A

Group Project Report On

"LOAN PREDICTION SYSTEM"

Submitted in Partial Fulfillment of the Requirements For the award of the Degree

of

BACHELOR OF TECHNOLOGY IN

Electronics and Computer Engineering(ECM)

By

B. PRUTHVI REDDY(19311A19C3)

SATYA LOHIT GOLI(19311A19D6)

K.ADHARSH KRUPAKAR(19311A19H4)

Under the Guidance of

V.SOWMYA

ASSISTANT PROFESSOR

DEPARTMENT OF ELECTRONICS & COMPUTER ENGINEERING

Sreenidhi Institute of Science & Technology

(Affiliated to JNT University, Hyderabad)

Yamnampet, Ghatkesar, Hyderabad – 501 301.

2021-2022

DEPARTMENT OF ELECTRONICS & COMPUTER ENGINEERING

SREENIDHI INSTITUTE OF SCIENCE & TECHNOLOGY

(AUTONOMOUS)

CERTIFICATE

This is to certify that the Group Project work entitled **"LOAN PREDICTION SYSTEM"**, submitted by **B.PRUTHVI REDDY(19311A19C3), SATYA LOHIT GOLI(19311A19D6),K.ADHARSH KRUPAKAR(19311A19H4)** towards partial fulfillment for the award of Bachelor's Degree in Electronics & Computer Engineering from Sreenidhi Institute of Science & Technology, Ghatkesar, Hyderabad, is a record of bonafide work done by them. The results embodied in the work are not submitted to any other University or Institute for award of any degree or diploma.

Internal Guide	**Project Coordinator**	**H.O.D**
V.SOWMYA	**V. SOUJENYA**	**Dr. D. MOHAN**
Assistant Professor	Assistant Professor	**Department of ECM**
Department of ECM	Department of ECM	

EXTERNAL EXAMINER

DECLARATION

This is to certify that the work reported in the present project titled **"LOAN PREDICTION SYSTEM**" is a record work done by us in the **Department of Electronics and Computer Engineering, Sreenidhi Institute of Science and Technology,** Yamnampet, Ghatkesar**.**

The report is based on the project work done entirely by us and not copied from any other source.

B.PRUTHVI REDDY (19311A19C3)

SATYA LOHIT GOLI (19311A19D6)

K.ADHARSH KRUPAKAR (19311A19H4)

ACKNOWLEDGEMENT

We convey our sincere thanks to **Prof. CV.Tomy,** Director and **Dr. Shiva Reddy**, Principal, Sreenidhi Institute of Science and Technology, Ghatkesar for providing resources to complete this project. We are very thankful to **Dr.D. Mohan**, Head of ECM Department, **V.Sowmya** Assistant Professor and **V.Soujenya** Assistant Professor of Sreenidhi Institute of Science and Technology, Ghatkesar for providing an initiative to this group project and giving valuable timely suggestions over our group project work and for their kind co-operation in the completion of the group project. We convey our sincere thanks to all the faculties of ECM department, Sreenidhi Institute of Science and Technology, for their continuous help, co-operation, and support to complete this project.

ABSTRACT

With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So in this paper we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. 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 onmost promising model (iv) Testing

CONTENTS

LIST OF FIGURES AND TABLES

1.INTRODUCTION

The immense increase in capitalism, the fast-paced development and instantaneous changes in the lifestyle has us in awe. Emi, loans at nominal rate, housing loans, vehicle loans, these are some of the few words which have skyrocketed from the past few years. The needs, wants and demands have never been increased this before. People gets loan from banks; however, it may be baffling for the bankers to judge who can pay back the loan nevertheless the bank shouldn't be in loss. Banks earn most of their profits through the loan sanctioning. Generally, banks pass loan after completing the numerous verification processes despite all these, it is still not confirmed that the borrower will pay back the loan or not [1]. To get over the dilemma, I have built up a prediction model which says if the loan has been assigned in the safe hands or not. Government agencies like keep under surveillance why one person got a loan and the other person could not. In Machine Learning techniques which include classification and prediction can be applied to conquer this to a brilliant extent. Machine learning has eased today's world by developing these prediction models. Here we will be using the fine techniques of machine learning – Decision tree algorithm to build this prediction model for loan assessment. It is as so because decision tree gives accuracy in the prediction and is often used in the industry for these models. 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.[3]

2.LITERATURE SURVEY

The immense increase in capitalism, the fast-paced development and instantaneous changes in the lifestyle has us in awe. Emi, loans at nominal rate, housing loans, vehicle loans, these are some of the few words which have skyrocketed from the past few years. The needs, wants and demands have never been increased this before. People gets loan from banks; however, it may be baffling for the bankers to judge who can pay back the loan nevertheless the bank shouldn't be in loss. Banks earn most of their profits through the loan sanctioning. Generally, banks pass loan after completing the numerous verification processes despite all these, it is still not confirmed that the borrower will pay back the loan or not. To get over the dilemma, I have built up a prediction model which says if the loan has been assigned in the safe hands or not. Government agencies like keep under surveillance why one person got a loan and the other person could not.In Machine Learning techniques which include classification and prediction can be applied to conquer this to a brilliant extent. Machine learning has eased today's world by developing these prediction models. Here we will be using the fine techniques of machine learning – Decision tree algorithm to build this prediction model for loan assessment. It is as so because decision tree gives accuracy in the prediction and is often used in the industry for these models.

3.IMPLEMENTATION

3.1 ARCHITECTURE

Algorithm steps:

Step 1: Read the dataset.

Step 2: Random Sampling is done on the data set to make it balanced.

Step 3: Divide the dataset into two parts i.e., Train dataset and Test dataset.

Step 4: Feature selection are applied for the proposed models.

Step 5: Accuracy and performance metrics has been calculated to know the efficiency for different algorithms.

Step6: Then retrieve the best algorithm based on efficiency for the given dataset.

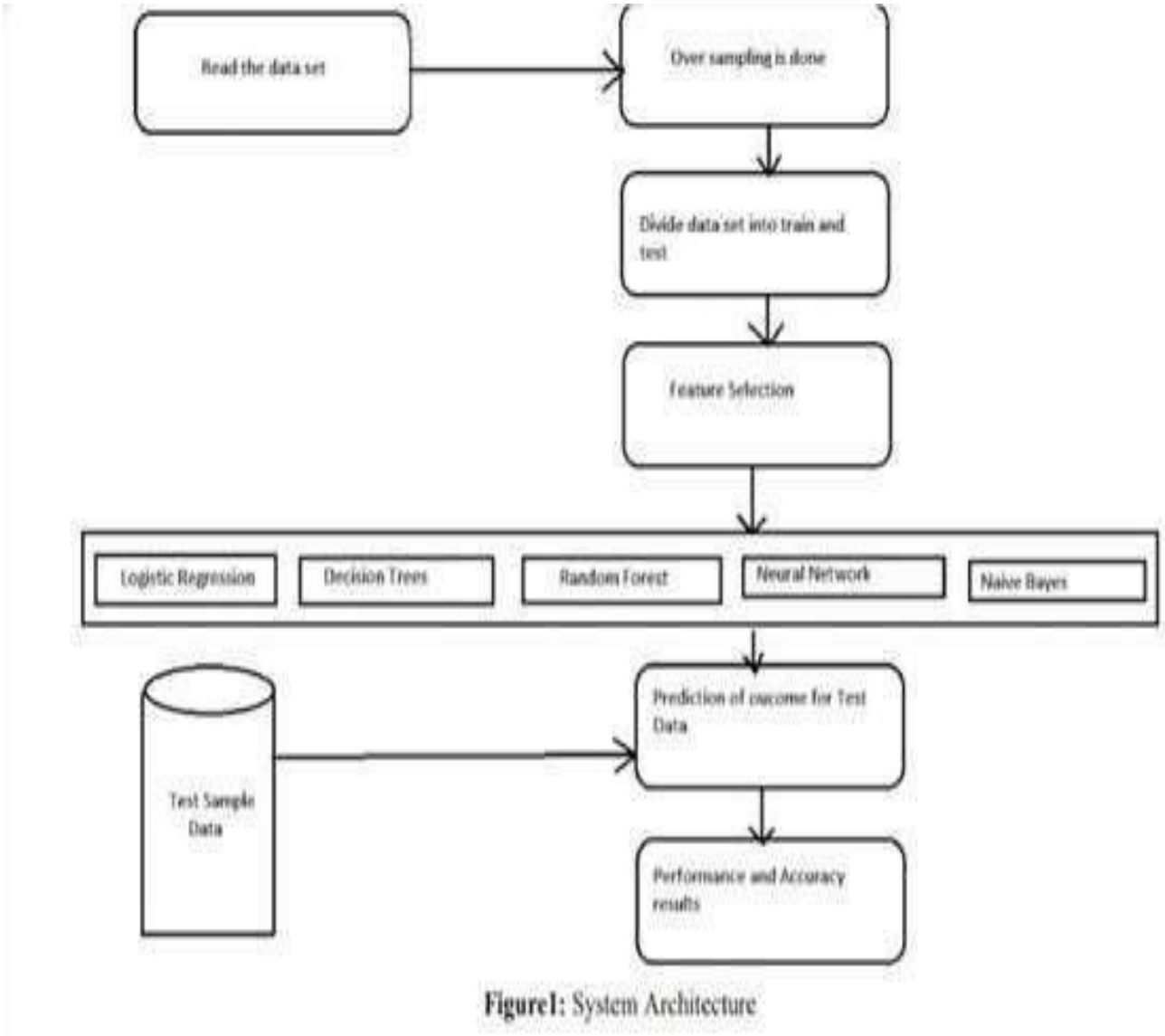


Figure1: System Architecture

3.2 DESCRIPTION OF DATA

- The data for our study was accessed from Kaggle Machine Learning repository.
- The data set includes 18000 records and 21 features.
- We aim to build classification models like random forest, svm , xgboost, cat boost and logic regression.
- These models would be trained on a dataset which will be engineered carefully after performing the feature engineering.

3.3 DESIGN

- Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.
- Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression.The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features.
- Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1

3.3.1 RANDOM FORESTS

Random forests or **random decision forests** is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted tree.. However, data characteristics can affect their performance.

Decision trees are a popular method for various machine learning tasks. Tree learning "come[s] closest to meeting the requirements for serving as an off-the-shelf procedure for data mining", say Hastie *et al.*, "because it is invariant under scaling and various other transformations of feature values, is robust to inclusion of irrelevant features, and produces inspectable models. However, they are seldom accurate".

In particular, trees that are grown very deep tend to learn highly irregular patterns: they overfit their training sets, i.e. have low bias, but very high variance. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance.[3]:587–588 This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model.

Forests are like the pulling together of decision tree algorithm efforts. Taking the teamwork of many trees thus improving the performance of a single random tree. Though not quite similar, forests give the effects of a k-fold cross validation.[2]

3.3.2 SUPPORT VECTOR MACHINE(SVM)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision

boundary or hyperplane:[4]

3.3.3 LOGISTIC REGRESSION

Logistic regression is an example of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring. An example of logistic regression could be applying machine learning to determine if a person is likely to be infected with COVID-19 or not. Since we have two possible outcomes to this question - yes they are infected, or no they are not infected - this is called *binary classification*. In this imaginary example, the probability of a person being infected with COVID-19 could be based on the viral load and the symptoms and the presence of antibodies, etc. Viral load, symptoms, and antibodies would be our factors (Independent Variables), which would influence our outcome (Dependent Variable).

The three types oflogistic regression

1. **Binary logistic regression** - When we have two possible outcomes, like our original example of whether a person is likely to be infected with COVID-19 or not.
2. **Multinomial logistic regression** - When we have multiple outcomes, say if we build out our original example to predict whether someone may have the flu, an allergy, a cold, or COVID-19.
3. **Ordinal logistic regression** - When the outcome is ordered, like if we build out our original example to also help determine the severity of a COVID-19 infection, sorting it into mild, moderate, and severe cases.

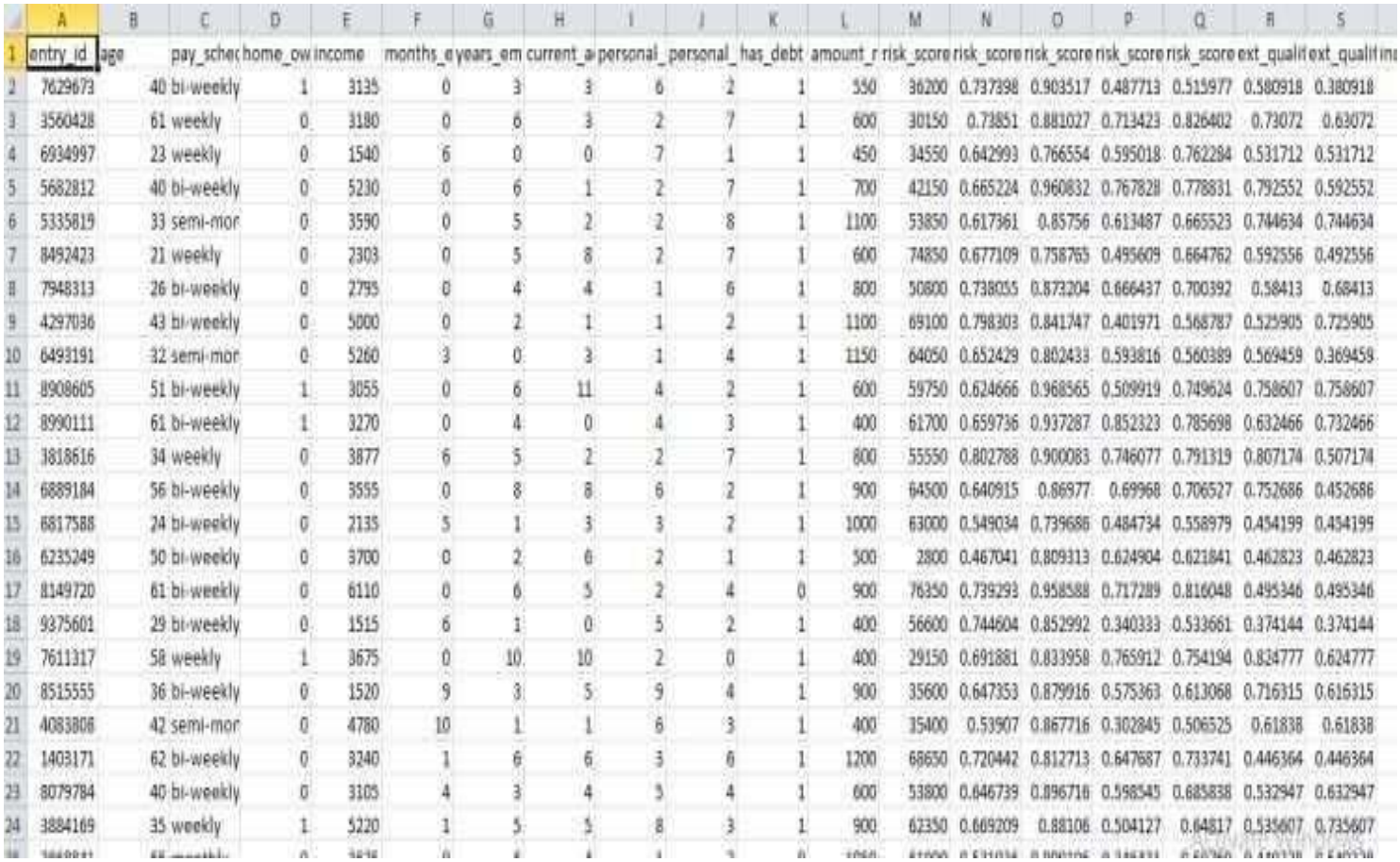
Training data assumptions forlogistic regression

Training data that satisfies the below assumptions is usually a good fit for logistic regression.

- The predicted outcome is strictly binary or dichotomous. (This applies to binary logistic regression).
- The factors, or the independent variables, that influence the outcome are independent of each other. In other words there is little or no multicollinearity among the independent variables.
- The independent variables can be linearly related to the log odds.
- Fairly large sample sizes.[1][2]

3.4 DATA SET

- The data for our study was accessed from Kaggle Machine Learning repository.
- The data set includes 18000 records and 21 features.
- We aim to build classification models like random forest, svm , xgboost, cat boost and logic regression.
- These models would be trained on a dataset which will be engineered carefully after performing the feature engineering.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	entry_id	age	pay_sche	home_ow	income	months_e	years_em	current_a	personal_	personal_	has_debt	amount_r	risk_score	risk_score	risk_score	risk_score	risk_score	ext_qualit	ext_qualit
2	7629673	40	bi-weekly	1	3135	0	3	3	6	2	1	550	36200	0.737398	0.903517	0.487713	0.515977	0.580918	0.380918
3	3560428	61	weekly	0	3180	0	6	3	2	7	1	600	30150	0.73851	0.881027	0.713423	0.826402	0.73072	0.63072
4	6934997	23	weekly	0	1540	6	0	0	7	1	1	450	34550	0.642993	0.766554	0.595018	0.762284	0.531712	0.531712
5	5682812	40	bi-weekly	0	5230	0	6	1	2	7	1	700	42150	0.665224	0.960832	0.767828	0.778831	0.792552	0.592552
6	5335819	33	semi-mor	0	3590	0	5	2	2	8	1	1100	53850	0.617361	0.85756	0.613487	0.665523	0.744634	0.744634
7	8492423	21	weekly	0	2303	0	5	8	2	7	1	600	74850	0.677109	0.758765	0.495609	0.664762	0.592556	0.492556
8	7948313	26	bi-weekly	0	2795	0	4	4	1	6	1	800	50800	0.738055	0.873204	0.666437	0.700392	0.58413	0.68413
9	4297036	43	bi-weekly	0	5000	0	2	1	1	2	1	1100	69100	0.798303	0.841747	0.401971	0.568787	0.525905	0.725905
10	6493191	32	semi-mor	0	5260	3	0	3	1	4	1	1150	64050	0.652429	0.802433	0.593816	0.560389	0.569459	0.369459
11	8908605	51	bi-weekly	1	3055	0	6	11	4	2	1	600	59750	0.624666	0.968565	0.509919	0.749624	0.758607	0.758607
12	8990111	61	bi-weekly	1	3270	0	4	0	4	3	1	400	61700	0.659736	0.937287	0.852323	0.785698	0.632466	0.732466
13	3818616	34	weekly	0	3877	6	5	2	2	7	1	800	55550	0.802788	0.900083	0.746077	0.791319	0.807174	0.507174
14	6889184	56	bi-weekly	0	3555	0	8	8	6	2	1	900	64500	0.640915	0.86977	0.69968	0.706527	0.752686	0.452686
15	6817588	24	bi-weekly	0	2135	5	1	3	3	2	1	1000	63000	0.549034	0.739686	0.484734	0.558979	0.454199	0.454199
16	6235249	50	bi-weekly	0	3700	0	2	6	2	1	1	500	2800	0.467041	0.809313	0.624904	0.621841	0.462823	0.462823
17	8149720	61	bi-weekly	0	6110	0	6	5	2	4	0	900	76350	0.739293	0.958588	0.717289	0.816048	0.495346	0.495346
18	9375601	29	bi-weekly	0	1515	6	1	0	5	2	1	400	56600	0.744604	0.852992	0.340333	0.533661	0.374144	0.374144
19	7611317	58	weekly	1	3675	0	10	10	2	0	1	400	29150	0.691881	0.833958	0.765912	0.754194	0.824777	0.624777
20	8515555	36	bi-weekly	0	1520	9	3	5	9	4	1	900	35600	0.647353	0.879916	0.575363	0.613068	0.716315	0.616315
21	4083808	42	semi-mor	0	4780	10	1	1	6	3	1	400	35400	0.53907	0.867716	0.302845	0.506525	0.61838	0.61838
22	1403171	62	bi-weekly	0	3240	1	6	6	3	6	1	1200	68650	0.720442	0.812713	0.647687	0.733741	0.446364	0.446364
23	8079784	40	bi-weekly	0	3105	4	3	4	5	4	1	600	53800	0.646739	0.896716	0.598545	0.685838	0.532947	0.632947
24	3884169	35	weekly	1	5220	1	5	5	8	3	1	900	62350	0.669209	0.88106	0.504127	0.64817	0.535607	0.735607

Fig2: data set

4.TRAINING THE MODEL

- The data for our study was accessed from Kaggle Machine Learning repository.
- The data set includes 18000 records and 21 features.
- We aim to build classification models like random forest, svm , xgboost, cat boost and logic regression.
- These models would be trained on a dataset which will be engineered carefully after performing the feature engineering.
- Data with various features are taken and data is trained and tested and based on which the models are applied for better accuracy.

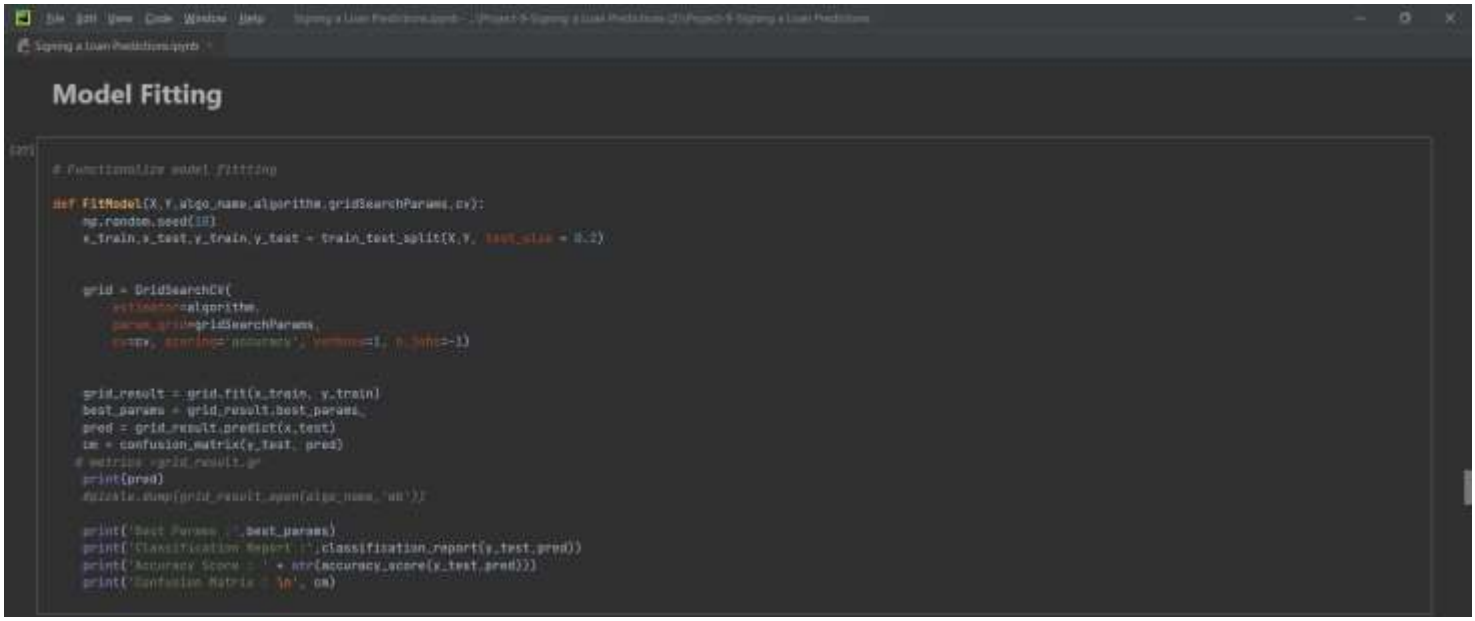


fig3:model fitting

5.RESULTS

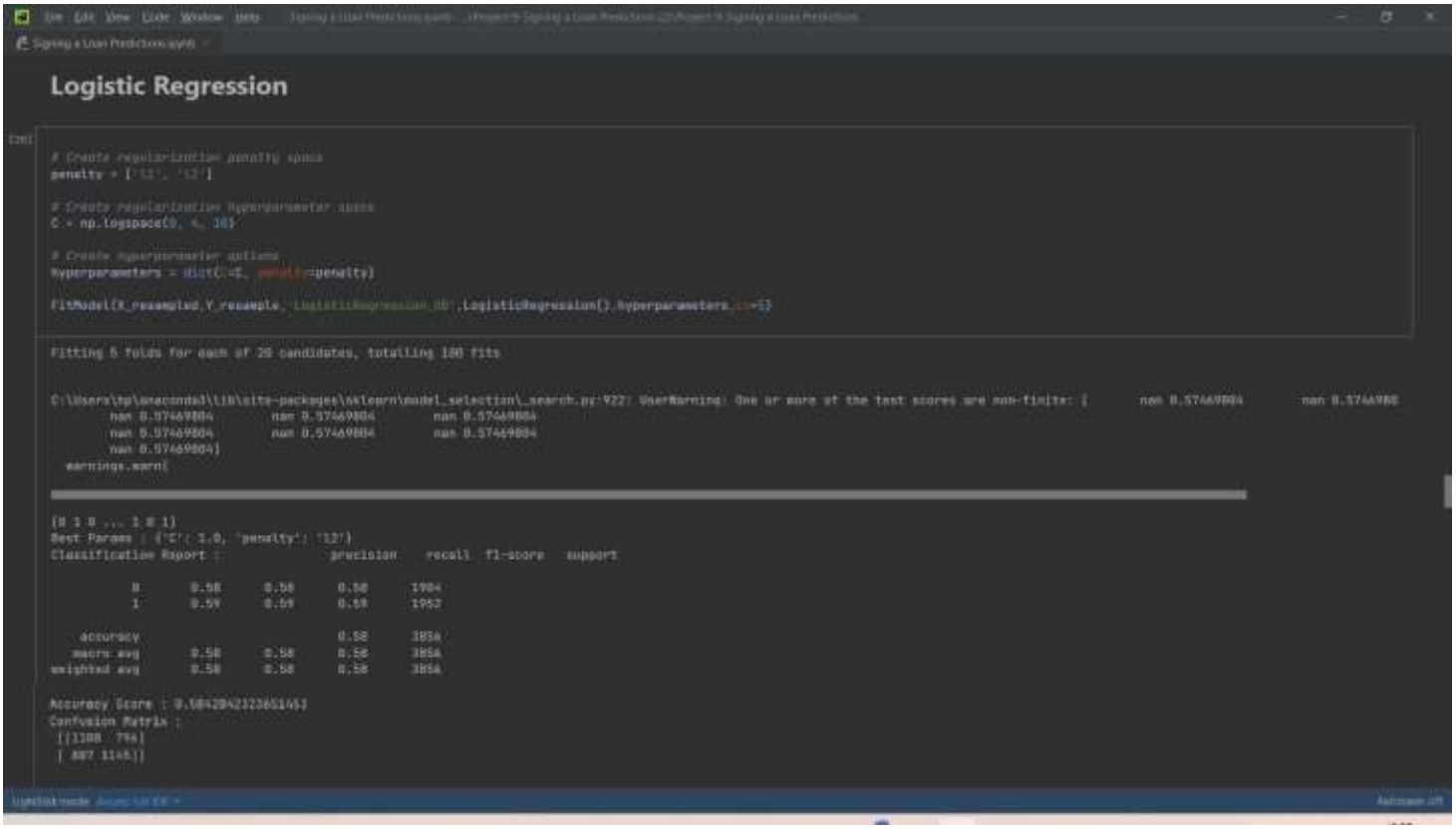


Fig4:logistic regression

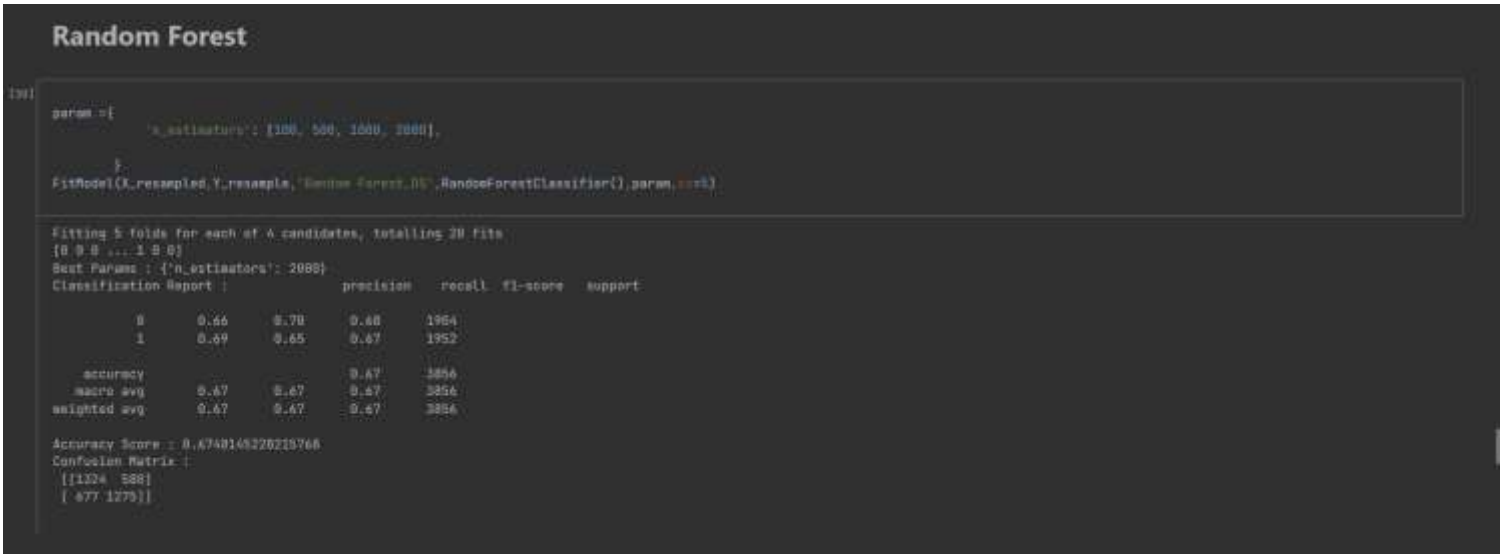


Fig5:random forest

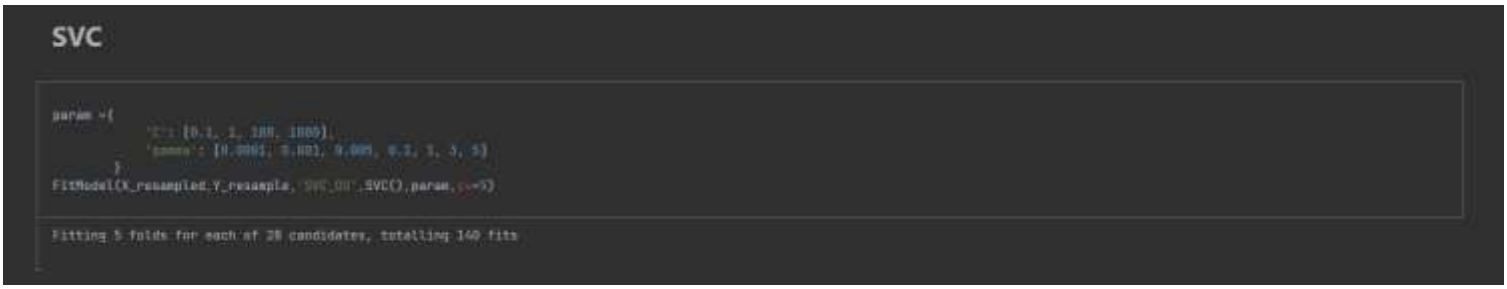

SVC
param ={
'C': [0.1, 1, 100, 1000],
'gamma': [0.0001, 0.001, 0.005, 0.1, 1, 3, 5]
}
FitModel(X_resampled,Y_resample,'SVC_GS',SVC(),param,cv=5)
Fitting 5 folds for each of 28 candidates, totalling 140 fits

Fig6:svm

6.CONCLUSION

Machine learning technique like Logistic regression, SVM and Random forest ,xgboost, were used for loan prediction system. The accuracy and error rate are used to evaluate the performance forthe proposed system. The accuracy for logistic regression, SVM and random forest classifier are 0.58, 0.63, and 0.68 respectively. By comparing all the method, found that random forest classifier is better than the logistic regression and SVM. From a proper analysis of positive points and constraints on the member, it can be safely concluded that the product is a considerably productive member. This use is working duly and meeting to all Banker requisites. This member can be freely plugged in numerous other systems. There have been mathematics cases of computer glitches, violations in content and most important weight of features is fixed in automated prophecy system, so in the near future the so – called software could be made more secure, trustworthy and dynamic weight conformation. In near future this module of prophecy can be integrated with the module of automated processing system. The system is trained on old training dataset in future software can be made resembling that new testing date should also take part in training data after some fix time.

7. FUTURE SCOPE

In future, this model can be used to compare various machine learning algorithm generated prediction models and the model which will give higher accuracy will be chosen as the prediction model. Furthermore, other emerging models of Machine Learning could also be studied to check for the accuracy rate resulted by them. Sentiment analysis though Machine Learning on how loans can be granted safely. Other deep learning based models can also be used for prediction purposes. For future work, deep learning models could be developed which consider financial news articles along with financial parameters such as aclosing price, traded volume, profit and loss statements etc., for possibly better results.

8.REFERENCES

[1] J.R. Quinlan, Induction of decision trees, Machine learning Springer,vol. 1, no. 1, pp. 81-106, 1086.

[2] Arun Kumar, Ishan Garg and Sanmeer Kaur, Loan Approval Prediction based on Machine Learning Approach.

[3] Nur Alfiyatin Adyan, Taufiq Hilman, Ema Febrita Ruth and FirdausMahmudy Wayan, "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization", International Journal of Advanced Computer Science and Applications, vol. 8, no.10, 2017.

[4] S.S. Keerthi and E.G. Gilbert, Convergence of a generalize SMO algorithm for SVM classifier design, Machine Learning, Springer, vol. 46, no. 1, pp. 351-360, 2002..

[5] Algorithm steps: Step 1: Read the dataset. Step 2: Random Sampling is done on the data set to make it balanced. Step 3: Divide the dataset into two parts i.e., Train dataset and Test dataset. Step 4: Feature selection are applied for the proposed models. Step 5: Accuracy and performance metrics has been calculated to know the efficiency for different algorithms. Step6: Then retrieve the best algorithm based on efficiency for the given dataset.