"Comparison Of Supervised Learning Techniques For Predicting Auto Insurance Claims"

Shreya Govil[1], Naman Gupta[2], Saachi Kaup[3]

[1]NMIMS University Computer Science Department Mumbai, India, Email: shreya.govil326@nmims.edu.in

[2]NMIMS University Computer Science Department Mumbai, India, Email: naman.gupta382@nmims.edu.in

[3]NMIMS University Computer Science Department Mumbai, India, Email: saachi.kaup146@nmims.edu.in

Abstract

The escalating prevalence and intensity of auto insurance claims have underscored the need for effective solutions to manage these claims efficiently. Machine learning has emerged as a promising technique to address this challenge, with the potential to enhance the prediction of claim frequency and enable insurance providers to more accurately assess the likelihood of claim submission for a specific policy. To improve the accuracy of our predictions, we employ an extensive dataset for model training, which mitigates the risk of biased analysis. Our approach not only streamlines the claims management process but also fosters more informed decision-making among insurance providers, ultimately enhancing the overall efficiency and effectiveness of the auto insurance industry. Compared to other models we have implemented; the support vector machine (SVM) has the highest degree of accuracy of 87.67%.

Keywords: Auto insurance, machine learning, SVM.

Introduction

Predicting auto insurance claims is vital for insurance companies to manage risks and make informed decisions. The growing volume of data generated by various sources has made this task increasingly challenging. Due to the complexity and volatility of the insurance industry, predicting auto insurance claims is a difficult endeavor. Insurance companies must analyze vast quantities of data to identify potential risks and accurately predict the likelihood of a claim.

In recent years, Machine learning algorithms, which can analyze large datasets and extract valuable insights, have emerged as a potential solution. This study explores the application of multiple machine learning algorithms, including supervised, unsupervised, and reinforcement learning, to predict auto insurance claims. We will assess their accuracy, strengths, and weaknesses, and provide recommendations for insurance companies to improve their predictive models.

Our research will utilize a large dataset of motor insurance claims, comparing the algorithms' accuracy, precision, and recall. Additionally, we will examine the impact of feature engineering, data pre-processing, and hyperparameter tuning on algorithm performance. The goal is to identify the most effective machine learning algorithms and techniques for predicting motor insurance claims, contributing to the development of more accurate and efficient predictive models in the insurance industry.

Literature Review

There are several studies on the use of supervised, unsupervised, and reinforcement learning algorithms in this field.

Supervised learning is a popular approach for fraud detection in insurance claim prediction. Many studies have found it effective for predicting future claim frequencies of policies [10]. In this process, fraudulent claims are identified while reports for genuine claims are generated and subjected to payment clearing [3]. Studies have used supervised learning for various insurance-related applications, such as predicting health insurance costs [6] and property and casualty insurance claims [4]. Predictive analytics has been used in life insurance for over 20 years, primarily for scoring disability claims on the probability of recovery and modeling mortality rates of applicants to improve underwriting decisions and profitability [4]. It has also been used in processing and extracting essential information from copious amounts of motor insurance claim data [9]. Ensemble classifiers, a type of supervised learning, have been used for insurance claim prediction [7].

Unsupervised learning has been used to gain insights into the insured's behavior and to detect fraudulent behavior [11]. Unsupervised Defect Prediction (UnSDP) models have also been utilized to predict software defects and provide software practitioners and researchers with guidance on the viability of using unsupervised prediction models [15]. Machine learning ensemble classifiers, such as XGBoost and TabNet [13, 16], have been used to predict insurance claims.

Reinforcement Learning (RL) is utilized in numerous fields, including the prediction of auto insurance claims and the prediction of insurance claims using a machine learning ensemble classifier. Several studies have examined the application of RL to insurance claim prediction. [20,19]. There have been numerous studies [21, 22] on RL and its use in insurance claim prediction as well as a variety of other applications.

Classifying unbalanced data is still full of challenges. The imbalanced data distribution led to the results of classification emphasis on the majority class. In another word, the accuracy for the minority class is often quite low. Traditional classification methods such as ANN, KNN, cannot solve the problem effectively. The experimental results showed that the classification accuracy of the minority class had a great improvement by using SVM-RBF Mode. [23] The RBF kernel in SVM can be useful for biased data by capturing complex non-linear relationships and allowing control over the smoothness of the decision boundary. This can prevent overfitting to the majority class and create a decision boundary that generalizes well to both majority and minority classes. Techniques like oversampling or under sampling can also be used in combination with SVMs to mitigate the effects of class imbalance.

In conclusion, this study provides insurance companies with valuable insights on the most effective machine learning algorithms and techniques for predicting motor insurance claims, contributing to the development of more accurate and efficient predictive models in the insurance industry.

Methodology

The aim is to investigate the efficacy of multiple classification and neural network approaches for predicting auto insurance claims. We will follow a step-by-step methodology comprising data preprocessing, cleaning, splitting, model selection, selecting weights, and performance evaluation to accomplish this.

The first step in the methodology is the preparation of data for analysis. We have compiled a large set of automobile insurance claims from various sources, including insurance companies and public datasets. The dataset includes various characteristics, such as the driver's age, gender, vehicle make and model, driving record, and other pertinent data.

We have performed data cleansing, which entails removing missing values, addressing outliers, and addressing inconsistencies. In addition, we have conducted data exploration to gain insight into the data distribution, correlation between features and to identify any data quality issues.

The next step is Feature Engineering, in which relevant features are extracted from the data and transformed into a suitable format for machine learning algorithms. We will employ domain expertise and statistical analysis to determine the most influential factors influencing the likelihood of an insurance claim.

Scaling, normalization, and encoding has also been performed to prepare the data for machine learning algorithms. To convert categorical variables to numeric values, we have used a variety of techniques, including one-hot encoding, label encoding, and binary encoding.

Selecting Appropriate Machine Learning Models for Predicting Auto Insurance Claims is the next step.

Models that were used are:

- Support Vector Machines (SVM): SVM is a potent classification algorithm that separates data into distinct classes by locating the optimal hyperplane. SVMs can handle both linear and nonlinear data and effectively manage high-dimensional data. In our research paper, we will use support vector machines (SVMs) to predict motor insurance claims based on driver age, gender, vehicle make and model, driving history, and other relevant data. To train the SVM models, we will use various kernels, including linear, polynomial, and radial basis function (RBF).
- Random Forest is an ensemble classifier that combines multiple decision trees to enhance the model's predictive performance. Random Forest is effective with high-dimensional data and can handle categorical and continuous features. In our research paper, we will use Random Forest to predict auto insurance claims based on a variety of characteristics. Multiple decision trees will be trained using bootstrapping and feature bagging, and predictions will be made using majority voting based on the ensemble of trained decision trees.
- Neural Networks: Neural Networks are a type of machine learning model whose structure and function are inspired by the human brain. Neural Networks are effective at handling complex and non-linear relationships between features and can handle both categorical and continuous features. In our research paper, we will use Neural Networks to predict automobile insurance claims based on a variety of characteristics. We will train the models using various architectures, including feedforward, convolutional, and recurrent Neural Networks. To prevent overfitting, we will also employ regularization techniques such as dropout and L2 regularization.
- Logistic regression: Logistic regression is a type of statistical model used for binary classification problems. It models the probability of a binary outcome (e.g., 0 or 1) using a logistic function. The model estimates the coefficients of the input features and uses them to predict the probability of the target class. It is widely used in various fields, such as finance, medicine, and social sciences.
- Naive Bayes: Naive Bayes is a probabilistic algorithm used for classification tasks. It is based on Bayes' theorem and assumes that the features are conditionally independent given the class label. The algorithm estimates the probability of each class given the input features and selects the class

with the highest probability as the prediction. Naive Bayes is widely used in text classification tasks such as spam filtering and sentiment analysis.

- SGD Classifier is an estimator that implements regularized linear models with stochastic gradient descent (SGD) learning. It is a popular algorithm for large-scale machine learning tasks as it updates the model's parameters incrementally and efficiently. The gradient of the loss is estimated for each sample at a time and the model is updated along the way with a decreasing strength schedule (learning rate). SGD Classifier can be used for binary and multiclass classification problems. It allows users to specify various hyperparameters such as the learning rate, penalty, and maximum number of iterations.

In selecting models, we considered multiple factors such as efficacy, interpretability, and computational complexity. Additionally, we ensured a good balance between the trade-off of bias and variance while choosing the models. To optimize the models' performance, we employed various techniques such as grid search, random search, and Bayesian optimization to adjust their hyperparameters for optimal results.

However, all models showed a significant bias towards the majority class. To address this issue, we implemented class weights to balance the class distribution in the data, ensuring the models no longer gave biased outputs. The weights were calculated based on the inverse frequency of the classes in the dataset.

After implementing class weights, we observed that only the SVM model's performance was significantly improved. This gave us a more usable model to work with, providing accurate predictions while balancing the class distribution. To evaluate the models' performance, we used the accuracy score as the primary metric. By doing so, we determined the effectiveness of each model in predicting the correct class labels.

The weights were calculated as follows:

No. of 1s in is_claim [x] = 3748

No. of 0s in is_claim [y] = 54844

Therefore, the ratio of x:y = 0.5:7.8, which is the weights we used for building the models.

We will evaluate the performance of the models using accuracy score.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Lastly, we will evaluate the performance of the models and compare their ability to predict auto insurance claims.

This methodology allows us to determine the optimal classification and neural network approaches for predicting auto insurance claims. We will provide insurance companies with recommendations to improve their predictive models and risk management strategies. We will also contribute to the development of more precise and effective insurance industry predictive models. This led to the development of multiple inferences on the topic that can be discussed later.

Here is a general flowchart of what was done:

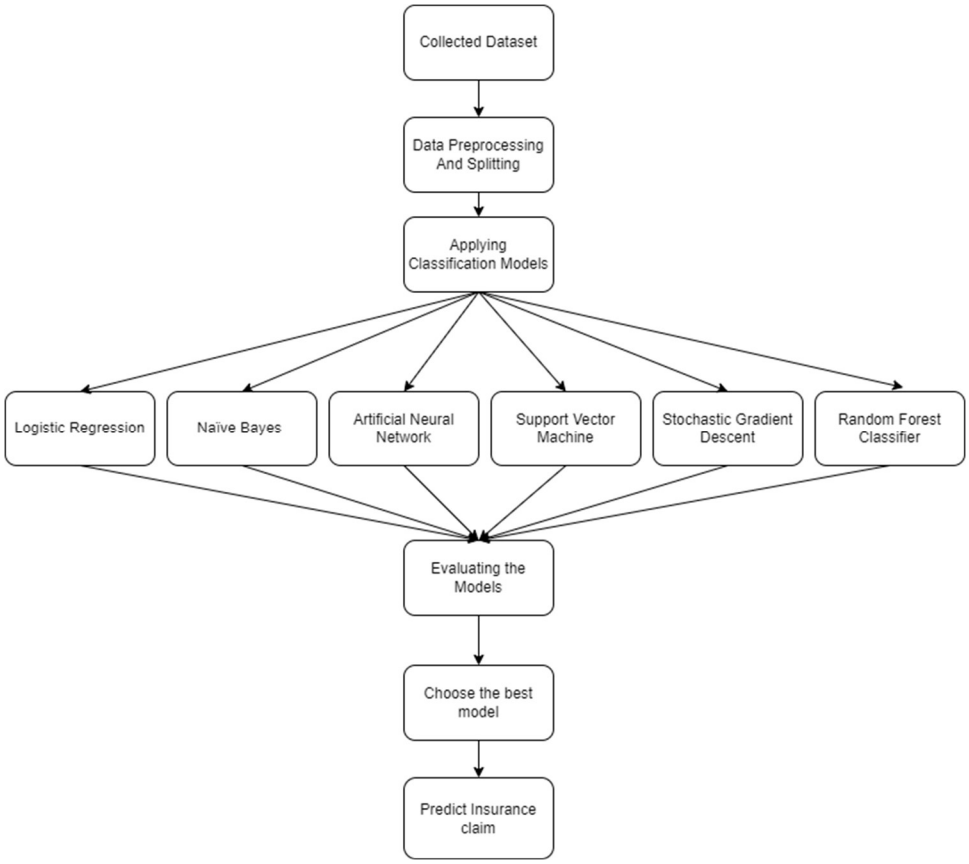

Fig-1

Observations and Analysis

Classification Model	Accuracy Score
Logistic Regression	93.67
Naïve Bayse (Gaussian)	93.67
KNN (Minkowski Distance n=5)	93.43
Random Forest (Gini) and Weighted	93.617
SVM (Without Weights)	93.84
SVM (With Weights), kernel = RBF	**87.67**
Balanced Random Forest Classifier	93.43
SGD Classifier	93.35
Artificial Neural Networks	93.548

Table-1

According to the results shown in the table, all the models, except for SVM with RBF kernel, displayed high bias. Since most of the data (93%) is heavily biased towards the "not claimed" class, these results imply that most of the data is falsely classified as negative. It is important to note that traditional

classification methods like ANN, KNN, were ineffective in solving this issue, as evident from the results.[23]

In contrast, only the RBF kernel SVM model provided accurate readings, with an impressive accuracy score of 87.67%. This finding highlights the importance of selecting appropriate models and techniques to deal with imbalanced datasets. By choosing the right model and using methods such as class weighting, we were able to overcome the issue of bias and obtain reliable results. It also falls in line with the expectations of the claims made in [23].

Conclusion

Our findings show that these techniques can produce highly accurate predictive models that can help insurance companies evaluate risks and set premiums. The SVM (Support Vector Machine) algorithm outperformed all of the models we tested. This is because the RBF (Radial Basis Function) kernel used by the SVM identifies non-linear relationships between financial variables, ultimately improving the accuracy of financial analysis models. Furthermore, these models can efficiently manage high-dimensional and complex data with both categorical and continuous features.

In practice, developing more accurate claim predictions allows insurers to reduce costs while increasing customer satisfaction. However, these models rely heavily on high-quality and relevant data, and they must be updated on a regular basis to ensure their continued effectiveness. Overall, this study highlights the potential of machine learning to provide valuable insights to the insurance industry, improving its ability to mitigate risks and better serve its customers.

Future work and Recommendations

Reinforcement learning, specifically Q-Learning, can be leveraged to optimize the claims process in insurance ether a claim should be approved or denied, or the optimal method for resolving a claim. This can help reduce costs for both the insurance company and policyholders.

To further improve predictive models, hybrid approaches can be employed that combine unsupervised, supervised, and reinforcement learning techniques. For example, clustering can group customers based on their likelihood of filing a claim, and then reinforcement learning can optimize the claims process for each group. This more targeted and individualized approach can provide greater accuracy and efficiency in predicting insurance claims.

Overall, the use of unsupervised and reinforcement learning has the potential to revolutionize the insurance industry by providing more accurate and efficient predictive models, leading to improved decision-making, cost savings, and a better customer experience.

References

[1] Morid MA, Kawamoto K, Ault T, Dorius J, Abdelrahman S. Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. AMIA Annu Symp Proc. 2018 Apr 16; 2017:1312-1321. PMID: 29854200; PMCID: PMC5977561. URL: https://pubmed.ncbi.nlm.nih.gov/29854200/

[2] Suhaimi, Nur Amalina Diyana & Suhaimi, Diyana & Abas, Hafiza. (2020). A SYSTEMATIC LITERATURE REVIEW ON SUPERVISED MACHINE LEARNING ALGORITHMS. 10. 1-24. URL: https://www.researchgate.net/publication/344869267_A_SYSTEMATIC_LITERATURE_REVIEW_ON_SUPERVISED_MACHINE_LEARNING_ALGORITHMS

[3] Boodhun, N., Jayabalan, M. Risk prediction in life insurance industry using supervised learning algorithms. *Complex Intell. Syst.* **4**, 145–154 (2018). URL: https://link.springer.com/chapter/10.1007/978-981-10-7512-4_98

[4] Boodhun, N., Jayabalan, M. Risk prediction in life insurance industry using supervised learning algorithms. Complex Intell. Syst. 4, 145–154 (2018). URL: https://link.springer.com/article/10.1007/s40747-018-0072-1

[5] Hanafy, Mohamed. (2021). Predict Health Insurance Cost by using Machine Learning and DNN Regression Models. International Journal of Innovative Technology and Exploring Engineering. Volume-10. 137. 10.35940/ijitee.C8364.0110321. URL: https://towardsdatascience.com/life-insurance-risk-prediction-using-machine-learning-algorithms-part-i-data-pre-processing-and-6ca17509c1ef

[6] Hanafy, Mohamed. (2021). Predict Health Insurance Cost by using Machine Learning and DNN Regression Models. International Journal of Innovative Technology and Exploring Engineering. Volume-10. 137. 10.35940/ijitee.C8364.0110321. URL: https://www.researchgate.net/publication/348559741_Predict_Health_Insurance_Cost_by_using_Machine_Learning_and_DNN_Regression_Models

[7] "Insurance Claim Prediction Using Machine Learning Ensemble Classifier | by Paul Wanyanga | Analytics Vidhya URL: https://medium.com/analytics-vidhya/insurance-claim-prediction-using-machine-learning-ensemble-classifier-14652907a65e

[8] "Focusing on predictive analysis, exploration and big data Follow More from Medium Zach Quinn in Pipeline: A Data Engineering Resource 3 Data Science Projects That Got Me 12 Interviews. And 1 That..." URL: https://medium.com/geekculture/insurance-claims-fraud-detection-using-machine-learning-78f04913097

[9] Endalew Alamir, Teklu Urgessa, Ashebir Hunegnaw and Tiruveedula Gopikrishna, “Motor Insurance Claim Status Prediction using Machine Learning Techniques” International Journal of Advanced Computer Science and Applications(IJACSA), 12(3), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120354 URL: https://thesai.org/Downloads/Volume12No3/Paper_54-Motor_Insurance_Claim_Status_Prediction.pdf

[10] " Supervised learning techniques in claims frequency modelling " URL: https://ifoadatascienceresearch.github.io/blog/supervised-learning-techniques-in-claims-frequency/

[11] Xiao Lin, Mark J. Browne, Annette Hofmann, Race discrimination in the adjudication of claims: Evidence from earthquake insurance, Journal of Risk and Insurance, 10.1111/jori.12386, 89, 3, (553-580), (2022). URL: https://onlinelibrary.wiley.com/doi/full/10.1111/jori.12359

[12] Dridi, Salim. (2021). Unsupervised Learning - A Systematic Literature Review. https://www.researchgate.net/publication/357380639_Unsupervised_Learning_-_A_Systematic_Literature_Review

[13] "Insurance Claim Prediction Using Machine Learning Ensemble Classifier | by Paul Wanyanga | Analytics Vidhya | URL: https://medium.com/analytics-vidhya/insurance-claim-prediction-using-machine-learning-ensemble-classifier-14652907a65e

[14] "Insurance Claims Fraud Detection Using Machine Learning" URL: https://medium.com/geekculture/insurance-claims-fraud-detection-using-machine-learning-78f04913097

[15] Ning Li, Martin Shepperd, Yuchen Guo, A systematic review of unsupervised learning techniques for software defect prediction, Information and Software Technology, Volume 122,2020,106287, ISSN 0950-5849, URL: https://www.sciencedirect.com/science/article/pii/S0950584920300379

[16] "Car Insurance Claim Prediction", URL: https://github.com/matheusboaro/car_insurance_claim_prediction.

[17] "Insurance Claim Prediction", URL: https://github.com/prathibha13/Insurance-Claim-Prediction

[18] Salvador, José & Oliveira, João & Breternitz, Mauricio. (2020). REINFORCEMENT LEARNING: A LITERATURE REVIEW (September 2020). 10.13140/RG.2.2.30323.76327. https://www.researchgate.net/publication/344930010_REINFORCEMENT_LEARNING_A_LITERATURE_REVIEW_September_2020

[19] Dridi, Salim. (2021). Reinforcement Learning - A Systematic Literature Review. https://www.researchgate.net/publication/357380640_Reinforcement_Learning_-_A_Systematic_Literature_Review

[20] Thrun, Sebastian & Littman, Michael. (2000). A Review of Reinforcement Learning. AI Magazine. 21. 103- 105. https://www.researchgate.net/publication/220605601_A_Review_of_Reinforcement_Learning

[21] Gallistel, Charles. (2006). Review: reinforcement learning. Journal of Cognitive Neuroscience. 11. 126-134. https://www.researchgate.net/publication/292740571_Review_reinforcement_learning

[22] Sharma, Dr. Rashmi & Prateek, Manish & Sinha, Ashok. (2013). Use of Reinforcement Learning as a Challenge: A Review. International Journal of Computer Applications. 69. 28-34. 10.5120/12105-8332. https://www.researchgate.net/publication/272864640_Use_of_Reinforcement_Learning_as_a_Challenge_A_Review

[23] Ding, Lei & Watada, Junzo & Chew, Lim & Ibrahim, Zuwairie & Jau, Lee & Khalid, Marzuki. (2010). A SVM-RBF method for solving imbalanced data problem. ICIC Express Letters. 4. 2419-2424. https://www.researchgate.net/publication/236678815_A_SVM-RBF_method_for_solving_imbalanced_data_problem/citation/download