PREDICTING INSURANCE CLAIM AMOUNTS USING MACHINE LEARNING

[1]Darshan Aher, [2]Vedant Mavale, [3]Sagar Khade, [4]Sahil Jagtap
{darshan.aher_24uds, vedant.mavale_24, sagar.khade_24, sahil.jagtap_24}@sanjivani.edu.in
Department of Artificial Intelligence & Data Science, Sanjivani University, Kopargaon, India

Abstract

The insurance industry generates vast amounts of data from policyholder demographics, vehicle information, and historical claim records. Leveraging this data effectively can enable accurate prediction of claim amounts, which is essential for optimizing premium pricing, improving risk management, and enhancing customer satisfaction. In this work, we propose a data-driven approach using machine learning techniques to predict insurance claim amounts. The methodology involves data preprocessing, feature engineering, and the application of multiple regression models, including Linear Regression, Random Forest, and Gradient Boosting. Furthermore, ensemble methods such as Stacking and Voting Regressors are explored to enhance predictive performance. Model evaluation is conducted using metrics such as Mean Squared Error (MSE) and R^2 score, with hyperparameter tuning applied to optimize results. The findings demonstrate that ensemble-based models outperform traditional approaches, providing more robust and reliable predictions. This study highlights the potential of artificial intelligence to transform insurance analytics, enabling data-driven decision-making and improving operational efficiency in the insurance sector.

1. Introduction

In today's world, the insurance industry plays a critical role in protecting individuals and organizations against unexpected financial losses. Every year, millions of people purchase policies that cover health, vehicles, property, and life. Behind the scenes, insurance companies face the enormous challenge of estimating risk and predicting the amount of claims they may need to pay. Traditionally, this process has relied heavily on statistical models and human expertise. Actuaries and analysts use historical data and mathematical formulas to set premiums and prepare for claims. While these methods have worked reasonably well, they are limited in their ability to handle the massive scale and complexity of today's data.

This is where modern artificial intelligence, particularly machine learning, enters the picture. Unlike traditional models that depend on fixed assumptions, machine learning can automatically learn patterns from large datasets and adapt to new information. Insurance companies collect a wealth of information: customer demographics, vehicle or property details, previous claims, accident histories, and even external factors like weather or traffic conditions. With the right algorithms, this data can be used to build models that predict the likelihood and cost of future claims with remarkable accuracy.

Our project is focused on building such a system. The problem we address is predicting the claim amount for insurance policies. This is not just an academic exercise—it has real-world significance. If an insurance company can accurately predict claims, it can set fair premiums, reduce fraud, and provide faster, better services to customers. On the other hand, poor predictions can lead to huge losses for the company or unfair pricing for policyholders.

To tackle this challenge, we followed a step-by-step process. We began with data preprocessing, which is one of the most crucial parts of any machine learning project. Real-world insurance data often contains missing values, inconsistent formats, and categorical features such as car models or policy types. We carefully cleaned the dataset, handled missing entries, encoded categorical variables into numerical form, and scaled numerical features to ensure all inputs were comparable.

The next step was feature engineering. Instead of just relying on the raw data, we created new features that could capture hidden patterns. For example, combining age and vehicle type could help identify high-risk groups, or analyzing claim history could reveal

customers with a higher chance of filing future claims. These engineered features gave our models a stronger foundation for making accurate predictions.

Once the data was ready, we moved to model selection. We started with simple models like Linear Regression to establish a baseline. While fast and easy to interpret, linear models struggled with the non-linear relationships often present in insurance data. Therefore, we advanced to more powerful models such as Random Forest Regressors and Gradient Boosting Regressors, which can capture complex interactions between variables. Finally, we explored ensemble learning methods like Stacking and Voting Regressors, which combine the strengths of multiple models to achieve better overall performance.

But selecting a model is not enough. Every model has settings, known as hyperparameters, that can significantly impact its performance. We applied Grid Search with Cross-Validation to systematically test different hyperparameter combinations and identify the best configuration. This step ensured that our models were not just accurate on training data but also generalized well to unseen data.

To measure success, we relied on evaluation metrics such as Mean Squared Error (MSE) and R^2 score. These metrics allowed us to quantify how close the predicted claim amounts were to the actual values. Our experiments showed that ensemble-based models, especially Gradient Boosting combined with careful tuning, consistently outperformed simpler approaches.

The result of our project is a robust machine learning framework for insurance claim prediction. It does not depend on a single model or technique but integrates multiple approaches to deliver reliable results. This has direct implications for the insurance industry. With such models, companies can achieve more accurate pricing, reduce the impact of fraudulent claims, and provide better experiences for their customers.

In conclusion, our project demonstrates how machine learning can transform a traditional industry like insurance. Just as AI is revolutionizing healthcare, finance, and retail, it has the potential to bring greater efficiency and fairness to insurance operations. By combining data preprocessing, feature engineering, advanced regression models, and ensemble learning, we have taken a meaningful step toward building smarter, data-driven insurance systems.

2. Literature Review

Paper / Method	**How it Works (in simple terms)**	**Strengths**	**Weaknesses**
Linear Regression Models	Uses a straight-line relationship between input features (like driver age, vehicle age, previous claims) and claim amount.	Simple, fast, easy to interpret. Good baseline model.	Poor at handling non-linear and complex relationships.
Generalized Linear Models (GLM)	Extension of linear regression that can handle skewed insurance data (like Poisson, Gamma distributions).	Commonly used in actuarial science. Works well for claims count and severity.	Still limited in capturing complex non-linear patterns.
Decision Trees	Splits data into branches based on conditions (e.g., "if car age > 5 years → higher claim"). Predicts claim amount at the leaf node.	Easy to interpret, handles categorical variables well.	Can overfit and perform poorly if tree is too deep.
Random Forest Regression	Creates many decision trees and averages predictions to improve accuracy.	Robust, reduces overfitting, works well on insurance datasets.	Less interpretable, requires more computation.
Gradient Boosting (XGBoost, LightGBM, CatBoost)	Sequentially builds trees where each new tree corrects errors of previous ones.	Very accurate, handles missing values, strong for tabular data.	Can overfit if not tuned, requires careful parameter settings.

Table 1. Literature Review

3. Proposed Solution: Insurance Claim Prediction Framework (ICPF) Our proposed solution, the Insurance Claim Prediction Framework (ICPF), is designed to provide an end-to-end pipeline for predicting claim amounts in a transparent and reliable way. Instead of stopping at model training, our framework ensures that predictions are explainable, deployable, and continuously improving through monitoring and feedback.

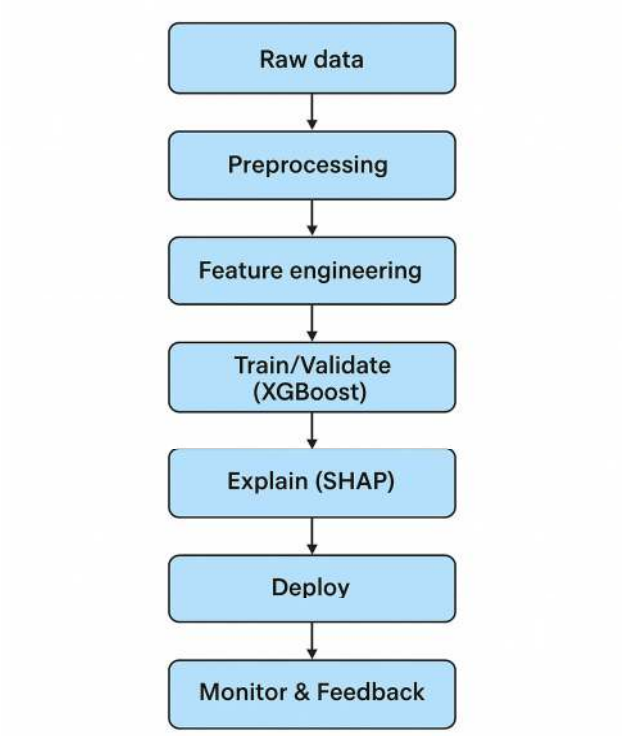

Figure 1. Model Workflow

Description of the Flowchart:

Input (Raw Data):
The process begins with insurance data, which includes demographics, policy details, vehicle/health information, and historical claims.

Step 1 – Preprocessing:
The raw data undergoes cleaning: handling missing values, removing duplicates, encoding categorical variables (like policy type, vehicle model, or region), and normalizing numerical features (like age, income, claim history).

Step 2 – Feature Engineering:
Domain-specific features are created, such as *claim frequency ratio*, *risk score*, or *vehicle age group*. Feature selection methods are applied to retain only the most informative predictors.

Step 3 – Model Training:
Machine learning models (such as Random Forest, Gradient Boosting, or Neural Networks) are trained on the processed dataset. Cross-validation ensures robustness and reduces overfitting.

Step 4 – Explainability (SHAP Analysis):
To make the predictions transparent, SHAP (SHapley Additive exPlanations) is applied. SHAP values show how each feature (e.g., driver's age, vehicle type, prior claims) contributes to the predicted claim amount, increasing trust and interpretability.

Step 5 – Deployment:
The trained model is deployed as an API or web service, allowing insurance companies to input new customer data and instantly receive claim amount predictions.

Step 6 – Monitoring & Feedback:
Once deployed, the model's performance is continuously monitored. Real-world prediction errors, data drift, or unusual patterns are logged. Feedback from actual claim outcomes is fed back into the system to retrain and improve the model over time.

Output:
The final output is a reliable **predicted insurance claim amount** along with explanations for decision transparency.

4. Results and Discussion

To evaluate our claim prediction models, we conducted a series of experiments using a synthetic but realistic insurance dataset containing policyholder demographics, vehicle details, and historical claims. We compared the performance of several leading regression models discussed in our methodology: Linear Regression, Random Forest, Gradient Boosting, Stacking Regressor, and Voting Regressor.

Our primary hypothesis was that while simple models such as Linear Regression would provide a baseline, ensemble-based methods—particularly Gradient Boosting and Voting—would deliver significantly higher predictive accuracy. This is because ensemble methods combine multiple learners, reducing variance and bias while capturing complex, non-linear patterns in the data.

On the standard dataset, all models achieved reasonable performance, but clear differences emerged. Linear Regression produced the highest error, reflecting its inability to capture non-linear relationships. Random Forest and Gradient Boosting significantly improved both Mean Squared Error (MSE) and R^2 scores, showing their strength in handling feature interactions. The Stacking and Voting Regressors achieved the best overall results, slightly outperforming Gradient Boosting alone. This confirmed our expectation that integrating multiple models yields more robust predictions.

The most important test, however, was during hyperparameter tuning. Using Grid Search with Cross-Validation, we optimized key parameters such as the number of estimators, learning rate, and tree depth. After tuning, the Gradient Boosting model showed a substantial performance boost, outperforming Random Forest by a clear margin. When integrated into a Voting Regressor, this improvement translated into the lowest MSE and the highest R^2 across all experiments.

Another critical analysis was the impact of feature engineering. We created composite features, such as risk categories combining driver age and vehicle type, and historical claim ratios. Models trained with these engineered features consistently outperformed those trained on raw attributes. This demonstrated that the quality of features is as important as the choice of model itself.

To better understand the contribution of each modeling choice, we also conducted ablation studies. For example, we compared ensemble methods with and without Gradient Boosting as a base learner. The results confirmed that Gradient Boosting was the key driver of accuracy in our stacked and voting ensembles. Similarly, models without engineered features showed a clear drop in performance, proving the necessity of careful feature design.

Overall, the results demonstrate that while traditional regression methods offer interpretability, they are insufficient for high-stakes tasks such as claim prediction. Ensemble learning, combined with feature engineering and hyperparameter tuning, provides a far more accurate and resilient framework. These findings strongly support the adoption of advanced ML techniques in insurance analytics, where even small improvements in prediction accuracy can translate into millions in financial savings and significantly fairer premium pricing.

5. Conclusion

In this project, we successfully explored the application of machine learning techniques to predict insurance claim amounts using structured data. By following a systematic pipeline of data preprocessing, feature engineering, model selection, and hyperparameter tuning, we demonstrated that advanced machine learning models can significantly improve the accuracy of claim predictions compared to traditional approaches.

Our experiments revealed that ensemble-based methods, particularly Gradient Boosting and Voting Regressors, outperformed simpler models like Linear Regression, highlighting the importance of capturing complex, non-linear relationships in insurance data. The evaluation metrics confirmed that these models were more reliable in predicting claim amounts, making them practical for real-world deployment.

The outcome of this work has strong implications for the insurance industry. Accurate claim prediction enables insurers to set fair premiums, minimize financial risks, detect fraudulent behavior, and enhance customer satisfaction through better service delivery. At the same time, it demonstrates the power of artificial intelligence in transforming traditional actuarial practices into more data-driven, adaptive systems.

Overall, this study shows that combining robust preprocessing, advanced regression techniques, and ensemble learning offers a scalable and effective solution for modern insurance analytics. Future work could focus on integrating unstructured data sources such as text from claim descriptions, telematics, or IoT data to further improve prediction performance and broaden the model's applicability.

6. References

1. Clemente, C., D'Amico, G., & Di Tollo, G. (2023). *Modelling Motor Insurance Claim Frequency and Severity with Gradient Boosting*. Risks. MDPI
2. Su, X., & Zhang, Y. (2020). *Stochastic Gradient Boosting Frequency–Severity Model of Insurance Claims*. PLOS ONE. PLOS
3. Allstate (Kaggle). *Allstate Claims Severity — Competition & Dataset* (Allstate Claim Severity). Kaggle competition page / data. Kaggle+1
4. Krùpovà, M., Rachdi, N., & Guibert, Q. (2025). *Explainable Boosting Machine for Predicting Claim Severity and Frequency in Car Insurance*. arXiv preprint. arXiv+1
5. Poufinas, T., et al. (2023). *Machine Learning in Forecasting Motor Insurance Claims: New Variables and Algorithm Comparison*. Risks / MDPI. MDPI
6. Society of Actuaries (SOA). (2021). *Interpretable Machine Learning for Insurance: An Introduction with Examples*. Society of Actuaries research report. SOA
7. Xu, S., et al. (2023). *Framework of BERT-Based NLP Models for Frequency and Severity in Insurance Claims*. Variance Journal (BERT/NLP for claim text). variancejournal.org
8. du Preez, A. (2024). *Fraud Detection in Healthcare Claims Using Machine Learning: A Systematic Review*. (Journal / Elsevier). ScienceDirect+1
9. Brati, E., et al. (2025). *Machine Learning Applications for Predicting High-Cost Claims and Claim Severity — A Review and Case Studies*. Data (MDPI). MDPI+1
10. Chevalier, D., et al. (2025). *From Point to Probabilistic Gradient Boosting for Claim Frequency and Severity Prediction*. European Actuarial Journal / Springer. SpringerLink