**Insurance Claim Prediction using Machine Learning**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**Team**

**JENIN L**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**Nov 2022**

**List of Figures**

**1. Dataset Distribution**

**2. Preprocessing Overview**

**3. Model Performance Comparisons**

**4. ROC Curve for Model Accuracy**

**5. Feature Importance Graph**

**List of Abbreviations**

**- ML - Machine Learning**

**- DL - Deep Learning**

**- AUC - Area Under Curve**

**- ROC - Receiver Operating Characteristic**

**Table of Contents**

**1. Abstract  
2. Problem Definition  
3. Introduction  
4. Literature Survey  
5. Data Collection and Preprocessing  
6. Methodology  
7. Results  
8. Conclusion  
9. References**

**Abstract**

This project develops a predictive model to assess insurance claims, leveraging advanced machine learning techniques. The dataset, sourced from Kaggle, includes 58,592 records with 44 features capturing policy details, policyholder demographics, vehicle specifications, and various safety features. Key features include Policy Tenure, Age of Car, Age of Policyholder, Vehicle Segment, Fuel Type, and several binary safety indicators such as Electronic Stability Control (ESC) and Brake Assist. The target variable, is\_claim, is binary, indicating whether a claim was made (1) or not (0).

**1. Problem Definition**

**1.1 Overview**

Insurance companies face challenges in managing and predicting claims accurately, which can lead to resource misallocation and financial losses.

**1.2 Problem Statement**

To build a predictive model that forecasts insurance claims using machine learning, thereby assisting insurance companies in optimizing resource allocation and improving the accuracy of claim assessments.

**2. Introduction**

In the insurance industry, predicting potential claims efficiently is crucial for effective risk management and cost optimization. This project explores the use of machine learning techniques to develop a predictive model that aids in assessing the likelihood of claims based on policy details, customer demographics, and vehicle information

**3. Literature Survey**

Previous studies have demonstrated the effectiveness of machine learning models in insurance claim predictions. Techniques such as Logistic Regression and Random Forest have shown promising results, but complex data patterns often require more robust methods like XGBoost. XGBoost has been widely recognized for its high performance in structured data tasks, making it suitable for this project.

# Data Collection and Preprocessing

Total Records: 58,592  
  
Features: 44 columns  
  
**Key Feature Categories:**  
Policy Details: policy\_tenure, policy\_id  
Vehicle Information: age\_of\_car, make, segment, fuel\_type, etc.  
Policyholder Information: age\_of\_policyholder, area\_cluster  
Safety Features: Binary features like is\_esc, is\_brake\_assist, is\_power\_steering  
  
**Target Variable:**  
is\_claim: Binary variable indicating claim status (1 = Claim made, 0 = No claim).

# Data Cleaning

Missing Values:  
Managed missing data by using median imputation for numerical values and mode imputation for categorical values.  
  
**Outlier Removal:**Employed z-score analysis to remove extreme outliers in continuous variables such as age\_of\_policyholder and policy\_tenure.

# Feature Engineering

**Encoding:**  
Applied one-hot encoding for categorical features like fuel\_type, segment, and model.  
  
**Normalization:**Standardized numerical features for improved model performance.

# Data Splitting

Training Set: 80% (46,873 records)  
Testing Set: 20% (11,719 records)

# Methodology

# Exploratory Data Analysis (EDA)

Correlation Analysis:  
Identified strong positive correlations between policy\_tenure and is\_claim.  
  
Visualization:  
Histograms and Box Plots were used to understand feature distributions.  
Correlation Heatmap to examine relationships among variables.

# Model Selection

Implemented and compared the following models:  
  
Logistic Regression: Baseline model with an accuracy of 75%.  
Random Forest: Improved accuracy to 87%, but limited interpretability.  
XGBoost: Achieved the best performance with 91% accuracy.

# Hyperparameter Tuning

Utilized Grid Search CV to tune XGBoost parameters:  
  
Parameters Tuned:  
max\_depth: [3, 5, 7]  
learning\_rate: [0.01, 0.1, 0.3]  
n\_estimators: [100, 200, 500]  
subsample: [0.6, 0.8, 1.0]  
  
Best Configuration:  
max\_depth=5, learning\_rate=0.1, n\_estimators=200, subsample=0.8

# Results

# Model Performance Comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
| Logistic Regression | 75% | 72% | 70% | 71% | 0.78 |
| Random Forest | 87% | 84% | 82% | 83% | 0.91 |
| XGBoost (Tuned) | 91% | 89% | 88% | 88% | 0.93 |

# Feature Importance Analysis

The top features impacting claim likelihood included:  
  
Policy Tenure  
Age of Car  
Age of Policyholder  
Vehicle Segment

# ROC Curves

The XGBoost model showed superior performance, achieving an AUC score of 0.93.

# Web Application Interface with Streamlit

# Application Features

Real-Time Predictions: Users can enter policyholder and vehicle data to get immediate claim predictions.  
Feature Visualizations: Displays the impact of important features on predictions.  
User-Friendly Interface: Accessible design for non-technical insurance professionals.

# Deployment

Platform: Deployed using Streamlit Cloud.  
Accessibility: Publicly accessible link for end-users.  
Security: No personal data is stored.

# Conclusion

The project successfully built a highly accurate claim prediction model using XGBoost, achieving 91% accuracy with extensive hyperparameter tuning. The Streamlit app provides a practical tool for insurance professionals to assess claim risks, making it valuable in real-world insurance settings.

# Future Work

Expand Feature Set: Incorporate additional features such as customer credit scores or geolocation data.  
Optimize Deployment: Explore containerization for better scalability.  
User Feedback Mechanism: Integrate feedback to refine predictions further.

# References

Kaggle Insurance Dataset: Link  
XGBoost Documentation: [Link](https://xgboost.readthedocs.io/)