Predictive Modeling of Auto Insurance Claims

1st Cyril Odiwuor
College of Computing
Grand Valley State University
Grand Rapids, USA
odiwuorc@mail.gvsu.edu

2nd Patrick Odongo
College of Computing
Grand Valley State University
Grand Rapids, USA
odongow@mail.gvsu.edu

3rd Nixon Okiomeri
College of Computing
Grand Valley State University
Grand Rapids, USA
okiomern@mail.gvsu.edu

4th Jacinta Babu
College of Computing
Grand Valley State University
Grand Rapids, USA
babuj@mail.gvsu.edu

***Abstract*—This paper presents a project proposal for developing a data mining pipeline to predict total auto insurance claim amounts based on the number of claims made. The insurance industry in the United States represents a significant sector of the economy, with auto insurance being mandatory in most states and affecting nearly every driver. By creating an accurate predictive model for insurance claims, this project aims to contribute to more efficient risk assessment methodologies that could potentially benefit both insurance companies and consumers. The approach involves applying various regression techniques to the Auto Insurance Total Claims dataset, comparing their performance, and identifying the most effective predictive model. Through this project, we aim to gain insights into the relationship between claim frequency and total payment amounts, potentially uncovering patterns that could help stakeholders make more informed decisions.**

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

I. Introduction

The insurance industry in the United States represents a significant sector of the economy, with auto insurance being mandatory in most states and affecting nearly every driver. By creating an accurate predictive model for insurance claims, this project aims to contribute to more efficient risk assessment methodologies that could potentially benefit both insurance companies and consumers.

The approach involves applying various regression techniques to the Auto Insurance Total Claims dataset, comparing their performance, and identifying the most effective predictive model. The project is motivated by the real-world applications of such predictive models in premium pricing, risk assessment, and financial planning within the insurance sector. Through this project, the team aims to gain insights into the relationship between claim frequency and total payment amounts, potentially uncovering patterns that could help stakeholders make more informed decisions.

II. Related Work

Predictive modeling in insurance has seen significant advancements in recent years, with researchers exploring increasingly sophisticated techniques:

Li [1] applies deep learning models to auto insurance claim prediction, demonstrating superior performance over traditional statistical methods when dealing with complex nonlinear relationships. The current project differs by focusing on comparing a broader range of regression algorithms on a simpler dataset to identify the most efficient approach.

Blier-Wong et al. [2] explore the use of machine learning for claims reserving in property and casualty insurance, emphasizing interpretability alongside predictive accuracy. While their focus is on reserves, this project targets claim amount prediction, though it shares their concern for model interpretability.

Abakarim et al. [3] investigate the use of ensemble methods specifically for insurance fraud detection, finding that incorporating domain knowledge significantly improves model performance. The current project will adapt similar ensemble approaches but focus on claim amount prediction rather than fraud detection.

Alomair [4] conducts a comparative study of traditional and machine learning approaches for insurance pricing, providing a methodological framework that will be partially adapted for the evaluation strategy in this project. This project specifically focuses on the relationship between claim frequency and total payment amounts.

III. Data Plan

This project will use the Auto Insurance Total Claims dataset (auto-insurance.csv) available in the MachineLearningMastery repository. This dataset contains 63 observations with two primary variables: the number of claims made and the corresponding total payment for those claims. The data was originally collected by the Federal Insurance Administration for risk assessment purposes.

This dataset was selected because it provides a clear regression problem with direct relevance to the insurance industry in America. Its simplicity makes it ideal for comparing different regression techniques without the complexity of extensive feature engineering, while still offering meaningful insights into insurance risk assessment.

For preprocessing, the following steps are anticipated:

1) Checking for and handling any missing values, though the dataset is reported to be complete
2) Performing exploratory data analysis to understand the distribution of claims and payments
3) Checking for outliers that might significantly impact model performance
4) Splitting the data into training and testing sets (using an 80/20 split)

5) Normalizing or standardizing the features if required by specific algorithms
6) Generating additional polynomial features to explore non-linear relationships

IV. Implementation Plan

The data mining pipeline will consist of the following key components:

A. Data Acquisition and Preprocessing

- Loading the Auto Insurance Total Claims dataset
- Performing exploratory data analysis to understand data distributions
- Applying necessary preprocessing steps identified in the data plan
- Splitting the data into training and testing sets

B. Feature Engineering

- Generating polynomial features to capture potential non-linear relationships
- Creating interaction terms if beneficial
- Implementing feature scaling as needed for various algorithms

C. Model Development and Training

- Implementing multiple regression techniques:
  - Linear Regression (baseline)
  - Polynomial Regression
  - Ridge and Lasso Regression
  - Support Vector Regression
  - Random Forest Regression
  - Gradient Boosting Regression
  - XGBoost
  - LightGBM
- Training each model on the training dataset
- Tuning hyperparameters using cross-validation

D. Model Evaluation and Comparison

- Applying each model to the test dataset
- Calculating and comparing performance metrics
- Visualizing prediction results
- Analyzing model residuals
- Assessing model explainability using SHAP values

The implementation will primarily use Python with the following libraries:

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

V. Evaluation Plan

The models will be evaluated using several regression performance metrics:

1) Mean Absolute Error (MAE)
2) Root Mean Squared Error (RMSE)
3) R-squared (coefficient of determination)
4) Mean Absolute Percentage Error (MAPE)

The primary success metric will be RMSE, as it penalizes larger errors more heavily, which is important in the insurance context where large prediction errors could have significant financial implications.

K-fold cross-validation (with k=5) will be used to ensure robust evaluation and to avoid overfitting. The models will be compared against each other, with linear regression serving as the baseline. A holdout validation set (20% of the data) that will not be used during model development will provide a final, unbiased evaluation of the best-performing model.

To further assess model robustness, sensitivity analysis will be conducted by introducing artificial noise to the test data and observing how model performance degrades. Additionally, model explainability will be evaluated using SHAP (SHapley Additive exPlanations) values to understand which features contribute most to the predictions.

VI. Plan for Group Collaboration

The team will meet twice weekly: once via Zoom on Mondays (7-8 PM) for planning and progress updates, and once in-person on Thursdays (4-6 PM) at the library for collaborative work sessions. Between meetings, the team will coordinate asynchronously through a dedicated WhatsApp Group for sharing updates, asking questions, and posting resources.

For code collaboration, GitHub will be used with the following workflow:

1) Main branch for stable, reviewed code
2) Feature branches for individual components
3) Pull requests with code reviews before merging into main

The dataset will be stored in the GitHub repository, and Google Colab will be used for collaborative coding sessions. Each team member will be responsible for implementing and evaluating specific regression algorithms, while collectively contributing to data preprocessing, analysis, and final report writing.

VII. Timeline

Table I outlines the project timeline and key milestones.

VIII. Conclusion

This project proposal outlines a comprehensive approach to developing predictive models for auto insurance claims. By applying various regression techniques to a focused dataset, we aim to identify the most effective methods for predicting total claim amounts based on claim frequency. The insights gained from this project could potentially contribute to more efficient risk assessment methodologies in the insurance industry, benefiting both providers and consumers. Through careful

TABLE I
PROJECT TIMELINE

Week	Goals
Week 9 (Mar 4-10)	• Complete detailed EDA • Finalize data preprocessing steps • Implement baseline linear regression model
Week 10 (Mar 11-17)	• Implement polynomial regression • Implement regularization techniques (Ridge, Lasso) • Begin feature engineering
Week 11 (Mar 18-24)	• Implement SVR and tree-based models • Submit progress report (Mar 24) • Begin hyperparameter tuning
Week 12 (Mar 25-31)	• Implement XGBoost and LightGBM models • Finalize hyperparameter tuning • Begin comprehensive model evaluation
Week 13 (Apr 1-7)	• Complete model evaluation • Implement SHAP for model explainability • Begin drafting final report
Week 14 (Apr 8-14)	• Draft complete final report • Peer review of report • Finalize all code and documentation
Week 15 (Apr 15-18)	• Final polishing of report • Submit final report (Apr 18)

evaluation and comparison of different models, we hope to advance understanding of the relationships between claim patterns and financial outcomes in auto insurance.

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.