**A Project Report**

***on***

**An Implementation of XGBoost for Insurance Fraud Detection**

***in partial fulfillment***

***for the award of the degree of MSc-CA***

***Submitted by***

**Nachiket Chaudhari**

**22030142004**

**AFFILIATED TO SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY) (INDIA)**

**MSc-CA Batch 2022-2024**

Date: 10/04/2024

# COMPLETION CERTIFICATE

This is to certify that the project report titled “**An Implementation of XGBoost for Insurance Fraud Detection”** is the bonafide work of **Nachiket Chaudhari** PRN - 22030142004, MSc-CA Batch 2022-2024 at Symbiosis Institute of Computer Studies and Research (SICSR) who carried out the project work under my supervision. He has completed the project during

17 December 2023 to 10 April 2024

**Signature of the Mentor -**

**Name of the Mentor** – Prof Shirish Joshi

**Department Name and City –** SICSR, Pune

**Seal of the Company -**

**Acknowledgement:**

I am pleased to present our Report 'XGBoost for Insurance Fraud Detection' and take this opportunity to express my profound gratitude to all those individuals who assisted in the completion of this Report.

I extend my heartfelt appreciation to our college for providing us with excellent facilities that greatly contributed to the successful completion and presentation of this Report. Additionally, I express my sincere gratitude to my project Mentor, Prof. Shirish Joshi, whose guidance and expertise were invaluable in navigating various aspects and solutions related to the topic.

I am deeply grateful to Prof. Shirish Joshi for her continuous support and timely advice throughout the different phases of our Report. I also acknowledge her for providing us with the necessary facilities and serving as the project coordinator. I appreciate her support, patience, and confidence in our abilities, as well as for granting us flexibility in our working and reporting schedules.

Lastly, I would like to express my gratitude to everyone who contributed directly or indirectly to our report's completion.

**Abstract**

Fraudulent insurance claims constitute a huge economic burden, which consumes resources and leads to increased premiums on genuine policyholders. With the ever-changing ways fraudsters apply for insurance policies, it is difficult for traditional detection systems that often involve manual examination to keep up with them. The paper investigates how advanced analytics mainly machine learning and artificial intelligence (AI) can help in improving the detection of insurance fraud.   
Insurance fraud poses several hurdles that include the vast amount of data being generated and difficulty in anomaly identification. This study also addresses how these patterns can be extracted by using machine learning algorithms and identify certain characteristics associated with suspicious claims. Other techniques such as anomaly detection, supervised learning, unsupervised learning are covered in this article stressing on their strengths and limitations in this sphere.

Additionally, the abstraction will focus on the need for better data quality and feature engineering to generate better models, and acknowledge the limitations of the rules of AI-assisted fraud supported ethical and legal impediments which should be considered for this model to incorporate the acceptance. They explain how such detailed analysis can automate this analysis, reducing the human effort and cost associated with fraudulent claims Finally they present the benefits of using AI in effort created to detect insurance fraud is highlighted and concludes, and future opportunities for this technology are discussed

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Chapter Number** | **Title and sub-titles** | **Page Number** |
| 1 | Introduction   * 1. Objectives   2. Problem definition   3. Project Overview   4. Hardware Specifications   5. Software Specifications | 6  6  7  7  10  10 |
| 2 | System Design   * 1. Dataset   2. Flowchart   3. System Architecture   4. Algorithm Flow   5. Working and Methodology | 11  11  11  12  14  15 |
| 3 | System Implementation   * 1. Front-End Screenshot   2. Backend Screenshot   3. Improving Accuracy Screenshot   4. Testing Process | 16  16  17  17  18 |
| 4 | Conclusion | 19 |
| 5 | References | 20 |

1. **Introduction**

Today, it is acknowledged that insurance fraud is still a serious problem, as there are people who try to trick the insurance company with false claims or changes that not only result in financial loss but also affect the credibility of the insurance policy and by justice. As technology advances, so do the methods used by fraudsters; This means that traditional fraud detection methods no longer work. Consequently, sophisticated tools need to be developed in response to the significant number of these crimes that occur frequently. The purpose of this paper is to provide an overview of what insurance fraud is and to propose effective detection methods. Among the issues we will be discussing are the nature of fraud, what motivates the act of fraud, the challenges of catching fraud offenders and more, we are going to discuss the emerging trends that technological advances have had on fraud prevention efforts of positive influence. On the other hand, it is important that these methods are used ethically and within the guidelines for their use.

* 1. **Objectivities**

The goals of insurance fraud detection programs usually interlinked. First, the financial security of the company is based on measures preventing fraudulent claims and eliminating illegal payments. As a result, it enables better results in handling claims. Early detection of potential fraud allows users to concentrate their efforts on verifying proper claims process for honest clients. In addition, effective fraud detection ensures that premiums are fair for all policyholders. If frauds are uncontrollable, it can cause high premiums overall. A strong system solution increases the possibility of catching and thus catches any person from committing fraud, which in turn lowers the overall incentives for fraudulent activities. On other hand, dealing with fraud continuously is very critical since it guards a company against such harmful events or activities and can ruin its reputation and standard.

* 1. **Problem Definition**

One of the most complicated yet important parts in the insurance industry is fraudulent detection. Insurance fraud is deceitful actions intended to gain illicit advantages from insurance policies. Hence, it calls for well-established approaches that could identify claims made in bad faith for effective coverage fraud discovery and remediation. Even though few insurers have realized significant achievements while trying to implement systems for identifying fraud, these criminals do not sleep at anytime but rather constantly find new ways of exploiting vulnerabilities and escaping detection. On the other hand, relating it with fraud, more credit card transactions and claims from insurance bodies means that there is indeed need to introduce anti-fraud measures through continuous research and oversight as this has major importance. At its heart, however, these problems can be solved by incorporating all stakeholders, leveraging modern technological solutions such as AI and ML technology adoption which processes large quantities of data and following-up every single wave with insights obtained from datasets compiled during different points in time. Society must be able to engage in preventive action at an early stage and fully understand all factors leading up or causing possible consequences before any harm takes place.

**1.3 Project Overview**

The goal of our project is to broaden a sturdy Insurance Fraud Detection System (IFDS) to address the prevalence of fraudulent pastime, specially within the insurance zone. In doing so, it makes use of ultra-modern equipment which include gadget getting to know algorithms, statistical analysis and records mining techniques. IFDS does this with the aid of reading a wealth of coverage information together with regulations, claims, and beyond fraud instances to look for uncommon styles or suspicious conduct that may imply capacity fraud, further to externalities integration of information assets to further decorate the fraud detection machine, increasing its predictive functionality. Through real-time analysis and alert capabilities, dealers and investigators can make sure integrity in the insurance enterprise, construct believe with key stakeholders and respond quick to suspected fraud round.

1. **Data Source**

**Our supply dataset is to include honest and dubious Insurance Fraud Detection. If it takes place that manner, the skewed information approach will need to be taken into consideration for heading off model bias toward the majority class.**

**2. Data Preprocessing**

Before building a building, you first lay a solid foundation. The information cleansing manner is the start of this meticulous system. Like an demanding tiny stone for your shoe, any moderate inconsistencies or lacking portions from our dataset can regulate consequences. This means that we should ensure our records is smooth without such anomalies. At this factor we begin characteristic engineering which entails remodeling present variables into significant combos designed to highlight capability fraud sports. Think of placing collectively proof like a detective could: something like this will be suspicious if it were simply one transaction after every other with no pause for breath. But now not all evidence or capabilities in records shout on the identical pitch. Some often overpower others due to their magnitudes handiest; this is in which normalizing of data will become vital because it create stability and guarantees each shred speaks similarly well approximately its issuer.

Lastly, we apprehend that actual transactions regularly a ways outnumber fraudulent ones in normal datasets. This imbalance could make our model develop a blind spot for fraud. To counteract this, we hired methods inclusive of the Support Vector Machine, Logistic Regression, and Random Forest to generate more records points skilfully, ensuring our version efficiently distinguishes among authentic and fraudulent movements.

**3. Model Selection and Training**

**In our pursuit to create a whole device for detecting insurance fraud, we have systematically skilled and developed numerous system getting to know models which includes help vector machines (SVM), random forests, logistic regressions, Naive Bayes classifiers, MLPs and XGBoost. Each emerge as selected mainly based totally on its fit for use in the insurance location towards fraud detection. Logistic Regression was selected inside the most important because it is straightforward and interpretable which is ideal for initial exploratory analysis. For Random Forests and SVMs, they had been preferred because they could version complicated non-linear relationships with information which makes them robust in class duties. On the opposite hand, Naïve Bayes provided a lightweight conditional opportunity model with appreciate to processing large datasets having specific capabilities as it employs a probabilistic approach. Multilayer Perceptron represented an interesting possibility of a neural community-based totally method able to assimilating complicated styles from excessive-dimensional statistics. Significantly, XGBoost got here out pinnacle among all models examined by the usage of attaining highest take a look at accuracy in evaluation to other algorithms that have been skilled. This may be majorly attributed to its ensemble’s form that involves incorporating gradient boosting strengths at the equal time as improving regularization techniques sooner or later giving better predictive Performance and ability to come across the fraud.**

**4. Evaluation Metrics**

**Precision:** High Precision shows that model is detecting very few genuine transactions as a fraudulent.

**Recall:** A High Recall shows that we are actually capturing almost all correct insurance details genuinely.

**F1-Score:** The comparison between precision and recall gets less valuable when there is a difference in the fraud associated with false positives and false negatives.

**AUC-ROC:**  A High AUC Indicates that model is performing well.

**6. Model Deployment:** Deploying our model through flask web applications. So, that all users can access our fraud detection model.

**1.4 Hardware Requirements:**

The hardware is an Intel core i5 processor with 16GB ram and a solid-state drive with 25GB of free space.

**1.5 Software Requirements:**

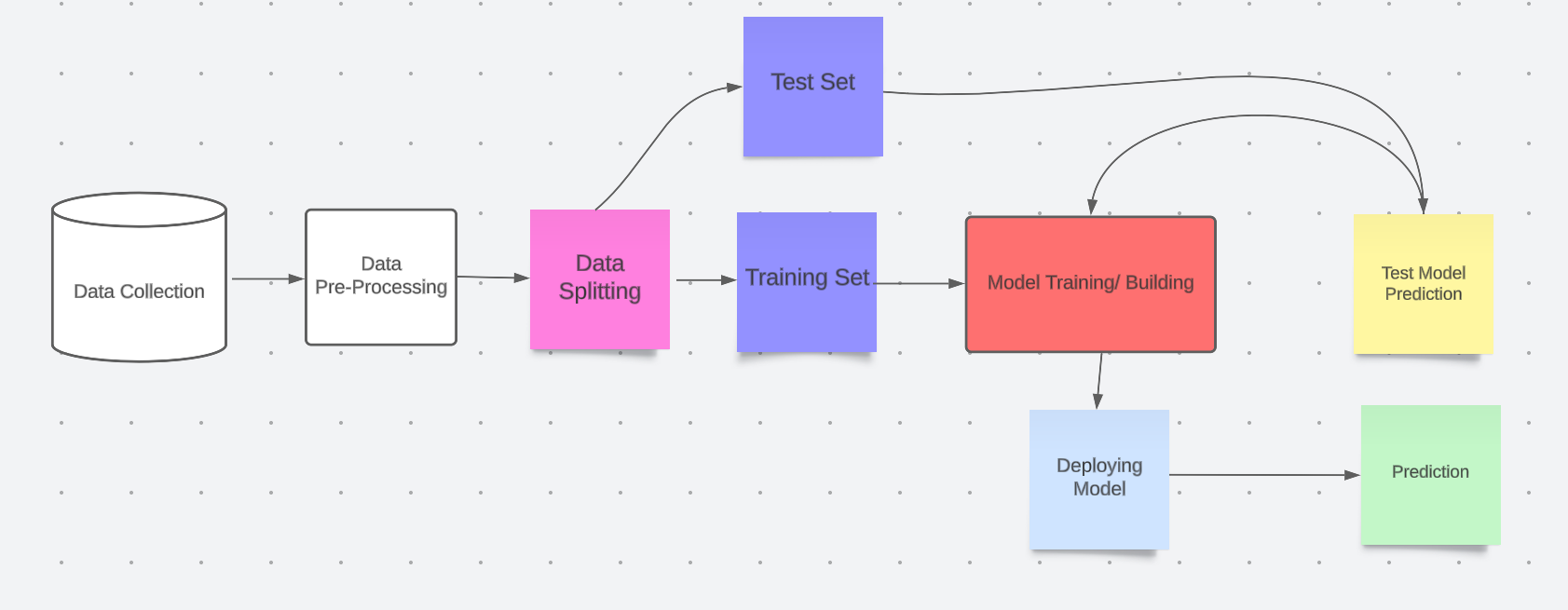
The Operating system is Windows 11 with a 21H2 update. The machine is running Python 3.10 in an Anaconda environment. The Python libraries sci-kit learn, XGBoost and matplotlib is used in this project.

1. **SYSTEM DESIGN**

**2.1 Dataset**

Various details make up the dataset for insurance fraud detection. These include policy information, claim amounts and descriptions of claims. It consists of both structured data in terms of numbers and categories as well as unstructured data in form of written descriptions. In order to effectively train systems for identifying fraud, the dataset incorporates instances of known fraudulent cases. Such instances help in building accurate fraud detection algorithms by detecting patterns from previous fraudulent activities. To adapt to new fraud tactics, it is essential to maintain an updated and accurate dataset. The data has to be cleaned and organized so that it can be used constructively for building models for detecting insurance frauds.

**2.2 Insurance Fraud Detection Flowchart**



**2.3 System Architecture**

**1. Data Collection Layer:**

Our system revolves around the Data Collection Layer, which integrates two major data sources. The first is the Transaction Data Source – the mixture of historical and real-time transactional data with a focus that presents finance operations with context. The second, the External Data Source, enhances the workings of the first source by incorporating external data on, for instance, blacklisted IP addresses, known fraudulent activities, etc. This approach allows the system to develop a unified picture of both internal transactions and external people.

**2. Data Processing and Storage Layer:**

Afterwards, the Data Processing & Storage Layer receives data. This is where the Data Preprocessing Module plays a critical role in ensuring data quality through cleaning, normalization of values, feature engineering and solving class imbalances to enable more analysis. In case their potential length poses any risk, efficient storage must be a management’s priority.

**3. Machine Learning Layer:**

The Machine Learning Layer takes in processed data from the prior stage as well as trains predictive models with updated ones over time. The Training Module continually updates models and thus ensures that systems remain “live” whereas Evaluation Module check for accuracy of these models against several metrics to guarantee predefined performance. Lastly, the Model Repository retains trained.

**4. Prediction & Decision Layer:**

The Prediction Layer serves as an interface between real-time transactions and actual system operation. With this, fraudulent activities can be detected timely since it undertakes continuous monitoring of ongoing dealings using its Real-time Prediction Module. Furthermore, it responds accordingly when risk arises such as raising alarms for major risks, notifying users/ administrators about the system and suspending that transaction if required

**5. Model Interpretation:**

How predictions are made by models should be appreciated. Understanding model behavior that helps during interpretation and decision-making process can be done using such technologies as importance analysis, partial dependence plots or model-specific visualization tools.

**6. Model Deployment:**

Deploying a model into production for real-time prediction of the model.

**7. Monitoring and Maintenance:**

To demonstrate the significance of persevered supervision of fashions in real-international putting, that is vital. In order to stay correct, as soon as clean statistics is supplied, fashions must be retrained or best-tuned hence. Continued overall performance tracking for idea go with the flow and updating of fashions are vital.

**8. Iterative Improvement:**

Machine gaining knowledge of involves a repetitive system. This method that there may be consistent room for improvement in phrases of accuracy, performance and general overall performance with time by way of learning from the performance of deployed fashions and user remarks

**2.4 Algorithm**

1. **Initiate**

This is the 1st stage since this is where the algorithm has to be initiated.

1. **Input real-time Insurance data.**

It is done by the system in real-time where the transaction data gathered instant exact when they are generated. Some instances of it comprise of; Months as Customer, Capital Loss, Incident Severity and etc.

1. **Data Pre-processing and Normalization**

The raw data should be cleaned and formatted correctly for it to be analysed, or in other words, it means getting rid of any errors or irrelevant data as well as ensuring that all records are in a consistent format which is quite conducive to accurate fraud analysis.

1. **Use the predictive model to indicate fraud:**

The data is now taken through a pre-trained machine learning model or any other algorithm and put through the suspicious and fraud looking Insurance data to compare it with what it has learned they pattern to be, or actual.

1. **Raise an alarm if you suspect fraud**

Once some Insurance data is determined to be a scam, the system raises the red flag. The red flag can be used to inform the Reporting analyst and administrators that. It is fraudulent. Which is then

**6. Put up warnings on UI for analyst/administrator checking:**

Then put up all the flags on one UI for human analyst or administrators. They will then be able to check through if it is actually

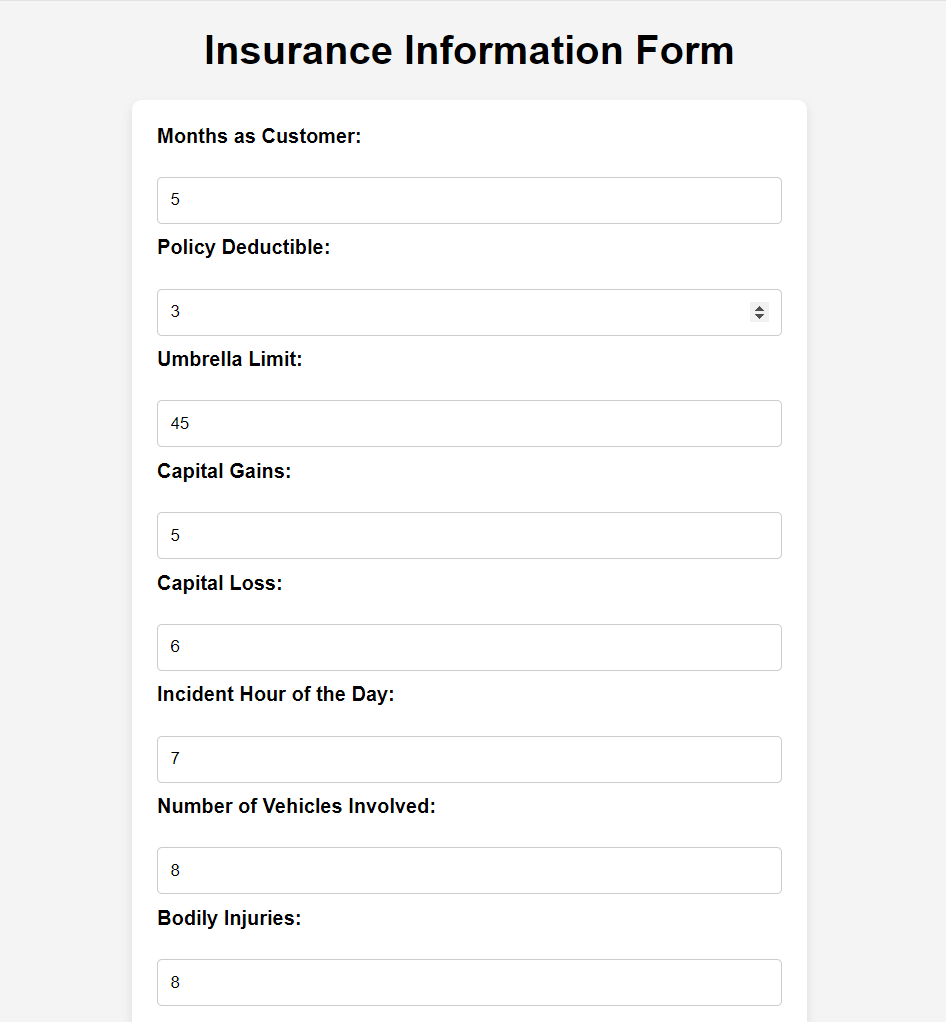
This shows the end of the algorithm and as such the process has been completed for that set of transactions in particular.

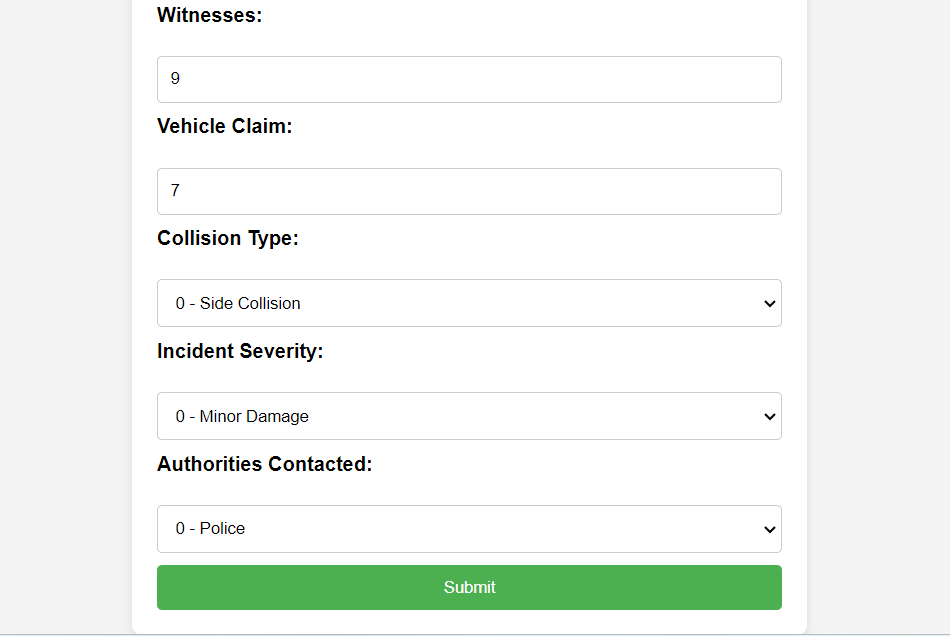
**7.End**

* 1. **Working and Methodology**
* In a list, for example, we’ll include insurance information: policy number, type of accident, person to call the police, witnesses.
* Before it is analysed this information has been preprocessed to remove outliers.
* For example, these are: Policy Number; Umbrella Limit; Authorities Contacted; Incident Type.
* Then a machine learning algorithm is trained that incorporates these features: the data includes non-fraudulent as well as fraudulent insurance data.
* Here is the machine learning algorithm, which has scored the Insurance Data to predict the probability that this occurrence is suspicious.
* Machine learning algorithm provides such score for Insurance Fraud as an Admission Criteria. Decision for insurance application may be approved, rejected or put in a queue for a manual review depending on the situation.

1. **System Implementation:**

**3.1 Front End Screenshots**

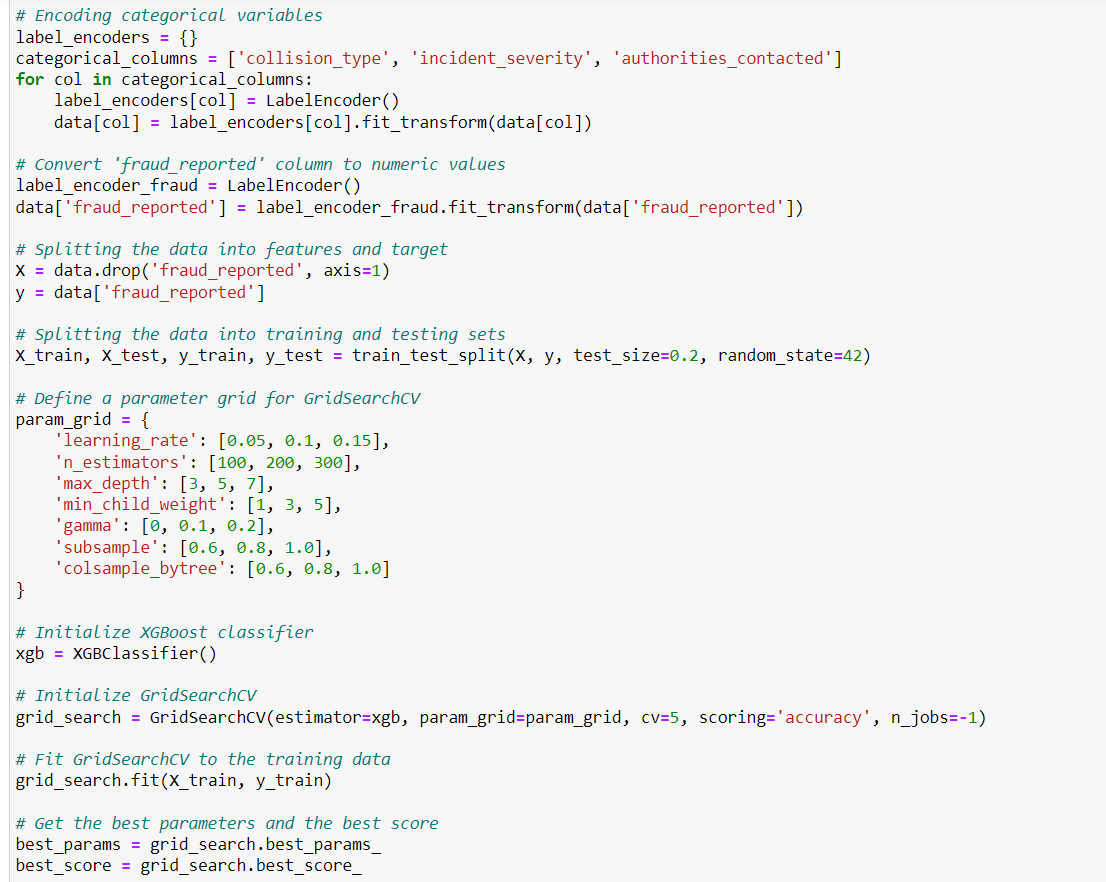
****

****

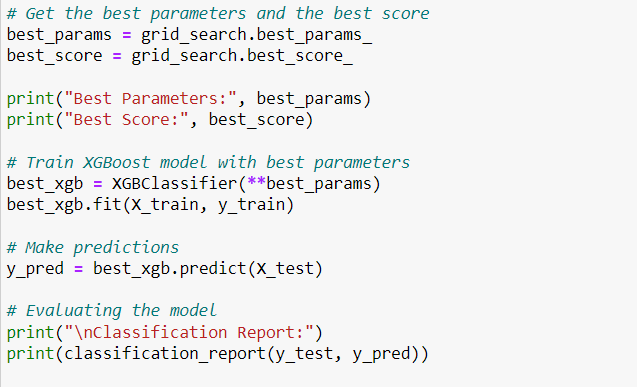
* 1. **Back End Screenshots**

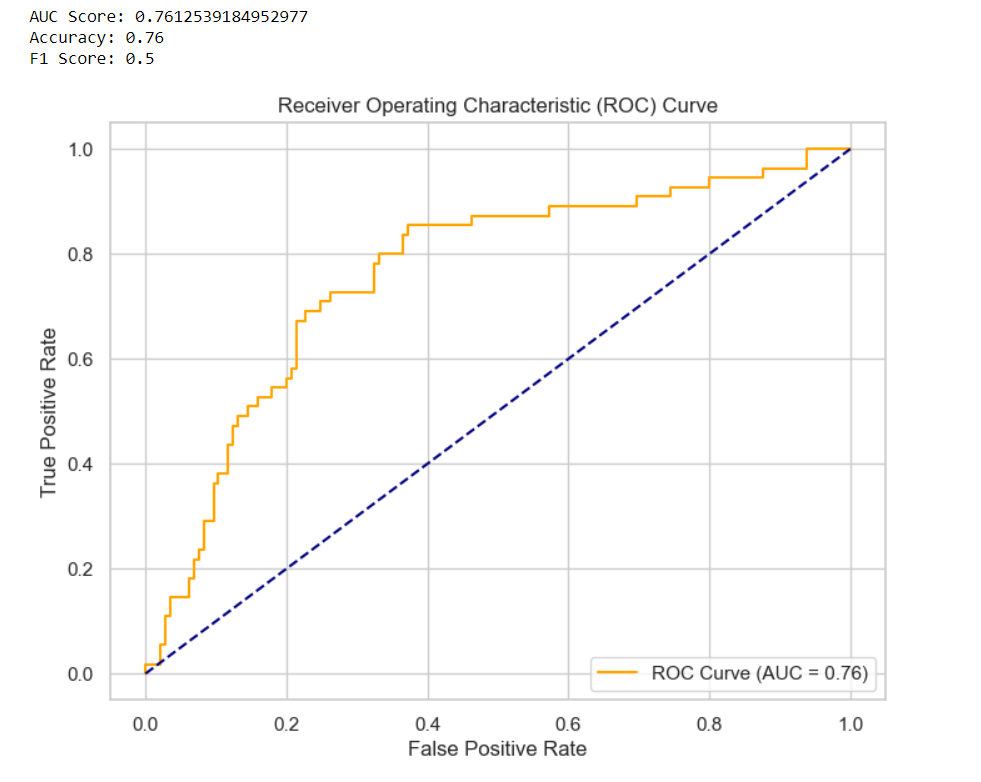
****

* 1. **Improving the Machine Learning Algorithm**

****

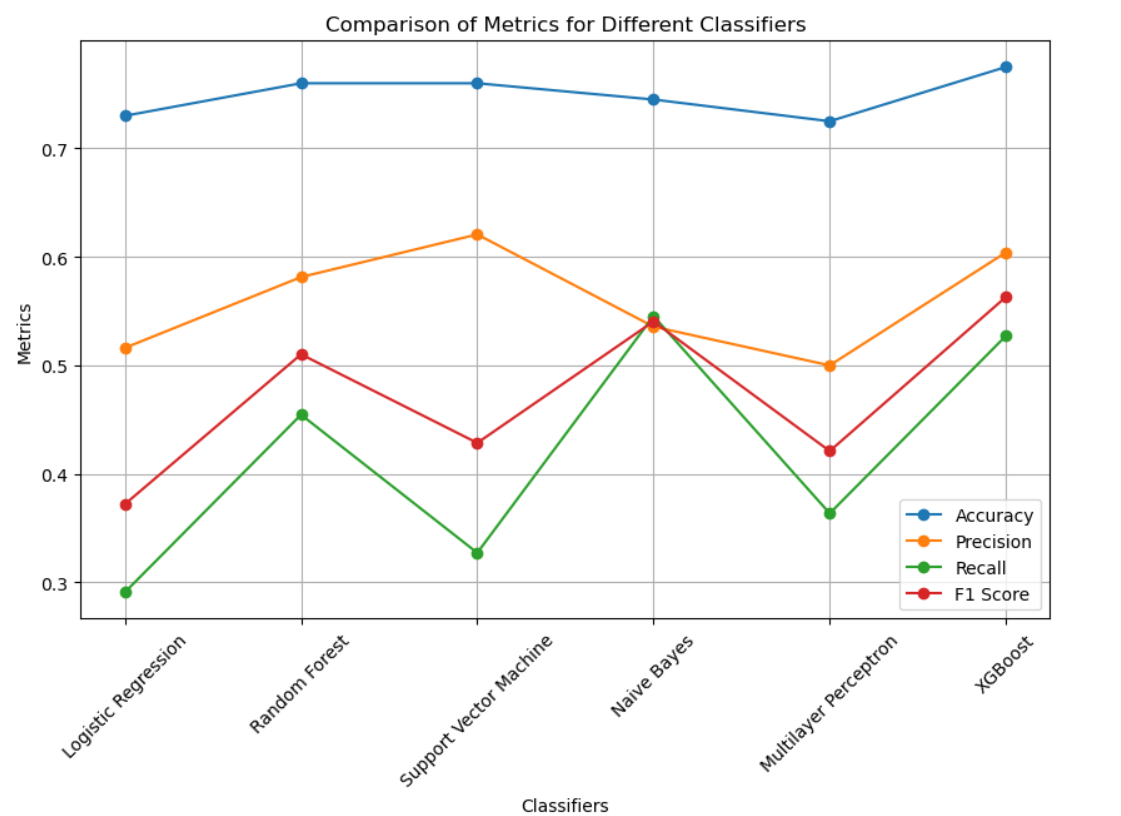
**3.4 Testing Process**

****



1. **CONCLUSION**

The fact that it combines many good features is what makes XGBoost outperform other techniques like Random Forest, Multilayer Perceptron (MLP), Support Vector Machines (SVM), Logistic Regression, and Naïve Bayes. This is achieved by the XGBoost ensemble learning approach which makes it be characterized by an average accurate predictability. This flexible working with non-linear data relationships and adoption of strong regularization methods in XGBoost helps to avoid overfitting which commonly occurs in machine learning applications. The conclusion made based on this new column subsampling technique and parallelization for efficient processing of large datasets can’t find a better algorithm in terms of efficiency while handling big data. Its hyperparameter tuning has also made xgboost adaptive across various types of datasets depending on a particular task requirement. So, one way to improve the overall classification performance is dynamically adjust-ing xgboost to optimize for minority classes. Additionally, xgboost has feature importance metrics that make results more interpretable and provide insights into underlying data patterns allowing stakeholders to take better decisions from them. As a result, all these factors are the ones that contribute towards making xgboost the most used algorithm among many machine



1. **References**

**[1]** Sithic, H. L., & Balasubramanian, T. (2013). Survey of insurance fraud detection using data mining techniques. *arXiv preprint arXiv:1309.0806*.

**[2]** Gomes, C., Jin, Z., & Yang, H. (2021). Insurance fraud detection with unsupervised deep learning. *Journal of Risk and Insurance*, *88*(3), 591-624.

**[3]** Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications*, *68*, 90-113.

**[4]** Schiller, J. (2006). The impact of insurance fraud detection systems. *Journal of Risk and Insurance*, *73*(3), 421-438.

**[5]** Gomes, C., Jin, Z., & Yang, H. (2021). Insurance fraud detection with unsupervised deep learning. *Journal of Risk and Insurance*, *88*(3), 591-624.

**[6]** Aslam, F., Hunjra, A. I., Ftiti, Z., Louhichi, W., & Shams, T. (2022). Insurance fraud detection: Evidence from artificial intelligence and machine learning. *Research in International Business and Finance*, *62*, 101744.

**[7]** Roy, R., & George, K. T. (2017, April). Detecting insurance claims fraud using machine learning techniques. In *2017 international conference on circuit, power and computing technologies (ICCPCT)* (pp. 1-6). IEEE.

**[8]** Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, *105*, 87-95.

**[9]** Hancock, J., & Khoshgoftaar, T. M. (2020, December). Performance of catboost and xgboost in medicare fraud detection. In *2020 19th IEEE international conference on machine learning and applications (ICMLA)* (pp. 572-579). IEEE.

**[10]** Averro, N. T., Murfi, H., & Ardaneswari, G. (2023). The Imbalance Data Handling of XGBoost in Insurance Fraud Detection. In *DATA* (pp. 460-467).

**[11]** Hajek, P., Abedin, M. Z., & Sivarajah, U. (2023). Fraud detection in mobile payment systems using an XGBoost-based framework. *Information Systems Frontiers*, *25*(5), 1985-2003.

**[12]** Zhang, Y., Tong, J., Wang, Z., & Gao, F. (2020, March). Customer transaction fraud detection using xgboost model. In *2020 International Conference on Computer Engineering and Application (ICCEA)* (pp. 554-558). IEEE.

**[13]** Mubarek, A. M., & Adalı, E. (2017, October). Multilayer perceptron neural network technique for fraud detection. In *2017 International Conference on Computer Science and Engineering (UBMK)* (pp. 383-387). IEEE.

**[14]** Rukhsar, L., Bangyal, W. H., Nisar, K., & Nisar, S. (2022). Prediction of insurance fraud detection using machine learning algorithms. *Mehran University Research Journal of Engineering & Technology*, *41*(1), 33-40.

**[15]** Kasasbeh, B., Aldabaybah, B., & Ahmad, H. (2022). Multilayer perceptron artificial neural networks-based model for credit card fraud detection. *Indonesian Journal of Electrical Engineering and Computer Science*, *26*(1), 362-373.

**[16] `**Rukhsar, L., Bangyal, W. H., Nisar, K., & Nisar, S. (2022). Prediction of insurance fraud detection using machine learning algorithms. *Mehran University Research Journal of Engineering & Technology*, *41*(1), 33-40.

**[17]** Li, Y., Yan, C., Liu, W., & Li, M. (2016, August). Research and application of random forest model in mining automobile insurance fraud. In *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)* (pp. 1756-1761). IEEE.

**[18]** Liu, C., Chan, Y., Alam Kazmi, S. H., & Fu, H. (2015). Financial fraud detection model: Based on random forest. *International journal of economics and finance*, *7*(7).

**[19]** Ghuse, N., Pawar, P., & Potgantwar, A. (2017). An improved approch for fraud detection in health insurance using data mining techniques. *no*, *5*, 27-32.

**[20]** Resende, P. A. A., & Drummond, A. C. (2018). A survey of random forest based methods for intrusion detection systems. *ACM Computing Surveys (CSUR)*, *51*(3), 1-36.

Top of Form