

INSURANCE CLAIMS PREDICTION USING MACHINE
LEARNING MODELS

BY

ENOCH CHUKWUEBUKA JEREMIAH

210202013

A PROJECT SUBMITTED TO THE
DEPARTMENT OF COMPUTER SCIENCE,
COLLEGE OF COMPUTING,
McPHERSON UNIVERSITY, SERIKI SOTAYO,
OGUN STATE, NIGERIA.

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE
AWARD OF BACHELOR OF SCIENCE (B.Sc) DEGREE IN
COMPUTER SCIENCE.

JULY, 2025.

CERTIFICATION

This is to certify that I ENOCH CHUKWUEBUKA JEREMIAH with matriculation number 210202013 of the Department of Computer Science, College of Computing, McPherson University, Seriki Sotayo, Ogun State carried out this project titled INSURANCE CLAIMS PREDICTION USING MACHINE LEARNING MODELS in partial fulfilment of the requirement for the award of Bachelor of Science in Computer Science of McPherson University, Ogun State, Nigeria.

Mrs Mercy Adebisi

Project Co-Supervisor's Name

Date

Dr. Emmanuel Ibam

Project Supervisor's Name

Date

Dr. Kayode Oladapo

HOD's Name

Date

DEDICATION

Firstly, I want to dedicate this project to God for His loving kindness and tendered mercies, and for sustaining me through the period of my schooling. I am also dedicating this project to my parents, Elder Enoch Ugwu and Elder Uche Ugwu and my brother Enoch Kenechukwu for their never-ending love, care, and support throughout my academic journey.

ACKNOWLEDGEMENTS

Special thanks to my project Supervisor Dr. Emmanuel Ibam for his relentless support and guidance through the journey of my project, I sincerely do not take it for granted.

I also want to acknowledge Mrs Mercy Adebisi, my project's co-supervisor for her support right from the maiden stage of this idea up to this very point, God bless you ma.

Further more, I also want to acknowledge the support and guidance of the lecturers in the college of computing and more specifically, the Department of Computer Science.

A token of appreciation goes to Mr Chukwuebuka Arueze, for his mentorship and guidance through my journey as I embarked on this project.

ABSTRACT

Insurance firms have the tendency to experience disruption in operational activities within a fiscal year due to improper resource management and the challenges around understanding the dynamics of claims which is an important aspect of this industry can be very overwhelming as the basis of a claim is unforeseen circumstances. This project suggests a web-based pipeline of machine learning models designed to address this issue using structured claims historical data. By implementing a range of classification and regression models including, Linear Regression, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost as well as a custom hybrid model combining the predictive strength of Random forest and XGBoost, this system is empowered to identify complex patterns in claims datasets and relationship between features to speculate claim values and possible categories.

Important techniques like data pre-processing, feature encoding and leakage prevention were applied to ensure the data is ready for prediction, while various performance evaluation metrics like accuracy, precision, recall, f1 score, mean absolute error, mean squared error and R-squared were implemented to assess the performance of the various models. To ensure transparency of predictions, model interpretability tools such as SHAP and LIME were integrated, helping users with the clear insights needed to understand the dynamics behind the prediction. The entire solution was deployed as a web-based application using Flask and hosted on Google Cloud, enabling users to upload their datasets and receive predictions in real time.

TABLE OF CONTENTS

TITLE PAGE	I
CERTIFICATION	II
DEDICATION	III
ACKNOWLEDGEMENTS	IV
ABSTRACT	V
TABLE OF CONTENTS	VI
LIST OF TABLES	VIII
LIST OF FIGURES	IX
CHAPTER ONE	1
INTRODUCTION	1
1.1 Background of the Study	1
1.2 Statement of Problem	4
1.3 Aim and Objectives	5
1.4 Methodology Overview	5
1.4.1 Research Design	5
1.4.2 Data Collection	6
1.4.3 Data Pre-processing	6
1.4.4 Model Development	6
1.4.5 Model Evaluation	7
1.4.6 Web Application Development	7
1.5 Scope of Study	8
1.6 Significance of Study	8
1.7 Organization of Work	8
CHAPTER TWO	10
LITERATURE REVIEW	10
2.1 Introduction	10
2.2 Review of Related Works	10
CHAPTER THREE	19
METHODOLOGY	19
3.1 Introduction	19
3.2 System Architecture	20
3.3 System Development	25

3.4 Tools and Frameworks Used	26
CHAPTER FOUR	27
IMPLEMENTATION	27
4.1 Introduction	27
4.2 System Setup and Environment	27
4.3 Testing the Prediction System	28
4.4 Model Implementation and Evaluation Using Claims Datasets.	31
CHAPTER FIVE	46
CONCLUSION AND RECOMMENDATION	46
5.1 Summary	46
5.2 Conclusion	47
5.3 Recommendation	47
5.4 Future work	47
REFERENCES	48

LIST OF TABLES

Table 1 Representation of Tools and Frameworks Used.....	28
--	----

LIST OF FIGURES

Figure 1 : Flowchart Representing Prediction Pipeline	20
Figure 2 : Summary Information on Dataset	31
Figure 3 : Attached Dataset and Indication of Target Column	32
Figure 4 : Data Processing Message Displayed	33
Figure 5 : Prediction Results and Best Model.	34
Figure 6 : Various Models and Their Performance Evaluation Metrics.	35
Figure 7 : Feature Importance Plot and Downloadable CSV File	36
Figure 8 : SHAP Explanations Plot Based on Best Model	37
Figure 9 : LIME Explanations and Downloadable CSV file	38
Figure 10 : LIME Explanation Values	39
Figure 11 : Information Summary on Dataset 2	40
Figure 12 : Prediction Sample Results Classification Task.	41
Figure 13 : Various Classification Models with Performance Evaluation Metrics.	42
Figure 14 : Feature Importance Plot Based Classification	43
Figure 15 : SHAP Explanations Plot Based on Best Model	44
Figure 16 : LIME Explanation Values for a Field in The NCID Dataset.	45

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Insurance is a form of financial agreement that makes provision for resources on the account of a potential future monetary loss. Insurance is a situation where an individual swaps an unsure monetary loss (risk) in exchange for a loss he/she is sure of (premium) (Pouffinas et al., 2023). Premium is an amount of money the insurer charges the insured and it should reflect the degree of risk, by implication, the higher the premium, the higher the risk (Xiaonan, 2023). An individual or entity in this case referred to as the insured, comes to a registered financial organization referred to as the insurer that specialises in this type of financial service on the basis of an agreement called the policy, with a sum of money called the premium whose frequency is also indicated in the policy to stand in place of a possible financial loss (David, 2015). An important challenge for the insurance sector is the correctly addressing the relationship between risks and premiums (Edossa, 2023). The primary aim of insurance firms is to be able to decide the right insurance premium for the insured to cover the agreed risk (Sun et al., 2024). The ideal method for premium analysis is by identifying the characteristics of the risk and then multiplying by the expected claim frequency (David, 2015). Performing accurate risk measurement and management is important to effectively manage claims and maintain financial stability (Hanafy & Ming, 2021).

As technological innovations advance, one of the objectives of science is to influence every area of life. Artificial intelligence is one breakthrough of technology whose aim is to develop machines that can be able to think and make decisions just like humans, though the entire concept is progressively improving daily (Hanafy & Ming, 2021). Machine learning is an aspect of Artificial Intelligence that explains the concept of the ability for computers or

machines learn and make predictions or decisions without direct programming, the performance of these machine increase over time as they as explored because the cycle of their learning increases with use (Poufinas et al., 2023).The cycle of operations in the insurance sector is hinged around claims. A claim is a request made by the insurer to the insurer on the premise of the agreement stated in the policy. The subject claims is an important insurance activity because they are the basis of the trade of the insurance product (Poufinas et al. 2023). In the insurance sector, the importance of predicting future claims cannot be overemphasized, as it helps keeps the insurance organization in business (Abdelhadi et al., 2020). Insurance companies should be able to speculate the total number of claims and how intense these claims can be in order to enable them set a reasonable cost for the product in their policy agreement (Nolan, 2024). The rapid growth in the frequency of claims and their severity have raised a need for more advanced methods for claims predictions (Burri et al., 2019). Traditional mathematical concepts like statistics have been employed in the insurance sector for claims predictions but these methods have fallen short of the ability the ability to accurately communicate the message of complex relationships and patterns in the insurance sector, asides that, vast datasets would be overwhelming for these traditional models (Segura-Gisbert et al., 2025). The traditional methods for forecasting claims have a tendency to repeat the patterns of claims in historical data (Paul, 2024). Machine learning algorithms have significantly improved the accuracy of predictions in-turn helping the sector in their operations (David, 2015). Machine learning models have the capacity to analyse previous claims data in large scales to identify the trends and speculate future patterns based on existent ones (So, 2024). Research carried out by Smith et al (2018) in the United States showed that when predictive models were implemented in the insurance sector, they reduced the errors of claims processing errors by 15% in a space of two years. Accurately predicting future claims will amongst many benefit help insurers in premium

adjustment and optimization. One relevant aim of insurance firms is their ability to determine the right insurance premium that the insured need to pay to cover a particular risk (Seng, Shin and Liu 2024). It is only in the benefit of insurance firms to be able to speculate claims ahead of their occurrence (Pouffinas et al. (2023)). Choosing the best model to perform predictions is very important. The cycle of operations in the insurance sector is hinged around claims. A claim is a request made by the insured to the insurer on the premise of the agreement stated in the policy. The subject claims is an important insurance activity because they are the basis of the trade of the insurance product (Pouffinas et al. 2023). In the insurance sector, the importance of predicting future claims cannot be overemphasized, as it helps keep the insurance organization in business (Abdelhadi et al., 2020). Insurance companies should be able to speculate the total number of claims and how intense these claims can be to enable them to set a reasonable cost for the product in their policy agreement (Nolan, 2024). The rapid growth in the frequency of claims and their severity have raised a need for more advanced methods for claims predictions (Burri et al., 2019). Traditional mathematical concepts like statistics have been employed in the insurance sector for claims predictions but these methods have fallen short of the ability to accurately communicate the message of complex relationships and patterns in the insurance sector, besides that, vast datasets would be overwhelming for these traditional models (Segura-Gisbert et al., 2025). The traditional methods for forecasting claims tend to repeat the patterns of claims in historical data (Paul, 2024). Machine learning algorithms have significantly improved the accuracy of predictions in turn helping the sector in their operations (David, 2015). Machine learning models have the capacity to analyse previous claims data in large scales to identify the trends and speculate future patterns based on existent ones (So, 2024). Research carried out by Smith et al (2018) in the United States showed that when predictive models were implemented in the insurance sector, they reduced the errors of claims processing errors by

15% in a space of two years. Accurately predicting future claims will amongst many benefit help insurers in premium adjustment and optimization. One relevant aim of insurance firms is their ability to determine the right insurance premium that the insured need to pay to cover a particular risk (Seng, Shin and Liu 2024). It is only in the benefit of insurance firms to be able to speculate claims ahead of their occurrence (Poufinas et al. (2023)). Choosing the best model to perform predictions is very important.

1.2 Statement of Problem

Accurately predicting insurance claims is very vital with respect to efficient risk management, premium analysis and the entire stability of finances of the insurance sector. However, traditional statistical models which have been used over time for claims prediction have been found with the inability to address the complexity and variety of recent insurance datasets. These traditional methods find it difficult capturing non-linear relationships, highly dimensioned data and transiting trends in the behaviour of claims.

It is on this basis that insurance firms experience significant challenges in accurately speculating future claims leading to potential under or over pricing of premiums or misallocation of resources. When these circumstances occur, it would affect both the profitability of the firms and its long-term sustainability.

Machine learning (ML) models have been found to possess the features capable of addressing the challenges posed by implementing traditional statistical methods, identifying complex trends in-turn improving the accuracy of predictions. Irrespective the present potential of ML models, it still possesses present hindrances for example selecting the models to implement, the relevance of features in predictions, the quality of data and explaining the outcomes of predictions.

This project intends to bridge these gaps by implementing various machine learning models

for insurance claims prediction, performance evaluation and overall develop a system to perform insurance claims prediction automatically.

1.3 Aim and Objectives

The aim of this project is to develop a web-based system that implements machine learning models for claims prediction in real time.

Furthermore, the objectives are as follows:

- a. To develop models with optimised ability to predict claims using features in insurance datasets.
- b. To evaluate the predictive strength of the models using performance metrics.
- c. To integrate objectives (a and b) into a web-based application, providing predictions and model comparison in real time.

1.4 Methodology Overview

This project implements six machine learning models, five models each for classification and regression tasks and one hybrid model. In order to achieve the above stated objectives, machine learning models will be evaluated to justify their strengths and weaknesses using accuracy and explainability metrics.

The following section will provide a summary of the pipeline process

1.4.1 Research Design

A comprehensive review of related works, their findings and limitations were carefully analysed, it is on this premise that this project was commenced. This project adopts an applied research design focused on implementing predictive analytics using supervised machine learning methods. The project entails both classification and regression tasks for the

prediction of insurance claims and developing a web-based application for practical implementation.

1.4.2 Data Collection

- a. This project implements publicly available insurance claims datasets.
- b. Data features include policyholder characteristics, accident characteristics, claim details and claim outcomes.

1.4.3 Data Pre-processing

- a. Data Cleaning: Handling missing values, duplicates and inconsistent data objects
- b. Feature engineering: Creating new features, encoding categorical variables and scaling numerical features.
- c. Data Splitting: Dividing the claims dataset into training and testing (70-30 split)

1.4.4 Model Development

- a. Regression Machine Learning models
 - i. Linear Regression
 - ii. Decision Tree
 - iii. XGboost
 - iv. Random Forest Regressor
 - v. Gradient Boosting
- b. Classification Machine Learning Models
 - i. Logistic Regression
 - ii. Decision Tree
 - iii. XGboost Classifier

iv. Random Forest Classifier

v. Decision Tree Classifier

C. Hybrid Model (Random Forest + XGBoost)

1.4.5 Model Evaluation

a. Evaluation Metrics:

i. Classification Metrics: Accuracy, Precision, Recall, and F1-Score.

ii. Regression Accuracy: MAE, RMSE, R-Squared Score.

b. Analysis of Feature Importance: Identification of key players influencing predictions using LIME and SHAP.

1.4.6 Web Application Development

a. Backend:

i. Developed using Flask (Python).

ii. Allows for implementation of a variety of models.

iii. Ability to detect the type of prediction (classification or regression)

b. Frontend:

i. HTML, CSS

ii. Data submission

iii. Prediction results

iv. Model Comparison

v. Data visualization

c. Integration:

i. RESTful API connects front and back ends together.

ii. Results are rendered in real-time available for user interaction.

d. Deployment

- i. The web application was deployed on google cloud hosting server.

e. Tools and Technologies

- i. Python (pandas, NumPy, scikit-learn, XGBoost, Matplotlib, seaborn)
- ii. Flask (Backend Framework)
- iii. HTML and CSS (Frontend Development)
- iv. Deployment Platform (Google Cloud)

1.5 Scope of Study

The scope of this study covers the insurance sector and its claim activities.

1.6 Significance of Study

- I. It helps in the management and evaluation of claims in the insurance sector.
- II. It helps to improve trust in predictions through model interpretability tools.
- III. It shows the practical application of the roles machine learning models can play in the insurance industry.

1.7 Organization of Work

Chapter One introduces the concept of the project, defining the basic terms and explaining the need for the project to be carried out.

Chapter Two contains a review of various literatures, papers that relate to this project in methodology or area of speciality were identified and summarised, their methodologies, findings and limitations were summarised in this chapter.

Chapter Three contains the methodology that was employed in carrying out the project, the various components and their interdependence to achieve the overall project.

Chapter Four contains the results obtained from testing the claims prediction web-based application, various datasets were tested and their results confirmed to ensure all functionalities are in operation.

Chapter Five the concluding chapter contains the recommendation and the future works on the project.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

The insurance industry experienced an unprecedented revolution with the integration of Machine Learning (ML) (Poufinas et al. (2023)). In this review, a collection of work in this field is being reviewed, applications of machine learning around claims management are analysed with its pros and cons. In terms of work between 2015 and 2025, a speedy development in these technology forms in terms of its application in insurance is seen in the review below.

2.2 Review of Related Works

Abdelhadi et al., (2020) focused on proposing a machine learning based methodology to predict the probability of auto insurance claims. It focuses on using real world auto insurance datasets and explores different classification models. The models implemented in this paper are Random Forest, Decision Trees, Naive Bayes and Support vector machines. It was discovered that XGBoost and Decision Trees achieved the top accuracy scores of 92.53% and 92.2% respectively. The paper attended to missing data using various imputations which increased the performance of the models and identified critical features relevant for the high accuracy predictions like driving history. The limitations identified in this research is that it did not implement ensemble or hybrid models, did not implement model interpretability techniques, did not deploy the model for real time predictions and cannot be used for model generalization. How this project addresses the imitations identified is by implementing hybrid model (Random Forest and Xgboost), implementing model interpretability using SHAP and LIME, Web-app deployment for real time prediction and handling leakage columns explicitly.

Burri et al., (2021) presents an empirical study on insurance claim prediction using several traditional and ensemble machine learning algorithms. The core goal is to analyse historical insurance data to predict whether a claim will occur (a classification task) and evaluate model performance. The models implemented in this paper were Decision tree, Random Forest, Naive Bayes, Support Vector Machine (SVM) and Logistic Regression. They also evaluated model performance using evaluation metrics like Accuracy, Precision, Recall and F1 Score. Data preprocessing methods like Feature encoding, missing value imputation and feature scaling were also carried out. Some of the limitations discovered were that it lacked model interpretability, no model deployment for real time prediction, no hybrid model implementation, no explicit addressing of leakage columns and it focuses on just classification tasks. My project addresses these limitations through Application of both regression and classification tasks Hybrid model approach present, Implementation of Interpretability metrics and Model deployment for real time prediction.

Mohamed and Ruixing (2021) proposed a variety of approaches for machine learning for processing and managing big-data insurance information. Specifically, processing complexity and high volumes of new-facets-bearing datasets in new insurance received specific consideration. Neural networks with 3 hidden layers, ReLU activation functions, Real-time data streaming integration and Ensemble methods combining multiple models were explored in this research. After implementing the methodology a 40% faster processing time compared to traditional methods, 85% accuracy in claim amount prediction were discovered and Successful handling of 1M+ insurance records with a reduced false positives by 32%. While reviewing this research it was discovered that it lacks model interpretability, the research did not explicitly handle data leakage columns, there was no hybrid model experimentation, only traditional metrics like accuracy and precision was used and the solution is purely technical and not user centered. This project addresses the limitations of the research through the use of

enhanced explainability metrics like SHAP and LIME for enhanced model interpretability, prevention of leakage columns from affecting predictions, the development of web-based system for both technical and non-technical users, the use of broader classification evaluation like F1 and Recall, the use of a methodology that explicitly explains all processes and the implementation of a Hybrid Model.

Henckaerts et al., (2022) explores how tree-based machine learning models—particularly Gradient Boosting Machines (GBM) and Random Forests (RF)—can enhance the pricing and risk modelling of insurance tariff plans, moving beyond the limitations of classical generalized linear models (GLMs). Gradient Boosting Machines and Random Forest were the models implemented in this research. Random forest demonstrated increased predictive performance and interpretability using variable importance and partial dependence plots. The limitations identified in this research is that only variable dependence and partial dependence plots are used for model interpretability, it focuses on claims modelling in one insurance line, focuses on Acquirers and insurance modelers, data includes tariff plan design and pricing insights and not claims and carried out research without deployment. This project addresses the limitations through interactive model interpretation, deployment of web-based prediction model for real time prediction, hybrid model performance comparison and practical application of claims prediction and not just a theoretical benchmark.

Sun et al., (2023) proposed a Transformer-based deep learning model for evaluating risk in auto insurance. It focuses on policyholders' risk assessment, combining the power of sequence modelling with model interpretability using SHAP. This paper used Actuarial Transformer (AT): Combined the transformer architecture and tree-based model with SHAP to aid model interpretability. After implementation it was discovered that AT performed better than the traditional models in risk prediction,

SHAP analysis identified Bonus-Malus scores (the history of the drivers) as a major feature

and shows the trade-off between accuracy and interpretability in deep learning models. The limitations identified in this research were high resource requirements, Transformer models remain harder to interpret than tree-based models, There was no deployment framework for real time prediction and performance greatly relies on access to sequential data. This project addresses the limitations of the research through improved model interpretability, Web-application deployment, Faster execution and lesser computational requirements and bulk evaluation using policy holders history.

Clemente et al., (2023) explored the application of new implementations of gradient boosting techniques in model development for claim count and claim cost work showed a significant improvement over statistical technique. This research implements a two-stage modelling approach: Frequency modelling using Poisson distribution and Severity modelling using Gamma distribution. A 25% reduction in prediction error compared to GLMs, identification of non-linear relationships in risk factors and model showing robust performance across different policy types were discovered from the implementation. The limitations identified from this research is that it focuses on only modelling frequency, implements only Gradient Boosting algorithms, lacks interpretability tools and is purely statistical analysis done in R without practical deployment. This project handles the limitations identified in the research by handling both classification and regression tasks, implementing of a wide variety of machine learning models for better predictive performance using enhanced explainability metrics like SHAP and LIME for enhanced model interpretability and the development of web-based system for both technical and non-technical users..

Pouffinas et al., (2023) compared a variety of algorithms for claims forecasting in a thoughtful evaluation of model selection and performance in an actuarial scenario.

This research compares multiple algorithms like the neural networks, random forest, gradient boosting, long-short term model networks, Time series model and cross-validation feature

engineering for temporal patterns Ensemble method combination. After implementation it was discovered that LSTM showed best performance for long-term predictions, Random Forests excelled in short-term forecasting and a 28% improvement over traditional forecasting methods. The limitations identified in this research is that it used limited evaluation metrics that limits the understanding of full model performance, there was no deployment framework for real time prediction, the study did not explore explainability metrics like SHAP and LIME, little attention was given to leakage columns that could create a bias. After identifying the limitations, this project addresses the limitations by exploring multiple machine learning models, implementing a wider range of performance evaluation metrics, web deployment for real time prediction, addressing model interpretability using SHAP and LIME for better prediction explanation and precisely handling leakage columns in data preprocessing.

Abdulkadir and Fernando (2023) proposes a deep learning-based solution to predict insurance claim outcomes (severity or likelihood). It focuses on addressing the complexity of patterns in claim data, which may be difficult for traditional machine learning models to capture. The research implemented deep neural network using TensorFlow/karas with a structured insurance dataset, categorical encoding, normalization and dropout layers to avoid overfitting and fully concerned with neural networks. The methodology improved the prediction of insurance claims. Deep learning models are usually harder to interpret when compared to other machine learning models. The limitations identified were that DNN's required higher computational resources to train the model, the results over simpler models aren't generally large and there was no implementation of the model for real time prediction. This project addresses the limitations of the paper by implementing lighter models that do not require high computational resources, implementing of SHAP and LIME for model interpretability and deployment inclusion for real time prediction.

Groen (2023) explores the use of machine learning techniques to optimize insurance premium pricing, based on real-world data from MS Amlin. The focus is on improving the accuracy of premium predictions, typically determined via traditional actuarial models. The models implemented are XGboost, Random Forest, Gradient Boosting, Elastic Net and Traditional GLM's as baseline. The limitations identified in this research is that it lacks integration of SHAP and LIME or any other interpretability framework, there is no information about real-world deployment or an automation pipeline, presence of potential data leakages were not properly analysed, it focuses on only regression tasks and limited model interpretability. This project addresses the limitations through deployment of model for real time prediction, explicit addressing of leakage columns to avoid model bias and implementation of both regression and classification tasks.

Li (2023). This paper investigates multiple machine learning models to identify the most effective algorithm for predicting whether an insurance claim will be made (classification) in the auto insurance domain. The emphasis is on model comparison using advanced evaluation strategies. The methodology used were Logistic Regression, Random Forest, Gradient Boosting, XGBoost, Support Vector Machine (SVM), K-Nearest Neighbors, performance evaluation metrics, Feature encoding, missing value imputation and feature scaling. The limitations identified in this research is that it did not implement interpretability tools, there was no model deployment, strictly experimental, model evaluation was independent, did not explicitly handle leakage columns and only one dataset was used. This project addresses the limitations through hybrid modelling, implementing model interpretability, model deployment for real time predictions and handling leakage columns that could cause model bias.

Banghee (2024) work deals with zero-inflation in claims, an intractable problem in insurance, and high-powered gradient boosting techniques. CatBoost, XGBoost, and LightGBM

comparisons generate useful information for practitioners. The researcher performed a comparative analysis of three boosting algorithms namely CatBoost, XGBoost and LightGBM. The implementation also used zero-inflated insurance claims data, implemented hyper-parameter tuning via grid search, applied 5-fold cross-validation and evaluated model performance using metrics like RMSE, MAE, and R-squared. It was discovered that CatBoost outperformed other algorithms with: 15% lower RMSE and 12% better prediction accuracy. LightGBM showed fastest training time, XGBoost demonstrated better handling of categorical variables and models interpretability was better with CatBoost. The limitations identified from this research is that it focuses only on Zero-inflated Regression Only, Implemented only Tree Boosting Models, Lacks interpretability tools, did not implement for real time predictions and lacks treatment of leakage columns. This project addresses the limitations in the research by covering both regression and classification tasks, implementing a diverse range of models not just tree boosting models, use of explainability metrics like SHAP and LIME and the development of a web-based system for real time prediction.

Joel (2024) discussed a future view of AI use in insurance is, with a review of improvements in price-setting and claims processing via automation. This paper reviewed the use of deep Learning for pattern recognition, natural language processing for document analysis and computer vision for damage assessment. A 60% reduction in claims processing time, 45% improvement in fraud detection, 90% accuracy in damage assessment and a significant cost reduction in claims handling was discovered. The limitations identified from this research is that it gives only a conceptual overview of the potential of AI in pricing and claims lacking technical implementation, lacks model comparison or benchmarking performance evaluation, lacks essential aspects like interpretability or ethical considerations, did not implement framework for deployment and lacks exploration of leakage columns. This project addresses

the identified limitations by exploring of both classification and regression tasks, implementing diverse machine learning models, integrating SHAP and LIME to aid local and instance-based interpretability to aid prediction explanation, developing a web-based system for real time prediction and the removal of leakage columns that could affect prediction.

Nolan (2024) explored the use of machine learning models to improve insurance underwriting, having its focus on risk assessment, policy approval and premium pricing. The paper implements four machine learning models namely Decision Trees, Logistic Regression, Support Vector Machines (SVM's) and Gradient Boosting. The research figured that Machine Learning Models increased the accuracy of underwriting by capturing complex risk factors for example telematics data and NLP for the analysis of texts, that Ensemble Methods for example Random Forests, Gradient Boosting performed better than single models in risk predictions. An improved accuracy with XGboost in predicting claim likelihood was also discovered. The limitations identified from this research is that it lacks discussion on model interpretability, the dataset researched on was sourced from a single regional insurer which may limit generalizability, focuses on underwriting and not post-policy events like claims processing or fraud detection and did not implement ensemble/hybrid techniques, just individual comparison among models. This project addresses the limitations identified from the research by focusing on claim prediction in real time and historical, implementing ensemble methods like Random Forest and XGBoost, implementing model interpretability using SHAP and LIME and Web-based deployment for real time predictions.

Mihaela (2025) started a work in leveraging statistics in automotive premium computation work. According to the research car insurance premiums could accurately be computed with generalized forms of machines, opening doors for complex forms of machines with statistics. The research implemented generalised linear models, focuses on optimizing premium values and performing statistical validation through cross validation. During the research a 20%

improvement in the accuracy of premiums was discovered, it identified the crucial factors that affect risks and established a background for further machine learning applications. The limitations discovered from this research is that no system was implemented for real time prediction, only analysis, the research did not explore a variety of Machine Learning Models, the reseach implemented GLM's which assume linear relationship between features and require manual feature engineering to account for interaction, it also did not explore model explainability and the research focuses on just regression tasks. This project addresses the limitations of the reserch by implementing machine learning models like Random Forest and XGboost that can capture complex relationships between features, use of ensemble and hybrid models can learn interactions automatically, implementing enhanced explainability metrics like SHAP and LIME for enhanced model interpretability, deploying a Web based system for real time predictions and exploring both regression and classification tasks.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

The third chapter of this report describes the methodological approach that has been used in the design, testing, implementation, and deployment of the Insurance Claims Prediction System. The goal of this project as stated in the maiden chapter is to develop a system that can predict regression or classification claims using machine learning models. The methodology entails all processes involved in the system development lifecycle.

The approach followed in this project work is a data-driven experimental methodology, bringing together both theoretical and practical frameworks to account for accuracy, transparency, and scalability. The pipeline includes:

- a. Cleaning and transforming raw insurance claims data.
- b. Methods to prevent data leakage which may affect the accuracy of prediction, excluding features that could cause a bias in the prediction.
- c. Identifying and selecting suitable machine learning algorithms based on the type of target variable which may be regression or classification.
- d. Implementing a hybrid model, a combination of Random Forest and XGBoost for better prediction.
- e. Implementing model interpretability tools like SHAP and LIME to improve the transparency and explainability of predictions.
- f. Developing a web-based application using Flask and deploying it on Google Cloud to ensure accessibility and maintainability.

3.2 System Architecture

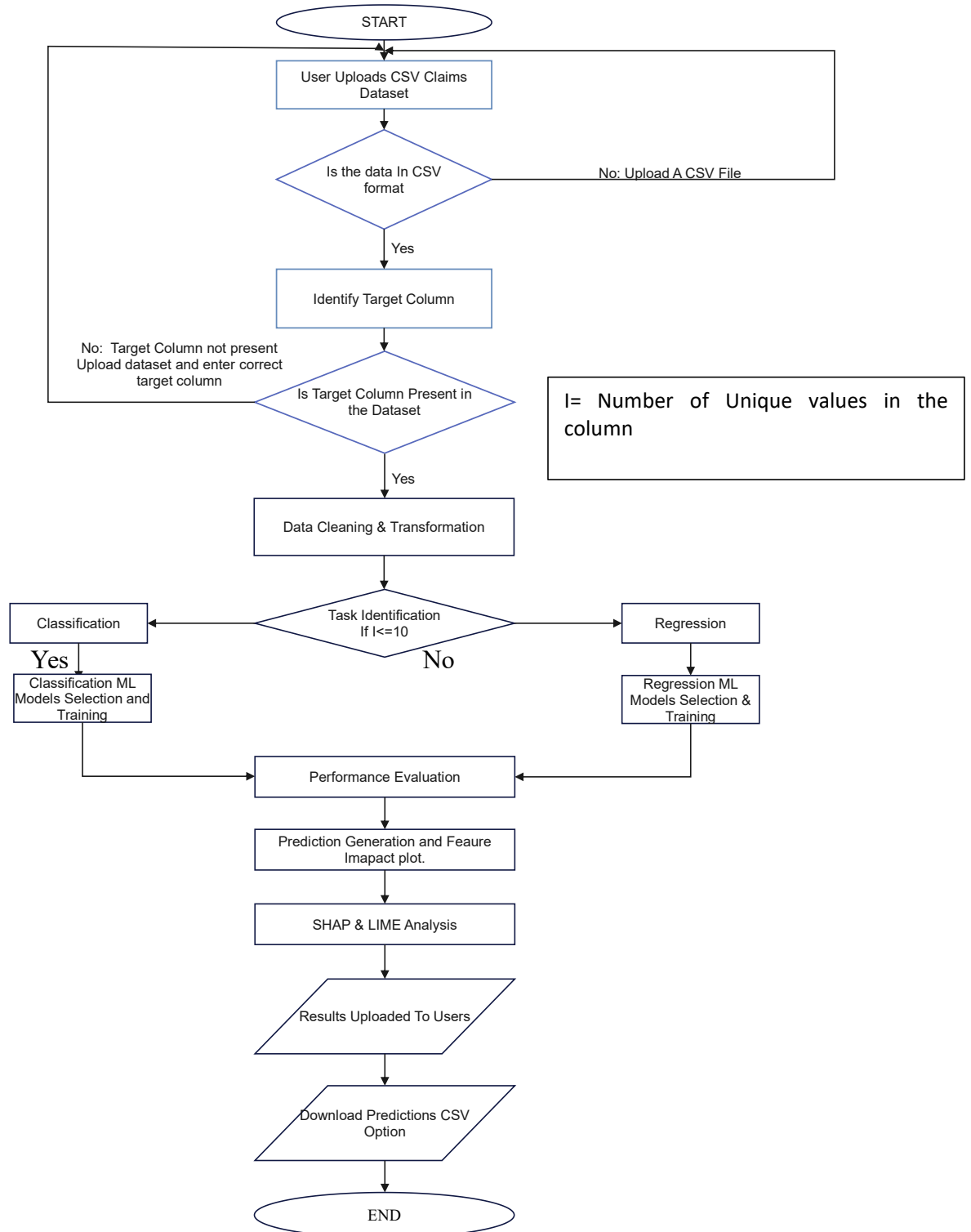


Figure 1: Flowchart Representing Prediction Pipeline

3.2.1 Start: The process begins when the user or data professional accesses the claims prediction system using the web interface.

3.2.2 User Uploads the dataset: The insurer is prompted to upload the firm's claims data in a CSV file format for the prediction to perform successfully.

3.2.3 Is the Data in CSV format: The claims system strictly supports only files in .csv format. It proceeds to the next step if the file format uploaded is in .csv extension and an error message that is displayed saying "Please upload a CSV file" if not in csv format.

3.2.4 Identify Target Column: For variability in the features within insurance claims datasets and various naming convention, If the inputted target column is present it proceeds to the next step, else it shows the error message "Target column not found".

3.2.5 Data Cleaning and Transformation: The claims system performs the following preprocessing tasks:

- i. Replacing missing or unknown values.
- ii. Dropping leakage columns that could bias the accuracy of predictions
- iii. Encode the categorical variables within the dataset using label encoding to aid prediction.
- iv. Replace missing numerical values with statistical methods like mean imputation.

3.2.6 Task Identification: The system automatically detects the type of task as stated in the target column. It identifies the task as categorical task if it has less than or equal to ten (10) unique values else it is identified as a regression task.

3.2.7 Model Selection and Training: Based on the selected tasks the various models are chosen and trained using the uploaded dataset.

For Regression and Classification Tasks, five models are available for implementation

A. Linear Regression

Linear regression is a basic regression model that often assumes a linear relationship between variables. For example, $B_0 + B_1x_1 + B_2x_2 + \dots + B_nx_n + e$. Usually used as baseline to compare

more advanced models and is easy to compute and interpret.

Advantages: It is simple, it can easily be interpreted and works well between linear relationships.

Disadvantages: Linear regression is inefficient in analysing non-linear and complex relationships, it is extremely sensitive to outliers.

B. Random Forest (Regressor and Classifier)

An ensemble learning model that constructs multiple decision trees and finds the average of their predictions to increase accuracy and reduce the over-fitting level. RF splits each component of the decision tree at the construction stage and randomly chooses a subset of the predictors then finally performs the regular split selection process on the chosen predictor subset (Edossa M. T., 2023). One use case of the FR is that it normally reduces the variations of individual trees by producing multiple trees.

Random forest represented by RF brings an improvement over bagging through a small modification in ornamenting trees (Gareth et al., 2013).

Application: The random forest regressor can capture non-linear relationships and handling categorical/continuous features effectively.

Advantages: It is immune to outliers and noise, and it provides feature importance, hence helping interpretability.

Disadvantages: It is computationally costly for large datasets. It might over-fit with deep trees if it is not finely tuned.

C. Gradient Boosting (Classifier and Regressor)

This is an ensemble method that builds trees sequentially, where each new tree makes an attempt to address the errors in the preceding one.

Advantages: It has high predictive power and handles non-linear relationships well and is tuneable with parameters like learning rate, depth, etc.

Disadvantages: It is computationally expensive and is sensitive to overfitting when not tuned properly.

D. Logistic Regression (Classification)

This is a statistical model that estimates the probability that an occurrence belongs to a certain class using a logistic function

Advantages: It is simple and interpretable. It works fine for linearly differentiated data. It is also fast to train.

Disadvantages: It usually assumes linearity between features and log-odds. It doesn't work too well with complex relationships in data.

E. Decision Tree (Classifier and Regressor)

This is a tree-based machine learning model that divides data based on the values of the features to make predictions, in turn forming a hierarchy of decisions. (Henckaerts et al., 2020)

Advantages: It is easy to interpret and visualise. It handles both numerical and categorical features. It requires minimal exploratory data analysis.

Disadvantage: It is suitable to overfitting on noisy data. It is unstable to little changes in data I.e. high variance.

F. XGBoost (Classifier and Regressor)

This is an optimized, upgradable version of the gradient boosting machine that makes use of regularized techniques and efficient computation (Clemente et al., 2023). The concept of “boosting” refers to a group of algorithms that can convert weak learners into strong learners to improve the predictive capabilities of a regression or a classification procedure (Edossa M. T. 2023).

Advantages: It is very efficient and fast. Its built-in regularization helps reduce overfitting. It handles missing values without stress.

Disadvantages: It is complex to tune properly and is less interpretable when compared to simpler models.

G. Hybrid Model (Random Forest + XGBoost)

For an increased accuracy, a custom ensemble model that combines predictions from Random Forest and XGBoost either by voting for classification tasks or averaging for regression tasks.

Advantages: It leverages strengths of both bagging and boosting. Often yields superior accuracy. Adds robustness to predictions.

Disadvantages: It has a higher computational cost. Increased complexity and reduced transparency.

3.2.8 Model Training and Evaluation

Each selected model is designed to be trained using the training split method. Performance is measured using:

- a. Classification Metrics: Accuracy, Precision, Recall, F1-score
- b. Regression Metrics: R squared, Mean Absolute Error and Mean Squared Error.

The model with the highest relevant metric (Accuracy or R squared) is selected as the best model, The model is then what is used to generate the prediction

3.2.9 Model Interpretability

- a. Implementing SHAP (SHapely Additive Explanations):

SHAP provides both global (the feature importance across all the fields in the dataset) and local (the feature importance across individual predictions) interpretations. It shows the contribution of individual features either positively or negatively towards the prediction.

- b, Implementing LIME (Local interpretable Model-Agnostic Explanations):

LIME helps to understand individual predictions by decimating the model locally with an easier linear model. Users can check out these explanations per instance or download them in CSV format for bulk review.

The Explainability metrics are further displayed as SHAP bar plots and LIME explanations on the web interface. These visualisations help insurers understand why a particular claim prediction outcome has the result it has, which helps to increase transparency, compliance and trust in the models.

3.2.10 Results Uploaded to Users: Predictions are then displayed to the insurers on the webpage with performance evaluation table containing a sample of predictions.

3.2.11 Download Predictions (CSV Option): Users are given the option of downloading the full results of the prediction which includes a clean dataset, the predicted values and SHAP or LIME outputs if requested

3.2.11 Stop: The prediction cycle is ended. The insurers can either choose to re-analyse by the reset button or uploading a new dataset

3.3 System Development

This section describes the method of integrating the machine learning models on the web that can be accessible to users.

3.3.1 Flask backend with model logic.

The backend is built using Flask, containing logic for:

Handling file uploads

Preprocessing datasets

Selecting and training models

Making predictions

Generating SHAP and LIME plots

Returning results and explanations to the frontend

3.3.2 Google Cloud deployment and CI/CD integration.

The application is deployed on Google Cloud App Engine, enabling accessibility and

scalability. GitHub is used for version control and continuous updates, ensuring smooth synchronization through pushes to the deployment repo.

3.3.3 Interface design and user experience considerations.

The frontend was designed using HTML allowing for a responsive and user-friendly experience. Progress indicators, alerts, performance tables, and visual charts make the system intuitive and accessible for non-technical users.

3.4 Tools and Frameworks Used

Table 1: Representation of Tools and Frameworks Used.

Category	Tool/Frameworks
Programming Language	Python
Web Framework	Flask
ML Libraries	Scikit-Learn, XGBoost
Model Interpretability	SHAP, LIME
Frontend	HTML
Deployment	Google Cloud Platform
Version Control	Git, GitHub
Plotting	Matplotlib
Integrated Development Environment (IDE)	Visual Studio Code

CHAPTER FOUR

IMPLEMENTATION

4.1 Introduction

This chapter explains the practical implementation of the Insurance Claims Prediction System, producing results from the various machine learning models available, and also evaluating their performances on the strength of various performance metrics. This chapter also shows systems features with respect to explaining the results of the prediction as well as the idea behind the deployment.

4.2 System Setup and Environment

The web-application was developed using Python (Flask framework for the backend), and the models are trained with Scikit-Learn and XGBoost. Visualizations and interpretability are treated using SHAP, LIME and Matplotlib. The system is then deployed temporarily on Google Cloud Platform; this allows users access the interface and perform independent predictions.

System's configuration

Language: Python 3.11.0

Framework: Flask (Backend)

Libraries: Scikit-learn, XGBoost, SHAP, LIME, Pandas, Matplotlib

Deployment: Google Cloud Run with CI/CD integration

Frontend: HTML and CSS

4.3 Testing the Prediction System

4.3.1 User Interface Experience:

The front end of the system was built to be intuitive:

Drag-and-drop CSV upload

Real-time display of sample predictions

Downloadable results

Model performance table

Feature importance visualizations

SHAP and LIME embedded plots

4.3.2 Preliminary Processes

The systems permit users to upload claims data in csv file format, identify and specify the target column, and automatically detect whether the specified column is a classification or regression task. After identification it then:

A. Performs Data Preprocessing (it handles missing values and encodes categorical features).

B. It prevents data leakage by ignoring columns that could affect the prediction accuracy negatively.

C. It trains multiple models which includes:

- a) Linear Regression
- b) Logistic Regression
- c) Decision Tree
- d) Random Forest
- e) Gradient Boosting
- f) XGBoost
- g) Hybrid Model (Random Forest +XGBoost)

D. It then selects the best model based on the appropriate performance models and produces

prediction.

4.3.3 Testing Classification Tasks

For classification tasks, performance is always evaluated using:

- A. Accuracy: Overall correctness of the model.
- B. Precision: How many predicted positives were truly positives.
- C. Recall: How many actual positives were correctly predicted.
- D. F1 Score: The balance between precision and recall.

4.3.4 Testing Regression Tasks

For regression tasks, performance is always evaluated using:

- A. Mean Absolute Error (MAE)
 - i. What it measures: The Average magnitude of the errors in predictions.
 - ii. Interpretation: Low MAE values indicate better model accuracy. For example, if $MAE = 400$, the model's predictions are, on average, 400 units off from the actual loss values. The smallest MAE across models preforms best at minimizing speculation errors. It is less sensitive to outliers than RMSE.
- B. R squared Score:
 - i. What it measures: The proportion of variance in the target variable explained by the model.
 - ii. The values range from 0 to 1
 - $R^2 = 1$: Perfect Prediction
 - $R^2 = 0$: Model explains none of the variance.
 - Negative values: Model performs worse than the predicting mean.

For example, if $R^2 = 0.55$: It shows that the model explains 55% of the variance in the loss data. Hence a higher R^2 Value the better explanatory power.

C. Mean Squared Error: MSE is a measure of the average of the squares of the errors (differences between predicted and actual values) in a set of predictions. It quantifies how well a model predicts the values in a dataset, with a lower MSE indicating better accuracy.

4.3.5 Model Interpretability:

a. SHapely Additive exPlanations (SHAP): SHAP attaches every feature a Shapely value which is its representation in the prediction. The aggregate of all Shapley values across the various predictions will provide an overall understanding of the importance of features (Sun et al., 2024). It would aid insurers understand which factors affect individual predictions, like why a claim is high for a specific policyholder. SHAP provides global and local explanations for model behaviour. The summary plot revealed which features had the greatest impact on predictions, helping insurers understand the reasoning behind each decision.

Legend Interpretation:

Red: Feature pushes the prediction higher.

Blue: Feature pushes the prediction lower.

X-axis: Strength of the effect.

Dots: Each represents one prediction instance

b. Local Interpretable Model-Agnostic Explanations (LIME): LIME (Local Interpretable Model-Agnostic Explanations) provided instance-specific insights. A bar chart explained why the system made a specific decision for a particular case.

Downloadable CSV Feature:

To enhance usability, the system generates a downloadable .csv file with LIME explanations for multiple records, summarizing:

Feature contribution per record, Positive/Negative influence Probability or impact.

4.4 Model Implementation and Evaluation Using Claims Datasets.

4.4.1 Data-Source: Insurance Fraud Dataset on Mendeley Data

<https://data.mendeley.com/datasets/992mh7dk9y/2>

Target Column: Total_claim_amount

Task Type: Regression

Data Description:

Dataset Size: 1,000 Records by 40 columns

File Name: insurance_claims.csv

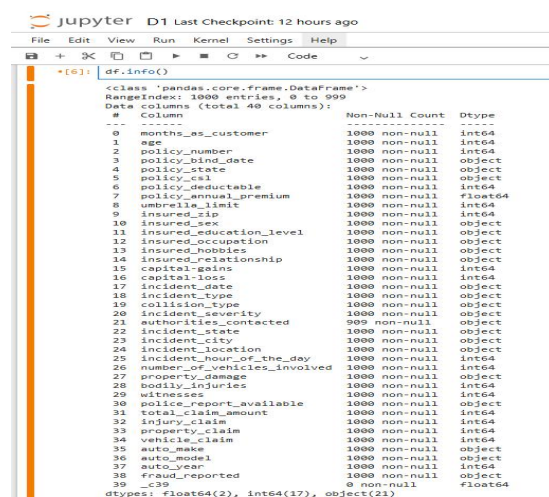
Data Quality: Highly complete with minimum missing values present. (99.77%)

Columns with Missing Values:

Authorities_Contacted: 9.1% (i.e. 91 out of 1000 records)

_c39: 100% (empty column).

The figure below shows the summary of the dataset, count and datatypes.



```

In [5]: df.info()

<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_deductible                    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               909 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)

```

Figure 2: Summary Information on Dataset

The figure below shows the attachment of the insurance claims csv file on the web-application. The dataset must be in csv file format as already explained in sections above. The target column is then specified in the format it is represented in the csv file, there is allowance for misplacement of uppercase or lowercase letters and white spaces but not

spelling errors or the likes. The refresh button clears predictive analysis from the frontend. Data is stored temporarily in a static folder as connection is in session and deleted after session for memory management.

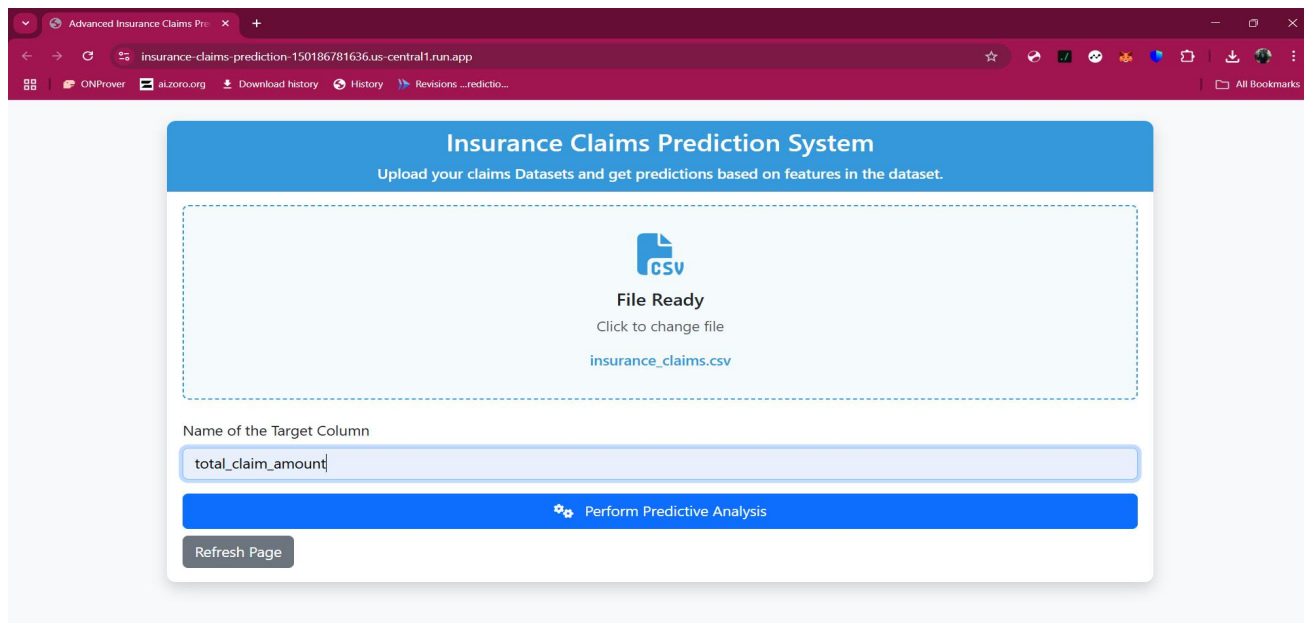


Figure 3: Attached Dataset and Indication of Target Column

The figure below shows the message that is displayed while predictive analysis occurs at the backend before results are produced.

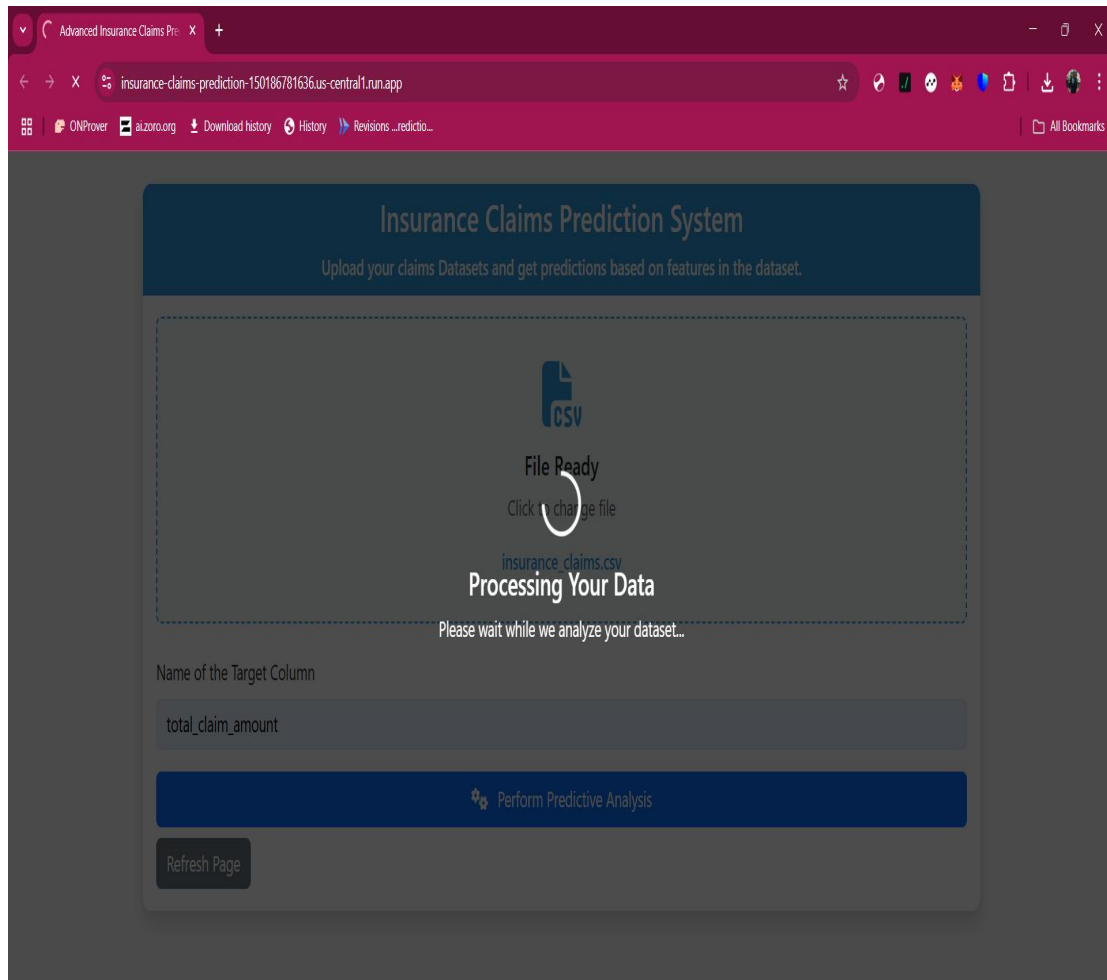
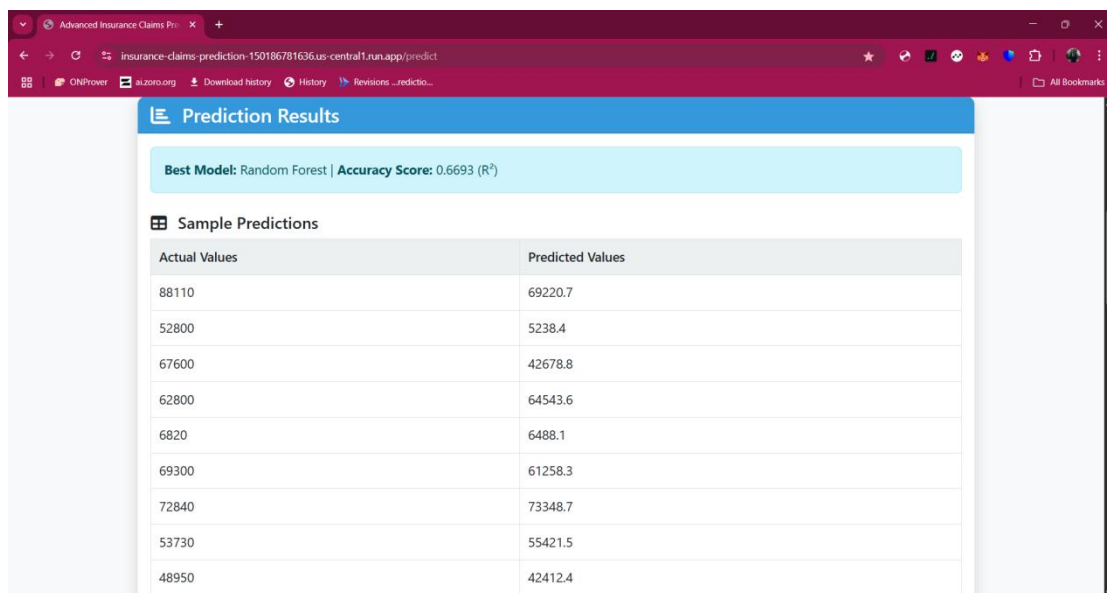


Figure 4: Data Processing Message Displayed

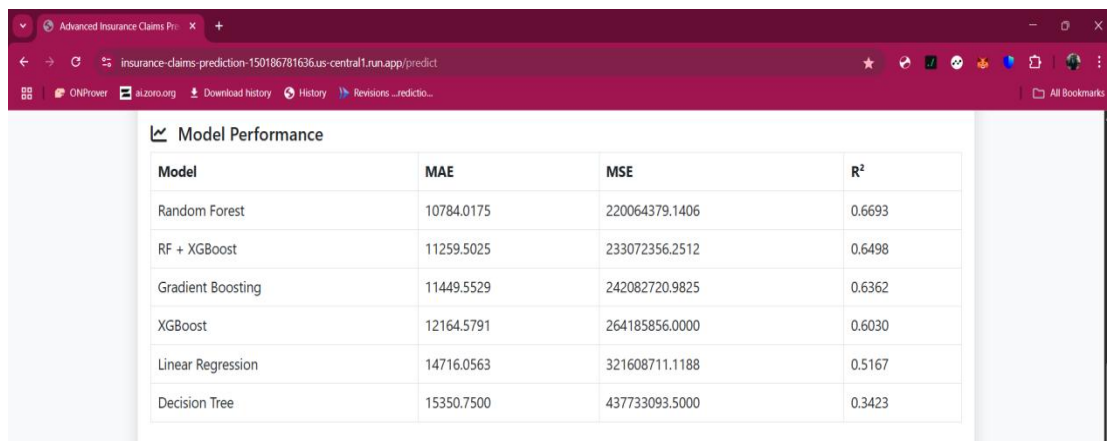
The figure below shows the sample of the predicted results when the data was supplied to the model. As already discussed in preceding chapters, the model uniquely identifies which task is to be performed and makes the choice of which model to train and implement based on the target column. The table shows the sample of the first ten rows of predictions for the user to access. The model that performs best is what is used to produce this prediction and the accuracy score and corresponding model is displayed on the web interface.



Actual Values	Predicted Values
88110	69220.7
52800	5238.4
67600	42678.8
62800	64543.6
6820	6488.1
69300	61258.3
72840	73348.7
53730	55421.5
48950	42412.4

Figure 5: Prediction Results and Best Model.

The figure below shows the various Models and how well they performed. As stated in the sections above, there are both classification and regression models. This particular dataset and its target column was identified to be a regression task, hence regression models. The various regression metrics are Mean squared error which tells you how far off your predictions are from the actual values — but it makes big errors look even bigger because it squares them, Mean absolute error which tells you the average difference between your predictions and the real values, no matter the direction (too high or too low), It treats all errors equally — big or small and R squared which tells you how well your model explains the data, a score of 1 means your model is perfect, and 0 means your model is no better than just guessing the average.



Model	MAE	MSE	R ²
Random Forest	10784.0175	220064379.1406	0.6693
RF + XGBoost	11259.5025	233072356.2512	0.6498
Gradient Boosting	11449.5529	242082720.9825	0.6362
XGBoost	12164.5791	264185856.0000	0.6030
Linear Regression	14716.0563	321608711.1188	0.5167
Decision Tree	15350.7500	437733093.5000	0.3423

Figure 6: Various Models and Their Performance Evaluation Metrics.

The figure below shows the relevance of each feature within the dataset that was uploaded and how relevant they were regarding the predictions. Considering that the prediction results rendered to the users are based on the best model, and so is the plot. Each bar on the plot represents the strength of the various features on the y axis and their corresponding values on the x axis. The longer the bar, the more important it is for making predictions. Green bars show that feature plays a positive role in the prediction and red bars imply that the feature plays a negative role in predictions. The downloadable csv file containing the cleaned uploaded dataset and the predicted values for each field can be accessed using the Download Predictions Button placed directly beneath the feature importance plot to enable users get the full scope of the predictions for each policy holder.

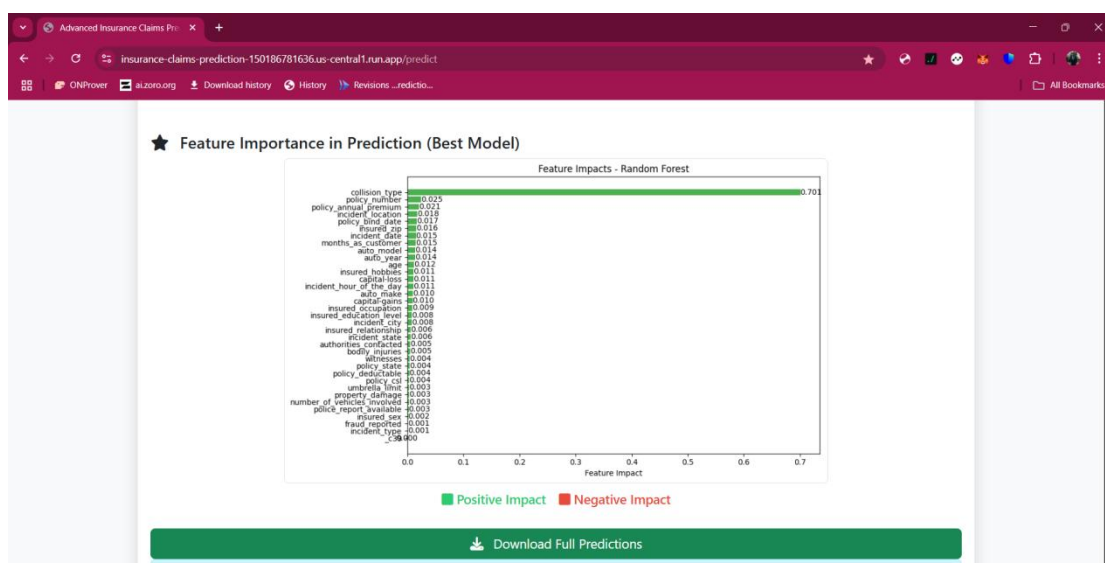


Figure 7: Feature Importance Plot and Downloadable CSV File

The figure below represents the Shapely additive explanations plot of the predictions. The chart shows how each feature pushes the prediction in either direction. The center line is usually the base line or neutral point. Red dots pointing right push the prediction towards the positive outcome, while the blue dots pointing left influence the prediction to the negative outcome. In this case, majority of the features are clustered around the baseline but having collision type at the top having longer bars indicating a stronger push in the left direction.

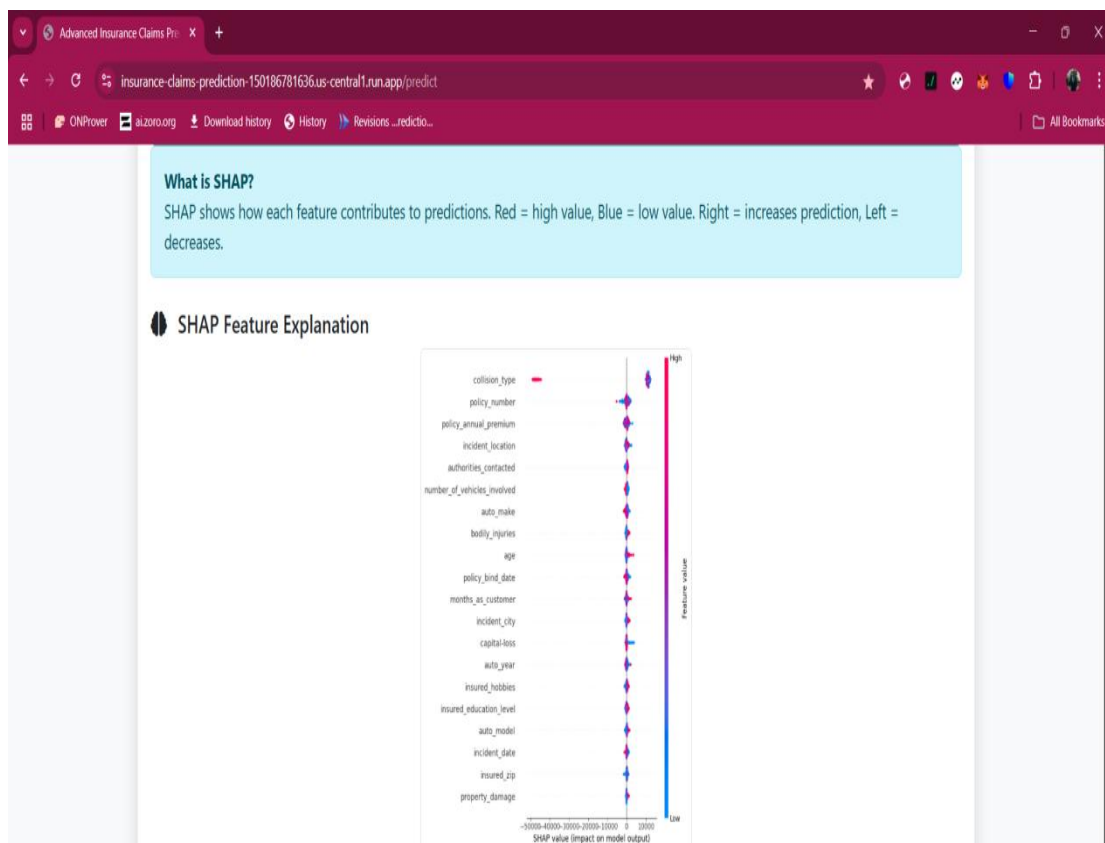


Figure 8: SHAP Explanations Plot Based on Best Model

The figure below shows the option to be able to download the lime explanations in a csv file for those that need additional insights, Since the LIME explanations are quite large, it cannot be displayed here, hence it is rendered on another web page that can be accessed using the view lime html report. When it is clicked, it redirects to the landing page that displays the lime results for one field.

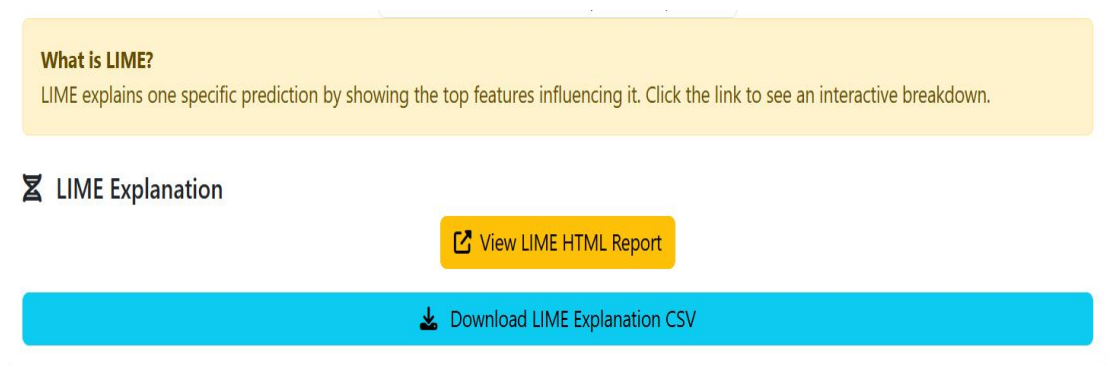


Figure 9: LIME Explanations and Downloadable CSV file

The figure below indicates the Local interpretable model-agnostic explanation values for one field in the dataset. LIME values explain in more detail the dynamics behind predictions. For this this lime explanation, the bar on the left corner indicates that the starting point is 3000.70 and final prediction 75430.50. This was the baseline the model worked with to arrive at the prediction. Each bar also represents each features contribution at achieving that particular prediction. Orange/red bars pointing right makes the model assume the value should be higher while blue bars pointing left suggests that the values should go lower or negative. Like in this case the orange features pushed the predictions up and the blue features reduced the prediction values.

Feature	Value
collision_type	1.00
policy_annual_premium	1137.02
policy_state	0.00
incident_location	297.00
authorities_contacted	0.00
auto_year	2003.00
incident_date	24.00
fraud_reported	0.00
incident_hour_of_the_day	22.00
policy_bind_date	930.00

Figure 10: LIME Explanation Values

4.4.2 Data-Source: Ncid Private Motor claims insurance dataset

Target Column: Claim Type

Task Type: Classification

Dataset size: 784 records by 4 columns

File name: ncid_private_motor_ultimate_claims.csv

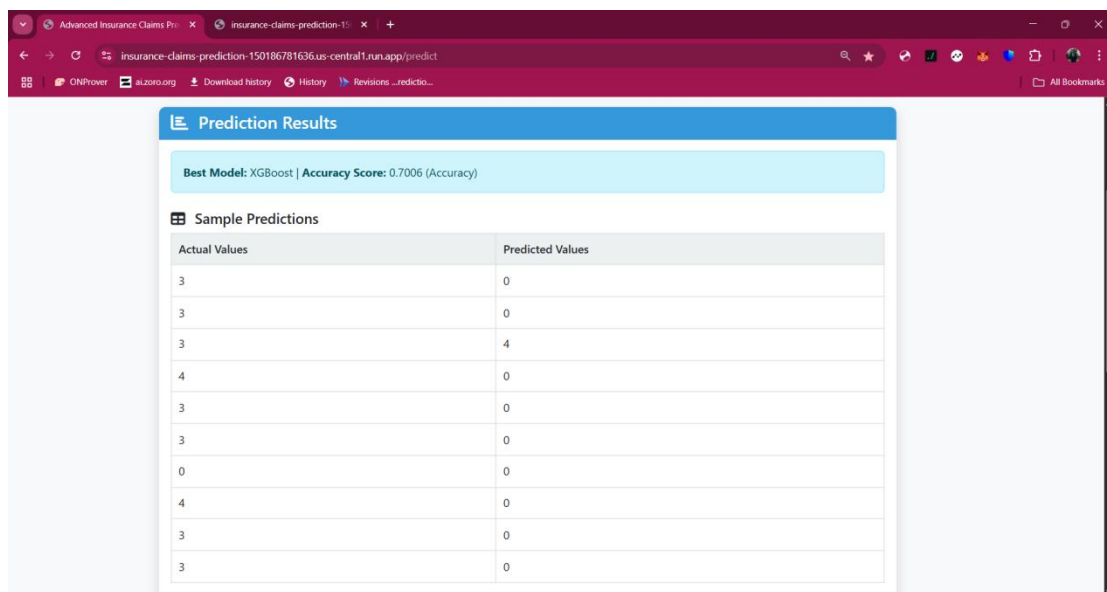
Data Quality: No missing Values (Completeness)

```
: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 784 entries, 0 to 783
Data columns (total 4 columns):
#   Column          Non-Null Count  Dtype
---  -
0   AccidentQuarter  784 non-null   int64
1   Measure          784 non-null   object
2   ClaimType        784 non-null   object
3   Value            784 non-null   float64
dtypes: float64(1), int64(1), object(2)
memory usage: 24.6+ KB
```

Figure 11: Information Summary on Dataset 2

The figure below shows the sample of the predicted results when the data was supplied to the model. As already discussed in preceding chapters, the model uniquely identifies which task is to be performed and makes the choice of which model to train and implement based on the target column. The table shows the sample of the first ten rows of predictions for the user to access. The model that performs best is what is used to produce this prediction, and the accuracy score and corresponding model is displayed on the web interface. This is classification task as seen and the best model as seen is the XGBoost Classifier with an accuracy score of 0.7 which is quite commendable.



Prediction Results

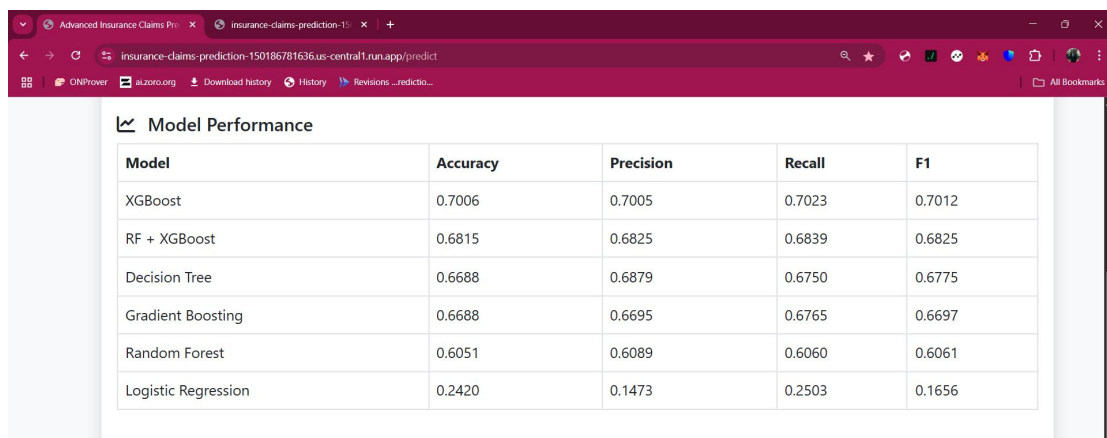
Best Model: XGBoost | Accuracy Score: 0.7006 (Accuracy)

Sample Predictions

Actual Values	Predicted Values
3	0
3	0
3	4
4	0
3	0
3	0
0	0
4	0
3	0
3	0

Figure 12: Prediction Sample Results Classification Task.

The figure below shows a tabular representation of the performance of the various classification models and their corresponding performance evaluation metrics. Accuracy explains how many time the model predicted right results overall in all cases, Precision explains how trustworthy predictions are, it focuses on the positive predictions and how often they are true, Recall evaluates how many actual positive scenarios the model was able to identify, F1 Score balances precision and recall, it is vital when catching all the real claims and not raising too many false alarms.



The screenshot shows a web browser window with a dark red header. The address bar displays the URL 'insurance-claims-prediction-150186781636.us-central1.run.app/predict'. Below the header, there is a section titled 'Model Performance' with a table containing five columns: Model, Accuracy, Precision, Recall, and F1. The table lists six models: XGBoost, RF + XGBoost, Decision Tree, Gradient Boosting, Random Forest, and Logistic Regression, each with its corresponding performance metrics.

Model	Accuracy	Precision	Recall	F1
XGBoost	0.7006	0.7005	0.7023	0.7012
RF + XGBoost	0.6815	0.6825	0.6839	0.6825
Decision Tree	0.6688	0.6879	0.6750	0.6775
Gradient Boosting	0.6688	0.6695	0.6765	0.6697
Random Forest	0.6051	0.6089	0.6060	0.6061
Logistic Regression	0.2420	0.1473	0.2503	0.1656

Figure 13: Various Classification Models with Performance Evaluation Metrics.

The figure below shows the relevance of each feature within the dataset that was uploaded and how relevant they were regarding the predictions. Considering that the prediction results rendered to the users are based on the best model, and so is the plot. Each bar on the plot represents the strength of the various features on the y axis and their corresponding values on the x axis. The longer the bar, the more important it is for making predictions. Green bars show that feature plays a positive role in the prediction, and red bars imply that the feature plays a negative role in predictions. The downloadable csv file containing the cleaned uploaded dataset and the predicted values for each field can be accessed using the Download Predictions Button placed directly beneath the feature importance plot to enable users get the full scope of the predictions for each policy holder.

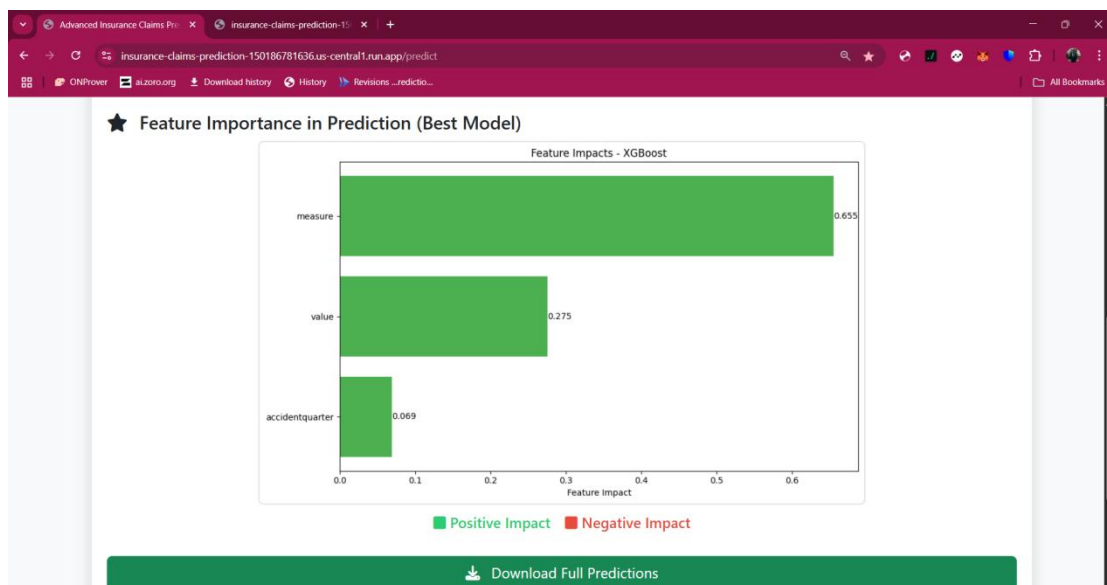


Figure 14: Feature Importance Plot Based Classification

The figure below represents the Shapely addictive explanations plot of the predictions. The chart shows how each feature pushes the prediction in either direction. The center line is usually the base line or neutral point. Red dots pointing right push the prediction towards the positive outcome, while the blue dots pointing left influence the prediction to the negative outcome and various classes.

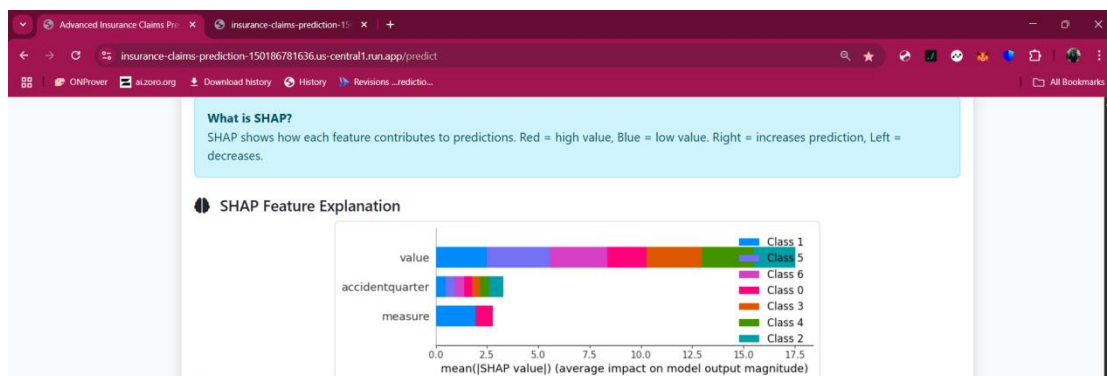


Figure 15: SHAP Explanations Plot Based on Best Model

The figure below indicates the Local interpretable model-agnostic explanation values for one field in the dataset. LIME values explain in more detail the dynamics behind predictions.

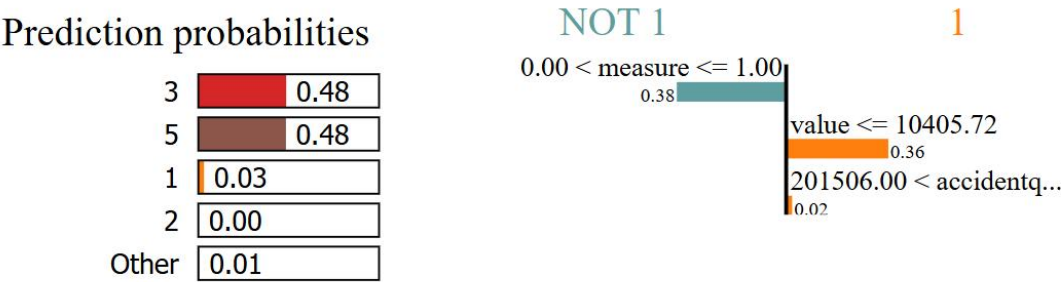


Figure 16: LIME Explanation Values for a Field in The NCID Dataset.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Summary

This project is centred on building a system that predicts insurance claims using machine learning models. The aim was to automate the act of speculating the value of insurance claims and at what intensity as treated by classification and regression models, using policy holders' historical data and features. Various classification/regression models including Linear regression, logistic regression, Decision Tree, Random Forest, Gradient Boosting XGBoost and a hybrid model (Random Forest + XGBoost).

These various tasks were further evaluated to measure the strength and accuracy of the predictions and compared using performance evaluation metrics depending on the type of task to be performed, Accuracy, precision, recall, F1 score, Mean squared error, r squared and mean absolute error were the metrics used to check and compare the various models.

After reviewing several papers, it was discovered that model explainability was either a recurrent issue or was consistently left out when carrying out insurance claims speculations.

It was on this premise that model explainability metrics like Local Interpretable Model-agnostic Explanations (LIME) and SHapely Addictive Explanations (SHAP) for both local and global predictions explanations that would help in providing meaningful insights on how various features influenced the results of the predictive models.

All these functionalities were then coupled together using various frame works and tools like Flask on Python for backend development, HTML, CSS and JavaScript for Front-end development and Google cloud for deployment to enable users upload claims datasets and get predictions in real time.

5.2 Conclusion

The results obtained from the tests carried out on the web-interface shows that machine learning models can efficiently predict insurance claims when they are supplied with quality datasets. The best model in terms of performance were discovered to be in a variety as how much a model would perform relatively depends on the dataset and the co-herent relationships between features. The use of LIME and SHAP helped in understanding the role features play in predictions, overall contributing to the trust of stakeholders and insurers in the predictive models.

Deploying the model on a web-application shows the practical application of this project, ultimately making it accessible by insurance companies or data professionals and those with little or no knowledge about machine learning models.

5.3 Recommendation

The benefits of exploring Artificial intelligence and Machine Learning cannot be overstated, it is then recommended that insurance firms adopt these methodologies for key operational activities within the insurance industry to enhance sustainability, productivity and maintain profitability as they carry out operational activities. It is important that insurance firm ensure that their data is managed properly and stored in the best possible format, this would enhance the perfoamnnce of the models by increasing their accuracy scores.

5.4 Future work

The project can be further enhanced to allow for increased productivity, effeciency and improved accuracy of models. Future works would include integrating the web interface into AWS cloud server for increased speed and memory management, implementing deep learning models for increased accuracy while still comparing across machine learning models for benefit of doubt and including time series models to predict peak claim periods.

REFERENCES

- Abdelhadi, S., Elbahnasy, K., & Abdelsalam, M. (2020). A PROPOSED MODEL TO PREDICT AUTO INSURANCE CLAIMS USING MACHINE LEARNING TECHNIQUES. *Journal of Theoretical and Applied Information Technology*, 98, 22. www.jatit.org
- Boehmke, B. & Greenwel, B. (2020). *Hands-on Machine Learning with R*. Taylor and Francis.
- Bolancé, C., Guillén, M., Nielsen, J. P., & Pelican, E. (2018). Statistical modeling of insurance claim severity: A review. *ASTIN Bulletin: The Journal of the International Actuarial Association*, 48(1), 215-252. doi:10.1017/asb.2017.28
- Burri, R. D., Burri, R., Bojja, R. R., & Buruga, S. R. (2019). Insurance claim analysis using machine learning algorithms. *International Journal of Innovative Technology and Exploring Engineering*, 8(6 Special Issue 4), 577–582. <https://doi.org/10.35940/ijitee.F1118.0486S419>
- Clemente, C., Guerreiro, G. R., & Bravo, J. M. (2023). Gradient Boosting in Motor Insurance Claim Frequency Modelling. *Atas Da Conferencia Da Associacao Portuguesa de Sistemas de Informacao*, 53–69. <https://doi.org/10.18803/capsi.v23.53-69>
- David, M. (2015). Auto Insurance Premium Calculation Using Generalized Linear Models. *Procedia Economics and Finance*, 20, 147–156. [https://doi.org/10.1016/s2212-5671\(15\)00059-3](https://doi.org/10.1016/s2212-5671(15)00059-3)
- Frees, E. W., Gao, J., & Rosenberg, M. A. (2011). Predicting the frequency and amount of health care expenditures. *North American Actuarial Journal*, 15(3), 377–392.
- Friedman, J. (2001). Greedy boosting approximation: a gradient boosting machine. *Annals of Statistics*, 29, 1189–1232.
- Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2 (3), 916-954.
- Garrido, J., Genest, C. & Schulz, J. (2016). Generalized linear models for dependent frequency

- and severity of insurance claims. *Insurance: Mathematics and Economics*, 70, 205-215.
- Goldstein, A., Kapelner, A. Bleich, J. & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24 (1), 44-65.
- Gschlößl, S., & Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal*, 3, 202–225.
- Hanafy, M., & Ming, R. (2021). Machine learning approaches for auto insurance big data. *Risks*, 9(2), 1–23. <https://doi.org/10.3390/risks9020042>
- Henckaerts, R., Côté, M.-P., Antonio, K., & Verbelen, R. (2020). Boosting insights in insurance tariff plans with tree-based machine learning methods. <http://arxiv.org/abs/1904.10890>
- Bishop, N. (2024). Application of machine learning techniques in insurance underwriting. *Journal of Actuarial Science*, 2(1), 1–13. <https://www.carijournals.org>
- Ong, S. H., Sim, S.Z., & Liu, S. (2025). The Modelling of Auto Insurance Claim-Frequency Counts by the Inverse Trinomial Distribution. *Journal of Risk and Financial Management*, 18(1), 7. <https://doi.org/10.3390/jrfm18010007>
- Paul, J. (2024). AI-Powered Data Analytics: Shaping the Future of Auto Insurance Pricing and Claims Processing. <https://www.researchgate.net/publication/386076125>
- Poufinas, T., Gogas, P., Papadimitriou, T., & Zaganidis, E. (2023). Machine Learning in Forecasting Motor Insurance Claims. *Risks*, 11(9). <https://doi.org/10.3390/risks11090164>
- Segura-Gisbert, J., Lledó, J., & Pavía, J. M. (2025). Dataset of an actual motor vehicle insurance portfolio. *European Actuarial Journal*, 15(1), 241–253. <https://doi.org/10.1007/s13385-024-00398-0>
- Shi, P., Feng, X., Ivantsova, A. (2015). Dependent frequency–severity modeling of insurance claims. *Insurance: Mathematics and Economics*, 64, 417–428.
- Shu, C., & Burn, D. H. (2004). Artificial neural network ensembles and their application in pooled

- flood frequency analysis. *Water Resources Research* 40, 1–10.
- Staudt, Y. and Wagner, J. (2021) Assessing the performance of random forests for modeling claim severity in collision car insurance. *Risks*, 9(3), 53.
- Su, X., Bai, M. (2020) Stochastic gradient boosting frequency-severity model of insurance claims. *PLoS ONE* 15(8): e0238000. (available at <https://doi.org/10.1371/journal.pone.0238000>).
- So, B. (2024). Enhanced Gradient Boosting for Zero-Inflated Insurance Claims and Comparative Analysis of CatBoost, XGBoost, and LightGBM. <http://arxiv.org/abs/2307.07771>
- Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing Auto Insurance Risk Evaluation With Transformer and SHAP. *IEEE Access*, 12, 116546–116557.
<https://doi.org/10.1109/ACCESS.2024.3446179>
- Wei, Y., Li, Z., & Li, L. (2019). Machine learning methods in predictive maintenance for vehicle systems: A review. *Applied Sciences*, 9, 3617. <https://doi.org/10.3390/app9173617>
- Xiaonan, L. (2023). Identifying the Optimal Machine Learning Model for Predicting Car Insurance Claims: A Comparative Study Utilising Advanced Techniques. *Academic Journal of Business & Management*, 5(3). <https://doi.org/10.25236/ajbm.2023.050317>