Identifying the probability of the Natural Disaster to help the Insurance Company to take decisions on providing Insurance

MSc Research Project
MSc Fintech

Dhanush Raju
Student ID: x22196757

School of Computing
National College of Ireland

Supervisor: Sean Heeney & Noel Cosgrave

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Dhanush Raju

Student ID: X22196757

Program: MSc Fintech **Year:** 2023 - 2024 ……………………

Module: Practicum

Supervisor: Sean Heeney & Noel Cosgrave

Submission Due Date: 12/08/2024 ……………………………………………………………………………………………………

Project Title: Identifying the probability of the Natural Disaster to help the Insurance Company to take decisions on providing Insurance

Word Count: 7547 **Page Count:** 22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Dhanush Raju ………

Date: 11/08/2024 ………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Identifying the probability of the Natural Disaster to help the Insurance Company to take decisions on providing Insurance

Dhanush Raju
X22196757

Abstract

The aim of this study is to determine the probability of natural disasters occurring in a specific area, with the purpose of aiding Insurance companies in making well-informed choices, The process involves in the analysis of an extensive historical data, meteorological data, and statistical models based on either the parent insurance company or the third party insurance providers. Historical data on past catastrophes or natural disaster is collected and analyzed with meteorological factors associated with the occurrence of disasters. Statistical models, such as logistic regression and other machine learning algorithms like xgboost, light gradient boosting (LGBM), random forest are used to predict the likelihood of future disasters. We propose a response system which helps in the decision making of the insurance companies to save the claim amounts for a risk management of the company. The findings offer different useful insights that can inform decision-making in insurance companies, allowing them to make required adjustments to pricing, coverage restrictions, and risk mitigation strategies. Here we make a propensity based triple stage system which uses LGBM model due to its high performance and light weight to predict the propensity of the natural disasters. While for detecting the suspicious fraud model, similar model with different hypermeter tuning is used. The third stage is response system which can be used to deploy and find those cases which are prune to high risk areas and no suspicious.

Keywords: Insurance, Risk Analysis, Machine Learning, Finance, Impact Assessment, Multimodal Risk Management, Impact Analysis

1 Introduction

Insurance is an Indispensable in contemporary society and trade as it can facilitates the different consolidation of “certain” risks and mitigates the financial repercussions of natural disasters and calamities. The increasing in frequency and severity of different kinds natural disasters exacerbate the inequality in insurance coverage by protecting and inflicting even greater financial damages on insurance firms, leading to higher premiums or the entire termination of services (**Sheehan, B., et., al, 2023**).

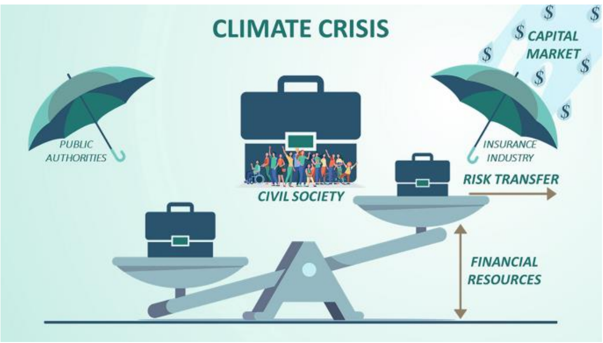


Figure 1: This image depicts the partnership between the public and private sectors in obtaining the necessary financial resources to improve the ability to withstand and recover from catastrophic risks in the field of financial management (Sheehan, B., et al., 2023).

This article introduces a framework for relationship between the insurance and DRM (Disaster Risk Management) communities that comprises of investigator and underwriters with the aim of attaining disaster risk management that is open, transparent, and ideal for both individuals and companies. The study emphasizes how important the insurance industry is to efficient disaster risk management (DRM) for risks related to financial loss and disaster-related concerns. This pipeline model identifies key areas where insurers can take a more proactive approach during natural disasters by implementing preventive measures to protect individuals at risk from climate change-related hazards and also the insurance providing companies for saving the risk losses with a high propensity (**Sheehan, B., et al., 2023**).

1.1 Research Objective

The need for Insurance companies to accurately assess and manage the risks associated with natural disasters is what spurred this research (Kunreuther, H., 1996). Natural disasters can have disastrous effects on economies and communities, causing large financial losses for both people and businesses. By using these risk models, insurance companies can improve their policies, rates, and risk management strategies by accurately determining the possibility of natural disasters in certain areas. (**Chen, C.W., Tseng, C.P., Hsu, W.K. and Chiang, W.L., 2012**). The objective of this research is to equip Insurance firms with the requisite tools and information to make informed assessments, drawing on prior studies and current advancements in the field of Artificial Intelligence. In the end, this will improve their ability to safeguard Insured assets and reduce financial losses in the case of a natural disaster. Artificial Intelligence may improve the precision of risk models by using different A.I plugins like sentiment analysis, terrorism analysis, flood prediction, and others.

Considering the foregoing discussion above, the research questions that will guide the course of this study's development and advancement are as follows:

Research Question: How the predicting the propensity of the natural disaster help in minimizing the risks of loss for any financial companies like insurance providers?

Research Area: In this process we are trying to make a staging model which will help the insurance companies minimize the loss by providing insurances to those places which are high change to fall in for a natural calamity which is an adverse for any claims options (in other words we are interested in making a response system).

In the subsequent chapters we will focus on the detailed research on the different literatures done towards achieving the risk analysis and risk management for different companies providing insurances or for risk analysis.

2 Related Work

The number of economic losses frequently exceed the costs of repairing and recovering the damaged property. For example, if a road bridge is demolished, it not only exhausts and reaches the financial assets required for its reconstruction, but also leads to a decline in economic advantages as correctly depicted (**Glenday, G., et.al. 2020**). As the duration and the process of recovery needed to repair the bridge or construct other traffic routes lengthens, the economic losses persistently escalate with different kinds of other resirce requirements. This raises the question of whether the government bodies (in respince to the things to get it done) could have designed and built a bridge that could have withstood the flood (**Jena, R., Pradhan, B., Beydoun, G., Al-Amri, A. and Sofyan, H., 2020**), or developed alternative transportation routes or methods (like using military pontoon bridges) in case the bridge failed, or established efficient repair or replacement capabilities. The goal of tis detailed research study is to bolster the economy ability to withstand anticipated natural disasters by adapting and services public investment appropriately (**Kalfin, Sukono, Supian, S. and Mamat, M., 2022**).

2.1 Research conducted on Identifying the Natural disaster

This paper by the researchers Cui and team (**Cui, P. et al., 2023**) presents a detailed and thorough examination and analysis of contemporary studies on natural hazards, with a specific emphasis on seven crucial areas like: the genesis, mechanism, and dynamics of natural hazards; evaluation of disaster risks; prediction; surveillance and early warning systems; risk management and post-disaster restoration; emergency response and rescue operations; and catastrophe prevention. The study looked at how various patterns exist in the meteorological sector of natural disasters and catastrophic risk management, as well as how to minimize current flaws and highlight important areas that require attention. The primary idea and research topic of this article (**Cui, P., et. al., 2023**) revolves around two key features. Firstly, it focuses on the scientific issues, such as cutting-edge scientific challenges and technology deficiencies, related to natural hazards and catastrophe risk in China between 2025 and 2035. Furthermore, it also suggests the establishment of a thorough and unified field of study dedicated to examining natural hazards and the risk of disasters.

Another article (**Aljohani, F.H., et. al., 2023**) lays up a enhanced response for better system for managing floods which is one of the most important risk management. To address the different kinds of shortcomings of conventional flood prediction systems, such as excessive latency and erroneous predictions, this framework as proposed employs and makes a collaborative the IoT and Machine Learning modules response system. The framework also makes sure that the part of flood detection and alert systems (flags) can be adjusted to different parts of a city and also globally, considering the unique topologies and conditions of each area, by integrating real-time data collected from multiple sensors with fog computing and cloud computing. This system's architecture consists of four layers: sensors, fog, the cloud, and applications for the usage. Reduced cloud space load and faster (computation power) threat categorization and warnings are both made possible by the different fog layer's local analysis of this data. The top three models were found to be DT, KNN, and RF, with the former achieving a model accuracy rate of over 99%.

The primary subject of investigation in disaster risk science, as highlighted by (**Shi et al., 2020**), is an enhanced management analysis and system for the disaster prediction. This system encompasses hazards, the geographical environment, and the units that are at danger of being affected. This system exhibits distinct regional characteristics, interconnectedness, coupling, and complexity. The geographical environment influences the severity of local dangers, which might subsequently change the distribution of damages. Regional multi-hazard events, catastrophe chains, and disaster compounds can have complex effects by either intensifying or reducing the intensity of hazards and changing the affected areas. The analysis in the H2020 ESPRESSO project, that is an acronym for Enhancing Synergies for Disaster Prevention in the European Union, was coordinated by the researchers (**Zuccaro, Leone, and Martucci, 2020**).

This project's major goal was to highlight specific study areas, new innovation requirements, and important objectives in the fields of climate change adaptation, disaster risk reduction, disaster risk management, and natural hazards. The projects that are now underway in Europe and throughout the world on climate change adaptation, disaster risk reduction, disaster risk management, and natural hazards have all been reviewed by the researchers. The purpose of research and innovation is to provide a comprehensive framework that can successfully manage the complexities of many disciplines. (**Zuccaro, Leone, and Martucci, 2020**) state that the objective of this framework is to offer a comprehensive perspective on the rising objectives.

Numerous thorough global risk assessments have been conducted to examine meteorological, climatological, and hydrological threats **(Ward, P.J., et al., 2020).** With an emphasis on floods specifically, these studies have included forecasts for the future as well as methods to reduce the danger of catastrophic events. However, there aren't many peer-reviewed studies that specifically address geological concerns in the global literature. Stochastic modeling techniques have been applied in recent studies on earthquake and tsunami risk to provide a comprehensive probabilistic evaluation of risk.

2.2 Research on machine learning enhanced natural disasters prediction system

Building reliable prediction models to lessen the impact of floods is the main goal of the research (**Hayder, I.M., Al-Amiedy, T.A., Ghaban, W., Saeed, F., Nasser, M., Al-Ali, G.A. and Younis, H.A., 2023**). By combining recurrent neural networks (RNN) with exponential smoothing-long short-term memory (ES-LSTM), a hybrid forecasting model was built to predict hourly precipitation. In order to further assess and classify precipitation, decision trees (DT) and artificial neural networks (ANN) were employed. The dataset utilized was sourced from the Australian Commonwealth Office of Meteorology. The method consisted of three steps: data preparation, prediction or forecasting, and decision-making or response to the predictions or forecasts. Part of getting the data ready for analysis was filling in missing values, doing data transformations, and standardizing the numbers. Predictions for time-series analysis were made using DT and ANN algorithms, while RNN and ES-LSTM were used for forecasting. With a MAPE of 3.17 and 6.42, respectively, ES-LSTM and RNN accomplished their goals. In contrast, DT arrived at an accuracy of 84.0 percent and ANN at 96.65% in their predictions.

The purpose of a study by (**Linardos, V., et. al., 2022**) is to provide a thorough examination of the research endeavors carried out since 2017, which concentrate on the advancement of machine learning (ML) and deep learning (DL) techniques like CNN, RNN etc. for the purpose of disaster management. The paper by (**Vinod, A.M., et al., 2022)** offers a thorough exploration of the most recent research and findings in the domains of data analysis, forecasting natural calamities, and the application of technology for executing management methods. This particular article explicitly analyzes the present condition of Industry 4.0, which encompasses cognitive computing. The primary aim of this article is to analyze the study principles that employ big data and data mining to detect and track patterns, which might facilitate predictive analysis for anticipating future disasters.

Information technology and communication development has enhanced the procurement of big data sets that are of immense value to scholars. This approach of its use in identification of abnormal situations has shown high benefits in management of disaster and emergency situations (**Sreelakshmi, S. and Chandra, S. V. , 2022**) mainly in pre and post event phases in the best interest of all the persons involved. Regarding flood management there can be the application of new and innovative methods and tools as well as the application of technology. These are flood forecasting, identification, surveying, management and rescue and rehabilitation (**Sreelakshmi, S. and Chandra, S. V. , 2022**). To reduce casuistry and the adverse influence of floods on the environment and economic system, it requisite accentuate techniques which are intended to minimize the consequences of floods and to counteract the associated disasters speedily. Therefore, the main focus in the subsequent endeavors should be aimed at fostering the collaboration between the knowledge in the disaster management, image processing techniques as well as in the application of Machine Learning algorithms to ensure the chances of proper and effective catastrophe management in all processes.

The study (**Wu, L., Ma, D. and Li, J., 2023**) set out to use a Data Envelopment Analysis (DEA) model to estimate the susceptibility of various mainland Chinese regions to natural catastrophes from 2006 to 2021. For this model, we included data on things like regional populations, GDP indicators, population density, GDP per square kilometer, and infrastructure investments related to water conservation, environmental management, and public facility management. The number of people impacted and the monetary losses directly attributable to natural catastrophes were both included in the results. The DEA-BCC model was chosen for its ability to handle many inputs and outputs and to compare the relative efficiency of homogeneous decision-making units (DMUs). Every year, official statistics from China were also incorporated into the study to ensure the accuracy and reliability of the data. A overall decrease in vulnerability was observed as one moved from West to Central and then East China, according to the results.

2.3 Research on Insurance Companies

The primary goals of this study **(Hussein, A.A. and Zoghlami, F., 2023)** are to identify the factors that lead Iraqi insurance companies to discontinue offering engineering insurance and to look into ways to bring engineering insurance back into project management. Additionally, the study aims to examine any potential benefits that bank loans could offer Iraqi insurance businesses. Improving the prediction of landslip displacement using advanced modelling techniques was the goal of the research (**Duan, G., Su, Y. and Fu, J., 2023**) conducted in the Baijiabao region. The study's results show that in order to achieve better accuracy by include data on reservoir water levels and rainfall, researchers compared traditional ARIMA models with LSTM models. The dataset contained information on rainfall, reservoir levels, and displacement from monitoring site ZG323 from November 2006 to December 2012. After applying cubic spline interpolation in the preprocessing step, ARIMA, univariate LSTM, and multivariate LSTM models were installed for prediction. The study conducted by Kumar, S., Rao, P., and Barai, M. in 2024 using a systematic literature review (SLR) technique.

A number of significant insights have emerged from the extensive research on the topic (**Poufinas, T., Gogas, P., Papadimitriou, T. and Zaganidis, E., 2023**) of auto insurance claim prediction using new factors and ML algorithms. First, compared to conventional metrics, a more nuanced understanding of claim trends is obtained by include elements like weather and new car sales as predictive variables. The results emphasise the importance of these variables in forming quarterly claim averages, especially the lagging impacts of new car sales and minimum temperatures. This method is useful for strategic planning and risk management because it increases the reliability of predictions and provides insurers with practical information about the effects of economic and seasonal factors on claims.

2.4 Summary

The number of previous studies available on natural hazards and disaster risk reduction, helped by various funding agencies such as the Natural Science Finance of China (NSFC), the Main Directions Finance of the CAS and many more have enlightened on ways to minimize risks and improve readiness for the impending disasters. These are comprehending the development and characteristics of natural disasters, how risky it is to be in an area where there is a possibility of natural disasters occurring, predicting future natural disasters, how to prevent losses in the event of natural disasters, and how to cope with natural disasters when they occur. The H2020 ESPREssO project and other similar projects are focused on establishing the priorities for research as well as offering the conceptual frameworks of disaster prevention. Machine learning and AI models give the possible solutions to management of disaster; such as disaster forecasting, early warning systems, and disaster assessment.

Table 1: Summary of some of the important limited researches done

Name	Author	Model	Dataset	Result
A Smart Framework for Managing Natural Disasters Based on the IoT and ML	Aljohani, F.H., Abi Sen, A.A., Ramazan, M.S., Alzahrani, B. and Bahbouh, N.M., 2023.	Random Forest, Decision Tree, K-Nearest Neighbor	Jeddah weather data (2009-2013, 2018-2022)	Over 99% accuracy achieved
An Intelligent Early Flood Forecasting and Prediction Leveraging Machine and Deep Learning Algorithms with Advanced Alert System	Hayder, I.M., Al-Amiedy, T.A., Ghaban, W., Saeed, F., Nasser, M., Al-Ali, G.A. and Younis, H.A., 2023.	Hybrid ES-LSTM, RNN, ANN, DT	Australian meteorology office data.	ES-LSTM best, 3.17 MAPE.
Assessment of the Regional Vulnerability to Natural Disasters in	Wu, L., Ma, D. and Li, J., 2023.	DEA-BCC for vulnerability assessment	Chinese official statistics 2006-2021	Regional vulnerability trends identified

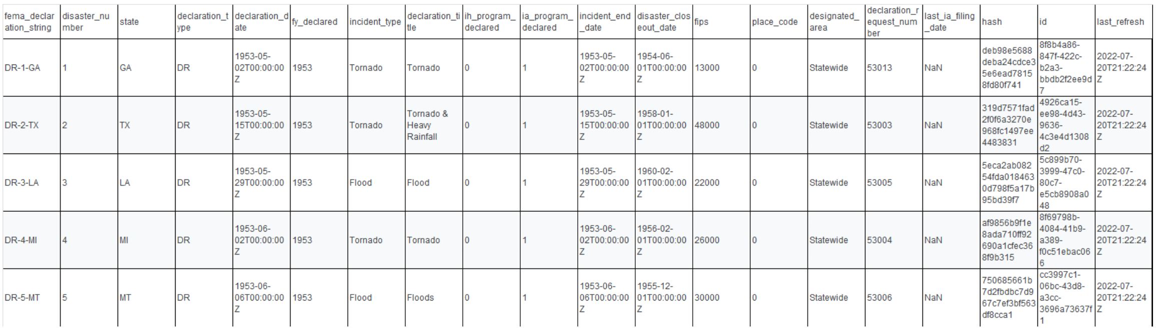
China Based on DEA Model				
Landslide Displacement Prediction Based on Multivariate LSTM Model	Duan, G., Su, Y. and Fu, J., 2023.	Multivariate LSTM	aijiabao displacement, rainfall, reservoir	Improved accuracy, precise displacement prediction
Machine Learning in Forecasting Motor Insurance Claims	Poufinas, T., Gogas, P., Papadimitriou, T. and Zaganidis, E., 2023.	Random Forest, XGBoost	Motor insurance claims data	Improved forecasting accuracy significantly

3 Research Methodology

According to the study, we discovered a need in the current technology for enhancing machine learning models' performance.Specifically, there is a need for a more advanced and versatile system that incorporates Intelligence and several modes of operation.
1. Provide an explainable artificial intelligence (AI) solution that can identify both the hazards and their underlying causes.
2. This will enhance the model's ability to accurately predict the real-time hazards associated with natural catastrophes and their influence on Insurance firms' risk exposure.

Dataset 1 – Natural Disaster: The dataset available at https://ourworldindata.org/natural-disasters. This dataset offers data on natural disaster incidents that have occurred globally, along with the resulting economic effects. Drought, Earthquake, Extreme temperature, Extreme weather, Flood, Impact, and Volcanic activity are among the natural calamities that are covered. It also contains information on the cumulative incidence of several catastrophic disasters. The period of time covered is 1900–2018.

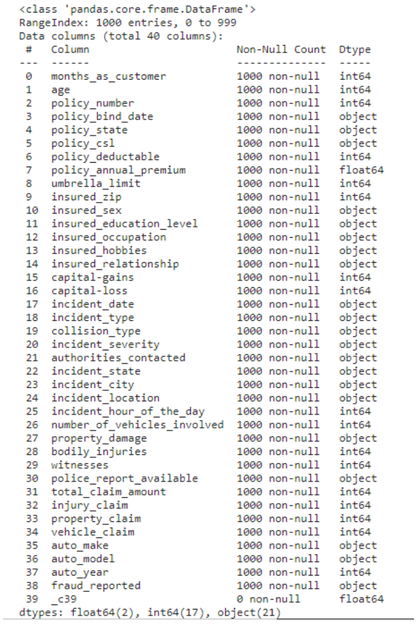
Figure 2: Natural Disaster Declarations Data sample



fema_declaration_string	disaster_number	state	declaration_type	declaration_date	fy_declared	incident_type	declaration_title	ih_program_declared	ia_program_declared	incident_end_date	disaster_closeout_date	fips	place_code	designated_area	declaration_request_number	last_ia_filing_date	hash	id	last_refresh
DR-1-GA	1	GA	DR	1953-05-02T00:00:00Z	1953	Tornado	Tornado	0	1	1953-05-02T00:00:00Z	1954-06-01T00:00:00Z	13000	0	Statewide	53013	NaN	deb98e5688deba24cdce35e6ead78158fd80f741	8f8b4a86-847f-422c-b2a3-bbdb2f2ee9d7	2022-07-20T21:22:24Z
DR-2-TX	2	TX	DR	1953-05-15T00:00:00Z	1953	Tornado	Tornado & Heavy Rainfall	0	1	1953-05-15T00:00:00Z	1958-01-01T00:00:00Z	48000	0	Statewide	53003	NaN	319d7571fad2f0f6a3270e968fc1497ee4483831	4926ca15-ee98-4d43-9636-4c3e4d1308d2	2022-07-20T21:22:24Z
DR-3-LA	3	LA	DR	1953-05-29T00:00:00Z	1953	Flood	Flood	0	1	1953-05-29T00:00:00Z	1960-02-01T00:00:00Z	22000	0	Statewide	53005	NaN	5eca2ab08254fda0184630d798f5a17b95bd39f7	5c899b70-3999-47c0-80c7-e5cb8908a048	2022-07-20T21:22:24Z
DR-4-MI	4	MI	DR	1953-06-02T00:00:00Z	1953	Tornado	Tornado	0	1	1953-06-02T00:00:00Z	1956-02-01T00:00:00Z	26000	0	Statewide	53004	NaN	af9856b9f1e8ada710ff92690a1cfec368f9b315	8f69798b-4084-41b9-a389-f0c51ebac066	2022-07-20T21:22:24Z
DR-5-MT	5	MT	DR	1953-06-06T00:00:00Z	1953	Flood	Floods	0	1	1953-06-06T00:00:00Z	1955-12-01T00:00:00Z	30000	0	Statewide	53006	NaN	750685661b7d2fbdbc7d967c7ef3bf563df8cca1	cc3997c1-06bc-43d8-a3cc-3696a73637f1	2022-07-20T21:22:24Z

Dataset 2 – Insurance Fraudulant Claims: In the context of this project, we have a dataset that contains the information on the client together with the specifics of the insurance policy. In addition to this, it contains the information pertaining to the incident that led to the filing of the claims. The dataset that has been provided to us has a total of 1000 rows and 40 columns. The titles of the columns, such as the policy number, the policy bind date, the yearly premium, the severity of the event, the location of the occurrence, the vehicle model, and so on.

Figure 3: Insurance Fraudulant Claims Data sample



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 40 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   months_as_customer           1000 non-null   int64  
 1   age                          1000 non-null   int64  
 2   policy_number                1000 non-null   int64  
 3   policy_bind_date             1000 non-null   object 
 4   policy_state                 1000 non-null   object 
 5   policy_csl                   1000 non-null   object 
 6   policy_deductable            1000 non-null   int64  
 7   policy_annual_premium        1000 non-null   float64
 8   umbrella_limit               1000 non-null   int64  
 9   insured_zip                  1000 non-null   int64  
 10  insured_sex                  1000 non-null   object 
 11  insured_education_level      1000 non-null   object 
 12  insured_occupation           1000 non-null   object 
 13  insured_hobbies              1000 non-null   object 
 14  insured_relationship         1000 non-null   object 
 15  capital-gains                1000 non-null   int64  
 16  capital-loss                 1000 non-null   int64  
 17  incident_date                1000 non-null   object 
 18  incident_type                1000 non-null   object 
 19  collision_type               1000 non-null   object 
 20  incident_severity            1000 non-null   object 
 21  authorities_contacted        1000 non-null   object 
 22  incident_state               1000 non-null   object 
 23  incident_city                1000 non-null   object 
 24  incident_location            1000 non-null   object 
 25  incident_hour_of_the_day     1000 non-null   int64  
 26  number_of_vehicles_involved  1000 non-null   int64  
 27  property_damage              1000 non-null   object 
 28  bodily_injuries              1000 non-null   int64  
 29  witnesses                    1000 non-null   int64  
 30  police_report_available      1000 non-null   object 
 31  total_claim_amount           1000 non-null   int64  
 32  injury_claim                 1000 non-null   int64  
 33  property_claim               1000 non-null   int64  
 34  vehicle_claim                1000 non-null   int64  
 35  auto_make                    1000 non-null   object 
 36  auto_model                   1000 non-null   object 
 37  auto_year                    1000 non-null   int64  
 38  fraud_reported               1000 non-null   object 
 39  _c39                         0 non-null      float64
dtypes: float64(2), int64(17), object(21)
```

3.1 Architecture Flow

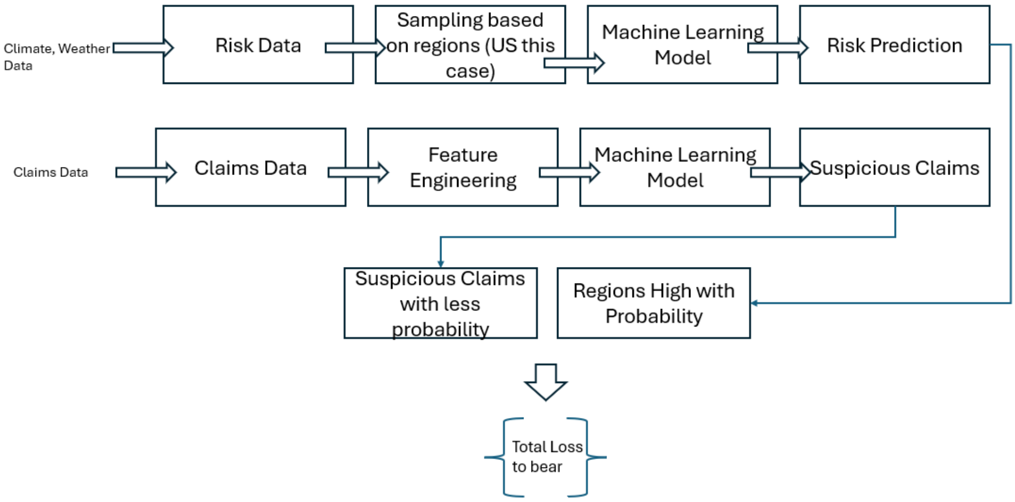


Figure 4: Dual Stages incorporating the Risk Management and Suspicious Claims prediction

Stage 1 - Risk Management: Disaster data preprocessing involves selecting relevant features from the disaster dataset, converting categorical variables into numerical values through one-hot encoding, and cleaning the data by removing unnecessary columns and handling missing values. After training the models, predictions are made on the preprocessed data.

Stage 2 - Suspicious Fraud Detection: For insurance data preprocessing, relevant features are selected from the insurance dataset. Categorical variables are converted into numerical values using one-hot encoding, and unnecessary columns are removed to reduce noise and improve model performance.

Modelling: After training the models, predictions are made on the preprocessed data. The fraud detection model is used to make insurance predictions, while the disaster classification model predicts disaster declarations.

Stage 3 - Risk Identification for minimizing loss: The next step is to match insurance claims with disaster declarations based on state and year. This involves iterating through the insurance and disaster data frames to compare state and year values. Matched records are compiled into a new data frame, which can be analyzed to investigate potential correlations between insurance fraud and disaster declarations.

4 Design Specification

The models and the techniques used in the Implementation.

4.1 Data Imbalance Techniques

In order to cope with datasets that are extremely skewed towards a particular class value, resampling is a common method that can be utilized. To do this, it is necessary to either decrease the number of cases from the majority class (also known as Under Sampling) or increase the proportion of instances from the minority class (also known as Over Sampling). Concerns Regarding Over-representation and Under-representation in Surveys Conducted on a Number of sample of Populations (**Thabtah, F., Hammoud, S., Kamalov, F., and Gonsalves, A., 2020**). Although these methods have their own drawbacks, they do help to maintain an equivalent amount of processes for the solution to work better. Oversampling can be carried out in a number different methods, including the following:

- Under-sampling is the practice of purposefully omitting records from the majority class;
- The most straightforward method is to replicate data from the minority class at random, notwithstanding the fact that this might result in overfitting.

Following techniques are used to do an extensive analysis of different sampling related to find out the best method that can be deployed in the real time. Below are descriptions of them:

1. **Random Under-Sampling**: To equalize the two sets of data, the oversampled samples are randomly eliminated. The removal of data from the dataset has a detrimental impact. Furthermore, another aspect to be negotiated is the specific distribution of classes.
2. **Random Over Sampling (R.O.S.):** enhances the likelihood of the minority class by duplicating its members at random and including them into the dataset.
3. **Stratification:** It is grouping together similar cases to increase the representation of certain groups, while maintaining a consistent ratio. The presence of oversampling causes the models to become overfit while this method keeps intact the population.
4. **Synthetic Minority Over Sampling Techniques (SMOTE)**: These techniques involve generating additional data points for the minority class by oversampling from

neighbouring cases. When working with data that has a large number of dimensions, this strategy is completely ineffective.

4.2 Cross Validation Techniques

In order to ensure accurate generalization of machine models, cross validation (CV) approaches are utilized (**Berrar, D., 2019**). The data is partitioned into training and testing sets using various techniques. Several techniques include:

a. **Hold Out Method (Fixed Ratio)**: Using this method, the ratio of the total data shape determines the test sample size, which in turn divides the data into two samples.
b. **K-Folds Cross Validation**: There are k folds produced in total. K-1 folds of data are used in the training phase, and the remaining k fold data is used for validation.

4.3 Machine Learning Models

4.3.1 XgBoost

XGBoost (extreme gradient boosting) is a machine learning framework that use gradient-boosted decision trees (GBDT) for its algorithms (**Chen, T. and Guestrin, C., 2016, August**). The system is specifically engineered to have the capability to expand in size and be spread over several locations. The best machine learning tool for regression, classification, and ranking tasks is XGBoost, which offers parallel tree boosting.
This is the reason why the algorithm is fast.

4.3.2 Random Forest

The Random Forest (**Rigatti, S.J., 2017**) creates a number of randomly week models in the form of decision trees which are then trained on randomly selected samples of the data in the form of bootstrapped aggregation. The reason for the random forest being a bootstrap ensemble is due to the random sampling of hyperparameters and samples.

4.3.3 Light Gradient Boosting

LightGBM is a gradient boosting framework created by Microsoft, as described in the work of (**Taha, A.A. and Malebary, S.J., 2020**). The software is open source, meaning its source code is freely available for anyone to see, modify, and distribute. It is also distributed, meaning it may be used on several computers or systems simultaneously. Additionally, it is specifically built to provide exceptional performance. This product is meticulously designed to enhance efficiency, scalability, and accuracy. Decision trees, which are carefully made to increase model efficiency and decrease memory usage, are the foundation of the system. Several state-of-the-art approaches are incorporated in the method, such as Gradient based One Side Sampling, which optimizes memory utilization and reduces training time by preferentially retaining instances with significant gradients during training.

5 Implementation

Detecting Insurance fraud is a difficult topic due to the numerous fraud schemes that may be used and the very low ratio of recognized frauds that is found in usual samples. When developing detection models, it is necessary to strike a balance between the amount of money saved because of loss prevention and the amount of money wasted on false alarms. This research looks not only the detection of the suspicious insurance fraudulent claims but looks into integrating a unique pipeline which,

Takes into the account the risk of identifying the natural disaster happening at a region that will help in taking a decision of whether the insurance company should give insurances to those probably affected regions or not. Or in other words anew input to the insurance company

If the decisions provided by the risk model is avoided and the insurance is sold, there are two things, either the claims are suspicious or correct claims. If the claims are suspicious, they will be in the looks of the insurance investigators or the underwriters. But if the claims are correct, they are those probable claims which could have been avoided using our risk models.

5.1 Stage 1 – Disaster Classification

The disaster declaration classification model is designed to predict the type of disaster declaration based on various features.

Data Collection - Importing necessary libraries for data manipulation, machine learning, and visualization. These libraries provide the tools needed for the entire machine learning pipeline, from data preprocessing to model evaluation. The first step in the process is loading and preprocessing the data. This is handled by a function that reads the CSV file using pandas and performs several cleaning steps.

Feature Engineering - Unnecessary columns that won't be used in the model are removed, reducing noise and improving model performance. Next, categorical variables are encoded using one-hot encoding, converting variables like state and incident type into binary columns that the machine learning models can understand. The target variable, 'declaration_type,' is also encoded into numerical values. After preprocessing, the data is prepared for modeling. The features (X) are separated from the target variable (y), and the data is split into training and testing sets. An 80-20 split is used, with 80% of the data for training and 20% for testing.

Model Building -The implementation includes several different classifiers, defined in a function that returns a dictionary of classifiers. This approach allows easy comparison of different machine learning algorithms on the dataset. The classifiers included might be RandomForestClassifier, GradientBoostingClassifier and others. The core of the implementation is the function that trains and evaluates the classifiers. For each classifier, it trains the model on the training data, makes predictions on the test data, calculates the accuracy, generates a classification report, and creates a confusion matrix. This function also visualizes the confusion matrix for each classifier, providing a clear picture of where each model excels or struggles.

Finally, a main function ties everything together. This function orchestrates the entire process, from data loading to final evaluation, making it easy to run the entire analysis with a single function call. By structuring the code this way, a flexible and reusable pipeline for disaster declaration classification is created. This design allows for easy experimentation with different preprocessing steps, classifiers, or evaluation metrics, and the modular design makes it simple to extend the code by adding new classifiers or implementing cross-validation.

5.2 Stage 2 – Insurance Fraud Detection

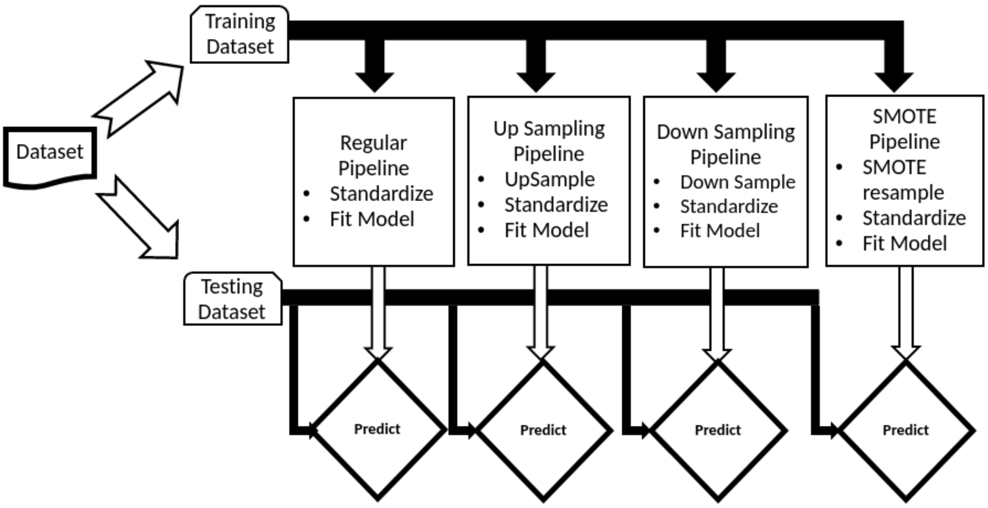


Figure 5 : Basic Implementation of the Fraudulent Claims detection pipeline (**Stage 2**)

Data Acquisition: To determine whether the claims data meets the minimal requirements for the model construction, it is first collected and examined.
The model building's minimum row count is one. In this instance, at least 1000 rows are taken into account. The publishing organization's low availability and data privacy are the cause of this. No ambiguities in the data, such as incorrect information or improper data encoding
Information Pre-processing: To understand the data, a thorough exploratory data analysis must be performed. The stages that are considered in this situation are as follows are exclusion of unnecessary features Imputation of Missing Values Finding Outliers , Data Scaling Feature Encoding. Understanding the Imbalance of the Data: The following sampling techniques needs to be done for making the transformation of the data to be balanced,

a. Random Over Sampling
b. SMOTE
c. Stratification
d. Random Under sampling

The Hold Out and K-Fold procedures are used to scale the dataset, encode it, and create training, validation, and out-of-sample datasets then the model is fitted and its parameters are learned using the training data. A few strategies to employ in order to improve learning are Optimizing Hyperparameters with Grid Search.
The Insurance fraud estimation model aims to predict whether a claim is fraudulent based on various features. The Implementation starts by importing necessary libraries for data manipulation, machine learning, and visualization, providing the tools needed for the entire pipeline, from data preprocessing to model evaluation. The first step in the process is loading and preprocessing the data. The `load_and_preprocess_data` function reads the dataset from a CSV file and performs several cleaning steps. It removes unnecessary columns to reduce noise and improve model performance. Categorical variables, such as 'property_damage' and 'police_report_available', are encoded to numerical values. The 'incident_date' column is processed to extract only the year, and one-hot encoding is performed on the 'incident_location' column to convert categorical data into binary columns.

After preprocessing, the data is prepared for modeling in the `prepare_data` function. This function separates the features (X) from the target variable (y) and splits the data into training and testing sets, using 80% for training and 20% for testing. The implementation includes several classifiers, defined in the `get_classifiers` function. This allows easy comparison of different machine learning algorithms on the dataset. The classifiers might include RandomForestClassifier, GradientBoostingClassifier, GaussianNB, LogisticRegression, SVC with different kernels, LGBMClassifier, and XGBClassifier. The core functionality is in the `train_and_evaluate` function. For each classifier, it trains the model on the training data, makes predictions on the test data, calculates accuracy, generates a classification report, and creates a confusion matrix. This function also visualizes the confusion matrix for each classifier, providing a clear picture of each model's strengths and weaknesses.

Finally, the `main` function integrates all the steps, from loading and preprocessing the data to training and evaluating the models. This function makes it straightforward to run the entire process with a single call, ensuring ease of use and reproducibility. By structuring the code this way, the pipeline becomes flexible and reusable for insurance fraud estimation. This design allows for easy experimentation with different preprocessing steps, classifiers, or evaluation metrics, and simplifies extensions such as adding new classifiers or implementing cross-validation.

5.3 Stage 3 – Matching both the stages 1 & 2

The Insurance fraud and disaster declaration matching model involve a series of steps, from loading models and data to making predictions and matching records. This stage is one of the most important validation step for the Insurance industry, mainly for the risk management department.

Step 1: Import Libraries – We start by importing necessary libraries for data manipulation, machine learning, and visualization. We use pandas for data handling, numpy for numerical operations, matplotlib and seaborn for potential visualizations, and joblib for loading pre-trained models.

Step 2: Load Models and Data – The `load_models_and_data` function is responsible for loading the pre-trained models and the datasets. This function loads the pre-trained fraud detection and disaster classification models, as well as the insurance claims and disaster declaration datasets.

Step 3: Preprocess Data – The `preprocess_insurance_data` and `preprocess_disaster_data` functions handle the preprocessing of the respective datasets. These functions select the relevant features and perform one-hot encoding on categorical variables to convert them into numerical format suitable for machine learning models.

Step 4: Make Predictions – The `make_predictions` function uses the pre-trained models to make predictions on the preprocessed data. This function uses the fraud detection model to predict fraudulent insurance claims and the disaster classification model to predict disaster declarations.

Step 5: Create Prediction DataFrames – The `create_prediction_dataframes` function combines the original data with the predictions. This function creates new dataframes that include the predictions from both models, which will be used for further analysis.

Step 6: Extract and Add State Information – The `extract_state_columns` and `add_state_column` functions handle the extraction and addition of state information. These functions extract the one-hot encoded state columns and create a single 'state' column for easier matching.

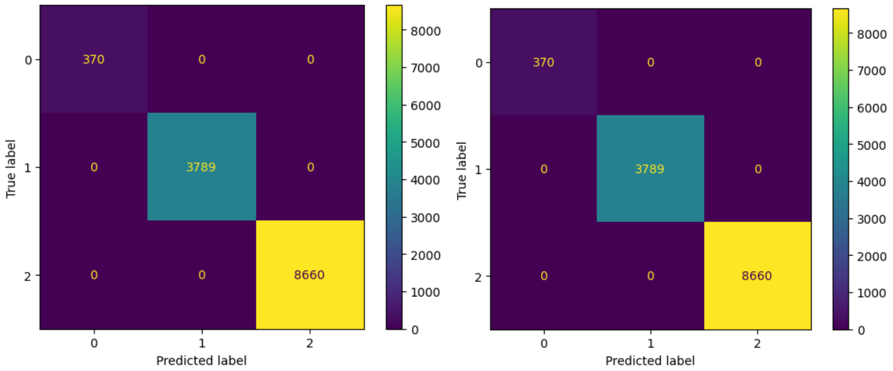
Step 7: Match Records – The `compare_states_dates` function matches insurance claims with disaster declarations based on state and year. This function iterates through the insurance and disaster dataframes, finding matches based on state and year (hardcoded to 2015 in this case).
Step 8: Main Function – Finally, the `main` function orchestrates the entire process. This function calls all the other functions in the correct order to load the data, preprocess it, make predictions, and find matches between insurance claims and disaster declarations.

6 Evaluation

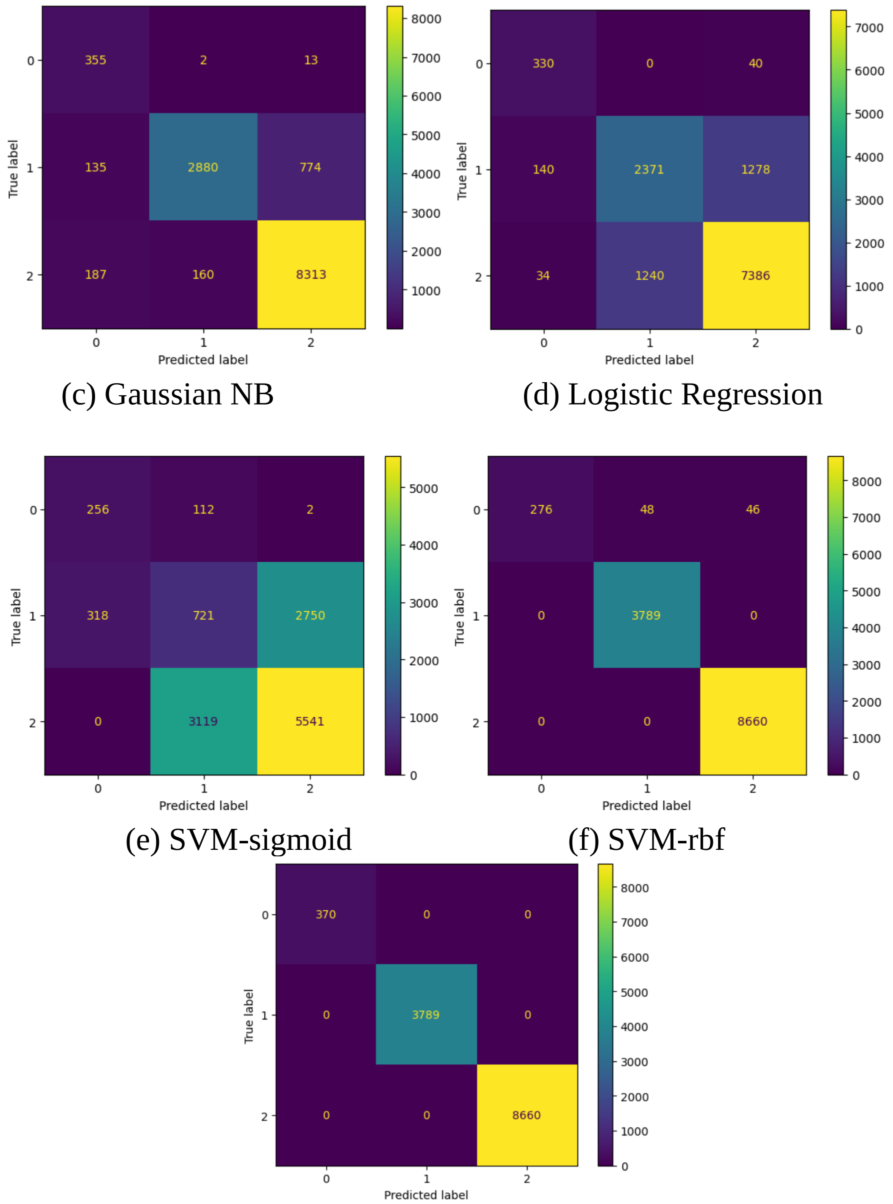
The entire coding structure can be broken down into the following steps. To use this script, you need to provide the correct file paths for the pre-trained models and datasets, then run the `main` function. The result is a dataframe of matched records, which could be further analyzed to investigate potential correlations between insurance fraud and disaster declarations. This implementation provides a flexible framework for matching insurance claims with disaster declarations, allowing for easy modifications or extensions to the matching criteria or analysis methods.

6.1 Disaster Severity Prediction - Case Study 1

For disaster declaration classification, the process involves loading the dataset, performing data cleaning and one-hot encoding, and separating features (X) from the target variable (y). The data is then split into training and testing sets using an 80-20 split. A set of classifiers, such as RandomForestClassifier and GradientBoostingClassifier, are defined and compared. The models are trained on the training data, predictions are made on the test data, and performance metrics such as accuracy, classification reports, and confusion matrices are generated.



(a) Random Forest (b) Gradient Boosting



(c) Gaussian NB

(d) Logistic Regression

(e) SVM-sigmoid

(f) SVM-rbf

(g) LGBM

Figure 6: Confusion matrix of the models.

The best models are found to be random forest, gradient boosting and light gradient boosting. Since the solution needs to be deployed in the servers as a backend response system, the computation and the space can one of the most important factor in deciding which models can be used for the deployment. In this situation, LGBM from its name can be found to be a light weighted model and can be used for the deployment. All the three classes was able to be predcted properly.

6.2 Insurance Fraud Detection - Case Study 2

XGBoost performed the best with 89.16% in the accuracy measure. For the Stratification over sampling SMOTE with Hyper parameters tuning using Grid Search, Decision Trees performance is as above. The accuracy is 0.81, precision as 0.81, recall as 0.83 and F1 score as 0.82. Out of 452 validation samples, 84 are missed classified and hence the precision is 81.14%, Recall is 83.9% and F1 score as 82.5%.

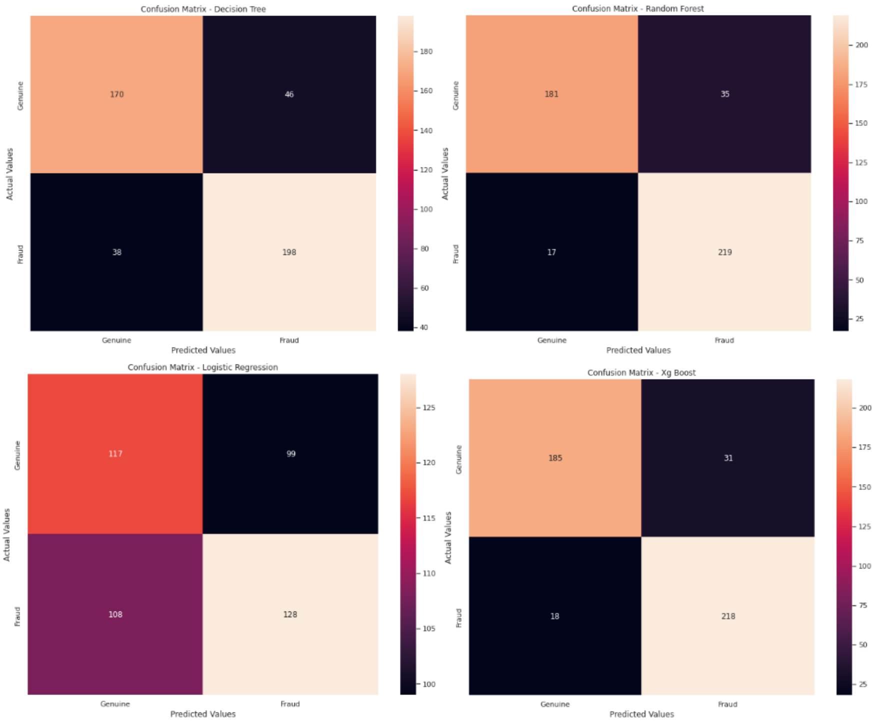


Figure 7: Confusion matrix of Decision Trees using the Stratification Over Sampling SMOTE and Hyper parameters Tuning using grid Search

For the Stratification over sampling SMOTE with Hyper parameters tuning using Grid Search, Random Forest performance is as above. Out of 452 validation samples, 52 are missed classified and hence the precision is 86.22%, Recall is 92.8% and F1 score as 84%. For the Stratification over sampling SMOTE with Hyper parameters tuning using Grid Search, Logistic Regression performance is as above. Out of 452 validation samples, 207 are missed classified and hence the precision is 56.4%, Recall is 54.23% and F1 score as 55.3%. For the Stratification over sampling SMOTE with Hyper parameters tuning using Grid Search, XGBoost performance is as above. The accuracy is 89.15%, Precision to be 87.55%, Recall being 92.37% and F1-score to be 89.89%.

6.3 Matching with Disaster with the Insurance - Case Study 3

Here we matched the states where the disaster took place and the year it took place and we streamlined the possibilities of the accidents that could have been affected by Disaster, which the insurance companies could avoid paying or reduce the claim amount. And those Insurances were saved in as a dataframe.

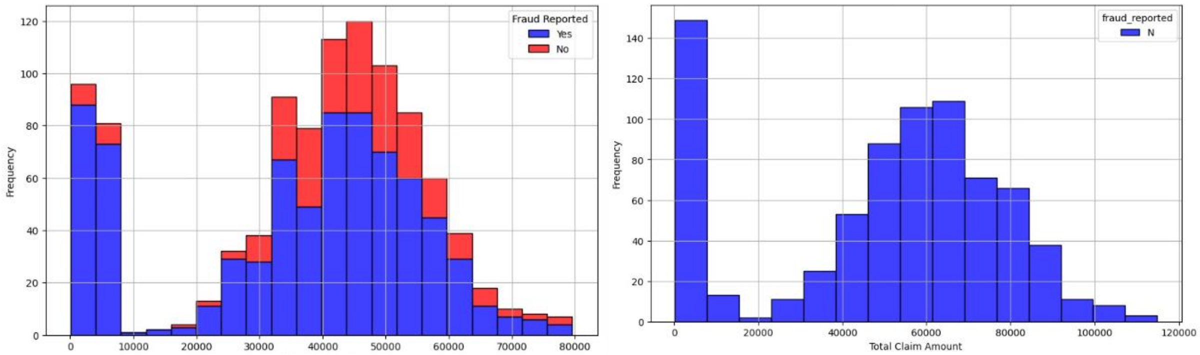


Figure 8: Total Claims and Total Claims amount pertaining to the claims reported

In this we found that the number of claims which has been reported as “Normal” which is one of the factor for the risk management to help in minimizing the loss for the Insurance company. Here only those claims are taking which are mostly falling for the natural disaster high probability scores. Now the total saving the company would have made was **$37M dollars**.

6.4 Discussion

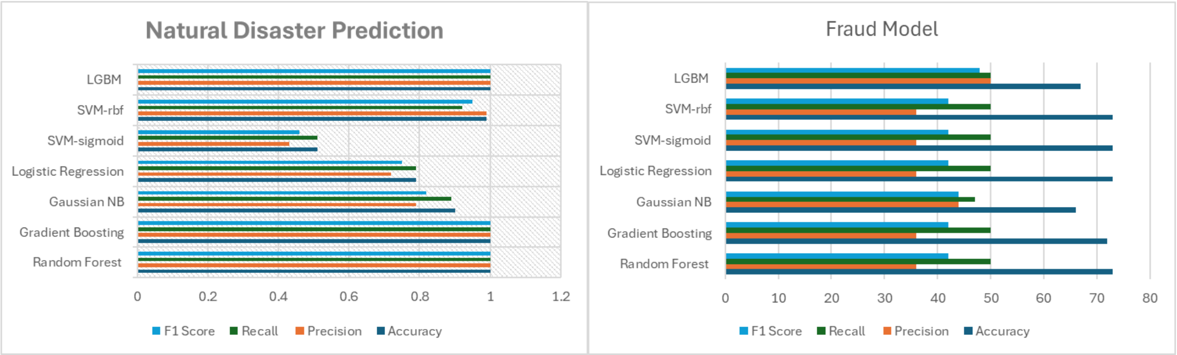


Figure 9: Comparison of the two models performance: (a) Using LGBM to forecast natural risk (b) Suspicious claims

When assessing models for backend deployment, Light Gradient Boosting Machine (LGBM) was found to be the most appropriate choice because of its lightweight design, which allows for efficient computation and storage. In order to detect insurance fraud, many classifiers were evaluated, such as Random Forest, Gradient Boosting, and Logistic Regression. Among them, Random Forest and Logistic Regression had the most effective performance overall. Nevertheless, LGBM exhibited the best accuracy, rendering it efficacious in forecasting dubious claims. However, the main emphasis is on precisely identifying claims that are not suspicious in order to avoid incorrect rejections. PyCaret improved model performance through hyperparameter adjustment, specifically demonstrating the usefulness of Logistic Regression in analyzing feature significance for decision boundaries.

The efficiency and speed of the Light Gradient Boosting Machine (LGBM) may be credited to its histogram-based algorithms, which optimize computation and memory utilization. LGBM is capable of effectively managing huge datasets and high-dimensional data. In addition, LGBM also provides support for parallel and distributed computing, which further improves its scalability and speed. The model's capacity to manage intricate linkages and interactions within the data, along with its strong ability to mitigate overfitting using regularization approaches, renders it a potent tool for tasks such as insurance fraud detection. These attributes enhance its exceptional accuracy and overall efficacy in forecasting results.

7 Conclusion and Future Work

To summarize, this study emphasizes the crucial importance of using modern machine learning to improve risk assessment and management in the Insurance sector. Insurers may enhance their capacity to forecast natural catastrophes and handle related risks by utilizing meteorological data and advanced algorithms like Random Forest, Gradient Boosting, and Light Gradient Boosting Machine (LGBM). The findings demonstrate that LGBM, specifically, excels in terms of both efficiency and accuracy, rendering it a desirable option for backend implementation in extensive applications. Moreover, the study showcases the capability of these models in identifying instances of insurance fraud, with XGBoost exhibiting exceptional performance in this field. By integrating these technologies, insurers may achieve more accurate risk assessment and implement proactive risk management techniques, leading to a decrease in financial losses. Subsequent investigations should prioritize the improvement of these models and the investigation of supplementary data sources to augment their prediction capacities. This technique will assist insurance firms in effectively managing the intricacies of natural catastrophe risks and detecting instances of fraud, so enhancing their ability to withstand challenges and improve their operational effectiveness. The bridging of these gaps in the future research will be significant in enhancing the operational practicality, and adaptability of these machine learning techniques in the insurance field.

References

Aljohani, F.H. *et al.* (2023) ‘A Smart Framework for Managing Natural Disasters Based on the IoT and ML’, *Applied Sciences*, 13(6), p. 3888. Available at: https://doi.org/10.3390/app13063888.

Balona, C. (2023) ‘ActuaryGPT: Applications of Large Language Models to Insurance and Actuarial Work’. Rochester, NY. Available at: https://doi.org/10.2139/ssrn.4543652.
Cui, P. *et al.* (2021) ‘Scientific challenges of research on natural hazards and disaster risk’, *Geography and Sustainability*, 2(3), pp. 216–223. Available at: https://doi.org/10.1016/j.geosus.2021.09.001.

Cui, P. *et al.* (2021) ‘Scientific challenges of research on natural hazards and disaster risk’, *Geography and Sustainability*, 2(3), pp. 216–223. Available at: https://doi.org/10.1016/j.geosus.2021.09.001.

Scholars@Duke publication: Climate-Informed Public Investment Management Diagnostic Framework; Climate Change Adaptation and Mitigation (no date). Available at: https://scholars.duke.edu/publication/1501687

Hayder, I.M. *et al.* (2023) ‘An Intelligent Early Flood Forecasting and Prediction Leveraging Machine and Deep Learning Algorithms with Advanced Alert System’, *Processes*, 11(2), p. 481. Available at: https://doi.org/10.3390/pr11020481.

Hussein, A. and Zoghlami, F. (2023) ‘The Role of Engineering Insurance in Completing Projects by Using Bank Loans:An Applied Study in a Sample of Iraqi Insurance Companies and Banks’, *International Journal of Professional Business Review*, 8, p. e0926. Available at: https://doi.org/10.26668/businessreview/2023.v8i1.926.

Jena, R. *et al.* (2020) ‘Seismic hazard and risk assessment: a review of state-of-the-art traditional and GIS models’, *Arabian Journal of Geosciences*, 13(2), p. 50. Available at: https://doi.org/10.1007/s12517-019-5012-x.

Kalfin *et al.* (2022) ‘Insurance as an Alternative for Sustainable Economic Recovery after Natural Disasters: A Systematic Literature Review’, *Sustainability*, 14(7), p. 4349. Available at: https://doi.org/10.3390/su14074349.

Kunreuther, H. (1996) ‘Mitigating disaster losses through insurance’, *Journal of Risk and Uncertainty*, 12(2), pp. 171–187. Available at: https://doi.org/10.1007/BF00055792.
Li, X., Jiang, Y. and Mostafavi, A. (2023) *AI-assisted Protective Action: Study of ChatGPT as an Information Source for a Population Facing Climate Hazards*. Available at: https://doi.org/10.2139/ssrn.4408290.

Linardos, V. *et al.* (2022) ‘Machine Learning in Disaster Management: Recent Developments in Methods and Applications’, *Machine Learning and Knowledge Extraction*, 4(2), pp. 446–473. Available at: https://doi.org/10.3390/make4020020.

Sheehan, B. *et al.* (2023) ‘On the benefits of insurance and disaster risk management integration for improved climate-related natural catastrophe resilience’, *Environment Systems and Decisions*, 43(4), pp. 639–648. Available at: https://doi.org/10.1007/s10669-023-09929-8.

Sreelakshmi, S. and Chandra, V. (2022) ‘Machine Learning for Disaster Management: Insights from past research and future implications’, in, pp. 1–7. Available at: https://doi.org/10.1109/IC3SIS54991.2022.9885494.

Taha, A.A. and Malebary, S.J. (2020) ‘An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine’, *IEEE Access*, 8, pp. 25579–25587. Available at: https://doi.org/10.1109/ACCESS.2020.2971354.

Thabtah, F. *et al.* (2020) ‘Data imbalance in classification: Experimental evaluation’, *Inf. Sci.*, 513(C), pp. 429–441. Available at: https://doi.org/10.1016/j.ins.2019.11.004.

Vinod, A. *et al.* (2022) ‘Natural Disaster Prediction by Using Image Based Deep Learning and Machine Learning’, in, pp. 56–66. Available at: https://doi.org/10.1007/978-3-030-84760-9_6.

Ward, P.J. *et al.* (2020) ‘Review article: Natural hazard risk assessments at the global scale’, *Natural Hazards and Earth System Sciences*, 20(4), pp. 1069–1096. Available at: https://doi.org/10.5194/nhess-20-1069-2020.

Wu, L., Ma, D. and Li, J. (2023) ‘Assessment of the Regional Vulnerability to Natural Disasters in China Based on DEA Model’, *Sustainability*, 15(14), p. 10936. Available at: https://doi.org/10.3390/su151410936.

Zuccaro, G., Leone, M.F. and Martucci, C. (2020) 'Future research and innovation priorities in the field of natural hazards, disaster risk reduction, disaster risk management and climate change adaptation: a shared vision from the ESPREssO project', *International Journal of Disaster Risk Reduction*, 51, p. 101783. Available at: https://doi.org/10.1016/j.ijdrr.2020.101783.