CAPSTONE PROJECT

LOAN ELIGIBILTY PREDICTION

MACHINE LEARNING MODEL IMPLEMENTATION

THARIKA HANSINI[REG NO: DSA_0302]
ML FOUNDATION COURSE
Conducted by Dialog Data Science Academy

Table of Contents

1. Introduction

One of the core business of a financial institution is money laundering. With a smooth process of document verification and the eligibility evaluation, particular financial institution may tend to grant a loan to their customers. Since there is a higher risk of defaulting the loan amount by particular customers, sometimes it should require to properly capture the most eligible customer group with minimum risk when granting a loan facility. Therefore, the goal of implementing below predictive model is to identify the customers who are capable of repaying the loan without any difficulties. Once implemented this predictive model, it would facilitate to predict whether a particular loan applicant is able to repay the loan or not. Therefore, this model implementation with highest accuracy will lead to provide much advantage to the Financial institution to minimize their credit risk and lend the money for most suitable applicant cluster.

Python Jupyter Notebook has been used to explore the given dataset and proceed the machine learning classification model implementation

2. Data

The Dream Housing Finance Company who deals with housing loans provided their dataset and that csv file has been obtained from the Kaggle Dataset repository. Below link can be used to obtain the relevant “Loan Eligible Dataset”.

https://www.kaggle.com/datasets/vikasukani/loan-eligible-dataset

The dataset contains the Loan Status (Target Variable) of the existing customer base with the feature variables which are associated in the online application form.

Below embedded csv file would be the dataset which has been used for the model implementation.

loan_data_set.csv

There are 12 feature variables and one target variable(Loan Status) in the above dataset.

Stated below the columns contains in the dataset.

Variable	Description	Variable Type	Data Type
Loan ID	Loan ID of the applicant	Categorical(Nominal)	Object
Gender	Gender of the applicant	Categorical(Nominal)	Object
Married	Married Status of the Applicant	Categorical(Nominal)	Object
Dependents	No of dependants of the applicant	Ordinal	Object
Education	Level of education of the applicant(whether	Ordinal	Object

	customer is a graduate or not)		
Self Employed	Whether customer is self-employed or not	Categorical(Nominal)	Object
Applicant Income	Income of the applicant	Numerical	Integer
Co Applicant Income	Income of the co applicant	Numerical	float
Loan Amount	Expected Loan value	Numerical	float
Loan Amount Term	No of Months which the loan have to repay	Numerical	float
Credit History	Whether the customer has repaid their doubts or not.	Numerical	float
Property Area	Area of the applicant	Ordinal	Object
Loan Status	Applicant is eligible for the loan or not.	Categorical(Nominal)	Object

3. Methodology

Python Jupyter Notebook has been used to explore the given dataset and proceed the model implementation. Since predicting loan eligibility as “Yes” or “No” is a binary classification, this model implementation would lead with some classification algorithms.

below stated libraries have been used throughout the model building process.

- Pandas
- Seaborn
- SkLlearn

Exploratory Data Analysis

Univariate Analysis

- 81% of the applicants are Males
- 65% of the Applicants are Married
- Around 58% of Applicants have no any dependants
- 78% of Applicants are Graduates
- 86% of applicants are not self employed
- Most of the applicants are from Semi urban Area

Bivariate Analysis

- Graduate applicants have higher chances of being selected for the loan approval
- Applicants with 0 dependents have higher chance of being selected for the loan approval
- Non self employed applicants have higher chance of being selected for the loan approval
- Applicants who have repaid their previous debts have higher chances of being selected for the loan approval
- Married applicants have higher of being selected for the loan approval
- Male applicants have higher of being selected for the loan approval

Missing Value Imputation

```
df.isnull().sum()
```

```
Loan_ID                0
Gender                13
Married                3
Dependents            15
Education              0
Self_Employed         32
ApplicantIncome        0
CoapplicantIncome      0
LoanAmount            22
Loan_Amount_Term      14
Credit_History        50
Property_Area          0
Loan_Status            0
Income_bin             0
Co_Income_bin        273
LoanAmount_bin        22
dtype: int64
```

Categorical and Numerical Variable missing value imputation:

Missing values has been replaced by mode of the particular variable for categorical variables whilst median is used to replace the missing values in numerical variables as mean cannot be used due to higher availability of outliers in the dataset for numerical variables.

Outlier Treatment

Log transformation has been applied for avoid the outlier impact for numerical variables and convert the same to normal distribution.

Feature Engineering

Below new features have been created in order to build most accurate model.

Total Income = Applicant Income + Co Applicant Income

Monthly Repayment = Loan Amount/Loan Amount Term

Correlation matrix and heat map is constructed and below variables (Applicant Income, Co Applicant income, Loan Amount, Loan Amount Term) have been removed from the dataset in order to get rid of the impact of highest correlation effect before passing the same to Model building process.

Label Encoding:

binary variables have used 1,0 replacement and variables which have more than 2 unique values have used the Label Encoder from sklearn package to encoding process of categorical variables in the dataset.

Model Building

Loan Status is the Target variable(Y) and below remaining variables have used as feature variables(X).

- Target variable(Y) – Loan Status
- Feature Variables(X)-
- Gender, Married, Dependents, Self-employed, Credit_History,Log_Loan_Amount,Total_Income_Log,Monthly_Repayment_Log

Log_Loan_Amount → Log Transformed Value of “Loan Amount” variable

Total_Income_Log → Log Transformed Value of “Total Income” variable

Monthly_Repayment_Log → Log Transformed Value of “Monthly Repayment” variable

Train Test Split Function have used and divided the dataset in to train and test sets (Test size is 30% from the initial dataset whilst training set is 70%)

Model Training Function has used and different Classification algorithms have been used to train the model. stated below the used algorithms,

- Logistic Regression
- Decision Tree
- Random Forest

Different algorithms with various hyper parameters have been used to train the model and best model has been chosen based on the model evaluation criteria to deploy the solution.

4. Results

One result is the setup of a model and training how the model works for different models and the accuracy of the results each model produce.

Below are the accuracies of the trained classification models.

	model_name	model	accuracy	precision	f1_score	roc_auc
0	lgr1	LogisticRegression(n_jobs=3, verbose=1)	0.854054	0.841060	0.842289	0.783776
1	rf1	(DecisionTreeClassifier(max_depth=10, max_feat...	0.827027	0.840278	0.818833	0.805315
2	rf2	(DecisionTreeClassifier(max_depth=20, max_feat...	0.827027	0.845070	0.820252	0.790140
3	rf3	(DecisionTreeClassifier(max_features='auto', r...	0.832432	0.846154	0.825193	0.800420
4	rf4	(DecisionTreeClassifier(max_depth=10, max_feat...	0.816216	0.838028	0.809018	0.800979
5	rf5	(DecisionTreeClassifier(max_depth=20, max_feat...	0.832432	0.851064	0.826524	0.810839
6	DT1	DecisionTreeClassifier(max_depth=10)	0.767568	0.837209	0.768166	0.722098
7	DT2	DecisionTreeClassifier(max_depth=8)	0.794595	0.843284	0.792275	0.739301
8	DT3	DecisionTreeClassifier(max_depth=5)	0.783784	0.835821	0.781342	0.740979
9	DT4	DecisionTreeClassifier(max_depth=4)	0.843243	0.838926	0.832216	0.750559

5. Conclusion

Model Evaluation

Best Model is the Logistic Regression Model with below accuracy metrics. accuracy is 85.40%while precision is 84.10%and f1Score is 84.22% and roc_auc is 78.37%

It is concluding that the Logistic Regression Classification model has been predict the loan eligibility for a particular applicant with 85% accuracy which is at the considerably satisfactory level.

6. Discussion

In this project, we learned how to implement a full end-to-end machine learning classification model, selecting its best model with the highest accuracy, to predict which applicants would be good enough to repay the loan. The logistic regression model which predicts the loan status has been implemented using 600 sample applicants and their associated features which has available in the online application form. In addition, some data pre-processing steps, such as adding new features and removing the highest correlated features and applying some coding techniques, were applied to the given raw data set before running the model, as it was necessary to achieve a particular satisfactory level of accuracy. If some more variables may have used as feature variables for model implementation, it might provide the much accuracy level than existing. This logistic regression classification model would allow to predict the loan status of a particular applicant with 85% accuracy without human involvement. It would save the time allocated to evaluate the eligibility of the loan status as a manual human task done in a particular financial institution. With enhancing the sample size and associated feature count this model can be fine-tuned for a specific level of accuracy than prevailing.