IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Credit Risk Assessment for Home credit group

[1]N.Manisha, [2]M.Riya Raj, [3]S.Manimala,[4]D.Soniya,[5]K. Sasi Kiran

B. Tech (IV-CSE) Students, Department of Computer Science and Engineering[1,2,3,4]

Assistant Professsor, Department of Mechanical Engineering[5]

Ace Engineering College,Hyderabad,Telangana,India

Abstract: Home Credit is an international Non-Banking Financial Institution (NBFC) founded in 1997 in the Czech Republic..Home Credit finances both the purchase of consumer goods through a cash-less loan transaction or other commodities, such as language courses, travel etc. through a cash loans. Home credit mainly focus on the people who has less credit and no credit history. There unit of measurement 307511 observations at intervals the dataset with 122 columns that represent every qualitative and quantitative attribute of those 67 columns have missing values. Out of 122, sixteen columns unit of measurement categorical and 106 unit of measurement numerical columns. there is a binary output variable that denotes “Delay in payments(1) or “No Delay in payments (0)”.

Index Terms - **Delay in payment , No delay in payment, Home Credit, Binary output**

I. INTRODUCTION

Home credit predict client repayment capabilities by making use of variety of various data including service transactional information Home Credit is currently using various statistical and machine learning methods to make these predictions successful. This will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful. These could check the customer capable to repay the loan in given time. This could check that the purchaser is capable of compensation are not rejected that loans unit of measurement given with a principal, maturity, and compensation calendar that will empower their purchasers to attain success.

II. DATA AND SOURCES OF DATA https://www.kaggle.com/c/home-credit-default-risk

-Data set link.

III. LITERATURE SURVEY

The general problem one encounters is that of finding effective methodologies and algorithms to produce mathematical or statistical descriptions (models) to represent the patterns, regularities or trends in the financial or business data. Conceptually this is not a new subject and in some ways it is the logical extension and generalization of the methods that have been used by statisticians for decades. For complex real-world data, where noise, non-linearity and idiosyncrasies are the rule, the best strategy is to take an interdisciplinary approach that combines statistics and machine learning algorithms. This type of interdisciplinary, data-driven computational approach, sometimes referred as Knowledge Discovery in Databases (Fayyad et al [1996], Simoudis et al [1996] ,Bigus [1996], Adrians and Santinge [1996]), is specially relevant today due to the convergence of three factors: I) Corporate and government financial databases, where all and every financial transaction can be stored, have growth in size, number and availability. Recent results on statistics, generalization theory, machine learning and complexity have provided new guidelines and deep insights into the general characteristics and nature of the model building/learning/fitting process Affordable computing resources including high performance multi-processor servers, powerful desktops and large The standardization of operating systems and environments (Unix, Windows NT/95 and Java) has facilitated the integration and interconnection of data sources, repositories and applications. There are many algorithms available for model construction so one of the main problems inpractice is that of algorithm selection or combination. Unfortunately it is hard to choose an algorithm a priory because one might not know the nature and characteristics of the dataset.

III. EXISTING SYSTEM

Traditionally there were two approaches for credit risk assessment i.e Early warning system and Risk decomposition and aggregation. Early warning systems rely on some failure-non-failure or problem-non-problem definition for the financial institution. This analysis is phenomenological since it only attempts to describe the failure of the whole institution without making any analytical assessment of the factors that produce the failure Risk decomposition and aggregation has its roots in the arrival of capital asset pricing models and the development of Contingent Claim Analysis. Risk decomposition and aggregation is an ambitious approach. Risk decomposition makes risk management easier since it provides the magnitude and the source of the risk; However, it requires much more information and calculations than an early warning system.

IV. PROPOSED SYSTEM

Our project work concentrates on evaluation of different machine learning classifier models to predict the credit risks associated with various borrowers of an institution. For this the major assessment parameters of the institution are taken as the predictor variables. XGBoost is a model which used in our project. On the other hand different statistical techniques like chi square test, 2 sample t-tests etc. are performed to determine the important features. Feature integration also implemented because of its high dimensionality. Normalization, one hot encoding and standard scalar methods have been executed to check the improvement of the model performance. KBest feature selection enacted to check the important features and PCA applied on the data because of its high dimensionality.

V. ALGORITHM

1 LOGESTIC REGRESSION

Logestic regression is an binary classification algorithm which comes under supervised learning. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. Logistic reg7ression is used for solving the classification problems. It permits you to make predictions from labelled information if the target (output) variable is categorical. Used as a result of having a categorical outcome variable violates the thought of spatiality in ancient regression. instead of building a mantic model for "Y (Response)" directly, the approach models “Log Odds (Y)”; thence the name provision or Logic. the foremost draw back with a straight line is that it isn't steep enough. at intervals the sigmoid curve, as you will see, you have low values for various points, then the values rise all of a pointy, once that you have got various high values.
Sigmoid function operates as: - s(x)=1/1e^-x Log(odds)=log(p/1-p)z=beta0+beta1 h(x)=sigmoid(z) h(x)1/1+e^(beta0+beta1)

2 DECISION TREE

A decision tree uses a tree-like model to make predictions. It resembles Associate in Nursing turned tree. it's to boot really reasonably like but you produce decisions in real life, you raise a series of inquiries to achieve a decision. A decision tree splits the information into multiple sets. Then, each of these sets is further split into subsets to achieve a decision. The topmost decision node throughout a tree corresponds to the foremost effective predictor referred to as the idea node. A node whereas not further branches is termed a leaf node. The leaf nodes represent the last word decisions. we have a tendency to square measure able to calculate that node is that the foundation by looking for the information gain, entropy, or Gini index. It suffers from high bias and high variance and will be a greedy learner.
Entropy E(s)= Σ-p(xi) log2 p(xi), Range is 0 to 1 lesser the score its best used in (ID3, C4.5, C5.0) Gini Index=1-Σp(xi)square, Range is 0 to 0.5 lesser the score its best used in (CART) Information Gain=H(s) Σ|v|/|s|(H(V), It should be high less entropy or Gini index information gain or vice versa.

3 Random Forest

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It uses the lowest principle of material with random feature option to kind a great deal of diverse trees. cacophonic a node throughout the event of a tree, the split that is chosen is not any longer the foremost effective split among all the choices. Instead, the split picked is that the most effective split among a random set of the choices. As a result of this randomness, the bias of the forest generally slightly can increase (with respect to the bias of 1 systematic tree). Random Forest will use sqrt root of n choices for classification and n/3 choices for regression whereas n being an entire vary of choices.

4 Bagging

Bagging is also known as bootstrap aggregation. It is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement. A bootstrap sample is created by sampling the given information set uniformly and with replacement. A bootstrap sample sometimes contains regarding 30-70% information from the information set. it is a parallel technique that means employment and testing square measure attending to be done parallelly and freelance of each totally different. fabric handles overfitting and reduces variance.

5 Boosting

Boosting is an ensemble learning method that combines a set of weak learners into a strong learner to minimize training errors. Boosting is also a sequent technique, where each ensuant model tries to correct the errors of the previous model. The succeeding model's unit of measurement obsessed to the previous model. Boosting provides misclassified samples higher weight. it is a thanks to boost weak learning algorithms (single tree) into a sturdy learning algorithm. employment and Testing unit of measurement sequent in Boosting. the aim of Boosting is to chop back Bias. it's attending to increase the overfitting at intervals the information.

6 ADA BOOST

Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances. It works on up the areas wherever the bottom learner fails. the bottom learner could be a machine learning rule that's a weak learner and upon that the boosting technique is applied to show it into a powerful learner. Any machine learning rule that accepts weights on coaching knowledge may be used as a base learner. AdaBoost works on up the areas wherever the bottom learner fails. the bottom learner could be a machine learning rule that's a weak learner and upon that the boosting technique is applied to show it into a powerful learner.

7 Gradient Boosting

Gradient boosting, each predictor corrects its predecessor's error. Gradient boosting involves three elements are : a loss function to be optimized, a weak learner to make predictions, an additive model to add weak learners to minimize the loss function. it's in a different way to grant additional importance to troublesome instances. At every iteration, the residuals are computed and a weak learner is fitted to those residuals.

8 XGBoost

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. XGBoost (Extreme Gradient Boosting) is associate optimized distributed gradient boosting library. It uses a gradient boosting (GBM) framework at the core. Yet, will higher than the GBM framework alone. XGBoost implements data processing and is quicker as compared to GBM. XGBoost has associate in-built routine to handle missing values. XGBoost tries various things because it encounters a missing worth on every node and learns that path to require for missing values in future.

VI. EXPERIMENTAL ANALYSIS

1 Class Imbalance

Oversampling the target variable by using SMOTE. Classification using category-imbalanced knowledge is biased in favor of the bulk class. Oversampling is that the method of changing the minority category into the bulk category. i.e., increase the minority category to majority category count. The artificial Minority Over-sampling Technique (SMOTE) is associate oversampling approach that makes artificial minority category samples. It doubtless performs higher than straightforward oversampling and it's wide used.

2 Statistical Analysis

Statistical tests were performed to envision whether or not the independent variables have a big relationship with the variable quantity, TARGET.

2.1 Chi-square Test

For the explicit Columns, a Chi-square take a look at of independence was performed with the target variable, TARGET that is additionally a categorical column.

2.2 Two-sample t test

For all the numeric variables, two-sample mismatched t-tests were performed between values of the variable for 2 categories of target variables to match their means.

3 PCA

Since the information is big, PCA is employed to envision whether or not this could improve our model performance. From the higher than results, it's ascertained that Accuracy, precision, recall, and f1-score are inflated when put next with the bottom model. when acting Cross-Validation for the PCA model with CV=5 and marking = "roc_auc", below the results.

```
[0.97176648 0.97093312 0.97008974 0.97173741 0.96994854]
Bias_error: 0.029104944712872283
VE: 0.0008681624871695027
```

Though we have a tendency to get smart results, we wish to undertake the "Select KBEST" technique for feature choice and proceed with the model building as a result of when implementing PCA on the dataset, our original options can develop into Principal parts. Principal parts remain the linear combination of our original options. Principal parts don't seem to be as clear and explicable as original options.

4 Select K -Best

Statistical tests like Chi-square and t-test for the dataset, the results of the take a look at shown that everyone the options are important with relevance the target variable. within the choose K-Best technique, we have a tendency to specify the quantity of options and also the technique returns the foremost important amongst them. We had number of trials with k=80, 90,100 and so on. We got good score for k=100 value. Although the scores will be more if we go beyond k=120, but there would be much variance error in the scores. So, choosing optimal k value can be done only through trial-and-error method. After performing Cross Validation for the model with CV=5 and scoring = "roc_auc" , below are the results.

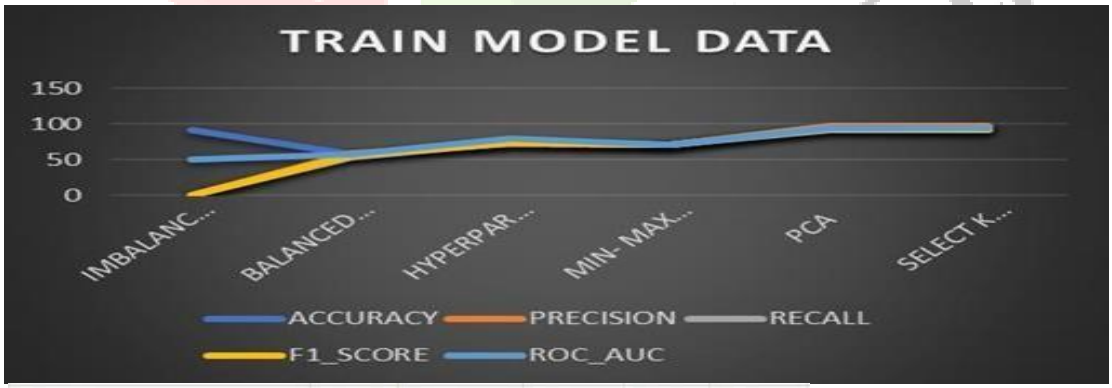
```
[0.97050352 0.96949795 0.96883082 0.97040282 0.9684298 ]
Bias_error: 0.030467017435086063
VE: 0.0009232839447959737
```

As our score are almost same as PCA, here we are able to interpret the features by backing them with statistical analysis. So going further we tried different ensemble methods on K Best Model to see if we can improve the score.

VII. RESULTS AND DISCUSSION

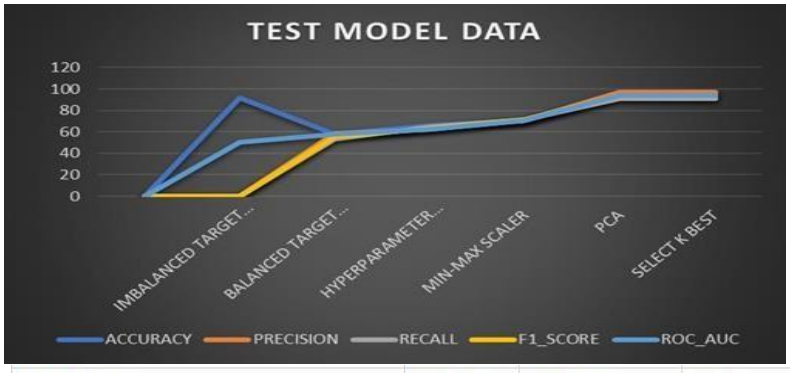
1.Tables and Graphs

TRAIN MODEL	ACCURACY	PRECISION	RECALL	F1_SCORE	ROC_AUC
IMBALANCED TARGET LOGISTIC REGRESSION	91.92	0	0	0	49.99
BALANCED TARGET LOGISTIC REGRESSION	57.7	58.5	53.6	55.9	57.8
HYPERPARAMETER TUNING FOR LR	72.1	72.6	74.2	73.1	79.5
MIN-MAX SCALER	70.6	70.2	71.4	70.8	70.6
PCA	94.3	97.3	91.1	94.1	94.3
SELECT K BEST	93.9	96.9	90.8	93.7	93.9



TEST MODEL	ACCURACY	PRECISION	RECALL	F1_SCORE	ROC_AUC
IMBALANCED TARGET LOGISTIC REGRESSION	91.92	NA*	0	0	50
BALANCED TARGET LOGISTIC REGRESSION	57.8	58.5	53.6	56	57.8
HYPERPARAMETER TUNING FOR LR	65.83	62.1	64.4	63.2	62.6
MIN-MAX SCALER	70.9	70.5	71.6	71.1	70.9
PCA	94.3	97.3	91.1	94.1	94.3
SELECT K BEST	94	97.1	90.7	93.8	94

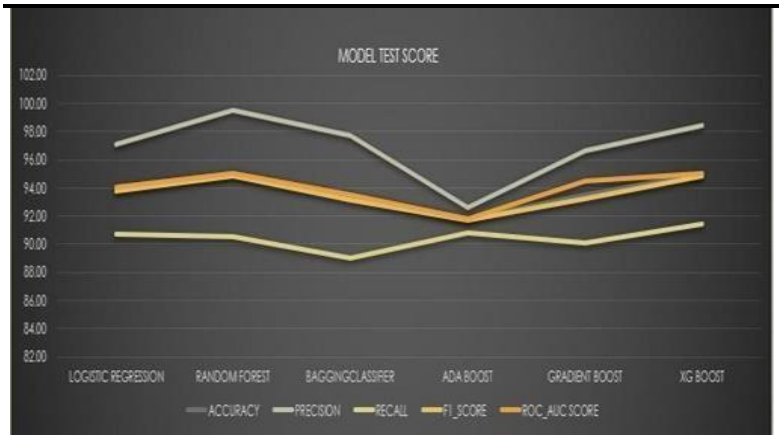
Note: It is an edge case, model hasn't predicted any positive cases due to class imbalance

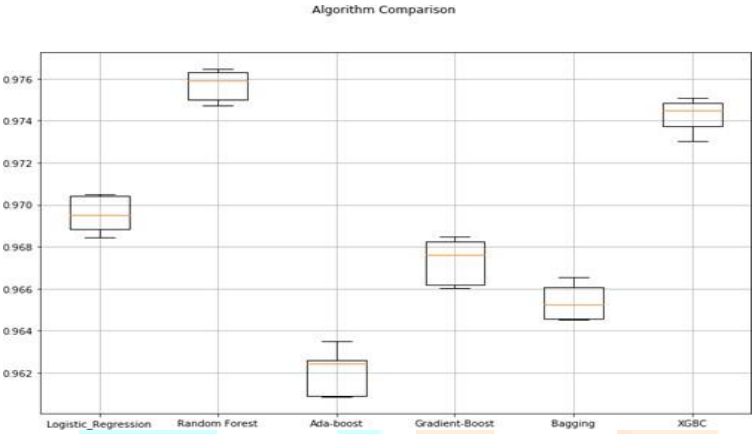


TRAIN MODEL	ACCURACY	PRECISION	RECALL	F1_SCORE	ROC_AUC
IMBALANCED TARGET LOGISTIC REGRESSION	91.92	0	0	0	49.99
BALANCED TARGET LOGISTIC REGRESSION	57.7	58.5	53.6	55.9	57.8
HYPERPARAMETER TUNING FOR LR	72.1	72.6	74.2	73.1	79.5
MIN-MAX SCALER	70.6	70.2	71.4	70.8	70.6
PCA	94.3	97.3	91.1	94.1	94.3
SELECT K BEST	93.9	96.9	90.8	93.7	93.9

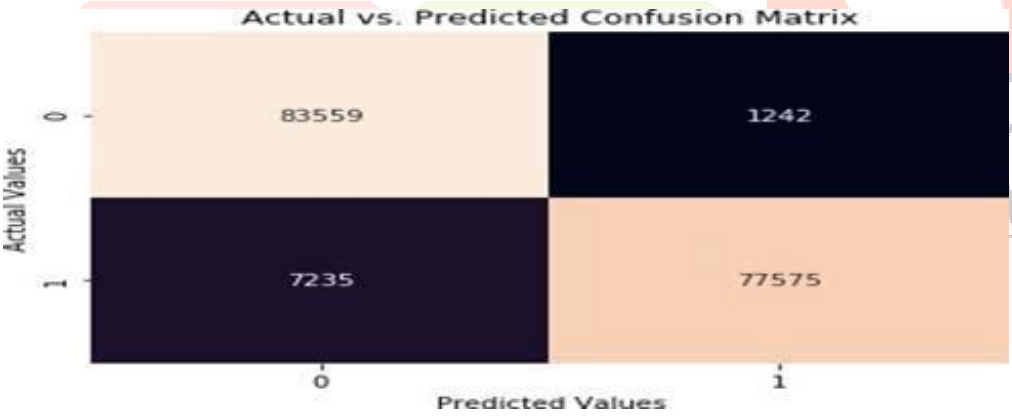
MODEL NAME(TRAIN)	ACCURACY	PRECISION	RECALL	F1_SCORE	ROC_AUC SCORE
LOGISTIC REGRESSION	93.92	96.90	90.75	93.72	93.92
RANDOM FOREST	99.99	100	99.99	99.99	99.99
BAGGING CLASSIFIER	99.31	99.97	98.64	99.30	99.31
ADA BOOST	91.74	100	99.99	91.67	91.74
GRADIENT BOOST	93.48	96.65	90.07	93.25	93.48
XG BOOST	95.42	98.85	91.90	95.25	95.42

MODEL NAME(TEST)	ACCURACY	PRECISION	RECALL	F1_SCORE	ROC_AUC SCORE
LOGISTIC REGRESSION	93.99	97.08	90.71	93.78	93.99
RANDOM FOREST	95.02	99.48	90.52	94.79	95.02
BAGGING CLASSIFIER	93.45	97.69	89.01	93.15	93.45
ADA BOOST	91.76	92.60	90.77	91.68	91.76
GRADIENT BOOST	93.48	96.67	90.05	93.24	94.48
XG BOOST	95.00	98.42	91.46	94.81	95.00





From above box plot comparison diagram, we can see that Random Forest and XGBC are better models amongst all of the models. But when we observe train and test scores of Random Forest and XGBC, we can see that XGBC model is performing better on both train and test data. So, we are finalizing best model as XGBC model.



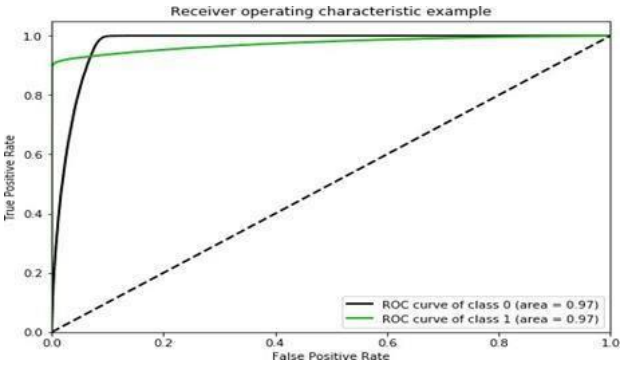


Fig: Roc_Auc curve for XGBC

VIII. CONCLUSION

Credit risk assessment is merely attainable by means that of activity. Machine learning models may be used as tools to live the credit risk exposure of assorted monetary establishments. With the proper prediction of credit risk, its management can become effective and economical. This project work concentrates on the analysis varied} machine learning classifier models to predict the credit risks related to various borrowers of an establishment. For this, the main assessment parameters of the establishment are taken because the predictor variables. There are several classifier models we've approached that are mentioned within the report. we will say once and for all that XG-Boost is that the model that performed well in our project. On the opposite hand, completely different applied math techniques just like the chi-square take a look at, a pair of sample t-tests, etc. are performed to work out the vital options. However, we've conjointly tried the K-Best technique to work out the feature importance. Feature integration has conjointly been enforced owing to its high spatial property. standardization, one-hot cryptography, and normal scalar ways are dead to visualize the advance of the model performance.

IX. REFERENCE

[1] https://www.kaggle.com/c/home-credit-default-risk - Data set link.
[2] Machine Learning by Tom M Mitchell.
[3] Hands on Machine Learning with Sckit-Learn and TensorFlow by Aurelien Geron.
[4] https://scikitlearn.org/stable/modules/generated/sklearn.decomposition.PCA.htmlPCA
[5] https://scikitlearn.org/stable/supervised_learning.html#supervisedlearning-Machine Learning Models
[6] https://scikitlearn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html- Feature Engineering
[7] https://docs.scipy.org/doc/scipy/reference/stats.html-statistical tests
[8] https://machinelearningmastery.com/smoteoversampling-forimbalanced-classification/-smote
[9]https://machinelearningmastery.com/hyperparameters-forclassification-machine-learning-algorithms
Hyperparameter tuning
[10] https://en.wikipedia.org/wiki/Decision_tree -Decision Tree
[11] https://en.wikipedia.org/wiki/Ensemble_learning - Ensemble Learning models
[12] https://towardsdatascience.com/cross-validation-and hyperparametertuning-how-to-optimise-your-machine learning-model13f005af9d7d?gi=1d33882a8888 - Hyperparameter Tuning