## INSURANCE CLAIM FRAUD PREDICTION WEB APP USING MACHINE LEARNING-FLASK WEB FRAMEWORK

## Mohamed Thanish.M1, Sam Aravind.R2, Balaji.S3, Murugavalli.S4

## 1,2,3 Student, Degree (Artificial Intelligence & Data Science), K.Ramakrishnan College of Technology, Trichy

## 4Assistance HOD (Department of Artificial Intelligence), K.Ramakrishnan College of Technology, Trichy

## [mohamedthanish14@gmail.com](mailto:mohamedthanish14@gmail.com),[samaravind.suv@gmail.com](mailto:samaravind.suv@gmail.com),[balajinandhasandy007@gmail.com](mailto:balajinandhasandy007@gmail.com),[murugavallis.ai@krct.ac.in](mailto:murugavallis.ai@krct.ac.in)

## --------------------------------------------------------------------------

## *ABSTRACT:* In recent years fraudulent insurance claims is the problem faced by many of the insurance companies which leads to huge financial loss yearly and insurance fraud has been since the beginning of the insurance organization. When a person makes a false claim in order to get benefits to which they are not entitled is known as an insurance fraud. So detection of an insurance fraud is a challenging problem for the insurance industry.so we chose to do a project based on the insurance claims fraud prediction using machine learning. These frauds have adverse consequences on society as the losses are settled down by increasing the premium cost of policy holders. Also the traditional claim investigation process being time consuming and tedious that generally leads to inaccurate results has been identified as main culprit. These project are going to identify potential fraud insurance claims and help insurance company to make more secured claim authentication & settlement by early identification of probable fraud claims. Solution of these problem is to Build a classification methodology AI based App to determine whether a customer is placing a fraudulent insurance claim using machine learning(ML) algorithms such as random forest classifier, XGBoost classifier, logistic regression, ensemble methods (namely bagging and boosting) with Support Vector Machines,  K-nearest neighbors and using HTML,FLASK for front end development. The aim is to identify fraud claims accurately within shorter period of time. Throughout the process data analysis is used to validate, clean and extract the relevant data. Hence, by using this framework insurance company can maintain its respectability in outside world and can also share trustworthy relationship with customers.

# INTRODUCTION

## 1.1 INSURANCE CLAIM FRAUD PREDICTION

The insurance market is a highly profitable market that moves large sums of money over the years. In Brazil alone, about 10.8 billion USD was paid in insurance policies in 2017 (Brazilian National Confederation of Insurance Companies, 2017). Similarly, frauds can bring huge losses to the companies: In the same year of 2017, the total value of all occurred claims was around 10.0 billion USD, while the value of proven frauds totaled 221.2 million USD (Brazilian National Confederation of Insurance Companies, 2017).Fraud in insurance is an unethical activity performed systematically to get some financial gain. These fraudulent claims present overpriced and large problem for insurance company leading to billions of dollars of needless expenses per year. Also due to some flaws in traditional process most of the companies are in search of some new technique to find fraud claims. So here we propose ML based automated framework employed with XGBoost algo to classify claims. We also compare the performance of XGBoost algo with other algorithms to obtain most accurate results.

## 1.2 PROBLEM STATEMENT

The goal of this project is to build a model that can detect auto insurance fraud. The challenge behind fraud detection in machine learning is that frauds are far less common as compared to legit insurance claims.

Insurance fraud detection is a challenging problem, given the variety of fraud patterns and relatively small ratio of known frauds in typical samples. While building detection models, the savings from loss prevention needs to be balanced with the cost of false alerts. Machine learning techniques allow for improving predictive accuracy, enabling loss control units to achieve higher coverage with low false positive rates.

**1.3 MOTIVATION AND PURPOSE**

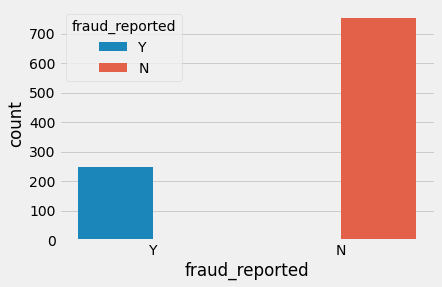
**MOTIVATION:** Early detection of fraud by employees or customers can significantly reduce industry costs and positively impact insurance premiums. By learning collected customer data, you can analyze and predict whether newly filed claims are inappropriate manipulation behavior.

**PURPOSE :** The purpose of this project is to analyze and detect the frauds in insurance claims with the help of data mining techniques like predictive models, rule learning, time series learning etc..,thereby helping insurance companies to savings from loss prevention needs to be balanced with the cost of false alerts.

**1.4 ANALYSIS & SCOPE**

**DATA ANALYSIS:**

In this project, we have a dataset which has the details of the insurance policy along with the customer details. It also has the details of the accident on the basis of which the claims have been made.

**Dependent variable:** Exploratory data analysis was conducted starting with the dependent variable, Fraud\_reported. There were 247 frauds and 753 non-frauds. 24.7% of the data were frauds while 75.3% were non-fraudulent claims.

**FIG 1 : Dependent variable**

**Correlations data analysis:** Heatmap was plotted for variables with at least 0.3 Pearson’s correlation coefficient, including the DV. Month as customer and age had a correlation of 0.92. Probably because drivers buy auto insurance when they own a car and this time measure only increase with age. Apart from that, there don’t seem to be many correlations in the data. There don’t seem to be multicollinearity problems except maybe that all the claims are all correlated, and somehow total claims have accounted for them. However, the other claims provide some granularity that will not otherwise be captured by total claims. Thus, these variables were kept.

**Visualizing varibales:** The value of fraud reported differs across hobbies of the customer seems like chess players and crossfitters have higher tendencies to fraud.

**SCOPE:**

The insurance fraud detection market is expected to register a CAGR of over 17.4% during the forecast period (2021-2026). Fraudulent claims in the insurance industry have steadily grown to be the single largest expense to property and casualty insurers, taking up to 10% of an insurer’s revenue. Accurate prediction gives a chance to reduce financial loss for the company. A major cause of  the increased costs are the payment errors made by the insurance companies while processing claims.

## MACHINE LEARNING

XGBoost Is A Fast And Powerful Machine Learning Algorithm, Which Has Become A Popular Choice For Online Machine Learning Competitions Due To Its High Efficacy. Its Complete Moniker Is Extreme Gradient Boosting. XGBoost Is An Enhanced Implementation Of Gradient Boosting In Terms Of Both Speed And Performance. XGBoost Is A High-Performance Decision-Tree-Based Ensemble Learning Method Usually Used For Supervised Learning. In Ensemble Learning,

## The random forest (RF) classifier is a supervised learning (SL) algorithm which you can use for regression and classification problems. It is among the most popular machine learning algorithms due to its high flexibility and ease of implementation.  it uses randomness to enhance its accuracy and combat overfitting, which can be a huge issue for such a sophisticated algorithm. These algorithms make decision trees based on a random selection of data samples and get predictions from every tree. After that, they select the best viable solution through votes.

## 2.SYSTEM SPECIFICATION

**2.1 H/W SYSTEM CONFIGURATION: -**

Processor – Intel Corei5,GPU

RAM - 4 GB (Min) Or Higher

Hard Disk - 20 GB

Keyboard – Standard Keyboard

Monitor – 20 Inch Color Monitor

## 2.2 S/W SYSTEM CONFIGURATION: -

Operating System : Windows 8 and above

Front End : PYTHON

Back End : MySQL

APP : web app

PLATFORM : VS CODE

## 2.3 PYTHON LIBRARIES

Numpy,Pandas,Matplotlib,Sklearn,Flask etc….

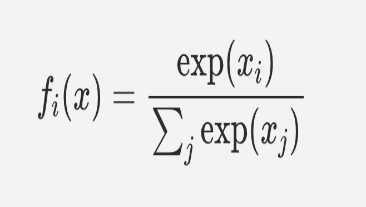
**3. SYSTEM ANALYSI****S - 3.1EXISTING SYSTEM**

An insurance claim fraud predicition is a challenging task since the factors Involved in pricing dynamically change overtime and make the price fluctuate. In the last decade, researcher have incorporated machine learning algorithms and Data mining strategies to better model observed prices but accurate results are not predicted since there is no day to day price comparison.

**ALGORITHMS USED:**

**LOGISTIC ALGORITHM**: Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

**NAÏVE BAYES ALGORITHM:** Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

**SOFTMAX REGRESSION ALGORITHM:**  The [Softmax function](https://en.wikipedia.org/wiki/Softmax_function) converts raw values (as an outcome of functions) into probabilities. In the function of softmax function was invented in 1959 by the social scientist R. Duncan Luce in context of choice models*.* Here is what the softmax function looks like:

**FIG 2:** **SOFTMAX REGRESSION DIAGRAM**

**SUPPORT VECTOR MACHINE:** Support Vector Machine or SVM is one of the most popular Supervised Learning (SL) algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors.

**3.2 PROPOSED SYSTEM**

Our system predicts whether the insurance claim is real or fake which means fraudulent. It considers the policy deductible , policy annual premium and the collision type to predict the premiums. XG Boost algorithm and Random forest algorithm have been used for the prediction, the output of which is obtained by performing voting among the class label obtained as result by each of these trees.

**RANDOM FOREST ALGORITHM:**

A Random Forest Algorithm is a supervised machine learning algorithm which is extremely popular and is used for Classification and Regression problems in Machine Learning. We know that a forest comprises numerous trees, and the more trees more it will be robust. Similarly, the greater the number of trees in a Random Forest Algorithm, the higher its accuracy and problem-solving ability.  Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

**XGBOOST ALGORITHM:**

The XGBoost (eXtreme Gradient Boosting) is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm that attempts to accurately predict a target variable by combining an ensemble of estimates from a set of simpler and weaker models. The XGBoost algorithm performs well in machine learning competitions because of its robust handling of a variety of data types, relationships, distributions, and the variety of hyperparameters that you can fine-tune. You can use XGBoost for regression, classification (binary and multiclass), and ranking problems.

## ****BOOSTING****

## Boosting is an ensemble learning technique to build a strong classifier from several weak classifiers in series. Boosting algorithms play a crucial role in dealing with bias-variance trade-off. Unlike bagging algorithms, which only controls for high variance in a model, boosting controls both the aspects (bias & variance) and is considered to be more effective.

**Below are the few types of boosting algorithms:**

AdaBoost,Gradient Boosting,XGBoost,CatBoost,Light GBM

## ****XGBoost****

## devLifeCycle_2XGBoost stands for eXtreme Gradient Boosting. It became popular in the recent days and is dominating applied machine learning and Kaggle competition for structured data because of its scalability.XGBoost is an extension to gradient boosted decision trees (GBM) and specially designed to improve speed and performance.

## ****ADABOOST:****

## AdaBoost is short for Adaptive Boosting. AdaBoost was the first successful boosting algorithm developed for binary classification. Also, it is the best starting point for understanding boosting algorithms. It is adaptive in the sense that subsequent classifiers built are tweaked in favour of those instances misclassified by previous classifiers. It is sensitive to noisy data and outliers. AdaBoost uses multiple iterations to generate a single composite strong learner. It creates a strong learner by iteratively adding weak learners. During each phase of training, a new weak learner is added to the ensemble, and a weighting vector is adjusted to focus on examples that were misclassified in previous rounds. The result is a classifier that has higher accuracy than the weak learner classifiers.

## ****GRADIENT BOOSTING:****

## Gradient boosting is one of the most powerful techniques for building predictive models, and it is called a Generalization of AdaBoost. The main objective of Gradient Boost is to minimize the loss function by adding weak learners using a gradient descent optimization algorithm.

**3.3 GOALS:**

Cost management will be attained and the insurance company will be reduced with the burden of fraud insurance claims. AI technology not only automates the fraud detection process but also identifies fraud patterns allowing early flagging and prompt response to any potential incidents.

## 4. ARCHITECTURAL DESIGN

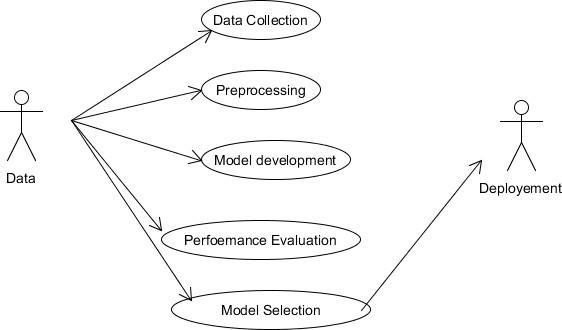
## 4.1 SYSTEM DESIGN

A system architecture is the conceptual model that defines the structure, behaviour, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviours of the system.

## FIG 3 : SYSTEM ARCHITECTURE

## 

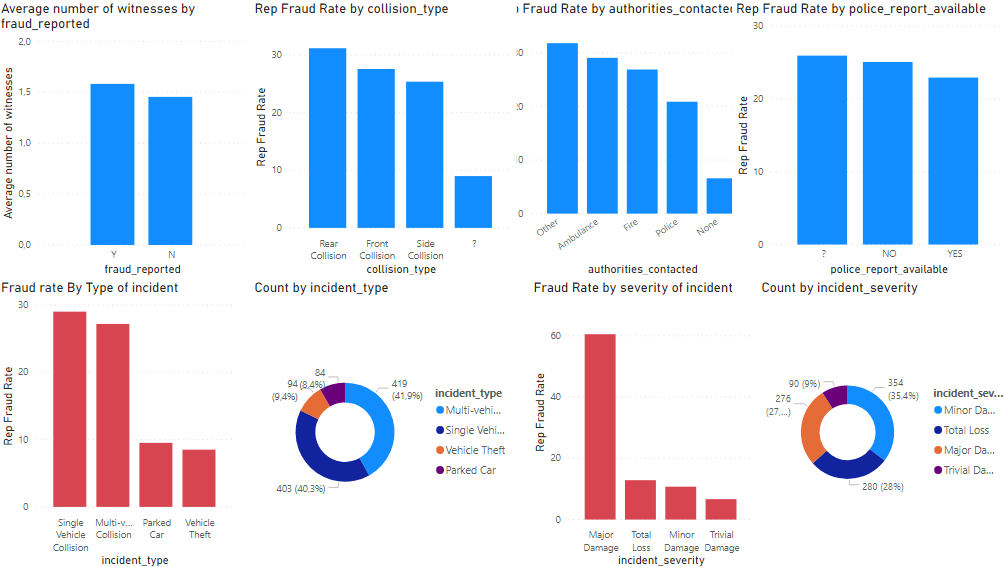
**4.2 USE CASE DIAGRAM:**

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis.Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases.

**FIG 4 :USE CASE DIAGRAM**

## 4.3 ACTIVITY DIAGRAM

**FIG 5: ACTIVITY DIAGRAM**

**5.MODULE DESCRIPTION:**

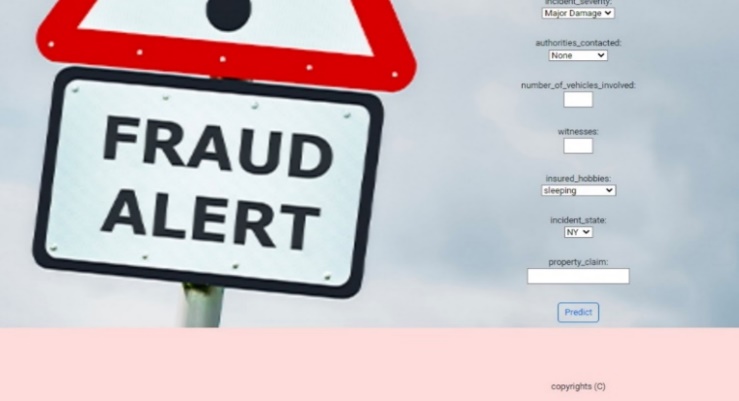
**5.1 DATA COLLECTION:**

Data collection is the process of gathering and measuring information from countless different sources. In order to use the data we collect to develop practical artificial intelligence (AI) and machine learning solutions, it must be collected and stored in a way that makes sense for the business problem at hand.

**5.2 DATA PREPROCESSING:**

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

**5.3 CREATION OF TREES USING C4.5 ALGORITHM**

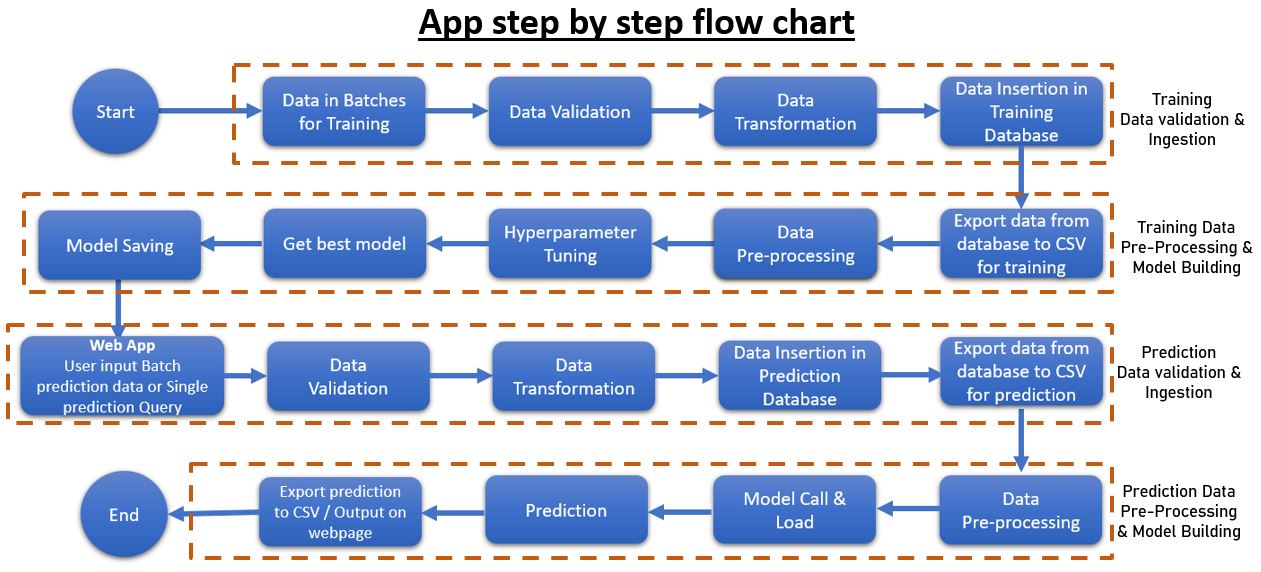
The trees that are built in our system are done so using C4.5 algorithm that is available on the internet. This algorithm uses gini index in order to find efficient split conditions for formation of daughter trees.

**5.4 MODEL SELECTION**

Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset.

**5.5 PERFORMANCE EVALUVATION**

In this process is used to implement the project Accuracy.

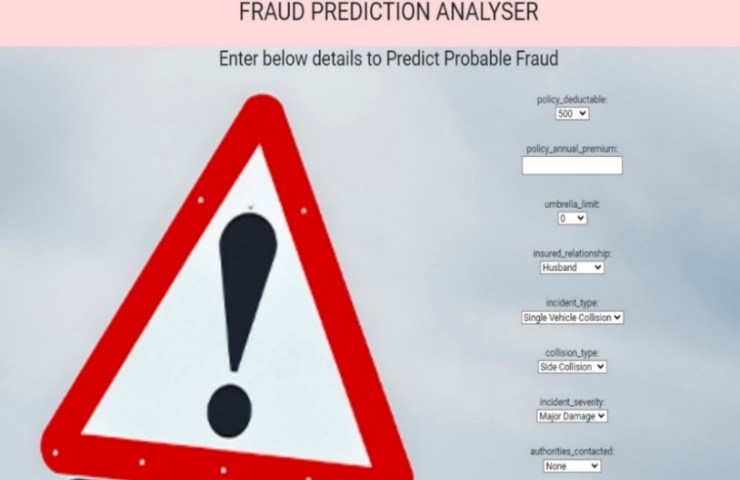
**6. PROJECT MODEL & DEPLOYMENT**

**FIG 9**

## 

## FIG 6: APP FLOWCHART

**FIG 7 : HISTOGRAM**

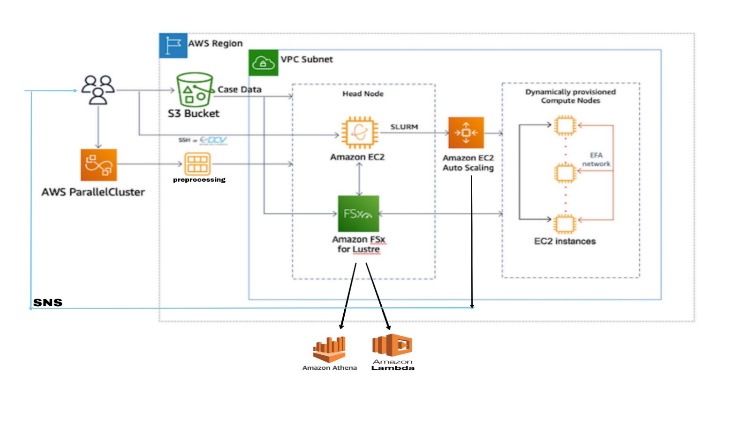


**FIG 8: WEB APP MODEL**

## https://blog.akscellenceinfo.com/wp-content/uploads/sites/2/2020/11/INSURANCE-CLAIM-FRAUD-PREDICTION-USING-MACHINE-LEARNING-1024x576.jpg

**6.1 DEPLOYMENT:**

Amazon Web Services (AWS) delivers reliable, scalable, and cost-effective computing resources on which to host our web applications. You can use the following AWS components alone or combined to host your application.

* Amazon Elastic Compute Cloud (Amazon EC2)
* Amazon Simple Storage Service (Amazon S3)
* Amazon Relational Database Service (Amazon RDS)
* Amazon CloudFront
* Amazon Simple Queue Service (Amazon SQS)
* ****Amazon DevPay

**FIG 10: AWS ARCHITECURE FOR THE PROJRCT**

**6.2 AWS SERVICES USED :**

**AWS PARALLELCLUSTER**

AWS ParallelCluster is an open source cluster management tool that makes it easy for you to deploy and manage High Performance Computing (HPC) clusters on AWS.

**AMAZON S3**

Amazon S3 to store and protect any amount of data for a range of use cases, such as data lakes, websites, mobile applications, backup and restore, archive, enterprise applications, IOT devices, and big data analytics.

**AMAZON EC2**

Amazon EC2 provides scalable computing capacity in the AWS Cloud. Using Amazon EC2 eliminates your need to invest in hardware up front, so you can develop and deploy applications faster.

**AMAZON EC2 AUTOSCALING**

It helps you maintain application availability and allows you to automatically add or remove EC2 instances according to conditions you define.

**AMAZON VPC**

Amazon VPC enables you to launch AWS resources into Amazon Virtual Private Cloud a virtual network and closely resembles a traditional network that you'd operate in your own data center, with the benefits of using the scalable infrastructure of AWS

**AMAZON FSx**

Amazon FSx makes it easy and cost effective to launch, run, and scale feature-rich, high-performance file systems in the cloud. In AMAZON FSx we use AWS LAMBDA & AWS ATHENA.

. AWS LAMBDA: It is built on the latest AWS compute, netwrking, and disk technologies to provide high performance and lower TCO.

. AWS ATHENA: It is an interactive query service that makes it easy to analyze data directly in Amazon S3 using standard SQL.

**7. CONCLUSION & FUTURE SCOPE**

**7.1 CONCLUSION**

Hence, we can conclude that machine learning is a most popular field among industry experts. After using this approach we can easily detect the potential frauds in the claims and also probability of fraud. By using this approach, we can early detect the frauds and this will be helpful in Reduction in fraud insurance investigation expenses,Lowers claim handling cost**,**Efficiently manages claims severity**,**Detection of early claims in the claim life cycle is paramount to managing overall claims costs.

An automated model to identify fraud claims in insurance industry. As explained in the results for XGBoost algorithm contain 94% which is the highest precision accuracy for fraud detection problem with machine learning data. Hence by the implementation of this model the insurance company can get accurate results in short duration of time. Thus, this automated framework can be used by any type of insurance company to reduce human labor and also to minimize the monetary loss in the insurance industry.

**7.2 FUTURE SCOPE:**

AI technology not only automates the fraud detection process but also identifies fraud patterns allowing early flagging and prompt response to any potential incidents. As the number of clients increases, claims adjusters are put under higher pressure and should either sacrifice the accuracy or the speed of the claims process. On the contrary, the more data machine learning algorithms receive, the faster they provide accurate results.

Next, predictive analytics delivers way more accurate results than a human agent can do. As a result of processing big data, digital tools have more information to make decisions with never before seen accuracy.By maximizing the use of technology and data analytics, insurers reduce the number of manual interventions in the claims management process. This reduces turnaround times and frees up insurance agents allowing them to focus on more valuable, high-impact tasks.

With more accurate fraud detection and reduced false positives made possible by AI technologies, insurers can decrease financial loss significantly. Also, by automating repetitive processes like fraud detection, you won’t need to increase your headcount as you scale up – which otherwise would have come with extra costs.

**8. REFERENCE**

1. Detecting Insurance Claims Fraud Using Machine Learning (ML) Techniques, 19 OCTOBER 2017, IEEE

2. Robust Fuzzy Rule Based Technique To Detect Fraud In Vehicle Claim , 21 JUNE 2018,IEEE

3. Insurance Fraud Detection Using Machine Learning, APR 2021,IRJET.

4. Enhancing Claims Handling Processes With Insurance Based Language Models, FEB 19 2019, IEEE

5. Detecting Fraudulent Motor Insurance Claims Using Support Vector Machines With Adaptive Synthetic Sampling Method. 19 NOVEMBER 2020, IEEE

6. Automobile Insurance Fraud Detection Using Supervised Classifiers, 21 NOVEMBER 2020, IEEE

7. Analysis Accuracy Of XGboost Model For Multiclass Classification - A Case Study Of Applicant Level Risk Prediction For Life Insurance, 10 FEBRUARY 2020,IEEE

8. A Data Mining Based Target Regression-Oriented Approach To Modelling Of Health Insurance Claims, 6 MAY 2021, IEEE