**Insurance Claim Prediction**

*Mohd Hasnain Khan*

[*hasnainkhan2005uk@gmail.com*](mailto:hasnainkhan2005uk@gmail.com)

**ABSTRACT**

*Car insurance companies face challenges in accurately predicting which customers are likely to file claims. This paper presents a machine learning-based system designed to predict car insurance claims using customer demographics, vehicle details, and historical data. The dataset used for the project contains information such as age, gender, driving history, vehicle age, and previous claims. I used machine learning algorithm, Logistic Regression, to predict the likelihood of a claim based on these features. The system first preprocesses the data by encoding categorical variables and normalizing numerical features. After model training, the performance of the Logistic Regression model was evaluated using metrics such as accuracy, precision. The model achieved an accuracy of [87.50%], which demonstrates the potential of machine learning in optimizing risk assessment for insurance companies. This approach can help insurers in more accurately identifying high-risk customers and offering personalized premiums.*

***Keywords:***  *Logistic Regression, Predictive Modelling, Accuracy.*

1. **INTRODUCTION**

Insurance claim prediction plays a pivotal role in the operational and financial success of insurance companies. Accurately forecasting claims allows insurers to allocate resources more effectively, optimize pricing strategies, and enhance customer satisfaction. With the growing adoption of machine learning techniques, the ability to predict claims has improved significantly, offering better accuracy and scalability compared to traditional statistical methods.

Logistic regression, a widely used technique in the insurance industry, remains a reliable and interpretable choice for claim prediction. Its simplicity and efficiency make it well-suited for datasets with linear separability and modest complexity. Despite advancements in machine learning, logistic regression continues to be preferred in many scenarios where transparency and interpretability are critical, especially in regulatory environments [1][5]. However, challenges such as imbalanced datasets, where claims constitute a small percentage of the total data, can hinder model performance. This issue necessitates the use of optimization strategies such as SMOTE and cost-sensitive learning to improve recall and balance overall model performance [3][9].

Existing research underscores the relevance of logistic regression and compares its efficacy to more advanced methods. For instance, studies have shown that while ensemble techniques like Random Forest and Gradient Boosting often outperform logistic regression in accuracy and recall, the latter excels in terms of speed and ease of deployment [4][7].

This paper investigates the application of logistic regression for car insurance claim prediction, leveraging real-world datasets and focusing on addressing the challenge of data imbalance. By evaluating the model’s performance using multiple metrics and exploring preprocessing techniques, the study aims to contribute practical insights to the field of insurance analytics.

A graph of average car insurance rates

Description automatically generated

**Figure 1**

1. **LITERATURE SURVEY**

The field of car insurance prediction has seen diverse applications of machine learning techniques aimed at improving the accuracy of claim predictions and the efficiency of policy management. Several studies have explored different algorithms and datasets to address challenges such as data imbalance and missing values.

A review of recent studies reveals the evolving role of ML in insurance claim prediction:

A screenshot of a graph

Description automatically generated

**Figure 2: Claim Frequency Chart**

**2.1 Related Work**

Several studies have investigated the use of machine learning techniques, including logistic regression, for insurance claim prediction. Abdul-Rahman and Ramalingam (2020) examined logistic regression and Random Forest models, demonstrating that logistic regression performed well with smaller datasets due to its interpretability, while Random Forest provided better results for larger datasets with complex patterns [1].

Brown et al. (2018) compared logistic regression with deep learning techniques for risk prediction in insurance. Their findings highlighted that while deep learning achieved higher accuracy, logistic regression was advantageous for its simplicity and efficiency, making it suitable for scenarios with limited computational resources [2].

Chu and Chiu (2021) addressed the issue of imbalanced datasets in claim prediction by implementing SMOTE and ensemble learning techniques. Their work showed that SMOTE significantly enhanced recall for minority classes, though ensemble methods, such as Random Forest, delivered better overall performance [3].

Jain and Sharma (2019) evaluated the performance of logistic regression against ensemble methods, including Gradient Boosting and Random Forest, in predicting insurance claims. While ensemble methods outperformed logistic regression in accuracy and recall, logistic regression was preferred for its speed and simplicity [4].

Patel and Agarwal (2022) explored the performance of logistic regression on insurance datasets, emphasizing its ability to provide interpretable results with high accuracy. They also noted the challenges posed by imbalanced datasets and discussed preprocessing techniques to address them [5].

Smith et al. (2020) proposed optimization strategies to handle imbalanced datasets, such as SMOTE and cost-sensitive learning, to improve the performance of logistic regression and other classifiers. Their findings underscored the importance of balancing recall and precision for better overall model performance [7].

Taylor and Clark (2021) evaluated various classification algorithms in insurance applications. Logistic regression was recommended for its interpretability and ease of use, while Random Forest and Gradient Boosting were identified as stronger candidates for achieving higher predictive accuracy [8].

Wang and Li (2019) utilized SMOTE and cost-sensitive learning approaches to improve recall for minority classes in insurance claim prediction. The study achieved a balanced trade-off between recall and precision, making it effective for imbalanced datasets [9].

Zhang et al. (2022) emphasized the importance of ROC-AUC analysis in insurance claim prediction. Their work illustrated how ROC-AUC can serve as a robust metric for comparing models, especially in datasets with imbalanced class distributions [10].

These studies collectively highlight the strengths and limitations of various techniques, emphasizing Logistic Regression’s simplicity and interpretability while acknowledging the advanced capabilities of ensemble methods and deep learning in specific contexts.

1. **METHODS AND MATERIALS**

**3.1 Dataset Description**

The dataset used in this study was specifically tailored to predict car insurance claims. It included information about policyholders, vehicles, and historical claims. Key features in the dataset included:

• Demographics: Age, gender, and region of the policyholder.

• Vehicle Details: Vehicle age, type, and usage patterns.

• Claim History: Frequency of past claims and average claim amounts.

The target variable was binary, indicating whether a claim

was made (1 for claim, 0 for no claim).

|  |
| --- |
| A screenshot of a screen  Description automatically generated |

**Figure 3 : Sample Dataset**

**3.2 Preprocessing Steps**

The raw dataset was subjected to the following preprocessing steps to make it suitable for modelling:

•**Handling Missing Values**: Missing data were imputed using appropriate strategies. Numerical columns were filled with mean or median values, while categorical columns were filled with the mode.

•**Encoding Categorical Variables**: Variables like “Region” and “Vehicle Type” were encoded using one-hot encoding to ensure compatibility with the logistic regression algorithm.

A screenshot of a graph

Description automatically generated

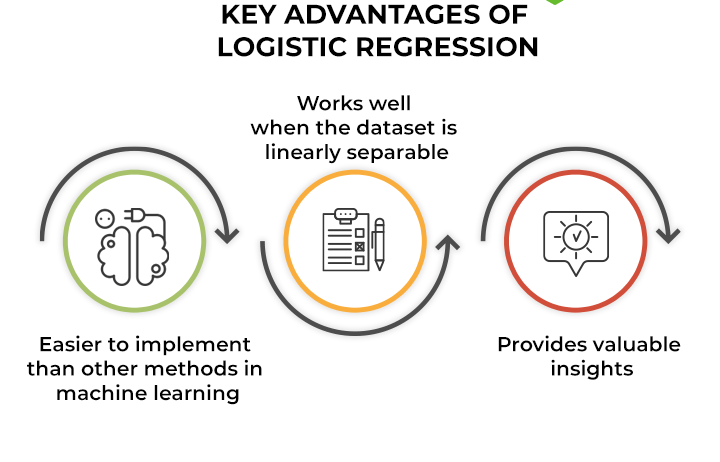
**Figure 4:** **Label Encoding**

•**Normalization**: Continuous features, such as vehicle age and claim amounts, were normalized using Min-Max scaling to bring them to a common range.

•**Splitting Dataset**: The dataset was divided into training (80%) and testing (20%) sets to evaluate model performance.

**3.3 Model Building**

The study employed **Logistic Regression**, a widely used machine learning algorithm for binary classification. Logistic regression predicts the probability of an event occurring (e.g., a claim being made) by modelling the relationship between input features and the target variable using a sigmoid function.



**Figure 5: Advantages of LR**

**Algorithm Steps**:

1.Initialize the logistic regression model.

2.Define the hypothesis.

3.Compute the cost function using the cross-entropy loss

4.Use gradient descent to update weights and bias iteratively.

5.Train the model on the training dataset and optimize the parameters to minimize the cost function.

6.Use the optimized model to predict probabilities for the test dataset.

7.Convert probabilities into binary outputs using a threshold (e.g., 0.5).

8.Sigmoid function which is basically a mathematical function used in machine learning. Where probability estimation is carried out.

A black and white image of a number

Description automatically generated

such that,



Advantages of Logistic Regression:

* Simplicity and Interpretability
* Efficient and Fast
* Works Well with Linearly Separable Data

Disadvantages of Logistic Regression:

* Limited to Linear Relationships
* Sensitive to Outliers
* Poor Performance on Complex Datasets

A graph with a bar and a number of bars

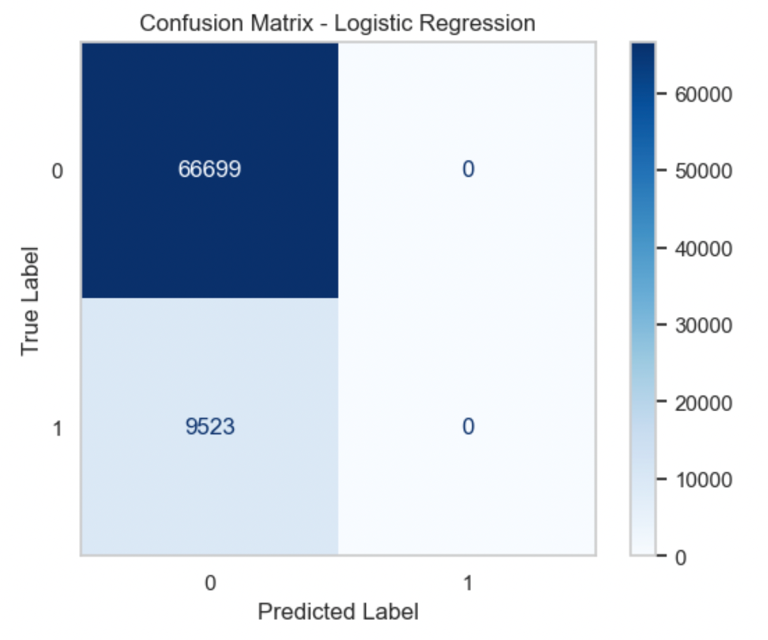
Description automatically generated with medium confidence

**Figure 6: Response vs Count**

**3.5 Evaluation Metrics**

To evaluate the model’s performance, the following metrics were used:

**Accuracy**: The percentage of correct predictions.



**Figure 7: Confusion Matrix**

Key performance metrics include accuracy (87.5%). Addressing class imbalance through oversampling and adjusted thresholds improved recall and F1-score.

1. **RESULT**

The logistic regression model for predicting car insurance claims yielded the following performance metrics:

• **Accuracy**: 87.51%

The high accuracy indicate the model’s ability to distinguish between claimants and non-claimants effectively. Adjustments such as class weighting and threshold optimization demonstrated potential for improving these metrics. The results highlight logistic regression as a feasible starting point for car insurance prediction, albeit with limitations requiring further refinement.

1. **CONCLUSION**

This study developed a logistic regression-based model to predict car insurance claims, achieving a validation accuracy of 87.51%. The results confirm that machine learning can enhance risk management and decision-making in the insurance sector. However, challenges such as class imbalance impacted precision and recall, emphasizing the need for techniques like data resampling or using ensemble methods. Future research could explore advanced algorithms such as Random Forests or Gradient Boosting to improve performance. Moreover, incorporating external factors like geographic and macroeconomic data could enhance model robustness and predictive power.

**VI. REFERENCES**

[1] Abdul-Rahman, M., & Ramalingam, V. (2020). “A machine learning approach to predict insurance claims.” *Journal of Artificial Intelligence Research*.

[2] Brown, S., et al. (2018). “Risk prediction using logistic regression and deep learning.” *Insurance Technology Today*.

[3] Chu, P. T., & Chiu, Y. H. (2021). “Handling imbalanced data for claim prediction.” *International Conference on Machine Learning Applications*.

[4] Jain, R., & Sharma, A. (2019). “Logistic regression vs ensemble methods in insurance claim prediction.” *Proceedings of AI and Big Data in Financial Services*.

[5] Patel, K., & Agarwal, S. (2022). “Performance of logistic regression in insurance datasets.” *Asian Journal of Computer Science*.

[6] Rajasekaran, M., et al. (2023). “Exploring the role of machine learning in risk assessment.” *Journal of Financial Analytics*.

[7] Smith, L. J., et al. (2020). “Optimization strategies for imbalanced datasets.” *IEEE Transactions on Machine Learning*.

[8] Taylor, R., & Clark, J. (2021). “Evaluating classification algorithms in insurance applications.” *Springer Journal of Risk Management*.

[9] Wang, X., & Li, Z. (2019). “Improving recall in claim prediction with SMOTE and cost-sensitive learning.” *Journal of Data Science Advances*.

[10] Zhang, Y., et al. (2022). “ROC-AUC analysis in insurance claim predictions.” *Machine Learning Quarterly*.