Project

on

Prediction on Approval of Bank Loan

Made By:-

Name- Pritish Sheth

Email-pritish.sheth@gmail.com

Date – 22/10/2020

Introduction

The Project “Prediction of Approval of Bank Loan” is a Machine Learning Project which uses different algorithms to predict whether the user will get Loan on basis of the data he provides. This project can be used by the banking sectors to complete the Loan Approval tasks at a faster rate as it will not require any Human Efforts of checking and comparing all the data given by the user and whether the given data is satisfying the conditions for Loan Approval. This Project can help banks to save their time and increase the efficiency of the employees.

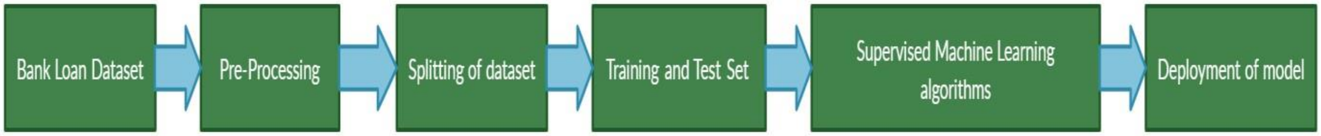
This Project has wide use in areas of Banking Sectors or the companies which provide loans to people. This Project can also help in creating some websites which can help people get to know whether they can get a Loan by just sitting at their places and not wasting their precious time waiting in queue at the banks.

Understand and Define the Problem

In this project the main motive is to Predict whether the person will be Approved for getting the Loan or not. The Problem statement was that we had to use Machine Learning algorithm to predict the eligibility of a person for Loan. (Scope).Dataset of 13 features which had total 614 data points in it was provided which contained customers data like Loan ID, ID, Gender, Married, Education, Self Employed etc. which was used to train our model and predict when new data is provided.

In this Project different libraries like Pandas , Matplotlib , Seaborn are used to work with given data and also to plot various graphs and heat maps. We have also used SKLEARN Library for using different ML algorithms like Logistic Regression, Decision Trees, Random Forest, KNN, and Support Vector Machine (SVM) to predict the bank loan status. We have used this 5 models and found out which model has the best accuracy on predicting the status. Limitations of this project are it would just give an idea about the approval, it would not be totally correct because accuracy is not 100%.

Steps Used in this Project



Dataset Preparation and Preprocessing

This is the stage in the project where we need to collect, clean, preprocess, the data using different techniques to make our model more efficient and accurate.

Data collection

There are various ways and resources to collect data for project. The data used in this project was in format of CSV, but it can be present in any other forms also.

Data Visualization

The data used in ML projects are huge, so it makes difficult to analyze statistically. So Data Visualization is used to visualize the given data and analyze it. Here libraries like Matplotlib, Seaborn are used to visualize the data in different forms.

Labeling

Labeling is an important part of data preparation as if labeling is not there its not possible for ML algorithms to use Supervised Classification method to classify data and predict. Labeled data gets classified easily using ML algorithms.

Data Selection

There are various columns in given data so we need to select specific data which is required for our project and separate them, by dropping the non useful data.

Data Preprocessing

The purpose of preprocessing is to convert raw data into a form that is useful in training and testing the model. The structured and clean data produces more precise results. In short, good quality data when fed to the model, it produces better results.

The Preprocessing technique includes data formatting, cleaning, and sampling techniques.

***Data Formatting*:** The data may come from different sources. Hence, it needs to be standardized.

***Data cleaning*:** In this process, the noise in the data is removed and inconsistencies are fixed. The missing values in the data are filled with mean attributes. The outliers in the data are either removed or corrected. And if the data is categorical then the missing data is filled by mode.

***Data anonymization*:** In ML projects, privacy is important. If the data contains sensitive private information, the concerned attribute needs to be removed or anonymized. It would be better if that private data is removed.

Data Transformation

In this stage, the data is transformed into the form which is appropriate for machine learning. The scaling and normalization is usually used to transform the data.

Scaling: The different attributes in the dataset may have different ranges i.e. data values may vary over different values. Scaling is used to correct this problem.

Feature Extraction: Some of the existing features are combined to create new features which are useful for ML modeling.

Dataset Splitting

The given dataset is split into two parts: training, testing. The ration of training and testing sets is typically 80 to 20 percent. The training part used to train our model and testing part is used to test our model and find accuracy so that we can do some changes if accuracy is less.

Model Training

In this stage, the training data is fed to the ML algorithm to build and train our model. The purpose of training is to develop a model.

Model Testing and Evaluation

The goal of this step is to develop the simplest, reliable and efficient model. This requires model tuning. We can use different type of models to find out which gives the best output and accuracy and use that model for the project.

Improving Predictions with Ensemble Methods

In most cases, the data scientists create and train one or more models. Then they select the best performing one.

Like Random Forest, data scientists also like to combine (ensemble) various models for prediction. Ensemble methods provide better results.

There are three ways to combine models:

***Stacking*:** In this case, usually used to combine models of different types. The aim of this method is to reduce generalization error.

***Bagging*:** In this case, the models of the same type are combined in sequential manner. The training dataset is split into subsets. Then the models are trained on each of these subsets.
Ultimately, the prediction is based on combining the result using mean or majority voting. The bagging reduces model over fitting.

Boosting: In this case, the data scientists use subsets of data to train moderately performing models. The prediction is based on the majority voting principle. Every next model is trained on a subset received from the performance of the previous model (particularly the emphasis is put on misclassified data points).

Deployment of Model

When the reliable model is selected and validated, the model is put into production. Model Deployment means putting the model in use.

In most cases, the deployment is done by translating the Model written in Python language to another languages like Java, C, C++, PHP etc. Then the Alpha and Beta testing is done.

There are various ways of deploying the model. The deployment way in which this project should be used is:

Stream Learning based Deployment: In this type, the model works dynamically. This means that the ML model keeps on improving and updating by itself through the continuously changing data fed to it. This will help in continuously train our model by feeding data which will help in increasing the accuracy of our model and also increase the efficiency at every step.

Conclusion and Further Development

In this project it was observed that KNN model gives better accuracy than other models. Also this project helped me understand the Confusion Matrix in a better way. This project helped to learn how to plot different data on different graphs and get information out of those graphs. It helped me to use different types of libraries, so that I work upon the data given and increase the efficiency of the model.

In future I would like bring some of changes which can help me to make my project work more practical and better. I would like to train my model using more huge data so that it can predict more accurately and also learn other different models of ML which can also increase the efficiency of my project.