SAS® GLOBAL FORUM 2021

Paper 1020-2021

Reliability of individual predictions estimation for claim risk scoring models

Yuriy Chechulin, Jina Qu, Hudson Mahboubi, Terrance D'souza
Workplace Safety and Insurance Board (WSIB)

ABSTRACT

The Workplace Safety and Insurance Board (WSIB) is an independent agency that administers compensation and no-fault insurance for workplaces in Ontario, Canada. The WSIB develops claim risk scoring models using SAS® to identify high-risk claims. Standard methods for assessing predictive model accuracy estimate an overall predictive power of the model (i.e. accuracy, C-statistic) averaged across all test observations. Even if the model is very accurate, on average, across all observations, it does not mean that individual predictions have the same reliability across the data. If an individual prediction for a person who has been injured at work has low reliability, the default action suggested by the model (i.e. triage into high versus low-risk group) could be ignored and additional human intervention can be taken. We use a model-independent approach to estimate the reliability of individual predictions for two example models: Claim Segmentation (binary outcome model, estimated using Random Forest) and Case Based Reserve (discrete time survival analysis model). The explored approach is remarkably accurate at separating reliable from unreliable individual predictions based on Receiver Operating Characteristic (ROC) curves, C-statistic and calibration curves calculated on the test (unseen) data. Identification of low reliability for individual predictions for claim risk scoring models may help to avoid automatic action based on artificial intelligence (AI) prediction for unreliable predictions. It can then trigger an in-depth human intervention for additional data collection and the development of an appropriate return-to-work plan.

INTRODUCTION

Standard methods for assessing a predictive model's accuracy estimate an overall predictive power of a given model (such as accuracy, sensitivity, specificity, C-statistic, mean squared error [MSE], etc.) averaged across all test observations. Although these estimates evaluate model performance by summarizing the error contributions of all test examples, they provide no local information about the expected error of an individual prediction for a given unseen example [3]. Even if a given model is very accurate, on average, across all observations, it does not mean that individual predictions have the same reliability across the data. We can broadly define reliability as correctness of each individual prediction (i.e. is this particular individual prediction indeed a true positive or a true negative outcome). Information on reliability of individual predictions is very important in high-risk settings, such as health care. Individual prediction reliability estimates can be directly shown to users to help them gauge whether they should trust the AI system. This is crucial when a model's prediction influences important decisions such as a medical diagnosis [5]. While a given stroke risk model can be very accurate on average, physicians need to know whether or not to trust a given prediction for a particular patient. If individual prediction reliability is low, a physician may use other available means to create a care plan for the patient. The same is

true in worker compensation settings: if an individual prediction for a person who has been injured at work has low reliability, the default action suggested by a given predictive model (for example, triage into high-risk versus low-risk group) can be ignored and additional human intervention can be taken to collect additional data and used to develop an appropriate return-to-work plan. A method is needed that identifies beforehand when an injured person belongs to a subgroup where the predictive model in question has reduced performance. We can define predictions for patients who belong to these poorly performing subgroups as unreliable because they correspond to misleading statements about a given patient's risk [1].

In general, there are two broad classes of approaches to estimate the reliability of individual predictions: model-dependent and model-independent [1, 2, 3, 4, 5]. Model-dependent methods generally report prediction confidence intervals that are calculated using least squares estimation or by estimating the uncertainty in learned model parameters. The drawback of these approaches is they mandate the use of a particular type of classifier. Model-independent approaches can be used with a variety of different predictive models, irrespective of the approach used to develop or train the model. Most model-independent approaches involve retraining the predictive model using an enhanced data set that contains the original training set supplemented with new, unclassified data examples, where class labels for the unlabeled data are assigned based on the model's predictions. The model's performance before and after retraining are used to estimate the reliability of the predicted classes for the new data [1]. The resulting "unreliability score" can be computed for any risk model [1].

METHODS

In this analysis we are testing the reliability of estimation for individual predictions for two example predictive models:

- Example 1 – Claim Segmentation model predicts probability of being on loss of earnings (LOE) benefits at three month duration from accident date using limited information only available at time of claim registration (binary outcome model, estimated using Random Forest in SAS Enterprise Miner).
- Example 2 – Case Based Reserve model predicts expected duration on LOE benefits up to 12 month duration (restricted mean survival time [RMST] discrete time survival analysis model, estimated using logistic regression in SAS Enterprise Guide).
    - For the purposes of discrete time survival analysis model, we use survival probability at a 12 month duration versus an observed 12 month duration outcome effectively converting the survival model into a binary prediction problem to simplify and make the assessment of this example with other studied examples consistent.

We use a model independent approach to estimate the reliability of individual predictions. This approach involves retraining the predictive model using an enhanced data set that contains the original training set supplemented with new, unclassified data examples, where class labels for the unlabeled data are assigned based on the initial model's predictions. It is then evaluated using the difference between the two estimated probabilities for each individual prediction. We use 60 per cent of the data in the original data set and we add 40 per cent additional data for the enhanced data set. We calculate then the unreliability metric as $U(x) = |\hat{y}1 - \hat{y}2|$ for each individual prediction.

Evaluation of the unreliability metric is done on test (unseen) data. We use sampling that is stratified by duration outcome flag to create training and test data sets and preserve the same prevalence of the duration outcome as the original data in training and test samples. We build plots of the unreliability metric versus predicted probability from the main (full) model (100 per cent of the test data), ROC curves for the main (full) model and ROC curves for subsets of the test data labelled as reliable predictions and unreliable predictions (using the arbitrary cutoff of the top third percentile of the unreliable metric distribution to define an unreliability flag), calculate corresponding area under ROC curve (C-statistic), as well as build calibration (goodness-of-fit) curves for both reliable and unreliable subsets. We repeat this exercise for both examples.

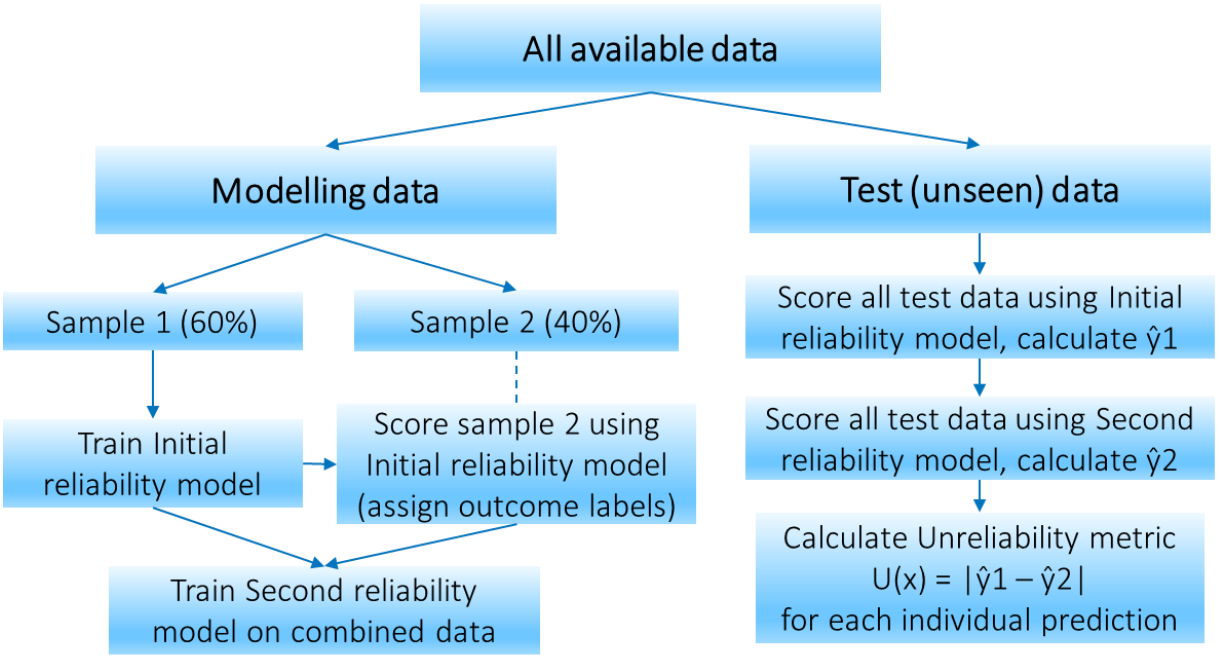


Figure 1. Methods diagram

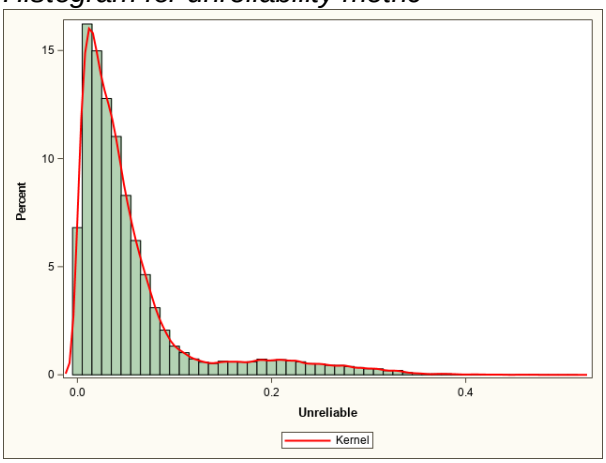
Methods diagram in Figure 1 shows schematic process of deriving unreliability metrics for each individual prediction.

RESULTS

EXAMPLE 1

Example 1 is a claim segmentation model predicting probability of being on loss of earnings (LOE) benefits at a three month duration from the accident date using limited information only available at time of claim registration (binary outcome model, estimated using Random Forest).

Histogram for unreliability metric



Unreliability metric vs. probability scatter plot

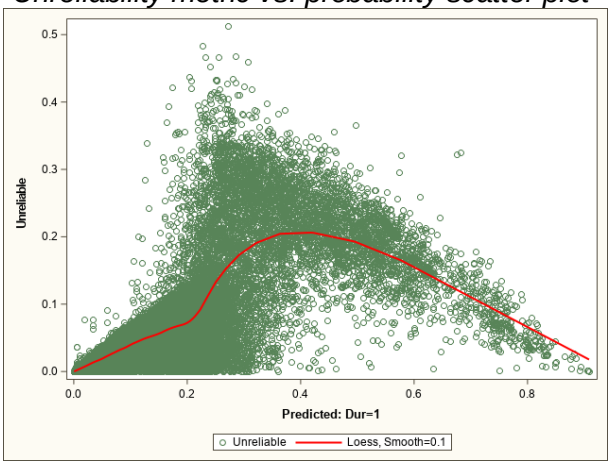
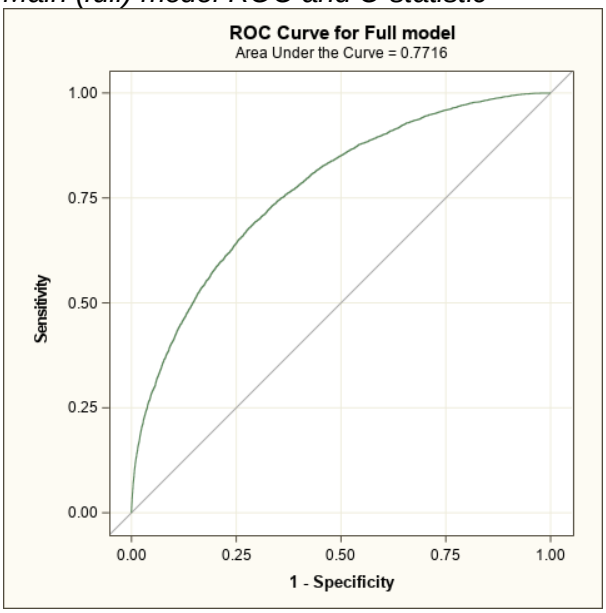


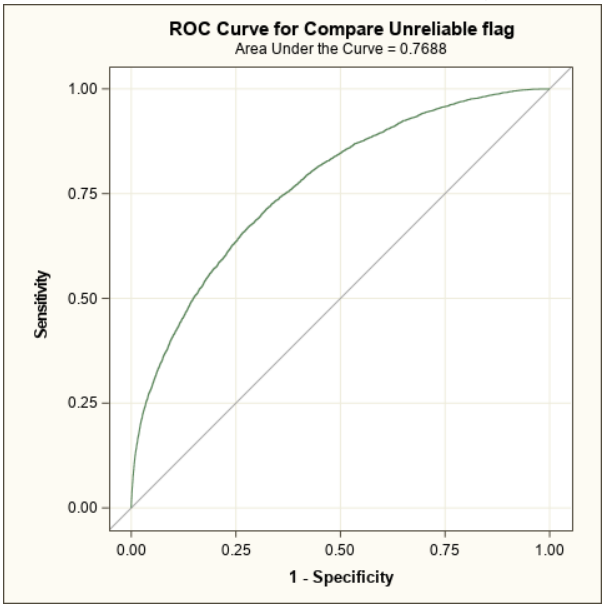
Figure 2. Histogram for unreliability metric and Scatter plot of unreliability metric vs. probability (Example 1)

Figure 2 shows that the unreliability metric distribution is moderately skewed and the range of values is not large. The scatter plot of the unreliability metric versus estimated probability from the main (full) model reveals that unreliable observations are concentrated mostly around uncertain probability region (in the middle of probability range).

Main (full) model ROC and C-statistic



ROC and C-statistic for unreliable flag = 0



ROC and C-statistic for unreliable flag = 1

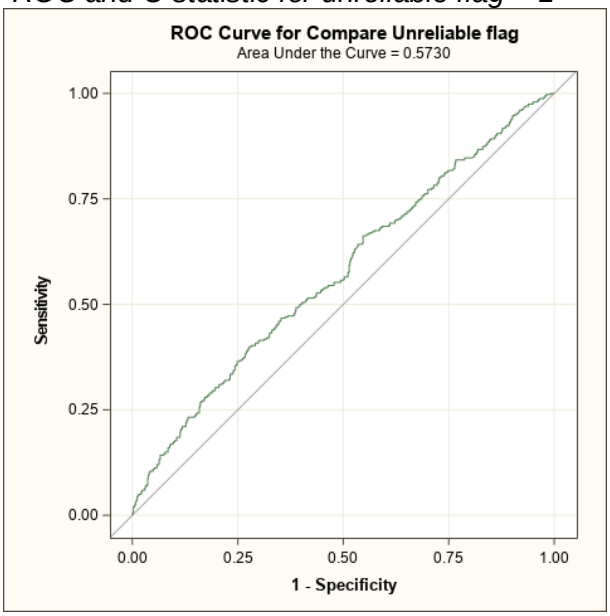
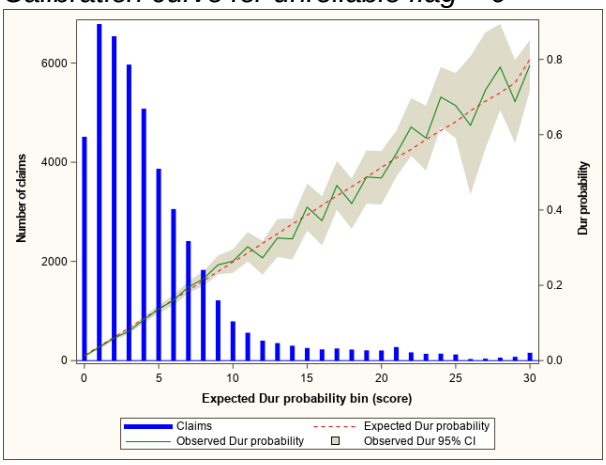


Figure 3. ROC and C-statistic (area under the curve) for Main (full) model and Unreliable and Reliable predictions (Example 1)

Figure 3 shows that the ROC curve and C-statistic for reliable observations (unreliable flag = 0) are very close to that of the main (full model). We can see substantial deterioration in ROC and C-statistic for observations identified as unreliable (unreliable flag = 1), which is very close to random choice (to C-statistic 0.5).

Calibration curve for unreliable flag = 0



Calibration curve for unreliable flag = 1

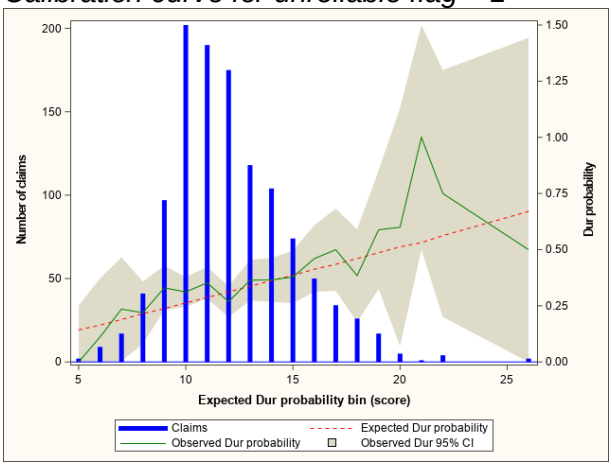


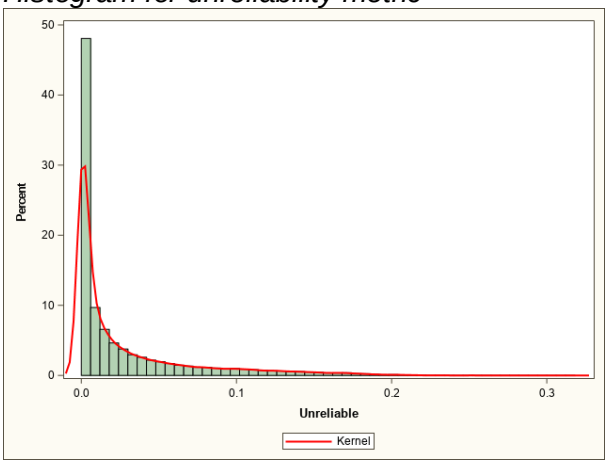
Figure 4. Calibration curves for unreliable and reliable predictions (example 1)

In Figure 4, we see that a subset of data identified as reliable observations shows good calibration (the observed probability line follows the expected probability line closely). The unreliable observations are not well calibrated (the two lines depart quite a bit).

EXAMPLE 2

Example 2 is a Case Based Reserve model predicting expected duration on LOE benefits up to 12 month duration (RMST discrete time survival analysis model).

Histogram for unreliability metric



Unreliability metric vs. probability scatter plot

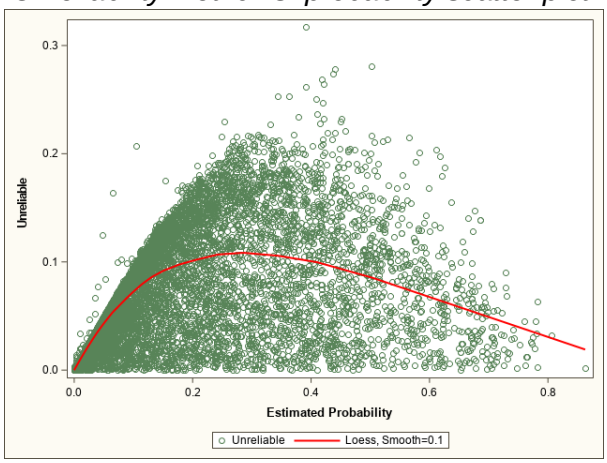
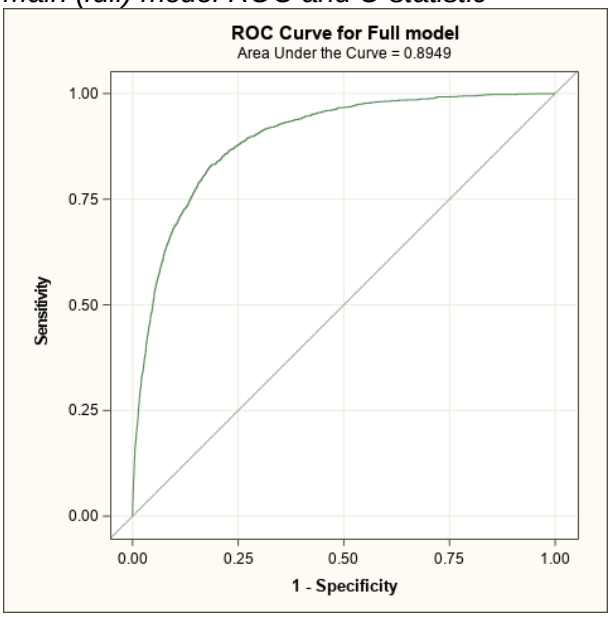


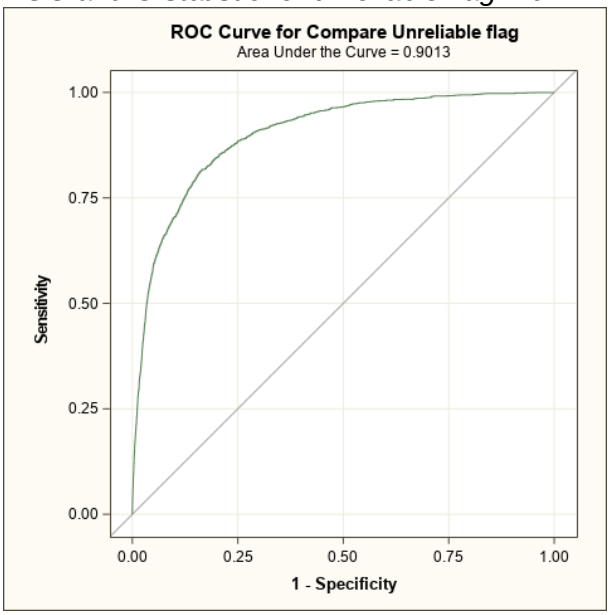
Figure 5. Histogram for unreliability metric and scatter plot of unreliability metric vs. probability (Example 2)

Figure 5 shows that the unreliability metric distribution is moderately skewed and the range of values is not large. The scatter plot of the unreliability metric versus estimated probability from the main (full) model reveals unreliable observations are concentrated mostly around uncertain probability region (in the middle of the probability range).

Main (full) model ROC and C-statistic



ROC and C-statistic for unreliable flag = 0



ROC and C-statistic for unreliable flag = 1

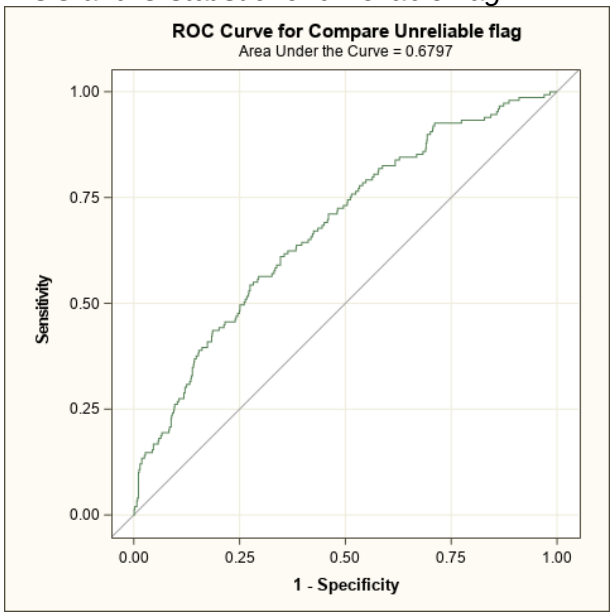
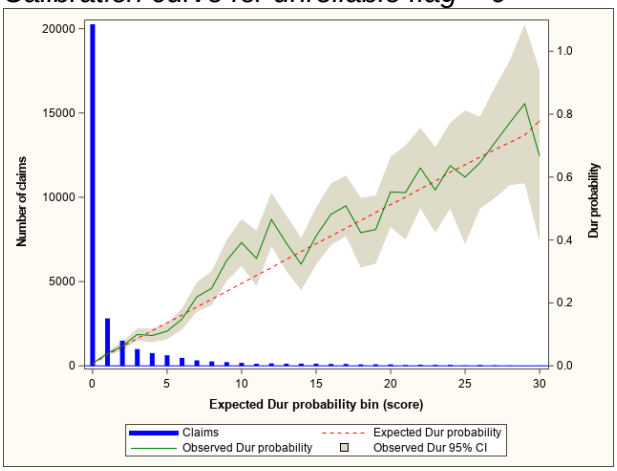


Figure 6. ROC and C-statistic (Area under the curve) for Main (full) model and unreliable and reliable predictions (Example 2)

An interesting observation in Figure 6 is that the ROC curve and C-statistic for reliable observations (unreliable flag = 0) is slightly better than that of main (full model). By removing unreliable observations, we make the reliable subset effectively a better model. We can see substantial deterioration in ROC and C-statistic for observations identified as unreliable (unreliable flag = 1).

Calibration curve for unreliable flag = 0



Calibration curve for unreliable flag = 1

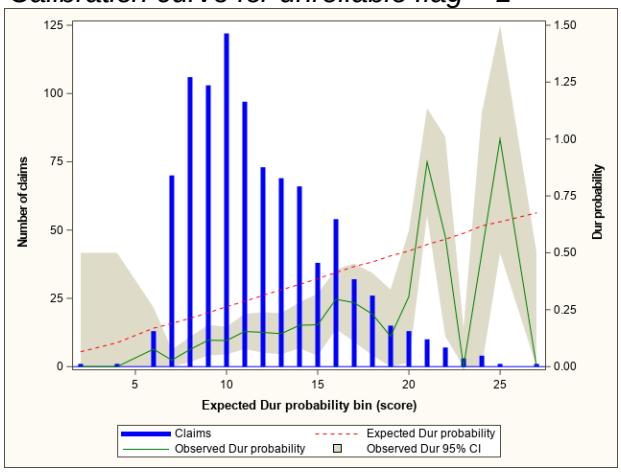


Figure 7. Calibration curves for unreliable and reliable predictions (Example 2)

We see in Figure 7 that a subset of data identified as reliable observations shows good calibration (the observed probability line follows the expected probability line closely). The unreliable observations are not well calibrated (the two lines are far apart from each other).

CONCLUSION

The explored approach allows identification of individual predictions with low reliability. As we see from a plot of unreliability metric versus main (full) model estimated probability, it identifies unreliable observations primarily around uncertain probability region (around 0.5 probability, or prevalence of the outcome in the data). This approach is remarkably accurate at separating reliable from unreliable individual predictions based on ROC curves, C-statistic and calibration curves calculated on the test (unseen) data. C-statistic for reliable prediction is the same or even better than for a full model (which makes a resulting model relatively more accurate on the remaining trusted individual predictions), and it is very close to random choice (C-statistic close to 0.5) for predictions identified as unreliable. Calibration curves for predictions identified as unreliable also show poor calibration.

We arbitrarily used the top third percentile of the unreliability metric distribution to flag individual predictions as unreliable. The cutoff could be increased or decreased based on the required sensitivity of a given claim risk prediction model and the availability of resources for intensive human follow up intervention on predictions identified as unreliable.

Identification of low reliability for individual predictions for claim risk scoring models may help to avoid automatic action based on AI prediction for unreliable sub-groups of predictions. It can then trigger an in-depth human intervention in additional data collection which can be used in the development of an appropriate return-to-work plan.

REFERENCES

1. Myers, P.D., Ng, K., Severson, K. et al. Identifying unreliable predictions in clinical risk models. npj Digit. Med. 3, 8 (2020). https://doi.org/10.1038/s41746-019-0209-7
2. Bosnić, Z., Kononenko, I. An Overview of Advances in Reliability Estimation of Individual Predictions in Machine Learning. Intelligent Data Analysis (2009). 13(2):385-401.
3. Bosnić, Z., Kononenko, I. Comparison of approaches for estimating reliability of individual regression predictions. Data Knowl. Eng. (2008), doi:10.1016/j.datak.2008.08.001.
4. Kononenko, I. et al. Explanation and Reliability of Individual Predictions. Informatica 37 (2013) 41–48.
5. Jiang, H. et al. To trust or not to trust a classifier. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

ACKNOWLEDGMENTS

The authors would like to thank Lorne Rothman (SAS Institute), Dragos Capan and Oliver Jovanovski (Manulife) for their thoughtful comments and peer review of the draft paper. The authors also would like to thank Andrea Concil (WSIB) for her contribution to the overall readability of the paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Yuriy Chechulin, Senior Statistician, Predictive Modelling
Advanced Analytics Branch
Corporate Business Information & Analytics Division
Strategy, Analytics & People Cluster
Workplace Safety and Insurance Board of Ontario, Canada
Yuriy_Chechulin@wsib.on.ca