REPORT ON LOAN PREDICTION BY MACHINE LEARNING MODEL

SUBMITTED BY

SHUVAM PRASAD SHARMA

E. KOUSHIK KUMAR SHUBAM KUMAR MADHU PULLURI

ABSTRACT

With the advancement in technology, there are so many enhancements in the banking sector also. The number of applications is increasing every day for loan approval. There are some bank policies that they have to consider while selecting an applicant for loan approval. Based on some parameters, the bank has to decide which one is best for approval. It is tough and risky to check out manually every person and then recommended for loan approval. In this work, we use a machine learning technique that will predict the person who is reliable for a loan, based on the previous record of the person whom the loan amount is accredited before. This work's primary objective is to predict whether the loan approval to a specific individual is safe or not.

CONTENTS Page No.

1. [INTRODUCTION 4](#_TOC_250007)
2. [IMPLEMENTATION DETAILS (Modules) 5](#_TOC_250006)
3. [PROPOSED ALGORITHM 7](#_TOC_250005)
4. [DISCUSSION BY SAMPLE CODE 8](#_TOC_250004)
5. [RESULTS OBTAINED 12](#_TOC_250003)
6. [CONCLUSION 12](#_TOC_250002)
7. [FUTURE ENHANCEMENT 13](#_TOC_250001)
8. [REFERENCES 13](#_TOC_250000)

# INTRODUCTION

Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this Paper is to provide quick, immediate and easy way to choose the deserving applicants. Dream housing Finance Company deals in all loans. They have presence across all urban, semi urban and rural areas. Customer first apply for loan after that company or bank validates the customer eligibility for loan. Company or bank wants to automate the loan eligibility process (real time) based on customer details provided while filling application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and other. This project has taken the data of previous customers of various banks to whom on a set of parameters loan were approved. So the machine learning model is trained on that record to get accurate results. Our main objective of this project is to predict the safety of loan. To predict loan safety, the SVM and Naïve bayes algorithm are used. First the data is cleaned so as to avoid the missing values in the data set.

The main objective of this paper is to predict whether assigning the loan to particular person will be safe or not. This paper is divided into four sections (i)Data Collection

(ii) Comparison of machine learning models on collected data (iii) Training of system on most promising model (iv) Testing

* 1. PROBLEM DEFINITION

Banks, Housing Finance Companies and some NBFC deal in various types of loans like housing loan, personal loan, business loan etc in all over the part of countries. These companies have existence in Rural, Semi Urban and Urban areas. After applying loan by customer these companies validates the eligibility of customers to get the loan or not. This project provides a solution to automate this process by employing machine learning algorithm. So the customer will fill an online loan application form. This form consist details like Sex, Marital Status, Qualification, Details of Dependents, Annual Income, Amount of Loan, Credit History of Applicant and others.

* 1. PROPOSED MODEL

This system predict whether the loan is approve or reject . This system refers the following things or ways.

* + - Data Collection
    - Data Pre-processing (Data Cleaning)
    - Model Selection
    - Model Evaluation Classification
    - Result (output)

# IMPLEMENTATION DETAILS (Modules):

* + 1. Loan Dataset :

Loan Dataset is very useful in our system for prediction of more accurate result. Using the loan Dataset the system will automatically predict which costumer’s loan it should approve and which to reject. System will accept loan application form as an input. Justified format of application form should be given as an input to get processed.

* + - * Dataset collection: The dataset is collected from Kaggle.

In this model prediction, we collected two datasets to form a final dataset. <https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset>

In this link, we collected datasets.

* + 1. Determine the training and testing data:

Typically, Here the system separate a dataset into a training set and testing set, most of the data use for training and a smaller portions of data is use for testing. after a system has been processed by using the training set, it makes the prediction against the test set.

* + 1. Data cleaning and processing:

In Data cleaning the system detect and correct corrupt or inaccurate records from database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying or detecting the dirty or coarse data. In Data processing the system convert data from a given form to a much more usable and desired form i.e. make it more meaningful and informative.

* 1. Models used :

Decision Tree:

* + - Decision tree is a type of supervised learning algorithm(having a pre-defined target variable) that is mostly used in classification problems. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on the most significant splitter/differentiator in input variables.
    - Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of

resultant sub-nodes. In other words, we can say that purity of the node increases with respect to the target variable.

SVM:

* + - In this approach, each data item is plotted in a ndimensional space, where n represents the number of features with each feature represented in a corresponding co- ordinates. A hyper plane is determined to distinguish the classes (possibly two) based on their features.

Random Forest:

* + - Random Forest is a tree-based bootstrapping algorithm wherein a certain no. of weak learners (decision trees) is combined to make a powerful prediction model.
    - For every individual learner, a random sample of rows and a few randomly chosen variables are used to build a decision tree model.
    - Final prediction can be a function of all the predictions made by the individual learners.
    - In the case of a regression problem, the final prediction can be the mean of all the predictions.

Logistic Regression:

* + - In statistics, the **logistic model** (or **logit model**) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one.

Gaussian Naive Bayes:

* + - Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. Naive Bayes are a **group of supervised machine learning classification algorithms** based on the Bayes theorem. It is a simple classification technique, but has high functionality.

Hyperparameter tuning:

* + - A Machine Learning model is defined as a mathematical model with a number of parameters that need to be learned from the data. By training a model with existing data, we are able to fit the model parameters.
    - However, there is another kind of parameters, known as Hyperparameters, that cannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. These parameters

express important properties of the model such as its complexity or how fast it should learn.

GridSearchCV:

* + - In GridSearchCV approach, machine learning model is evaluated for a range of hyperparameter values. This approach is called GridSearchCV, because it searches for best set of hyperparameters from a grid of hyperparameters values.

# PROPOSED ALGORITHM:

The following shows the pseudo code for the proposed loan prediction method

1. Import Dependencies
2. Data collection
3. Data preprocessing
4. Handling missing values
5. Dropping the missing values
6. Label encoding (Converts strings to numeric values)
7. Splitting the data
8. Applying the model training and model accuracy by:
   * RandomForestClassifer,
   * DecisionTreeClassiffier,
   * Logistic Regression,
   * SVM
9. Scaling of Dataset
10. Finding the best model using GridSearch CV
11. Model Prediction

* Google Colab link:

[https://colab.research.google.com/drive/1b\_yf9MWTw2992ceOfhuJJWEJsCqchuwf?](https://colab.research.google.com/drive/1b_yf9MWTw2992ceOfhuJJWEJsCqchuwf?usp=sharing) [usp=sharing](https://colab.research.google.com/drive/1b_yf9MWTw2992ceOfhuJJWEJsCqchuwf?usp=sharing)

# DISCUSSION BY SAMPLE CODE:

#Import dependencies

import numpy as np import pandas as pd import seaborn as sn

from sklearn.model\_selection import train\_test\_split from sklearn import svm

from sklearn.ensemble import RandomForestClassifier

from sklearn.model\_selection import GridSearchCV,RandomizedSearchCV from sklearn.linear\_model import LogisticRegression

from sklearn.naive\_bayes import GaussianNB from sklearn.tree import DecisionTreeClassifier

#Data cleaning and Data preprocessing

loan\_data\_1=pd.read\_csv("/content/train\_u6lujuX\_CVtuZ9i (1).csv") loan\_data\_1.head()

loan\_data\_2=pd.read\_csv("/content/test\_lAUu6dG.csv") loan\_data\_2.head()

loan\_data = pd.concat((loan\_data\_1,loan\_data\_2), ignore\_index=True) loan\_data

loan\_data.drop('Loan\_ID’,axis=1,inplace=True) loan\_data.head()

#Statastical measures

loan\_data.describe() loan\_data.shape

#Number of missing values in each columns

loan\_data.isnull().sum()

#Dropping missing values loan\_data=loan\_data.dropna() loan\_data.isnull().sum()

#Label encoding

loan\_data.replace({'Gender':{'Male':1,'Female':0},'Dependents':{'3+':4}

,'Education':{'Graduate':1,'Not Graduate':0},

'Married':{'No':0,'Yes':1},'Self\_Employed':{'Yes':1,'

No':0},

'Property\_Area':{'Urban':0,'Rural':1,'Semiurban':2},'Loan\_Status':{'Y': 1,'N':0}},inplace=True)

loan\_data.tail()

X=loan\_data.drop('Loan\_Status',axis=1) y=loan\_data['Loan\_Status']

#Splitting the data

X\_train,X\_test,y\_train,y\_test=train\_test\_split(X,y)

#Training and evaluating the dataset with different models

rfc=RandomForestClassifier(n\_estimators=100) rfc.fit(X\_train,y\_train) rfc.score(X\_test,y\_test)

dtc=DecisionTreeClassifier() dtc.fit(X\_train,y\_train) dtc.score(X\_test,y\_test)

lg=LogisticRegression() lg.fit(X\_train,y\_train) lg.score(X\_test,y\_test)

clf=svm.SVC(kernel='linear')

clf.fit(X\_train,y\_train) clf.score(X\_test,y\_test)

print(rfc.score(X\_test,y\_test),dtc.score(X\_test,y\_test),lg.score(X\_test

,y\_test),clf.score(X\_test,y\_test))

#Scaling of dataset or Standardisation

from sklearn.preprocessing import StandardScaler scaler=StandardScaler()

X\_train\_s=scaler.fit\_transform(X\_train)

X\_train\_s

#Finding best model by using GridSearch CV

clf=GridSearchCV(svm.SVC(gamma='auto'),{ 'C':[1,10,20],

'kernel':['rbf','linear']

},cv=5)

clf.fit(X\_train\_s,y\_train) clf.cv\_results\_

df = pd.DataFrame(clf.cv\_results\_) df

df[['param\_C','param\_kernel','mean\_test\_score']]

clf.best\_params\_

clf.best\_score\_

model\_params={ 'svm':{

'model': svm.SVC(gamma='auto'), 'params':{

'C':[1,10,20],

'kernel':['rbf','linear']

}

},

'logistic\_regression':{

'model': LogisticRegression(C=5,solver='liblinear', multi\_class

='auto'),

'params':{

'C':[1,10,20]

}

},

'),

'random\_forest':{ 'model':RandomForestClassifier(n\_estimators=100,criterion='gini

'params':{

'n\_estimators':[1,10,20]

}

},

'naive\_bayes\_gaussian':{ 'model':GaussianNB(), 'params':{}

}

,'decision\_tree':{ 'model':DecisionTreeClassifier(), 'params':{}

}

}

scores=[]

for model\_name, mp in model\_params.items(): clf=GridSearchCV(mp['model'], mp['params'], cv=5) clf.fit(X\_train\_s,y\_train)

scores.append({

'model': model\_name, 'best\_score': clf.best\_score\_, 'best\_params':clf.best\_params\_

})

scores

df=pd.DataFrame(scores, columns=['model','best\_score','best\_params']) df

#Model prediction

cutomer\_data=[1,1,1,1,0,4583,1508.0,128.0,360.0,1.0,1]

data=np.array(customer\_data).reshape(1,-1) data

#Prediction test by SVM model

prediction=clf.predict(data) prediction

#Prediction test by Decision Tree model

prediction=dtc.predict(data) prediction

#Prediction test by Random Forest model

prediction=rfc.predict(data) prediction

# RESULTS OBTAINED:

|  |  |
| --- | --- |
| Algorithms | Model accuracy |
| Random Forest | 0.766 |
| Decision Tree | 0.741 |
| Logistic Regression | 0.766 |
| SVM | 0.766 |

# CONCLUSION

* + The predictive models based on Logistic Regression, Decision Tree and Random Forest, SVM give the accuracy as 76.66%, 74.16%,76.66%,76.67

respectively.

* + By parameter tuning with GridSearch CV for the given dataset, the accuracy of model based on Random Forest is high.
  + The best model for our loan prediction model is Random Forest.

# FUTURE ENHANCEMENT

Among three the accuracy of DT algorithm is best for prediction of loans. In future the Decision Tree algorithm can be applied on other datasets available for loan approvals to further investigate its accuracy. A rigorous analysis of other machine learning algorithms other than these three can also be done in future to investigate the power of machine learning algorithms for loan approval prediction.

# REFERENCES

1. PhilHyo Jin Do ,Ho-Jin Choi, “Sentiment analysis of real life situations using location, people and time as contextual features,” International Conference on Big Data and Smart Computing (BIGCOMP), pp. 39–42. IEEE, 2015.
2. Bing Liu, “Sentiment Analysis and Opinion Mining,” Morgan Claypool Publishers, May 2012.
3. Bing Liu, “Sentiment Analysis: Mining Opinions, Sentiments, and Emotions,” Cambridge University Press, ISBN:978-1-107-01789-4.
4. Shiyang Liao, Junbo Wang, Ruiyun Yu, Koichi Sato, and Zixue Cheng, “CNN for situations understanding based on sentiment analysis of twitter data,” Procedia computer science, 111:376–381, 2017.CrossRef.
5. K I Rahmani, M.A. Ansari, Amit Kumar Goel, “An Efficient Indexing Algorithm for CBIR,”IEEE- International Conference on Computational Intelligence Communication Technology, 13-14 Feb 2015.
6. Gurlove Singh, Amit Kumar Goel ,”Face Detection and Recognition System using Digital Image Processing” , 2nd International conference on Innovative Mechanism for Industry Application ICMIA 2020, 5-7 March 2020, IEEE Publisher