# Loan-Prediction

Loan prediction by Tejaswini Shetty and Riyanna Lynn Dsouza

**Introduction**

Loans are the core business of banks.The main profit comes directly from loan's interest.The loan companies grant a loan after an intensive process of verification and validation.

However they still don't have assurance if the applicant is able to repay the loan with no difficulties.so we have build a predictive model to predict if an applicant is able to repay the loan amount or not.

We have made use of jupyter notebook for the implementation and we have made use of various algorithms along with graphs to so as to make the predicted result more accurate.

**Table of Contents**

|  |  |
| --- | --- |
| **Sl.no** | **topic** |
| **1** | Libraries used |
| **2** | Getting the system ready and loading the data |
| **3** | Preprocessing the data |
| **4** | Exploratory Data Analysis |
| **5** | Model Building |
| **i** | Decision Tree |
| **ii** | Random Forest |
| **iii** | Logistic Regression |
| **6** | Confusion Matrix |

**1. Libraries Used**

* **Numpy -** NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.
* **Pandas** - Pandas is mainly used for data analysis. Pandas allows importing data from various file formats such as comma-separated values, JSON, SQL, Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features.
* **matplotlib.pyplot** - Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy
* **seaborn** - is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

**2. Getting the system ready and loading the data**

We have used python programming language along with the following list of libraries

->pandas

->Seaborn

->sklearn

->matplotlib.pyplot

Data

We have two CSV files : Test,loan\_data\_set,sample file

Test file contains all the independent variables,but not the target variable.

loan\_data\_set(train file) which contains the data used for prediction

sample file contains the format in which we have to submit out prediction

Reading data

df = pd.read\_csv("loan\_dataa\_set.csv")

**2.Preprocessing the data**

Here we look for any null values or Nan(not a number) in the data set by using df.isnull().sum() and then by using the dropna() function we can drop the rows having null values

Nan can be replaced either by mean or median of that repective column.once all the null values are eliminated we can go to the next step that is understanding the data

**3. Exploratory Data Analysis**

We can understand the data better by ploting various graphs between the different attributes.

Graphs are of different types:

1. Bar graph

2.Histogram

3.Line Graph

We have made use of bar graph for finding the dependencies between the different attributes

**4.Model Building**

**Decision Tree** - The first algorithm we implemented was decision tree . decision tree is a supervised learning technique that can be used for both classification and regression problems

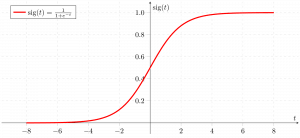
but mostly preffered for solving classification problem.It is a tree structured classifier.

the accuracy percentage we got was 71.16564417

**Random Forest** - Random Forest is a flexible,easy to use machine learning algorithm.It is also one of the most used algorithms because of its simplicity and diversity.Random Forest is a supervised learning algorithm.The 'forest' it builds is an ensemble of decision trees ,usually trained with the bagging method.

The accuracy percentage we got was 77.30061496

**Logistic Regression** - Logistic Regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable,although many more complex extensions exits

The accuracy percentage we got was 78.52760761

**5. Confusion Matrix**

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.it is relatively simple to understand.It is an NxN matrix used for evaluating the performance of a classification model.where N is the target values with those predicted by the machine learning model.

**Conclusion**

Out of the three algorithms we have used to train the data set we have noticed that logistic regression has given the best accuracy percentage compared to random forest and decision tree.