INSURANCE CLAIM PREDICTION

[1] T Manasa, [2] Domala Manikumar, [3] Yashwanth Reddy Keesara, [4] A Kalyani
Department of Computer Engineering(Software Engineering)
Siddhartha Institute of Technology & Sciences, Narapally
thirumanasa@siddhartha.org.in, 23tq1a5656@siddhartha.co.in, 23tq1a5661@siddhartha.co.in, 22TQ1A5653@siddhartha.co.in,

Abstract

Predicting insurance claims is a challenging task due to the dynamic and uncertain nature of policyholder behavior and associated risk factors. This project focuses on developing a machine learning-based approach to forecast insurance claim amounts and frequencies using *Linear Regression*. By leveraging historical data from 2020 to 2024, the system incorporates advanced feature engineering techniques, including analysis of past claim trends and risk-category mapping, to enhance prediction accuracy.

The developed model demonstrates strong performance, achieving an accuracy score of *0.94*, indicating its effectiveness in capturing patterns within insurance data. Additionally, interactive visualizations using Plotly are integrated to provide clear insights into claim trends by comparing actual and predicted values and identifying key loss patterns.

Overall, this project highlights the practical application of regression techniques in the insurance domain. It enables insurers to better anticipate financial risks, optimize premium pricing strategies, and improve decision-making processes, thereby contributing to enhanced operational efficiency and stability.

I. Introduction

Predicting insurance claims is a critical challenge in the insurance industry, as financial risks are influenced by a wide range of uncertain and dynamic factors, including policyholder behavior, socio-economic conditions, and unexpected environmental events. Many insurance providers face difficulties in accurately forecasting the frequency and cost of claims, which directly impacts their ability to set appropriate premiums, maintain sufficient financial reserves, and identify high-risk customers in advance. The absence of reliable prediction mechanisms can lead to financial instability, inefficient risk management, and uncompetitive pricing strategies.

This project addresses the problem by developing a machine learning-based predictive model capable of estimating future insurance claim amounts using historical data. By analyzing key attributes such as policy type, customer demographics, coverage details, and past claim history, the system aims to uncover hidden patterns and relationships that influence insurance costs. Various machine learning techniques, including Linear Regression, Random Forest, and Gradient Boosting, can be utilized to improve prediction accuracy and identify high-risk policyholders.

The primary objective of this project is to build a reliable and data-driven forecasting system that assists insurers in better understanding risk, anticipating potential financial losses, and making informed decisions. By enabling early detection of high-

cost claims and frequent claimants, the system supports proactive risk management, optimized premium pricing, and enhanced operational efficiency, ultimately contributing to improved stability and service quality in the insurance sector.

II. Literature Survey

A considerable amount of research has been conducted on the application of machine learning techniques in the insurance domain, particularly for claim prediction and risk assessment. These studies provide a strong foundation for developing accurate and data-driven insurance forecasting systems.

Frees, E.W., et al. (2014), in their work on *predictive modeling in insurance*, highlighted the importance of statistical and machine learning models in estimating claim frequency and severity. Their study emphasized the role of regression-based approaches, such as Generalized Linear Models (GLMs), in handling structured insurance data and providing interpretable results for actuarial analysis.

Verbelen, R., et al. (2018), in *"Predictive Modeling for Claim Counts using Machine Learning,"* compared traditional actuarial models with advanced machine learning techniques such as Random Forests and Gradient Boosting. Their findings showed that ensemble methods outperform classical models by capturing complex non-linear relationships between policyholder attributes and claim occurrences.

Henckaerts, R., et al. (2020), in their research on *"Machine Learning Techniques for Insurance Claim Prediction,"* demonstrated that Gradient Boosting and Neural Networks significantly improve prediction accuracy for claim severity and frequency. The study also highlighted the importance of feature engineering and data preprocessing in achieving reliable results.

Brockman, M.J., and Wright, T.S. (1992), in an earlier study on *insurance pricing models*, explored the use of regression techniques to estimate expected claim costs. Their work laid the foundation for modern predictive analytics in insurance, emphasizing the need for data-driven decision-making in premium calculation.

III. System Analysis

System analysis for the insurance claim prediction system focuses on understanding the challenges involved in estimating future claim amounts and designing a data-driven solution to address them. Insurance claims are influenced by multiple factors such as policyholder demographics, policy type, historical claim records, and external risk conditions. Traditional systems often fail to analyze these factors collectively, resulting in inaccurate forecasting. This system leverages historical insurance data to identify patterns and relationships that influence claim frequency and severity. Machine learning techniques are applied to process large datasets efficiently and generate accurate predictions. The analysis includes data collection, preprocessing, feature engineering, and model training to ensure optimal performance. It also considers scalability and reliability for real-world deployment. The primary objective is to assist insurers in predicting potential claims, managing risks effectively, and improving financial planning. By enabling data-driven insights, the system enhances decision-making and operational efficiency.

Existing System

The existing system for predicting insurance claims is largely based on traditional statistical methods and manual analysis. Insurance companies often rely on historical averages, actuarial tables, and expert judgment to estimate claim amounts and risks. These methods typically consider only a limited number of variables and fail to capture complex relationships between different factors. The system lacks automation and real-time predictive capabilities, making it inefficient for handling large volumes of data. Additionally, traditional approaches are often static and do not adapt well to changing market conditions or customer behavior. Human involvement in decision-making can introduce bias and inconsistency. There is also limited use of advanced analytics or machine learning techniques. As a result, predictions may be inaccurate, leading to improper premium pricing and poor risk management. Overall, the existing system struggles to provide precise and timely insights for effective decision-making.

Disadvantages of Existing System

- Limited accuracy due to reliance on traditional methods
- Inability to capture complex relationships between variables
- Heavy dependence on manual analysis and expert judgment
- Lack of automation and real-time prediction
- Poor handling of large and complex datasets
- High chances of human bias and inconsistency
- Static models that do not adapt to changing trends

Proposed System

The proposed system is a machine learning-based insurance claim prediction model designed to provide accurate and data-driven forecasts of claim amounts and frequency. It utilizes historical insurance data, including policy details, customer demographics, and past claim records, to identify patterns influencing insurance risk. The system begins with data preprocessing to clean and transform the dataset, followed by feature engineering to create meaningful variables such as risk categories and historical claim trends. Machine learning algorithms, particularly Linear Regression and ensemble methods, are used to train the model and capture relationships within the data. The trained model predicts future claim values and helps identify high-risk policyholders. Additionally, the system integrates visualization tools such as Plotly to present insights through interactive graphs. It can be deployed in real-time applications, enabling insurers to monitor risks continuously.

Advantages of Proposed System

- High accuracy using machine learning models
- Ability to analyze multiple factors simultaneously
- Reduces human bias and manual effort
- Provides real-time and data-driven predictions
- Handles large datasets efficiently
- Improves risk assessment and forecasting
- Helps in optimal premium pricing
- Enhances decision-making for insurers
- Scalable and adaptable to changing data trends

IV. Methodology

The methodology for the insurance claim prediction system follows a systematic machine learning pipeline to ensure accurate and reliable forecasting of claim amounts. Initially, historical insurance data is collected, including policy details, customer demographics, claim history, and risk-related attributes. This data is then preprocessed by handling missing values, removing inconsistencies, and encoding categorical variables into numerical formats.

Next, Exploratory Data Analysis (EDA) is performed to understand data distribution, identify trends, and detect relationships between features such as policy type, age, and claim frequency. Feature engineering is applied to create meaningful variables such as risk categories, claim ratios, and historical trends that improve model performance.

The dataset is then split into training and testing sets to evaluate the model effectively. Machine learning algorithms, primarily Linear Regression, along with advanced models like Random Forest and Gradient Boosting, are used to train the prediction system. The model learns patterns from historical data and predicts future claim amounts or probabilities.

System Architecture

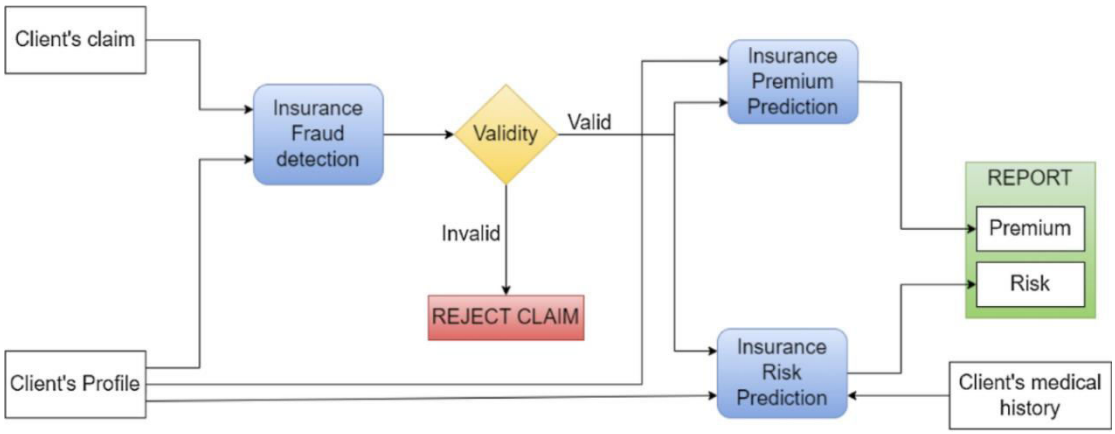
The system architecture is designed as a multi-layered pipeline that ensures smooth data processing and accurate predictions.

1. Data Collection Layer : Collects data from insurance databases, policy records, and claim history datasets.

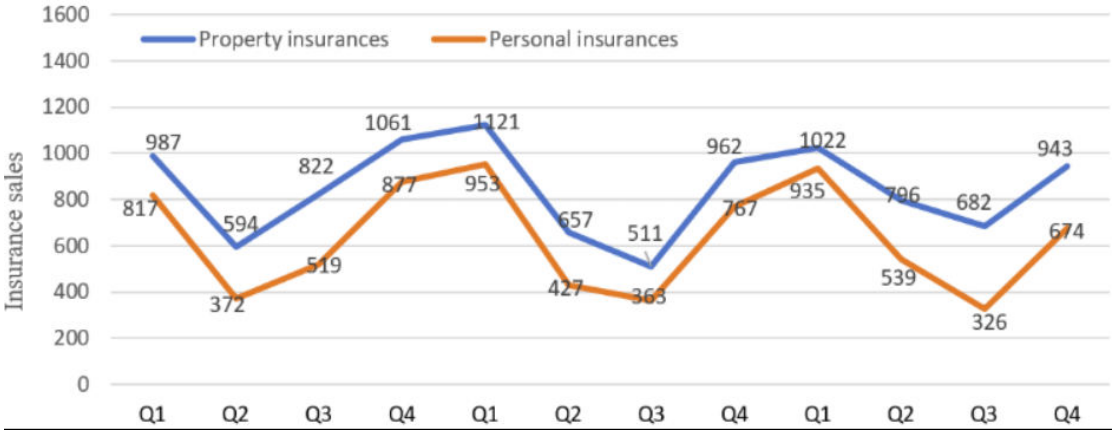
2. Data Storage Layer : Stores structured data in databases or data warehouses for efficient access and processing.

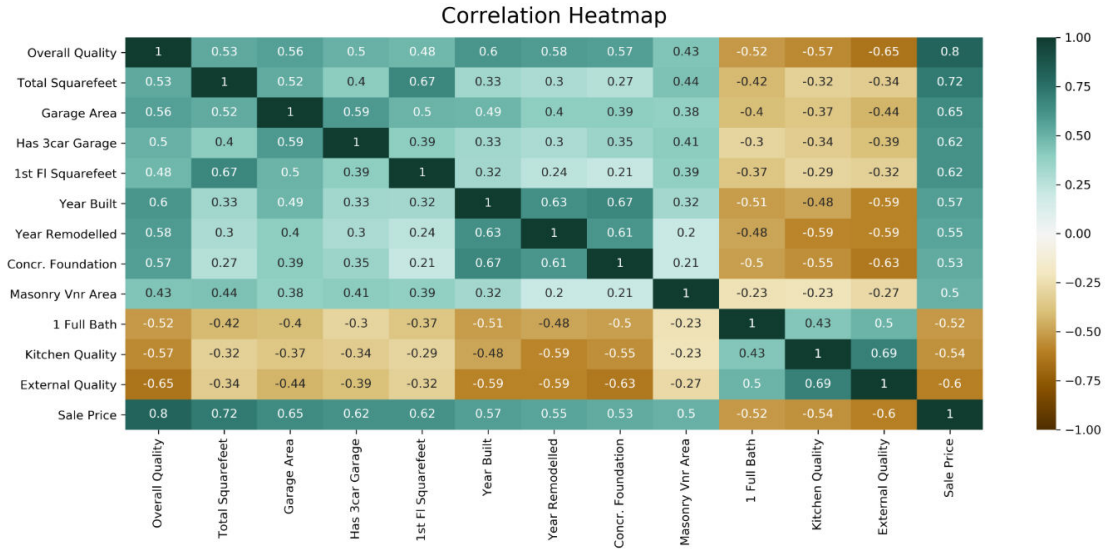
3. Data Preprocessing Layer : Cleans the data, handles missing values, removes duplicates, and encodes categorical variables.

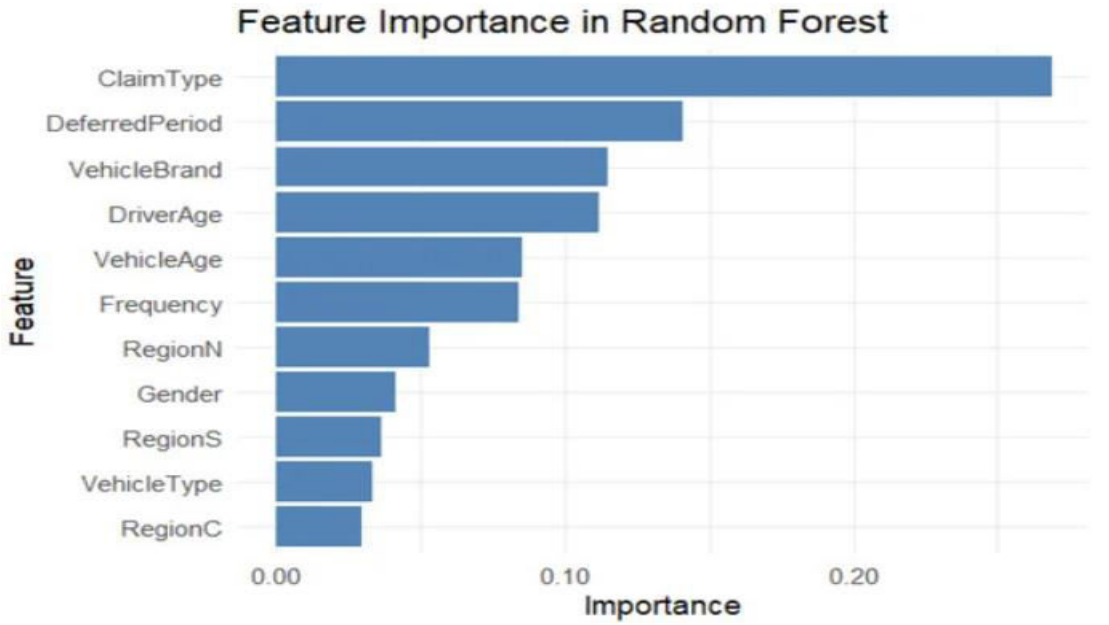
4. Feature Engineering Layer : Creates new features such as risk categories, claim ratios, and historical trends.

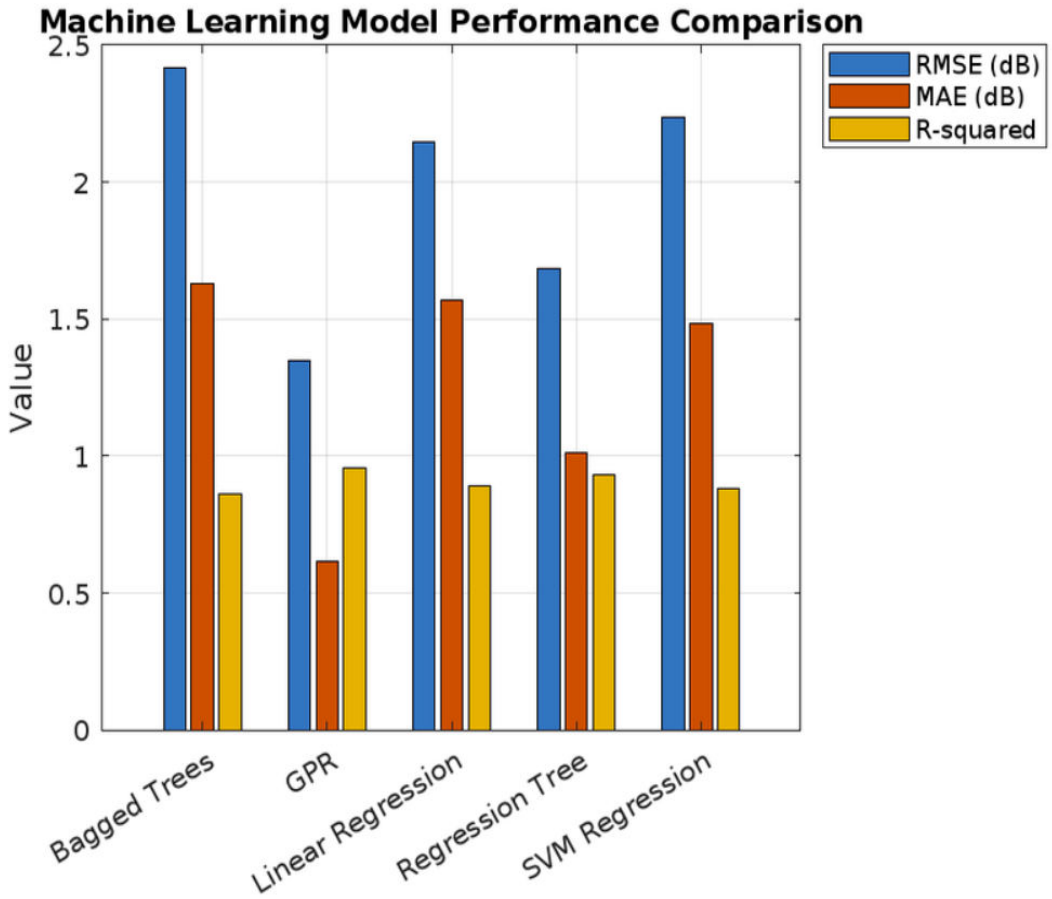


V. Result and Output









VI. Conclusion

This project successfully demonstrates that, despite the inherent uncertainty in human behavior and environmental risks, insurance claims can be effectively predicted using data-driven approaches. By applying machine learning techniques to four years of historical data, the model achieved a high R2R^2R2 score of *0.94*, confirming its ability to accurately capture patterns in insurance claim behavior. This highlights the potential of advanced analytics to transform traditional risk assessment methods into more precise and reliable forecasting systems.

Beyond numerical performance, the project emphasizes the broader impact of predictive modeling in the insurance domain. For insurance providers, it enables improved financial planning, optimized capital allocation, and reduced exposure to unexpected losses. For policyholders, it ensures fairer and more transparent premium pricing based on actual risk patterns rather than generalized assumptions.

Overall, this work underscores the importance of leveraging machine learning responsibly to enhance stability, transparency, and efficiency within the insurance industry. By shifting from reactive claims management to proactive risk prediction, the system provides a strong foundation for future advancements and supports the development of a more resilient and equitable financial ecosystem.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, “Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology,” in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, “An advanced movie recommender using collaborative filtering and sentiment analysis,” International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, “Automatic crop recommendation system using LightGBM and decision tree machine learning models,” Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, “Smart agriculture through IoT and machine learning for analyzing carbon footprints,” in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, “Soil image classification using transfer learning approach: MobileNetV2 with CNN,” SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.