PROJECT TITLE : LOAN DATA SET USING MACHINE LEARNING ALGORITHMS

Project Title : Loan data set using machine learning algorithms
Mail-id : moezzema10@gmail.com
GitHub : https://github.com/sayyedamoezzema

ABSTRACT :

Banks are making major part of profits through loans. Though lot of people are applying for loans.It's hard to select the genuine applicant, who will repay the loan. While doing the process manually, lot of misconception may happen to select the genuine applicant. Therefore we are developing loan prediction system using machine learning, so the system automatically selects the eligible candidates. This is helpful to both bank staff and applicant. The time period for the sanction of loan will be drastically reduced.

Loan prediction analysis uses specific parameters about a loan applicication to determine whether or not the loan should get approved .Approved loans usually have a good credit histoy,decent application income, and reliability in other factors.banks use statsical and manual method to verify these factors and decide about the applicant's loan status.

In this datatset use loan data information such as gender, self employed, dependents, education, applicantincome, loan status and credit history etc. We will preprocess the data by cleaning it and performing feature engineering to extract meaningful information from the raw data.

In this dataset set we will used multiple classification algorithms such as logistic regression, SVM(support vector machine) classifier, KNN(K-nearest neighbours) classifier and naive bayes classifier. This algorithms used to which one is best fitted and accuracy is good.

Finally, we will concluded that the performance of our all model using metrics such as accuracy, precision, recall, and F1-score.

1.INTRODUCTION :

The company seeks to automate (in real time) the loan qualifying procedure based on information given by customers while filling out an online application form.It is expected that the development of ML model that can help the company predict loan approval in accelerating decision-making process for determing whether an applicant is eligible for a loan or not.

A loan is the core business part of banks. The main portion the bank's profit is directly come from the profit earned from the loans. Though bank approves loan after a regress process of verification and testimonial but still there's no surety whether the chosen hopeful is the right hopeful or not. This process takes fresh time while doing it manually. We can prophesy whether that particular hopeful is safe or not and the whole process of testimonial is automated by machine literacy style. Loan Prognostic is really helpful for retainer of banks as well as for the hopeful also.

1.1 About the data :

So train and test dataset would have the same columns except for the target column that is “Loan Statuts”.

Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/Female
Married	Applicant married(Y/N)
Dependents	Number of dependents
Education	(Graduate/Under graduate)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
Loan amount	Loan amount in thousand
Loan_amount_term	Term of loan in months
Credit_History	Credit history meets guidelines
Property_Area	Urban/semi urban/rural
Loan status	Target loan approved(Y/N)

Preprocessing :

The collected data may conatin missing values that may lead to inconsistency.To gain better result data need to be preprocessing and so it will better the effectiveness of the algorithms.

Train model on training data set :

Now we should train the model on the training dataset and make for soothsaying the test dataset.We can divide our train dataset into two tract train and testimony.We can train the model on this training part and using that make predict for the testimony part.In this way we can validate our soothsaying as we’ve the true soothsaying for the true soothsayings for the testimony part(which we don’t have for the test dataset).

Correlation attributes :

Grounded on the correlation among attributes it was observed more likely to pay back their loans. The attribute that are individual and significant can include property area, education, loan status, credit history, which is since by insight it’s considered as important. The correlation among attributes can associated using box in python platform.

Model evaluation :

Evaluate the performance of the model on separate test dataset to determine its accuracy, precision, recall ,support and F1-score.

2.Data exploration and analysis :

In this data set we done some exploratory data analysis .we have plotted some barplot and box plot Loan_status vs Education.

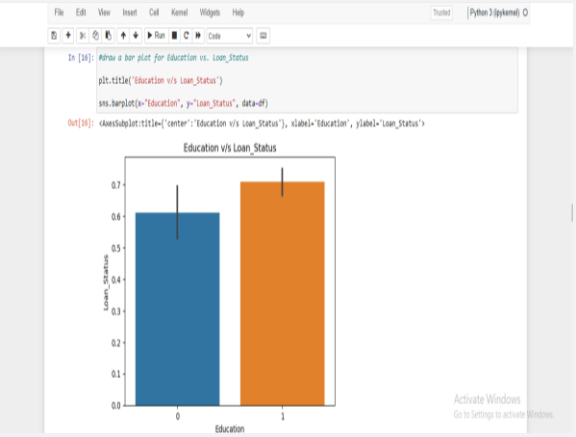

Fig 2.1: Loan_status vs Education

Fig 2.2 : Loan status vs loan amount

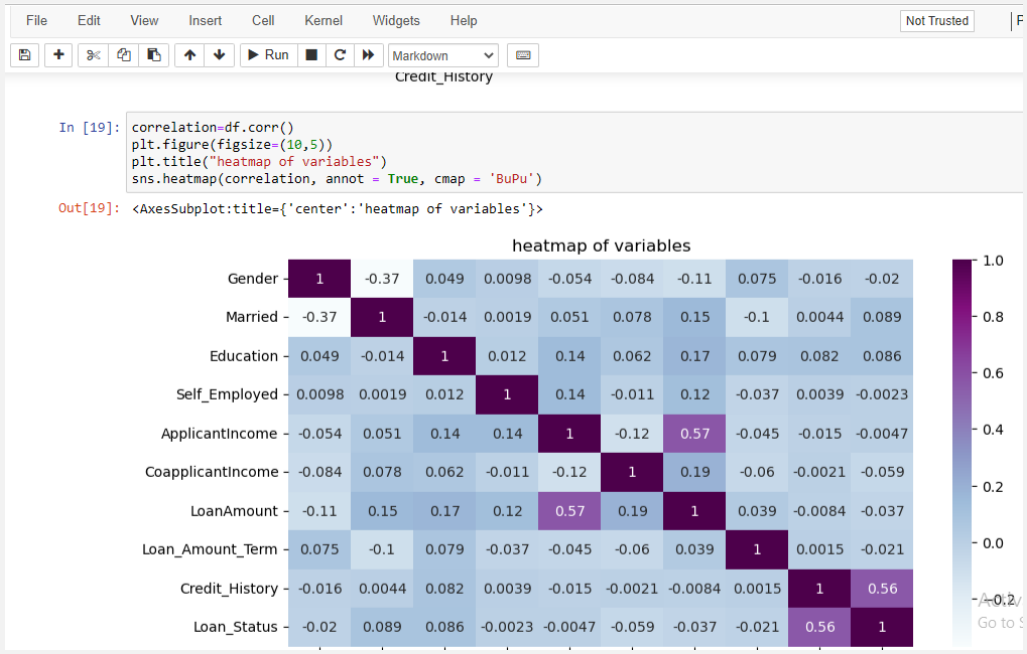

Fig 2.3 : Heat map of related data set

3 . Experiments :

We used loan dataset in this we have categorical variable for ex- yes , no then we then we cannot directly provide it to machine it won't be able to understand so ML algorithm involves lots of mathematical calculation so in this case we will use label encoder because we have features in target variable Loan_status Y/N and male , female. In our given dataset there is null values so we need to drop that missing values .After this stage, complete the data cleaning and manipulation process on the dataset.

3.1 Evaluation Metrics :

Accuracy score is not usually the best form of evaluation in machine learning. Evaluation metrics where we build the model and apply that model to our given dataset that means how well a model is able to make prediction or classifications based on a set of input data. The following evaluation metrics have been choosen.

3.1.1 Precision :

Precision shows the ratio of true positives to the sum of true positive and false positives.

$$\text{Precision} = \frac{\text{True\_positives}}{\text{True\_posoitives} + \text{False\_posoitives}}$$

From the above equation,It is clear that for a good model.False_positives should be as small as possible.precision lies between 1(good) and 0(bad).

3.1.2 Recall :

This is the ration of true positives to the sum of true positives and false negatives.

$$\text{Recall} = \frac{\text{True\_positives}}{\text{True\_positives} + \text{False\_negatives}}$$

Here, for a good model,Flase_negatives should be as small as possible.Recall also lies between 1(good) and 0(bad).

3.1.3 Accuracy :

Accuracy summarieses the whole modle. It is the ratio of the correctly classified prediction to the entire prediction.

$$\text{Accuracy} = \frac{\text{Correct\_Predictions}}{\text{All\_Predictions}}$$

3.1.4 F1 Score :

The F1 score seems to be the most suitable metrics for evaluating imbalanced data problems.This is because it in corporate both the precision and recall scores.

$$\text{F1\_Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.2 procedure

3.2.1 any further processing?

Further data processing such as to use standardscaler is to improve the performance of the machine learning models. Many machine learning algorithms such as regression are sensitive to the scale of input feature. when some features have a larger scale than other so it can lead to biased results and reduced performance. By StanderScaler, we ensure that all feature have similar scale and thus avoid this issue. StandardScaler is typically used as a preprocessing step before training a machine learning model. It involves fitting the StandardScaler on the training data and then transforming both the training and test data using the same scaler.

This ensures that the scaling is consistent across the training and test data. The main use of StandardScaler is to standardize the features of a dataset to improve the performance of machine learning models, especially those that are sensitive to the scale of the input features.

3.2.2 Models used:

There are several models used in machine learning classification, but for this use case we have label data i.e. target column contains loan_status yes,no. Some of the popular one is Logistic Regression, SVM(support vector machine)claasifier, KNN (K-Nearest Neighbour)claasifier and naive bayes. and check which model gives best suitable accuracy and which model we can deploy further.

4.Result :

In the below table we compare of the multiple algorithms that I used.In our dataset we have used different algorithms like logistic regression, SVM(support vector machine)classifier, KNN(k-nearest neighbour)classifier, and Navie bayes that accuracy as below.

Classifier	Class	Labels	PREC	REC	F1-SCORE	SUPPORT	ACC
Logistic Regression	Loan data using ML	1 (Y) 0(N)	0.85 0.83	0.57 0.95	0.68 0.89	30 66	83.33%
SVM(support vector machine) classifier	Loan data using ML	1 (Y) 0(N)	0.89 0.83	0.57 0.97	0.69 0.90	30 66	84.37%
KNN (k-nearest neighbours)classifier	Loan data using ML	1 (Y) 0(N)	0.10 0.91	0.33 0.69	0.15 0.78	9 87	65.62%
Naive bayes	Loan data using ML	1 (Y) 0(N)	0.87 0.86	0.67 0.95	0.75 0.91	30 66	86.45%

5. DISCUSSION :

5.1 Justify four algorithms :

Machine learning (ML) is a branch of artificial intelligence (AI) that enables computers to “self-learn” from training data and improve over time, without being explicitly programmed. Machine learning algorithms are able to detect patterns in data and learn from them, in order to make their own predictions.

ML algorithms are mainly divided into three categories firs one in supervised machine learning algorithm and second one is unsupervised and third one is Reinforcement machine learning algorithm.

supervised machine learning algorithms are a type of machine learning algorithm that learns to make predictions based on labeled training data. In supervised learning, the algorithm is trained using a set of inputoutput pairs, where the input is the data and the output is the corresponding label or target value that we want the algorithm to predict. The goal of the algorithm is to learn a mapping between the inputs and outputs, so that it can make accurate predictions on new, unseen data.

There are many different types of supervised learning algorithms, including:

Regression: Regression algorithms are used to predict a continuous value, such as a price or a temperature.

Classification: Classification algorithms are used to predict a categorical value, such as whether an email is spam or not.

5.2 Logistic regression :

Logistic regression is a type of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring.

Logistic regression is used to solve classification problems, and the most common use case is binary logistic regression, where the outcome is binary (yes or no). In the real world, you can see logistic regression applied across multiple areas and fields.

- In health care, logistic regression can be used to predict if a tumor is likely to be benign or malignant.
- In the financial industry, logistic regression can be used to predict if a transaction is fraudulent or not.
- In marketing, logistic regression can be used to predict if a targeted audience will respond or not.

5.3 Support Vector Machines (SVMs):

SVMs are a type of algorithm that learn to separate the data into different classes based on a hyperplane that maximizes the margin between the classes.in our dataset we use SVC. SVC (Support Vector Classification) is a type of supervised learning algorithm used in machine learning for binary and multi-class classification.

It is based on the concept of finding the optimal hyperplane that separates the different classes in a high-dimensional space. In SVC, the training data is used to find the optimal hyperplane that maximizes the margin, which is the distance between the hyperplane and the closest data points of each class.

The hyperplane is chosen such that it maximizes the margin and correctly classifies he training data. In cases where the data is not linearly separable, SVC uses a kernel function to map the input data into a higher-dimensional space where it is more likely to be linearly separable. The most common kernel functions used in SVC are the linear, polynomial, and radial basis function (RBF) kernels.

The hyperparameters of the SVC algorithm, such as the regularization parameter (C) and the kernel parameters, can be tuned using cross-validation or other model selection techniques to optimize the performance of the algorithm on the given dataset. SVC has been shown to be effective in a wide range of applications, including image classification, text classification, and bioinformatics.

5.4 KNN(k-nearest neighbour):

In KNN classification, the K nearest neighbours is found based on a distance metric, such as Euclidean or Manhattan distance, and the class of the new observation is assigned based on the majority class of the K neighbors.

In KNN regression, the K nearest neighbours is used to calculate the average or median value, which is assigned as the predicted value for the new observation. The choice of the value of K is an important parameter in KNN. A larger value of K makes the algorithm more robust to outliers and noise, but can lead to overfitting in some cases.

A smaller value of K makes the algorithm more sensitive to noise and may result in overfitting. The optimal value of K can be determined through cross-validation or other model selection techniques.

5.5 Naive bayes:

Naive Bayes is a type of probabilistic classification algorithm used in machine learning. It is based on Bayes' theorem, which describes the probability of an event occurring based on prior knowledge of conditions that might be related to the event. Naive Bayes is called "naive" because it assumes that the features used for classification are independent of each other, which is often not the case in practice.

In Naive Bayes, the algorithm builds a probabilistic model of the training data, assuming that the distribution of each feature is independent of the other features. Then, for a new observation, the algorithm calculates the probability of each class given the observed values of the features, using Bayes' theorem. The class with the highest probability is assigned as the predicted class for the new observation.

here are three main types of Naive Bayes classifiers: Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes. Gaussian Naive Bayes is used for continuous numerical data, Multinomial Naive Bayes is used for discrete count data, such as text data, and Bernoulli Naive Bayes is used for binary data.

The Naive Bayes algorithm is computationally efficient and requires relatively small amounts of training data to achieve high accuracy. It has been used in a variety of applications, such as text classification, sentiment analysis, spam filtering, and medical diagnosis. However, its performance may suffer if the assumption of independence between features is strongly violated, or if the training data is imbalanced or noisy.

6.CONCLUSION :

From a proper analysis of analysis this system can be used perfect for detection of client who are eligible for approval of loan.The software is working perfect and can be used for all banking requirements.This system can be easily uploaded in any operating system.Since the technology is moving toward online,this system has more scope of upcoming days .This system is more secure and reliable.Since we have used different classification algorithms the system return good accuracy results.There is no issue if there are many no of customer applying for loan.This system accepts data for N no. of customers.In future we can add more algorithms to this system for getting more accurate results.

References :

1.Dataset from (https://www.kaggle.com/datasets/burak3ergun/loan-data-set)

2.”what is machine learning”by https://monkeylearn.com/machine-learning

3.”logistic regression”by https://www.capitalone.com/tech/machine-learning/what-is-logistic-regression

4. "Introduction to Machine Learning with Python" by Andreas C. Müller and Sarah Guido

5. "Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

6. "Machine Learning Yearning" by Andrew Ng
"Bayesian Reasoning and Machine Learning" by David Barber