Predicting Credit Card Fraud, Employee Attrition, and Customer Churn: Using A Machine Learning Approach

Ketaki Dabekar

Data Analytics
National College of Ireland
Dublin,Ireland
x22149619@student.ncirl.ie

***Abstract*—This study focuses on data mining and machine learning techniques for credit card fraud detection, employee attrition, and customer churn. Credit card fraud is a significant problem for financial institutions, while employee attrition and customer churn can have a significant impact on an organization's finances and reputation. The objectives of this study are to develop predictive models to identify fraudulent transactions, employees at risk of leaving the company, and factors contributing to customer churn. We used a variety of techniques, including Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine algorithms, to analyze the datasets.Four distinct machine learning modeling techniques will be used in this research on three data sets.The study demonstrates the potential of machine learning in addressing real-world problems and provides insights for businesses to make informed decisions.**

For the credit card fraud detection dataset,the results showed that logistic regression models had the highest accuracy and AUC scores. For the employee attrition dataset,random forest classifiers and SVM models performing slightly better. For the customer churn dataset,the results showed that random forest classifiers had the highest accuracy and AUC scores, while logistic regression models had the highest precision and recall scores.

***Index Terms*—Credit card fraud,Fraud detection,Machine learning,Classification,Logistic regression, Decision trees, Random forests,SVM,Employee attrition, Employee turnover,Human resources,Customer churn,Customer retention,Customer behavior**

I. Introduction

Machine learning is a subbranch of artificial intelligence (AI) [1] that automatically develops and learns from past data. Machine learning is increasingly being used in our daily lives.

The increasing use of electronic payment systems, coupled with the growth of online commerce, has led to a rise in credit card fraud. Financial institutions are constantly seeking ways to identify fraudulent transactions and prevent losses. On the other hand, employee attrition and customer churn are critical issues that can have a significant impact on an organization's finances and reputation.

Data mining and machine learning techniques have emerged as effective tools for identifying patterns and insights in large datasets. In this study, we use these techniques to analyze three datasets: credit card fraud detection, employee attrition, and customer churn.

The objectives of this study are to develop predictive models to identify fraudulent transactions, employees at risk of leaving the company, and factors contributing to customer churn. By doing so, we aim to demonstrate the effectiveness of data mining and machine learning techniques in detecting fraud, predicting employee attrition, and identifying factors that contribute to customer churn.

The rest of this paper is organized as follows. In the Methodology section, we tools and techniques which I am applying on dataset.In dataset section i outline steps which i am taking on dataset before apply algorithm. In the Results section, I present the findings of our analysis, including the performance of each algorithm on each dataset. Finally, we discuss the implications of our results in conclusion and offer suggestions for future research.

II. Research questions

1. RQ1 – Do the different implementations of identically-named machine learning techniques perform exactly the same? If not, what are the outstanding implementations of the identically-named machine learning techniques for specific evaluation measures? That is, if different implementations of the identically-named techniques perform differently, which implementation is better than the others in each dataset? This will identify the best-performing implementation of each.

2. RQ2 – How do the datasets employed in this work differ from each other?Specifically, how the three datasets employed are different from each other in terms of their characteristics which can impact the effectiveness of machine learning techniques?

III. INITIAL LITERATURE REVIEW

A. Customer Churn Prediction:

Credit card fraud is a significant issue in the financial industry, costing billions of dollars annually. Many studies

have focused on developing effective fraud detection methods.Customer churn is a major problem for companies as it leads to a loss of revenue and customers. Several studies have been conducted to predict customer churn using various techniques such as decision trees, logistic regression, and artificial neural networks [4]. In recent years, machine learning algorithms such as random forest, support vector machine, and gradient boosting have been used to predict customer churn with high accuracy [5].

For example, Wangetal. (2021) proposed a novel deep learning-based approach for detecting credit card fraud, which achieved high accuracy and outperformed traditional machine learning models. Similarly, Liuetal. (2020) developed a hybrid model using a self-organizing map and a support vector machine for credit card fraud detection. They reported significant improvements in fraud detection rates compared to traditional methods.

B. Employee Attrition Prediction:

Employee attrition is a major concern for organizations as it not only leads to a loss of talent but also incurs a significant cost in terms of recruitment and training of new employees. Several studies have been conducted to predict employee attrition using various techniques such as logistic regression, decision trees, and artificial neural networks [6]. In recent years, machine learning algorithms such as random forest, support vector machine, and gradient boosting have been used to predict employee attrition with high accuracy [7].

For example, Krishnan and Viswanathan (2017) used machine learning algorithms such as logistic regression, decision trees, and random forests to predict employee attrition. They reported high accuracy rates and identified several key factors that contribute to employee attrition, including job satisfaction and salary.

C. Credit Card Fraud Detection:

Credit card fraud has become a major issue for banks and credit card companies. The losses incurred due to fraud have increased significantly in recent years. Several studies have been conducted to detect fraud using various techniques such as rule-based systems, neural networks, and decision trees [8]. However, with the advancement in machine learning techniques, researchers have started using machine learning algorithms to detect fraud. In recent years, logistic regression, decision tree, random forest, and support vector machine algorithms have been used to detect fraud with high accuracy [9].

For example, Adewumi and Odusanya (2021) proposed a hybrid model that combines machine learning algorithms such as logistic regression, decision trees, and random forests for customer churn prediction. They reported high accuracy rates and identified several key factors that contribute to customer churn, including customer satisfaction and service quality.

Overall, the literature suggests that machine learning algorithms such as logistic regression, decision trees, random forests, and support vector machines can be effective for credit card fraud detection, employee attrition prediction, and customer churn prediction. These algorithms can analyze large volumes of data and identify patterns that are difficult for humans to detect.

IV. METHODOLOGY

Machine learning is a branch of Artificial Intelligence referring to systems, which learn to solve a set of prediction or detection tasks, based on learning from data. Machines can be trained in three broad ways of supervised, unsupervised, and reinforcement training methodologies [1]. ML algorithms enable analysis of massive quantities of information that is otherwise not possible or extremely time-consuming for manual analysis by humans.[3] The performance of ML systems can be quantitatively assessed using different error metrics.

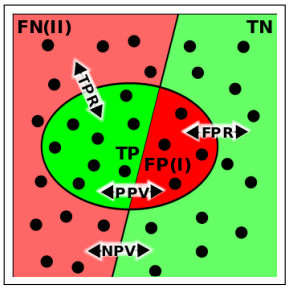


Fig. 1. Binary Classification

In all three datasets, we are dealing with binary classification tasks where we need to predict whether a credit card transaction is fraudulent, whether an employee is likely to leave the company, or whether a customer is likely to churn.Fig 1. shows binary classification image. Therefore, the main goal of these datasets is to develop models that can accurately predict the target variable based on the available features.I have decided to apply Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and Support Vector Machine (SVM) algorithms on dataset. Because all are classification algorithms commonly used in machine learning.They fall under the category of supervised learning algorithms, which means they learn to predict output values based on input data, by being trained on labeled examples.Specifically, Logistic Regression is a linear algorithm that predicts the probability of a binary outcome. Decision Tree Classifier and Random Forest Classifier are tree-based algorithms that can handle both binary and multi-class classification problems. SVM is a non-linear algorithm that finds a decision boundary that maximally separates the different classes.

Each dataset should implement eight models.The data were analyzed by implementing the Logistic Regression algorithm on the dataset using two different libraries, namely Library 'Sklearn' and Library 'statsmodels'. We trained and evaluated on two different models, one using Library Sklearn and the other using Library statsmodels.For Decision Tree Classifier algorithm implemented using Sklearn and XGBoost libraries.For Random Forest Classifier algorithm implemented

using Scikitlearn and XGBoost libraries.Similar libraries implemented to SVM algorithm.The same process followed by the rest of the datasets.All of these algorithms have been tested, and the results are measured in terms of accuracy, classification report, confusion matrix, precision, recall, f1-score, support, AUC, and ROC curve.

In this research, we have used CRISP-DM (Cross Industry Standard Process for Data Mining) method. Basically CRISP-DM means business and data understanding, data preparation, modeling, evaluation and deployement.It can be essential to go back and forth between the various phases because the order of the steps is not specific, Fig. 2 shows the flow chart of the same.

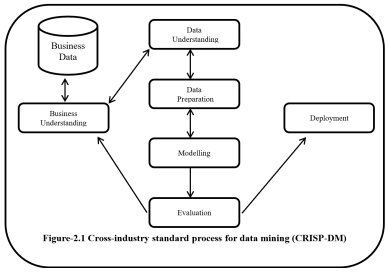


Fig. 2. CRISP-DM Flow Chart

This work was primarily conducted using Jupyter Notebook on a Windows operating system. The Excel worksheet is used for data loading and the Python script utilized for coding purposes.

V. Customer churn prediction - I

The first stage of the CRISP-DM process in data mining and machine learning projects. The Customer Churn Prediction dataset is a collection of customer information from a telecommunications company.Data is selected from the source scikit-learn website.[2] Dataset contains number of records:7043, number of column 21. The dataset is in CSV format. Fig.3. shows the description of each column.

	Column Name	Description
0	customerID	Customer ID
1	gender	Customer gender (Male/Female)
2	SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)
3	Partner	Whether the customer has a partner or not (Yes, No)
4	Dependents	Whether the customer has dependents or not (Yes, No)
5	tenure	Number of months the customer has stayed with the company
6	PhoneService	Whether the customer has a phone service or not (Yes, No)
7	MultipleLines	Whether the customer has multiple lines or not (Yes, No, No phone service)
8	InternetService	Customer's internet service provider (DSL, Fiber optic, No)
9	OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)
10	OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)
11	DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)
12	TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)
13	StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)
14	StreamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)
15	Contract	The contract term of the customer (Month-to-month, One year, Two year)
16	PaperlessBilling	Whether the customer has paperless billing or not (Yes, No)
17	PaymentMethod	Payment method of the customer (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
18	MonthlyCharges	The amount charged to the customer monthly
19	TotalCharges	The total amount charged to the customer
20	Churn	Whether the customer churned or not (Yes or No)

Fig. 3. Column Description

A. Business and Data Understanding :

The goal of this dataset is to predict whether or not a customer will churn, or discontinue their service with the company.The dataset includes various customer attributes such as demographic data, tenure, service usage, and contract information, which can be analyzed to identify patterns and predictors of customer churn. The model will help the business to identify customers who are at risk of churn, enabling targeted retention strategies and reducing overall churn rate.

The target variable which I am using for all models is "Churn". The data types of the columns include integers, floats, and strings.The dataset contains 26.5% of customers that are having churned.The dataset has no missing values, and the attributes have different data types, including categorical, numerical, and binary.In Fig.4. I am showing statistical summary of numeric columns.

	tenure	MonthlyCharges	TotalCharges
count	7043.000000	7043.000000	7043.000000
mean	32.371149	64.761692	2279.581350
std	24.559481	30.090047	2264.729447
min	0.000000	18.250000	0.000000
25%	9.000000	35.500000	394.000000
50%	29.000000	70.350000	1393.600000
75%	55.000000	89.850000	3786.100000
max	72.000000	118.750000	8550.000000

Fig. 4. Statistical summary of numeric columns

B. Data preparation :

Data preparation is a crucial step in any data mining or machine learning project, as it involves transforming raw data into a format that can be used for analysis.

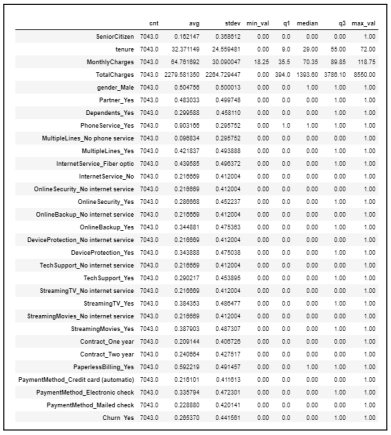
- Data cleaning : This involves removing or correcting any incorrect, or irrelevant data in the dataset."Totalcharges" is given as object datatype but it is float datatype so I just convert that into float. Then there is issue with many rows that they are not showing appropriate value for Totalcharges column so I just compute the formula and derived values for few rows.I dropped "customerID" column from dataset because it is not needed in model building.
  I used SimpleImputer library so this library impute missing values in datasets with mean, median, or most frequent values.I used this only on Totalcharges column because this column contains blank values.
- Data transformation and Data encoding : Hot encoding (also known as one-hot encoding) is a technique used to convert categorical data into a format that can be easily understood and processed by machine learning algorithms.In this process, each category is converted into a binary feature that has a value of 1 or 0, representing the presence or absence of the category in the data. This is necessary because most machine learning algorithms cannot work with categorical data directly.The major columns in the dataset contain categorical data,therefore I choose this technique and implement on dataset.Fig.5. shows columns which i is used for hot encoding. In that

image we can recognize that using hot encoding method it is providing me separation of column according to categorical values.

- Feature selection : I used RFE which stands for Recursive Feature Elimination, which is a feature selection method used in machine learning to select the most important features from a dataset.The algorithm used recursively elimination technique. We need to pass the number of features to be selected it can be done by manually or it can be determined automatically using cross-validation. I passed N values as 20. That is it will take 20 columns further for model building and evaluation.

C. EDA :

Exploratory Data Analysis (EDA) is an important step in any data mining and machine learning project.



	cnt	avg	stdev	min_val	q1	median	q3	max_val
SeniorCitizen	7043.0	0.162147	0.368612	0.00	0.0	0.00	0.00	1.00
tenure	7043.0	32.371149	24.559481	0.00	9.0	29.00	55.00	72.00
MonthlyCharges	7043.0	64.761692	30.090047	18.25	35.5	70.35	89.85	118.75
TotalCharges	7043.0	2279.581350	2264.729447	0.00	394.0	1393.00	3786.10	8550.00
gender_Male	7043.0	0.504756	0.500013	0.00	0.0	1.00	1.00	1.00
Partner_Yes	7043.0	0.483033	0.499748	0.00	0.0	0.00	1.00	1.00
Dependents_Yes	7043.0	0.299588	0.458110	0.00	0.0	0.00	1.00	1.00
PhoneService_Yes	7043.0	0.903166	0.295752	0.00	1.0	1.00	1.00	1.00
MultipleLines_No phone service	7043.0	0.096834	0.295752	0.00	0.0	0.00	0.00	1.00
MultipleLines_Yes	7043.0	0.421837	0.493888	0.00	0.0	0.00	1.00	1.00
InternetService_Fiber optic	7043.0	0.439585	0.496372	0.00	0.0	0.00	1.00	1.00
InternetService_No	7043.0	0.216669	0.412004	0.00	0.0	0.00	0.00	1.00
OnlineSecurity_No internet service	7043.0	0.216669	0.412004	0.00	0.0	0.00	0.00	1.00
OnlineSecurity_Yes	7043.0	0.286668	0.452237	0.00	0.0	0.00	1.00	1.00
OnlineBackup_No internet service	7043.0	0.216669	0.412004	0.00	0.0	0.00	0.00	1.00
OnlineBackup_Yes	7043.0	0.344881	0.475363	0.00	0.0	0.00	1.00	1.00
DeviceProtection_No internet service	7043.0	0.216669	0.412004	0.00	0.0	0.00	0.00	1.00
DeviceProtection_Yes	7043.0	0.343888	0.475038	0.00	0.0	0.00	1.00	1.00
TechSupport_No internet service	7043.0	0.216669	0.412004	0.00	0.0	0.00	0.00	1.00
TechSupport_Yes	7043.0	0.290217	0.453895	0.00	0.0	0.00	1.00	1.00
StreamingTV_No internet service	7043.0	0.216669	0.412004	0.00	0.0	0.00	0.00	1.00
StreamingTV_Yes	7043.0	0.384353	0.486477	0.00	0.0	0.00	1.00	1.00
StreamingMovies_No internet service	7043.0	0.216669	0.412004	0.00	0.0	0.00	0.00	1.00
StreamingMovies_Yes	7043.0	0.387903	0.487307	0.00	0.0	0.00	1.00	1.00
Contract_One year	7043.0	0.209144	0.406726	0.00	0.0	0.00	0.00	1.00
Contract_Two year	7043.0	0.240664	0.427517	0.00	0.0	0.00	0.00	1.00
PaperlessBilling_Yes	7043.0	0.592219	0.491457	0.00	0.0	1.00	1.00	1.00
PaymentMethod_Credit card (automatic)	7043.0	0.219101	0.413513	0.00	0.0	0.00	0.00	1.00
PaymentMethod_Electronic check	7043.0	0.335794	0.472301	0.00	0.0	0.00	1.00	1.00
PaymentMethod_Mailed check	7043.0	0.228880	0.420141	0.00	0.0	0.00	0.00	1.00
Churn_Yes	7043.0	0.265370	0.441561	0.00	0.0	0.00	1.00	1.00

Fig. 5. Statistical summary of categorical column.

Fig.5. shows summary statistics charts its contains mean, unique, standard deviation, minimum and maximum, quartiles (Q1,Q2,Q3) values for categorical variable in the dataset.

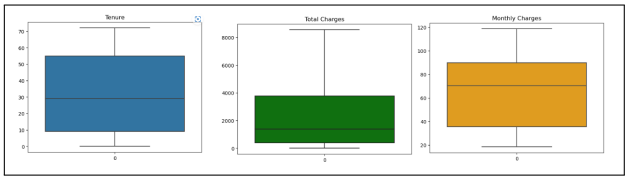


Fig. 6. Outliers analysis

I did outliers analysis with IQR method on numeric columns. There are only 3 column which are in numeric format so i applied outlier analysis only on those column.In Fig.6. shows the particular box plot we can see there is no any outliers in the dataset.

The Fig 7. graph shows the count of customers who churned (left side of the x-axis) and those who did not churn (right side of the x-axis). The height of each bar represents the number of customers in each category. From the graph, we can see that the number of customers

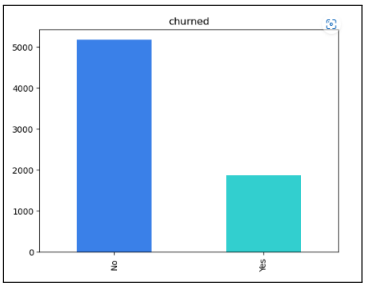


Fig. 7. Target variable distribution

who did not churn is much higher than the number of customers who did churn.

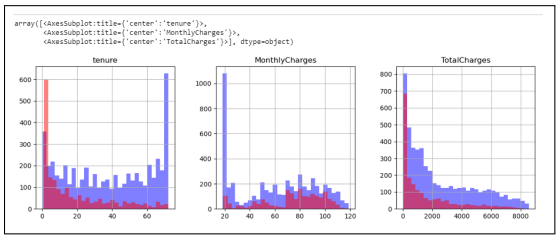


Fig. 8. Distributions of numerical features

I just checked distributions of numerical features in Fig 8. in relation to the target variable. We can observe that the greater TotalCharges and tenure are the less in the probability of churn.A right-skewed distribution for MonthlyCharges and TotalCharges can make the predictive model biased towards the majority class (i.e., non-churned customers). This is because the majority of non-churned customers may have lower MonthlyCharges and TotalCharges.To address this issue, one possible approach is to normalize the data using encoding technique to make the distribution more symmetric. This can help the model to capture the patterns in the minority class (i.e., churned customers) better and improve its predictive performance.

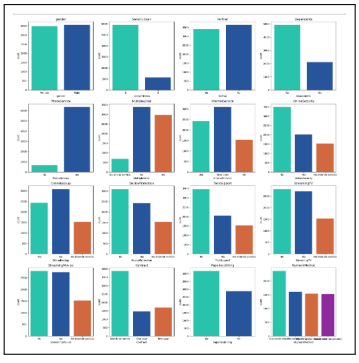


Fig. 9. Categorical feature distribution I

The bar charts (Fig.9) show the count of observations for each category of the respective feature. The resulting visualization allows for quick comparison of the

distribution of each categorical feature, as well as easy identification of the most and least frequent categories.

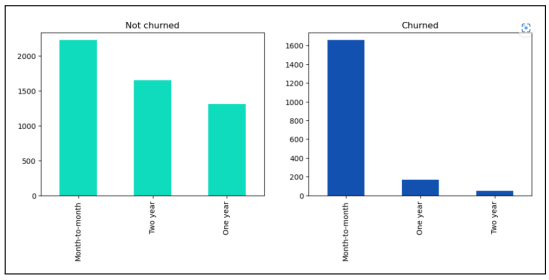


Fig. 10. Categorical feature distribution II

From the plots,(Fig.10.) we can observe that customers who have a month-to-month contract are more likely to churn compared to those with one or two-year contracts. Therefore, contract type may be an important predictor of customer churn.

These two graphs (Fig.11) show the frequency count of customer contracts for one-year and two-year contract periods.We can observe that customers with a two-year contract are much less likely to churn than those with a one-year contract, indicating that longer contract periods may lead to increased customer loyalty.

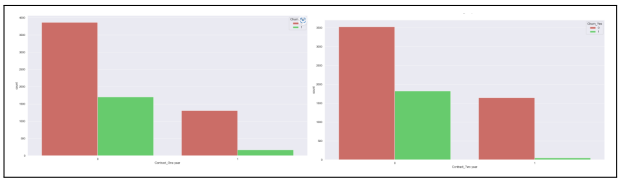


Fig. 11. Categorical feature distribution III

D. Model Training and Evaluation :

Data set is divided into the test and train split using sklearntrain_test_split subset in 80:20 ratio.

- Model 1 Logistic Regression:

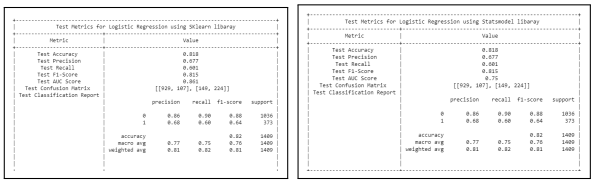


Fig. 12. SKlearn VS Stats models score

Fig 12. Shows scores between models.Both libraries provide similar accuracy scores for both training and testing datasets, as well as similar classification reports and confusion matrices.Overall, both libraries produce similar results for logistic regression.In this case, both models have similar accuracy, precision, recall, and F1-score values. However, the AUC-ROC value of the sklearn logistic regression model is higher (0.861) than the statsmodels logistic regression model (0.750), indicating that the sklearn model performs better at distinguishing between positive and negative classes.As the model seems to be perfectly fitting the training data, but not generalizing well to new data. On the other hand, the sklearn model seems to be more balanced in its performance, with a more realistic recall score that reflects its ability to predict both classes.Based on the AUC-ROC value,It is conclude that the sklearn logistic regression model performs better than the statsmodels logistic regression model on the given dataset.

- Model 2 Decision Tree Classifier:

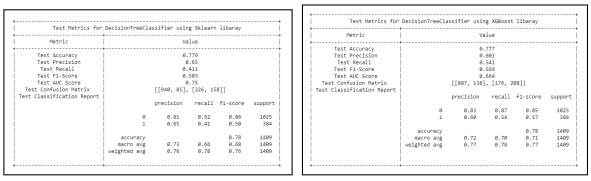


Fig. 13. SKlearn VS XGBoost models score

Fig 13. Shows scores between models.It seems that the XGBoost is performing better than the SKlearn for the given dataset.The XGBoost algorithm has a higher accuracy score for both the training and testing datasets, as well as higher precision, recall, and F1 scores for both datasets. The XGBoost algorithm also has a higher AUC score for the training dataset, indicating better discrimination between the positive and negative classes.However,for the testing dataset, the AUC score is higher for the SKlearn. This may suggest that the decision tree algorithm is better at generalizing to new, unseen data.

- Model 3 Random Forest Classifier: Fig 14. Shows scores

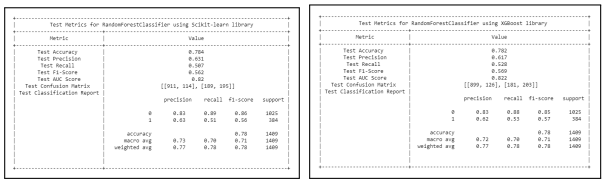


Fig. 14. SKlearn VS XGBoost models score

between models.Both algorithms have similar accuracy on the testing set (around 78%), but SKlearn has slightly higher precision and F1_score values for the positive class (1), and thus a slightly higher recall. However, XGBoost has a higher AUC score on the training set (0.95) compared to SKlearn (0.97), although the AUC score on the testing set is similar for both algorithms (around 0.82).On the testing set, both models have similar accuracy and AUC scores, but the SKlearn has higher precision while the XGBoost model has higher recall.

- Model 4 SVM: Fig 15. Shows scores between models.Comparing both the algorithms' results, we can see that XGBoost performs better than SVM in this case, with a higher training and testing accuracy. Moreover, the AUC

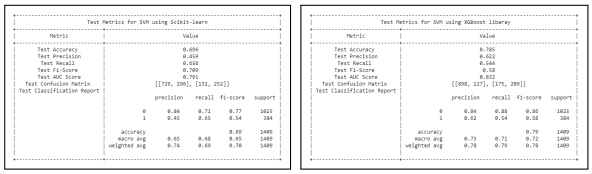


Fig. 15. Scikit-learn and XGBoost models score

score of the XGBoost model is higher than SVM, which indicates that the XGBoost model has a better ability to distinguish between positive and negative classes.

E. Comparison between algorithm implementations :

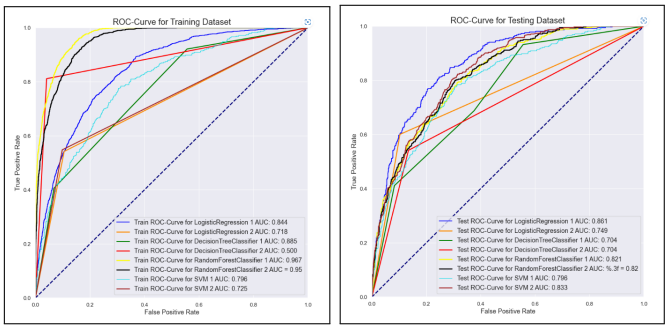


Fig. 16. ROCCurve on test and train dataset.

This graph (Fig 16.) shows the ROC curves for different classification models.The closer the AUC is to 1, the better the classifier is at distinguishing between positive and negative samples.In the training set ROC curve graph, all the classifiers are performing well.The RandomForestClassifier 1 and 2 have the highest AUC value followed by the DecisionTreeClassifier 1,then LogisticRegression 1.In the testing set ROC curve graph, the LogisticRegression 1 has the highest AUC value followed by SVM 2, then RandomForestClassifier 1.

VI. Employee Attrition prediction - II

The Employee Attrition dataset is a collection of employee information from an particular organization.Data is selected from the source github.com website. Dataset contains number of records:1470, number of column 35. The dataset is in CSV format.Fig. 17 shows description of each columns.

	Column	Description
0	Age	Age of the employee
1	Attrition	Whether the employee has left the company or not
2	BusinessTravel	Frequency of the employee travelling for business purposes
3	DailyRate	Daily rate of pay
4	Department	Department of the employee
5	DistanceFromHome	Distance of employee from home to work in miles
6	Education	Education level of the employee
7	EducationField	Field of education of the employee
8	EmployeeCount	Number of employees in the company
9	EmployeeNumber	Employee identification number
10	EnvironmentSatisfaction	Employee satisfaction level with the work environment
11	Gender	Gender of the employee
12	HourlyRate	Hourly rate of pay
13	JobInvolvement	Level of involvement of the employee with the job
14	JobLevel	Job level of the employee
15	JobRole	Role of the employee in the company
16	JobSatisfaction	Employee satisfaction level with the job
17	MaritalStatus	Marital status of the employee
18	MonthlyIncome	Monthly income of the employee
19	MonthlyRate	Monthly rate of pay
20	NumCompaniesWorked	Number of companies the employee has worked for
21	Over18	Whether the employee is over 18 years of age or not
22	OverTime	Whether the employee works overtime or not
23	PercentSalaryHike	Percentage increase in salary
24	PerformanceRating	Performance rating of the employee
25	RelationshipSatisfaction	Employee satisfaction level with relationships at work
26	StandardHours	Standard hours of work
27	StockOptionLevel	Level of stock option
28	TotalWorkingYears	Total years of work
29	TrainingTimesLastYear	Number of times the employee was trained last year
30	WorkLifeBalance	Employee satisfaction level with work-life balance
31	YearsAtCompany	Number of years the employee has worked at the company
32	YearsInCurrentRole	Number of years the employee has been in the current role
33	YearsSinceLastPromotion	Number of years since the employee was last promoted
34	YearsWithCurrManager	Number of years the employee has been with the current manager

Fig. 17. Column Description

A. Business and Data Understanding :

The Employee Attrition prediction dataset typically contains information about employees such as age, gender, education level, job level, job satisfaction, work-life balance, job involvement, job role, years at company, and salary among others. The dataset is used to predict the likelihood of an employee leaving the company based on these factors.To gain a better understanding of the dataset, it's important to analyze the different columns and their descriptions. This can help in identifying any patterns or correlations between the different variables and employee attrition.

The objective of such a prediction is to help organizations take proactive measures to retain their valuable employees and reduce employee turnover.Attrition is a critical issue for any business.The target variable which I am using for all models is "Attrition".And rest column considered as independent column.The dataset does not contain any missing values,duplicate and NA values.The attributes have different data types, including categorical, numerical, and binary.

B. Data preparation :

Data preparation for the employee attrition dataset may involve the following steps:

- Data cleaning : The EmployeeCount, EmployeeNumber, Over18, StandardHours columns only have one unique value, it doesn't provide any meaningful information for analysis so I dropped those columns without affecting the analysis results.Check outliers analysis on the numeric columns.Fig 18. shows box plot of numeric columns. We can see MonthlyIncome column contains few outlier so remove those outliers from dataset I used IQR method.

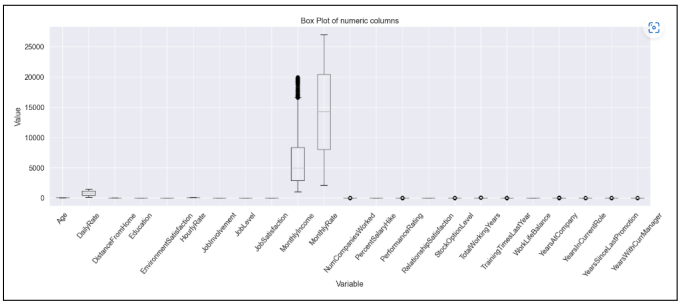
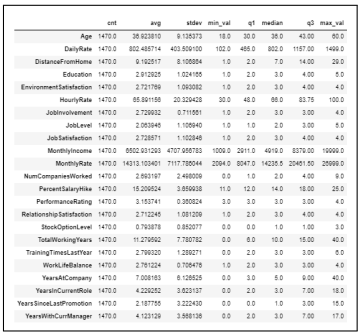


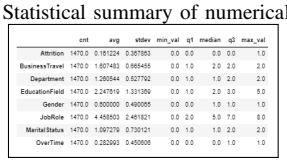
Fig. 18. Outliers analysis

- Data transformation and Data encoding : I convert categorical features into numeric values because most of the tools works with numbers.So for this i used Scikitlearn label encoding to encode character data into numeric values. There are 8 number of categorical columns in the dataset so i converted all those categorical column using Scikitlearn label encoding technique.Fig 20. shows categorical features and it's Statistical summary.
- Feature selection : It is the process of choosing the best features that can be used in the predictive modeling. For Feature selection i used mutual_info_classif function from the scikitlearn library.This function used to compute



	cnt	avg	stdev	min_val	q1	median	q3	max_val
Age	1470.0	36.923810	9.135373	18.0	30.0	36.0	43.00	60.0
DailyRate	1470.0	802.485714	403.509100	102.0	465.0	802.0	1157.00	1499.0
DistanceFromHome	1470.0	9.192517	8.106864	1.0	2.0	7.0	14.00	29.0
Education	1470.0	2.912925	1.024165	1.0	2.0	3.0	4.00	5.0
EnvironmentSatisfaction	1470.0	2.721769	1.093082	1.0	2.0	3.0	4.00	4.0
HourlyRate	1470.0	65.891156	20.329428	30.0	48.0	66.0	83.75	100.0
JobInvolvement	1470.0	2.729932	0.711561	1.0	2.0	3.0	3.00	4.0
JobLevel	1470.0	2.063946	1.106940	1.0	1.0	2.0	3.00	5.0
JobSatisfaction	1470.0	2.728571	1.102846	1.0	2.0	3.0	4.00	4.0
MonthlyIncome	1470.0	6502.931293	4707.956783	1009.0	2911.0	4919.0	8379.00	19999.0
MonthlyRate	1470.0	14313.103401	7117.786044	2094.0	8047.0	14235.5	20461.50	26999.0
NumCompaniesWorked	1470.0	2.693197	2.498009	0.0	1.0	2.0	4.00	9.0
PercentSalaryHike	1470.0	15.209524	3.659938	11.0	12.0	14.0	18.00	25.0
PerformanceRating	1470.0	3.153741	0.360824	3.0	3.0	3.0	3.00	4.0
RelationshipSatisfaction	1470.0	2.712245	1.081209	1.0	2.0	3.0	4.00	4.0
StockOptionLevel	1470.0	0.793878	0.852077	0.0	0.0	1.0	1.00	3.0
TotalWorkingYears	1470.0	11.279592	7.780782	0.0	6.0	10.0	15.00	40.0
TrainingTimesLastYear	1470.0	2.799320	1.289271	0.0	2.0	3.0	3.00	6.0
WorkLifeBalance	1470.0	2.761224	0.706476	1.0	2.0	3.0	3.00	4.0
YearsAtCompany	1470.0	7.008163	6.126525	0.0	3.0	5.0	9.00	40.0
YearsInCurrentRole	1470.0	4.229252	3.623137	0.0	2.0	3.0	7.00	18.0
YearsSinceLastPromotion	1470.0	2.187755	3.222430	0.0	0.0	1.0	3.00	15.0
YearsWithCurrManager	1470.0	4.123129	3.568136	0.0	2.0	3.0	7.00	17.0

Fig. 19. Statistical summary of numerical column.



	cnt	avg	stdev	min_val	q1	median	q3	max_val
Attrition	1470.0	0.161224	0.367863	0.0	0.0	0.0	0.0	1.0
BusinessTravel	1470.0	1.607483	0.885456	0.0	1.0	2.0	2.0	2.0
Department	1470.0	1.260544	0.527792	0.0	1.0	1.0	2.0	2.0
EducationField	1470.0	2.247619	1.331369	0.0	1.0	2.0	3.0	5.0
Gender	1470.0	0.600000	0.490065	0.0	0.0	1.0	1.0	1.0
JobRole	1470.0	4.458503	2.461821	0.0	2.0	5.0	7.0	8.0
MaritalStatus	1470.0	1.097279	0.730121	0.0	1.0	1.0	2.0	2.0
OverTime	1470.0	0.282993	0.450606	0.0	0.0	0.0	1.0	1.0

Fig. 20. Statistical summary of categorical column.

the mutual information between a set of features and a target variable.This function returns a score for each feature, which can be used to rank the features.The mutual information score is higher if the feature is highly dependent on the target variable, and lower if the feature is independent of the target variable. This score can be used to select the most relevant features for classification models.From mutual_info_classif function selected ”15” number of columns for model building purpose.

C. EDA :

EDA can provide insights into factors that contribute to employee attrition and help identify potential predictors.Fig 21 shows numerical features and it’s Statistical summary.

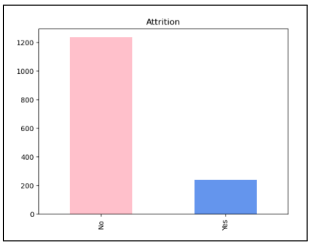


Fig. 21. Target variable distribution

The graph (Fig.21) gives an idea of the proportion of employees who have left the company compared to those who have stayed. From the graph, we can see that the number of employees who have left the company is relatively smaller than the number of employees who are still with the company.

This graph(Fig.22) shows the density plot of age for churned and non-churned employees. The red curve represents the density of age for employees who have churned, and the blue curve represents the density of age for employees who have not churned.From the graph, we

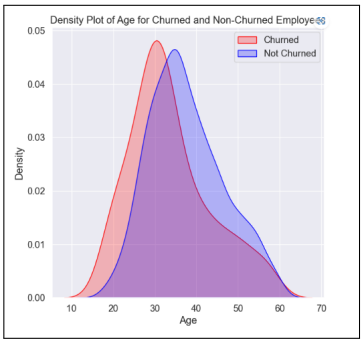


Fig. 22. Density Plot of Age for Churned and Non-Churned Employees

can see that the age distribution of churned employees is skewed towards the younger age group compared to the age distribution of non-churned employees. The peak of the red curve is around 28-30 years, while the peak of the blue curve is around 35-38 years. This indicates that younger employees are more likely to churn compared to older employees.

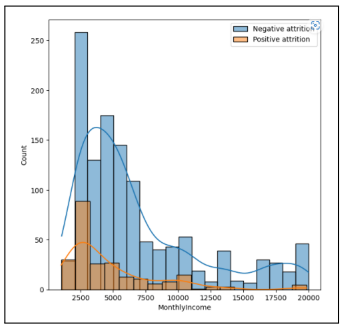


Fig. 23. Monthly income for employees

The graph(Fig.23.) helps to visually compare the two distributions and see if there are any noticeable differences in the monthly income range of employees who left the company compared to those who stayed.we can see that the density of employees with negative attrition is higher in the middle range of monthly income, while the density of employees with positive attrition is higher in the lower range of monthly income. This suggests that employees with lower monthly income are more likely to leave the company.

The graph (Fig.24.)shows the count of employees who have left the company for each job role. The x-axis shows the job roles and the y-axis shows the count of employees. From observation top three roles facing attrition 62 number of employees who are likely to quit belong to Laboratory Technician group.57 number of employees belong to Sales Executive group.47 number of employees belong to Research Scientist group.

This graph(Fig.25.) shows the number of employees who work overtime and those who don’t, separated by attrition status.We can observe that there are more employees who do not work overtime than those who do. Additionally, the

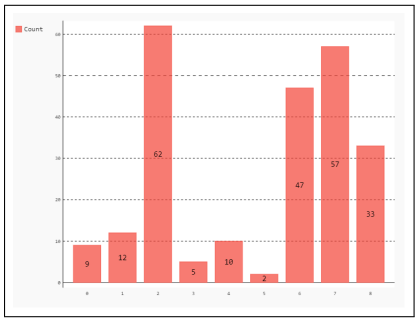


Fig. 24. Count of employees for each job role

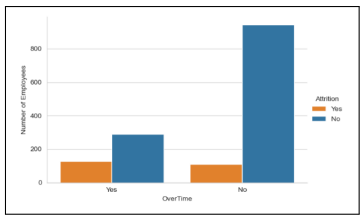


Fig. 25. Categorical feature distribution

number of employees who experienced attrition is higher for those who work overtime compared to those who don't work overtime. This indicates that there may be a relationship between working overtime and employee attrition.

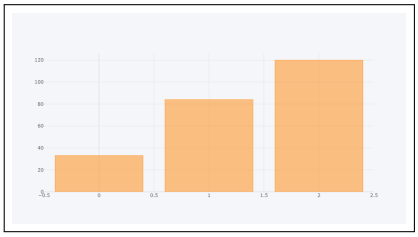


Fig. 26. Categorical feature distribution

The graph(Fig.26.) shows the count of employees for each category of Marital Status - single, married, or divorced. It seems that a large portion of employees who had positive attrition were single, followed by married employees, and then divorced employees.

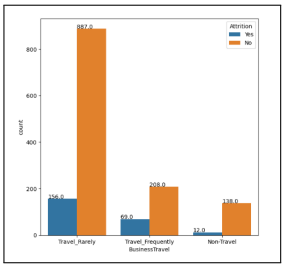


Fig. 27. Categorical feature distribution

The graph(Fig.27.) indicates that the majority of employees in the dataset travel rarely for business, and a smaller proportion travel frequently. Among those who do not travel, the proportion of employees who left the company is relatively low, while among those who travel frequently, the proportion of employees who left the company is relatively high. This suggests that frequent business travel may be a factor that contributes to employee attrition.

D. Model Training and Evaluation :

Data set is divided into the test and train split using sklearntrain_test_split subset in 80:20 ratio.

- Model 1 Logistic Regression:

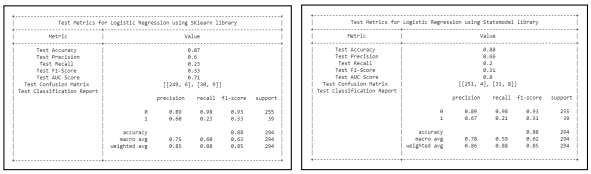


Fig. 28. Sklearn and Stats-model score

Fig 28. Shows scores between models.Both the sklearn and the statsmodel seem to have very similar results with only slight differences in their evaluation metrics.The accuracy scores for both models are very close sklearn having an accuracy more than statsmodel model.The precision,recall,F1 scores are also quite similar for both model.In terms of the confusion matrices, both models seem to be able to predict non-default cases (class 0) very well with very few false positives. However, both models have a relatively high number of false negatives for the default cases (class 1).Overall, both models seem to have similar performance on this dataset, but the statsmodel model has slightly better performance in terms of precision.

- Model 2 Decision Tree Classifier:

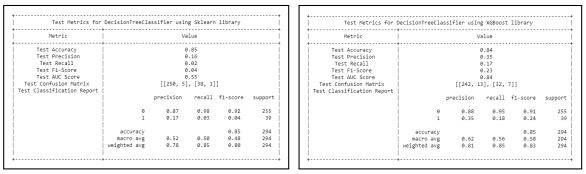


Fig. 29. Sklearn and XGBoost models score

Fig 29. Shows scores between models.Based on the results model provided, it seems that the Sklearn and XGBoost models have similar accuracy scores on the testing data, with the Sklearn having a slightly higher score.However, when it comes to precision, recall, and F1 scores, the XGBoost model outperforms the Sklearn. The XGBoost model has higher precision, recall, and F1 scores for the positive class (1), indicating that it is better at identifying the positive class than the Sklearn.In terms of the confusion matrices, the Sklearn correctly identifies more TN, while XGBoost correctly identifies more TP. This suggests that the Sklearn may be better at

identifying the negative class (0) while XGBoost is better at identifying the positive class (1).it appears that the XGBoost model outperforms the Decision Tree Classifier in terms of identifying the positive class.

- Model 3 Random Forest Classifier: Fig 30. Shows scores

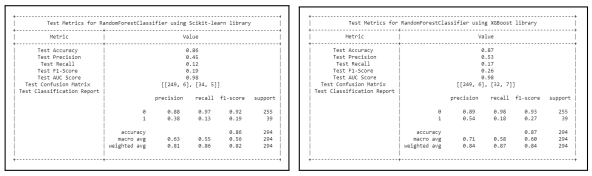


Fig. 30. Sklearn and XGBoost models score

between models.Based on the evaluation metrics, the XGBoost model appears to perform slightly better than the Sklearn model.The XGBoost model has a slightly higher accuracy, precision, recall, F1 score, and AUC score on the test set. The confusion matrix also shows a slightly better performance for the XGBoost model.Therefore, if you are comparing these two models, it seems that the XGBoost model is the better than Sklearn model.

- Model 4 SVM: Fig 31. Shows scores between models.Based on the metrics provided, it appears that the XGBoost classifier outperforms the Scikit-learn classifier on all the evaluation metrics.The XGBoost classifier achieved a higher accuracy score and higher precision, recall, and F1-score for the positive class. Additionally, the XGBoost classifier achieved a higher AUC score which indicates better overall performance.Therefore, it seems that the XGBoost classifier is a better choice for this problem compared to the SVM classifier.

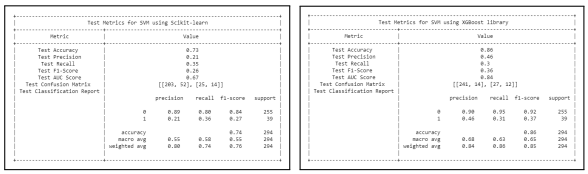


Fig. 31. Scikit-learn and XGBoost models score

E. *Comparison between algoruthm implementation :*

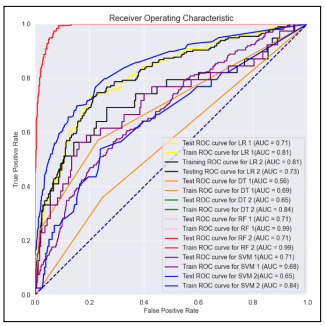


Fig. 32. ROC Curve for train and test data

From the graph (Fig.32),the train Random Forest 1 and 2 has the highest AUC value.The train SVM 2 model and the train Logistic Regression 1 model also performs well.The test decision tree classifier 1 model, the test SVM 2 and test decision tree classifier 2 models perform relatively poorly compared to the other models.Overall, the graph provides a useful visualization of how well each model is able to classify the data.

VII. Credit Card Fraud Detection Dataset - III

The Credit Card Fraud Detection dataset is a public dataset containing credit card transactions made by European cardholders over a two-day period in September 2013. The dataset consists of 284,807 transactions, of which 492 are fraudulent. The dataset is highly imbalanced, with fraudulent transactions accounting for only 0.172% of the total transactions. Fig.33. Shows statistical summary of each column.

	count	mean	std	min	25%	50%	75%	max
Time	284807.0	9.481386e+04	47488.145955	0.000000	54201.500000	84692.000000	139320.500000	172792.000000
V1	284807.0	3.918649e-15	1.958696	-56.407510	-0.920373	0.018109	1.315642	2.454930
V2	284807.0	5.682686e-16	1.651309	-72.715728	-0.598550	0.065486	0.803724	22.057729
V3	284807.0	-8.761736e-15	1.516255	-48.325589	-0.890365	0.179846	1.027196	9.382558
V4	284807.0	2.811118e-15	1.415869	-5.683171	-0.848640	-0.019847	0.743341	16.875344
V5	284807.0	-1.552103e-15	1.380247	-113.743307	-0.691597	-0.054336	0.611926	34.801666
V6	284807.0	2.040130e-15	1.332271	-26.160506	-0.768296	-0.274187	0.398565	73.301626
V7	284807.0	-1.698953e-15	1.237094	-43.557242	-0.554076	0.040103	0.570436	120.589494
V8	284807.0	-1.893285e-16	1.194353	-73.216718	-0.208630	0.022358	0.327346	20.007208
V9	284807.0	-3.147640e-15	1.098632	-13.434066	-0.643098	-0.051429	0.597139	15.594995
V10	284807.0	1.772925e-15	1.088850	-24.588262	-0.535426	-0.092917	0.453923	23.745136
V11	284807.0	9.289524e-16	1.020713	-4.797473	-0.762494	-0.032757	0.739593	12.018913
V12	284807.0	-1.803266e-15	0.999201	-18.683715	-0.405571	0.140033	0.618238	7.848392
V13	284807.0	1.674888e-15	0.995274	-5.791881	-0.648539	-0.013568	0.662506	7.126883
V14	284807.0	1.475821e-15	0.958598	-19.214325	-0.425674	0.050601	0.493150	10.526766
V15	284807.0	3.501098e-15	0.915316	-4.498945	-0.582884	0.048072	0.648821	8.877742
V16	284807.0	1.392460e-15	0.876253	-14.129855	-0.468037	0.066413	0.523296	17.315112
V17	284807.0	-7.466538e-16	0.849337	-25.162799	-0.483748	-0.065678	0.399675	9.253526
V18	284807.0	4.258754e-15	0.838176	-9.498746	-0.498850	-0.003636	0.500807	5.041069
V19	284807.0	9.019919e-16	0.814041	-7.213527	-0.456299	0.003735	0.458949	5.591971
V20	284807.0	5.128845e-16	0.770925	-54.497720	-0.211721	-0.062481	0.133041	39.420904
V21	284807.0	1.473120e-16	0.734524	-34.830382	-0.228395	-0.029450	0.186377	27.202839
V22	284807.0	8.042109e-16	0.725702	-10.933144	-0.542350	0.006782	0.528554	10.503090
V23	284807.0	5.282512e-16	0.624460	-44.807735	-0.161846	-0.011193	0.147642	22.528412
V24	284807.0	4.456271e-15	0.605647	-2.836627	-0.354586	0.040976	0.439527	4.584549
V25	284807.0	1.426896e-15	0.521278	-10.295397	-0.317145	0.016594	0.350716	7.519589
V26	284807.0	1.701640e-15	0.482227	-2.604551	-0.326984	-0.052139	0.240952	3.517346
V27	284807.0	-3.662252e-16	0.403632	-22.565679	-0.070840	0.001342	0.091045	31.612198
V28	284807.0	-1.217809e-16	0.330083	-15.430084	-0.052960	0.011244	0.078280	33.847808
Amount	284807.0	8.834962e+01	250.120109	0.000000	5.600000	22.000000	77.165000	25691.160000
Class	284807.0	1.727486e-03	0.041527	0.000000	0.000000	0.000000	0.000000	1.000000

Fig. 33. Description of columns

A. *Business and Data Understanding :*

Our aim is to build a classification model to detect fraudulent transactions and to minimize the false negatives.We need to ensure that our model's accuracy is high enough to detect as many fraudulent transactions as possible, while keeping the false positive rate low to avoid mistakenly flagging genuine transactions as fraudulent.

The dataset contains a total of 31 features, of which all but two features (Time and Amount) have been transformed using PCA.The 'Class' feature is the target variable, where 1 denotes a fraudulent transaction and 0 denotes a non-fraudulent transaction.

B. *Data preparation :*

To build a classification model following steps were take place :

- Data cleaning : Dataset is already well-prepared and cleaned.There are still a few things that we need to take care of before proceed.Imbalanced data is a common issue in fraud detection datasets, where the number of non-fraudulent transactions is much higher than the number of fraudulent transactions.So for this issue i applied

under sampling technique means build a sample dataset containing similar distribution of normal transactions and Fraudulent Transactions.Dataset contains outlier in amount column so I removed all outliers from dataset using quantile and IQR technique. Fig.34 shows box plot aftering removing outlier from dataset.

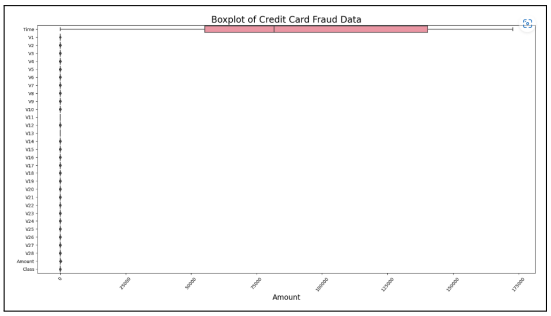


Fig. 34. Outlier Analysis

- Data transformation and Data encoding: It's important to scale the numerical features in order to ensure that all features are on the same scale.The StandardScaler class from the scikitlearn library is used to transform the data.Standardization scales the data to have a mean of 0 and a standard deviation of 1, which can improve the performance models,
- Feature selection : In the case of credit card fraud detection, it is crucial to identify the most relevant features that can help the model differentiate between fraudulent and non-fraudulent transactions. To perform feature selection i used Recursive Feature Elimination with Cross-Validation and pass CV value as 5.

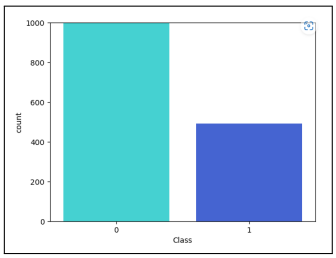


Fig. 35. Target variable distribution

C. EDA :

This Fig 35. plot can help us understand the class distribution in the dataset. In this case, we can see that the number of non-fraudulent transactions is much higher than the number of fraudulent transactions.

The plot(Fig. 36. and 37.) provides a visual comparison of the distribution of transaction amounts for fraudulent and non-fraudulent transactions. It suggests that fraudulent transactions tend to have smaller transaction amounts compared to non-fraudulent transactions.

The x-axis represents the time of the transaction in seconds, while the y-axis represents the amount of the

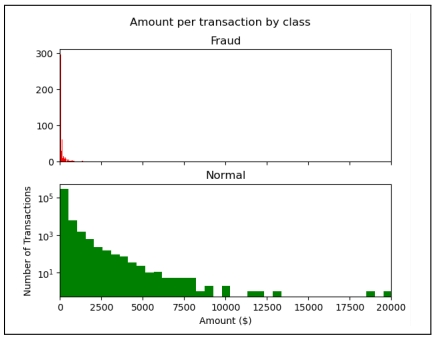


Fig. 36. Amount per transaction by class

transaction in dollars. The plot helps us visualize whether fraudulent transactions occur more often during certain time frames or whether there is a pattern in the amount of the fraudulent transactions.

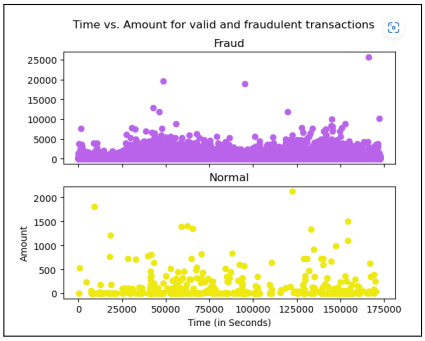


Fig. 37. Time vs. Amount for valid and fraudulent transactions

The histogram(Fig. 38.) shows each numerical variable in the credit card dataset, with the y-axis scaled to a logarithmic scale.The generated histograms can help identify the shape of the distribution of each variable, detect any outliers or unusual patterns, and get an overall sense of the data.Here I only mentioned 4 images but i have plotted histogram for all columns in code.

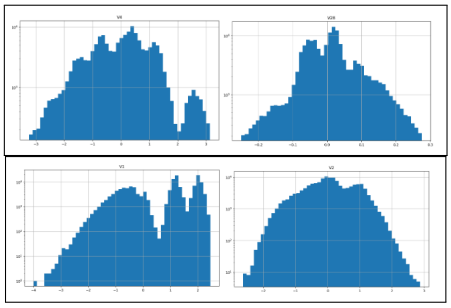


Fig. 38. Distribution of each variable

The below graph(Fig.39) is a heatmap that shows the correlation between each pair of features in the credit card fraud detection dataset. The correlation values are color-coded and displayed in each cell of the heatmap. The range of correlation values is from -1 to 1, where a value

of -1 indicates a perfect negative correlation between the two features, a value of 1 indicates a perfect positive correlation, and a value of 0 indicates no correlation between the two features. The darker shades of green indicate higher positive correlation, while the darker shades of red indicate higher negative correlation. The purpose of this graph is to identify which features are highly correlated with each other and to see if there is any multicollinearity between them, which can affect the performance of the machine learning model.

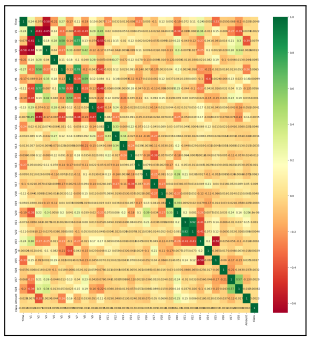


Fig. 39. Heatmap

D. Model Training and Evaluation :

Data set is divided into the test and train split using sklearntrain_test_split subset in 80:20 ratio.

- Model 1 Logistic Regression:

```
Accuracy score on Test Data :  0.9187817258883249
Precision scores  on Test Data : 0.9555555555555556
Recall scores  on Test Data: 0.8775510204081632
F1 scores on Test Data: 0.9148936170212767
Classification report on Test Data :
              precision    recall  f1-score   support

           0       0.89      0.96      0.92        99
           1       0.96      0.88      0.91        98

    accuracy                           0.92       197
   macro avg       0.92      0.92      0.92       197
weighted avg       0.92      0.92      0.92       197

Confusion matrix score on Test Data :
 [[95  4]
 [12 86]]
Log loss on Test Data: 2.805196044204527
```

```
Test Accuracy: 0.883248730964467
Precision scores  on Test Data : 0.9310344827586207
Recall scores  on Test Data: 0.826530612244898
F1 scores on Test Data: 0.8756756756756757
Testing Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.94      0.89        99
           1       0.93      0.83      0.88        98

    accuracy                           0.88       197
   macro avg       0.89      0.88      0.88       197
weighted avg       0.89      0.88      0.88       197

Testing Confusion Matrix:
 [[93  6]
 [17 81]]
Log loss on Test Data: 4.032470328261558
```

Fig. 40. Scikit-learn and XGBoost models score

Fig 40. Shows scores between models.The Logistic Regression model implemented using sklearn seems to perform better than the one implemented using stats library for the dataset. The accuracy, precision, recall, and F1 scores are all higher for the model implemented using sklearn. Additionally, the AUC score for the testing dataset is also higher for the model implemented using sklearn. Therefore, it can be concluded that the Logistic Regression model implemented using sklearn is more accurate for this particular dataset.

- Model 2 Decision Tree Classifier:

Fig 41. Shows scores between models.The Xgboost model appears to be more accurate than the Scikit-learn model. The Xgboost model has a higher test accuracy score compared to the Scikit-learn model. Additionally, the Xgboost model has a lower test log loss and a higher AUC score on testing data compared to the Scikit-learn

```
Test Accuracy: 0.8883248730964467
Test confusion matrices:
 [[84 15]
 [ 7 91]]
Test precision scores: 0.8584905660377359
Test recall scores: 0.9285714285714286
Test F1 scores: 0.8921568627450982
Test Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.85      0.88        99
           1       0.86      0.93      0.89        98

    accuracy                           0.89       197
   macro avg       0.89      0.89      0.89       197
weighted avg       0.89      0.89      0.89       197

Test Log loss on : 3.8571831200481626
```

```
Test accuracy scores: 0.9137055837563451
Test confusion matrices:
 [[93  6]
 [11 87]]
Test precision scores: 0.9354838709677419
Test recall scores: 0.8877551020408163
Test F1 scores: 0.9109947643979057
Test Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.94      0.92        99
           1       0.94      0.89      0.91        98

    accuracy                           0.91       197
   macro avg       0.91      0.91      0.91       197
weighted avg       0.91      0.91      0.91       197

Training Log loss : 2.980527899990167
```

Fig. 41. Scikit-learn and XGBoost models score

model.In terms of precision, recall, and F1 scores, the Xgboost model performs better for class 0, while the Scikit-learn model performs better for class 1. However, since both models have similar performance in these metrics, the overall better performance of the Xgboost model based on accuracy, log loss, and AUC score suggests that it is the more accurate model for this particular dataset.

- Model 3 Random Forest Classifier: Fig 42. Shows scores

```
Test Accuracy: 0.9086294416243654
Test Precision: 0.9347826086956522
Test Recall: 0.8775510204081632
Test F1 Score: 0.9052631578947369
Test AUC Score: 0.968872397443826
Test Confusion Matrix:
 [[93  6]
 [12 86]]
Test Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.94      0.91        99
           1       0.93      0.88      0.91        98

    accuracy                           0.91       197
   macro avg       0.91      0.91      0.91       197
weighted avg       0.91      0.91      0.91       197

Test Log loss : 3.155851638035399
Testing AUC score: 0.968872397443826
```

```
Test Accuracy: 0.9035532994923858
Test Precision: 0.9157894736842105
Test Recall: 0.8877551020408163
Test F1 Score: 0.9015544041450776
Test Confusion Matrix:
 [[91  8]
 [11 87]]
Train Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.92      0.91        99
           1       0.92      0.89      0.90        98

    accuracy                           0.90       197
   macro avg       0.90      0.90      0.90       197
weighted avg       0.90      0.90      0.90       197

Test Log loss : 3.331183493821039
```

Fig. 42. Scikit-learn and XGBoost models score

between models.Based on the evaluation metrics above mentioned, the Random Forest model seems to be slightly more accurate than the XGBoost model.The Random Forest model has a higher Test Accuracy, Test Precision, Test F1 Score, and Test AUC Score compared to the XGBoost model. However, the XGBoost model has a slightly higher Test Recall score.Random Forest is typically easier to use and requires less tuning compared to XGBoost, which can be computationally expensive and requires more hyperparameter tuning.

- Model 4 SVM: Fig 43. Shows scores between mod-

```
Test Accuracy:  0.9187817258883249
Test Precision:  0.9555555555555556
Test Recall:  0.8775510204081632
Test F1 Score:  0.9148936170212767
Test Confusion Matrix:
[[95  4]
 [12 86]]
Test Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.96      0.92        99
           1       0.96      0.88      0.91        98

    accuracy                           0.92       197
   macro avg       0.92      0.92      0.92       197
weighted avg       0.92      0.92      0.92       197

Test Log loss : 2.805196044204527
```

```
Test Accuracy: 0.9035532994923858
Test Precision: 0.9157894736842105
Test Recall: 0.8877551020408163
Test F1 Score: 0.9015544041450776
Test Confusion Matrix:
[[91  8]
 [11 87]]
Test Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.92      0.91        99
           1       0.92      0.89      0.90        98

    accuracy                           0.90       197
   macro avg       0.90      0.90      0.90       197
weighted avg       0.90      0.90      0.90       197

Test Log loss : 3.331183493821039
```

Fig. 43. Scikit-learn and XGBoost models score

els.Based on the performance metrics which provided above, it appears that the Xgboost model has better performance than the SVM model on dataset. The Xgboost model has higher accuracy, precision, and F1 score, as well as a lower log loss. The recall score is slightly

lower for the Xgboost, but overall it seems that the Xgboost model is a better choice for this particular task. Additionally, the AUC score is higher for the Xgboost model, indicating that it has better overall performance in terms of correctly identifying both true positives and true negatives.

E. Comparison between algoruthm implementation :

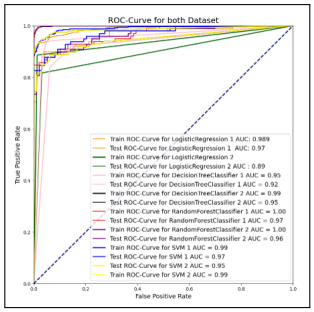


Fig. 44. ROC curve for train and test dataset

The above graph i.e. Fig 44 is a ROC curve for several models used to classify the data.From the graph, we can observe that most of the models have performed well on both the train and test datasets with AUC scores greater than 0.8. The best-performing model is RandomForestClassifier train 1 and 1, with an AUC score of 1 on the train dataset. The worst-performing model is LogisticRegression 2, with an AUC score of 0.89 on the test dataset.We can also observe that there is a trade-off between the True Positive Rate and False Positive Rate, and we can choose a threshold value that balances the two rates.

RESULTS AND DISCUSSION

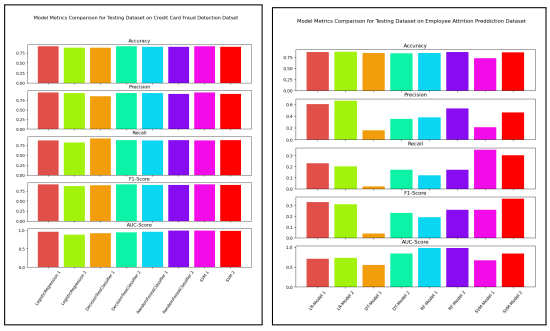


Fig. 45. Score comparision between 3 dataset - I

The graph Fig. 45 and 46 shows the comparison of different ML models on various metrics used for evaluating the performance of the models on the testing dataset of 3. For performance evaluation I used testing accuracy,precision,recall,AUC, F1-score for each model. And in graph we can easily see the overall picture and score of 3 dataset for 8 implementation.

The 3 datasets - credit card fraud detection, employee attrition, and customer churn datasets - differ from each other in terms of their characteristics.These differences can impact the effectiveness of ML techniques. The credit card dataset have imbalanced classes, as the number of fraudulent transactions is likely to be lower than the number of legitimate transactions.The employee attrition and customer churn dataset includes missing values,categorical and numerical variables that need to be transformed into numerical values.Overall,imbalanced classes in the credit card fraud detection dataset i am used under sampling to balance the classes. Similarly, missing values in the employee attrition and customer churn datasets i used imputation techniques and to convert categorical values into numeric i used standard encoding library and Hot encoding technique to fill in missing values before ML algorithms applied.

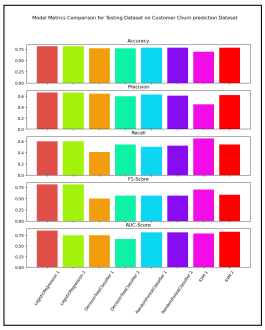


Fig. 46. Score comparision between 3 dataset - II

CONCLUSION AND FUTURE WORK

The graph Fig. 45 and 46 shows or it appears that the credit card fraud detection dataset has the highest performance across all models, with the highest accuracy, precision, recall, f1_score, and AUC. The employee attrition dataset has good performance, with high accuracy and precision, and reasonable recall, f1_score, and AUC. The customer churn prediction dataset has the lowest performance of the three, with relatively low precision, recall, and f1_score.In terms of horizontal comparison, we can see that LogisticRegression and RandomForestClassifier models consistently perform well in all three datasets. SVM and DecisionTree models have mixed performance.In terms of vertical comparison, we can see that the RandomForestClassifier 2(XGBoost) model performs the best overall in terms of AUC, while the LogisticRegression 1(Sklearn) model performs the best overall in terms of accuracy, precision, recall, and F1-score across all datasets.

Continuously improving feature engineering techniques can help in extracting more meaningful insights from the data, which can improve the performance of the models.Credit card fraud detection, employee attrition, and customer churn are all time-sensitive problems. Therefore, it is essential to monitor these datasets in real-time and to develop models that can update their predictions as new data becomes available.

REFERENCES

[1] By:IBM Cloud Education, "What is machine learning?,"IBM,15-Jul2020.[Online]. Available: https://www.ibm.com/cloud/learn/machinelearning

[2] Available :https://huggingface.co/datasets/scikit-learn/churn-prediction/raw/main/dataset.csv.
[3] S.C. Bagley, H. White and B.A. Golomb, "Logistic Regression in the Medical Literature:Standards for Use and Reporting, with Particular Attention to one Medical Domain", Journal of Clinical Epidemiology, Vol. 54, No. 10, pp. 979-985, 2001.
[4] Kim, J., and Lee, J. (2005). Mining customer value from web data with Kohonen's self-organizing map. Expert Systems with Applications, 28(1), 241-250.
[5] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., and Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211-229.
[6] Mohamed, M., and Faisal, S. (2013). Predicting employee turnover using artificial neural network approach. Procedia Technology, 11, 491-497.
[7] Bhattacharyya, S., and Chakraborty, S. (2017). Employee turnover prediction using machine learning algorithms: A case study from the Indian IT industry. Journal of Business Research, 70, 1-9.
[8] Bhattacharya, S., and Bose, I. (2005). A rule-based expert system for credit card fraud detection. Decision Support Systems, 40(1), 81-91.
[9] Phua, C., Lee, V. C., Smith-Miles, K., and Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119.