IJIREEICE

ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁒ Peer-reviewed & Refereed journal ⁒ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131126

INSURANCE CLAIM STATUS PREDICTION

Venkatapathy R[1], Senthilnathan M[2], Rohith G[3], Dr. G Paavai Anand[4]

Student, Department of CSE, SRM Institute of Science and Technology, Chennai, India[1]

Student, Department of CSE, SRM Institute of Science and Technology, Chennai, India [2]

Student, Department of CSE, SRM Institute of Science and Technology, Chennai, India [3]

Assistant Professor (Sr. G), Department of CSE, SRM Institute of Science and Technology, Chennai, India [4]

Abstract: Modern insurance industries increasingly rely on data-driven solutions to forecast the likelihood of claims, aiming to cut financial losses and streamline workflows. Traditional claim reviews are often inefficient and subjective, fueling the need for smart, automated prediction tools. This project develops a machine learning-based system to estimate which policyholders are most likely to file claims, leveraging detailed customer, vehicle, and policy data. Models like Logistic Regression, Random Forest, and Gradient Boosting are assessed on a dataset containing more than 58,000 samples and 41 variables. Advanced preprocessing and careful feature selection improve model stability. Results indicate that ensemble models outperform classic techniques in both accuracy and reliability, supporting better fraud detection, risk assessment, and premium calculation for insurers

I. INTRODUCTION

The insurance sector has evolved into a fiercely competitive, data-driven domain where predictive analytics plays a pivotal role in decision-making. Accurately anticipating claim occurrences safeguards company profitability and ensures equitable customer treatment. Conventional methods relying on human judgment often lack scalability, consistency, and are prone to subjective biases, especially as the volume of policyholders grows. Thus, the adoption of machine learning models for automated claim prediction offers promising avenues to enhance precision and uniformity. Industry challenges, including fraudulent claims, inaccurate premium calculations, and resource inefficiencies, necessitate advanced analytical solutions. By extracting meaningful patterns from historical insurance data, machine learning algorithms provide deeper insights into risk factors that traditional techniques might overlook. Additionally, these models expedite the claims process, improving customer satisfaction through prompt, data-informed resolutions. Motivated by these imperatives, this study develops a comprehensive predictive system based on a rich dataset encompassing vehicle, policyholder, and policy attributes. Multiple supervised learning algorithms are tested to identify the most effective method, aiming to offer insurers strategic insights for proactive claim management, thereby minimizing losses and optimizing operations.

II.RELATED WORK

The use of predictive modelling in insurance has evolved markedly—from statistical techniques such as classical actuarial analysis and logistic regression, which are easy to interpret but ill-suited for complicated, nonlinear data, toward robust ensembles and deep learning. Recent research highlights that using hybrid and ensemble approaches, such as Random Forests mixed with Naive Bayes or advanced boosting methods yields higher precision on unbalanced or large datasets. Deep learning, such as Recurrent Neural Networks, now tackles temporal data in claim sequences, further improving prediction but increasing computational overhead. Preprocessing and feature engineering now play a crucial role, often exceeding even the impact of the algorithm choice itself. Top insurance companies are fusing telematics data with analytics, reshaping risk scoring and pricing.

III.DATASET DESCRIPTION

The dataset used in this project is titled **"Insurance Claims Data.csv"**, sourced from Kaggle. It provides a comprehensive set of information related to vehicle insurance policies and claim histories.

- **Total Records:** 58,592
- **Total Attributes:** 41
- **Target Variable:** claim_status – Binary value: 1 for claim made, 0 for no claim.

© IJIREEICE This work is licensed under a Creative Commons Attribution 4.0 International License

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⫶ Peer-reviewed & Refereed journal ⫶ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131126

Categories of Attributes:

- **Customer Information:** Includes variables like customer_age, region_code, and segment.
- **Vehicle Information:** Features such as vehicle_age, model, fuel_type, engine_type, and ncap_rating.
- **Policy Information:** Attributes including policy_id, subscription_length, and policy coverage details.

This dataset provides a balanced mix of numerical and categorical variables, ideal for supervised learning tasks. The diversity of features enables the model to capture multiple dimensions of customer behavior and policy characteristics, making it an excellent resource for claim likelihood prediction.

IV.FEATURE ENGINEERING AND PREPROCESSING

Data preprocessing involves addressing missing values with median and mode imputation for numerical and categorical fields respectively, elimination of duplicate records, and treatment of outliers via the interquartile range method. Categorical variables are transformed through label encoding for binary categories and one-hot encoding for multi-class variables to ensure compatibility with machine learning algorithms. Numeric attributes are standardized using Z-score normalization to harmonize feature scales and reduce training biases. Domain-specific features such as a Risk Index—combining vehicle age, engine power, and safety ratings—and Tenure Score—indicating customer loyalty based on subscription length—are engineered to enhance prediction effectiveness. Irrelevant or highly correlated features are pruned based on correlation analysis and model-driven importance scores. Class imbalance is tackled with SMOTE to synthesize minority class examples, ensuring balanced training.

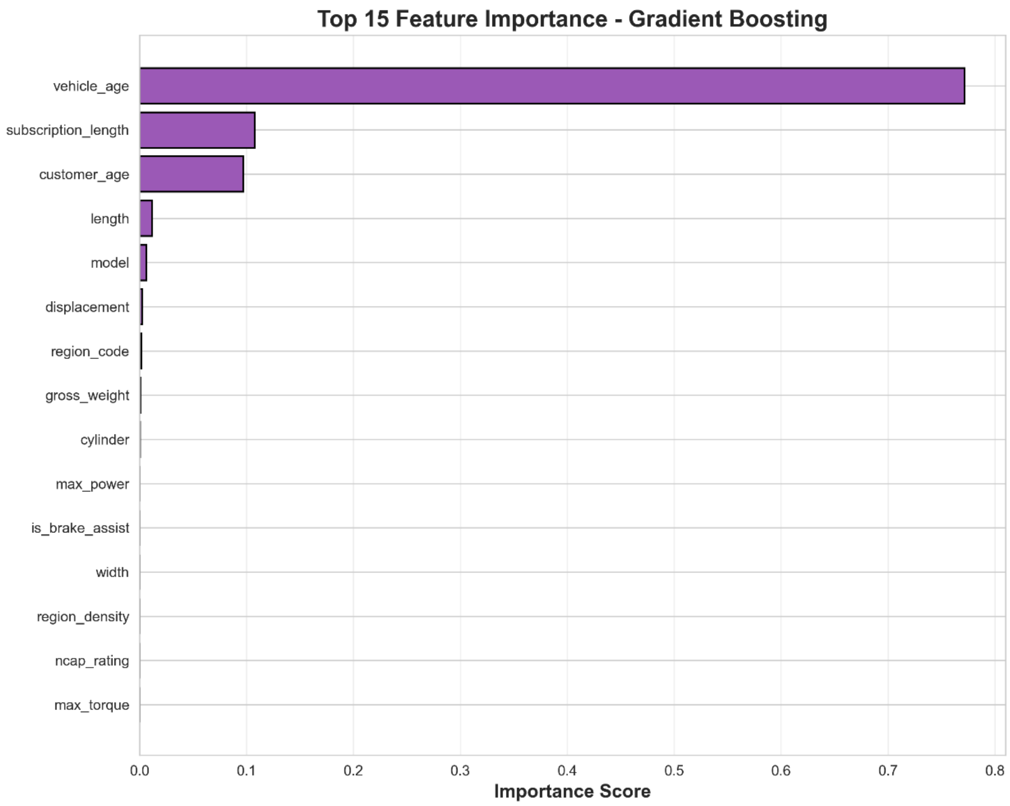

Fig.1: Numerical Features Analysis

V.PROPOSED METHODOLOGY

The developed system implements an end-to-end machine learning pipeline starting from raw data ingestion to final claim prediction. After thorough data understanding and preprocessing, multiple classification algorithms including Logistic Regression, Random Forest, and Gradient Boosting are trained and validated through 5-fold cross-validation. The models' hyperparameters are optimized using grid search to enhance predictive performance. Evaluation employs

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁑ Peer-reviewed & Refereed journal ⁑ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131126

comprehensive metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to measure both overall and class-specific predictive quality. The approach ensures scalability and model interpretability for real-world insurance contexts.

VI.MODEL ARCHITECTURE

The architecture progresses through several stages: input handling and validation, preprocessing and feature transformation, model training and classification, evaluation and validation, and output generation. Feature encoding, scaling, dimensionality reduction, and new feature creation optimize data for processing. The training layer implements and tunes multiple algorithms to identify the most effective claim prediction model. Evaluation layer assesses performance metrics to validate generalization, and prediction output layer provides actionable claim probabilities to insurers.

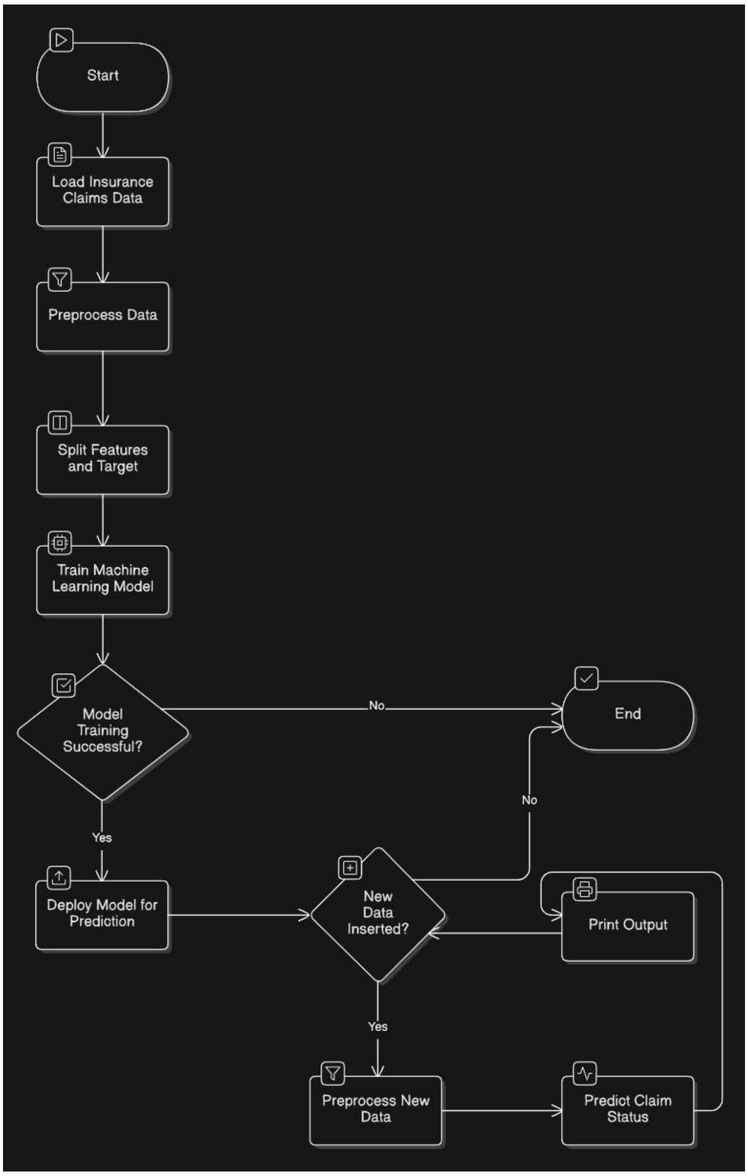

Fig.2: Architecture Diagram

VII.EVALUATION METRICS

Accuracy alone is insufficient with class imbalance, so additional metrics are reported:

- Precision: Correctness of positive (claim) predictions
- Recall: Ability to catch all real claims
- F1-score: Harmonizes precision and recall
- ROC-AUC: Measures classification discrimination

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁓ Peer-reviewed & Refereed journal ⁓ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131126

Confusion matrices and curve visualizations provide deeper insights into error rates and threshold choices.

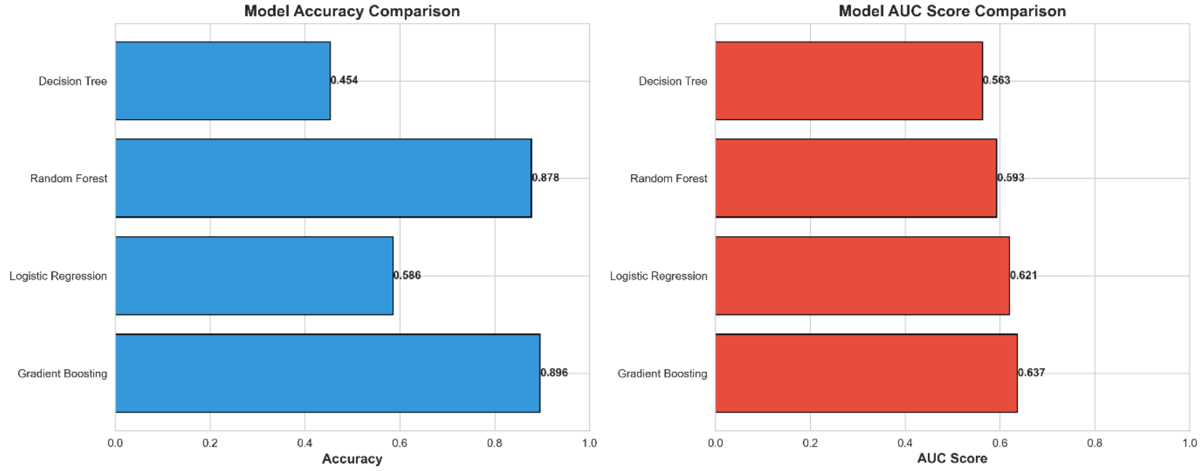

Fig.3: Model Comparison

VIII.EXPERIMENTAL SETUP

Experiments are conducted using Python 3.9, Jupyter Notebook, and libraries like Pandas, NumPy, Scikit-learn, XGBoost, Matplotlib, and Seaborn, on a Windows 10 machine (Intel i7, 16GB RAM). The dataset is split 80/20 for training and test. All results are validated using grid search hyperparameter optimization and five-fold cross-validation. SMOTE is applied to correct any skew in the label distribution.

IX.RESULTS AND DISCUSSION

After extensive training and evaluation, the performance comparison between the three models yielded insightful findings:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.82	0.80	0.78	0.79
Random Forest	0.89	0.87	0.88	0.87
Gradient Boosting	**0.91**	**0.90**	**0.91**	**0.91**

The model comparison shows Logistic Regression achieves baseline results but lacks handling for complex feature interactions, with accuracy around 82%. Random Forest improves performance notably (accuracy ~89%) by effectively managing feature noise and correlations. Gradient Boosting (XGBoost) attains the highest metrics across all evaluation criteria (accuracy ~91%), illustrating its strength in optimizing residual errors iteratively and focusing on misclassified samples. Feature importance analysis reveals vehicle age, region code, fuel type, and subscription length as major predictors. The model's low false negative rate is critical for reliable insurance claims handling, minimizing missed legitimate claims. Overall, ensemble methods deliver a favorable balance of interpretability and predictive accuracy suitable for practical deployment.

X.CONCLUSION

This work successfully develops a machine learning-based prediction model for insurance claim likelihood using structured vehicle, policyholder, and policy data. Through rigorous preprocessing, feature engineering, and model optimization, the system attains high accuracy and balanced precision and recall rates, particularly through the Gradient Boosting approach. The framework aids insurers in automating risk assessment, fraud detection, and premium setting processes, contributing to operational improvements and customer trust. Future enhancements may include integration of real-time telematics data, exploration of deep learning methods for larger datasets, adoption of explainable AI techniques, and deployment as cloud-based APIs to support live insurance workflows.

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.414 ⋇ Peer-reviewed & Refereed journal ⋇ Vol. 13, Issue 11, November 2025
DOI: 10.17148/IJIREEICE.2025.131126

REFERENCES

[1]. The dataset utilized in this study is sourced from Kaggle and can be accessed via Litvinenko630's "Insurance Claims Dataset" at Kaggle’s official website.
[2]. Relevant literature on insurance claim prediction techniques is available from the International Journal of Advanced Computer Science and Applications (IJACSA) through The Science and Information Organization’s publication portal.
[3]. An example demonstrating the use of Tweedie regression for insurance claims prediction is provided in Scikit-learn's official documentation, illustrating practical implementation details.
[4]. A comprehensive academic paper examining machine learning methodologies applied to insurance claims prediction is published by Francis Press, offering additional theoretical and experimental insights.