This report investigates two machine learning models to determine the future expected cost of a given automobile insurance policy that can be used as part of Intact’s pricing strategy. It compares the high-level features of each model and fits a basic implementation for each model to a test dataset to gauge comparative performance. The report then makes a model recommendation based on the comparison and the performance results.

The first model is a generalized linear model or ‘GLM’, which is the existing dominant model for costing policies in the P&C industry and seeks to determine a linear relationship between the rating variables and the distribution of the random future loss variable. The second model is a gradient boosted decision tree model, using the more complicated ‘XGBoost’ algorithm that seeks to use iterative decision trees to predict the expected future cost of a policy to a high degree of accuracy.

While both models fulfill the intended purpose of insurance costing, the XGBoost model outperforms the GLM in the test dataset and thus predicts the policy cost much more accurately. It does however struggle with interpretability and analysis as it is a more complicated model then the GLM and therefore makes inferring business decisions from its results more challenging.

The recommendation is to deploy the XGBoost model to determine policy cost as the gains from a pricing strategy with more accurate costing information outweighs the extra analysis needed to infer information from the model. As Intact is in the position to invest in cutting edge modelling techniques, it should move ahead with implementing XGBoost to remain ahead of competing insurers using less accurate models.

Insurance pricing:

The business of insurance is one of risk management. In property and casualty or P&C insurance this risk is related to automobile and homeownership risks, both in personal and commercial sectors. Whenever a driver signs up for auto insurance, they represent a risk towards the insurance company, and it is thus the job of the insurance company to price their policies so that they can cover the risk for their policyholder along with expenses and profit. Estimating the risk posed by any given policyholder is a difficult job, and represents a large amount of the effort that goes into pricing insurance.

The premium for a policy that would theoretically cover the expected future risk posed by that policyholder is known as the costing premium. This costing premium is then combined with an estimate of the expenses associated with the policy, and then a provision for profit. Some companies may also adjust premiums of individual policies to achieve different business goals in the context of the whole book of business, such as ensuring a certain book size or customer experience. The accuracy of this calculation is very important, as ultimately an insurance company that can better predict and price risks will outcompete other insurers in the market by attracting the most profitable and less risky customers with low rates while leaving the risky, expensive customers to their competitors.

Because of this it’s clear that having a accurate prediction model for insurance risk is extremely important for the long-term health of the business. Insurers put a large amount of effort into collecting data for use in pricing policies, and spend substantial resources on researching and developing new pricing models to remain ahead of competitors. Most insurers in Canada usually use a form of generalized linear models or GLMs when pricing products – which due to its properties end up being a natural fit for insurance. However, as more research has gone into machine learning and artificial intelligence more companies are looking into more advanced prediction models, such as the XGBoost algorithm which is a more advanced machine learning method that may be more accurate then a GLM. As companies look into these new methodologies, it is valuable to analyze the differences between GLMs and XGBoost, as well as fit a version of both models to simulated data and compare their accuracy.

Modelling Options:

Generalized Linear Models

Generalized Linear Models or GLMs are the current standard for P&C insurance pricing in Canada. We know for any policy our unknown true costing premium (designated by ‘’) is a random variable of some distribution, and we designate as the mean of . We also have rating variables that we are trying to use to determine the costing premium. In a generalized linear model, we assume our costing premium can represented as the follows:

Where are unknown parameters, is some error term and is some ‘link’ function that connect the output of the linear equation with the mean . Our aim when modelling a GLM is to estimate the parameters , which we can then use to estimate the costing premium for any combination of variables.

Typically, we assume usually follows a ‘Tweedie’ distribution where the number of claims is Poisson distributed and the size of claims is Gamma distributed. Moreover, the ‘link’ function is usually the natural logarithm, giving us our estimated costing premium of:

This gives a useful property where the predicted costing premium can never be negative, which makes sense in the context of insurance.

One critical advantage of GLMs is that it is very easy to interpret, as it is easy to determine how the predicted costing premium will respond to a change in rating variable by examining the parameter of the rating variable. This is a valuable property especially for automobile insurance, where pricing schemes must be justified to local regulators.

The main disadvantage is our initial assumption – that our costing premium can be represented as a linear combination of the rating variables. This may not be true – in fact in real data it is almost certainly not true – and we just hope that it does not bias our model too much. The ability to model more complicated relationships in one of our motivations to find a better model structure.

XGBoost:

The second possible model is the eXtreme Gradient Boost or XGBoost model. Gradient boosting is a method for transforming a relatively simple base model into a more complicated model. First, we fit a base model to the data, giving us:

Where is the simple base model fit to the rating variables , is the true costing premium for that policy, and is the error difference between the two. We now fit another version of the base model to , essentially using a model to predict the error to our first model. We then combine these two models to get a combined model by adding them as follows:

Where is determined to be the value that makes as accurate as possible. Note that this process can then be repeated, until the resulting model do not become any more accurate.

XGBoost applies this gradient boosting approach with some minor changes to make the resulting model more accurate on data the model has not been trained on. It should be noted that gradient boosting is not a model in and of itself but rather a method to transform an initial weak model into a stronger one. Thus, in order to apply XGBoost an initial model must be chosen.

Typically, gradient boosting is done with decision trees. A decision tree simply decides on a prediction value using a series of simple if/else statements similar to a flowchart. These trees are usually built in way to try and make the resulting prediction as accurate to the training data as possible, but as the model structure is very simple it can be prone to both failing to capture important information while also capturing irrelevant or incorrect information. Applying the XGBoost algorithm using the decision tree model however can result in a very robust final model capable of learning much more complicated information. The capability to learn information free of assumptions such as linearity is one reason why in theory the XGBoost algorithm should outperform a GLM in most situations.