Combining Predictions of Auto Insurance Claims

Chenglong Ye, Lin Zhang, Mingxuan Han, Yanjia Yu, Bingxin Zhao, Yuhong Yang

August 29, 2018

Abstract

This paper aims at achieving better performance of prediction by combining candidate predictions, with the focus on the highly-skewed auto insurance claim cost data. We analyze a version of the Kangaroo Auto Insurance company data, and incorporate different combining methods under five measurements of prediction accuracy. The results show: 1) When there exists an outstanding prediction among the candidate predictions, the phenomenon of the "forecast combination puzzle" may not exist. The simple average method may perform much worse than the more sophisticated combining methods; 2) The choice of the prediction accuracy measure is crucial in defining the best candidate prediction for "low frequency and high severity" datasets. For example, Mean Squared Error (MSE) does not distinguish well different combining methods, the MSE values of which are very close to each other; 3) The performances of different combining methods can be drastically different. 4) Overall, the combining approach is valuable to improve the prediction accuracy for insurance claim cost.

1 Introduction

The average countrywide insurance expenditure keeps rising from year to year. Analyzing insurance data to predict future insurance claim cost is of enormous interest to the insurance industry. Particularly, accurate prediction of claim cost is fundamental to the determination of policy premiums, preventing potential loss of customers due to loss of profit or overcharge.

Non-life insurance data is distinct from common regression data due to its "low frequency and high severity" characteristic, i.e., the distribution of the claim cost is highly right-skewed and has a large point mass at zero. This paper focuses on model averaging/combining methods to improve the prediction accuracy in non-life insurance data.

Researchers have developed various methods to analyze insurance data in recent decades. In the 1960s, Bailey & Simon (1960) developed the minimum

bias procedure as an insurance pricing technique for multi-dimensional classifications. However, the minimum bias procedure lacks the statistical evaluation of the model. See Feldblum & Brosius (2003) for a detailed overview about the minimum bias procedure and its extensions. Late in the 1990s, the generalized linear models (GLM) framework (Nelder & Wedderburn 1972) was applied to model the insurance data, which is nowadays the standard method in the insurance industry for modeling the claim cost. Jørgensen & Paes De Souza (1994) proposed the classical compound Poisson-Gamma model, which assumes the number of claims to follow a Poisson distribution and be independent of the average claim cost that has a Gamma distribution. Gschlößl & Czado (2007) extended the approach and allowed for dependency between the number of claims and the claim size by a fully Bayesian approach. Smyth & Jørgensen (2002) used double generalized linear models for the case where we only observe the claim cost but not the frequency. Many authors proposed methods on insurance ratemaking under different frameworks other than GLM, including quantile regression (Heras et al. 2018), hierarchical modeling (Frees & Valdez 2008), machine learning (Kašćelan et al. 2015, Yang et al. 2016), copula model (Czado et al. 2012), and spatial model (Gschlößl & Czado 2007).

Given many useful statistical tools/models, empirical evidence has showed that model combining in general is a robust and effective way to improve predictive performance. Many works in the model combination literature have been done and succeeded to improve the prediction accuracy given different models, which can be different types of models or models of the same type but with different tuning parameters. For instance, Wolpert (1992) proposed Stacked Generalization to take prediction results from first layer base learners as meta-features to produce model-based combining forecasts in the second layer. A gradient boosting machine (Friedman 2001), known as greedy function approximation, suggests that a weighted average of many weak learners can produce an accurate prediction. Yang (2001) proposed adaptive regression by mixing, a weighted average method for regression that can work well with unknown error distribution. Hansen & Racine (2012) proposed Jackknife model averaging, a linearly weighting average of linear estimators searching the optimal weight of each base regression models. We refer readers to Wang et al. (2014) for a detailed literature review on model combining theory and methodology.

In the specific context of insurance data, however, little has been done on combining predictions, except Ohlsson (2008), Sen et al. (2018), which focused on applying model averaging methods to cluster the categories of predictors that have too many levels, such as the car brand. To our best knowledge, no previous work has been done on combining different modeling procedures to improve the prediction accuracy on highly-skewed insurance data. Given the obvious importance of accurate prediction of insurance claim cost, researches that fill in the gap are valuable.

Our paper focuses on combining multiple predictions to improve the prediction accuracy over individual model/predictions. We investigate how different model averaging methods perform under different measures of prediction accuracy for "low frequency and high severity" data.

More specifically, we try to answer several interesting questions: Do combining methods improve over the best candidate prediction for insurance data? Is the so called "forecast combination puzzle" still relevant when dealing with insurance data? Under different measurements of prediction accuracy, which combining method works the best? We carry out a real data analysis in this work. Twelve forecasters participated in building their models to predict the claim cost of each insurance policy. Based on their predictions, we apply different model averaging methods to obtain new predictions with the goal of higher prediction accuracy. Different measurements of prediction accuracy are considered due to the existence of various constraints in practice. For example, a good prediction should not only identify the most costly customer but also provide the correct scale of the claim amount. Specifically, five measurements are included in our paper: mean absolute error, root mean squared error, rebalanced root mean squared error, the relative difference between the total predicted cost and the actual total cost (named SUM), and Normalized Gini index.

The remainder of this paper begins with the general methodology in Section 2, with data summary in Section 2.1, project description in Section 2.2, and measurements of performance in Section 2.3. Section 3 describes the performances of the predictions provided by the forecasters. The results of the model combination are in Section 4. We end our paper with a discussion in Section 5.

2 General Methodology

In this section, we give a detailed description of our research methodology.

2.1 Data Summary

The dataset (De Jong et al. 2008) we used is based on one-year vehicle insurance policies written in 2004 or 2005 and called Kangaroo Auto Insurance company data. The original data set is downloadable from R package "insuranceData". We add random noise to each variable. The perturbed data is available upon request. There are 67,856 policies and ten variables in this dataset. The variable information is presented in Table 1.

Variable	Description	Variable	Description
veh_value	Vehicle value	gender	The gender of the driver
veh_body	The type of the vehicle body	area	Driver's area of residence
veh_age	The age group of the vehicle	agecat	Driver's age group
claimcst0	Total amount of the claims	exposure	The covered period
numclaims	Number of claims	**clm**	Indicator if a vehicle has at least a claim

Table 1: Variable description of Kangaroo dataset. The variables in bold are the response variables, which are directly related to the claim cost. All the other variables are treated as the predictors.

2.2   Project Description

1. (Data Process) The dataset is split into three parts following the Kaggle style: 22610 observations for **T**raining, 22629 observations for **V**alidation and 22617 observations for **H**oldout. All the variables in the training set are available for model construction while the response variables in the validation and holdout sets are hidden.

2. (Prediction) Twelve forecasters came up with their methods to predict the "total amount of claims" (*claimcst0*) for the validation and holdout sets. After the submission of their predictions, each forecaster was provided with the feedback (the measurements of prediction accuracy based on the validation set) of his/her own prediction and the best prediction. Then they adjusted their model accordingly and submitted an updated version of their predictions. We term these predictions by the forecasters as *base predictions*, or more specifically *base predictions before feedback* and *base predictions after feedback*.

3. (Model Combining) With all the twelve base predictions in step 2 as the candidate predictions for combining, we apply different model averaging methods and assess their performances on a subset of 5000 observations from the validation set to train the weights for combining.

4. (Evaluation) Finally, an overall evaluation (based on the holdout set) of the performance of all the predictions (base predictions before feedback, base predictions after feedback, predictions using model combining methods) in steps 2 and 3 is conducted.

Remark 1. It is worth pointing out that 94% of the claim costs are zeros (no claims) in the training set. We present a histogram and a Lorenz curve (Lorenz 1905) (the cumulative proportion of the claim amount) of the training set in Figure 1. For the non-zero claims, the distribution is right-skewed and heavy-tailed. There is a massive spike at 0 with frequency at 21076, which is not plotted in Subfigure 1b for space limitation.

Remark 2. In steps 2 and 3, although there are three response variables, the goal is to predict the total amount of the claims, *claimcst0*. The performances of 12 base predictions evaluated on the 5000 random observations in step 3 are similar to the performances on the holdout set.

2.3   Measurement of Prediction Accuracy

Let n be the number of policies and subscript i correspond to the i-th policy. Denote y_i as the claim cost and $\hat{y}_i$ as the predicted claim cost for the i-th policy. We consider the following five measurements of the prediction accuracy of $\{\hat{y}_i\}_{i=1}^n$.
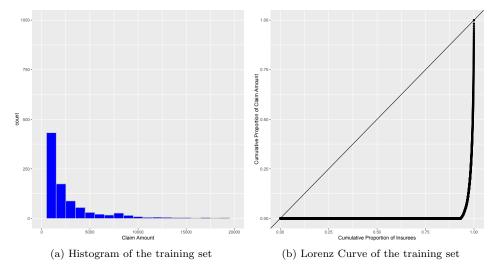
(a) Histogram of the training set   (b) Lorenz Curve of the training set

Figure 1: Data summary of the training set

Gini index   Gini index (Gini 1912) based on ordered Lorenz curve is a well-accepted tool to evaluate the performance of the predictions, but there exist many variants of Gini index. The one we utilize here is slightly different from those considered in Frees et al. (2014). For a sequence of numbers $\{s_1, ..., s_n\}$, let $\mathrm{R}(s_i) \in \{1, ..., n\}$ be the rank of s_i in the sequence in an increasing order ($\mathrm{R}(s_i) < \mathrm{R}(s_j)$ if $s_i < s_j$, given no ties exist; the tie-breaking method will be discussed later in Remark 4). Then the normalized Gini index is referred to as

$$
G = \frac{\frac{\sum_{i=1}^{n} y_i \mathrm{R}(\hat{y}_i)}{\sum_{i=1}^{n} y_i} - \sum_{i=1}^{n} \frac{n-i+1}{n}}{\frac{\sum_{i=1}^{n} y_i \mathrm{R}(y_i)}{\sum_{i=1}^{n} y_i} - \sum_{i=1}^{n} \frac{n-i+1}{n}}. \tag{1}
$$

Remark 3. With definition (1), the Gini index depends on prediction $\{\hat{y}_i\}$ only through their relative orders. With some easy algebra we have $\sum_{i=1}^{n} y_i \mathrm{R}(y_i) \geq \sum_{i=1}^{n} y_i \mathrm{R}(\hat{y}_i)$ and $\sum_{i=1}^{n} y_i \mathrm{R}(y_i) + \sum_{i=1}^{n} y_i \mathrm{R}(\hat{y}_i) \geq (n+1)\sum_{i=1}^{n} y_i$, with $\sum_{i=1}^{n} [y_i \mathrm{R}(y_i) / \sum_{i=1}^{n} y_i] - \sum_{i=1}^{n} (n-i+1)/n > 0$. Therefore, we have $-1 \leq G \leq 1$, where the equality holds at $\mathrm{R}(y_i) = \mathrm{R}(\hat{y}_i)$ or $\mathrm{R}(y_i) + \mathrm{R}(\hat{y}_i) = n+1$, respectively.

Remark 4. Unlike other measurements we consider, a prediction with larger Gini index (closer to 1) is favored. To break the ties when calculating the order, we set $R(y_i) > R(y_j)$ if $y_i = y_j$, $i < j$.

Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)
Root Mean Squared error and mean absolute error are defined as $\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}$

and $\frac{1}{n}\Sigma_{i=1}^{n}|y_i - \hat{y}_i|$ respectively.

Whatever the determination of the policy premiums is, the insurance company needs to make profits and thus cares about the difference between the total cost and the predicted total cost. Below we consider two measurements of prediction accuracy that take the overall scale of the prediction into consideration.

Rebalanced Root Mean Squared Error (Re-RMSE) Let $\lambda = \frac{\Sigma y_i}{\Sigma \hat{y}_i}$ be the scale parameter with which the scaled total predicted cost is equal to the actual total claim cost. Then the rebalanced root-mean-squared error is defined as $\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - \lambda\hat{y}_i)^2}$, i.e. the root mean squared error of the scaled prediction $\lambda\hat{y}_i$.

SUM Error Here we define (relative) SUM error as $\Sigma_{i=1}^{n}(\hat{y}_i - y_i)/\Sigma_{i=1}^{n}y_i$, the relative difference between the total predicted cost and the actual total cost. SUM error is a way to measure the deviance of the predicted claim cost from the actual claim cost.

3 Performances of base predictions

The 12 base predictions can be categorized into two types. One type was based on separate predictions of the number of claims (frequency) and the claim cost (severity). This method typically generates predictions with zero values. The other type is to directly predict the claim cost, which typically produces many small non-zero valued predictions of the claim cost. Four out of the 12 base predictions belong to the first type (separate predictions).

Tables 2 and 3 show the performances of the 12 base predictions before and after feedback. Some forecasters performed differently before and after feedback. For example, forecaster 8 (A8) has really inferior performance in Gini index (-0.17) before feedback, but improved it to 0.235 with the feedback. No prediction outperformed all its competitors in every measurement of prediction accuracy. For instance, A5 has the largest/worst RMSE among all the predictions while its Gini index is the largest/best, with other Gini indices no more than 0.24. We also provide the estimated standard error of MAE, RMSE and Re-RMSE to assist the understanding of their reliability respectively. The MAE values of predictions are closely related to SUM. Since the response $\{y_i\}_{i=1}^{n}$ contains too many zeros, a prediction $\{\hat{y}_i\}_{i=1}^{n}$ will have relatively small MAE if $\max\{y_i\}$ is small, such as A1 with SUM around -1. For the SUM error, most predictions have negative values except A5. Specifically, the SUM errors of A1 and A2 almost reach -1. Not surprisingly, we look at each predicted value of A1 and A2, and find that all the predicted values are less than 10. It means the corresponding predictions have a very small scale so that they are unreasonable even with their acceptable performance on MAE and Gini. This is also the reason why we consider more than one measurement of prediction accuracy.

Forecasters	MAE	RMSE	Re_RMSE	Gini	SUM
A1	**149.93**(7.49)	1136.00(65.71)	1125.29(65.53)	0.2033	-1.00
A2	154.08(7.48)	1135.36(65.72)	1125.54(65.45)	0.2097	-0.97
A3	271.00(7.26)	1125.42(65.55)	1125.37(65.51)	0.1678	**-0.05**
A4	269.81(7.26)	1125.30(65.43)	1125.36(65.38)	0.2113	**-0.05**
A5	203.43(8.39)	1278.72(59.85)	1160.99(59.16)	**0.9555**	0.27
A6	270.39(7.26)	1125.18(65.29)	1125.15(65.22)	0.1473	-0.07
A7	270.11(7.41)	1132.11(65.70)	1139.20(64.92)	-0.0054	-0.62
A8	267.72(7.54)	1150.19(64.41)	1309.15(59.39)	-0.1710	-0.67
A9	268.75(7.26)	**1124.43**(65.44)	**1124.44**(65.38)	0.2309	**-0.05**
A10	254.64(7.35)	1137.27(64.57)	1139.17(64.40)	0.0882	-0.07
A11	270.07(7.26)	1124.87(65.37)	1124.88(65.31)	0.2132	**-0.05**
A12	205.93(7.55)	1156.21(64.33)	1222.92(61.57)	0.1502	-0.46
Best_base	149.93(7.49)	1124.43(65.44)	1124.44(65.38)	0.9555	-0.05

Table 2: Performance of base predictions before feedback

Forecasters	MAE	RMSE	Re_RMSE	Gini	SUM
A1	**149.93**(7.49)	1136.00(65.71)	1125.41(65.57)	0.1956	-1.00
A2	154.08(7.48)	1135.36(65.72)	1125.54(65.45)	0.2092	-0.97
A3	271.00(7.26)	1125.42(65.55)	1125.37(65.51)	0.1678	**-0.05**
A4	269.81(7.26)	1125.23(65.46)	1125.25(65.41)	0.1942	**-0.05**
A5	203.43(8.35)	1271.88(57.76)	1156.88(58.40)	**0.9553**	0.27
A6	270.39(7.27)	1125.55(65.20)	1125.59(65.12)	0.1328	-0.07
A7	270.11(7.26)	1125.29(65.33)	1125.37(65.27)	0.2163	**-0.05**
A8	267.72(7.26)	1124.76(65.46)	1124.69(65.40)	0.2350	-0.07
A9	268.75(7.26)	**1124.43**(65.44)	**1124.44**(65.38)	0.2309	**-0.05**
A10	254.64(7.30)	1126.36(65.59)	1125.99(65.45)	0.1354	-0.19
A11	270.07(7.26)	1124.87(65.37)	1124.88(65.31)	0.2132	**-0.05**
A12	205.93(7.38)	1129.29(65.78)	1129.91(65.36)	0.1510	-0.55
Best_base	149.93(7.49)	1124.43(65.44)	1124.44(65.38)	0.9553	-0.05

Table 3: Performance of base predictions after feedback

4   Model Combining

Many modern model combining methods adopt a linear combination of the candidate models. Usually, model combining has two goals. Following the terms in (Yang 2004, Wang et al. 2014), they are (i) combining for improvement and (ii) combing for adaptation. By model combining for improvement, we hope to combine the base models to exceed the prediction performance of any base model. On the other hand, model combining for adaptation targets at capturing the best model (usually unknown) among the base models. In this paper, both goals are of interest.

Let $\mathbf{y} = \{y_i\}_{i \in Holdout}$ denote the response vector for the holdout set. Denote $\mathbf{f} = (\mathbf{f}_1, ..., \mathbf{f}_{12})$ as the matrix with each column representing a base prediction for the holdout set. Let $\mathbf{f}_c$ denote the combined prediction. We consider two types of model combining methods: (i) combining the 12 base predictions and (ii) combining all the subsets of the 12 base predictions. More specifically, for the first type, $\mathbf{f}_c = \sum_{i=1}^{12} \theta_i \mathbf{f}_i$, with the weight $\theta = (\theta_1, ..., \theta_{12})^T$ obtained by different combining methods. For the second type, we first fit a linear regression of $\mathbf{y}$ on a subset, say m_i, of the 12 base predictions. Let $\hat{\mathbf{y}}_{m_i}$ be the corresponding prediction of $\mathbf{y}$ on the subset m_i, where we have $2^{12} = 4096$ possible choices of subsets, i.e., $i = 1, ..., 4096$. We then combine all such predictions, i.e., $\mathbf{f}_c = \sum_{i=1}^{4096} w_i \hat{\mathbf{y}}_{m_i}$, where $w = (w_1, ..., w_{4096})$ is the corresponding vector of weights.

4.1 Combining the 12 base predictions

Simple Average:

1. Simple Average (SA): $\theta_i \equiv \frac{1}{12}$, $\forall i = 1, ..., 12$.

2. Simple Average without A5 (SA$_{(-5)}$): $\theta_5 = 0$, $\theta_i \equiv \frac{1}{11}$, $\forall i \neq 5$.

Remark 5. Simple average is easy to apply and the most basic procedure in model combining. Although it is simple, more often than not, simple average has a better or similar performance than other complicated methods, which is known as the "forecast combination puzzle". However, we are curious about its performance in our case where a dominant prediction exists among the base predictions. Also, we consider the simple average among all the base predictions except the dominant one (A5).

Linear Regression   Treat the base predictions $\mathbf{f} = (f_1, ..., f_{12})$ as the regressors and y as the response, we conduct a linear regression. The estimated coefficients become the corresponding weights for the combination of predictions.

1. LR-AIC: Fit a linear regression model of $\mathbf{y}$ on each subset of the 12 base predictions. Pick the model with the smallest AIC and use its corresponding estimated coefficients as the weights.

2. Data-driven Linear Regression (LR-D): Choose all the significant predictors at significance level $\alpha = 0.05$ after fitting the full model. Then fit the smaller model only with selected predictors and the estimated coefficients would be the weights.

3. Constrained Linear Regression (LR-C): Fit a linear regression with y on $\mathbf{f}$ with the constraints that all the coefficients are non-negative and have a sum of 1.

Quantile Regression (QR) and Gradient Boosting (GB): Fit a quantile regression model and a gradient boosting regression model respectively with 12 predictions as features and y as the response. Then the estimated coefficients will be the weights.

Remark 6. The quantile regression predicts the median (when quantile is equal to 0.5) of the response rather than the mean of the response. In this case, we also use the estimated coefficients as the weights in combination. The reason why we consider quantile regression is that quantile regression does not require the assumption of normality for error distribution. Also, quantile regression is more robust to outliers, since significant changing of specific values does not influence the median.

Adaptive Regression by Mixing (ARM): Adaptive regression by mixing is a combining strategy by data splitting and cross-assessment, which is proposed by Yang (2001). The weighting by ARM is proved to capture the best rate of convergence among the candidate procedures for regression estimation. It is combining for adaptation, hence we call this method ARM-A, which combines general regression procedures with 12 predictions as the features and y as the response to obtain the weight.

Remark 7. This method combines general candidate regression procedures. The advantage is that under mild conditions, the resulting estimator is theoretically shown to perform optimally in rates of convergence without knowing which of the candidate methods works the best. Also, simulation results show that ARM works better than AIC and BIC when the variance is not very small.

4.2 Combining based on all the subsets of 12 base predictions

In this scenario, two methods are considered:

1. Simple Average for all subsets (SA-S): $w_1 = ... = w_{4096} = \frac{1}{4096}$

2. ARM-I: We use the ARM algorithm to combine the 4096 predictions for **I**mproving the prediction accuracy, hence named ARM-I.

4.3 Performance of combining

Tables 4 and 5 summarize the performance of the combined predictions under five measurements of prediction accuracy. Among all the combining methods, QR performs well in all measures but SUM. Note that Gini is only related to the order of predictions, but SUM concerns more about the scale of the total cost of the claims. On the other hand, ARM_A and ARM_I overall have small SUM errors while performing reasonably well in Gini index. ARM_I has better performance than ARM_A concerning all the measurements. Similarly, average weighting based on all the subsets SA_a overall performs better than SA.

One reason is that combining through all the subsets receives more information than combining through the twelve base predictions. However, it is more time-consuming for combining through all subsets. When the number of base predictions is large, it may even be computationally infeasible. One should take consideration the practical cost when conducting combining methods based on all the subsets. On the other hand, we may pay much higher price in including all the subsets than just consider the candidate predictions, under which case combining all the subsets may not have an advantage.

Given a specific measurement of prediction accuracy, when there is a dominant base prediction, such as A5 with respect to Gini index, it is hard to achieve the goal of combining for improvement. For measurements where there is no dominant base prediction, such as MAE, RMSE, Re-RMSE and SUM, we can improve the performance through combining methods. Specifically, for MAE and RMSE, we have approximately 10% relative improvement (from best base to combining best). For Re-RMSE and SUM, the improvement is 25% and 15% respectively. Unfortunately, we have almost zero improvements in Gini index although the Gini index for quantile regression is a little larger than the best base before feedback.

We also conducted a paired two-sided two-sample t-test comparing each combining method and the best base prediction under the MAE, RMSE and Re-RMSE measures. A superscript $*$ in the tables shows that the result of the t-test is significant and the performance of combined prediction is better than that of the best individual. The significance of the t-tests further verifies the statistically significant improvement of prediction performance by the combining methods.

If we compare the two tables, we can see that although some forecasters change their predictions significantly, the combining results are not much affected. This is because most of the base predictions (more importantly the base predictions with better predictive performance) have little change after feedback. Most importantly, those base predictions with better performances have not changed too much after feedback. From this perspective, compared to the base predictions, model combining methods are more stable than a single method in the predictive modeling.

Predictions	MAE	RMSE	Re-RMSE	Gini	SUM
Best_base	149.93(7.49)	1124.43(65.44)	1124.44(65.38)	0.95547	-0.05
LR-D	206.18(6.91)	1059.43*(63.83)	1059.70*(63.85)	0.95436	-0.009
LR-AIC	210.61(6.91)	1059.95*(63.69)	1060.11*(63.74)	0.94021	**-0.005**
LR-C	210.04(6.90)	1059.01(63.75)	1058.57(63.85)	0.95518	0.016
GB	138.06*(7.33)	1110.74*(66.03)	**999.79***(66.03)	0.94489	-0.903
SA	213.77(7.16)	1098.26*(65.88)	1087.16*(65.89)	0.85341	-0.315
SA$_{(-5)}$	229.87(7.33)	1126.77*(65.61)	1126.08(65.29)	0.20382	-0.286
SA-S	230.13(6.94)	1068.85*(65.64)	1068.20*(65.61)	0.92143	-0.020
QR	**135.19***(7.08)	1073.91*(65.85)	1160.99(65.97)	**0.95554**	-0.730
ARM_A	234.18(7.01)	1071.16*(65.64)	1070.862*(65.62)	0.92168	-0.009
ARM_I	208.99(6.90)	**1058.76***(63.93)	1058.98*(63.87)	0.94824	-0.009

Table 4: Performance of the combined predictions before feedback. The first row shows the best base prediction on different measurements of prediction accuracy. The highlighted values in each column indicate the best combining method for each measure among all the combining methods we conduct.

Predictions	MAE	RMSE	Re_RMSE	Gini	SUM
Best_base	149.93(7.49)	1124.43(65.44)	1124.44(65.38)	0.95529	-0.05
LR-D	206.10(6.91)	1059.02*(63.85)	1059.21*(63.81)	0.95460	**-0.007**
LR-AIC	210.22(6.90)	1059.17*(63.74)	1058.93*(63.79)	0.95328	0.009
LR-C	209.86(6.90)	1058.59(63.77)	1058.15(63.89)	0.95490	0.018
GB	138.08*(7.33)	1111.24*(66.03)	**999.69***(62.40)	0.94249	-0.906
SA	224.30(7.14)	1097.56*(65.89)	1089.69*(65.74)	0.86502	-0.236
SA$_{(-5)}$	241.01(7.31)	1125.64(65.62)	1124.67(65.40)	0.23067	-0.189
SA-S	229.26(6.94)	1068.99*(65.68)	1068.18*(65.64)	0.91735	-0.024
QR	**135.05***(7.12)	1078.95*(65.97)	1156.88(58.40)	**0.95529**	-0.762
ARM_A	234.05(6.97)	1068.01*(63.84)	1068.45*(63.82)	0.93023	0.013
ARM_I	209.04(6.90)	**1057.59***(64.11)	1057.79*(64.04)	0.95202	-0.012

Table 5: Performance of the combined predictions after feedback

5   Conclusion and Discussion

We start this section with answering the questions raised in the introduction.

Do combining methods improve over the best individual prediction when there is a dominant candidate prediction? From the results of our analysis, it is hard to achieve the goal "combining for improvement" when there is a dominant candidate prediction. One reason may be that the predictive power of the dominant prediction is weakened by general combining methods. But it does not exclude the possibility that some combining methods unknown to us can improve the predictive performance over the best candidate prediction.

Does "forecast combination puzzle" still exist in our project for

insurance data? We conclude that simple average is not as competitive as other combining methods. Specifically, the Gini index of SA is the smallest compared to other combining methods in our results. The set of base predictions is of great importance when considering simple averaging. When there exists a dominant prediction for a particular measure, such as Gini index in our data analysis, simply averaging all the base predictions may even lead to the deterioration of the performance. In that case, we need a combining method that learns better from the data.

Under different measurements of prediction accuracy, which combining method works the best? When researchers and insurance companies are concerned with different aspects of a prediction, their preferences differ accordingly. For the criteria we took into consideration, most combining methods improve the performance of the best base prediction. For instance, after feedback, gradient boosting method reduces the most of the Re-RMSE; ARM_I has the smallest RMSE, and QR defeats all other methods on MAE. The measurement is crucial in the nonconventional data analysis such as the insurance data. If it is difficult to determine one single measurement, we should at least check different measurements of prediction error that are of interest in practice.

Insurance data is nonconventional and little work has been done for the current model combining literature on such data. A well-known advantage of model combining is its stability. As mentioned in Section 4.3, though some base predictions have changed significantly after feedback, the performance of model combining does not alter much. From the perspective of the Gini index, the dominant one does not change much after feedback. Most weighting methods assign higher weights to this particular prediction A5. As long as A5 stays unchanged (or similar), the performance of combining method in Gini index will stay similar.

In our data analysis, the details of the twelve base models are unknown. If two models are built based on the base method but with different parameters, which may lead to high correlation between the two predictions. It is also of potential interest to study whether the details of the models will improve the performance of combining methods. Additionally, a model combining method that assigns weights according to a specific measurements (a reasonable measurement with respect to the data type) of prediction error is worth investigation.

References

Bailey, R. A. & Simon, L. J. (1960), 'Two studies in automobile insurance ratemaking', *ASTIN Bulletin: The Journal of the IAA* **1**(4), 192–217.

Czado, C., Kastenmeier, R., Brechmann, E. C. & Min, A. (2012), 'A mixed copula model for insurance claims and claim sizes', *Scandinavian Actuarial Journal* **2012**(4), 278–305.

De Jong, P., Heller, G. Z. et al. (2008), *Generalized linear models for insurance data*, Vol. 10, Cambridge University Press Cambridge.

Dimakos, X. K. & Di Rattalma, A. F. (2002), 'Bayesian premium rating with latent structure', *Scandinavian Actuarial Journal* **2002**(3), 162–184.

Dunn, P. (2001), 'Likelihood-based inference for tweedie exponential dispersion models', *Unpublished PhD Thesis, University of Queensland* .

Dunn, P. K. & Smyth, G. K. (2005), 'Series evaluation of tweedie exponential dispersion model densities', *Statistics and Computing* **15**(4), 267–280.

Dunn, P. K. & Smyth, G. K. (2008), 'Evaluation of tweedie exponential dispersion model densities by fourier inversion', *Statistics and Computing* **18**(1), 73–86.

Feldblum, S. & Brosius, J. E. (2003), The minimum bias procedure: A practitioner's guide, *in* 'Proceedings of the Casualty Actuarial Society', Vol. 90, pp. 196–273.

Frees, E. W. J., Meyers, G. & Cummings, A. D. (2014), 'Insurance ratemaking and a gini index', *Journal of Risk and Insurance* **81**(2), 335–366.

Frees, E. W. & Valdez, E. A. (2008), 'Hierarchical insurance claims modeling', *Journal of the American Statistical Association* **103**(484), 1457–1469.

Friedman, J. H. (2001), 'Greedy function approximation: a gradient boosting machine', *Annals of statistics* pp. 1189–1232.

Gini, C. (1912), 'Variabilità e mutabilità: Contributo allo studio delle distribuzioni e delle relazioni statistiche: Tipogr. di p'.

Gschlößl, S. & Czado, C. (2007), 'Spatial modelling of claim frequency and claim size in non-life insurance', *Scandinavian Actuarial Journal* **2007**(3), 202–225.

Haberman, S. & Renshaw, A. E. (1996), 'Generalized linear models and actuarial science', *The Statistician* pp. 407–436.

Hansen, B. E. (2008), 'Least-squares forecast averaging', *Journal of Econometrics* **146**(2), 342–350.

Hansen, B. E. & Racine, J. S. (2012), 'Jackknife model averaging', *Journal of Econometrics* **167**(1), 38–46.

Heras, A., Moreno, I. & Vilar-Zanón, J. L. (2018), 'An application of two-stage quantile regression to insurance ratemaking', *Scandinavian Actuarial Journal* pp. 1–17.

Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. (1999), 'Bayesian model averaging: a tutorial', *Statistical science* pp. 382–401.

Jørgensen, B. & Paes De Souza, M. C. (1994), 'Fitting tweedie's compound poisson model to insurance claims data', *Scandinavian Actuarial Journal* **1994**(1), 69–93.

Kašćelan, V., Kašćelan, L. & Novović Burić, M. (2015), 'A nonparametric data mining approach for risk prediction in car insurance: a case study from the Montenegrin market', *Economic Research-Ekonomska Istraživanja* **29**(1), 545–558.

Lorenz, M. O. (1905), 'Methods of measuring the concentration of wealth', *Publications of the American statistical association* **9**(70), 209–219.

Nelder, J. A. & Wedderburn, R. W. M. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society. Series A (General)* **135**(3), 370–384.
URL: *http://www.jstor.org/stable/2344614*

Ohlsson, E. (2008), 'Combining generalized linear models and credibility models in practice', *Scandinavian Actuarial Journal* **2008**(4), 301–314.

Phua, C., Alahakoon, D. & Lee, V. (2004), 'Minority report in fraud detection: classification of skewed data', *Acm sigkdd explorations newsletter* **6**(1), 50–59.

Sakamoto, Y., Ishiguro, M. & Kitagawa, G. (1986), 'Akaike information criterion statistics', *Dordrecht, The Netherlands: D. Reidel* p. 81.

Sen, H., Adrian, O. & Brendan, M. T. (2018), 'Motor insurance claim modelling with factor collapsing and bayesian model averaging', *Stat* **7**(1), e180. e180 sta4.180.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.180*

Smyth, G. K. & Jørgensen, B. (2002), 'Fitting tweedie's compound poisson model to insurance claims data: dispersion modelling', *ASTIN Bulletin: The Journal of the IAA* **32**(1), 143–157.

Timmermann, A. (2006), 'Forecast combinations', *Handbook of economic forecasting* **1**, 135–196.

Wang, Z., Paterlini, S., Gao, F. & Yang, Y. (2014), 'Adaptive minimax regression estimation over sparse ℓ_q-hulls', *Journal of Machine Learning Research* **15**, 1675–1711.
URL: *http://jmlr.org/papers/v15/wang14b.html*

Wolpert, D. H. (1992), 'Stacked generalization', *Neural networks* **5**(2), 241–259.

Yang, Y. (2001), 'Adaptive regression by mixing', *Journal of the American Statistical Association* **96**(454), 574–588.

Yang, Y. (2004), 'Combining forecasting procedures: some theoretical results', *Econometric Theory* **20**(1), 176–222.

Yang, Y., Qian, W. & Zou, H. (2016), 'Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models', *Journal of Business & Economic Statistics* **43**, 1–45.

Yuan, Z. & Yang, Y. (2005), 'Combining linear regression models: When and how?', *Journal of the American Statistical Association* **100**(472), 1202–1214.

Zou, H. & Yang, Y. (2004), 'Combining time series models for forecasting', *International journal of Forecasting* **20**(1), 69–84.