determine the best possible predictions.
- Based on this, we will try to make recommendations for the same.

## Data Snapshot

| loan_id | no_of_de | educatior | self_emp | income_a | loan_amc | loan_terr | cibil_scor | residentia | commerc | luxury_as | bank_ass | loan_status |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | Graduate | No | 9600000 | 29900000 | 12 | 778 | 2400000 | 17600000 | 22700000 | 8000000 | Approved |
| 2 | 0 | Not Grad | Yes | 4100000 | 12200000 | 8 | 417 | 2700000 | 2200000 | 8800000 | 3300000 | Rejected |
| 3 | 3 | Graduate | No | 9100000 | 29700000 | 20 | 506 | 7100000 | 4500000 | 33300000 | 12800000 | Rejected |
| 4 | 3 | Graduate | No | 8200000 | 30700000 | 8 | 467 | 18200000 | 3300000 | 23300000 | 7900000 | Rejected |
| 5 | 5 | Not Grad | Yes | 9800000 | 24200000 | 20 | 382 | 12400000 | 8200000 | 29400000 | 5000000 | Rejected |
| 6 | 0 | Graduate | Yes | 4800000 | 13500000 | 10 | 319 | 6800000 | 8300000 | 13700000 | 5100000 | Rejected |
| 7 | 5 | Graduate | No | 8700000 | 33000000 | 4 | 678 | 22500000 | 14800000 | 29200000 | 4300000 | Approved |
| 8 | 2 | Graduate | Yes | 5700000 | 15000000 | 20 | 382 | 13200000 | 5700000 | 11800000 | 6000000 | Rejected |
| 9 | 0 | Graduate | Yes | 800000 | 2200000 | 20 | 782 | 1300000 | 800000 | 2800000 | 600000 | Approved |
| 10 | 5 | Not Grad | No | 1100000 | 4300000 | 10 | 388 | 3200000 | 1400000 | 3300000 | 1600000 | Rejected |

***Note:***

- CIBIL Score is a 3-digit numeric summary of your credit history and ranges from 300 to 900. The closer your score is to 900, the better your credit rating is.

## Metadata

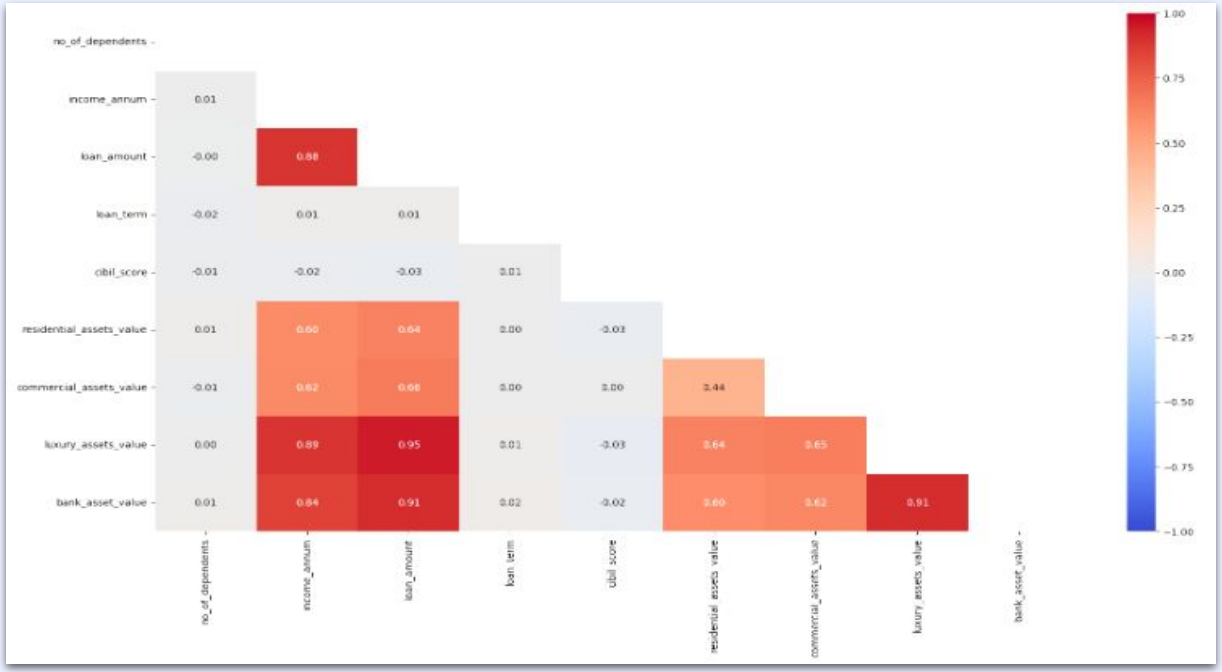
- **Loan id**
- **No of Dependents**
- **Educational Status**
- **Self Employment Status**
- **Income Per Annum**
- **Loan Amount**
- **Loan Term**
- **Cibil Score**
- **Residential Assets Value**
- **Commercial Assets Value**
- **Luxury Assets Value**
- **Bank asset value**
- **Loan Status**

# 02

# Exploratory Data Analysis

# Exploratory Analysis - Numerical Variables

## Correlation Matrix




## Key Observations

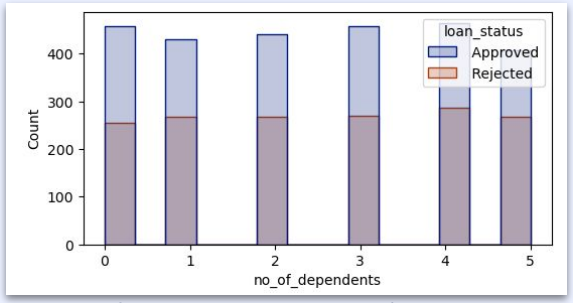
| |
|---|
| **loan_amount + income_annum** |
| **income_annum + luxury_assets_value** |
| **income_annum + bank_asset_value** |
| **loan_amount + luxury_assets_value** |
| **loan_amount + bank_asset_value** |
| **bank_asset_value + luxury_assets_value** |

***Note:***

- Due to the skewed distribution of variables like loan_amount, commercial_asset_value etc, the log transformation of the variables were considered here.
- Point to note: We do not see any correlation between the CIBIL score and any other variable

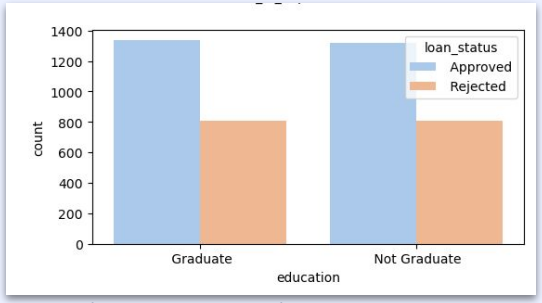
# Exploratory Analysis - Predictors vs Loan Status

## No of dependents vs Loan status



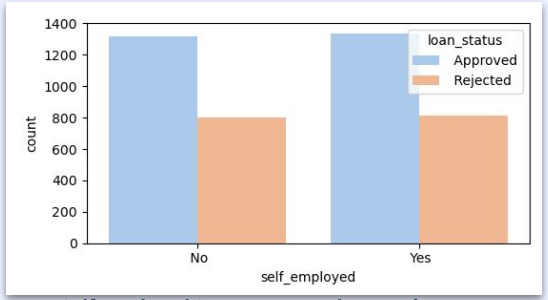

*No of dependents seem to have no impact on loan status*

## Education vs Loan status



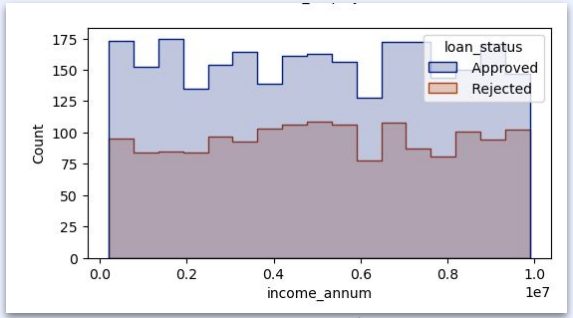

*Education seem to have no impact on loan status*

## Self Employed vs Loan status



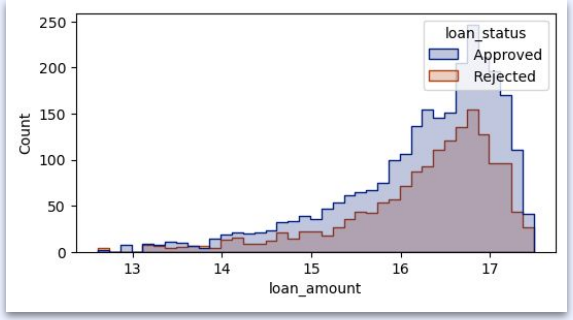

*Self Employed Status seem to have no impact on loan status*

## Income Annum vs Loan status



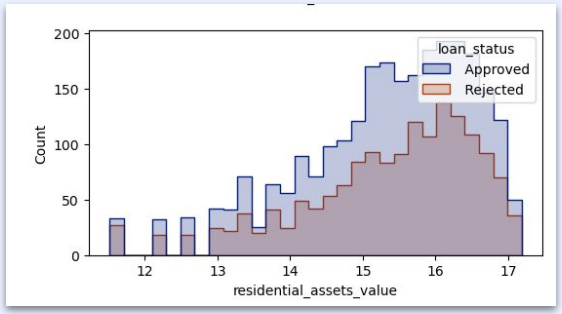

*Annual Income seem to have no impact on loan status*

## Loan Amount vs Loan status




*Loan Amount seem to have no impact on loan status*

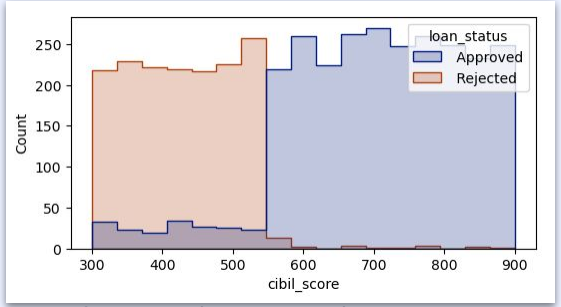
## Residential Asset vs Loan status




*Residential Asset Value seem to have no impact on loan status*

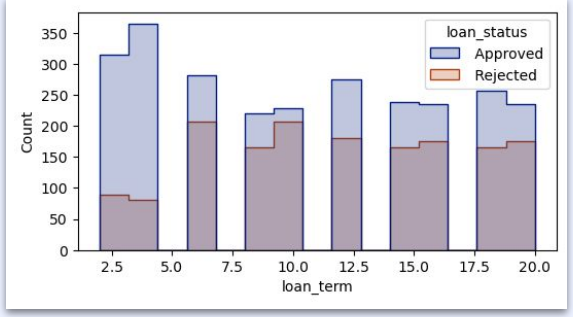
# Exploratory Analysis - Predictors vs Loan Status (Contd)

## CIBIL Score vs Loan status



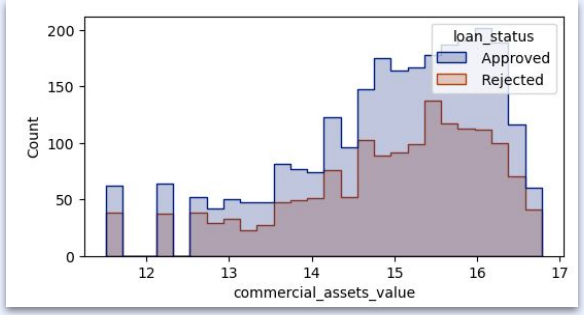

*If CIBIL Score is above a certain value, then chances of loan approval increases significantly*

## Loan Term vs Loan status



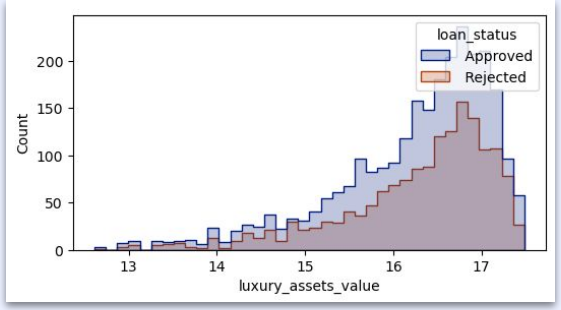

*As Loan Term increases, the chances of Loan Rejection increases*

## Commercial Asset vs Loan status



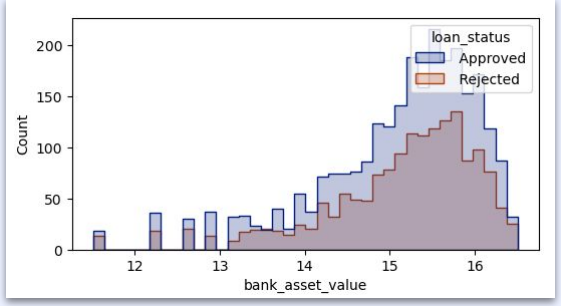

*Commercial Asset Value seem to have no impact on loan status*

## Luxury Asset vs Loan status




*Luxury Assets Value seem to have no impact on loan status*

## Bank Asset vs Loan status




*Bank Asset Value seem to have no impact on loan status*

## Key Takeaway Point

***Loan Term seems to have an impact on the loan approval status.***

***CIBIL Score seems to have an impact on the approval status***

***Note:***
Looking at the distribution of Approval and Rejects in the dataset, the baseline accuracy for anu model should be 62%.

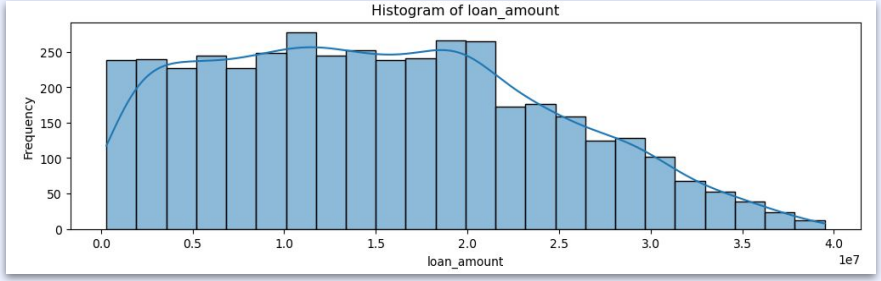
# 03

# Feature Engineering

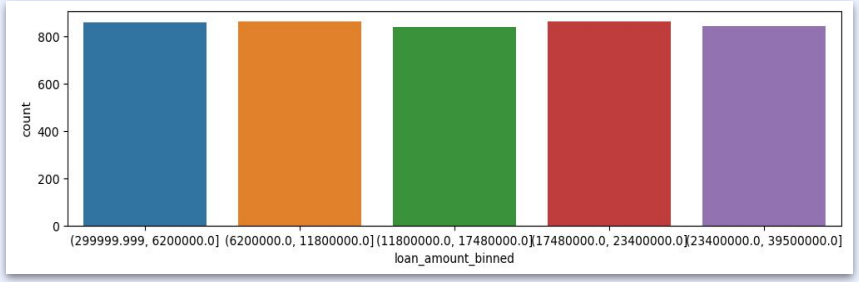
# Feature Engineering

## 1. Bucketing of Continuous Variables

**Before Transformation**




**After Transformation**




**For the variables such as Loan Amount, Residential Assets Value, Commercial Assets Value, the distribution is right skewed. So, we segment the variables into bins for future analysis.**

## 2. Removal of Unnecessary Variables

| | |
|---|---|
| **Loan ID** | **No business value since system field** |
| **Luxury Assets Value, Bank Asset Value** | **Fields have a high correlation with other fields - Loan Amount and Income Amount** |

# Feature Engineering

## 3. One Hot Encoding

### Before Transformation

| commercial_assets_value_binned |
|---|
| (8800000.0, 19400000.0] |
| (1000000.0, 2600000.0] |
| (2600000.0, 5000000.0] |
| (2600000.0, 5000000.0] |
| (5000000.0, 8800000.0] |

### After Transformation

| commercial_assets_value_binned_(1000000.0, 2600000.0] | commercial_assets_value_binned_(2600000.0, 5000000.0] | commercial_assets_value_binned_(5000000.0, 8800000.0] | commercial_assets_value_binned_(8800000.0, 19400000.0] |
|---|---|---|---|
| 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 |

For variables like Education, Self Employed, commercial assets value binned etc, each categorical value is converted into a new categorical column with the binary value of a 0 or 1.

## 4. Variable Scaling using Z-scale Normalization

Numerical Variables such as Loan Amount are standardized to have a zero mean and unit standard deviation

The numerical values are scaled as per z = ( x - mean) / standard_deviation. This is done to bring all the features to a similar scale, so that they contribute equally to the analysis.

# 04

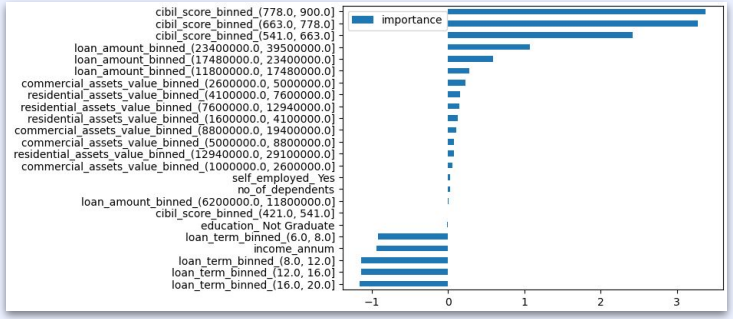
# Modeling

# Modeling

## Logistic Regression

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.96 | 0.93 | 0.95 | 798 |
| 1 | 0.89 | 0.94 | 0.92 | 483 |
| accuracy | | | 0.93 | 1281 |
| macro avg | 0.93 | 0.94 | 0.93 | 1281 |
| weighted avg | 0.94 | 0.93 | 0.93 | 1281 |

- The AUC for Logistic Regression is : **0.935300**
- The time to run the Logistic Regression model is 0.137 seconds.

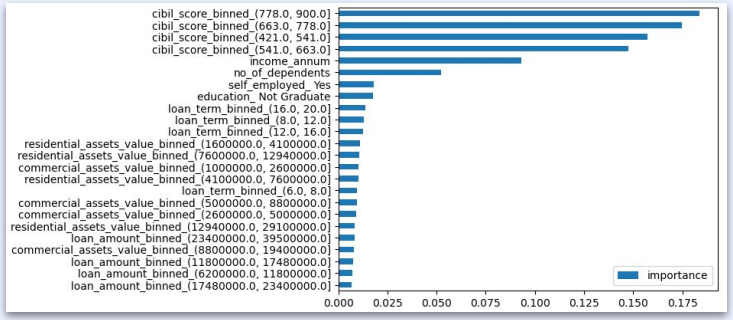



- CIBIL score of 540+ is most significant in the approval of the loan.
- Loan Term negatively affects the Loan Approval Status. This is in alignment with our exploratory analysis.

## Decision Tree

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.94 | 0.91 | 483 |
| 1 | 0.96 | 0.93 | 0.94 | 798 |
| accuracy | | | 0.93 | 1281 |
| macro avg | 0.92 | 0.93 | 0.93 | 1281 |
| weighted avg | 0.93 | 0.93 | 0.93 | 1281 |

- The AUC for Decision Tree is : **0.93404**
- The time to run the Decision Tree model is 0.13869 seconds.



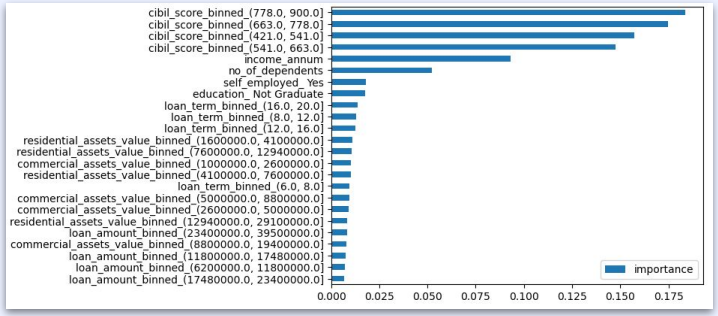

- Decision Tree also lists Annual Income, no_of_dependents. This is not aligned with the initial analysis.

# Modeling

## Random Forest

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.93 | 0.92 | 483 |
| 1 | 0.96 | 0.94 | 0.95 | 798 |
| accuracy | | | 0.94 | 1281 |
| macro avg | 0.93 | 0.93 | 0.93 | 1281 |
| weighted avg | 0.94 | 0.94 | 0.94 | 1281 |

- The AUC for Random Forest is : **0.93451**
- The time to run the Random Forest model is 0.699 seconds.

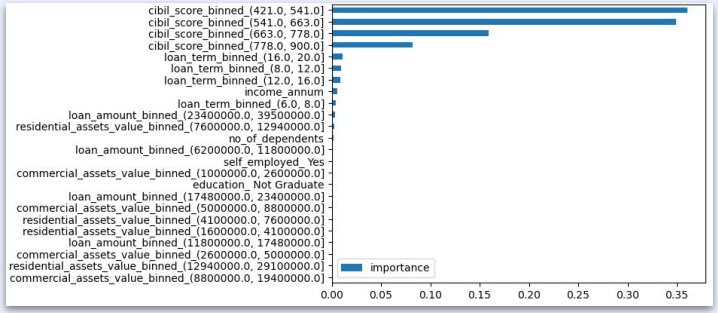



Similar to Decision Tree, Random Forest also considers annual income, no of dependents, self employed, education along with CIBIL Score and Loan Term as the most significant indicators.

## Gradient Boosting

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.95 | 0.92 | 483 |
| 1 | 0.97 | 0.93 | 0.95 | 798 |
| accuracy | | | 0.94 | 1281 |
| macro avg | 0.93 | 0.94 | 0.94 | 1281 |
| weighted avg | 0.94 | 0.94 | 0.94 | 1281 |

- The AUC for Gradient Boosting is : **0.94028**
- The time to run the Gradient Boosting model is 1.880 seconds.




Gradient Boosting only considers CIBIL Score and Loan Term as the significant features.

# Modeling

## K Nearest Neighbours

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.91 | 0.85 | 0.88 | 483 |
| 1 | 0.91 | 0.95 | 0.93 | 798 |
| accuracy | | | 0.91 | 1281 |
| macro avg | 0.91 | 0.90 | 0.90 | 1281 |
| weighted avg | 0.91 | 0.91 | 0.91 | 1281 |

- The AUC for KNN is : **0.89977**
- The time to run the KNN model is 0.179 seconds.

## Neural Network

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.91 | 0.93 | 0.92 | 483 |
| 1 | 0.96 | 0.94 | 0.95 | 798 |
| accuracy | | | 0.94 | 1281 |
| macro avg | 0.93 | 0.94 | 0.93 | 1281 |
| weighted avg | 0.94 | 0.94 | 0.94 | 1281 |

- The AUC for NN is : **0.98242**
- The time to run the NN model is 2.88 seconds.

## Key Takeaway Points

Neural Network has the highest AUC at 98%.
However, it also has the greatest run-time compared to all the models.

CIBIL Score and Loan Term are the most important predictors. Following this, features like education, self employed, loan amount and no of dependents also have some significance.

05

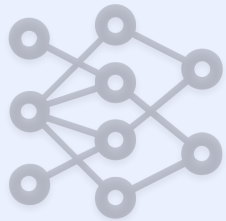
# Conclusion

## Conclusion

| Model Type | Accuracy (%) | F-1 Score | | Run Time (Secs) |
|---|---|---|---|---|
| | | Approval | Rejection | |
| Logistic Regression | 93 | 0.92 | 0.95 | 0.137 |
| Decision Tree | 93 | 0.94 | 0.91 | 0.139 |
| Random Forest | 94 | 0.95 | 0.92 | 0.699 |
| Gradient Boosting | 94 | 0.95 | 0.92 | 1.880 |
| KNN | 91 | 0.93 | 0.88 | 0.179 |
| Neural Network | **98** | **0.95** | **0.92** | **2.88** |

# Observations and Recommendations

| | |
|---|---|
| 01 | **CIBIL Score** or the Credit Score is the most important predictor for the loan approval status. It has been seen that a CIBIL score > 540 significantly increases the chances of loan approval. This is in alignment with EDA. |
| 02 | Loan Term is also an important predictor for loan approval status. It has a negative relationship with the chances of loan approval i.e. it the loan term increases then the chances of approval decreases. This is in alignment with EDA. |
| 03 | Factors like education, self employed,were initially not considered to be significant during Exploratory Data Analysis. But, the machine learning algorithm reveals that these factors are significant. This is most probably due to the interaction with other variables. |
| 04 | **Neural Network** gives the best accuracy with this data set.<br>It has also been tested that using Neural Network with a selected list of only significant predictors gives an AUC score of **97.8%**.<br>If all the predictors had been selected the AUC score was seen to be 98.2%.<br>Thus, there is no major drop in accuracy if we use only the significant indicators. |



# Thanks!

**Do you have any questions?**