European Journal of Electrical Engineering and Computer Science
Vol 8 | Issue 5 | October 2024
ISSN 2736-5751

RESEARCH ARTICLE

Ordered Lorenz Regularization (OLR): A General Method to Mitigate Overfitting in General Insurance Pricing via Machine Learning Algorithms

Justin Morgan and Yanzhen Qu*

ABSTRACT

When machine learning algorithms are used to determine the price of general insurance, they can sometimes overfit the data. This overfitting can lead to problems for both customers and insurance companies. To address this issue, we've developed a new approach called Ordered Lorenz Regularization (OLR). We have tested OLR on general insurance data. The results have demonstrated that OLR is successful in reducing overfitting. Additionally, when we use OLR for pricing general insurance, it helps establish the lowest and highest possible premiums.

Keywords: General insurance, Machine Learning, overfitting, pricing.

Submitted: August 12, 2024
Published: October 04, 2024

10.24018/ejece.2024.8.5.646

Colorado Technical University, USA.

Corresponding Author:
e-mail: yqu@coloradotech.edu

1. Introduction

Insurance providers frequently use machine learning [ML] to develop risk-based pricing for general insurance. Risk-based pricing involves segmenting the customer base by insurance risk and determining the cost of providing each segment with insurance coverage [1]. Finer segmentation provides opportunities for competitive advantage. However, finer segmentation also invites overfitting [2]. Overfitting can result in inaccurate, inadequate, and excessive insurance premiums.

Ordered Lorenz curves were introduced to measure the effectiveness and value of risk-based insurance pricing models [3]. Here, we extend ordered Lorenz curves to postulate a new method, called Ordered Lorenz Regularization [OLR], for mitigating overfitting when using ML to price general insurance. Ordered Lorenz Regularization modifies the predictions of existing algorithms for pricing general insurance, and it applies generally to such algorithms. Experimental results indicate that OLR tempers inaccurate premium predictions and imposes minimums and maximums that serve to alleviate concerns about inadequate and excessive premiums.

The remainder of this article is organized as follows. Section 2 discusses related works, including ordered Lorenz curves and existing methods for mitigating overfitting. The problem statement, hypothesis, and research question used to guide this study are provided in Section 3. Section 4 provides the derivation of OLR and details the design science approach used to test its effectiveness. Experimental results are presented in Section 5. Section 6 concludes this work by providing context for applying OLR to pricing general insurance with machine learning and suggestions for further research.

2. Related Works

2.1. Ordered Lorenz Curves and Lift Curves

Actuaries designed lift curves to measure the economic value of predictive models when applied to risk-based insurance pricing [4]. Goldburd *et al*. [5] detailed several types of lift curves. This work relies on the loss ratio chart.

Frees *et al*. [6] introduced the ordered Lorenz curve. Ordered Lorenz curve construction parallels lift curve construction by first ordering data instances [insurance contracts] by the relative risk of each instance as determined by a risk-based pricing algorithm. From there, both lift curves and ordered Lorenz curves seek to measure the value of the pricing algorithm based on policy premiums and claim outcomes. The parallels between lift curves and ordered Lorenz curves form the basis of OLR.

2.2. *Overfitting Mitigation Methods*

Researchers have developed a multitude of methods for mitigating overfitting. Early stopping applies to iterative fitting algorithms. The method works by tracking model performance on training data versus testing data by learning iteration [7]. The iteration on which model accuracy diverges for the training and testing data determines the optimal machine learning parameters. Penalized regression adds terms to the fitting algorithm to facilitate eliminating or shrinking regression coefficients [8]. Ridge regression applies the product of a hyperparameter and the sum of the squares of the regression coefficients to reduce the magnitude of the fitted coefficients [9]. In addition to shrinking regression coefficients, Lasso regression can drop features via a penalty defined by the sum of the absolute values of the regression coefficients [10]. Elastic nets combine ridge regression and LASSO by incorporating the penalty terms from both [9].

Pruning methods help avoid overfitting decision trees [11]. Tree ensembles, such as random forests and XGBoost, reduce overfitting by fitting and combining the results of many decision trees [12]. The dropout method for neural networks applies concepts like those used for decision trees. Dropout thins neural networks by randomly removing neurons and their connections [13]. The process is applied repeatedly, and the resulting set of thinned neural nets is then combined to produce a single prediction algorithm.

Several researchers have incorporated actuarial credibility theory into machine learning to reduce overfitting. Ohlsson and Johansson [14] introduced Buhlmann-Straub credibility into generalized linear models [GLMs] using the GLM offset term with an iterative algorithm. Klinker [15] demonstrated that linear mixed models can approximate actuarial credibility in the context of GLMs. Diao and Weng [16] built credibility into both branching and node predictions for simple decision trees.

Dimensionality reduction techniques combine or remove redundant features and include principal component analysis and various filter methods [17]. Filters are an example of feature selection, which aims to identify the most effective subset of available features [18]. Principal component analysis, like other feature extraction techniques, projects the existing features onto a lower-dimensional feature space [17].

Data augmentation approaches adjust the training data [7]. Data expansion works by adding random noise to the data set or producing additional training instances based on the distribution of the available data. Instance reduction reduces the noise in the training data by adding or removing records [19].

3. Problem Statement, Hypothesis Statements, and Research Questions

3.1. *Problem Statement*

No generally applicable method exists for mitigating the overfitting of machine learning algorithms when pricing general insurance.

3.2. *Hypothesis Statement*

Ordered Lorenz curves can be extended to develop a general method for mitigating the overfitting of machine learning algorithms when pricing general insurance.

3.3. *Research Question*

Is OLR effective at mitigating the overfitting of machine learning algorithms when pricing general insurance?

4. Methodology

We applied the design science methodology to developing and evaluating OLR. Design science studies postulate artifacts for solving problems and then test the effectiveness of those artifacts [20]. The remainder of this section provides the derivation of OLR and the process used for testing.

4.1. *Artifact Construction*

Assume a sample of N insurance contracts. Further assume that the sample is sorted, from lowest to highest, based on the output of an underlying ML algorithm with s_i being the algorithm's output for contract i where i = 1, 2, ..., N and $s_1 <= s_2 <= ...s_N$. Let c_i denote contract i in the sorted sample, p_i be the premium for c_i under the current rating plan, and l_i be the insurance losses that occurred on c_i. Then denote P = sum(p_i) and L = sum(l_i). For each contract c_i in the sorted sample, calculate its x-coordinate on the ordered Lorenz curve as sum (p_j, j <= i)/P. Similarly, calculate the y-coordinate as sum (l_j, j <= i)/L.

Let r_b be the loss ratio for the subset of contracts c_1, c_2...., c_i, which can be calculated as y × L/(x × P) where x and y are the coordinates of the ordered Lorenz curve corresponding to c_i. Similarly, let r_t be the loss ratio for the subset of contracts c_{i+1}, c_{i+2}, ... c_N, which can be calculated as $(1-y) \times L/[(1-x) \times P]$. Empirical evidence suggests r_t/r_b is constant for all points (x, y) that comprise the ordered Lorenz curve. Therefore, write (1) as:

$$[(1-y) \times L/[\,(1-x) \times P\,]]/[\,y \times L/(x \times P)] = g \qquad (1)$$

where g is a constant.

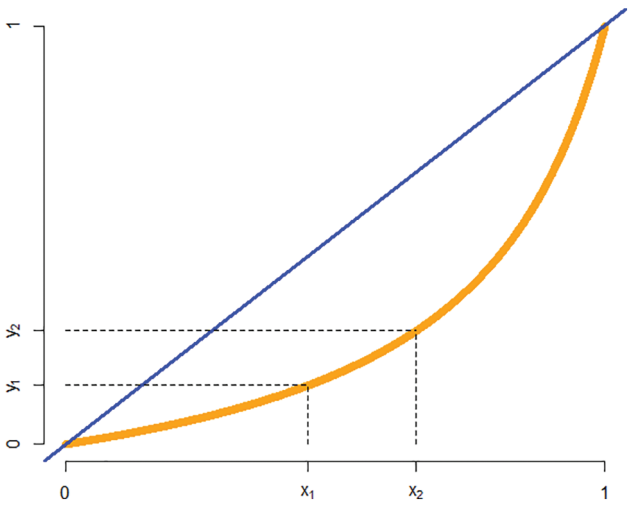

Fig. 1. Ordered Lorenz curve.

Solving (1) for y provides a parametric representation of the ordered Lorenz curve as shown in (2).

$$y = x/[g + (1 - g) \times x] \tag{2}$$

The loss ratio, R, for any subset of contracts between x_1 and x_2 with corresponding y_1 and y_2, as shown in Fig. 1, can be calculated as $(y_2 - y_1) \times L/[(x_2 - x_1) \times P]$. Substitution of (2) for y_2 and y_1 yields, after some algebra, (3).

$$R = (L/P) \times g/([g + (1 - g) \times x_2] \times [g + (1 - g) \times x_1]) \tag{3}$$

Note that (L/P) in (3) is the overall loss ratio for the sample of contracts used to construct the ordered Lorenz curve (Fig. 1).

The familiar k-bar lift curve frequently used to measure the quality of pricing models can be derived from the parameterized ordered Lorenz curve through the vectorization of (3), i.e., replace x_1 and x_2 with v_1 and v_2, respectively, where $v_1 = [0, 1/k, 2/k, \ldots, (k - 1)/k]$ and $v_2 = [1/k, 2/k, \ldots, 1]$. Dividing the heights of the lift curve bars by L/P produces risk-based pricing relativities for the customer segments corresponding to each bar. Those pricing relativities are constrained, or regularized, indirectly by the parameterized ordered Lorenz used in their derivation. The degree of pricing refinement can then be controlled via the parameter k. Alternatively, taking the limit as $x_2 - x_1$ goes to zero in (3) and dropping the overall loss ratio term, L/P, produces the regularized risk-based pricing relativity as a function of x, as stated in (4):

$$r(x) = g/[g + (1 - g) * x] \wedge 2 \tag{4}$$

Because the range for x is (0,1], (4) produces pricing relativities in the range [1/g, g].

Utilizing (4) requires determining the x and g parameters. The g parameter can be estimated by fitting (2) to the empirical ordered Lorenz curve constructed on a test data set not used in fitting the underlying ML algorithm. That approach, as shown in Fig. 2, was utilized for this study. Alternatively, k-fold cross-validation can be employed with an estimate of g computed for each fold. The individual estimates can then be combined by taking the mean or median. The x inputs required by (4) can be obtained by fitting a function x = f(s) where s represents the underlying ML algorithm's prediction and x represents the horizontal axis of the ordered Lorenz curve constructed from training, test, or cross-validation data.

4.2. *Artifact Evaluation*

We used the brvehins1 data set available from CASdatasets [21] to test OLR. The brvehins1 data set represents risk characteristics and claim outcomes for a sample of Brazilian private passenger autos during the year 2011. The data provider split brvehins1 into six subsets of 393,071 records each and labeled them as brvehins1a, brvehins1b, ..., brvehins1e. Vehicles with identical risk characteristics were aggregated by the data provider, with approximately 1,262,000 vehicle-year exposures for each data subset.

Fig. 3 provides a flow chart of our approach to testing OLR. Using subset brvehins1e, we constructed a GLM to predict auto collision loss ratios. The information content of the feature set was boosted to levels common in general insurance pricing models through two data adjustments. First, we recalculated the premium component of the loss ratios to remove the effects of embedded rating plans that had been constructed on the available feature set. Second, we introduced a synthetic feature by ordering the data via the target variable, indexing the ordered records with

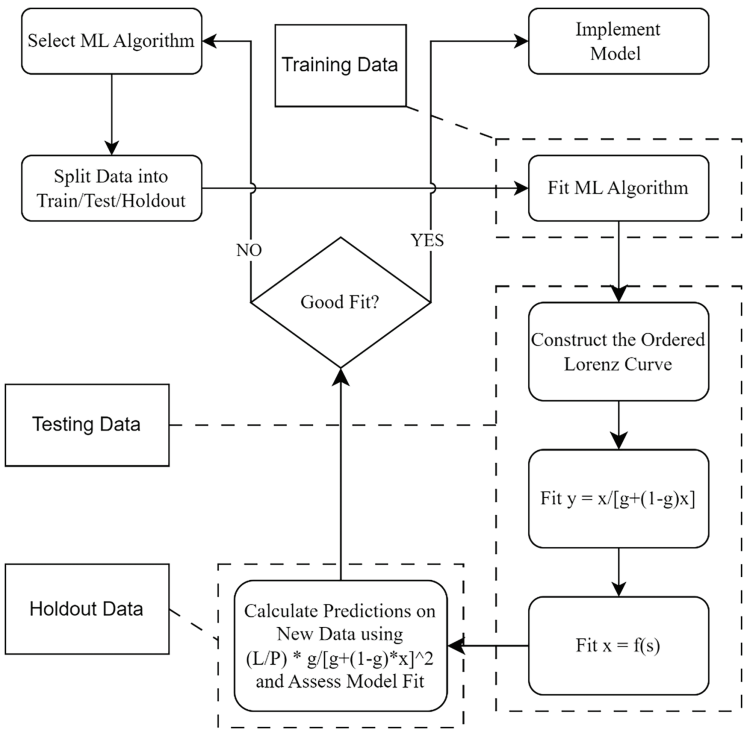

Fig. 2. Ordered Lorenz Regularization.

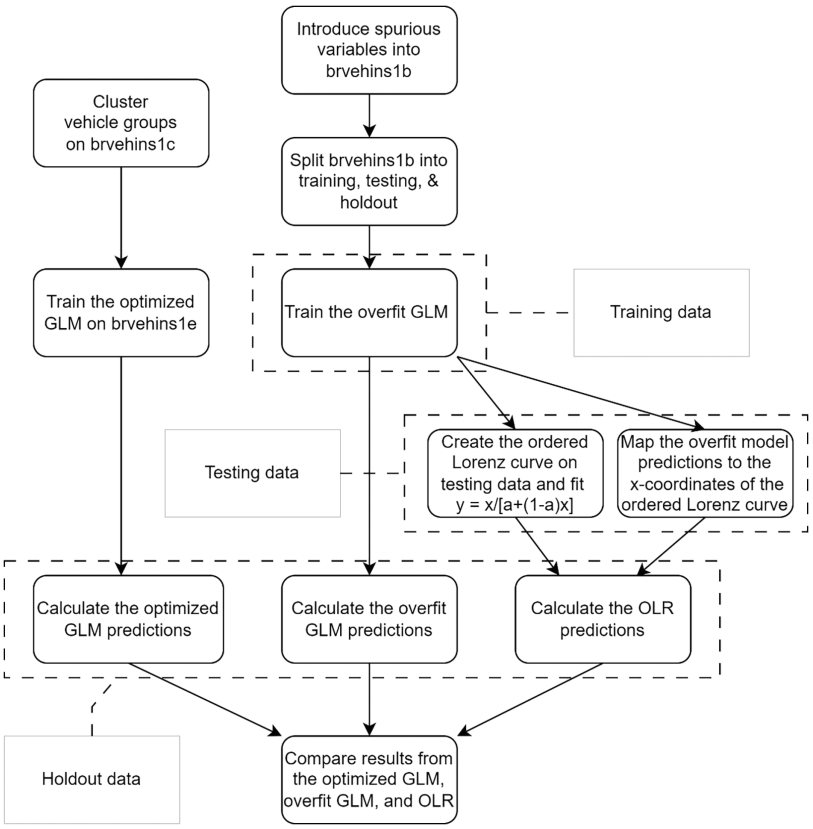

Fig. 3. OLR Testing Procedure.

TABLE I: Model Fit Statistics

Model	RMSE	MAE	Tweedie deviance
Optimzed GLM	8.46	1.17	71.59
OLR	8.48	1.20	71.88
Overfit GLM	8.50	1.21	72.30
Improvement	50.0%	25.0%	59.2%

positive integers, and then converting the index to a series of Normal random variables with means equal to the index values and standard deviations equal to the index values divided by 3.5. The GLM fit was then optimized using common statistical and actuarial methods for building general insurance GLM pricing models.

Turning to subset brvehins1b, we made the same data adjustments to boost the information content of the feature set. Then, we introduced overfitting by adding spurious features into brvehins1b and limiting the training data through a 5%/47.5%/47.5% split of brvehins1b into train/test/holdout sets. The linear predictor of the optimized GLM was expanded to include the spurious features and then the model coefficients were refit on the brvehins1b training set to obtain an overfit GLM. The application of OLR to the overfit GLM produced a prediction set that could be compared to the prediction sets of the optimized GLM and the overfit GLM.

5. Results

The root mean squared error [RMSE], mean absolute error [MAE], and Tweedie deviance statistics for the optimized GLM, overfit GLM, and OLR are displayed in Table I. The Improvement statistics in the bottom row of Table I represent the portion of the linear distances between the overfit GLM and the optimized GLM statistics that are bridged by the OLR statistics, calculated as [overfit GLM statistic–OLR statistic]/[overfit GLM statistic–optimized GLM statistic].

Fig. 4 shows scatterplots of the overfit GLM and OLR predictions versus the optimized GLM predictions. The dashed horizontal lines at 0.3 and 3.3 represent the OLR minimum and maximum, respectively. The ratio of the maximum to the minimum, 11.0, provides the relativity between the lowest insurance premium charged and the highest. The relativity seems reasonable for private passenger collision coverage where the risks are rather homogeneous and partial collision claims are common.

Nearly 95% of the optimized GLM predictions fall in the range [0, 2]. Fig. 5 focuses on the region from Fig. 4 defined by the [0, 2] range of the optimized GLM predictions. Again, the dashed horizontal lines represent the minimum and maximum imposed by OLR. Notice that all the OLR predictions in Fig. 5 are pulled below the maximum.

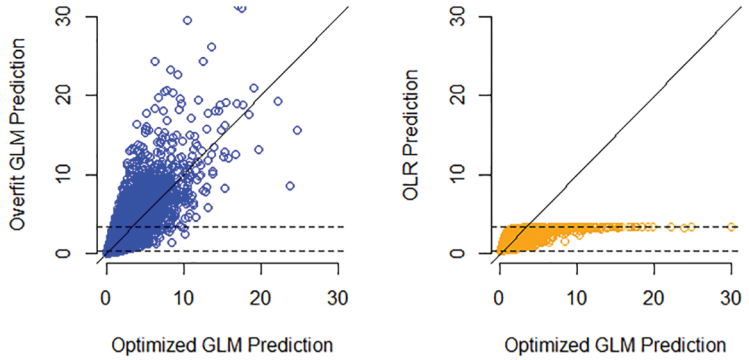

Fig. 4. Scatterplots of the effects of OLR.

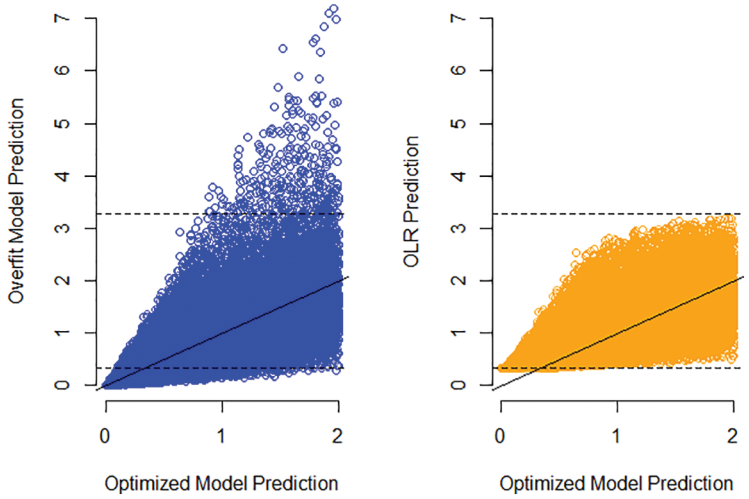

Fig. 5. Scatterplots of the effects of OLR–focused view.

6. Conclusion

Our experiment indicated that OLR is effective at mitigating the overfitting of machine learning algorithms applied to general insurance pricing. Furthermore, OLR sets upper and lower bounds on the predictions. The upper bound may alleviate the concerns of regulators, insurers, and consumers regarding the contributions of overfitting to unfair premiums and insurance affordability problems. The lower bound implies a minimum premium. Setting minimum premiums is a common general insurance practice because some level of residual risk remains even as the exposure tends to zero.

The application of OLR is independent of the underlying ML algorithm used for insurance pricing. Ordered Lorenz Regularization only requires the underlying algorithm to order insurance contracts based on risk. As such, OLR can be applied in addition to existing methods designed to mitigate overfitting. Further research might investigate the effectiveness of OLR when used in combination with such methods.

The data used for this study represents just one type of general insurance, one geography, and one period. More research is needed to determine whether OLR applies more broadly to general insurance pricing. Additionally, simulated data would offer the advantage of a known baseline against which to compare the effectiveness of OLR versus the approach of constructing an optimized model as used in this study.

Conflict of Interest

Authors declare that they do not have any conflict of interest.

References

[1] Powell L. Risk-based pricing of property and liability insurance. *J Insur Regul*. 2020;1:1–22.
[2] Brady J, Brockmeier DR. Bias-variance tradeoff: a property-casualty modeler's perspective. *Variance: Casualty Actuar Soc*. 2018;13(2):207–32.
[3] Henckaerts R, Côté M-P, Antonio K, Verbelen R. Boosting insights in insurance tariff plans with tree-based machine learning methods. *N Am Actuar J*. 2021;25(2):255–85.
[4] Tevet D. Exploring model lift: is your model worth implementing. *Actuar Rev*. 2013;40(2):10–3.
[5] Goldburd M, Khare A, Tevet D, Guller D. Generalized linear models for insurance rating. *Casualty Actuarial Society, CAS Monographs Series*, 2016;5.
[6] Frees EW, Meyers G, Cummings AD. Summarizing insurance scores using a Gini index. *J Am Stat Assoc*. 2011;106(495):1085–98.
[7] Ying X. An overview of overfitting and its solutions. *J Phys: Conf Series*. 2019;1168(2):022022.
[8] Cherlin S, Howey RA, Cordell HJ editors. Using penalized regression to predict phenotype from SNP data. *BMC Proceedings*. Springer, 2018.
[9] Lever J, Krzywinski M, Altman N. Points of significance: regularization. *Nat Methods*. 2016;13(10):803–4.
[10] Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B: Stat Methodol*. 1996;58(1):267–88.
[11] Zhou X, Yan D. Model tree pruning. *Int J Mach Learn Cybern*. 2019;10:3431–44.
[12] Mienye ID, Sun Y. A survey of ensemble learning: concepts, algorithms, applications, and prospects. *IEEE Access*. 2022;10:99129–49.

[13] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15(1):1929–58.
[14] Ohlsson E, Johansson B. Combining credibility and GLM for rating of multi-level factors. *Casualty Actuarial Society Discussion Paper Program Casualty Actuarial Society-Arlington*, Virginia, 2004, pp. 319.
[15] Klinker F editor. Generalized linear mixed models for ratemaking: a means of introducing credibility into a generalized linear model setting. *Casualty Actuarial Society E-Forum, Winter 2011 Volume 2*, 2010.
[16] Diao L, Weng C. Regression tree credibility model. *N Am Actuar J*. 2019;23(2):169–96.
[17] Salam MA, Azar AT, Elgendy MS, Fouad KM. The effect of different dimensionality reduction techniques on machine learning overfitting problem. *Int J Adv Comput Sci Appl*. 2021;12(4):641–55.
[18] Li Y, Li T, Liu H. Recent advances in feature selection and its applications. *Knowl Inf Syst*. 2017;53(3):551–77.
[19] Al-Akhras M, El Hindi K, Habib M, Shawar BA. Instance reduction for avoiding overfitting in decision trees. *J Intell Syst*. 2021;30(1):438–59.
[20] Elragal A, Haddara M. Design science research: evaluation in the lens of big data analytics. *Systems*. 2019;7(2):27.
[21] Dutang C, Charpentier A, Dutang MC. Package 'casdatasets'. 2020. Available from: https://www.openmlorg/search.