Loan Application status Prediction

Banking sector is growing day by day, started from the bartering system to exchange of money; we all see evolution of banking world or financial world. Banking is an industry that handles cash, credits, and other financial transactions. Banks provide a safe place to store extra cash and credit. They offer savings accounts, certificates of deposit, and checking accounts. Banks use these deposits to make loans.

This definition points to the three primary activities of a commercial bank which distinguish it from the other financial institutions.

=> Deposit taking

=> lending

=> Funds Remittance

We know that, RBI regulates all the banking operations in India. As banks earn their profit on loans i.e., lending money by charging a huge rate of interest on their customers. In past era there were no kyc of customers, no rules and no regulations and this the main reason when fraud started in banks in terms of loans, insurance and many other services too. In this article we are going to talk about loans as to whom we can give loan or not, is the major decision needs to be taken by banks. The main reason behind it that banks faced and still facing huge losses due to these mistakes, and it directly affects GDP of the country and it applies for all other countries too.

From the past as exchange of money operation took place, the fraud rate in terms of loan application is very huge due to unutilization of technology and unawareness of the consequences stated by the government. Let's speak about some of the main reason for loan rejection of the applicants as:

- ⇨ Low credits score
- ⇨ Lower income
- ⇨ Inaccurate details in application form
- ⇨ Job instability
- ⇨ Too many pending loans
- ⇨ Not eligible for loans

There are many more reason behind it.

Different types of loans are:

- ⇨ Home loan
- ⇨ Vehicle loan
- ⇨ Personal loan
- ⇨ Gold loan
- ⇨ Loan against properties, share, atc
- ⇨ Business loan

- ⇨ And many more….

When we talk about loans, most important factor while taking loan is your CIBIL score. Credit information bureau (India) Limited (CIBIL) is a credit bureau or credit information company, engaged in maintaining the records of all the credit-related activities of companies as well as individuals, including credit cards and loans. It ranges from 300 to 900, if your score is in the range is 300-550, your loan application will get rejected. If it is between 550-700, there are some chances you will get your loan, and if it range 700-900, chances of loan application will get approved is very high. So as a growing Data scientist, it's our duty to help the banks to showcase your skills as Artificial intelligence and machine learnings plays an important role in BFSI sector.

Role of machine learning in loans:

"Machine learning has flipped the script on traditional lending, allowing for more accurate and faster decisions by shifting traditional decision-making from analysis of individuals to analysis of trends and patterns. The result for lenders? More repeat business, and lower operational costs,"

In finance, machine learning algorithms are used to detect fraud, automate trading activities and provide financial advisory service to investors. Machine learning can analyze millions of data sets within a short time to improve the outcomes without being explicitly programmed. The dataset on which we worked consist of 614 rows and 13 columns with following details:-

- ⇨ Loan_ID
- ⇨ Gender
- ⇨ Married
- ⇨ Dependents
- ⇨ Education
- ⇨ Self_Employed
- ⇨ ApplicantIncome
- ⇨ CoapplicantIncome
- ⇨ Loan_Amount
- ⇨
- ⇨ Loan_Amount_term
- ⇨ Credit History
- ⇨ Property_Area
- ⇨ Loan Status

Analysis of problem statement:-

(https://github.com/dsrscientist/DSData/blob/master/loan_prediction.csv)

"so, in the dataset we have details of different customers and we have to predict that, the loan of the applicant will get approved or not?"

To begin with data analysis, we imported required libraries as matplotlib.pyplot and seaborn which gives us graphical representation and relationship of different columns present in the dataset. Some important observation from the analysis of the data are:

- ⇨ There are missing value in the dataset, we replace the missing value in the dataset by mean of the columns for some and by 0 for some columns, as missing values affect model accuracy
- ⇨ Dataset is the combination of string and integer data type.
- ⇨ Loan is approved for male category is more as compare to female category. And the same for unapproved loan
- ⇨ Loan approval rate is high as well as low for married category
- ⇨ Loan approval rate is high as well as low for graduated category
- ⇨ Loan approval and disapproval is high and low respectively for people who are not self employed.
- ⇨ Percentage of loan approval is high as well as low for people with o dependents.
- ⇨ For credits history as 1, loan gets approved and percentage is also high, although it is nearly same for credit rating as 0 and 1 for loan disapproval.
- ⇨ Loan approval rate is highest for semi-urban category as well as the lowest and disapproval.
- ⇨ While predicting any model, one thing always need to take care of is handling class imbalance problem, but in the dataset class is properly balanced.
- ⇨ Main focus on statistical summary as it indicates us about mean, median, maximum value, minimum value, and standard deviation for each column individually and we get insight about missing values, outliers, skewness for individual column.
- ⇨ Checking correlation of all columns with each other by using heatmap for graphical representation of the dataset
- ⇨ Outliers are removed from the dataset by using zscore technique by keeping threshold as 3
- ⇨ We lost 6% of the data after removing of outliers and we saved dataset into new data frame as df_new.
- ⇨ To check the distribution of the data, we plot distplot as for most of the columns, we saw skewness.

"we performed EDA of the dataset by using different plot as countplot, heatmap, boxplot and to check data distribution we used distplot".

Preprocessing of data:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real – world data is often incomplete, inconsistence, and/or lacking in certain behaviour or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

Here, in the dataset maximum columns are of string data type, as EDA were already done, we converted the string data type into integer data type as machine learning models can only work with numeric data. So for that, we used label encoder to encode the required data and then by using standard scaler operation, data scaling were done to get great model accuracy. Dataset is related to real world, so we used standard scaler.

Building Machine Learning Models:

A machine learning model is built by learning and generalizing from training data, then applying that acquired knowledge to new data it has never seen before to make predictions and fulfill its purpose. Lack of data will prevent you from building the model, and access to data isn't enough.

Following Models and libraries were imported:

=> from sklearn.linear_model import LogisticRegression

=> from sklearn.metrics import accuracy_score

=> from sklearn.svm import SVC

=> from sklearn.tree import DecisionTreeClassifier

=> from sklearn.ensemble import RandomForestClassifier

=> from sklearn.model_selection import train_test_split

=> from sklearn.model_selection import GridSearchCV

=> from sklearn.model_selection import cross_val_score

=> from sklearn.metrics import
accuracy_score,confusion_matrix,classification_report

Model Accuracies for different machine learning models are as followed:

- ⇨ Logistic Regression – 79.6%
- ⇨ Support vector classifier – 79.64%

- ⇨ Decision tree classifier – 63.41%
- ⇨ Random forest classifier – 76.42%

Always take a note of one thing, we are getting above model accuracies due to overfitting or underfitting of the data which in future will affect our model performance. So, to overcome this problem, always check the cross validation score as it help us to choose perfect model for prediction purpose. After applying cross validation, we came to know that LogisticRegression model performing best among other models as the difference between accuracy and cross val score is less for Logistic Regression.

Here, while building any model for prediction, always go for Hyper parameter tuning as it will increase our model accuracies by using the best parameters(max-iter, penalty and solver).

After applying hyper parameter tuning with the best parameters, we got the final accuracy of the model as, 79.67%.

Importing joblib to dump the model and then to load the SVC model as follows,

joblib.dump(Final_mod,"Loan_Prediction.pkl")

LG_mod=joblib.load("Loan_Prediction.pkl").

An article by:

Name :- Durgesh rana

Email:- durgeshrana480@gmail.com

Batch no: - 0422

Datatrained