Case Study for Insurance Modeling

Jinhuizi (Jeanne) Fu

May 12, 2024

1 Introduction

This case study will focus on pricing strategies for a commercial auto line of business. The goal of the study is to incorporate Machine Learning to improve pricing accuracy, while maintain explainability.

Throughout the life cycle of a insurance pricing project, there are in general six steps, and we will discuss in more detail how to use ML in each step:

- Scoping (Set up a goal for the project, success criteria and/or metrics, resource, cost, etc.)
- Data Preparation (Data sourcing, EDA, data cleaning)
- Feature Engineering (Transformation, variable selection)
- Modeling (Benchmarking, feature importance ranking, model, validation)
- Implementation (Deploy model for business end-users, testing)
- Monitoring (and prepare for model refreshing)

This case study will be focusing on Data Preparation, Modeling and Implementation. I will discuss about potential risks and how to manage them in the end.

2 Data Preparation

In general, when preparing a modeling dataset for a pricing model, we will put together both internal and external data sources, and prepare for modeling data.

Internal data source:

- Policy data: account/policy information, exposure, location, primary usage of vehicle, business industry, etc.
- Loss data: (Target, loss history). Loss linkage may be needed.
- Vehicle information: vehicle weight, vehicle age, vehicle cost, etc. May use external VIN decoding service if needed.

External data source:

- Driver information (credit history, police driving record, etc.)
- Credit history (for the business)

In this case, I downloaded a toy dataset as a quick sample to explore different modeling methods. The source is from this website: https://data.mendeley.com/datasets/5cxyb5fp4f/1

Things to do when I have more time:

- Explore missing values. Check why they are missing, either back fill with distribution, or define new "Missing" category.
- Check distribution and transform if needed

3 Feature Engineering

A lot of explorations could be done here. Include but not limited to:

- Encoding catergorical variables
- Normalization for certain variables. For example, normalize historical claim count by a variable accounting for policy size, will help isolate the feature from account size.
- Explore feature interaction
- Bin numerical variables, and/or explore polinomial trend

4 Model Training and Validation

* Split Training, Test, Holdout (or Cross Validation)

* Selection of Target variables: Frequency and Severity VS. Pure Premium

Things to think about:

- Different coverages
- Outlier in pure premium
- Need to develop loss (alternative is to use policy year as control variable)
- Need to trend

4.1 AutoML for model Benchmarking

AutoML (automated machine learning) is a framework to run machine learning models automatically. For insurance pricing, because of regulation restrictions, GLM is still the most used model structure due to its simplicity and explainability. However, AutoML framework can be used as a first step in pricing models.

A few ways to use AutoML in pricing model:

- Model benchmarking: AutoML can run multiple machine learning models at one time, so we can compare the performance of different models.
- Variable importance ranking: Shapley plot can be used to explore variable importance.
- Partial dependence plot: like traditional one-way plot for each features VS target variable, we can examine the upward/downward trend when the feature increases/decreases.

Output from my sample run can be checked out here: $\mathrm{h2o}_{a}utoml.ipynb$

4.2 AutoML for model variable importance (Shapeley Value)

This is a beeswarm plot, that summarizes the entire distribution of SHAP values for each feature:

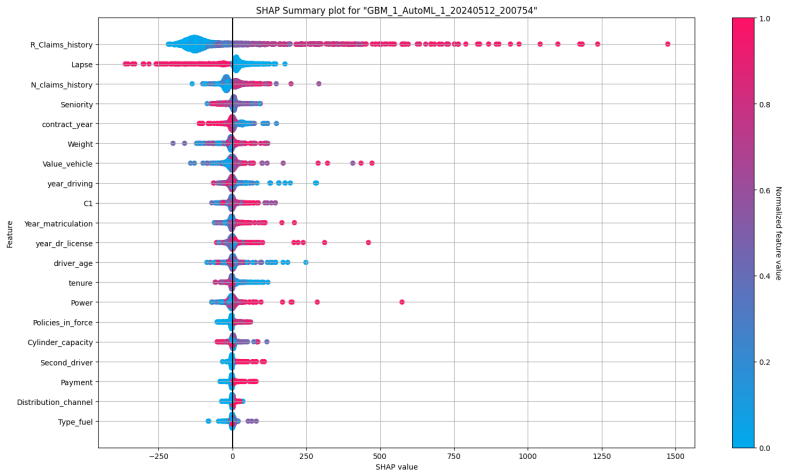

Figure 1: Shapley Plot

- SHAP value: calculated for each record (row), SHAP values are based on game theory and assign an importance value to each feature in a model. Features with positive SHAP values positively impact the prediction, while those with negative values have a negative impact. The magnitude is a measure of how strong the effect is.
- Feature importance is ranked by mean absolute SHAP value. In the plot, normalized claim count history is ranked most important.
- Horizontal axis is SHAP value for each record. Each dot is a record, and they "pile up" vertically when different dots have the same value.

5 Potential Risks and How to Manage Them

Explainability The final model could still be developed with GLM. ML benchmarking could be used to compare model performance.

Overfitting Machine learning models tend to over-fit more than traditional models (e.g. GLM). GLM with regularization (elastic net) will have the capability to select features and prevent overfitting. Cross-validation can be used to validate model and prevent overfitting.

6 Conclusion

Machine learning methods could be integrated into different steps when developing an insurance princng model. It could be used in feature engineering like encoding, normalization, and interaction, etc. AutoML is good for benchmarking and comparing different model architectures. AutoML outputs feature importance ranking, and partial dependency plot, that could be used to help with modeling explainability.