CREDIT RISK ASSESSMENT FOR HOME CREDIT GROUP

NAME :- VIVEK GUTTI
ROLLNO :- CB.EN.P2AID20021

GMAIL ID:-vivekgutti@gmail.com

NAME :- K.V.SHASHANK
ROLLNO :- CB.EN.P2AID20027

GMAIL ID:-
shashankkv1008@gmail.com

Abstract— Home Credit is an international Non-Banking Financial Institution (NBFC) founded in 1997 in the Czech Republic. The company operates in 14 countries and focuses on lending primarily to people with little or no credit history. There are 307511 observations in the dataset with 122 columns that stand for both qualitative and quantitative attributes of which 67 columns have missing values. Out of 122, 16 columns are categorical and 106 are numerical columns. There is a binary output variable that denotes “Delay in payments” (1) or “No Delay in payments (0)”.

Keywords—non banking, financial institution, delay, no delay in payments.

I. INTRODUCTION

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data--including Telco and transactional information--to predict their clients' repayment abilities. 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. PROJECT JUSTIFICATION

1) Conservative credit risk management policies, fast loan decisions and reasonable loan pricing achieve this balance of protecting loan portfolios while keeping bank customers satisfied with the institution.

2) The objective of this project is to predict the home loan credit risk for the financial institution. The project will enable the bank to reduce their risk of loan loss by gaining an apt understanding its customer base, thus minimizing the loss of capital for the financial institution while reaping optimal profit.

3) By analyzing the customer features such as transaction history, annual income, demographics etc., and the bank will be able to estimate the risk of the loan repayment.

II. ALGORITHMS USED

1. Logistic Regression: -

The most common use of logistic regression models is in binary classification problems. Logistic Regression is a supervised classification model. It allows you to make predictions from labelled data, if the target (output) variable is categorical. Used because having a categorical outcome variable violates the assumption of linearity in normal regression. Instead of building a predictive model for "Y (Response)" directly, the approach models “Log Odds (Y)”; hence the name Logistic or Logit. The main problem with a straight line is that it is not steep enough. In the sigmoid curve, as you can see, you have low values for a lot of points, then the values rise all of a sudden, after which you have a lot of high values.

Sigmoid function 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.DecisionTrees: -

Decision tree uses a tree-like model to make predictions. It resembles an upside-down tree. It is also very similar to how you make decisions in real life: you ask a series of questions to arrive at a decision. A decision tree splits the data into multiple sets. Then, each of these sets is further split into subsets to arrive at a decision. The topmost decision node in a tree which corresponds to the best predictor called root node. A node without further branches is called a leaf node. The leaf nodes represent the final decisions. We can calculate which node is the root by finding out the information gain, entropy or Gini index. It suffers with high bias and high variance and is 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 used for ensemble of decision trees. It uses base principle of bagging with random feature selection to create more diverse trees. Splitting a node during the construction of a tree, the split that is chosen is no longer the best split among all the features. Instead, the split picked is the best split among a random subset of the features. As a result of this randomness, the bias of the forest usually slightly increases (with respect to the bias of a single non-random tree). Random Forest will use sqrt root of n features for classification and n/3 features for regression while n being total number of features.

4.Bagging: -
Bagging stands for Bootstrap Aggregation. Bootstrapping means creating bootstrap samples from a given data set. A bootstrap sample is created by sampling the given data set uniformly and with replacement. A bootstrap sample typically contains about 30-70% data from the data set. It is a parallel process that mean training and testing will be done parallelly and independent of each other. Bagging handles over fitting and reduces variance.

5.Boosting: -
Boosting is a sequential process, where each subsequent model attempts to correct the errors of the previous model. The succeeding models are dependent on the previous model. Boosting gives misclassified samples higher weight. It is a method to boost weak learning algorithm (single tree) into strong learning algorithm. Training and Testing are sequential in Boosting. The purpose of Boosting is to reduce Bias. It may increase the over fitting in the data.

6.Ada Boost: -
AdaBoost (Adaptive Boosting) works on improving the areas where the base learner fails. The base learner is a machine learning algorithm which is a weak learner and upon which the boosting method is applied to turn it into a strong learner. Any machine learning algorithm that accept weights on training data can be used as a base learner. AdaBoost works on improving the areas where the base learner fails. The base learner is a machine learning algorithm which is a weak learner and upon which the boosting method is applied to turn it into a strong learner.

7.Gradient Boosting: -
Gradient boosting doesn't modify the sample distribution. Instead of training on a newly sample distribution, the weak learner trains on the remaining errors (so-called pseudo-residuals) of the strong learner. It is another way to give more importance to the difficult instances. At each iteration, the residuals are computed and a weak learner is fitted to these residuals**.**

8.XGBoost: -
XGBoost (Extreme Gradient Boosting) is an optimized distributed gradient boosting library. It uses gradient boosting (GBM) framework at core. Yet, does better than GBM framework alone. XGBoost implements parallel processing and is faster as compared to GBM. XGBoost has an in-built routine to handle missing values. XGBoost tries different things as it encounters a missing value on each node and learns which path to take for missing values in future.

III. EXPERIMENTAL ANALYSIS (TABLES AND GRAPHS)

1.Min-Max Scalar: -
Min-max scaler for non-normalized features, get dummies for all features and feature integration.

2.PCA: -
2.1 Since the data is huge, PCA is used to see whether this can improve our model performance. From the above results it is observed that Accuracy, precision, recall and f1-score are increased when compared with base model. After performing Cross Validation for PCA model with CV=5 and scoring = "roc_auc", below are the results.

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

2.2 Although we got good results, we want to try with "Select KBEST" method for feature selection and proceed with model building because after implementing PCA on the dataset, our original features will turn into Principal Components. Principal Components are the linear combination of our original features. Principal Components are not as readable and interpretable as original features.

3.Select K-Best: -

Statistical tests such as Chi square and t-test for the dataset, the result of test shown that all the features are significant with respect to target variable. In Select K-Best method we specify the number of features and the method returns the most significant amongst them.

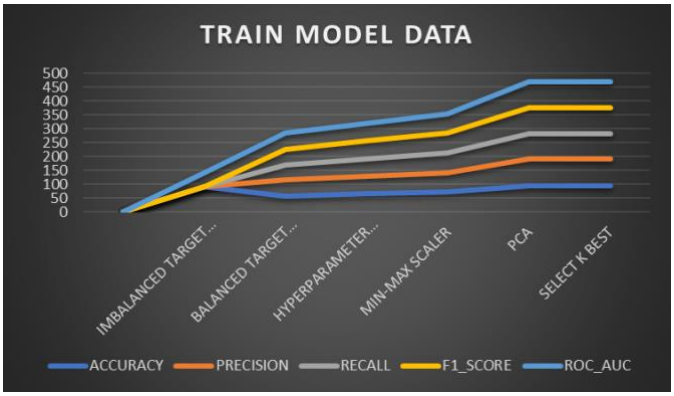
We had number of trials with k=80, 90,100 and so on. We got good score for k=100 value. Although the scores will be more if we go beyond k=120, but there would be much variance error in the scores. So choosing optimal k value can be done only through trial and error method. After performing Cross

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
```

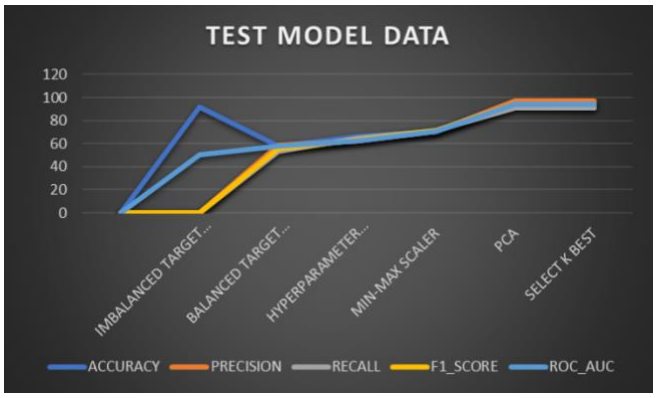
As our scores are almost same as PCA, here we are able to interpret the features by backing them with statistical analysis. So going further we tried different ensemble methods on K Best model to see if we can improve the score.

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

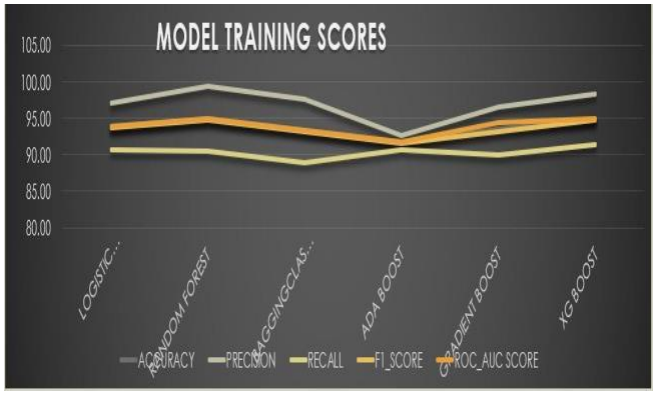


TEST MODEL	ACCURACY	PRECISION	RECALL	F1_SCORE	ROC_AU
IMBALANCED TARGET LOGISTIC REGRESSION	91.92	NA*	0	0	
BALANCED TARGET LOGISTIC REGRESSION	57.8	58.5	53.6	56	
HYPERPARAMETER TUNING FOR LR	65.83	62.1	64.4	63.2	
MIN-MAX SCALER	70.9	70.5	71.6	71.1	
PCA	94.3	97.3	91.1	94.1	
SELECT K BEST	94	97.1	90.7	93.8	

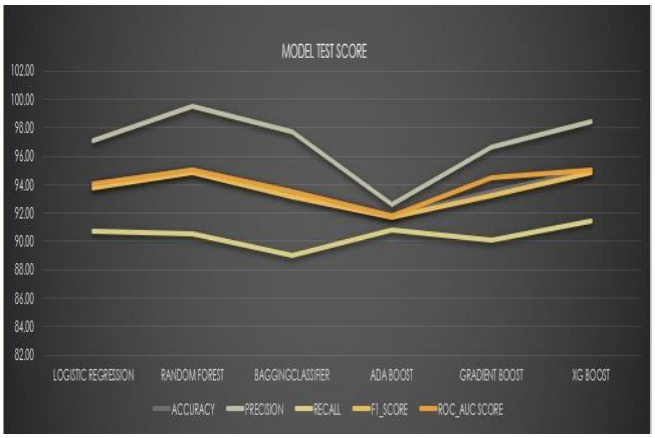
NA* : - It is an edge case, model hasn't predicted any positive cases due to class imbalance



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



CONCLUSION

Credit risk assessment is only possible by means of measurement. Machine learning models can be used as tools to measure the credit risk exposure of various financial institutions. With the correct prediction of credit risk, its management will become effective and efficient. This 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. There are many classifier models we have approached which are discussed in the report. We can say conclusively that XG-Boost is the model that performed well 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. However we have also tried K-Best technique to determine the feature importance. Feature integration has also been 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. K-Best feature selection was enacted to check the important features and PCA was applied on the data because of its high dimensionality. Bagging and Boosting methods has been tried to check whether it improves the score or not. After comparing the performance of all the models it is concluded that XGBoost is the best model which is performing well. This project opens the doors for further research in the credit risk using the deep learning models and there is a scope of smoothing the process of decision making in credit risk.

REFERENCES

[1] https://www.kaggle.com/c/home-credit-default-risk - Data set link.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html-PCA

[3] https://scikit-learn.org/stable/supervised_learning.html#supervised-learning-Machine Learning Models

[4] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html- Feature Engineering

[5] https://docs.scipy.org/doc/scipy/reference/stats.html- statistical tests

[6] https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/-smote

[7] https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms - Hyperparameter tuning

[8] https://en.wikipedia.org/wiki/Decision_tree - Decision Tree

[9] https://en.wikipedia.org/wiki/Ensemble_learning - Ensemble Learning models

[10] Machine Learning by Tom m Mitchell.

[11] Hands on Machine Learning with Sckit-Learn and TensorFlow by Aurelien Geron.

[12] https://towardsdatascience.com/cross-validation-and-hyperparameter-tuning-how-to-optimise-your-machine-learning-model-13f005af9d7d?gi=1d33882a8888 - Hyperparameter Tuning