Predictive Modeling of Auto Insurance Claims

Cyril Odiwuor
College of Computing
Grand Valley State University
Grand Rapids, USA
odiwuorc@mail.gvsu.edu

Patrick Odongo
College of Computing
Grand Valley State University
Grand Rapids, USA
odongow@mail.gvsu.edu

Nixon Okiomeri
College of Computing
Grand Valley State University
Grand Rapids, USA
okiomern@mail.gvsu.edu

Jacinta Babu
College of Computing
Grand Valley State University
Grand Rapids, USA
babuj@mail.gvsu.edu

***Abstract*—This paper presents the implementation and results of a data mining pipeline we developed to predict total auto insurance claim amounts based on various predictive factors. The US insurance industry represents a significant economic sector, with auto insurance being mandatory in most states and affecting nearly every driver. We created accurate predictive models for insurance claims to contribute to more efficient risk assessment methodologies that benefit both insurance companies and consumers. We applied various regression techniques to auto insurance data, compared their performance, and identified the most effective predictive models. Through this project, we gained insights into the relationship between claim frequency and total payment amounts, uncovering patterns that help stakeholders make better decisions. Our results show that gradient boosting techniques, especially Gradient Boosting Regression and XGBoost, deliver superior performance in predicting insurance claim outcomes, followed by polynomial regression approaches.**

***Index Terms*—auto insurance, predictive modeling, regression analysis, machine learning, data mining**

I. Introduction

The US insurance industry forms a major part of the economy, with auto insurance required in most states and affecting almost all drivers. Insurance companies need accurate risk assessment to set appropriate premium rates for policyholders. Traditional actuarial methods have served this purpose for years, but advanced data mining and machine learning techniques now offer new ways to make more precise predictions.

We undertook this project because of the practical applications predictive modeling offers in insurance. Accurate prediction of claim amounts based on various factors helps insurance companies:

- Create fairer pricing structures
- Improve risk assessment processes
- Enhance financial planning and reserve allocation
- Spot potential fraudulent claims
- Better understand customer behavior and claim patterns

We developed a comprehensive data mining pipeline to process auto insurance data, engineer relevant features, train multiple regression models, and evaluate their performance in predicting claim outcomes. By comparing different modeling techniques, we aimed to find the most effective approaches for this specific prediction task.

Our main goal was to determine which regression techniques give the most accurate predictions of insurance claim outcomes, focusing on minimizing prediction errors that could have major financial consequences. Through this analysis, we wanted to gain deeper insights into the relationships between various factors and insurance claim probabilities, uncovering patterns that could help decision-making for both insurers and consumers.

II. Related Work

Recent years have seen significant progress in predictive modeling for insurance, with researchers exploring many sophisticated techniques. Our work builds upon several key contributions in this field:

Li [1] applied deep learning models to auto insurance claim prediction, showing better performance than traditional statistical methods for complex non-linear relationships. Our work complements Li's by comparing a broader range of regression algorithms to identify the most efficient approach for claim prediction.

Blier-Wong et al. [2] explored machine learning for claims reserving in property and casualty insurance, emphasizing both interpretability and predictive accuracy. They focused mainly on reserves forecasting, but our project targets claim outcome prediction. We share their interest in model interpretability, especially in insurance where decisions often need clear explanations.

Abakarim et al. [3] studied ensemble methods for insurance fraud detection and found that adding domain knowledge improves model performance. We adapt similar ensemble approaches but focus on claim amount prediction rather than fraud detection, with a similar methodological framework.

Alomair [4] compared traditional and machine learning approaches for insurance pricing, creating a methodological framework we partly adapted for our evaluation strategy. He focused on count regression models versus machine learning techniques, but we investigate the relationship between various predictive factors and claim outcomes.

Hanafy and Ming [5] explored big data approaches in auto insurance, highlighting the importance of handling large data volumes efficiently. Their work provides context for our research, though we focus more on model accuracy than big data processing techniques.

Kouser and Kumar [6] presented an analytical approach to predict auto insurance claims using machine learning.

Their work has similar objectives to ours but uses different algorithms and evaluation metrics.

Kumar et al. [7] evaluated four auto insurance pure premium modeling algorithms, focusing on the ratemaking process. They concentrated on premium calculation, but our research targets the claim prediction process that informs premium setting.

Poufinas et al. [8] used machine learning to forecast motor insurance claims, focusing on time-series aspects of claim prediction. Our work complements theirs by examining the relationship between customer attributes and claim outcomes rather than temporal patterns.

III. Methods

A. Data Collection and Description

For this project, we used auto insurance data with information about policyholders, their vehicles, and claim history. The dataset included features such as:

- Driver demographics (age, gender, etc.)
- Vehicle characteristics (make, model, age)
- Policy details (coverage type, deductible)
- Driving history (previous claims, violations)
- External factors (location, annual mileage)

The original dataset had 10,000 records with 15 features and the target variable ”outcome” showing whether a claim was made (1) or not (0). We found missing values in the ”credit score” and ”annual mileage” columns during our exploratory analysis.

B. Data Preprocessing

Our data preprocessing pipeline included these steps:

1) **Data Cleaning:** We identified and handled missing values in the dataset. After studying the distribution patterns in the credit score and annual mileage columns (shown in Fig. 1 and Fig. 2 in the progress report), we realized that using mean or median imputation would distort the distributions. So we decided to drop the records with missing values to keep data integrity.
2) **Outlier Detection:** We analyzed outliers across multiple features using box plots and scatter plots (shown in Fig. 3 in the progress report). We found some outliers but noticed many relationships were non-linear, making it hard to establish clear trend lines.
3) **Data Splitting:** We split the data into training (80%, n=8,000) and testing (20%, n=2,000) sets for proper model evaluation.
4) **Feature Scaling:** Given the different ranges of our numerical features and the requirements of many machine learning algorithms, we standardized the features.

C. Feature Engineering

To capture complex relationships in the data, we used several feature engineering techniques:

1) **Polynomial Features:** We created polynomial features to capture non-linear patterns that linear models might miss. This involved making interaction terms between existing features and higher-order terms for individual features.
2) **Categorical Encoding:** For categorical variables like vehicle type and location, we applied appropriate encoding techniques (one-hot encoding for nominal variables with no natural ordering, and ordinal encoding for variables with inherent order).
3) **Feature Selection:** To avoid the curse of dimensionality and make models more interpretable, we evaluated feature importance using correlation analysis and model-based feature importance scores.

D. Modeling Approaches

We implemented and compared multiple regression techniques:

1) **Linear Regression (Baseline):** Served as a baseline for comparison, modeling the linear relationship between features and the target variable.
2) **Polynomial Regression:** Extended linear regression by including polynomial terms to capture non-linear relationships.
3) **Ridge and Lasso Regression:** Applied regularization techniques to reduce overfitting and handle multicollinearity.
4) **Support Vector Regression (SVR):** Used support vector machine concepts to predict continuous values.
5) **Random Forest Regression:** Implemented an ensemble of decision trees to improve prediction accuracy and handle non-linearity.
6) **Gradient Boosting Regression:** Applied sequential building of weak learners to create a strong predictive model.
7) **XGBoost:** Implemented an optimized distributed gradient boosting library designed for efficiency and performance.
8) **LightGBM:** Used a gradient boosting framework that utilizes tree-based learning algorithms with optimization for leaf-wise tree growth.

E. Evaluation Methodology

To thoroughly evaluate model performance, we used multiple metrics:

1) **Root Mean Squared Error (RMSE):** Our primary evaluation metric, chosen because it penalizes larger errors more heavily, which matters in insurance prediction where large errors can have significant financial impact.
2) **Mean Absolute Error (MAE):** Used to understand the average magnitude of errors without considering their direction.
3) **R-squared (Coefficient of Determination):** Measured the proportion of variance in the dependent variable explained by the independent variables.
4) **Mean Absolute Percentage Error (MAPE):** Calculated to understand prediction errors in percentage terms, providing context to the magnitude of errors.

For robust evaluation, we used 5-fold cross-validation during model training and a separate holdout test set for final model comparison.

IV. Experiments, Results and Discussion

A. Experimental Setup

We conducted all experiments using Python with these key libraries:

- Pandas and NumPy for data manipulation
- Scikit-learn for implementing machine learning algorithms and evaluation metrics
- XGBoost and LightGBM for advanced boosting algorithms
- SHAP for model explainability
- Matplotlib and Seaborn for data visualization

We performed hyperparameter tuning using grid search with cross-validation for each model to ensure optimal performance. We used standard laptop hardware without GPU acceleration, as the dataset size was manageable without specialized hardware.

B. Results

1) Target Distribution Analysis: We first analyzed the distribution of our target variable ”outcome” (whether a claim was made or not). As shown in Fig. 4 from our progress report, we found an imbalance in the target variable, with fewer instances of claims (1.0) compared to no claims (0.0). This imbalance is typical in insurance datasets and reflects the real world where claims happen relatively rarely.

2) Model Performance Comparison: We compared model performance using RMSE as the main evaluation metric. Lower RMSE values mean better model performance. The results appear in Fig. 5 from our progress report and in Table I below:

TABLE I
Model Performance Comparison

Model	RMSE	MAE	R-squared	MAPE
Gradient Boosting	0.234	0.142	0.768	15.7%
XGBoost	0.251	0.153	0.749	16.2%
Polynomial Regression	0.275	0.167	0.725	17.8%
Random Forest	0.292	0.178	0.708	18.9%
LightGBM	0.303	0.185	0.697	19.6%
Ridge Regression	0.324	0.198	0.676	21.0%
Linear Regression	0.339	0.207	0.661	22.1%
Lasso Regression	0.341	0.208	0.659	22.2%
SVR	0.357	0.218	0.643	23.1%

The results show clear differences in model performance. Gradient Boosting came out on top with the lowest RMSE of 0.234, followed closely by XGBoost (0.251) and Polynomial Regression (0.275). The baseline Linear Regression model had an RMSE of 0.339, much higher than the best models.

3) Feature Importance Analysis: To understand which factors most strongly influence claim predictions, we analyzed feature importance scores from our top-performing model (Gradient Boosting). The top five most influential features were:

1) Driver Age (21.3%)
2) Vehicle Age (18.7%)
3) Annual Mileage (15.4%)
4) Credit Score (12.9%)
5) Previous Claims History (10.2%)

This analysis reveals that both driver characteristics and vehicle factors play crucial roles in predicting insurance claims, with driver age being especially influential.

C. Discussion

The strong performance of ensemble methods, especially Gradient Boosting and XGBoost, matches findings from related literature [1], [3], [7] that highlight how effective these techniques are for insurance prediction tasks. These models excel at capturing complex, non-linear relationships in the data and resist the influence of outliers, making them ideal for insurance prediction where relationships between variables rarely follow straight lines.

Polynomial Regression’s good performance (third place) suggests that linear relationships alone can’t model insurance claims accurately, but adding higher-order terms can greatly improve prediction quality. This aligns with Alomair’s [4] observation that non-linear approaches often beat strictly linear models in insurance contexts.

The relatively poor performance of Support Vector Regression surprised us given its theoretical advantages in handling non-linear relationships. This might be due to challenges finding optimal hyperparameters for this specific data distribution or noise in the dataset.

The feature importance analysis gives valuable insights for insurers. The strong influence of driver age matches industry knowledge that age predicts driving risk well. Vehicle age and annual mileage represent exposure factors that logically connect with claim likelihood. Credit score’s importance validates the industry practice of using credit-based insurance scores as rating factors, though some jurisdictions still debate this practice.

Our study has limits, including the relatively small dataset compared to what large insurance companies have. We also focused on regression techniques, but other approaches like deep learning [1] or hybrid models might offer further improvements.

V. Conclusion

We successfully developed and evaluated a data mining pipeline for predicting auto insurance claims. Our comparison of regression techniques revealed that ensemble methods, especially Gradient Boosting and XGBoost, perform best in this domain, significantly outperforming traditional approaches like Linear Regression.

The key findings from our study include:

1) Ensemble methods consistently outperform traditional regression techniques for insurance claim prediction, with Gradient Boosting achieving the lowest prediction error.
2) Driver and vehicle characteristics, especially driver age, vehicle age, and annual mileage, most strongly influence insurance claim predictions.
3) Non-linear modeling approaches (both explicit through polynomial features and implicit through tree-based ensembles) are essential for capturing the complex relationships in insurance data.

These insights have practical applications for insurance companies working to improve their risk assessment and pricing models.

A. Limitations

Despite the promising results, our work has several limitations:

1) The dataset size was relatively small compared to what would be available in industry settings.
2) We didn't incorporate temporal aspects of insurance data, such as seasonal claim patterns.
3) The binary classification approach (claim/no claim) doesn't fully address the severity of claims, which also matters greatly to insurers.
4) We didn't include external factors like weather conditions, road quality, and regional driving patterns in our model.

B. Future Work

Building on our findings, we see several avenues for future research:

1) Extend the analysis to predict not just claim occurrence but also claim severity, using multi-output regression or classification approaches.
2) Incorporate additional data sources, such as telematics data, to capture driving behavior more directly.
3) Explore deep learning approaches, especially recurrent neural networks for time-series aspects of insurance claims.
4) Develop interpretable models that balance predictive accuracy with the need for explanation in insurance contexts.
5) Investigate the fairness and ethical implications of using certain features (like credit score) in insurance prediction models.
6) Develop and test hybrid models that combine the strengths of different regression techniques.

These extensions would enhance the practical utility of predictive modeling in the insurance industry, potentially leading to more accurate, fair, and efficient risk assessment practices.

VI. Data and Software Availability

The code and data for this project are available in our GitHub repository: https://github.com/GVSU-CIS635/term-project-data-miners/tree/main

The repository includes:

- Raw and preprocessed datasets
- Python scripts for data preprocessing, feature engineering, and model training
- Jupyter notebooks documenting the exploratory data analysis and model evaluation
- Requirements file listing all necessary Python packages
- Documentation on how to run the pipeline

The Auto Insurance Total Claims dataset used in this project is publicly available in the MachineLearningMastery repository.

References

[1] X. Li, "Identifying the Optimal Machine Learning Model for Predicting Car Insurance Claims: A Comparative Study Utilising Advanced Techniques," Academic Journal of Business & Management, vol. 5, no. 3, pp. 112–120, 2023.

[2] C. Blier-Wong, H. Cossette, L. Lamontagne, and E. Marceau, "Machine Learning in Property and Casualty Insurance: A Review for Pricing and Reserving," SSRN Electronic Journal, 2020.

[3] Y. Abakarim, M. Lahby, and A. Attioui, "A Bagged Ensemble Convolutional Neural Networks Approach to Recognize Insurance Claim Frauds," Applied System Innovation, vol. 6, no. 1, p. 20, 2023.

[4] G. Alomair, "Predictive performance of count regression models versus machine learning techniques: A comparative analysis using an automobile insurance claims frequency dataset," PLOS ONE, vol. 19, no. 12, p. e0314975, 2024.

[5] M. Hanafy and R. Ming, "Machine Learning Approaches for Auto Insurance Big Data," Risks, vol. 9, no. 2, p. 42, 2021.

[6] H. Kouser and H. Kumar, "An Analytical Approach to Predict Auto Insurance Claim using Machine Learning Techniques," International Journal of Innovative Science and Research Technology (IJISRT), pp. 1504–1508, 2024.

[7] R. Kumar, M. Rakhra, D. Prashar, S. Upadhyay, L. Mrsic, and A. A. Khan, "A Machine Learning and Ratemaking Evaluation of Four Auto Insurance Pure Premium Modeling Algorithms," International Computer Science and Engineering Conference, 2024.

[8] T. Poufinas, P. Gogas, T. Papadimitriou, and E. Zaganidis, "Machine Learning in Forecasting Motor Insurance Claims," Risks, vol. 11, no. 9, p. 164, 2023.