# Predicting Insurance Policy Costs Using Machine Learning: A Data Science Project Report

Saba Madadi
Data Science Course Project

August 25, 2025

## Abstract

Insurance providers rely heavily on accurate cost estimation to ensure balanced pricing, manage risk, and sustain profitability. This report documents a machine learning approach to predicting insurance policy costs based on heterogeneous customer data. The workflow covers data preprocessing, exploratory data analysis (EDA), feature engineering, and the development of predictive models. A log-transformation of the target variable and the Root Mean Squared Logarithmic Error (RMSLE) are used to manage the skewed cost distribution and provide robust evaluation. The results demonstrate that regularized linear models such as Ridge and Elastic Net regression achieve competitive performance when supported by thoughtful preprocessing and careful encoding of categorical variables. Furthermore, the inclusion of temporal and interaction features significantly improves model stability and accuracy. The study highlights both the challenges and opportunities of applying machine learning in actuarial contexts and suggests future directions involving ensemble and hybrid models.

## 1 Introduction

Insurance companies must design pricing strategies that are equitable for customers and sustainable for the business. Policy cost prediction is central to this task. Traditional actuarial models rely on limited parametric assumptions and small sets of features, which may overlook complex relationships in the data. By contrast, modern machine learning enables the use of rich feature spaces, automated handling of missingness, and robust evaluation techniques. This project explores these opportunities by building predictive models for policy cost using a provided dataset as part of a data science course assignment. The emphasis is not only on achieving predictive accuracy but also on developing a principled, reproducible pipeline that balances complexity and interpretability.

## 2 Dataset and Preprocessing

The dataset comprises numerical, categorical, and temporal attributes related to customers and their policies. The target variable, policy cost, is highly skewed, with most customers holding relatively inexpensive policies and a small number of clients incurring extremely high costs. To address this imbalance, the natural logarithm of cost (after adding one) was used as the modeling target. This transformation stabilized residuals and ensured meaningful use of RMSLE during evaluation.

Preprocessing steps were guided by two goals: ensuring consistent numerical inputs for algorithms and preserving information relevant to prediction. Missing values in numerical fields were imputed with robust central statistics such as medians. For categorical features, rare categories were grouped under an "other" label to avoid instability. Features with ordinal structure, such as financial ratings, were encoded to reflect inherent order, while small nominal features were represented using one-hot encodings. High-cardinality features were treated using out-of-fold mean target encoding, which captured conditional effects while avoiding leakage. Finally, scaling was applied where model assumptions required standardized input ranges.

## 3 Exploratory Data Analysis

Exploratory analysis was essential to understanding both the structure of the data and the challenges of prediction. Figure 1 shows the original policy cost distribution, which is dominated by low-value policies and characterized by a long right tail. This reinforces the need for logarithmic transformation. Figure 2 illustrates the distribution of the log-transformed target, which appears much more symmetric and suitable for modeling.

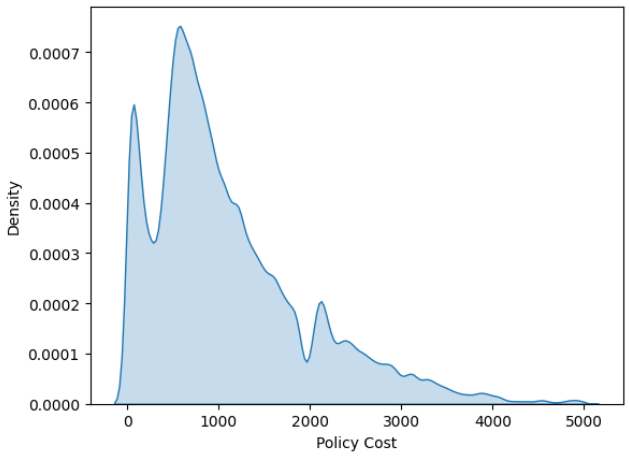


Figure 1: Density distribution of policy cost before transformation, showing strong skew and long right tail.

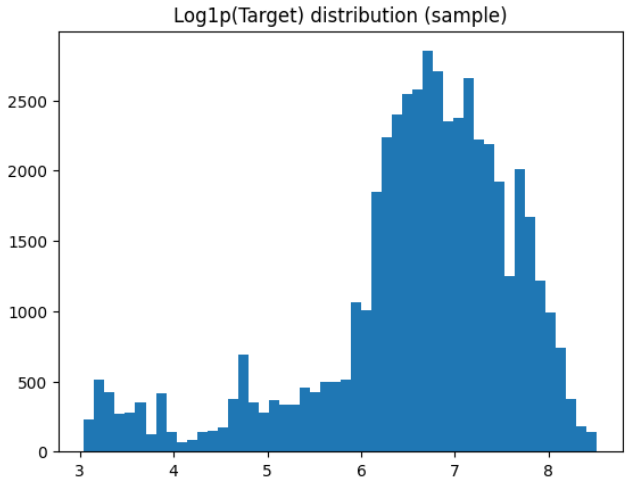


Figure 2: Distribution of log-transformed policy cost (log1p), which is considerably more symmetric.

Correlation analysis (Figure 3) revealed moderate relationships between variables such as yearly earnings and policy cost, while other predictors like prior claims, dependent count, and automobile age displayed weaker but potentially nonlinear effects. This guided subsequent feature engineering decisions.

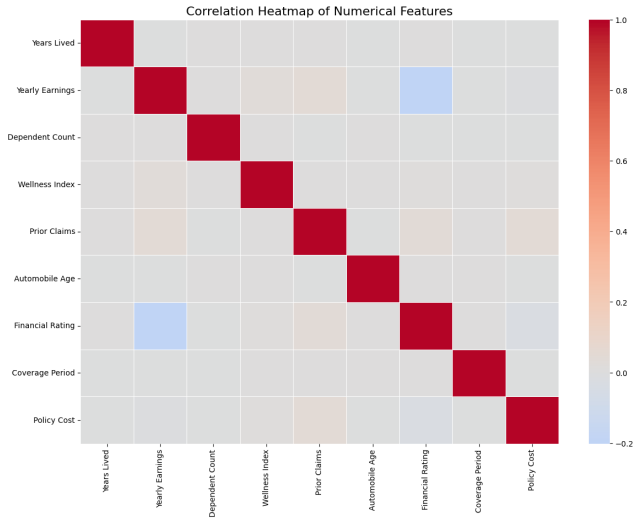


Figure 3: Correlation heatmap of numerical features, highlighting dependencies relevant to prediction.

Feature variability was also inspected (Figure 4), where features such as yearly earnings and wellness index demonstrated significant spread. This variability provided useful signal for modeling, although skewed distributions necessitated transformations.

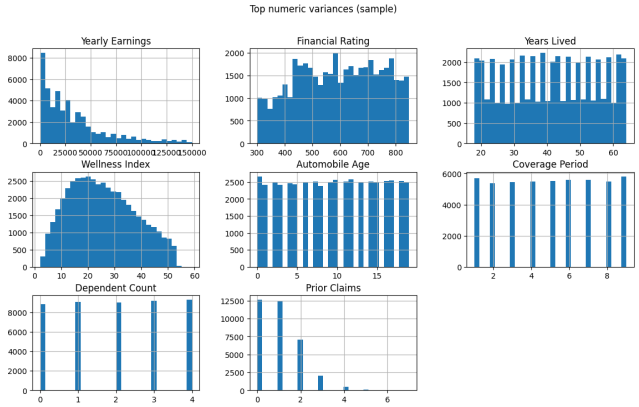


Figure 4: Sample distributions of selected numeric features, showing high variance across the dataset.

## 4 Feature Engineering

Feature engineering sought to extract meaningful patterns while controlling complexity. Temporal features were derived from policy coverage dates, including year, month, day of the week, and an indicator for seasonality. Additionally, policy age was computed as the number of days since coverage began, capturing potential temporal drift between older and newer policies.

To reduce skewness, selected numerical variables were log-transformed, while others were discretized into quantile-based bins, allowing linear models to approximate nonlinear effects. Interaction terms were created

where domain knowledge suggested synergy, such as between earnings and dependent count. High-cardinality categorical variables were encoded through smoothed target encoding in an out-of-fold manner, ensuring unbiased estimation. These engineered features substantially enriched the representation of the dataset and improved the models' ability to generalize.

# 5 Modeling Approach

Modeling began with a ridge regression baseline, which offered a simple and interpretable reference point. Elastic net regression was then introduced to leverage both L1 and L2 regularization, achieving a balance between sparsity and stability. Each model was implemented within a composite pipeline that included preprocessing, encoding, and scaling, guaranteeing consistency across validation folds.

Evaluation relied on five-fold cross-validation, producing out-of-fold predictions that were compared against observed targets. Hyperparameters were tuned selectively, prioritizing computational feasibility. The primary evaluation metric, RMSLE, was chosen for its robustness to outliers and interpretability in the insurance context.

# 6 Results

The results showed consistent improvements over naive baselines. Ridge regression stabilized variance but lacked flexibility in capturing complex interactions. Elastic net regression, by combining feature selection with shrinkage, produced lower RMSLE values and proved more resilient to multicollinearity. Feature engineering played a decisive role: temporal variables reduced error, log transformations stabilized distributions, and target encoding enhanced categorical representation. In contrast, aggressive one-hot encoding of high-cardinality features increased model variance without significant gains.

# 7 Discussion

The analysis suggests that the predictive structure of policy cost is driven by both temporal and socioeconomic attributes. The inclusion of coverage start date features emphasized that risk patterns evolve seasonally and over time. Target encoding demonstrated its value in handling categorical variables with many levels, provided that careful cross-validation prevents leakage. Although linear models delivered competitive performance in this study, future extensions could benefit from tree-based ensembles such as gradient-boosted decision trees, which naturally capture nonlinearities without extensive feature design. Moreover, stacking linear and nonlinear models offers a promising avenue for further improvement.

# 8 Conclusion and Future Work

This project documented the complete path from raw data to validated predictive models for insurance policy cost estimation. By integrating preprocessing, feature engineering, and model evaluation into a coherent pipeline, we achieved robust and interpretable results. The work highlights the importance of temporal features, careful handling of categorical data, and rigorous validation in avoiding overfitting. Future research should expand hyperparameter tuning, explore ensemble approaches, and incorporate external data sources such as macroeconomic indicators to further refine predictions. Such improvements would enhance both predictive accuracy and the practical utility of the models in real-world actuarial practice.

# References

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer. Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.