A Study into the performance of various Machine Learning Algorithms applied to three Finance related datasets

Karan veer Singh
Msc in Data Analytics
National College Of Ireland
Dublin, Ireland
x20146248

***Abstract*—Data mining is a process to convert raw data into usable information. Machine learning is the study of algorithms that develop themselves automatically through the use of data. The main goal of data mining and machine learning is to extract knowledge from the integration of information received from data. In this paper, 3 large data set subjected to the financial and banking industry are used and with the help of data mining and machine learning algorithms and following CRISP-DM methodology research question related to that data set is answered. Also, different machine learning methods are used to make predictions and evaluation methods to compare between the methods. The first data set is related to Credit card default, Naive Bayes and Random forest classification techniques are used with recall values of 90 percent and 38 percent respectively, factors like repayment status of credit, total bill of credit, and balance of customers are the most important factors influencing prediction of default. The second data set is credit card fraud, for this dataset SVM performs better than Logistic Regression and SMOTE gives better performance in comparison to undersampling and oversampling. For third data set bank customer churn XGBoost classifier in comparison to k-nearest neighbors (KNN) algorithm performs better with an accuracy of 86.20 percent and 80.82 percent respectively. For XGBoost classifier ROC curve is better then k-nearest neighbors (KNN).**

***Index Terms*—Credit Card Default, Credit Card fraud, Credit Card churn,CRISP-DM, Supervised, Random forest,SVM,XGBoost classifier,k-nearest neighbors (KNN),Logistic Regression,Naive Bayes,SMOTE.**

I. Introduction

Many companies have started to systematically capture massive amounts of historical data documenting their processes, goods, and consumers over the last decade. Simultaneously, scientists and engineers across a wide range of disciplines have been capturing increasingly complex experimental data sets. The field of data mining seeks to answer the question of how to best use historical data to uncover general patterns and enhance decision-making [1]. Even though machine learning algorithms are central to the data mining process, it is worth noting that the procedure also includes few other crucial steps like data maintaining, cleaning, formatting, visualization, statistical analysis, and summarization of the human expertise to construct inputs to learning algorithm and evaluate the result.

Thousands of years have passed since retailers used credit to assist their customers in financing transactions. Companies built on the concept of a promissory note in the mid 19th century by including a tangible object which can be used to quickly distinguish their customer accounts. Some fall into the category of coins or jewels displaying the trader's name and emblem, as well as the buyer's account number. Credit cards initially functioned similarly to earlier medals, stamps, and labels [2]. Traders would simply mark the card, which is familiar to anybody who recalls how common credit card transactions were in the 1990s. By the early 1990s, though, several cards had a magnetic stripe on the back that might be read by advanced electronic equipment that was cutting-edge at the time. By the end of the 1990s era of EMV smart chips came into the picture, this technology was commonly used throughout Europe for many years. Wireless payment systems are being increasingly incorporated into devices, phones, and several other portable platforms. Finally, many experts envision the time when biometric authentication would enable customers to charge transactions with their fingerprints or retinal scans, rather than having to carry around a card or other device containing their account details.

One of the early applications of old machine learning was in credit risk modeling, which aims to predict default risk using financial data. Identifying fraudulent transactions using conventional traditional processes is time-consuming and costly, so the introduction of machine learning has made manual methods obsolete. Financial fraud is a rising threat in the finance sector, business organizations, and the state, with far-reaching implications. Fraud is described as criminal fraud to obtain monetary benefit. The data mining methodology is one of the most well-known tools for detecting credit fraud. Credit card fraud prevention is the method of classifying fraudulent activity into two categories: authentic (legit) and fraudulent transactions.

- In this paper for Credit Card Default data set [3] after prepossessing like feature engineering and feature selection and analyzing the data-set, I will try to find out

which are the strongest predictors of default payment? and compare different machine learning techniques based on their performance. Naive Bayes and Random forest classification techniques are used in this data-set as machine learning techniques to answer the research question. For evaluation confusion matrix and also cross-validation is used. Pycaret automated machine learning library is used to compare different machine learning techniques on the basis of their performance.

- For Credit card fraud data set [4] it is undergone principal component analysis transformation and data is heavily skewed, with the fraud class accounting for just 0.172 percent of all transactions. In this paper, I try to find out which sampling method should be used in machine learning algorithms to get better performance.SVM and logistic regression machine learning algorithms are also used. To check performance after applying different sampling methods like under-sampling, over-sampling, and SMOTE (Synthetic Minority Oversampling Technique) Random forest algorithm is used.
- In customer bank churn data set [5] after analyzing the data and removing the outliers using an interquartile range for age column, feature engineering is performed and data type of few columns are corrected. For this dataset research question is to find out which machine learning algorithm among eXtreme Gradient Boosting (XGBoost) and k-nearest neighbors (KNN) algorithm perform better in terms of accuracy to predict customer churn.

II. Related Work

In the research paper [6], imbalanced data set is used to produce an interpretable model for risk prediction of the credit card. To minimize the effect of varying values min Max normalization is used and one-way analysis of variance is used to check the significance of the result of the model. 66.9% accuracy on imbalanced Taiwan data set is increased to 89% by the proposed method in research paper [6]. For accuracy for South German client data set is improved from 70.7% to 87.1% and Belgium client from 65% to 87% respectively. In this paper, the authors observed oversampling techniques SMOTE performs better than under-sampling techniques and gradient boosting decision tree (GBDT) model method gives the best results in comparison to other methods.

In another study, [7] by Riddhi, Gourab, and Rahul objective of this study is to identify the strong features which are important for prediction, to implement different techniques for class imbalanced data set, and to predicted default. weight of evidence (WOE) and information value (IV) is useful feature selection. This Decision tree, random forest, and Naïve Bayes classifier are used as modeling techniques. different techniques such as random undersampling, random oversampling and hybrid sampling, etc. are used to overcome the imbalanced data.

Comparison of different machine learning approaches to find cost-effective methods for credit card default prediction. Shantanu and Benjaman in research paper [8] discussed 7 machine learning methods out of which only random forest and artificial neural network perform better. Among artificial neural networks and random forest, the model has the most cost-effective prediction factors having discriminate variables like marital status, gender, education, etc. are considered important in the study other fields like limit, billing, and payments are observed to be more important. Matthew's correlation coefficient is used to evaluate model performance.

Another study [9] to solve the problem to predict credit card default using real-world data is done by Huei-Wen Teng and Michael Lee. They reviewed a total of 5 machine learning techniques namely k-nearest neighbors, decision trees, boosting, support vector machine, and neural networks to solve problems to predicate credit card default. According to the study decision trees perform best in terms of the learning curve in comparison to other machine learning algorithms used on data-set having 29,999 instances with 23 features. Support vector machine performs the worst, not only it yields the lowest accuracy but also runs slow. Some limitations in the study are only one data set is used, multiple datasets can be used to check the robustness of the model. This paper only uses accuracy as a major for the evaluation of the machine learning algorithms, other methods like precision, recall, F1 score, and agency can also be used in addition to standard measures.

Fraud is spreading with the development of IT and communication channel. Even though prevention mechanisms which are using an industry like CHIP and PIN are not able to prevent the fraudulent use of credit card because among the most common form of fraud uses virtual POS and Mail orders. In this study, [10] by Y Sahin and F Duman they use classification model support vector machine(SVM) and Decision tree to detect fraud and compare the performance of both using data taken from banks data-warehouse with relevant permissions. Credit card fraud can be of many types it can be online fraud application fraud or simple theft. A credit card can be used at millions of places it can be online places such as an E-Commerce website, shopping site, or offline place like supermarkets due to this vast usage of credit cards it is difficult to track the pattern for the fraudulent transaction. Many techniques can be used to detect fraud broadly divided into supervised and unsupervised learning. They have used overall seven alternative models based on SVM and decision tree model and compare the result among them. According to the study, [10] decision tree outperforms SVM based model. Accuracy rates are used to compare these models.

Many papers related to credit card fraud are available online. Another study [11] by John O. Awoyemi , Adebayo O. Adetunmbi, and Samuel A. Oluwadare, used k-nearest neighbors, naive Bayes, and logistic regression to predict credit card fraud. The work is implemented in Python language. Accuracy of 54.86% for logistic regression, 97.92% for k-nearest neighbors, and 97.69% for naive Bayes achieved. According to the research, k-nearest neighbors performed better than naive Bayes and logistic regression method. A combination of differ-

ent techniques is used to overcome the problem of imbalanced data set like undersampling and oversampling. That they die subjected to three techniques. Matthews correlation coefficient is also used to compare the results. The study can also include other sampling methods to improve the performance of the model. Also, meta-classifiers and learning approaches for dealing with highly imbalanced data set can be used.There are three main contributions in the article [12] by Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, and Gianluca Bontempi.. First, they suggest a formalization of the fraud-detection issue with the aid of research assistance that accurately defines the operational conditions of FDSs (fraud-detection system) that evaluate large flows of credit card transactions daily. They also show how to use the most effective performance metrics for fraud detection. Second, they implement and test a novel learning technique for addressing class imbalance and confirmation lag. Thirdly, they describe the effects of class imbalance and idea drift in an actual dataset comprising more than 75 million payments approved over three years in their studies.

In this study [13], they use 12 classification techniques in a real-world credit card holders' scenario. A predictive churn model was built using data for the commercial bank which are in majority in China. In addition, the prediction accuracy of classification algorithms based on Multi-Criteria Decision techniques like PROMETHEE II and TOPSIS is compared. They suggested additional research may be conducted in a variety of directions such as several other MCDM (Multi-Criteria Decision-Making Methods) methods that can also be used. According to their findings, the most effective methods for credit card churn study are logistic regression and the J48 algorithm. Another study [14] by T.Vafeiadis, K.I.Diamantarasb, G.Sarigiannidisa, and K.Ch.Chatzisavvas show a comparative analysis of the most common machine learning models applied to the complex issue of predicting customer churn in the telecom sector. In the first part of the experiment, they applied all Machine learning models and evaluated them using cross-validation. For the second phase, they use boosting to improve the performance of the model. According to the study they got nearly 97 percent and an F-measure of over 84 percent with the SVM model using AdaBoost.This research clarified the efficiency and the benefit of using boosting techniques. In this [15] study they used CRISP methodology for predictive modeling of customer churn in the banking sector domain and to get a better understanding of churner characteristics. They provide the characteristics of features that are related to customer churn. Limitation to study is the dataset as it is related to bank's database associated privacy issues arise and data collection.

This paper [16] provides a structured overview and study of different machine learning and data mining algorithms and their application in process related to the credit card which is mainly divided into 4 categories, first is customer relationship management, second is repayment prediction, third is new customer selection and last is fraud detection. This study can help credit card issuers to have the solution to their problems related to churn, fraud, and default. It is suggested in the paper for credit card issuers to have a systematic framework for credit card process as to gain trust they must have best fraud detection method, for this cause data mining and machine learning algorithm used has to be best. They suggest the model data selection should be appropriate and the time period of the dataset should be suitable. They did not discuss the limitation of machine learning algorithms used in the study it will help to understand models better.

Author's in this paper [17] used,k-nearest neighbors algorithm and outlier detection methods to get the best solution to credit card fraud detetion problem.These methods have been shown to reduce false alarm rates and improve fraud detection levels. According to the study to identify and deter fraudulent transactions, any of these approaches can be used.The outlier identification system aids in credit card fraud detection while demanding less memory and calculation also outlier detection operates fast and well on large datasets.According to study as compared to other organized anomaly detection methods and power methods, empirical findings determine that the k-nearest neighbors algorithm is superior.The method is accurate and adequate.

The paper [18] examines the use of both manual and automated classification methods of data mining, provides an overview of the entire development process, and compares various machine learning approaches. Logistic regression, Logistic regression, random forests, and support vector machines, support vector machines, and random forests. Cross-validation on the training set was used to select the models and their best-fit parameters. The highest output was obtained by random forests, according to their findings.SVM and logistic regression work similarly based on performance having AUC slightly lower than random forest. The outcomes revealed a very high number of valid orders allowed and a false alarm percentage of just about 2 percentage, based on the assumption that 80 percent of the orders in the testing collection with the lowest scores will be automatically approved and the remainder tested manually. The likelihood of refusing a fraudulent order, on the other hand, was only about 59 percent. Nonetheless, the chargeback rate will be 0.88 percent, well below the industry average of 1.6 percent for foreign orders in North America.

In this paper [19], a behavior-based classification system is used based on Support Vector Machines for credit card fraud detection. According to this study, SVM delivers a distinctive solution. By practicing the kernel, they gain SVM flexibility in the form of a threshold for separating the data. Such qualities make the SVM bring out the classification problem in this complex domain and also it yields a good result. The proposed method in this paper gives higher accuracy of detection and is also scalable for handling large volumes of transactions.

This survey [20] is divided into two sections. In the first section, they look at studies that use conventional machine learning models to predict fraud, which often uses traditional transnational features. In the second part, they look into more advanced techniques for user authentication such as psychological and behavioral to identify the individual based on his

or her unique behavior. The key finding is supervised learning strategies were used more commonly than unsupervised and the most widely used fraud detection machine learning technique is naive Bayes and SVM. Also as per the survey most recent and common user authentication are password, biometric, and combination of both.

III. Data Mining Methodology and Evaluation

Data mining is a process of mining or discovering knowledge from data. The aim of using data mining methodology is to come up with an informative conclusion and perform the task needed in the process. CRISP-DM (Cross-industry process for data mining) is followed in this report to get an insightful conclusion. Various steps include in the cross-industry process for data mining are understanding of business which involves determining business objectives, determine data mining goals, and producing a project plan. The second step in the methodology is data understanding which includes processes like describing data, explore data, verify details, and quality of data. Data preparation is stage three in this process which involves data cleaning, construction, and integration of required data. Stage four modeling is the most important step of CRISP-DM (Cross-industry process for data mining) methodology it includes modeling techniques, generates test design-build, and assesses models. The next step Evaluation of the model, process review, and last is deployment.

This section of the report is divided into 3 sections dedicated to three selected data set and involve all the step applied from CRISP-DM to answer research questions and evolution of models applied to each dataset.

Fig. 1. CRISP-DM Methodology

A. Credit Card Default Data-set:

There can be many reasons customers default credit card payments and to identify when a user is going to default, analysis has been conducted on data set containing information of default payment, history of payment demographic factors like gender, education, marital status age of credit card clients. There is a total of 25 columns and 30,000 rows with the target or independent variable "default.payment.next.month" signifying one as default and 0 is not default. The below table is created to get a detailed description of attributes.

Column	Description
ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars
SEX	Gender (1=male, 2=female)
EDUCATION	(1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY_0	Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
PAY_2	Repayment status in August, 2005 (scale same as above)
PAY_3	Repayment status in July, 2005 (scale same as above)
PAY_4	Repayment status in June, 2005 (scale same as above)
PAY_5	Repayment status in May, 2005 (scale same as above)
PAY_6	Repayment status in April, 2005 (scale same as above)
BILL_AMT1	Amount of bill statement in September, 2005 (NT dollar)
BILL_AMT2	Amount of bill statement in August, 2005 (NT dollar)
BILL_AMT3	Amount of bill statement in July, 2005 (NT dollar)
BILL_AMT4	Amount of bill statement in June, 2005 (NT dollar)
BILL_AMT5	Amount of bill statement in May, 2005 (NT dollar)
BILL_AMT6	Amount of bill statement in April, 2005 (NT dollar)
PAY_AMT1	Amount of previous payment in September, 2005 (NT dollar)
PAY_AMT2	Amount of previous payment in August, 2005 (NT dollar)
PAY_AMT3	Amount of previous payment in July, 2005 (NT dollar)
PAY_AMT4	Amount of previous payment in June, 2005 (NT dollar)
PAY_AMT5	Amount of previous payment in May, 2005 (NT dollar)
PAY_AMT6	Amount of previous payment in April, 2005 (NT dollar)
default.payment.next.month	Default payment (1=yes, 0=no)

Fig. 2. Data description table for Credit Card Default data-set

All the data is fetched in CSV format and stored to variable df in Jupiter notebook using the read.csv() function. Missing values are checked using the isnull() function, no missing value is found in the data set. Few columns name is changing to more understandable and user-friendly terms like Sex is changed to 'Gender', 'Pay_0' column name is changed to 'Pay_1' to put it in the sequence of other columns. Target column 'default.payment.next.month' is changed to more readable name Default. The new column 'Repayment_sum' having some of all the payments repayment status from April to September and 'Bill_sum' which has the amount of bill statement from April to September are added to the df. More feature engineering is applied to columns 'marriage', 'age', and 'education' attributes. For education, it has values 1 for Graduate School, 2 for University, 3 for high school, 4 others, and 5 for others. Few values like 0 and 6 are also present having unknown value so they are grouped into category 5 itself using apply() function. Similarly, for the 'marriage' column it has a value of 3 for others 1 for married 2 for single. For value having 0, it is grouped with 3 others and 'Age' is

converted into five categories one for customer age between 20 and 32 fourth age between 30 to 43 for age between 50 to 60 on five for others. Further binary values are generated for education, age, and marriage attribute using get_dummies() function, and dummy values are added to df, and old values are dropped. Feature engineering is complete and now data is analyzed using describe function. The average value of the credit card limit is 167, 484. Education level is mostly graduate school that is one and University that is two more. Most of their clients are married or single. Averages of customers 35. To check the distribution of the default data plot.bar() function is used. Correlation with the target column is shared with the cor function. Pay_1, Pay_2, and replacement_sum are highly correlated with defaulters.

Since some variables like limit_bal, total_bill, and total_sum have a wide range which may affect the variable, In our model we use the minmaxscalar() function to reduce values in range 0 - 1 and avoid biases due to a large number of variables.

Describe function is used to get the insight of data distribution.

- The average value for the amount of credit card limit is 167,484.
- Most of the clients are either married or single.
- Average age of clients is 35.5 years, with a standard deviation of 9.2.

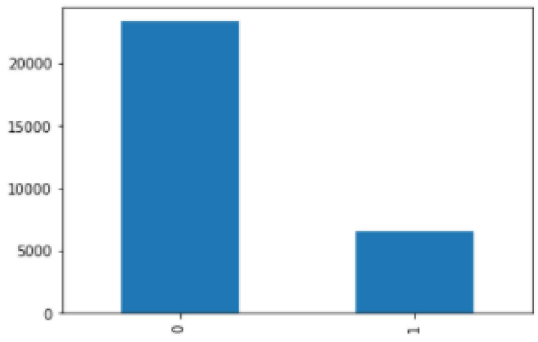

Fig. 3. Default distribution

- Since a value of 0 indicates that a payment is not defaulted and a value of 1 indicates that a payment is defaulted, the mean of 0.221 indicates that 22.1 percent of credit card contracts will default next month as shown in Fig 3.

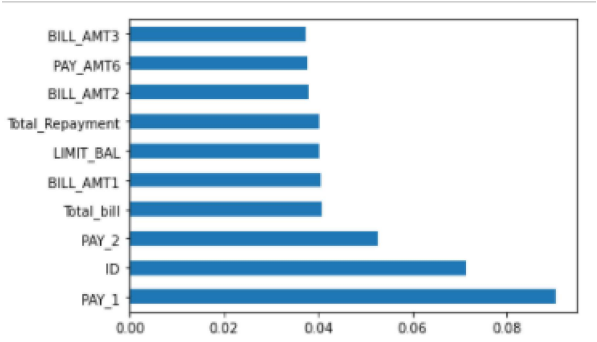

Fig. 4. Feature importance plot

- Factors like repayment status of credit, totalbill of credit, and balance of customers are the most important factors influencing prediction of defaults shown in fig 4. Highly correlated column are dropped so that it will not impact model performance.

```
[[6555  433]
 [1246  766]]
              precision    recall  f1-score   support

           0       0.84      0.94      0.89      6988
           1       0.64      0.38      0.48      2012

    accuracy                           0.81      9000
   macro avg       0.74      0.66      0.68      9000
weighted avg       0.80      0.81      0.79      9000

Accuracy 81.34 %
```

Fig. 5. Evaluation for Randomforest Model

Confusion matrix is plot with classification_report library in sklearnmetrics for evaluation of model's Naive Bayes and Random forest classification techniques as shown in Fig 5 and Fig 6 it is showing accuracy of 81.34 percent for random forest and 37.38 percent for Naive Bayes but recall value of Naive Bayes is better then random forest i.e 89 percent.Naive Bayes performs better as focus is on recall i.e true positive rate, which is more sensitive to the true positives than accuracy.cross_val_score library(Cross Validation) from sklearn.model_selection is also used to evaluate machine learning.

```
[[1592 5425]
 [ 220 1763]]
              precision    recall  f1-score   support

           0       0.88      0.23      0.36      7017
           1       0.25      0.89      0.38      1983

    accuracy                           0.37      9000
   macro avg       0.56      0.56      0.37      9000
weighted avg       0.74      0.37      0.37      9000

Accuracy 37.28 %
```

Fig. 6. Evaluation for NB Model

For evaluating the model, we focused on recall and F1 score. More focus on these matrices because it is more important to us, as banks forecast the likelihood of default, they need to be extremely accurate. This means there are consumer defaults, the model should be able to correctly predict the result. We need a large number of true positives (actual '1', predicted '1') and a small number of false negatives (actual '1', predicted '0'). As recall is a ratio of true positive and the sum of true positive and false negative it should be large as possible. The precision of the model is not our priority as if the model predicts '0' in place of '1' it won't cost much to do that.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.8220	0.7727	0.3666	0.6759	0.4751	0.3791	0.4050	0.2010
ridge	Ridge Classifier	0.8214	0.0000	0.3566	0.6798	0.4676	0.3726	0.4007	0.0530
gbc	Gradient Boosting Classifier	0.8197	0.7741	0.3538	0.6714	0.4630	0.3670	0.3944	1.4560
lda	Linear Discriminant Analysis	0.8193	0.7566	0.3748	0.6565	0.4769	0.3774	0.3992	0.1610
ada	Ada Boost Classifier	0.8181	0.7717	0.3298	0.6779	0.4435	0.3499	0.3826	0.3840
rf	Random Forest Classifier	0.8147	0.7589	0.3590	0.6416	0.4601	0.3588	0.3811	0.7410
et	Extra Trees Classifier	0.8052	0.7314	0.3647	0.5932	0.4514	0.3412	0.3562	0.6840
lr	Logistic Regression	0.7830	0.6691	0.0212	0.7226	0.0412	0.0292	0.0974	0.6560
knn	K Neighbors Classifier	0.7603	0.5988	0.1644	0.3926	0.2315	0.1170	0.1327	1.4120
dt	Decision Tree Classifier	0.7234	0.6115	0.3857	0.3746	0.3800	0.2020	0.2021	0.1390
svm	SVM - Linear Kernel	0.6718	0.0000	0.3009	0.2818	0.2279	0.0607	0.0751	0.2070
qda	Quadratic Discriminant Analysis	0.4902	0.5229	0.5812	0.2381	0.3324	0.0335	0.0389	0.2500
nb	Naive Bayes	0.3399	0.6865	0.9171	0.2391	0.3793	0.0467	0.1078	0.0450

Fig. 7. Pycaret Library output for model comparison

Pycaret is automated machine learning library help us to do a lot of automated features compared to all the life cycle of a project like a feature engineering, features selection, handling outlier and more. It is an open-source library. We used it to compare different classification model performance and get a holistic idea of all classification models. This automated library can be used to know the best model according to the need of research as shown in Fig 7. Naive Bayes is the best model to get a higher recall value and hyper tuning of the model is done to get better performance.

B. Credit Card Fault Data set:

This data set contains transactions made by European credit cardholders. We have 492 fraud transactions out of 284,807 transactions that occurred in the last two days. This data set is heavily skewed, with the fraud class accounting for just 0.172 percent of all transactions. Only numerical input variables are present in this data set that have undergone principal component analysis transformation. due to confidentiality concerns, the data set does not have any further background information. time and amount are not transformed through principal component analysis transformation. The time contains seconds laps between transactions. The amount is transaction size. Overall, 31 columns are present, feature class take values '1' and '0', where '1' represents fraud and '0'normal transactions.

Fig. 8. Table for data description

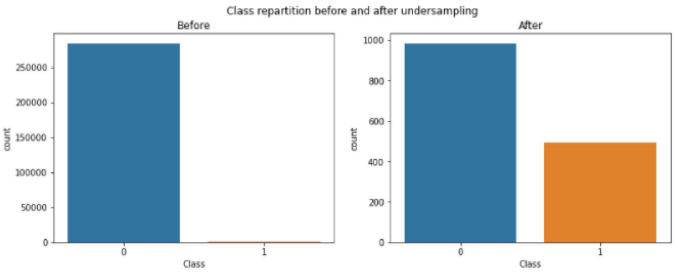

Fig. 9. Table for data description

Before applying data mining and machine learning algorithms, necessary preprocessing is performed on the dataset. Although data is undergone PCA(Principle component analysis) but time and amount columns are not. Time is dropped as not required for this study and from amount data new variable std_Amount is created using the standardized value of amount with the help of StandardScaler function from sklearn.To overcome the problem of imbalance dataset random undersampling is applied before modeling.

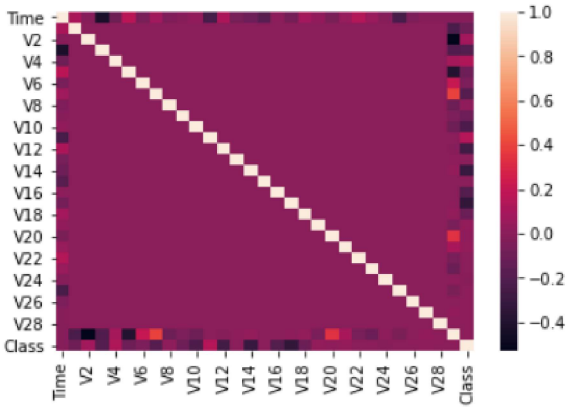

Fig. 10. Table for data description

Column	Description
Row Number	corresponds to the record (row) number and has no effect on the output.
CustomerId	contains random values and has no effect on customer leaving the bank.
Surname	the surname of a custome
CreditScore	Credit score of customer
Geography	a customer's location
Gender	Customer gender
Age	Ager of customer
Tenure	refers to the number of years that the customer has been a client of the bank.
Balance	Balance of customer
NumOfProducts	refers to the number of products that a customer has purchased through the bank.
HasCrCard	denotes whether or not a customer has a credit card. This column is also relevant, since people with a credit card are less likely to leave the bank.
IsActiveMember	active customers are less likely to leave the bank
EstimatedSalary	as with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.
Exited	whether or not the customer left the bank. (0=No,1=Yes)

Fig. 11. Table for data description

The dataset is split into train and test samples in which the test set comprises 30 percent of total data. Train data set is used to train the model and evaluation is done on the test dataset. Logistic Regression is used to predict independent variables with an accuracy of 99.91 percent but the recall rate is low 56 percent. For SVM with a recall of 65 percentage and accuracy of 99.94 percentage. Further other sampling techniques are used like oversampling and SMOTE (Synthetic Minority Oversampling Technique) and modeling is performed for the random forest to compare different sampling techniques. As observed SMOTE performed best of all sampling techniques.Only disadvantage to using Synthetic Minority Oversampling Technique is it took too much time to run.

C. Churn for Bank Customers Data-set:

Churn control is such an important part of keeping valuable customers, the ability to predict customer churn is essential. This dataset consists of information regarding bank customers such as customer id, surname,creditscore , geography, gender, age, tenure, balance, numofproduct, hascrcard, isactive member, estimatedsalary, and exited. Out of these 13 variables, 3 are string variable surname, geography, and gender, others are numerical. All variables have 10000 rows. There is no missing value in the dataset. Columns that are no impacting the research are dropped from data frame. After visualization of the data set, it is observed that customers are mostly from France, but there are also a significant number of customers from Germany and Spain.No of customers from Germany and France are almost the same. The gender of customers is overwhelmingly male, the usual tenure of the customer is 1-9 years. Most of the customer has a credit card and count of customer exited are very less.

Outliers are detected for the columns using the Interquartile concept, Interquartile range (IQR) is measured for each column by checking the difference between the Q3 and Q1 values of the particular column. Rows having values out of the Interquartile range (IQR) are dropped from the data frame. It is observed columns like Credit Score, Age, Num Products are having maximum outliers. Overall 432 rows having outliers are removed from the data frame.

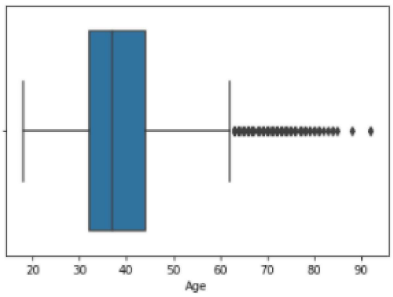

Fig. 12. Table for data description

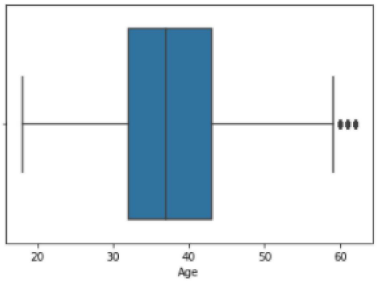

Fig. 13. Table for data description

Feature engineering is performed by implementing the dummies process to convert the value of variables based on the

values they receive. To check which model performers best among eXtreme Gradient Boosting (XGBoost) and k-nearest neighbors (KNN) algorithm. Firstly data frame is divided into two one with dependent variables and another with the independent variable. For this dataset independent variable is Exited having a value of 0 for the customer not exited and 1 for customer exited. From sklearn.model_selection train_test_split library is used for Train-Test Separation and data is divided into X_train and y_train as training data on the test size of 80 percent to X_test and y_test with data size 20 percent.

```
----------------------------
KNeighborsClassifier:
Accuracy: 80.8255%
----------------------------
GradientBoostingClassifier:
Accuracy: 86.2069%
```

Fig. 14. Table for data description

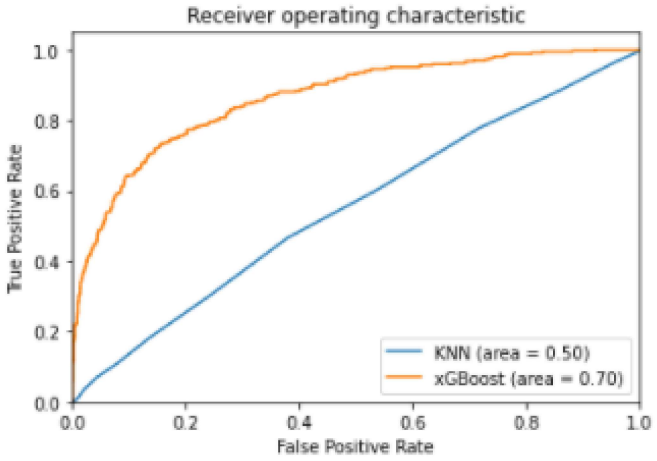

Fig. 15. Table for data description

GradientBoostingClassifier() with default parameters are used as a model and trained with X_train and y_train values of the data frame and after prediction confusion matrix is polt for evaluation getting True Positive: 174, True Negative: 1487, False Positive: 67 and False Negative: 186. For evaluating the efficacy of the model cross-validation is used for accuracy having 10 iterations and getting 0.8608 mean accuracy for gradient boosting model. Similarly, the same values of train and test split are used to train the k-nearest neighbors (KNN) algorithm to compare both models. further GridSearchCV is used to select the best hyperparameter for the k-nearest neighbors (KNN) algorithm to improve model performance. After it, a comparison between the model is performed by plotting ROC (Receiver operating characteristic) and got better curve for gradient boosting model as shown in Fig 15.

IV. Conclusions and Future Work

The examination of Six different algorithms on three different datasets is performed. The factors that play a significant role in prediction are identified and it helped to answer the research questions after analysis. It is observed that the principle of operating of each algorithm differs from the others. The same independent variables cannot be used in each model considering every algorithm works differently.

For credit card default dataset Naive Bayes and Random Forest classification techniques is used with recall values of 90% and 38%, respectively. The most important variables affecting estimation of default are repayment status of credit, total bill of credit, and consumer balance.The second data set is credit card fraud, for this dataset SVM performs better than Logistic Regression and SMOTE gives better performance in comparison to undersampling and oversampling. In contrast to k-nearest neighbours (KNN), the XGBoost classifier performs better with an accuracy of 86.20 percent and 80.82 percent, respectively, for this data collection of bank customer churn. The ROC curve of the XGBoost classifier is superior to that of the k-nearest neighbours classifier (KNN).Bank customer churn prediction model performance is better for eXtreme Gradient Boosting (XGBoost classifier) in comparison to k-nearest neighbors (KNN) algorithm with an accuracy of 86.20 percent and 80.82 percent respectively as shown in figure 14.

The scope of this project is limited to default parameters of data mining and machine learning approaches used in this paper, different parameterization techniques can be used to improve model performance. Others aspects of the finance and banking domain for security such as biometric and customer behaviors can be included to get better results. Different other machine learning algorithms like neural networks and deep learning etc., fine-tuning and other sampling methods can be studied in the future.

References

[1] T. M. Mitchell, "Machine learning and data mining," *Communications of the ACM*, vol. 42, no. 11, pp. 30–36, 1999.

[2] T. Wolters, "" carry your credit in your pocket": The early history of the credit card at bank of america and chase manhattan," *Enterprise & Society*, pp. 315–354, 2000.

[3] https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset.

[4] https://www.kaggle.com/mlg-ulb/creditcardfraud.

[5] https://www.kaggle.com/mathchi/churn-for-bank-customers.

[6] T. M. Alam, K. Shaukat, I. A. Hameed, S. Luo, M. U. Sarwar, S. Shabbir, J. Li, and M. Khushi, "An investigation of credit card default prediction in the imbalanced datasets," *IEEE Access*, vol. 8, pp. 201 173–201 198, 2020.

[7] R. Luthra, G. Nath, and R. Chellani, "A review on class imbalanced correction techniques: A case of credit card default prediction on a highly imbalanced dataset."

[8] S. Neema and B. Soibam, "The comparison of machine learning methods to achieve most cost-effective prediction for credit card default," *Journal of Management Science and Business Intelligence*, vol. 2, no. 2, pp. 36–41, 2017.

[9] H.-W. Teng and M. Lee, "Estimation procedures of using five alternative machine learning methods for predicting credit card default," *Review of Pacific Basin Financial Markets and Policies*, vol. 22, no. 03, p. 1950021, 2019.

[10] I.-C. Yeh and C.-h. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2473–2480, 2009.

[11] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," in *2017 International Conference on Computing Networking and Informatics (ICCNI)*. IEEE, 2017, pp. 1–9.

[12] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection: a realistic modeling and a novel learning strategy," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 8, pp. 3784–3797, 2017.

[13] G. Wang, L. Liu, Y. Peng, G. Nie, G. Kou, and Y. Shi, "Predicting credit card holder churn in banks of china using data mining and mcdm," in *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 3. IEEE, 2010, pp. 215–218.

[14] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9, 2015.

[15] A. Keramati, H. Ghaneei, and S. M. Mirmohammadi, "Developing a prediction model for customer churn from electronic banking services using data mining," *Financial Innovation*, vol. 2, no. 1, pp. 1–13, 2016.

[16] P. Wongchinsri and W. Kuratach, "A survey-data mining frameworks in credit card processing," in *2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. IEEE, 2016, pp. 1–6.

[17] N. Malini and M. Pushpa, "Analysis on credit card fraud identification techniques based on knn and outlier detection," in *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*. IEEE, 2017, pp. 255–258.

[18] N. Carneiro, G. Figueira, and M. Costa, "A data mining based system for credit-card fraud detection in e-tail," *Decision Support Systems*, vol. 95, pp. 91–101, 2017.

[19] V. Dheepa and R. Dhanapal, "Behavior based credit card fraud detection using support vector machines," *ICTACT Journal on Soft computing*, vol. 2, no. 07, p. 2012, 2012.

[20] N. Yousefi, M. Alaghband, and I. Garibay, "A comprehensive survey on machine learning techniques and user authentication approaches for credit card fraud detection," *arXiv preprint arXiv:1912.02629*, 2019.