# A Push Towards Leverageing GBMs in Personal Auto Insurance Pricing Models

Helen Moses
*Department of Statistics*
*University of Michigan*
Ann Arbor, United States
hemoses@umich.edu

***Abstract*—Rate making in insurance relies on statistical models with accurate prediction capabilities. Due to regulatory constraints, the current gold standard for rate making is Linear Models (LMs). I built an LM with the package statsmodels and a Gradient Boosted Machine (GBM) with XGBoost to showcase the advantages of GBMs in premium prediction. Additionally, I show that with the aid of Shapley Additive Explanation (SHAP) values, GBM predictions can be explained. Due to the robust, automated, and explainable reality of GBMs, regulators should reconsider their current exclusion from rate making.**

## I. Introduction

While working with a property and casualty insurance company, I learned that one of the hot topics in modeling is whether or not gradient boosted machines (GBMs) should be utilized for pricing models. Currently, regulators do not allow GBMs due to their black box nature. However, statisticians have found that GBMs are capable of capturing more nuanced relationships between features compared to multiple linear regression models (LMs), despite LMs being the current gold standard. Blier-Wong et al. hypothesized that learning models such as GBMs outperform LMs due to their unique ability to go beyond capturing linear relationships and capture much more complex interactions between features and the outcome [1]. Insurance premiums are set based on many different complicated factors, so it makes sense that this would be an applicable setting for this kind of problem. Dugas et al. found that within the application of rate making in auto insurance, neural networks have lower MSEs compared to linear models for training data, testing data, and validation data [2]. Clemente, Guerreiro and Bravo looked specifically at the performance of GBMs compared to LMs in the prediction of claim frequency (rather than ratemaking) and found that the GBMs are superior [3].

This project aims to replicate the results that GBMs can achieve more accurate pricing predictions than LMs. There are many different types of tree based models and boosting techniques. I leveraged XGBoost to create my GBM due to the finding that it outperforms Adaboost "unpublished" [4]. As the main critique of GBMs is their black box nature, this project will also go further to demonstrate how GBMs can be demystified if they were to be used in pricing models.

Code Base: https://github.com/hemoses27/stats507

## II. Methods

### A. Data

The dataset used for this analysis is available on kaggle under the name Auto Insurance Dataset [5]. The dataset is in a tabular format with 9,134 observations and 24 variables. Each observation represents an individual customer with an auto insurance policy. The variables include information about the customer (education, gender, marital status, income, etc.), the specified policy (monthly premium, number of policies, type of coverage, months since inception, etc. ), the vehicle (vehicle class and vehicle size), and the claim data (total claim amount and months since last claim). The dataset is fairly clean and contains no missing values.

For more reliable results, I split the data into a training and testing dataset using the scikit-learn package [6]. It was a random 80, 20 split resulting in 7,308 observations in the train set and 1,827 observations in the test set.

### B. Modeling

As the goal of this project is to determine the relative performance capabilities of an LM and a GBM in auto insurance price prediction, the outcome for both models is the continuous variable 'Monthly Premium Auto'. I built the LM with the package statsmodels [7]. To find the most accurate version of the LM, I utilized backwards selection. Once I removed all of the insignificant predictors, I added polynomial terms to the continuous variables if it resulted in an increased $R^2$ and a decreased Akaike Information Criterion (AIC). The final model is represented in (1). Although the model is technically interpretable, it is very convoluted with all of the polynomial terms added.

I utilized the package XGBoost to build the GBM [8]. The first iteration of the model contained all of the available features in the dataset. However, the final model only contains the top seven most important features. I used the built in feature importance function to determine the impactful features. The features I included are Customer Lifetime Value, Income, Number of Policies, Total Claim Amount, Coverage, Employment Status, and Vehicle Class. The categorical features were one-hot encoded so that they could be properly handled by the XGBRegressor. Once the features were finalized, I tuned

the hpyerparameters to improve the model performance with respect to RMSE and $R^2$. The final model has 500 estimators, a max depth of 4, and a learning rate of 0.05. Unlike the LM, there is no closed form equation, hence why regulators do not currently allow them in insurance pricing models.

$$
\begin{aligned}
Y = \beta_0 &+ \beta_{\text{Edu,College}}\,\mathbf{1}\{\text{Education} = \text{College}\} \\
&+ \beta_{\text{Edu,Doctor}}\,\mathbf{1}\{\text{Education} = \text{Doctor}\} \\
&+ \beta_{\text{Edu,HS}}\,\mathbf{1}\{\text{Education} = \text{High School or Below}\} \\
&+ \beta_{\text{Edu,Master}}\,\mathbf{1}\{\text{Education} = \text{Master}\} \\
&+ \beta_{\text{VS,Med}}\,\mathbf{1}\{\text{Vehicle Size} = \text{Medsize}\} \\
&+ \beta_{\text{VS,Small}}\,\mathbf{1}\{\text{Vehicle Size} = \text{Small}\} \\
&+ \beta_{\text{Male}}\,\mathbf{1}\{\text{Gender} = M\} \\
&+ \beta_{\text{Married}}\,\mathbf{1}\{\text{Marital Status} = \text{Married}\} \\
&+ \beta_{\text{Single}}\,\mathbf{1}\{\text{Marital Status} = \text{Single}\} \\
&+ \beta_1\,\text{CLV} + \beta_2\,\text{CLV}^2 + \beta_3\,\text{CLV}^3 \\
&+ \beta_4\,\text{Income} + \beta_5\,\text{Income}^2 \\
&+ \beta_6\,\text{MonthsSinceInception} \\
&+ \beta_7\,\text{TotalClaim} + \beta_8\,\text{TotalClaim}^2 + \beta_9\,\text{TotalClaim}^3 \\
&+ \beta_{10}\,\text{TotalClaim}^4
\end{aligned}
\tag{1}
$$

### C. Gradient Boosted Machine Explainability

Shapley Additive Explanation (SHAP) can be used to demystify the black box nature of a GBM [9]. The summary plots provide insight on the global SHAP values. Features appearing at the top of the plot indicate high importance, whereas features at the bottom of the plot indicate less importance. Additionally, color coding within the plot shows how the value of the feature impacts the prediction. Waterfall plots provide insight into local SHAP values. They can be used to obtain information about how a prediction for a specific observation was made based on the feature values.

## III. Results

### A. Models

To evaluate the performance of the LM and GBM, I used RMSE, $R^2$, and a visualization of prediction results compared to the actual monthly premium values. For each of these measures, the GBM greatly outperforms the LM. Table I shows that the GBM has a much higher $R^2$ compared to the LM, indicating that the GBM captures a lot more of the outcome's explainability. Additionally, Table I highlights that the GBM has a much lower RMSE compared to the LM. A much lower RMSE is further evidence that the GBM captures a more comprehensive fit.

Fig. 2 contains a visualization of what the enhanced prediction capabilities look like in practice. Points along the red line indicate a perfect prediction. The GBM contains many predictions along the line, and few that stray far from the line. The LM is the opposite where few predictions fall on the line and many stray from the line. Additionally, Fig. 1 highlights that the LM is very limited, and most prediction results are concentrated within 50 to 150 despite actual premiums being concentrated within 50 to 200.

TABLE I
Model Performance Metrics

| Model | RMSE | $R^2$ |
|---|---|---|
| LM Test | 21.55 | 0.614 |
| GBM Test | 3.85 | 0.988 |
| LM Train | 23.85 | 0.587 |
| GBM Train | 3.07 | 0.992 |

### B. Gradient Boosted Machine Explainability

Section III-A, provides overwhelming evidence that the GBM has far better price prediction capabilities compared to the LM. However, regulators will never approve a black box model. As such, the secondary aim of this project is to demystify the black box nature of GBMs. Figs. 3, 4, and 5 provide a visualization that explains how the features impact the prediction.

Fig. 3 shows that the three most impactful features are whether the coverage is basic, whether the car is an SUV, and the customer's lifetime value. We can see that basic coverage policies are associated with smaller predictions, vehicles classified as SUVs are associated with larger predictions, and lower customer lifetime values are associated with lower premium predictions.

Figs. 4 and 5 give two unique examples of how the model determined specific premium predictions. In Fig. 4 we see that there are many factors lowering the prediction, but the fact that it is not basic coverage brings back up a bit. In Fig. 5 we see that there are a few factors lowering the prediction, but it is ultimately driven up due to it being a premium coverage policy.

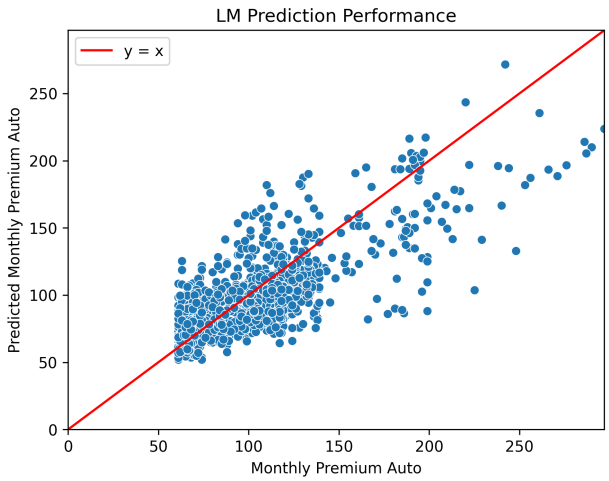


Fig. 1. LM predicted monthly premium vs. actual monthly premium for testing dataset.

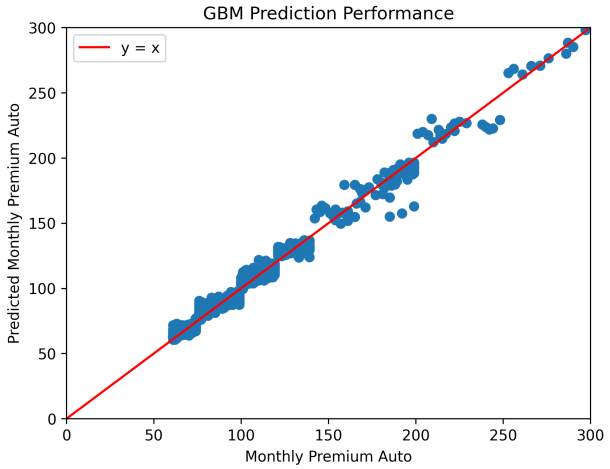


Fig. 2. GBM predicted monthly premium vs. actual monthly premium for testing dataset.

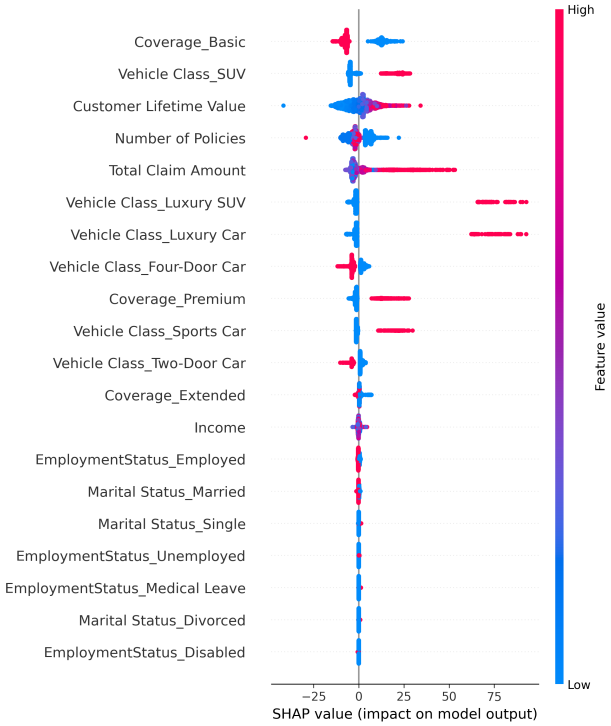


Fig. 3. Summary plot of global SHAP values for GBM.

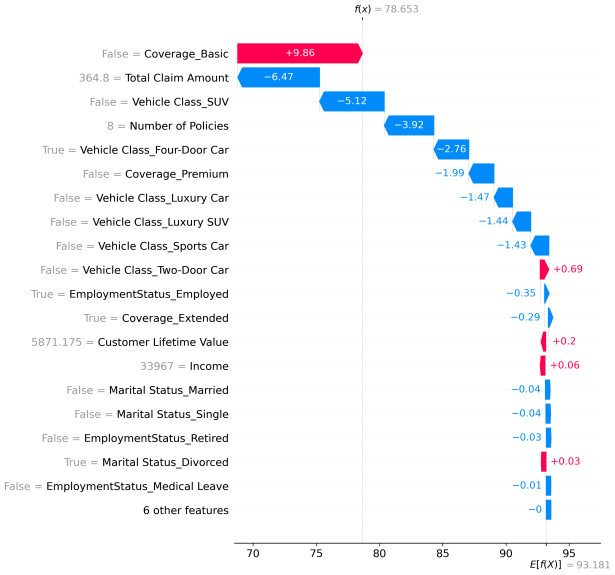


Fig. 4. Waterfall plot of local SHAP values for observation 1 in the training dataset.

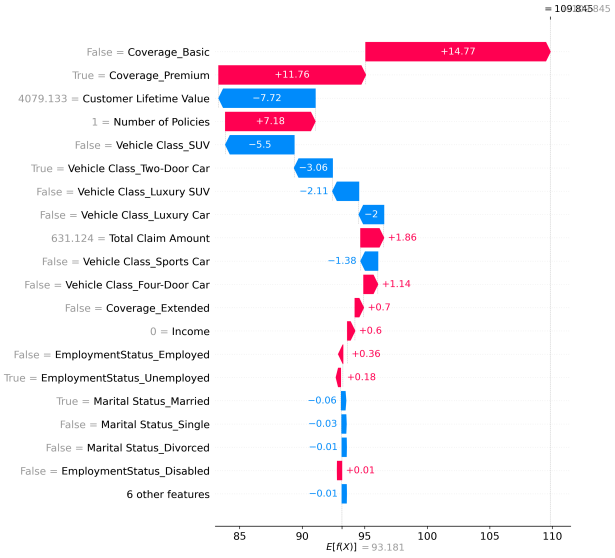


Fig. 5. Waterfall plot of local SHAP values for observation 158 in the training dataset.

## IV. Conclusion

Pricing models are necessary within the insurance industry so that underwritters can identify the appropriate premium for each policy. Currently, insurance industries are forced to use LMs due to regulations prohibiting GBMs. Unfortunately, as shown in section III-A, the LMs have very weak premium prediction capabilities. This results in less optimum rate making. Fortunately, GBMs pose a superior alternative and are much more explainable than regulators realize.

Linear models must be manually built, which requires a lot of time and knowledge from the statistician. Determining the appropriate interactions to include in the model requires significant domain knowledge that is likely unattainable. GBMs, however, are capable of picking up the feature interactions on their own in the training phase. As shown in section III-A, this results in very accurate premium predictions.

Moreover, section III-B highlights that SHAP values are very useful in demystifying the black box nature of GBMs. With just a few plots, we can see exactly how and why the GBM made its predictions. Due to these findings, regulators must revaluate whether GBMs should be used in insurance rate making.

## REFERENCES


[1] C. Blier-Wong, H. Cossette, L. Lamontagne, and E. Marceau, "Machine learning in p&c insurance: a review for pricing and reserving," Risks, vol. 9, no. 1, pp. 4, 2021.

[2] C. Dugas, B. Yoshua, C. Nicolas, V. Pascal, D. Germain, and F. Christian, "Statistical learning algorithms applied to automobile insurance ratemaking," Arlington: Casualty Actuarial Society Forum, pp. 179-213, 2003.

[3] C. Clemente, G. R. Guerreiro, and J. M. Bravo, "Modeling motor insurance claim frequency and severity using gradient boosting," Risks, vol. 11, no. 9, pp. 163, 2023.

[4] A. Ferrario and R. Hammerli, "On boosting: theory and application," 2019, unpublished.

[5] J. Singh, 2021, "Auto Insurance Dataset," kaggle. [Online] Available: https://www.kaggle.com/datasets/singhnproud77/auto-insurance-dataset.

[6] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Version 1.7.2, scikit-learn Developers, 2011. [Online]. Available: https://scikit-learn.org/

[7] S. Seabold and J. Perktold, "statsmodels: Econometric and Statistical Modeling with Python," Version 0.14.0, statsmodels Developers, 2010. [Online]. Available: https://www.statsmodels.org/.

[8] T. Chen and C. Guestrin, "XGBoost: Scalable and Flexible Gradient Boosted Trees," Version 3.1.2, XGBoost Developers, 2016. [Online]. Available: https://xgboost.ai/

[9] S. M. Lundberg and S.-I. Lee, "SHAP: SHapley Additive exPlanations," Version 0.50.0, SHAP Developers, 2017. [Online]. Available: https://shap.readthedocs.io/

[10] Code base available at: https://github.com/hemoses27/stats507