Predicting the Future Transaction from Large and Imbalanced Banking Dataset

Sadaf Ilyas[1], Sultan Zia[2]
Zaib un Nisa[5]
University of Lahore
Gujrat Campus, Gujrat, Pakistan

Umair Muneer Butt[3]
Universiti Sains Malaysia
University of Lahore
Gujrat Pakistan

Sukumar Letchmunan[4]
Universiti Sains Malaysia
Penang, Malaysia

Abstract*—Machine learning (ML) algorithms are being adopted rapidly for a range of applications in the finance industry. In this paper, we used a structured dataset of Santander bank, which is published on a data science and machine learning competition site (kaggle.com) to predict whether a customer would make a transaction or not? The dataset consists of two classes, and it is imbalanced. To handle imbalance as well as to achieve the goal of prediction with the least log loss, we used a variety of methods and algorithms. The provided dataset is partitioned into two sets of 200,000 entries each for training and testing. 50% of data is kept hidden on their server for evaluation of the submission. A detailed exploratory data analysis (EDA) of datasets is performed to check the distributions of values. Correlation between features and importance of characteristics is calculated. To calculate the feature importance, random forest and decision trees are used. Furthermore, principal component analysis and linear discriminant analysis are used for dimensionality reduction. We have used 9 different algorithms including logistic regression (LR), Random forests (RF), Decision tree (DT), Multilayer perceptron (MLP), Gradient boosting method (GBM), Category boost (CatBoost), Extreme gradient boosting (XGBoost), Adaptive boosting (Adaboost) and Light gradient boosting (LigtGBM) method on the dataset. We proposed LighGBM as a regression problem on the dataset and it outperforms the state-of-the-art algorithms with 85% accuracy. Later, we have used fine-tune hyperparameters for our dataset and implemented them in combination with the LighGBM. This tuning improves performance, and we have achieved 89% accuracy.

Keywords*—Machine Learning (ML); banking; Santander; transactions; prediction; imbalanced; unbalanced; skewed; hyperparameter; oversampling; undersampling; EDA; dimensionality reduction; PCA; LDA; LR; RF; DT; MLP; GBM; CatBoost; XGBoost; AdaBoost; LigtGBM

I. Introduction

The invention of ML has become an essential advancement in technology, which has opened several doors to success for all areas. As technology changes rapidly, it is necessary to keep oneself up-to-date in this ever-changing era. The speed of technological advancement has progressed exponentially. Presently, the opportunity has arrived, and machines are getting ready to make judgments at their very own, and it's all because of artificial intelligence or particularly ML [1].

The development of skills guides us to live in an epoch where the achievement is not just about the present but to excel; you have to think like a chess player. The financial sector is one of the substantial areas in terms of value as well as the production of large size of data in seconds. Furthermore, it needs real-time decision making according to the situation [2]. To understand this, we take the simplest example of traditional banking systems, in which one has to visit the bank and give a check to the cashier to withdraw money during the banking hours. In this process, every single transaction was made/processed by humans. But now, with the start of e-banking, the situation has become very complicated. No doubt, e-banking has opened new horizons and helped us in many ways (e.g., ATMs, online money transfer, online shopping, etc.) to make it possible for making transactions whenever we want and wherever we are. This causes an exponential increase in transacted data produced by banking (no time limitation). Over time, e-banking services evolve and increase day by day (Fig. 1) [1].

This new system also causes some problems, e.g., Security, authentication, vulnerability, extensive data, no time limit, monitoring, etc. [3]. To protect the system from these problems seems near to impossible for humans in terms of time, cost, and efficiency. The need for intelligent machines was felt because humans could not achieve the processing of millions of transactions in seconds. To accomplish such goals, ML has been introduced in banking to make decisions on the spot on behalf of humans even when humans are not supervising them. This is not the end. This necessity has become a powerful tool, and much more is being achieved by using it.

This paper consists of techniques to handle class imbalance while improving the accuracy of prediction. Rest of the paper is organized as Related work, Problem statement, Methodology (including data analysis, techniques to handle imbalance and implementation of classification algorithms, dimensionality reduction & resampling), Results, finding Hyperparameters and combined the best-performed algorithm (based on results), Conclusion and Findings & Finally Future Work.

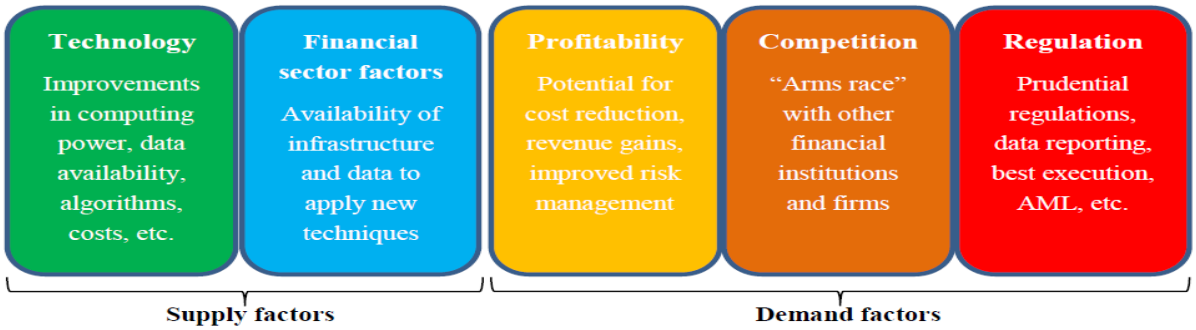

Fig. 1. AI and Machine Learning Factors for Financial Adoption [1].

II. Related Work

Since the ML has been introduced in banking, it has revolutionized this vital sector. The idea of using ML in banking was first given to make financial decisions using neural networks by Hawaly et al. [2], which laid the foundation for the automation of the banking sector. Nowadays, machine learning is not only used for automation, but also for predicting the future, finding customer loyalty, information on standard recovery rates [3], learning optimal coverage rates[4], modeling investor sentiment [5], fraud detection [6], detecting the stock price [7, 8], client maintenance, programmed credit endorsement, extortion discovery, showcasing and hazard the executives in the financial part [9]. Banks keep up a lot of data about their clients. This information can be utilized to make and keep up an unmistakable relationship and association with clients to manage them exclusively for specific items or banking offers. Information mining [10] has turned out to be prevalent for illustrative and prescient applications in financial procedures.

Reverberating its future advantage to banking, various early investigations have additionally shown up in the journal of banking and finance during the 1990s, which investigated the potential for ML to improve loaning choices and credit chance management. Altman et al. [11] connected neural systems to order Italian firms dependent on the probability of budgetary trouble, while Varetto [12] based on this examination by applying hereditary learning calculations to a similar subject. Later research into account, diaries has maintained the attention on the forecast; however, it has moved towards profound learning methods, and other propelled ML procedures. These ongoing applications include: the comprehension of default recuperation rates cheng and cirillo [3] learning ideal choice supporting standards Nian et al. [4] demonstrating financial specialist feeling Renault [5] and the discovery of stock value advancement dependent on request books Kercheval and zhang, 2015 [13]. As fund diaries have made provisional strides to recognize the capability of the new methods of ML in an account, different orders have been attempting progressively strident endeavors to apply ML ways to deal with budgetary information. To some extent, this is because of the appeal of the exhaustive, organized, and effectively open, information accessible in money. Examine on ML and account outside of money diaries are surpassed by an extensive various ML and fund explore in fund diaries.

There can be many factors supporting the use of ML just like other disciplines are using, but being a financial institute, there are also some supply and demand factors. The organization's essential reports are fiscal summaries that dismiss its monetary status doer [14]. The financial proclamation is the fundamental reason for choice-making concerning an immense number of speculators, leasers, and different people needing bookkeeping data, just as a robust articulation of business execution, financial status, and the social obligation of recorded organizations and otc organizations [15].

Be that as it may; lately, instances of deceitful financial proclamations have turned out to be progressively genuine (wells [16]; Spathis et al. [17]; yeh et al. [18]). There have been numerous instances of false financial articulations in us & Taiwan since the Asian financial crisis in 1997. Enron case in 2001, the Abit computer, pro comp, info disc, and summit technology cases in 2004 in Taiwan and WorldCom case in 2003 in us. Given these episodes, it has turned out to be essential to have the option to distinguish fraudulent conduct before its event. Information mining is a critical apparatus for managing complex information investigation and classification. It identifies profitable occasions that are covered up in a lot of information for examination and abridges the information in an organized model to give a reference to essential leadership. Information mining has numerous different capacities, for example, classification, affiliation, grouping, and estimating [19].

Some explorers have been made in the field of client weakening and maintenance investigation in banking segments. A few examinations uncover that the most significant factors impacting client decision are compelling and proficient client administrations, quickness and quality administrations, assortment of administrations offered and little e-administration charges, web-based financial offices, wellbeing of assets and the accessibility of innovation-based service(s), low financing cost on advance, advantageous branch area, picture of the bank, well administration, and generally bank condition [20, 21]. Then again, the client is the center of their activity, so sustaining and holding them are significant for their prosperity. Numerous looks into were hung on client maintenance just as client weakening investigation lift is utilized as an appropriate measure for regular loss examination, supported credulous baye's system, specific baye's system, neural network system, and the above classifiers techniques [22]. Their first spotlights were on constant loss investigation

utilizing the lift. It can be determined by taking a gander at the total targets threshold up to p% as a level of everything being equal and isolating by p%. An agitate model with a prophetic presentation and membership set built to help machines [23]. They demonstrated that help vector machines show excellent speculation execution when connected to uproarious showcasing information. The model beats a strategic relapse just when the suitable parameter-determination method is compared, and the arbitrary woodlands outperform SVMs. Soeini et al. [24] chose 300 records of clients and by using clementine software for customer churn prediction. In this paper, statistic factors are used to decide the ideal number of bunches in k-implies grouping and assessed double characterization techniques (quest, c5.0, chaid, cart, Baye's systems, neural network systems) that foresee clients agitate. Goonetilleke et al. [25], in his paper, used DT and NN to create a model that anticipates stir. The presented models are evaluated using the roc curve and AUC deem. They likewise embraced cost touchy learning methodologies to address imbalanced class marks and unequal misclassification costs issues. [26] talked about business bank client stir forecast dependent on the svm model, and utilized an arbitrary testing strategy to improve the svm model, considering the unevenness attributes of client informational collections.

An examination explored determinants of client agitate in the Korean portable media communications administration showcase dependent on client exchange and charging information. Their investigation characterizes variations in a client's position from dynamic to no-use or suspended as partial surrender and from effective use to beat as complete deserting. Results demonstrate that a client's status change clarifies the connection between beat determinants and the likelihood of churn [27]. A Neural Network (NN) based system on dealing with foreseeing client stir in membership of remote cell administrations. Their aftereffects of analyses show that neural network-based methodology can anticipate client stir with precision over 92% [28]. A scholastic database was built between 2000– 2006 covering 24 diaries and proposes a grouping plan to arrange the articles. Nine hundred items were recognized and checked on for their immediate significance in applying information mining strategies to CRM. They found that the examination territory of client maintenance got most research consideration, and order and affiliation models are the two regularly utilized models for information mining in CRM [29]. A study on the idea of data mining and customer relationship management in sorted out banking and retail enterprises was additionally talked about [30].

Banking fraud is a profound term having many types. Zhou et al. [31] have studied several traditional machine learning algorithms, for fraud detection, they choose an improved version of gradient boosting decision tree within bankcard enrollment on the mobile device based payment for use in a real system, namely, XGBoost. K. Seeja and m. Zareapoor [6] proposed an intelligent model that can efficiently detect credit card fraud of data sets of anonymous and highly unbalanced transactions for credit card transactions using frequent extraction of sets of elements.

To find the fraudulent in the incoming transaction of a particular client is closest, an adjustment algorithm is also proposed so that a decision can be made accordingly. Furthermore, they used SVM, NB, KNN& RF on the same data, and when compared, the proposed model performed better than all of these. Another plus point for this model was that it could handle class imbalance too. Cheng et al. [32] proposed a fraud detection framework based on CNN that has intended to record the intrinsic patterns of fraud behavior extracted from tagged data. The data of abundant transactions are represented by a matrix of functions, to which a convolution neural network is applied to identify a series of underlying patterns for each sample. On the other hand, [33] the researcher has conducted research and analyzed fraudulent data by applying DT, BBN, SVM, and ann. In this research, the author has used two stages of statistical treatment to produce more accurate results, unlike other single-stage algorithms.

Many efficient and effective techniques have been presented based on the decision tree for classification [34] and prediction. There are many algorithms to build a decision tree model [35]. The first stage consists of the most influential variable selection by using two primary decision tree techniques, cart & chaid. Whereas, in the second stage DT (cart & chaid), BBN, SVM, and ANN are used to detect fraudulent transactions. Sezer et al. Proposed an image recognition technology to predict technical stock patterns [7]. The proposed model is a CNN model that performs five functions, dataset pre-processing (extract/transform), data labeling, image formation, and CNN performance. The goal of this practice is to determine the best fit for the buy, sell, and hold points in the time series of the associated stock prices. However, [8] the researcher has used online data to make a knowledge base by combining with ensemble methods, e.g., Neural network regression ensemble (NNRE), support vector regression ensemble (SVRE), boosted regression tree (BRT) and random forest regression (RFR) to predict short-term stock prices.

Moro et al. [36] worked with a vast arrangement of information gathered from 2008 to 2013 from a Portuguese retail bank, including the ongoing money related emergency. They broke down many capacities identified with the qualities of the bank's customers, items, and economics. Because of a choice of self-loader works that were researched in the displaying period of their strategy, performed with information before July 2012, the informational collection was diminished to 22 capacities. They likewise thought about four dm models (logistic regression (LR), decision trees (DT), neural network (NN) and support vector machine (SVM)) utilizing two measurements (region of the collector working bend (AUC) and territory of the total bend of lift) from which NNexhibited the best outcomes (AUC = 0.8 and lift = 0.7), with the goal that 79% of the endorsers could become to by choosing the best-positioned customers.

As per turban et al. [37], business intelligence incorporates models, devices, databases, applications, and philosophies to utilize information to help the choices of business chiefs. Information mining is a business insight innovation that utilizations information-driven models to remove helpful learning, that is, examples of intricate and enormous information accumulations [38]. Chitra and subashini [39] have utilized a few information extraction calculations for client maintenance, programmed credit endorsement,

misrepresentation location, advertising, and hazard the executives in the financial division. They have distinguished a few strategies and models to improve client maintenance and misrepresentation discovery.

Hu [22] applied information mining strategies to assist retailers with weariness examination to recognize the scope of customers with a high likelihood of acting. He utilized the choice tree (DT), guileless invigorated bayesian system, particular bayesian system, the neural system as an information mining model. Ghosh et al. [40] from a company that implements a specific "bonus program". 2 algorithms have been used (Naive Bayes improved with ADA and c4.5 developed with ADA with CF) that too generates probability. Tool for credit scoring is designed to speed up loaning judgments, while possibly preventing incremental risk with the help of machine learning. In any case, money lenders or banks are increasingly going to extra, amorphous and semi-organized information means, containing online life action, usage of cell phone and instant message action, to get a more nuanced perspective on reliability, and improve the score exactness of credits.

Implementing AI computations to this set of stars of new information has empowered assessment of subjective factors, for example, employment conduct and readiness to pay. The ability to use additional information on such measures takes into consideration a more significant, faster, and inexpensive partition of mortgagor quality and ultimately stimulates a quicker credit choice [41]. Ling and li [9] utilized information-digging procedures for direct promoting in three informational collections from three distinct sources. The primary informational index was for the advancement of an advance item in Canada. The following informational collection was from a considerable disaster protection organization, and the third informational index was 260 s. A great deal of work has been finished utilizing ML in banking, and many have utilized distinctive managed, solo and troupe strategies to accomplish their objectives, be that as it may, some have used to consolidate regulated and unaided procedures, for example, e.g., Wang et al. [42] have proposed a half breed procedure for credit scoring of the clients.

Table I shows the details about the banking-related tasks and the algorithms used to perform these tasks.

TABLE. I. MACHINE LEARNING TECHNIQUES USED IN BANKING TO ACHIEVE DIFFERENT GOALS

Task	Methodology	Year	Reference
Fraud Detection & Prevention	(SVM, KNN, NB, RF) , CNN, CNN, Clustering	2014,2016, 2016, 2013	[6], [32],[33], [39]
Stock Prices	CNN, (NNRE, SVRE, BRT, RFR)	2018,2018	[7], [8]
Risk Management	Bayesian Model	2017	[43]
Crisis Management	LR, DT, NN, SVM	2011	[36]
Customer Attrition	DT	2005	[22]
Banks Retailing		2010, 2005	[37], [38]
Customer Retention	DT	2013	[39]
Credit Approval	(DT, SVM, LR), DT	2015	[39], [44]
Marketing	AB, NB	1998	[9]

III. PROBLEM STATEMENT

A US-based bank named Santander has recently conducted a competition with the help of an ML and data science competition site KAGGLE. The bank shared its data by keeping the attributes hidden. This is a binary classification competition; in this competition, we need to predict whether a customer will do the transaction or not. The dataset that is shared by the organizer has numeric data fields.

IV. WORK / METHODOLOGY

In this research we have worked on a recent dataset of a bank, analyzed it for distributions of mean, standard deviation, skewness, and kurtosis, check for feature correlations and importance and tried to reduce dimensions of data by implementing principal component analysis (PCA) as well as linear discriminant analysis (LDA). Furthermore, to tackle class imbalance, a variety of methods, including evaluation metrics, a variety of different classifiers, and resampling, are used. Moreover, to increase the overall prediction, hyper-parameters are also computed for the best performing algorithm in combination with the most suitable evaluation metric for our dataset.

All the implementations are done using python 3.7.1 on anaconda3 & Jupiter notebook 5.7.4. Also, the machine used in this whole process is a Fujitsu core i3 2nd generation. All the algorithms are taken from sci-kit learn [45].

A. Dataset

The dataset [46] was taken from the Santander transaction prediction competition from the KAGGLE site. Dataset set has two different files, consisting of 200,000 entries each. The difference in both files was that one file has the target too, to use for training the model while the other file didn't have a target variable that can be used for making predictions. To check the prediction results, one can upload the results on the competition site to see how accurate the predictions are made. Both the datasets had an id column as well as 200 variables to make predictions while the training dataset one extra column consisting of the target. According to the information, the dataset has 800,000 entries in total, which is quite large [47].

B. Pre-Processing

Data pre-processing is a process that manipulates raw data for further implementations on the dataset. It involves handling incompleteness, inconstancy, and errors in data. In real-world

data are usually imperfect, varying, duplicate, and noisy. It may have lacking attribute values or may contain aggregate data only. There can be errors or outliers in data that may change affect the overall distribution of data, repeated values, and differences in attribute names can also be problematic. As these issues of data may lead us to wrong predictions for this, we use data pre-processing because it is an excellent approach to resolve such matters before occurring. It was seen while pre-processing that the dataset is entirely clean, with not a single missing value in both train & test datasets. No inconsistency, redundancy, or noise found. Furthermore, all the attributes have floating-point (numerical) values, and they're not a single categorical attribute.

C. Exploratory Data Analysis (EDA)

Exploratory data analysis, well-known as EDA, is a tactic for examining datasets to summarize their core features, often with realistic approaches. EDA is used for getting summarized insights about data to provide us in-depth knowledge about is which in the future we can use in the modeling task. These insights can be graphical as well as non-graphical. But graphics give us more details as compared to non-graphic approaches; graphical representations of data are preferred more. It is tedious and time-consuming to look at a value of attributes or a whole spreadsheet and conclude the vital characteristics of the data. It might also be annoying and overwhelming to extract insights by seeing ugly numbers. EDA techniques have been developed as an aid in this situation. Chris Chatfield [48] strongly encouraged to use of EDA to describe the data and formulate the model based on this data.

In this phase, we started by checking the min and max of data. Min and max refer to the minimum value as well as the maximum value. It was discovered that the datasets have the data types far more than the requirements of the data, hence consuming more memory not just on disk, but it requires more time to process the data (Fig. 2 and Fig. 3). To handle this data more effectively and to reduce the memory and time cost, data types have been changed with the help of memory reduction function.

This worked well and helped in saving almost 50% of the memory occupied by both datasets, as this process repeated once again but with the other dataset, to be used for testing. Reducing the memory cost will also reduce the computation power & time required. Table II and Table III are given below showcase the dataset along with the memory before and after memory reduction of both train and test datasets.

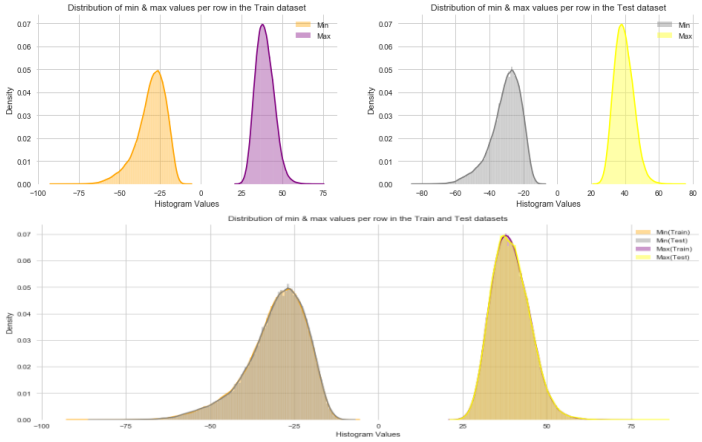

Fig. 2. Distribution of Min and Max Values Per Row in Train & Test Datasets.

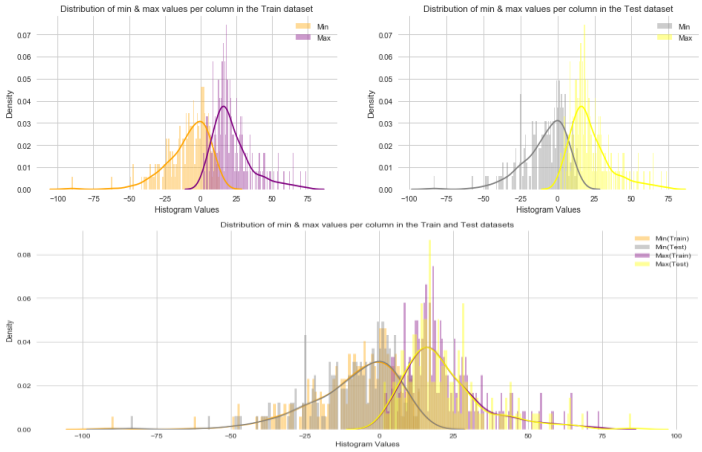

Fig. 3. Distribution of Min and max Values Per Column in Train and Test Datasets.

TABLE. II. REDUCING MEMORY OF TRAINING DATASET BY CHANGING DATA TYPES

Training Dataset				
Columns	Before Preprocessing		After Preprocessing	
	Data types	Memory Utilized	Data types	Memory Utilized
ID_code (1)	Object	308.2276153564453 MB	Object	154.30458068847656 MB
Variables (200)	Float64		Float32	
Target (1)	Int64		Unit8	

TABLE. III. REDUCING MEMORY OF TESTING DATASET BY CHANGING DATA TYPES

Testing Dataset				
Columns	Before Preprocessing		After Preprocessing	
	Data types	Memory Utilized	Data types	Memory Utilized
ID_code (1)	Object	306.7017364501953 MB	Object	154.1138458251953 MB
Variables (200)	Float64		Float32	

Furthermore, a complete analysis of datasets was conducted in which datasets have been analyzed for four major components of descriptive statistics; mean, standard deviation (sd), skewness and kurtosis, and their distribution in the train as well as test dataset, which showed that both datasets are almost identical. But an important thing finds out during analysis was that the distribution of class 0 and 1 is skewed, making this data imbalanced.

D. Feature Correlation

Feature correlation is another statistical method that can check whether the features are related to each other and, if so, how strongly these are associated. The name "correlation" denotes a universal relation among variables. Frequently, correlation is considered as the paramount step to understand these relations for building better statistical models. It can provide support in foreseeing one variable from another. It shows the existence of a contributory association, which is also used as a critical point and a base for several other modeling methods.

During analyzing the correlation between variables, we find out that there is not much correlation among these, neither positive nor negative, and the variables of both datasets are independent. Variables that are correlated, either positive or negative, are known as the dependent variable. Figures depict the correlation between the variables of each train & test datasets that how many variables are correlated. The scale on the right shows the correlation level and color. As there is not too much correlation or least correlation, that is why the color of the variables are in black. The only light color in figures is the correlation of a variable with itself (Fig. 4 and Fig. 5).

E. Feature Importance

Datasets were also analyzed to get knowledge about the essential features of these, which participates more in the decision-making process. For these two different techniques were used to see the vital features of data as well as the working of classification methods on the same data. Decision tree and random forest classifiers were selected for this task. By finding the importance of features, it was found out that "var_110" is an essential feature when working with a random forest classifier while the "var_86" has the least importance. On the other hand, when working with a decision tree to find out essential features, a tree was formed, showing 6 levels starting from root to leaf nodes. The decision tree finds "var_81" is the most important one hence build the complete tree by selecting it as the root node.

F. Dimensionality Reduction

Variables or features of data are also known as dimensionality, and dimensionality reduction refers to dropping the no. Of variables and finding the least no. Of variables to analyze data more effectively because the complexity increases with the increase of dimensions. In ML, the problems caused by working with a higher number of aspects are known as the curse of dimensionality. So to diminish the complication and make it easier to work on a dataset having a large no. Of the dimensions, different methods are used. In this regard, a well-known technique is principal component analysis (PCA) that lets us review and to envision the facts in a dataset holding interpretations defined by multiple intercorrelated measurable variables by defining new variables or dimensions called principal components.

- Principal component analysis (PCA): The foremost objective of PCA is to recognize guidelines along which the deviation in the data is most. It is valuable when we have got data on a large number of features and trust that there is some redundancy among these features. Redundancy, in this case, refers to that more than one variable is correlated might measure the same construct. Based on this, we consider that we can shrink the observed features into a relatively small number of principal components, which will participate for maximum deviation in the observed features [49]. These newly created variables relate to a linear mixture of the original variables. These variables will always smaller in numbers as compared to the originals.

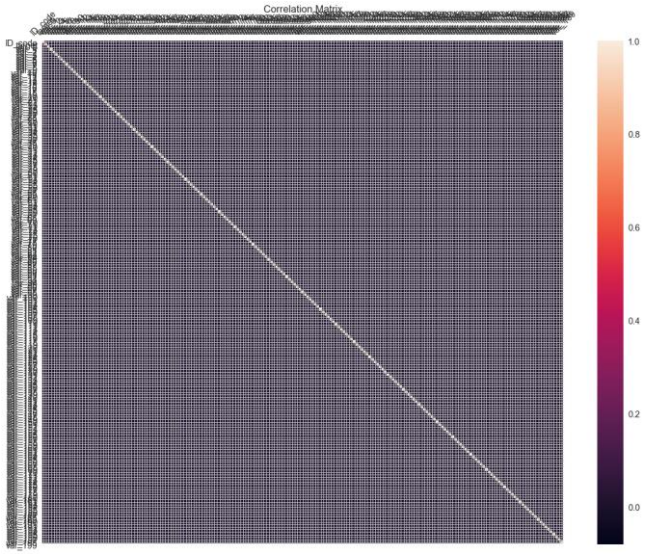

Fig. 4. Correlation between Features of Train Dataset.

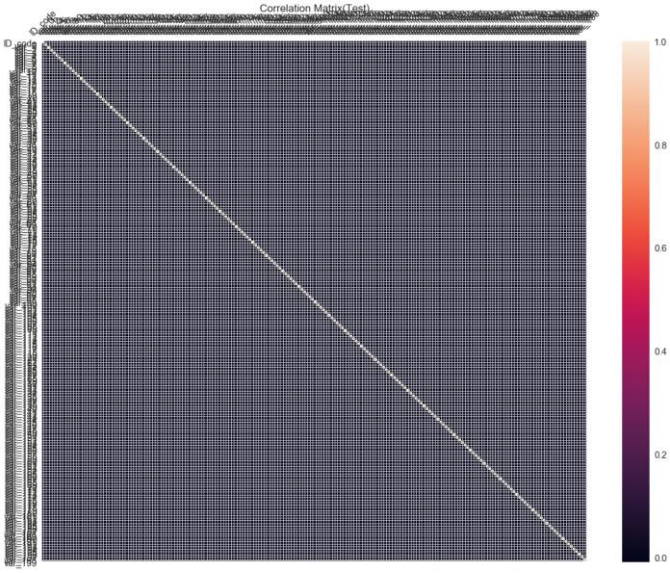

Fig. 5. Correlation between Features of the Test Dataset.

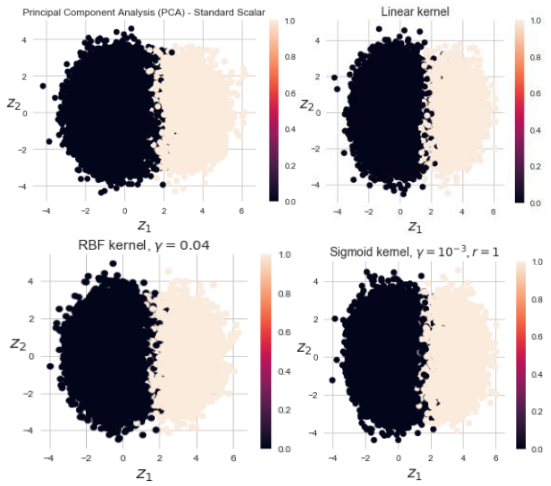

Fig. 6. Finding the Principal Component by using 4 different Techniques.

As can be seen, none of the applied PCA techniques could separate the "class0" & "class1" properly, not even reducing the dimensions to the least. Hence our PCA failed to meet the goal help us reducing the proportion to assist us in finding the principal components (Fig. 6).

- Linear discriminant analysis (LDA): It is worth mentioning here that after failing in achieving optimum goals for PCA, the researcher has tried another method to reduce the dimensionality, which is linear discriminant analysis (LDA) [50]. Both LDA and PCA are direct transformation techniques, but PCA is an unsupervised technique contrary to LDA, supervised. The LDA is quite similar to PCA. But as PCA helps in getting the axes of components that maximize the deviation of data, LDA helps in finding the axes that maximize the parting among classes [51]. LDA makes predictions assuming Gaussian distribution by taking the mean value for every type and considers variants [52]. The goal is to map a dataset with small dimension space but with good class separation to avoid overfitting along with reducing the computational cost. The plus point of LDA has over PCA is that it tackles overfitting too.

Unfortunately, just like PCA, LDA also couldn't help us in feature subspace, which could maximize the separability of our classes. During analyzing the data, another thing was noticed that this is a binary problem as the target may have only two conditions, either the client will make the transaction (1) or not (0), but there was an imbalance (Fig. 7).

G. Class Imbalance

It was found while analyzing data that this dataset has only 2 classes 0 & 1, making it as a binary problem the dataset is imbalanced. As both classes are not proportionate, there are only two classes, and both classes were not properly balanced, as out of 200,000 entries, 179,902 (89.95 %) had the target as 0, while the rest of the entries had the target value 1 which hardly make 10.05 % of aggregate data. The graph below shows the distribution of the training dataset. Class imbalance, as its name depicts, is the unbalanced distribution of classes [53] [54] (Fig. 8).

To address this problem and to get the maximum performance out of any algorithm when predicting, a variety of techniques are used:

- Collecting more data: The major and leading step in avoiding class imbalance is to collect as much data as possible. But in our scenario, the data is already quite enough, as well as is the property of a bank and cannot be collected directly due to several reasons, e.g. Lack of resources, lack of access to clients and lack of information about the parameters or variables.
- Evaluation metrics: Another way to avoid this problem is to consider the performance measure, which can help us not to prevent biases and predicting and understanding the actual performance of classifiers [54]. These performance measures include area under the curve (AUC) and confusion matrix etc. As the class "0" have 90% of the population, if a classifier does nothing and says that 100% data belongs to class "0", we still get 90% accuracy. So, accuracy is not a good measure while working with class imbalance. On the other hand, the confusion matrix shows the results as:

True positives (TP)

Identified as positive that is positive

False positives (FP)

Identified as positive that is negative

True negatives (TN)

Identified as negative that is negative

False negatives (FN)

Identified as negative that is positive

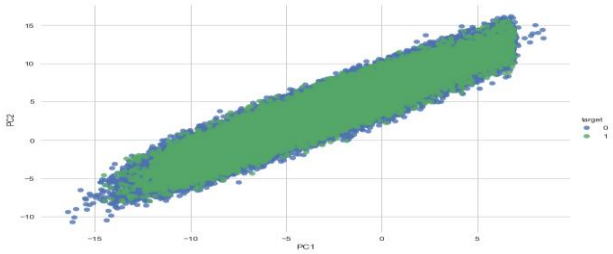

Fig. 7. Linear Discriminant Analysis for Santander Dataset.

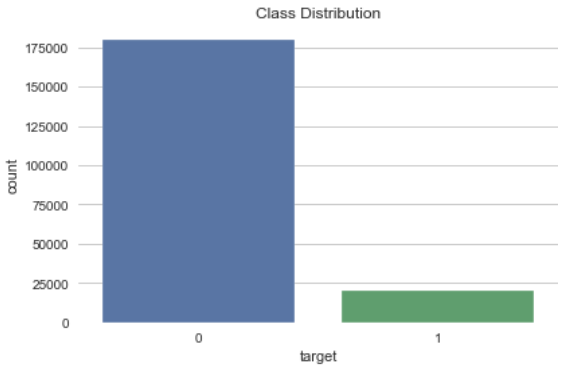

Fig. 8. Bar Graph Representation of the Class Distribution.

Based on these outcomes, we can check the effectiveness of our model by using recall, precision & f1 score. Recall is the number of true positives (TP) divided by the number of true positives (TP) plus the number of false negatives (FN). Recall is also known as the true positive rate (TPR) & sensitivity. Whereas, precision is defined as the number of true positives (TP) divided by the number of true positives (TP) plus the number of false positives (FP). Precision is also called true negative rate (TNR) & specificity (Eq. 1 & 2).

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

There is a trade-off in the metrics when we increase the recall, precision decreases. To find an optimum combination of precision and recall, we can combine the two parameters using what is called the f1 score (Eq. 3).

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (3)$$

The outcomes of the binary classification can be represented visually through confusion matrix (cm), receiver operating characteristics (roc) & area under the curve (AUC). Cm depicts the actual and predicted labels from a classification problem as a matrix, having actual classification on one side and predicted on the other (Table IV).

On the other hand, roc & AUC are graphical representations of true positive rate (TPR) & false-positive rate (FPR). Roc curve plots the true positive rate (TPR) versus the false positive rate (FPR) as a function of the model's threshold for classifying a positive and AUC metric is used to calculate the general performance of a classification model based on the area under the roc curve (Fig. 9).

- Classification algorithms: Classification methods are a vital part of data mining & machine learning applications. We chose different classifiers because of their different approaches to solve a problem. Every classifier cannot perform best on every type of dataset. A classifier outperforming others on a dataset is not meant that this classifier can outperform the others on any other datasets too.

So, we took some well-known classifiers (including simple as well as ensemble methods) to classify whether a customer will make a transaction or not. These classifiers include logistic regression (LR), random forests (RF), decision tree (DT), multilayer perceptron (MLP), gradient boosting method (GBM), category boost (CatBoost), extreme gradient boosting (XGBoost), AdaBoost and light gradient boosting process (LighGBM). Some of these gave a relatively good performance, while others hardly crossed the 50% accuracy. Among all these, LighGBM outperformed every other classifier. Initially, we did not specify any parameter for classification. We applied the most straightforward implementation to check the behavior of classifiers on our dataset. The roc curves of the algorithms mentioned above are shown (Figs 10-18).

When applied LighGBM, initially, it responded as a regressor and gave floating-point values instead of clear cut 0' or 1's. It is because when we imported LighGBM, we used it as it is, just like other algorithms. But the AUC was remarkably higher than all the previous classifiers applied to our dataset. Also, as the predictions did not classify the data in 0's or 1's, instead, it gave the floating-point predictions like regression; other metrics could not be calculated.

TABLE. IV. CONFUSION MATRIX

Confusion matrix		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

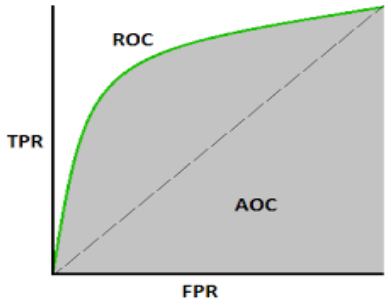

Fig. 9. The Area under the Curve (AUC).

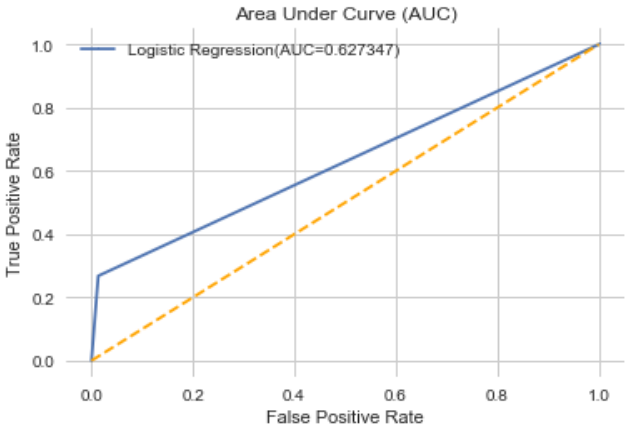

Fig. 10. AUC for Logistic Regression.

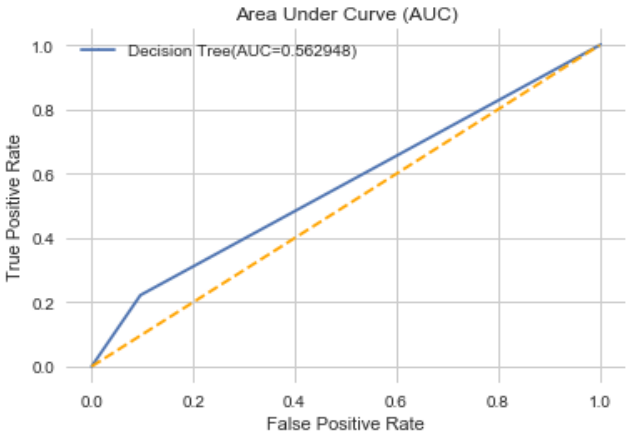

Fig. 11. AUC for Decision Tree.

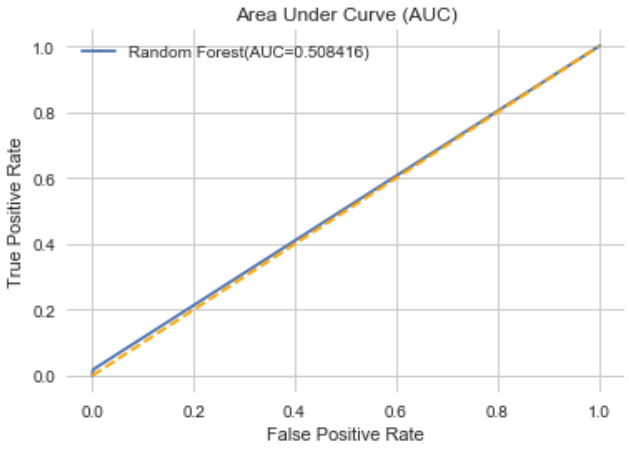

Fig. 12. AUC for Random Forest.

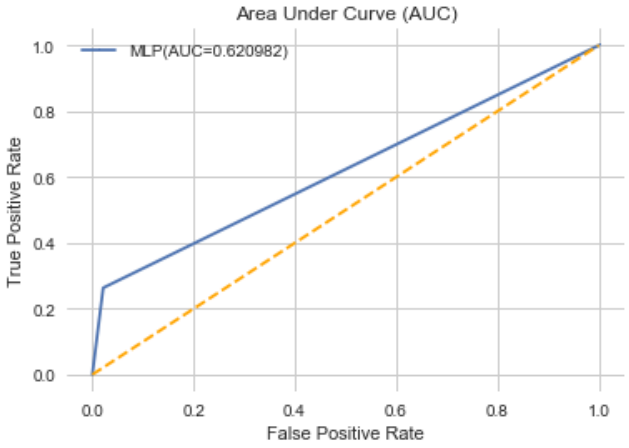

Fig. 13. AUC for Multi-Layered Perceptron.

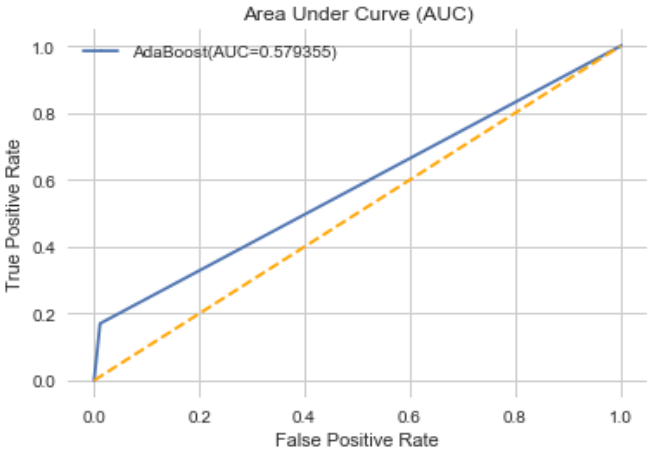

Fig. 14. AUC for AdaBoost.

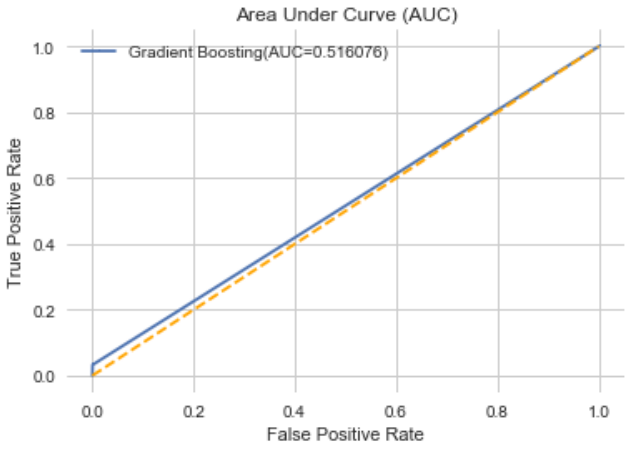

Fig. 15. AUC for Gradient Boosting Method.

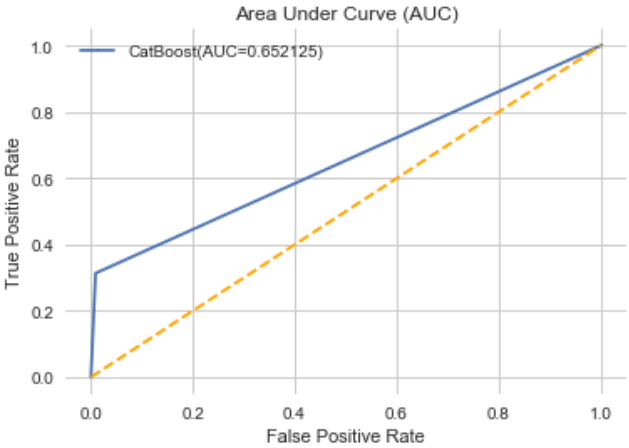

Fig. 16. AUC for CatBoost.

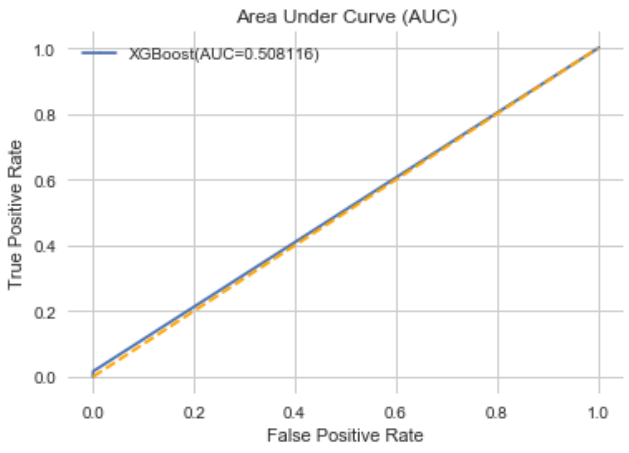

Fig. 17. AUC for XGBoost.

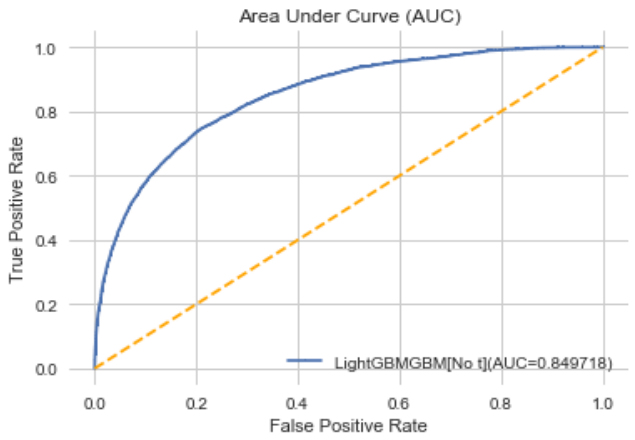

Fig. 18. AUC for LighGBM.

To calculate the values for other metrics and classify data in 0's & 1's, the researcher finds out the threshold after thorough analysis as well as researching and finds out the optimum value for threshold, which was round about 0.1025. By putting this value as threshold converted the predictions to classes 0's and 1's and out the precision, accuracy, recall, and f1 score, but the AUC dropped to 0.748172 (Fig. 19).

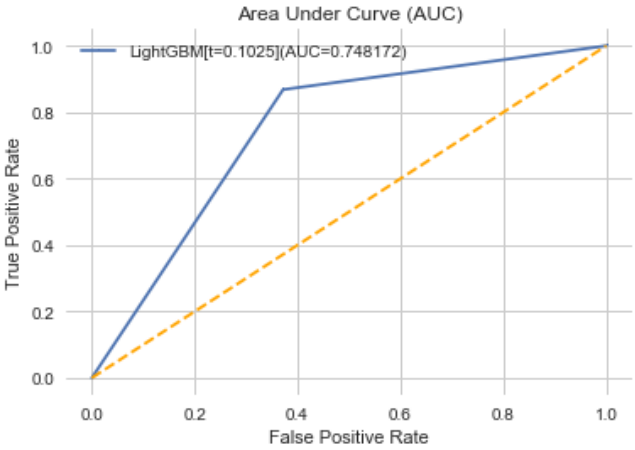

Fig. 19. AUC for LighGBM (Threshold=0.1025).

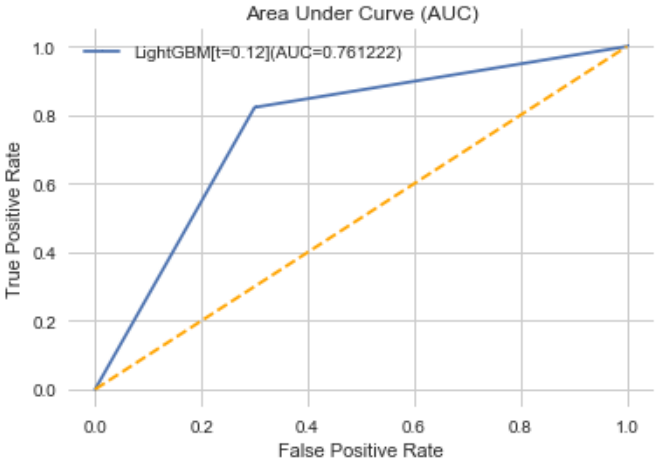

Fig. 20. AUC for LighGBM (Threshold=0.1200).

Further improvement in threshold value ended up on 0.1200 and gave somehow relatively good AUC (0.761222) but still less than without the threshold value (Fig. 20).

To get the binary output, we tried the LighGBM classification method too, which didn't perform well (Fig. 21).

One important thing we noticed here was that the problem was a binary problem; we had to predict the class of data as 0 or 1. But when we uploaded this file to the competition site, it still entertained the submission file, and even we got 85% AUC when submitted. So, as our ultimate goal was to maximize AUC, we kept using LighGBM as regressor because out of all these algorithms tried, LighGBM produced the best results, and the difference was remarkably high. The reason behind this is that LighGBM has a different structure, which leads it outperforming every other classifier tried on this dataset, including the other boosting methods too.

- Resampling: Resampling is a process of removing the imbalance of the classes either by increasing the number of smaller classes equal to the larger class or by decreasing the number of observations from larger classes to be equal to the smaller class. These are called undersampling & oversampling.
    - In undersampling, some of the observations from the majority class are deleted randomly to match the count with the minority class (Fig. 22).
    - In Fig. 23 results of LighGBM on the undersampled dataset can be seen.

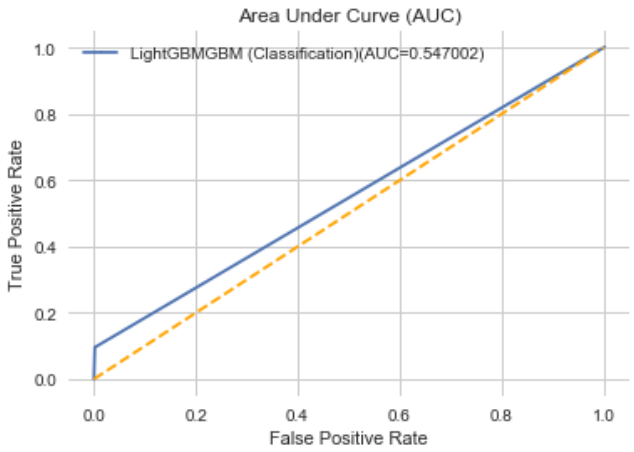

Fig. 21. AUC for LighGBM (Classification).

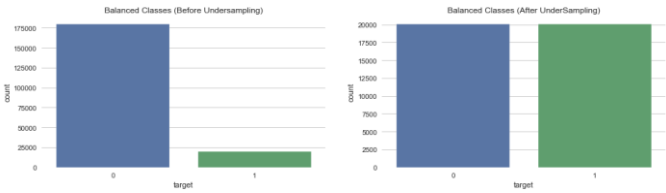

Fig. 22. Comparison of Imbalanced and Balanced (undersampled).

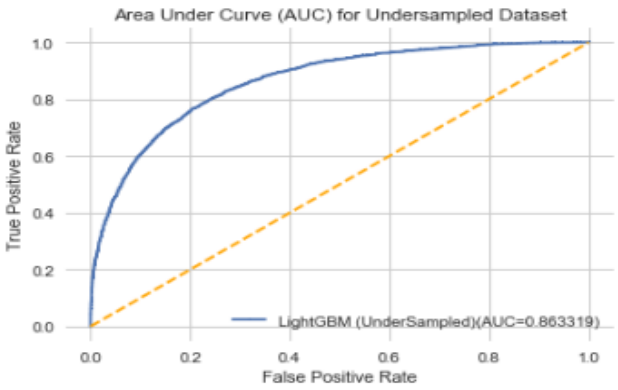

Fig. 23. LighGBM on the undersampled Dataset.

- Oversampling: On the other hand, oversampling is a process that is a little more complicated than undersampling. It is the process of increasing the observations of minority class to a level that matches the majority class (Fig. 24). This can be done by randomly copying the existing observations or generating artificial data that arbitrarily make a sample of the attributes from observations in the minority class. The frequently used technique is called synthetic minority over-sampling method or smote [55]. In simple terms, it looks at the feature space for the smaller class data points and considers its k nearest neighbors.
    - In Fig. 25 results of LighGBM on the undersampled dataset can be seen.

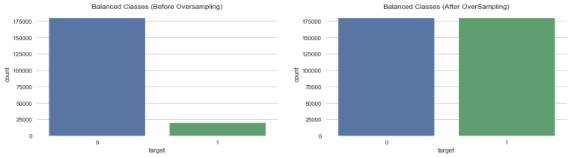

Fig. 24. Comparison of Imbalanced and Balanced (Oversampled).

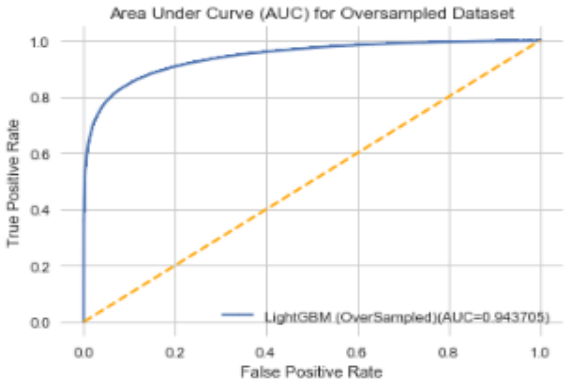

Fig. 25. LighGBM on the Oversampled Dataset.

V. Results

When initially implemented the classifiers, we got the results, as shown in Table II. Among all the nine classifiers, some performed better than others while some hardly crossed the random guessing limit or 50% accuracy. When we submitted the results of these predictions to the website for cross-checking, we got almost the same results there. Among all these random forests performed the least giving "0.50006" while LighGBM outperformed all classifiers tried in this research. LighGBM produced floating-point values as a target or solved it as a regression problem, which gave an excellent AUC (85.0000 %), but we could not find the training score, validation score, precision, recall & f1 score. For this, we went to set the threshold value to convert the target values to either 0 or 1. After thoroughly analyzing the predictions, we got that the threshold is somewhere near to "0.1". So we tried some values and observed the difference we got some excellent results (0. 748172 %) when we set the threshold (t=0.1025) while (76.1222 %) setting the value t=0.1200. These results of LighGBM were still better than every other classifier but significantly less than the results of LighGBM without setting the threshold value.

As we know that the AUC is the trade-off between the precision and recall, one increases when the other decreases. The results of precision and recall show that the implementations other than LighGBM show the precision values higher than recall except for decision tree, which had both the precision and recalls almost the same. But contrary to this, when implemented the LighGBM, we got the high recall values than precision. Getting that LighGBM is working well on our dataset, we selected this for further implementation on resampled datasets. When created resampled datasets, we first implemented the LighGBM once again now on the undersampled dataset. We got an increase in AUC by getting the (86.3319 %). The results were quite good, but we decided to implement LighGBM on an oversampled dataset too, which further gave us a 4.47% increase in AUC, and our AUC reached an excellent 94.3705 %. But when uploaded this result to the website, the result was slightly more than 73% (73.039%). This is because the dataset was synthetically oversampled by smote, which led our model to overfit it (Fig. 26 and Table V).

VI. Hyper-Parameters

To further improve the accuracy, we decided to select parameters, and instead of randomly guessing and manually adding the values, we decided to find hyperparameters for LighGBM. We have chosen these parameters and a range of benefits to test. This code was supposed to run for 1000 times, but it is about testing permutations and combinations on a large dataset, which is quite hectic and time-consuming. It took a long time to just run for 56 times out of 1000, giving about 89% AUC (Table VI).

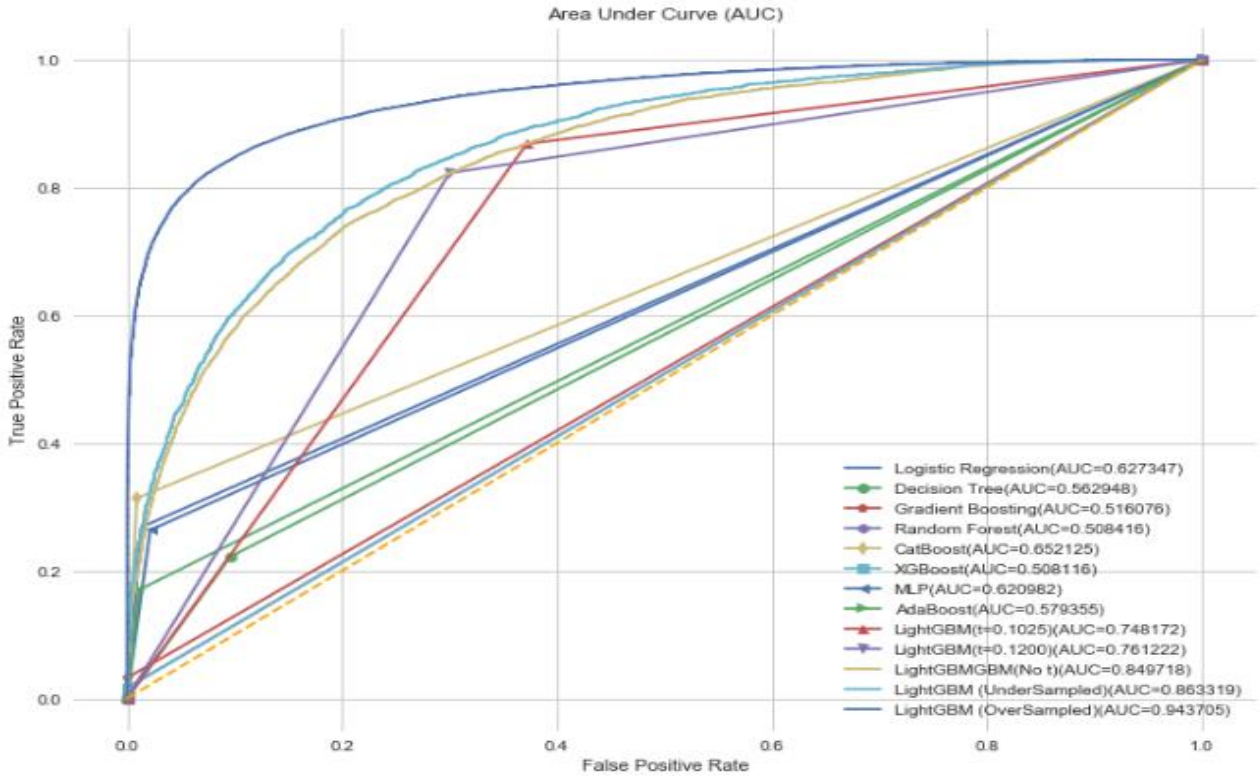

Fig. 26. Comparison Oversampled and undersampled Implementation with all Techniques.

TABLE. V. RESULTS OF CLASSIFICATION ON SANTANDER DATASET

Results of implementation of ML algorithms on Santander dataset (train=200,000, test 200,000)								Submission results
Name	**Training**	**Validation**	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Roc AUC**	
Logistic regression (LR)	0.913927	0.913573	0.915600	0.687597	0.268056	0.385735	**0.627347**	**0.63159**
Decision tree (DT)	1.000000	0.832773	0.837200	0.202383	0.219907	0.210781	**0.562414**	**0.56514**
Gradient boosting(GBM)	0.903620	0.902160	0.903820	0.852632	0.032774	0.063121	**0.516076**	**0.51807**
Extreme gradient boosting (xgb)	0.901627	0.900760	0.902620	0.920455	0.016387	0.032200	**0.508116**	**0.50933**
Category boost (CatBoost)	0.935487	0.921140	0.923900	0.790306	0.313372	0.448790	**0.652125**	**0.65006**
Multi layers perceptron(MLP)	0.927993	0.892720	0.906380	0.565174	0.229820	0.326765	**0.605211**	**0.61171**
Random forest (RF)	0.985287	0.899267	0.901140	0.500000	0.014971	0.029071	**0.506664**	**0.50006**
Adaboost	0.907300	0.905893	0.907660	0.620206	0.170140	0.267027	**0.579355**	**0.57989**
LighGBM (classifier)	0.916753	0.907507	0.909080	0.861566	0.095691	0.172251	**0.547002**	**0.55095**
LighGBM as regressor								
LighGBM Without threshold (no t)	-	-	-	-	-	-	**0.850000**	**0.85201**
LighGBM Threshold(t): 0.1025	-	-	0.651960	0.203935	0.868096	0.330280	**0.748172**	**0.85201**
LighGBM Threshold(t): 0.1200	-	-	0.712160	0.231242	0.822375	0.360980	**0.761222**	**0.85201**
LighGBM implementation of resampled data								
Under-sampled dataset	-	-	-	-	-	-	**0.863319**	**0.86417**
Over-sampled Dataset	-	-	-	-	-	-	**0.943705**	**0.73039**

TABLE. VI. RESULTS OF LIGHGBM WITH HYPERPARAMETERS ON SANTANDER DATASET

Results of Implementation of ML Algorithms on Santander Dataset (Train=200,000, Test 200,000)								Submission Results
Name	**Training**	**Validation**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC AUC**	
LighGBM (with HP)	-	-	-	-	-	-	**0.887997**	**0.88856**

Here is the set of best values for these parameters.

Space (best): ['bagging_freq': 3, 'bagging_seed': 100000, 'boost_from_average': 'false', 'boosting_type': 'dart', 'class_weight': none, 'colsample_bytree': 0.4, 'learning_rate': 0.45, 'max_bins': 60000, 'max_depth': 1, 'metric': 'auc', 'min_child_samples': 35, 'min_data_in_leaf': 8, 'min_sum_heassian_in_leaf': 4, 'n_estimators': 209.0, 'num_iteration': 280000, 'num_leaves': 45, 'objective': 'binary', 'reg_alpha': 4.4094144078689945, 'reg_lambda': 1.0182413699039161, 'seed': 100000, 'subsample_for_bin': 340000, 'tree_learner': 'serial', 'verbosity': 1].

VII. CONCLUSION AND FINDINGS

Working on the imbalanced dataset and by implementing 9 different basic, advanced and ensemble classification algorithms on the Santander customer transaction prediction dataset provided by the Kaggle, we find out that selecting the metrics is the first and foremost step to know what exactly we want to get from the classifier when working on Imbalanced data. After that, we can select different classifiers, but we find out that LighGBM performed better on this particular dataset. While working with a large dataset, undersampling can perform well, but the AUC will remain less than as compared with the AUC of hyperparameters find from the original dataset; on the other hand, oversampling or, more specifically, smote may lead to overfitting. Furthermore, the performance of a classifier can be increased by finding the hyperparameters. Finding the best parameters randomly and manually is a long, sturdy, and impossible task. Even by using libraries to find the best parameter values is a hectic and time-consuming job. There can be too many combinations of values to try as well as the size of the dataset matters a lot. When finding hyperparameters, undersampling can be considered, but oversampling, in this case, is not recommended due to a lot of time as well as computation requirement. Moreover, the problem was described as binary (classification) but could also be solved by regression as regression results were also accepted.

VIII. FUTURE WORK

In the future, we'll implement LightGBM on some other structured imbalanced datasets to find out the scalability. Furthermore, to find more specific and best hyperparameter values, we'll re-run this code on the latest machine to improve performance.

REFERENCES

[1] F. S. Board, "Artificial intelligence and machine learning in financial services," November, available at: http://www. fsb. org/2017/11/artificialintelligence-and-machine-learning-in-financialservice/(accessed 30th January, 2018), 2017.

[2] D. D. Hawley, J. D. Johnson, and D. Raina, "Artificial neural systems: A new tool for financial decision-making," Financial Analysts Journal, vol. 46, pp. 63-72, 1990.

[3] D. Cheng and P. Cirillo, "A reinforced urn process modeling of recovery rates and recovery times," Journal of Banking & Finance, vol. 96, pp. 1-17, 2018.

[4] K. Nian, T. F. Coleman, and Y. Li, "Learning minimum variance discrete hedging directly from the market," Quantitative Finance, vol. 18, pp. 1115-1128, 2018.

[5] T. Renault, "Intraday online investor sentiment and return patterns in the US stock market," Journal of Banking & Finance, vol. 84, pp. 25-40, 2017.

[6] K. Seeja and M. Zareapoor, "FraudMiner: A novel credit card fraud detection model based on frequent itemset mining," The Scientific World Journal, vol. 2014, 2014.

[7] O. B. Sezer and A. M. Ozbayoglu, "Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach," Applied Soft Computing, vol. 70, pp. 525-538, 2018.

[8] B. Weng, L. Lu, X. Wang, F. M. Megahed, and W. Martinez, "Predicting short-term stock prices using ensemble methods and online data sources," Expert Systems with Applications, vol. 112, pp. 258-273, 2018.

[9] C. X. Ling and C. Li, "Data mining for direct marketing: Problems and solutions," in Kdd, 1998, pp. 73-79.

[10] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques: Elsevier, 2011.

[11] E. I. Altman, G. Marco, and F. Varetto, "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience)," Journal of banking & finance, vol. 18, pp. 505-529, 1994.

[12] F. Varetto, "Genetic algorithms applications in the analysis of insolvency risk," Journal of Banking & Finance, vol. 22, pp. 1421-1439, 1998.

[13] A. N. Kercheval and Y. Zhang, "Modelling high-frequency limit order book dynamics with support vector machines," Quantitative Finance, vol. 15, pp. 1315-1329, 2015.

[14] W. H. Beaver, "Financial ratios as predictors of failure," Journal of accounting research, pp. 71-111, 1966.

[15] P. Ravisankar, V. Ravi, G. R. Rao, and I. Bose, "Detection of financial statement fraud and feature selection using data mining techniques," Decision Support Systems, vol. 50, pp. 491-500, 2011.

[16] J. T. Wells, Occupational fraud and abuse: Obsidian Publishing Company, 1997.

[17] C. Spathis, M. Doumpos, and C. Zopounidis, "Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques," European Accounting Review, vol. 11, pp. 509-535, 2002.

[18] C.-C. Yeh, D.-J. Chi, and M.-F. Hsu, "A hybrid approach of DEA, rough set and support vector machines for business failure prediction," Expert Systems with Applications, vol. 37, pp. 1535-1541, 2010.

[19] J. W. Seifert, "Data mining and the search for security: Challenges for connecting the dots and databases," Government Information Quarterly, vol. 21, pp. 461-480, 2004.

[20] O. Aregbeyen, "The determinants of bank selection choices by customers: recent and extensive evidence from Nigeria," International Journal of Business and Social Science, vol. 2, pp. 276-288, 2011.

[21] K. O. Siddiqi, "Interrelations between service quality attributes, customer satisfaction and customer loyalty in the retail banking sector in Bangladesh," International Journal of Business and Management, vol. 6, p. 12, 2011.

[22] X. Hu, "A data mining approach for retailing bank customer attrition analysis," Applied Intelligence, vol. 22, pp. 47-60, 2005.

[23] K. Coussement and D. Van den Poel, "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques," Expert systems with applications, vol. 34, pp. 313-327, 2008.

[24] R. A. Soeini and K. V. Rodpysh, "Evaluations of Data Mining Methods in Order to Provide the Optimum Method for Customer Churn Prediction: Case Study Insurance Industry," in International Conference on Information and Computer Applications, 2012, pp. 290-297.

[25] T. O. Goonetilleke and H. Caldera, "Mining life insurance data for customer attrition analysis," Journal of Industrial and Intelligent Information, vol. 1, 2013.

[26] B. He, Y. Shi, Q. Wan, and X. Zhao, "Prediction of customer attrition of commercial banks based on SVM model," Procedia Computer Science, vol. 31, pp. 423-430, 2014.

[27] J.-H. Ahn, S.-P. Han, and Y.-S. Lee, "Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry," Telecommunications policy, vol. 30, pp. 552-568, 2006.

[28] A. Sharma, D. Panigrahi, and P. Kumar, "A neural network based approach for predicting customer churn in cellular network services," arXiv preprint arXiv:1309.3945, 2013.

[29] E. W. Ngai, L. Xiu, and D. C. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," Expert systems with applications, vol. 36, pp. 2592-2602, 2009.

[30] P. S. Raju, D. V. R. Bai, and G. K. Chaitanya, "Data mining: techniques for enhancing customer relationship management in banking and retail industries," International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, pp. 2650-2657, 2014.

[31] H. Zhou, H.-f. Chai, and M.-l. Qiu, "Fraud detection within bankcard enrollment on mobile device based payment using machine learning," Frontiers of Information Technology & Electronic Engineering, vol. 19, pp. 1537-1545, 2018.

[32] K. Fu, D. Cheng, Y. Tu, and L. Zhang, "Credit card fraud detection using convolutional neural networks," in International Conference on Neural Information Processing, 2016, pp. 483-490.

[33] S. Chen, "Detection of fraudulent financial statements using the hybrid data mining approach," SpringerPlus, vol. 5, p. 89, 2016.

[34] A. K. Pujari, Data mining techniques: Universities press, 2001.

[35] J. R. Quinlan, "Simplifying decision trees," International journal of man-machine studies, vol. 27, pp. 221-234, 1987.

[36] S. Moro, R. Laureano, and P. Cortez, "Using data mining for bank direct marketing: An application of the crisp-dm methodology," in Proceedings of European Simulation and Modelling Conference-ESM'2011, 2011, pp. 117-121.

[37] E. Turban, R. Sharda, and D. Delen, "Decision Support and Business Intelligence Systems (required)," Google Scholar, 2010.

[38] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann, 2016.

[39] K. Chitra and B. Subashini, "Data mining techniques and its applications in banking sector," International Journal of Emerging Technology and Advanced Engineering, vol. 3, pp. 219-226, 2013.

[40] S. Ghosh, A. Hazra, B. Choudhury, P. Biswas, and A. Nag, "A Comparative Study to the Bank Market Prediction," in International Conference on Machine Learning and Data Mining in Pattern Recognition, 2018, pp. 259-268.

[41] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," European Journal of Operational Research, vol. 247, pp. 124-136, 2015.

[42] B. Wang, Y. Kong, Y. Zhang, D. Liu, and L. Ning, "Integration of Unsupervised and Supervised Machine Learning Algorithms for Credit Risk Assessment," Expert Systems with Applications, 2019.

[43] P. Cerchiello, P. Giudici, and G. Nicola, "Twitter data models for bank risk contagion," Neurocomputing, vol. 264, pp. 50-56, 2017.

[44] M. Islam and M. Habib, "A data mining approach to predict prospective business sectors for lending in retail banking using decision tree," arXiv preprint arXiv:1504.02018, 2015.

[45] Scikit-learn, "Scikit-learn," 2019.

[46] Kaggle, "Santander Customer Transaction Prediction - Can you identify who will make a transaction?," 2019.

[47] H. Khosravi and E. Kabir, "Introducing a very large dataset of handwritten Farsi digits and a study on their varieties," Pattern recognition letters, vol. 28, pp. 1133-1141, 2007.

[48] C. Chatfield, "Exploratory data analysis," European journal of operational research, vol. 23, pp. 5-13, 1986.

[49] KNIME, "Dimensionality Reduction Techniques," 05/12/2015 2015.

[50] S. Raschka, "Linear Discriminant Analysis," Aug 3, 2014 2014.

[51] A. M. Martínez and A. C. Kak, "Pca versus lda," IEEE transactions on pattern analysis and machine intelligence, vol. 23, pp. 228-233, 2001.

[52] Medium, "Dimensionality Reduction(PCA and LDA)," 10 March, 2019 2019.

[53] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," Progress in Artificial Intelligence, vol. 5, pp. 221-232, 2016.

[54] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in Data mining and knowledge discovery handbook, ed: Springer, 2009, pp. 875-886.

[55] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," Journal of artificial intelligence research, vol. 61, pp. 863-905, 2018.