Project Summary

Batch details	Post Graduate Program – Data Science & Engineering (Pune Nov'22)
Team members	1. Jash Gaddam 2. Shweta Suryavanshi 3. Harshada Karande 4. Jitendra Dixit 5. Parul Bhaiya
Domain of Project	Finance and Risk Analysis
Proposed project title	Online Bank Account Fraud Detection: Machine Learning Model a way forward.
Group Number	07
Team Leader	Jash Gaddam
Mentor Name	Mr. Jatinder Bedi

Date: 26-02-2023

Signature of the Mentor

Signature of the Team Leader

Table of Contents

OVERVIEW

The Bank Account Fraud Dataset (BAF) is a data suite published by NeurIPS (neural information processing systems foundation) 2022 with an aim to provide of synthetic accounts replicating real-world online bank account application fraud activity records. Synthetic identity fraud is a major challenge that is faced by financial institutions where fraudsters create synthetic and relatively believable accounts using false personal identifiable information (PII). Banks and Fintech Sector occasionally regard rule-based systems as being easier to operate and safer to execute. The incorporation of well-trained machine learning algorithms that can analyze huge volumes of information and identify patterns from the same can aid in fraud prevention for businesses.

BUSINESS PROBLEM STATEMENT

Banking industry has been undergoing major digital switch overs to provide enhanced services to the consumers, application for opening a new account is one of them. However, due to a fast-paced advancement in technologies that come with their loopholes, fraudsters have been using sophisticated techniques to commit fraud making it difficult for traditional rule-based systems to detect and prevent fraudulent activities. Banking establishments are highly affected as this type of activities directly harm the bank – consumer relationship and come with negative consequences for both the bank as well as the individuals involved. The banks may have to face situations such as loss of potential new consumers, loss of trust, reputation damages, and any substantial monetary loss for reparations of these situations. The consumers on the other hand may suffer from identity theft, monetary losses in some cases, service dissatisfaction, and sentiment affliction in some cases.

1. **Business Problem Understanding**

   In today's digitalized world, numerous financial institutions have gone through a digital transition and have begin providing various services that can be accessible for the consumers through the internet and/or web-based application for an easier and time saving method. However due to an increase in data breaches, there has been an adverse effect which has caused a rise in fraudulent activities such as new account fraud. New bank account fraud is a crucial problem faced by banking and fintech sector where fraudsters, scammers, money mules have been creating fake accounts by using breached data information of individuals and committing fraudulent activities. Although there have been certain rule–based systems implemented for control of these activities, financial institutions are turning towards artificial intelligence and machine learning techniques to build a solid, attested and secure system for detection of such activities

2. **Business Objective**

   To provide the bank/ financial institutes/financial government establishments (RBI) by developing an efficient and fair machine learning model for detection of fraudulent

actions by checking authenticity of information provided during online application by users, minimize financial losses for the bank that can incur, and ensure that legitimate applicants are not unfairly denied access to financial services.

3. Approach

Bank Fraud can happen in multiple ways such as loan application, credit card activation, money laundering and many more once the account is activated. Fencing the problem at the registration stage by building a system that can adequately differentiate between duplicitous and non-duplicitous account application will majorly help in fraud prevention.

The team aims to build a reliable model using machine learning algorithms that can be implemented for fair detection of fraud.

4. Conclusions

The goal of this project is to develop and evaluate a machine learning model to detect fraudulent bank account activities using the BAF dataset while ensuring fairness and privacy. the team will make use of the BAF dataset to develop and train machine learning models using various algorithms, considering the controlled types of bias, class imbalance, and temporal nature of the data. The team will evaluate model performance using defined performance metrics and fairness measures and optimize the model for better results.

TOPIC SURVEY IN BRIEF

A high probability of bank account fraud exists if and when a financial institution incorrectly identifies the validity of the information provided by an applicant. A new account fraud is a kind of malpractice that takes place when an applicant with malicious intent strategically manages to approve the account application for a new account. They do so by making use of information from sources such as breached data, phishing messages or any other of the sort. Fraud frequently involves a variety of events or incidents that involve consistent misconduct. Although most fraud cases are not identical, they might seem and sound similar.

Nowadays, rule-based systems are the main tool used by the banking sector to address these fraudulent practises. Rule based systems rely on identifying a known fraud pattern. They use traditional methods of correlation, statistics for their solutions which require complex and time – consuming investigations. It is limited to a certain set of rules and may miss out certain activities that might be fraudulent as the way fraudsters commit fraud is not always in the same patterns.

Several financial institutions are seeking for an automated body of work that might help reduce the challenges caused by the fraud and have been interested in the developments in artificial intelligence and machine learning systems and their performances. The fraud schemes are becoming increasingly complex, however a well-trained model might just be the solution, that could quickly and efficiently help tackle these activities.

CRITICAL ASSESSMENT OF TOPIC SURVEY

Machine learning algorithms can significantly contribute to the development of a system that might aid in problem solving by helping to build a reliable model that might yield favourable outcomes. Based on the data that is currently available, the features and attributes can be investigated, trained, and tested. The right methods and the

relevant data will be used to build a model whose performance can be assessed and validated against the demands of the issue.

Once a system built with a model that showcases an efficient performance is established, the institution and the clients will both considerably benefit from taking action to avert fraud at the registration stage itself. The financial institution can improve customer interactions, offer better services, and guard against fraud-related losses. On the other side, customers will receive improved services with the certainty of better protection for their accounts, financial transactions, and personal data.

METHODOLOGY

Data understanding-

The dataset contains of 10 lakh rows and 32 columns. It has no missing values. There are 5 categorical columns and 27 numerical columns. The target variable has imbalanced data (fraud_bool).

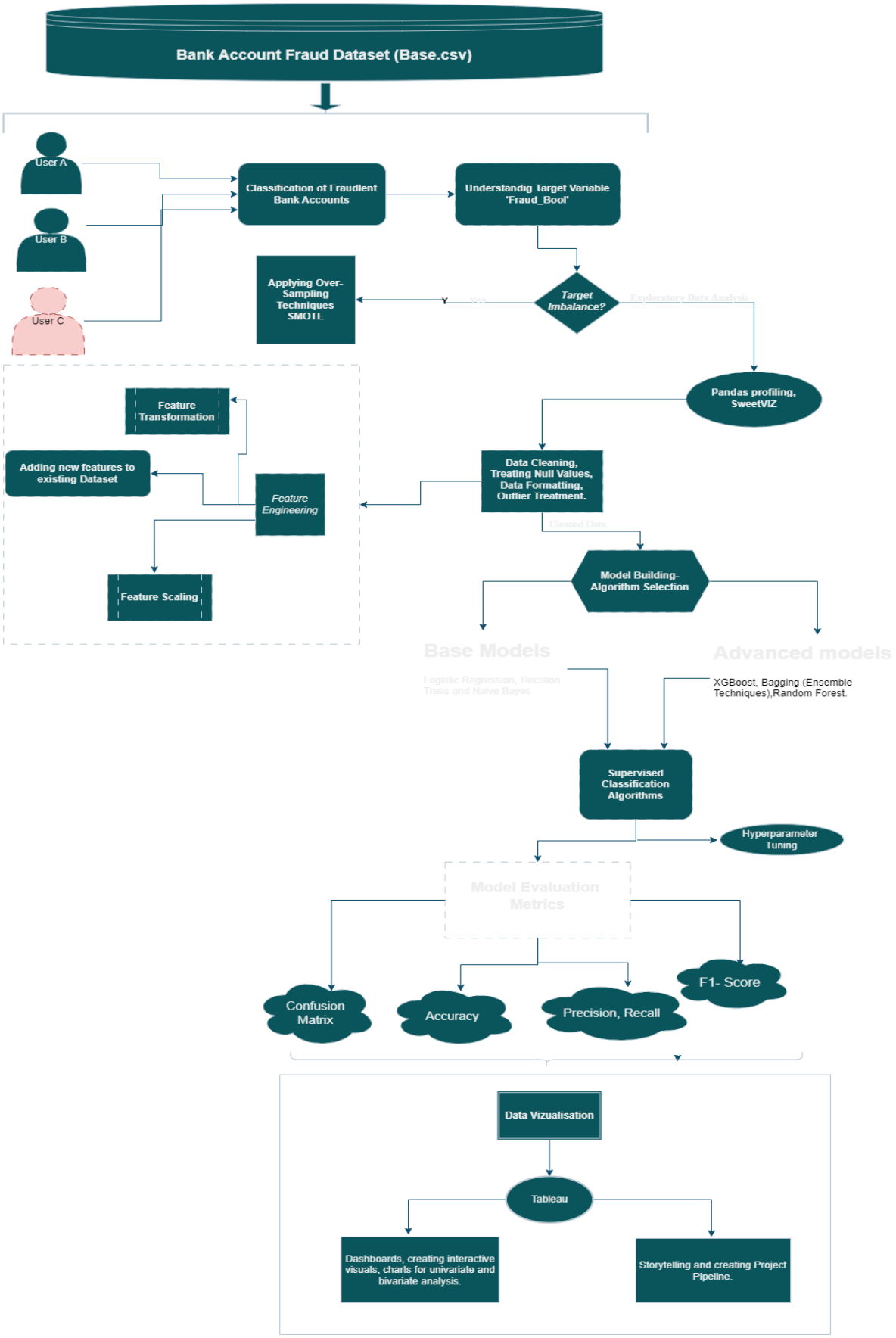
The information regarding each column is as below -

Fraud_bool	Represented in binary values of 0 and 1. 0 displaying a negative outcome showcasing there is no fraudulent activities in the application. 1 showcases a affirmation of fraudulent activities being present.
Income	The annual income of applicants is shown in a range of 0.1 to. 0.9. 0.1 being the lowest and 0.9 being the highest class of income.
Name_email_similarity	Displays the similarity between the name of the applicants and their email address. The values are displayed in numeric format ranging between 0 to 1. Higher values indicating higher similarity.

Prev_address_months_count	The values in this column represent the count of months of previous residency of the applicant from the day of application. The range goes from -1 to 383.
Current_address_month_count	Shows the count of months for current address residency of the applicant. The range goes from -1 to 428.
Customer_age	Ages of the customers in decades form ranging from 10 to 90.
Days_since_request	number of days since the application request. Ranges from 4 to 79.
Intended_balcon_amount	Initial amount transferred for application of accounts. The range is between -15 to 113.
Payment_type	The type of credit payment plan utilized by the applicant. There are 5 anonymized types of values.
Zip_count_4w	The count of applications made in the same zip code. Ranges from 1 to 6700.
Velocity_6h	average number of applications made per hour in last 6 hours. Ranges between -17 to 16716.
Velocity_24h	average number of applications made per hour in 24 hrs. ranges between 1300 to 9507.
Velocity_4w	average number of applications made per hour in the last 4 weeks. The range lies between 2825 to 6995.
Bank_branch_count_8w	count of applications in selected branch of the bank in 8 weeks. Ranges between 0 to 2385.
Date_of_birth_distinct_emails_4w	count of applicants with the same date of birth who have applied in the last 4 weeks. The range goes from 0 to 39.

Employment_status	Employment status of the applicants represented in 7 different pseudonymous values.
Credit_risk_score	The internal score of credit risk during application that is set up with respect to the standards and regulations of the bank. Ranges between -170 to 389.
Email_is_free	showcased in Boolean values of 0 and 1. 0 being free domain of application email and 1 being paid domain of application email.
Housing_status	The housing status of applicants, for privacy reasons have been anonymized. There are 7 categories.
Phone_home_valid	validity of the phone number being provided by the applicant in binary values 0 and 1. 0 representing invalid and 1 representing valid.
Phone_mobile_valid	validity of the mobile number provided by the applicant in binary values 0 and 1. 0 being invalid and 1 being valid.
Bank_months_count	The count of months of the previous account held by the applicant if any.
Has_other_cards	Holding card(s) from the same company, represented in binary values 0 and 1. 0 showcasing applicant not holding any other card, 1 being positive affirmation of a card being held by the user.

Project Flow-


Bank Account Fraud Dataset (Base.csv)
User A
User B
User C
Classification of Fraudlent Bank Accounts
Understandig Target Variable 'Fraud_Bool'
Target Imbalance?
Applying Over-Sampling Techniques SMOTE
Exploratory Data Analysis
Pandas profiling, SweetVIZ
Data Cleaning, Treating Null Values, Data Formatting, Outlier Treatment.
Cleaned Data
Feature Transformation
Adding new features to existing Dataset
Feature Engineering
Feature Scaling
Model Building- Algorithm Selection
Base Models
Logistic Regression, Decision Tress and Naive Bayes
Advanced models
XGBoost, Bagging (Ensemble Techniques),Random Forest.
Supervised Classification Algorithms
Hyperparameter Tuning
Model Evaluation Metrics
Confusion Matrix
Accuracy
Precision, Recall
F1- Score
Data Vizualisation
Tableau
Dashboards, creating interactive visuals, charts for univariate and bivariate analysis.
Storytelling and creating Project Pipeline.

Exploratory Data Analysis -

-Our methodology will be to create and end-to end Supervised Learning problem -Classification algorithm to classify whether a person is a fraud or legit. With machine learning modeling, a system can predict patterns of previously fraudulent and suspicious activities and categorize into potential and Risky/ Fraudsters

-We must define which consumer behaviors are acceptable and which are questionable. The system can learn how to recognize unsafe user conduct thanks to the knowledge about anomalous financial transactions. The user's name, location, payment options, income, bank branch, and other details could all be included in the patterns.

-Financial companies are using artificially generated data points that mimic the under-represented class as one strategy to overcome the problem of imbalanced data (oversampling). SMOTE is a well-liked method (Synthetic Minority Over-sampling Technique).

Abnormal patterns selection-

We need to set what customer behaviors are good and what are suspicious. The abnormal financial transaction information allows the system to learn how to detect risky user activity. The patterns may include the user's identity, location, payment methods, Bank details, income, and other characteristics.

Feature Engineering-

-Transformations: The feature engineering transformation process involves modifying the predictor variable to raise the model's precision and effectiveness. By ensuring that all the variables are on the same scale and that the model is flexible enough to accept input from a range of data, for instance, it makes the model simpler to comprehend. In order to prevent any computational error, it increases the model's correctness and makes sure that all of the features are within the permitted range.

- Feature Extraction: By extracting additional variables from the raw data, feature extraction is an automated feature engineering method. This step's major goal is to decrease the amount of data so that it can be used and managed for data modeling more simply. We will extract important features and use them for model building.

Training an algorithm-

Provide the guidelines to teach your algorithm to distinguish between legitimate and fraudulent user activity using classification techniques like Logistic Regression, Decision Trees and Naive Bayes -- For creating a base model. The most popular approach to machine learning-based fraud detection in fintech is supervised learning. In this method, material that has been classified as excellent or bad will be used to understand previously followed patterns. It indicates that the relevant responses have already been assigned to data components.

-This model analyses predictive data, and the precision of the training data determines the accuracy of the model. A common technique for assessing the strength of cause and effect links between variables in data sets is logistic regression. It can be used to develop an algorithm that determines whether or not a transaction is "good."

Using examples of fraud, Decision tree may be taught to recognize abnormalities and can be used to develop a set of rules that simulates consumers' typical behavior. Decision tree ensembles can be used to create Random forest, which combine several weak classifiers into a single powerful classifier.

Building a model-

After training, we will have the ML model ready to detect fraud. We will later evaluate model performance using evaluation metrics – Recall, Precision etc.

Data Visualization-

Data visualization in the form of dashboards is the preferred method for many firms to evaluate and communicate information with the aim of making data more accessible and intelligible.

LIMITATIONS

Dataset Limitations-

This project is based on the BAF of dataset, which are synthetic datasets designed to simulate real-world bank account fraud. While the dataset is realistic and diverse, it may not capture all the nuances and complexities of actual fraud patterns. Therefore, the performance of the machine learning model developed in this project may not be generalizable to all types of fraud scenarios. The data also displays a highly skewed pattern in the target variable showcasing an imbalance. The data has been anonymized for certain features due to privacy reasons.

Limitations of Machine Learning Models-

Machine learning models are not perfect and can only make predictions based on the patterns present in the data. Therefore, the performance of the machine learning model developed in this project will depend on the quality and diversity of the data used to train it.

ASSUMPTIONS

Assumption of Data Privacy-

This project assumes that the privacy-preserving techniques used to protect the identity of potential applicants in the BAF dataset are effective. However, there is always a risk of re-identification or information leakage in any privacy-preserving technique, which may compromise the privacy of individuals in the dataset.

Assumption of Fairness-

This project assumes that the fairness metrics used to evaluate the machine learning model are appropriate and unbiased. However, fairness is a subjective concept, and different fairness metrics may lead to different results.

REFERENCES

The reference material-

a) https://arvidhoffmann.nl/Hoffmann_Birnbrich_2012.pdf

b) https://simility.com/blog/4-ways-fraudulent-accounts-affect-the-banking-industry/

c) https://nexocode.com/blog/posts/ai-based-fraud-detection-in-banking-and-fintech-use-cases-and-benefits/#:~:text=At%20the%20same%20time%2C%20advanced,the%20probability%20of%20such%20tendencies

d) https://jair.org/index.php/jair/article/view/10302

e) https://www.netguru.com/blog/fraud-detection-with-machine-learning-banking

Project Timeline

23/2/23	domain finalisation
26/2/23	project flow planning and synopsis.
5/3/23	data preparation/EDA
12/3/23	base model + model building
15/3/23	additional model building
18/3/23	interim submission
20/3/23	PowerPoint presentation
26/3/23	final model building & testing
27 March to 5 April	exam break
9/3/23	Testing and evaluation of model
12/3/23	final documentation
15/3/23	final presentation

Original owner of data	Sérgio Jesus (Owner) Jose Pombal (Editor) Pedro Saleiro (Editor)
Data set information	https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022?select=Base.csv
Any past relevant articles using the dataset	
Reference	https://github.com/feedzai/bank-account-fraud/blob/main/documents/datasheet.pdf
Link to web page	www.kaggle.com

****