WIP DRAFT

Towards Explainability of Machine Learning Models in Insurance Pricing

Author One[a], Author Two[b]

[a]Foo; [b]Bar

ARTICLE HISTORY
Compiled October 2, 2019

ABSTRACT
Abstract tbd

KEYWORDS
actuarial science; general insurance

1. Introduction

Risk classification for property & casualty (P&C) insurance rating has traditionally been done with one-way, or univariate, analysis techniques. In recent years, many insurers have moved towards using generalized linear models (GLM), a multivariate predictive modeling technique, which addresses many shortcomings of univariate approaches, and is currently considered the gold standard in insurance risk classification. At the same time, machine learning (ML) techniques such as deep neural networks have gained popularity in many industries due to their superior predictive performance over linear models (LeCun, Bengio, and Hinton 2015). In fact, there is a fast growing body of literature on applying ML to P&C reserving (Kuo 2018; Wüthrich 2018; Gabrielli, Richman, and Wüthrich 2019; Gabrielli 2019). However, these ML techniques, often considered to be completely "black box", have been less successful in gaining adoption in pricing, which is a regulated discipline and requires a certain amount of transparency in models.

If insurers can gain more insight into how ML models behave in risk classification contexts, it would increase their ability to reassure regulators and the public that accepted ratemaking principles are met. Being able to charge more accurate premiums would, in turn, make the risk transfer system more efficient and contribute to the betterment of society. In this paper, we aim to take a step towards liberating actuaries from the confines of linear models in pricing projects, by proposing a framework for explaining ML models for ratemaking that regulators, practitioners, and researchers in actuarial science can build upon.

The rest of this paper is organized as follows: Section 2 provides an overview of P&C ratemaking, Section 3 discusses model interpretability in the context of ratemaking and proposes specific tasks for model explanation, Section 4 describes current model

CONTACT Author One. Email: foo@bar.foo, Author Two. Email: bar@foo.bar

interpretation techniques and applies them to the tasks defined in the previous section, and Section 5 concludes.

2. Property and Casualty Ratemaking

2.1. *History of Ratemaking*

Early classification ratemaking procedures were typically univariate in nature. For example, (Lange 1966) notes that (at that time) most major lines of insurance used univariate methods based around the same principle: distributing an overall indication to territorial relativities or classification relativities based on the extent to which they deviated from the average experience.

(Bailey and Simon 1960) introduced minimum bias methods, which were expanded throughout the 60s, 70s, and 80s. As computing power developed, minimum bias began to give away to generalized linear models, with papers such as (Brown 1988) and (Mildenhall 1999) bridging the gap between the methods.

Arguably, generalized linear models predate minimum bias procedures by a significant margin. The term "Generalized Linear Model" was coined by (Nelder and Wedderburn 1972), but generalizations of least squares linear regression date back at least to the 1930s. Like minimum bias methods, GLMs did not become mainstream in actuarial science for some time. For example, the syllabus of basic education does not seem to include any mention of GLMs prior to (Brown 1988) in the 1990 syllabus for basic education. From there, GLMs seem to have received only passing mention until 2006 with the introduction of (Anderson et al. 2005) to the syllabus.

2.2. *Machine Learning in Ratemaking*

Paralelling the development of generalized linear models was the development of machine learning algorithms throughout the middle part of the 20th century. Detailed histories of machine learning may be found in sources such as (Nilsson 2009) and (Wang and Raj 2017). Consistent with GLMs, machine learning was relatively unpopular in actuarial science until the last ten years as computing power has become cheaper and more easily available and as machine learning software packages have obviated the need for developing analyses from scratch each time an analysis is performed. Due to the breadth of machine learning as a field, it is difficult to identify the first time it entered the CAS syllabus; however, cluster analysis (in the form of k-means) seems to have been first included in 2011 with (Robertson 2009). More recently, the MAS-I and MAS-II exams introduced in 2018 have included machine learning explicitly.

Within the area of ratemaking, machine learning is still in its infancy. A significant portion of machine learning applications to ratemaking has been in the context of automobile telematics, such as (Gao, Meng, and Wuthrich 2018), (Gao and Wuthrich 2018), (Gao and Wuthrich 2019), (Roel, Antonio, and Claeskens 2018), or (Wuthrich 2017). Presumably this focus has been a result of the high-dimensionality and complexity of telematics data, making it a field in which the unique abilities of machine learning techniques give a clear advantage over traditional approaches.

Outside of telematics, (Yang, Qian, and Zou 2018) uses a gradient tree-boosting approach to capture non-linearities that would be a challenge for GLMs. (Henckaerts et al. 2018) makes use of "generalized additive models" to improve predictions of GLMs. Many researchers, in an apparent effort to demonstrate the range of possibil-

ities and advantages of machine learning, have approached the topic by comparing many different machine learning algorithms within a single study, such as in (Dugas et al. 2003), (Noll, Salzmann, and Wuthrich 2018), (Spedicato, Dutang, and Petrini 2018). These studies make use of such varied techniques as regression trees, boosting machines, support vector machines, and neural networks.

2.3. *Ratemaking Process*

Regardless of the method employed for determining this risk of various classifications, the actual process of setting rate relativities typically involves some variation of the following steps:

1. Obtain relevant policy-level data
2. Prepare data for analysis
3. Perform analysis on the data, employing desired method or methods to estimate needed rate relativites
4. Select final rate relativities based on rate indications
5. Present rates to the reglator, including explanation of the steps followed to derive the rates
6. Answer questions from regulators regarding the method employed

The focus of this paper is on steps 5 and 6. In particular, rate regulators are concerned with whether rates are inadequate, excessive, or unfairly discriminatory. In many states, rate filings that exceed certain thresholds for magnitude of rate changes or filings that make use of new or sophisticated predictive models may be subject to intense regulatory scrutiny. In these cases, it is necessary to be able to explain the results of the modeling process in a way that is understandable without sacrificing statistical rigor.

It should be noted that communicating results is not simply a method of passing regulatory muster. Generating interpretable modeling output is an important - even essential - facet of model checking. Therefore, the techniques discussed in this paper may be viewed from the lens of providing useful information to regulators, but they should also be considered as part of a thorough vetting of any rating model.

3. Interpretability in the Ratemaking Context

(literature review, settle on a definition to work with, e.g. Doshi-Velez and Kim (2017), identify maybe 3 questions about a model to answer, tie to principles on p/c ratemaking)

Within the actuarial profession, Actuarial Standard of Practice 41 ("Actuarial Communications") notes that "...another actuary qualified in the same practice area [should be able to] make an objective appraisal of the reasonableness of the actuary's work as presented in the actuarial report." (aso 2010) Underlying this requirement is an assumption that the hypothetical other actuary qualified in the same practice area is adequately familiar with the relevant techniques employed. Although the syllabus of basic education is constantly changing, there has at times been an assumption that all techniques and assumptions that have ever been a part of the syllabus of basic education needn't be explained from first principles in general actuarial communications, and that an actuary practicing in the same field should be able to make an

objective appraisal of the results from the methods found in the syllabus. [This can be supported by reference to the edits to ASOP 38 over time - originally it was designed to be about all models, but they revised it to be about only models that incorporate specialized knowledge outside of the actuary's area of expertise... do we need this citation, though?] This is notable because, beginning with the introduction of the MAS-I and MAS-II examination in July of 2018, several machine learning models were formally included in the syllabus of basic education. These exams cover a wide range of topics, such as splines, clustering algorithms, decision trees, boosting, and principle components analysis. (cas 2018)

Nevertheless, machine learning poses something of a special challenge for ASOP 41 for several reasons. Machine learning models can be very ad hoc compared to traditional statistical models. Because many machine learning models are deterministic, they may not admit of standard metrics for model comparison (e.g., it's not straightforward to calculate an AIC over a neural network). In addition, machine learning methods are often combined into ensembles that may not be easily separated and that may, as a collection, cease to resemble a single standard version of a model. Complicating matters still further, machine learning models can be "black boxes" insofar as the final form of response curve cannot be easily predicted and may depend heavily on the available data (which may not, in turn, be available to the reviewer).

This last item raises a final interesting issue. Generalized linear models and their ilk are often fitted using one of a handful of standard and well-understood approaches (e.g., maximum likelihood estimation). However, this is not possible in general with machine learning models, as machine learning algorithms often use loss surfaces that are very complex such that it may not be feasible to calculate the global minimum of the surface. Certainly, closed form representations of the loss surfaces are not generally available. For this reason, the training phase of a machine learning model is, in many ways, just as important to one's understanding as the model form and the data on which the model is fitted. Because the final model result is inseparable from these three components (training method, model form, and data), it is not generally adequate to just know the method employed to make an objective appraisal of the reasonableness of the model. More information is necessary.

Of course, these comments only apply to the actuarial profession. Outside of the actuarial profession, communication of results may be more challenging. A 2017 survey conducted by the Casualty Actuarial and Statistical Task Force of the National Association of Insurance Commissioners found that the plurality of responding regulators identified "Filing complexity and/or a lack of resources or expertise" as a key challenge that impedes their ability to review GLMs or other predictive models. (nai 2017) Given that machine learning algorithms are generally regarded as more complex than GLMs, this implies that the challenge of communicating machine learning model results is significant.

In response to the same survey, 33 state regulators noted that it would be helpful or very helpful for the NAIC to develop information and tools to assist in reviewing rate filings based on GLMs, and 34 noted that it would be helpful to develop similar items to assist in reviewing "Other Advanced Modeling Techniques." One outgrowth of this need was the development of a white paper on best practices for regulatory review of predictive models. The white paper focuses on review of GLMs, particularly with respect to private passenger automobile and homeowners' insurance. Some of the guidance offered in this regard is therefore not strictly applicable to the review of machine learning models. For example, as previously noted, p-values are not a concept that translates well to deterministic machine learning algorithms. However, among the

guidance applicable to machine learning algorithms are the following (paraphrasing):

- understand the relationship between the inputs and the expected loss or expense differences in risk,
- determine that these input characteristics are not unfairly discriminatory, and
- determine the extent of premium disruption among policyholders and be able to explain the source of premium disruptions. (nai 2018)

Note that this list is not exhaustive of the guidance offered by any means, but that these three items are those that may present greater challenge for machine learning algorithms compared to GLMs.

Depending on the model, machine learning algorithms can be non-linear in the inputs. In areas where data are sparse, the model could generate unnatural response curves that could go undetected by a very high-level view of the model results. This leads to a situation where lift charts or tests on holdout data may indicate that the model is performing adequately, but it may be difficult to explain the reasons for a particular premium indicaton or a change in premium when policy characteristics change.

-For this, we recommend ____ (grid search?) -Because the model form may be difficult to represent in the form of a single equation (and harder still to evaluate), one method of evaluating the reasonableness of model results is local approximations of the impact of different rating variables (LIME?)

4. Applying Model Interpretation Techniques

(some definitions, e.g. global/model vs. local/instance, categorize questions accordingly)

4.1. *(answer each question)*

(for each question, propose a technique, mention alternative techniques, pros/cons, implement, interpret results)

5. Conclusion

(conclude)

References

2010. "Actuarial Standard of Practice No. 41 - Actuarial Communications." http://www.actuarialstandardsboard.org/wp-content/uploads/2014/02/asop041_120.pdf.

2017. "2017 Proceedings or the National Association of Insurance Commissioners." August.

2018. "Regulatory Review of Predictive Models 10/25/18 Exposure Draft." October. https://www.naic.org/documents/cmte_c_catf_exposure_predictive_model_white_paper.pdf.

2018. "Syllabus of Basic Education." 2018. https://www.casact.org/admissions/syllabus/ArchivedSyllabi/2018Syllabus.pdf.

Anderson, D., S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher, and N. Thandi. 2005. "A Practitioner's Guide to Generalized Linear Models." 4–39.

Bailey, Robert A., and LeRoy J. Simon. 1960. "Two Studies in Automobile Insurance Ratemaking." *ASTIN Bulletin* 1 (4): 192–217.

Brown, Robert L. 1988. "Minimum Bias with Generalized Linear Models." 187–217.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards A Rigorous Science of Interpretable Machine Learning." *arXiv:1702.08608 [cs, stat]* .

Dugas, Charles, Y. Bengio, Nicolas Chapados, P. Vincent, G. Denoncourt, and C. Fournier. 2003. "Statistical Learning Algorithms Applied to Automobile Insurance Ratemaking." .

Gabrielli, Andrea. 2019. *A Neural Network Boosted Double Over-Dispersed Poisson Claims Reserving Model*. SSRN Scholarly Paper ID 3365517. Rochester, NY: Social Science Research Network.

Gabrielli, Andrea, Ronald Richman, and Mario V. Wüthrich. 2019. "Neural Network Embedding of the Over-Dispersed Poisson Reserving Model." *Scandinavian Actuarial Journal* 1–29.

Gao, Guangyuan, Shengwang Meng, and Mario V. Wuthrich. 2018. "Claims Frequency Modeling Using Telematics Car Driving Data." *Scandinavian Actuarial Journal* .

Gao, Guangyuan, and Mario Wuthrich. 2019. "Convolutional Neural Network Classification of Telematics Car Driving Data." *Risks* 7: 6.

Gao, Guangyuan, and Mario V. Wuthrich. 2018. "Feature Extraction from Telematics Car Driving Heatmaps." *European Actuarial Journal* 8: 383–406.

Henckaerts, Roel, Katrien Antonio, Maxime Clijsters, and Roel Verbelen. 2018. "A Data Driven Binning Strategy for the Construction of Insurance Tariff Classes." *Scandinavian Actuarial Journal* 2018: 681–705.

Kuo, Kevin. 2018. "DeepTriangle: A Deep Learning Approach to Loss Reserving." *arXiv:1804.09253 [cs, q-fin, stat]* .

Lange, Jeffrey T. 1966. "General Liability Insurance Ratemaking." *Proceedings of the Casualty Actuarial Society* LIII: 26–53.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436.

Mildenhall, Stephen J. 1999. "A Systematic Relationship Between Minimum Bias and Generalized Linear Models." *Proceedings of the Casualty Actuarial Society* LXXXVI: 393–487.

Nelder, J. A., and R. W. M. Wedderburn. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society* 135 (3): 370–384.

Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence*. Cambridge University Press.

Noll, Alexander, Robert Salzmann, and Mario Wuthrich. 2018. "Case Study: French Motor Third-Party Liability Claims." *SSRN Electronic Journal* .

Robertson, J.P. 2009. "NCCI's 2007 Hazard Group Mapping." *Variance* 3: 194–213.

Roel, Verbelen, Katrien Antonio, and Gerda Claeskens. 2018. "Unraveling the Predictive Power of Telematics Data in Car Insurance Pricing." *Royal Statistical Society* .

Spedicato, Giorgio, Christophe Dutang, and Leonardo Petrini. 2018. "Machine Learning Methods to Perform Pricing Optimization: A Comparison with Standard Generalized Linear Models." *Variance* 12.

Wang, Haohan, and Bhiksha Raj. 2017. "On the Origin of Deep Learning." *arXiv:1702.07800v4 [cs.LG]* .

Wuthrich, Mario V. 2017. "Covariate Selection from Telematics Car Driving Data." *European Actuarial Journal* 7: 89–108.

Wüthrich, Mario V. 2018. "Machine Learning in Individual Claims Reserving." *Scandinavian Actuarial Journal* 2018 (6): 465–480.

Yang, Yi, Wei Qian, and Hui Zou. 2018. "Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models." *Journal of Business and Economic Statistics* 36: 456–470.

Appendix A. (data and models)

(since the actual model training isn't a primary topic, put it in appendix for now)