Regression with an Insurance Dataset

A Kaggle Playground Prediction Competition

Assignment 2

Dev D. Rabadia (NF1005560),

Denisse C. Cortes (NF1007936),

Rosario D. Torres (NF1001385),

Miko L. Tan (NF1008647)

Master of Data Analytics, University of Niagara Falls Canada

DAMO-510-4: Winter 2025 Predictive Analytics

Professor Ali El-Sharif

March 16, 2025

Statement of Purpose

The insurance sector is a vital part of the global economy, providing financial protection against unforeseen risks. It operates by pooling resources from policyholders to compensate those who experience covered losses. According to Statista Research Department (2025), the global insurance market was worth approximately $8 trillion in 2024 and is expected to increase substantially in the coming years.

Insurance companies assess risks, set premium rates, and manage claims to ensure both sustainability and profitability. With advancements in data science, insurers are increasingly leveraging predictive analytics, machine learning, and AI to improve risk assessment and premium calculation, detect fraudulent claims by identifying unusual patterns, enhance customer segmentation to offer personalized policies, and optimize claims processing through automation.

The purpose of this project is to develop and optimize predictive models for estimating insurance premiums based on customer characteristics and policy details. The dataset used includes variables such as age, gender, annual income, marital status, dependents, education, occupation, health score, location, policy type, previous claims, vehicle age, credit score, insurance duration, customer feedback, smoking status, exercise frequency, and property type. By applying machine learning techniques, the group aims to create a model that closely predicts premium amounts while ensuring generalizability and efficiency through enhanced feature selection and hyperparameter tuning.

This project is relevant to the insurance industry as it enhances the ability to offer fair and competitive pricing based on customer risk profiles. It also helps insurers improve their underwriting process and provide personalized pricing strategies. Throughout the model development process, different combinations of features and algorithms were tested, and multiple optimization techniques were implemented to achieve an optimal solution. The primary goals of this project are:

- To analyze the relationship between customer attributes and insurance premiums
- To develop models to predict insurance premiums
- To evaluate and select the most accurate model
- To provide insights for strategic decision-making for insurers

Scope of the Project

The project focused on several key areas to achieve its objectives. First, the group began with **Exploratory Data Analysis (EDA)** and **Data Preprocessing**, which included cleaning and handling missing data, identifying and treating outliers, and transforming categorical variables into numerical representations, including any necessary transformations that need to be applied to improve model performance. This phase ensured that the dataset was structured appropriately for machine learning applications.

Following this, the group conducted **Statistical Analysis and Tests**, where the relationships between various customer attributes and premiums were re-analyzed, and then the most appropriate features were identified and selected.

The **Model Development & Evaluation** phase explored multiple techniques with a focus on improving predictive accuracy. Hyperparameter tuning techniques were also applied to optimize model performance efficiently. To assess model effectiveness, the group evaluated the results using **Root Mean Squared Logarithmic Error (RMSLE)**, ensuring that the final model balanced accuracy, generalization, and computational efficiency. This model was then tested in the Testing dataset for Kaggle submission. Finally, the group focused on interpretation and strategic insights, identifying the most influential factors affecting premium predictions and providing recommendations for refining insurance pricing strategies.

Background Research and Literature

Predicting health insurance premiums has been a critical task for insurance companies, as it enables accurate risk assessment and appropriate pricing strategies. For the baseline model, the group used Linear Regression, a widely used approach in insurance premium prediction due to its simplicity, interpretability, and ability to provide continuous predictions. A study by Narayana, Yogesh, and Kowshik (2023) employed Multiple Linear Regression to estimate medical insurance premiums, incorporating factors such as age, BMI, and smoking status.

Recent advancements in machine learning have demonstrated that algorithms such as Decision Trees, Random Forest and Gradient Boosting outperform traditional regression techniques in insurance prediction tasks. These methods can capture non-linear relationships and complex feature interactions. Iqbal et al. (2024) explored machine learning methods for predicting health insurance costs, showing their effectiveness in enhancing risk assessment and pricing strategies. The study found that Gradient Boosting achieved an accuracy of 86.86%, outperforming other models (Iqbal et al., 2024).

Compared to traditional statistical methods, machine learning models provide greater flexibility in handling complex, non-linear patterns in data, making them well-suited for insurance pricing models. While Linear Regression remains useful for its interpretability, studies have shown that ensemble-based models, such as Random Forest and Gradient Boosting, consistently achieve lower error rates and improved generalization to new data.

In terms of feature importance, numerous studies have investigated the impact of different features in predicting insurance premiums (Iqbal et al., 2024; Narayana et al., 2023). Identifying these key features is crucial for making accurate predictions. Previous research has often highlighted factors such as age, BMI, smoking status, number of dependents, and geographic location as significant predictors (Iqbal et al., 2024). These studies emphasize the role of behavioral and demographic factors in pricing models, which has shaped modern machine learning approaches to risk assessment. The study from Iqbal et al. (2024) emphasized that smoking status and BMI are among the most influential factors affecting insurance premiums.

Design and Data Collection Methods

Design/Data Source

One of the datasets used for this project is derived from Kaggle’s Insurance Premium Prediction Dataset from Saravanan (2024), which simulates real-world insurance data for regression modeling. This synthetic dataset was created to facilitate data-driven approaches in estimating insurance premiums based on factors such as age, income, health status, and claim history (Saravanan, 2024).

The primary dataset for this project was a Kaggle competition dataset, which was generated using a deep learning model trained on the Insurance Premium Prediction dataset. While the feature distributions are like the original dataset, they are not identical (Reade & Park, 2024). The competition dataset was used as the main dataset for model development and evaluation.

Sampling Methods

The dataset was split into a training set of 1,200,000 records with 21 columns and a test set of 800,000 records with 20 columns, with the division aimed at ensuring robust model evaluation and generalization (Reade & Park, 2024). The **80-20 train-test split** was applied to assess model performance on unseen data, minimizing the risk of overfitting and ensuring generalization to new observations.

Data Model

The data model for this project is a regression-based model designed to predict insurance premium amounts. The model uses a variety of features, including numerical variables (such as age, annual income, and health score) and categorical variables (such as gender, marital status, and policy type) to estimate the target variable: the insurance premium amount.

To enhance predictive performance, multiple regression-based models were explored, including Linear Regression, Decision Trees, Random Forest, and XGBoost. While Linear Regression served as a baseline, ensemble models such as Random Forest and XGBoost were implemented to capture non-linear relationships and improve generalization.

The dataset contained skewed numerical features and missing values, which were handled through appropriate preprocessing techniques, including feature binning for Age, Income, and Insurance Duration to improve model interpretability and robustness. Additionally, outlier detection and transformation methods, such as Box-Cox transformations and capping extreme values, were applied to mitigate the impact of highly skewed features like Previous Claims and Annual Income.

Categorical features were transformed using one-hot encoding and ordinal encoding techniques, ensuring they were optimally structured for the chosen models. These preprocessing steps helped standardize the dataset, reducing noise and improving prediction accuracy.

To further enhance model performance, hyperparameter tuning was conducted using RandomizedSearchCV, which allowed efficient exploration of different parameter combinations, ultimately improving the accuracy and stability of the final model.

Analytical Approach

The analytical approach in this study involved a comprehensive exploratory data analysis (EDA) to understand the dataset's structure and distribution. Each variable was examined through summary statistics, missing value assessments, and distribution analysis to identify skewness and potential outliers.

Categorical features were analyzed for unique values and frequency distributions, while numerical variables were assessed using histograms, box plots, and statistical tests. To detect potential multicollinearity, Variance Inflation Factor (VIF) analysis was conducted, helping identify highly correlated features that could negatively impact model stability.

Correlation analysis was performed to examine relationships between independent variables and the target variable, **Health Insurance Premium Amount**. These insights served as a guide for data preprocessing decisions, ensuring that transformations, imputations, and feature selection strategies were effectively applied.

Additionally, outlier detection was conducted by analyzing feature distributions and identifying extreme values that could influence model training. However, handling these outliers was addressed separately during preprocessing.

This analytical process ensured that **data-driven decisions were made before feature transformation and modelling**, providing a strong foundation for effective preprocessing and model development.

Methodology/Strategies

Table 1: *Methodology/Strategies for Predictive Analytics*

Methodology	Strategy	Brief Description
Data Preprocessing	Handling Missing Values	Imputed missing numerical values using median imputation, while categorical were handled using median/mode imputation or an "Unknown" category.
	Feature Engineering & Transformation	Applied binning for Age, Annual Income, and Insurance Duration to enhance interpretability. Log and Box-Cox transformations were selectively applied to normalize skewed data – Previous Claims and Premium Amount. One-hot encoding and label encoding techniques were used for categorical variables – Gender, Marital Status, Education Level, Occupation, Location, Policy Type, Customer Feedback, Smoking Status, Exercise Frequency, and Property Type. Temporal patterns were also extracted from Policy Start Date.
	Handling Outliers	Applied Winsorization/capping and binning for addressing extreme values. These techniques reduced the influence of anomalies while maintaining valuable data distributions.
Statistical Analysis & Tests	Checking Data Distributions	Examined feature distributions using histograms and Q-Q plots to assess skewness. Applied Box-Cox transformation where necessary to normalize data.
	Feature Selection	Employed Recursive Feature Elimination (RFE), Random Forest Feature Importance to identify the most relevant predictors, reducing model complexity and preventing overfitting.
Model Development & Evaluation	Baseline Models	Established a baseline model with Linear Regression for performance comparison before advancing to complex models.
	Advanced Models	Implemented Decision Tree, Random Forest, and XGBoost to improve predictive accuracy, focusing on capturing non-linear relationships.
	Hyperparameter Tuning	Optimized model performance through RandomizedSearchCV to refine key parameters and prevent overfitting.
	Evaluation Metrics	Assessed model effectiveness using Root Mean Squared Logarithmic Error (RMSLE) as the primary metric, along with R-squared and Root Mean Squared Error (RMSE) for comparative evaluation.

Key Findings and Insights

This section presents key observations derived from data exploration, feature engineering, model evaluation, and their impact on insurance premium prediction. The findings align with each stage of the methodology applied.

Data Preprocessing Findings

- Age, Income, and Insurance Duration were binned into categories to improve interpretability and model stability.
- Label encoding was applied to Education Level and Exercise Frequency to ensure the correct sequence of values was maintained.
- In addition to one-hot encoding, the group mapped categorical values to their numeric equivalents to ensure compatibility with the model.
- The dataset contained skewed distributions, notably in Annual Income and Premium Amount, which could negatively impact model performance. To normalize these features, we applied the Box-Cox transformation. This transformation was preferred over log transformation because it automatically selects the best power transformation (λ) for variance stabilization (Hastie et al., 2021).

  Due to Kaggle's restrictions on using the invboxcox function, we manually applied an estimated λ to revert Premium Amount predictions. While this workaround introduced minor approximation errors, it did not significantly impact accuracy.
- Temporal patterns, such as month and year, were extracted from Policy Start Date.

Statistical Analysis and Tests Findings

- Recursive Feature Elimination (RFE) was used to identify and remove the least important features. This method iteratively eliminates weaker predictors while preserving model performance (Guyon et al., 2002).
- The Customer Feedback feature was removed after correlation analysis indicated a weak relationship with Premium Amount. This decision aligns with feature importance rankings derived from RFE, confirming its limited predictive value.

Model Development & Evaluation Findings

XGBoost outperformed all other models with an RMSLE of 0.4653, demonstrating the best predictive accuracy. Linear Regression performed the worst, with an RMSLE of 0.4725, due to its assumption of linear relationships in the data.

Table 1: *Training vs. Testing Accuracy Comparison (Overfitting Analysis)*

Model	Training RMSLE	Test RMSLE	Overfitting Risk & Observations
Linear Regression	0.4714	0.4725	Minimal overfitting (small RMSLE difference). However, poor performance due to an inability to model non-linear relationships, making it unsuitable for complex data patterns (Hastie et al., 2021).
Decision Tree	0.0057	0.6603	High overfitting risk (large RMSLE gap). Decision Trees tend to overfit by creating overly specific splits, leading to high variance. Pruning or setting depth constraints can mitigate this (Breiman et al., 1984).
Random Forest	0.4646	0.4659	Low overfitting risk (small RMSLE gap). Ensemble averaging reduces variance and improves generalization, though deep trees may still introduce residual overfitting (Breiman, 2001).
XGBoost	0.4641	0.4654	Minimal overfitting (closely matched RMSLE). Gradient boosting corrects errors iteratively and uses regularization (L1/L2) to prevent overfitting, ensuring strong generalization (Chen & Guestrin, 2016).

Table 2: *Model Results*

Model	RMLSE Score	Observations with Justifications

Linear Regression	0.4725	Struggled with non-linear patterns due to its assumption of a linear relationship between features (e.g., annual income, previous claims) and premiums. This limitation, well-documented by Hastie et al. (2021), results in higher residuals, particularly for outliers, reducing predictive accuracy.
Decision Tree	0.4676	Improved accuracy over Linear Regression but showed overfitting. Decision Trees exhibit low bias and high variance due to deep splits that fit training data too closely (Breiman et al., 1984), as evidenced by the larger training-test RMSLE gap.
Random Forest	0.4660	Outperformed Linear Regression with reduced overfitting compared to Decision Tree. However, the model retained some overfitting tendencies due to deep trees, despite ensemble averaging improving generalization (Breiman, 2001).
XGBoost	0.4653	Best-performing model, effectively capturing complex, non-linear relationships. XGBoost's gradient boosting framework minimizes errors iteratively and uses regularization to prevent overfitting (Chen & Guestrin, 2016), excelling in handling feature interactions (e.g., income and claims history).

- In terms of feature importance, Credit Score, Previous Claims, and Annual Income were the strongest predictors of premium amounts. Health Score and Policy Start Month had a moderate influence but were not primary drivers. Gender and Policy Type had minor influence but were retained for interpretability.
- Hyperparameter tuning using RandomizedSearchCV significantly improved model accuracy by optimizing key parameters, particularly in Decision Trees, Random Forest, and XGBoost.
- Feature selection methods such as RFE helped eliminate redundant features, improving the model's generalization ability.
- XGBoost's ability to handle complex relationships helped achieve better generalization, reducing the risk of overfitting seen in simpler models.

Business Impact and Conclusion

By leveraging machine learning, insurance companies can refine pricing strategies based on data-driven risk assessments. Our analysis confirms that Credit Score, Annual Income, and

Previous Claims history are the strongest predictors of premium amounts, indicating that insurers should offer tailored policies based on financial and claims history rather than solely on demographic factors such as gender or marital status.

Additionally, several key features in our model align with real-world insurance practices and significantly influence how premium amounts are determined:

- **Age:** In the insurance industry, age is a critical risk indicator. Younger policyholders often face higher health insurance premiums due to higher accident and health risk probabilities. Similarly, older individuals may have increased health insurance costs due to age-related medical risks.
- **Annual Income:** Higher-income individuals are often considered lower risk and may qualify for premium insurance products.
- **Previous Claims:** A history of frequent claims often leads to increased premiums. Past claims data can help predict future risk levels, implementing risk-based pricing to balance profitability and fairness.
- **Credit Score:** Lower credit scores are statistically linked to higher claims. This practice allows insurers to adjust premiums based on financial responsibility, reflecting real-world underwriting standards.
- **Insurance Policy Duration:** Long-term customers often qualify for discounts or lower premium adjustments, as data shows they are less likely to switch providers and may have more stable risk profiles.

The integration of advanced machine learning models, particularly XGBoost, significantly improved predictive accuracy, allowing for more precise premium estimations. By leveraging these insights, insurers can develop fairer pricing models and enhance their underwriting processes to maximize profitability while maintaining customer satisfaction.

The successful implementation of this predictive model provides significant benefits to the insurance industry, such as:

- **Enhanced Pricing Accuracy:** More precise premium predictions result in fairer pricing and increased customer satisfaction
- **Improved Risk Assessment:** Insurers are better equipped to evaluate the risk profile of customers and adjust premiums accordingly
- **Competitive Advantage:** Companies that leverage data-driven insights can optimize their offerings and gain an edge in the market.

In conclusion, this project has the potential to revolutionize premium calculation by integrating machine learning techniques into the pricing strategy. Moving forward, insurers can explore additional data sources, refine models with real-time risk assessments, and continuously improve their predictive capabilities.

Future research may also include extending the model by integrating real-time risk assessment, fraud detection mechanisms, and predictive modeling for claim probability. Additionally, incorporating external datasets such as medical records or real-time financial indicators could further improve risk profiling and premium pricing accuracy.

References

Iqbal, S. M., Ghatol, S. D., Jadhav, P. V., & Raspalle, N. D. (2024). *Health insurance cost prediction using machine learning*. International Research Journal of Engineering and Technology (IRJET), 11(4), 171–178. https://www.irjet.net/archives/V11/i4/IRJET-V11I4171.pdf.

Narayana, K. L., Yogesh, & Kowshik, P. (2023). *Medical insurance premium prediction using regression models*. International Journal for Research Trends and Innovation, 8(4), 1512–1517. https://ijrti.org/papers/IJRTI2304248.pdf.

Reade, W., & Park, E. (2024, November 30). *Regression with an insurance dataset* [Dataset]. Playground Series - Season 4, Episode 12. Kaggle. https://www.kaggle.com/competitions/playground-series-s4e12/.

Saravanan, G. (2024, August 15). *Insurance premium prediction* [Dataset]. Kaggle. https://www.kaggle.com/datasets/schran/insurance-premium-prediction/.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Chapman & Hall/CRC.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Hastie, T., Tibshirani, R., & Friedman, J. (2021). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed., corrected 12th printing). Springer.

scikit-learn developers. (n.d.). *User guide*. In scikit-learn 1.0.2 documentation. Retrieved from https://scikit-learn.org/stable/user_guide.html.

Statista Research Department. (2025, January 29). *Global insurance industry*. Statista. https://www.statista.com/topics/6529/global-insurance-industry/.

xgboost developers. (n.d.). *XGBoost documentation*. Retrieved from https://xgboost.readthedocs.io/en/latest/index.html.