UNIVERSITY OF WATERLOO

Department of Mathematics

Analysis of Different Machine Learning Models

for the Purposes of Insurance Costing

Intact Insurance

Toronto, ON M5G 0A1

**Prepared By**

Shea Cardozo

Student ID: 20659614

User ID: SACARDOZ

4A Statistics/Actuarial Science

December 15, 2019

December 16, 2019

Shivani Karani, Actuarial Analyst

Intact Insurance

Toronto, ON M5G 0A1

Dear Shivani,

As we agreed, I have prepared the enclosed report, “Analysis of Different Machine Learning Models for

the Purposes of Insurance Costing” for my third of five work term reports in my 4A term. I hereby

confirm that I received no help, other than what is mentioned above, in writing this report. I also

confirm that this report has not been previously submitted for academic credit at this or any other

academic institution.

The Rating Revolution Homeowners New Business team aims to innovate in the field of homeowner’s

insurance pricing by leveraging new machine learning techniques. My job as an Actuarial Analyst

required that I validate data to facilitate modelling and creating tools to visualize different modelling

scenarios. This report is an in-depth study of the implementation of the company’s new XGBoost

models.

The Faculty of Mathematics requests that you evaluate this report for command of topic and technical

content/analysis. Following your assessment, the report, together with your evaluation, will be

submitted to the Math Undergrad Office for evaluation on campus by qualified work report markers.

The combined marks determine whether the report will receive credit and whether it will be considered

for an award.

Thank you for your assistance in preparing this report.

Shea Cardozo

**Executive Summary**

This report investigates two machine learning models to determine the future expected cost of a given automobile insurance policy to be used as part of an insurance firm’s pricing strategy. It compares the high-level features of each model and fits each one to a simulated 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 property and casualty (or P&C) insurance industry. It works by determining a linear relationship between the rating variables and 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 claim severity much more accurately. It does however struggle with interpretability compared to GLMs which therefore makes inferring business decisions from its results more challenging.

The recommendation is to deploy the XGBoost model to determine policy cost because 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.

**Table of Contents**

Executive Summary………………………………………………………………………………………………. iii

Table of Contents………………………………………………………………………………………….………. iv

List of Figures…………………………………………………………………………….…………………………... v

Introduction…………………………………………………………………….……….………………………….... 1

Modeling Options…………………………………………………………………………………………………… 3

Generalized Lined Models………………………………………………….…………………….…. 3

XGBoost............................………………………………………………….…….…………….…… 5

Analysis……………................……………………………………………………….….…………………………. 7

Conclusion……………................…………………………………………………………………………………. 11

Recommendations……………................……………………………………………………….……………. 12

Bibliography............……………................……………………………….……………………….….……... 13

Acknowledgements……………..................……………………………………………………….…………. 14

**List of Figures**

Figure 1: An example of a simple 2D linear model……………………………………………....... 3

Figure 2: An example of a simple decision tree expressed as a flowchart.…………...... 6

Figure 3: Comparison of RMSE values among tested models..................................... 8

Figure 4: Lift curve comparison of all models and true values.................................... 9

Figure 5: Lift curve comparison of all models without true values............................. 10

**Introduction**

The business of insurance is one of risk management. In property and casualty or P&C insurance this risk is most commonly related to a policyholder’s home or car, 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 is clear that having an 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 investigate 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.

**Modeling Options**

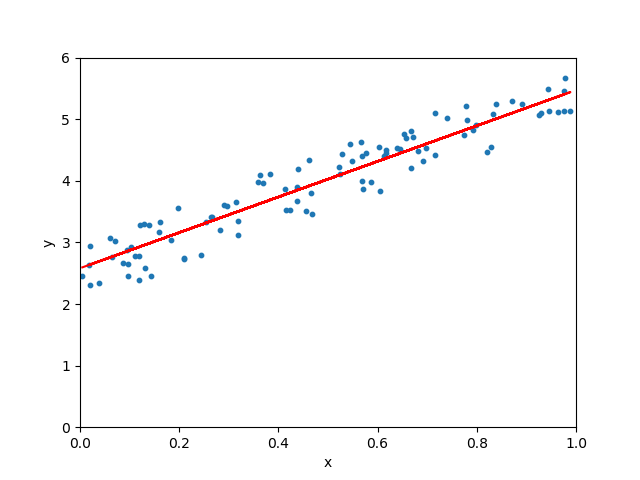
**Generalized Linear Models**

Figure 1: An example of a simple 2D linear model

Generalized Linear Models or GLMs are the current standard for P&C insurance pricing in Canada. For any policy the unknown true costing premium (designated by ‘’) is a random variable of some distribution, and is designated as the mean of . There are also rating variables that are used to determine the costing premium. In a generalized linear model, it is assumed the 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 . The aim when modelling a GLM is to estimate the parameters , which can then be used to estimate the costing premium for any combination of variables.

Typically, it is assumed usually follows a ‘Tweedie’ distribution, which is a compound distribution where the frequency of claims follows a ‘Poisson’ distribution and the severity of each individual claim follows a ‘Gamma’ distribution (with the total loss being equal to the frequency times the average severity of the claims). Moreover, the ‘link’ function is usually the natural logarithm, giving us an 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 as it allows decision-makers less familiar with the technical aspects of the model to still get an understanding of how it reacts to changes in rating variables. In addition, the multiplicativity of the model makes it easy for brokers and clients to adjust premiums quickly according to changes in rating variables.

The main disadvantage is the initial assumption – that the costing premium can be represented as a linear combination of the rating variables. In real data this assumption is almost certainly not true, causing a degree of bias in the model. The ability to model more complicated non-linear relationships in one of the 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, a base model is fitted 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. Another version of the base model is fitted to , essentially using a model to predict the error to the first model. The two models are then combined 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 does not become any more accurate.

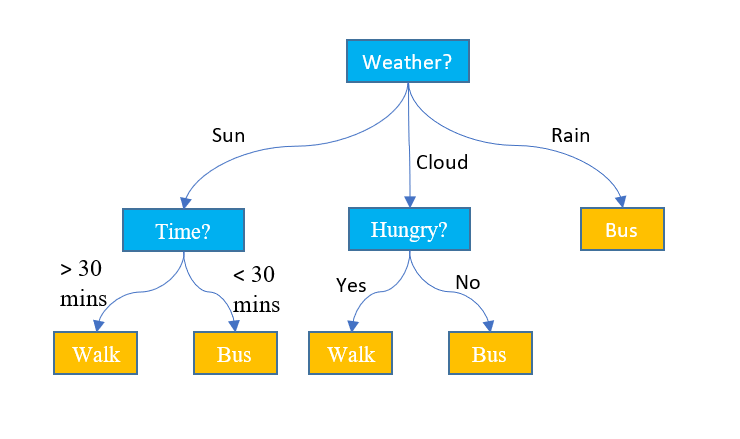
XGBoost applies this gradient boosting approach with some minor changes to less overfit to the data the model has been trained on so that it is better at extrapolating to unseen data. It should be noted that gradient boosting is not a model in and of itself but rather a method to transform an initial 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 like a flowchart.

Figure 2: An example of a simple decision tree expressed as 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.

**Analysis**

In order to compare both methodologies, there needs to be some sort of comparison of both methodologies to determine which model actually performs better when pricing insurance. To do this, both models will be fitted to a simulated set of insurance data to measure their model performance. The models are trained to target incurred Direct Compensation-Property Damage or DCPD severity for a given claim, which covers damage to your vehicle or its contents if another person was at fault for the accident. For the linear model, data with missing values is inputted in each column before training as the model. For the XGBoost model, no preprocessing is needed as the model can handle missing data and the model is simply fit to the raw data. Both models are put through a grid search, which essentially is an algorithm for determining the best parameters to use in model training, and then the final model is used to predict losses on a testing dataset that was not used in training. In addition to the linear model and the XGBoost model, a Null model is also created where the model always predicts the mean incurred DCPD loss in the training dataset for any given policy (essentially ignoring the rating variables). This is to compare the models against the simplest possible model, since if the model underperforms the Null it cannot be a very good model.

To compare the test prediction the root mean squared error or RMSE is used. RMSE is calculated as:

Where is the model prediction and is the true incurred loss for the ith policy. This metric can be thought of a measure of prediction accuracy, with a lower RMSE indicating a more accurate model. The resulting RMSEs for each model are as follows:

*Figure 3: Comparison of RMSE values among tested models*

|  |  |
| --- | --- |
| **Model** | **RMSE** |
| Null | 2688 |
| Linear | 2550 |
| XGBoost | 2427 |

As the results show, both the Linear and XGBoost models outperform the Null model. However, the XGBoost model also outperforms the Linear model significantly by around 5%, rougly by the same amount that the Linear model outperforms the Null model. This suggests that the XGBoost model is the most accurate model at predicting insurance losses.

In addition to the RMSE comparison, another way of visualizing model accuracy is through the use of a lift curve. Lift curves compare the true losses on the x-axis with the predicted losses on the y-axis. A model is said to have more ‘lift’ when its predictions follow an upward trend as the true losses increase, since it indicates it is correctly predicting larger loss amounts for policies that do end up having larger losses.

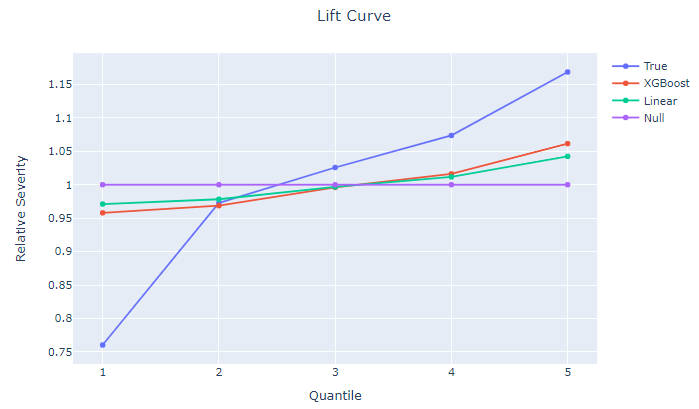


Figure 4: Lift curve comparison of all models and true values

Looking at the generated lift curve, it can be seen that the Null model has no lift at all – which makes sense as it predicts the same loss value regardless of policy. Comparatively, it’s clear that the True severity values have a large amount of lift, far more then any of the tested models.

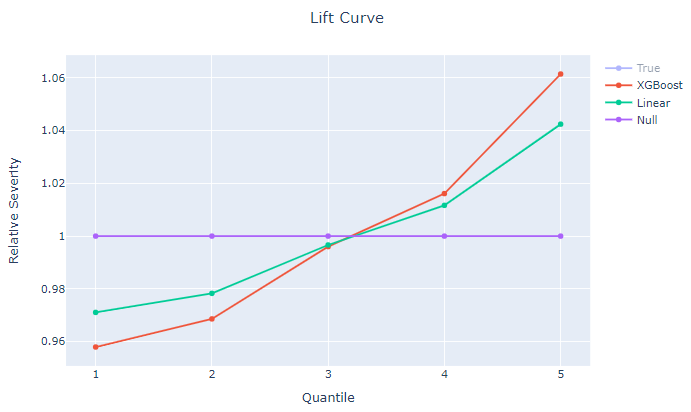


Figure 5: Lift curve comparison of all models without true values

Disabling the True value line to better compare the models, both the Linear and the XGBoost models have positive lift, but the XGBoost model clearly achieves more lift then the Linear model. Again, this indicates that the XGBoost model is doing a better job at sorting out policies with higher losses then the Linear model.

**Conclusion**

Both of the examined models outperform the Null model and are able to successfully predict insurance losses in the simulated data to an adequate degree. Therefore, both models would bring value to the insurance pricing process, as well as other applications that require predicting unknown policy values.

However, the XGBoost model seems to clearly outperform the Linear model when a measure of prediction accuracy such as RMSE is used to compare the model predictions. This suggests that the XGBoost model is better at modelling insurance losses than the Linear model and as a result, its use would lead to more accurate estimate of a policy’s risk.

A visual examination of the resulting lift curve shows that the XGBoost model also seems to do a better job than the Linear model at predicting losses. The XGBoost model clearly predicts riskier policies as having higher losses better than the Linear model does, generating more lift and again demonstrating better predictive capabilities.

**Recommendations**

Because of the clear advantage the XGBoost model provides over the Linear model at predicting loss amounts, it suggests that for that purpose insurance companies should look to migrating their current costing procedures towards XGBoost and possibly other machine learning algorithms. The benefits of more accurate costing prediction will lead to better insurance pricing where premiums better reflect the amount of risk a policyholder actually poses to the company, leading better drivers to pay less for their insurance.

However, there are some concerns with regard to the implementation of XGBoost that may require further addressing before such a model is put into production. While this report demonstrates that XGBoost may be superior to Linear models for the purposes of costing DCPD insurance, when applied to other applications (such as other coverages or lines of business) more analysis may be needed to determine whether XGBoost is the optimal algorithm for that purpose. In addition, for regulated lines of business such as Ontario auto insurance, all pricing models must be cleared with provincial regulators before going into use and as an XGBoost model is significantly more complicated then the Linear alternative it may required more work by insurance filers to justify the resulting changes in pricing to regulators. Again, more analysis may be need to investigate the feasibility of this implementation.

**Bibliography**

* Jake Hoare, *What is a Decision Tree?*, n.d; <https://www.displayr.com/what-is-a-decision-tree/>. Last accessed December 16, 2019.
* Animesh Agarwal, *Linear Regression using Python*, October 5 2018; <https://towardsdatascience.com/linear-regression-using-python-b136c91bf0a2>. Last accessed December 16 2019.
* FSCO, *What do the coverages mean?*, February 24 2016 <https://www.fsco.gov.on.ca/en/auto/brochures/Pages/what-the-coverages-mean-glossary.aspx>. Last accessed December 16 2019.
* Jason Brownlee, *A Gentle Introduction to XGBoost for Applied Machine Learning*, August 21 2019; <https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/>. Last accessed December 16 2019.
* Tianqi Chen, *XGBoost: A Scalable Tree Boosting System*, June 4 2016; <https://speakerdeck.com/datasciencela/tianqi-chen-xgboost-overview-and-latest-news-la-meetup-talk>. Last accessed December 16 2019.

**Acknowledgements**