Paper 2521-2018

Claim Risk Scoring using Survival Analysis Framework and Machine Learning with Random Forest

Yuriy Chechulin, Jina Qu, Terrance D'souza
Workplace Safety and Insurance Board of Ontario, Canada

ABSTRACT

The Workplace Safety and Insurance Board of Ontario is an independent trust agency that administers compensation and no-fault insurance for Ontario workplaces. Claim risk scoring can allow claims at most risk of prolonged duration to be identified. Early identification of such claims helps targeting them with interventions and tailored claim management initiatives to improve duration and health outcomes. Claim risk scoring is done using a discrete time survival analysis framework. Logistic regression with spline for time to better estimate the hazard function and interaction of a number of factors with time spline to properly address proportional hazard assumption is used to estimate the hazards and the corresponding survival probability (very sophisticated “conventional” model). In recent years, Machine Learning methods, including Random Forests (RF), started to gain popularity, especially when the emphasis of the modelling is accurate prediction. Comparison of the existing conventional model and RF Machine Learning algorithm implementation is presented. SAS Enterprise Miner® high-performance procedure HPFOREST was used for RF. RF parameters tuning using graphical analysis was explored. Time-specific percent response and lift charts, accuracy and sensitivity statistics were used to evaluate the predictive power of the models. RF achieved better performance in early stages of the claim life-cycle and was implemented.

INTRODUCTION

The Workplace Safety and Insurance Board of Ontario (WSIB) is an independent trust agency that administers compensation and no-fault insurance for Ontario workplaces. Claim risk scoring was undertaken to allow claims at most risk of prolonged duration to be identified. Early identification of such claims helps targeting them with interventions and tailored claim management initiatives to improve claim duration and health outcomes for injured workers.

For the purposes of the analysis, claim risk is defined as high probability of a claim to be on loss-of-earnings (LOE) benefits in the next month. Being off LOE benefits was used as indirect proxy for successful return to work. We use a discrete time survival analysis framework to model time-to-event (claim is off benefits) and two estimation methods: conventional logistic regression, and Machine Learning with Random Forest (RF). We discuss some of the advanced modelling features used in logistic regression to achieve a fairly sophisticated “conventional” model, as well as provide details on tuning some of the parameters for the competing estimation approach using RF. Comparison of the conventional model and RF Machine Learning algorithm implementation is presented.

METHODOLOGY

An injured workers cohort for the analysis was constructed for injury years 2013-2015 using de-identified WSIB administrative data. Since the interest was in claim durations up to and including one year (52 weeks), we used the necessary follow-up window to capture the claim outcome (on or off benefits).

A number of predictor variables were used in the analysis (see Table 1). Time-dependent variables are marked with an asterisk (the concept of a time-dependent variable is discussed later in the paper).

Name*	Description	Note
Acc_age or Age_group	Injured worker's age at accident	
Gender	Injured worker's gender	
GRP_CLM_SECTOR10	Industry sector	Grouped using sector Rate group
GRP_INJ20	Injury group	Grouped Nature of Injury and Part-of-Body codes
GRP_INJSTICK	Grouped Injury Stickman codes	
Source1 and Event1	Injury source and event codes	First digit of the code
GRP_FIRMSIZE	Grouped firm size	
Wage_grp	Grouped wage	Quartiles plus 90th percentile
Prior_claims	Prior claims flag	Within last 3 years
Prior_NEL	Prior claims with NEL flag	Non-economic loss (permanent impairment)
eAdj	e-Adjudication flag	Automatic claim adjudication
S2	Schedule 2 employer flag	Individual liability, larger mostly government employers (do not report firm size)
FLANGUAGE	Foreign language flag	English, French, or Other
NEL*	Non-economic loss (NEL) flag	Permanent impairment
NOC1	National Occupation Code (NOC)	First digit of the code
Partial_LOE*	Partial LOE benefits flag	Proxy for return to work on partial duties
RTW_ref*	Return-to-Work program referral flag	
SC_ref*	Specialty Clinic program referral flag	
Represent*	Employer or worker representative flag	
SIS	Serious Injury Program flag	
HC_IP*, HC_Psych*, HC_other*, Pain*, Opioid*	Inpatient care, Psych, other health care, presence of pain or opioid medication use	Flags for various health care services

*Time-dependent variables are marked with an asterisk.

Table 1. List of predictor variables used in the analysis

Categorical variables with too many levels to include (for example, industry mix with claim Rate group) were feature-engineered/binned into fewer levels. The problem with using too many levels in a regression modelling framework (for logistic regression) is that, first, it introduces too many degrees of freedom, which hinders the estimation, and second, some of the levels of the original categorical variable have too small sample sizes (issues with quasi-complete separation in logistic regression, etc.). First, we calculated the risk of the outcome (proportion on LOE benefits at 6 months) in each Rate group based on the whole study population, then we sorted the Rate groups in order of the risk, and binned Rate groups into 10 risk groups (GRP_CLM_SECTOR10) using quintile method (trying to keep about the same number of observations in each of ten groups). We employed the same method as above for grouping Nature of Injury and Part-of-Body codes into Injury mix groups (20 groups, GRP_INJ20).

Analysis of claim duration is a typical time-to-event analysis: best addressed with survival analysis framework, in our case its discrete-time variant (Allison, 2010). Each claim survival history was broken down into a set of discrete time units (weeks) that were treated as distinct observations. Then we created an expanded data set where each claim had as many records as there were "alive" time points, until this

claim is off benefits (claims were censored at 57 weeks of duration). We coded an outcome variable “Dur” as “1” for time periods when a claim is on LOE benefits and “0” when the claim gets off benefits (it allows a more logical interpretation of hazard ratios from the estimation using logistic regression: hazard ratios more than 1 show “negative” effect on duration, and less than 1, “positive”). Survival analysis allows proper modelling of time-dependent factors (factors that change over time). Table 2 shows an example of an expanded data set for discrete time survival analysis. It shows also an example of a time-independent variable, Gender (does not change over time), and a time-dependent variable, Partial LOE (may change over time; this is a flag for partial LOE benefits, which is an indirect proxy for return to work on partial duties).

Claim	Time (weeks from accident)	Gender	Partial LOE	Dur (outcome/target; on or off LOE benefits)
1	0	F	0	1
1	1	F	0	1
1	2	F	0	0
2	0	M	0	1
2	1	M	0	1
2	2	M	1	1
2	3	M	1	1
2	4	M	1	0

Table 2. Example of an expanded data set for discrete time survival analysis

First, we used a common approach to estimate whether an event did or did not occur in each time unit (week) using logistic regression model. In the survival model, interactions with time variable were used to address non-proportional hazard, as well as time itself was modelled using a spline effect to better estimate the hazard function.

SAS code below shows an example call to the LOGISTIC procedure. CLASS statement declares categorical variables. EFFECT statement specifies that we want to fit the natural cubic spline for time variable. MODEL statement specifies that we are modelling claim duration against the list of our variables; note that we also fit a number of interactions for time-dependent variables with our time spline. EFFECTPLOT statement asks for the plot of our fitted spline for time (see Figure 1); as can be seen, the effect is clearly non-linear, so spline for time is warranted. STORE statement stores our model as a binary file for future scoring (we will need to use the PLM procedure to score our data, since we used spline effects in the model). ODDSRATIO statement asks to produce hazard ratios as an example of one of the dependent variables (partial LOE, or proxy for return to work on partial duties) in this case. Since this variable was interacted with time, we need to ask for odds ratios (in fact, these are hazard ratios due to the discrete time survival analysis framework we employ) at different time points (weeks of duration). Table 3 shows the estimated hazard ratios for this time-dependent variable and, as can be seen, the hazard changes over time for the Partial LOE effect (in this way we address the non-proportional hazard assumption). Claims that survived to a given time point and have Partial LOE (return to work on partial duties) have a lower hazard of being on LOE benefits in the next time period than claims that are on full LOE (fully off work), and this hazard decreases over time / claim life cycle. In other words, injured workers who are already on partial duty are likelier to fully return to work in the next time period than are workers who are not at work at all, which makes sense.

```
ods graphics on;
proc logistic data=dur_surv descending;
      class Age_group(ref='1') Gender(ref='F')
      GRP_CLM_SECTOR10(ref='0') GRP_INJ20(ref='01')
      GRP_INJSTICK(ref='1') GRP_FIRMSIZE(ref='8') NOC1(ref='7')
      FLANGUAGE(ref='1') source1(ref='5') event1(ref='2')
      Wage_grp(ref='Q1') Prior_claims(ref='0') / param=ref;
      effect Time_spl = spline(Time / basis=tpf(noint)
      naturalcubic knotmethod=equal(5));
      model Dur = Age_group Gender GRP_CLM_SECTOR10 GRP_INJ20
      GRP_INJSTICK GRP_FIRMSIZE Wage_grp Prior_claims Prior_NEL
      eAdj S2 SIS FLANGUAGE NOC1 source1 event1 NEL Partial_LOE
      RTW_ref RTW_fail SC_ref Represent HC_IP HC_other HC_Psych
      Pain Opioid
      Time_spl
      Partial_LOE*Time_spl RTW_ref*Time_spl RTW_fail*Time_spl
      SC_ref*Time_spl Represent*Time_spl HC_IP*Time_spl
      HC_other*Time_spl HC_Psych*Time_spl Pain*Time_spl
      Opioid*Time_spl SIS*Time_spl;
      effectplot fit(x=Time) / noobs link;
      store crs.dur_surv_model;
      oddsratio Partial_LOE / at(time=4 8 13 17 22 26 34 42 52);
run;
ods graphics off;
```

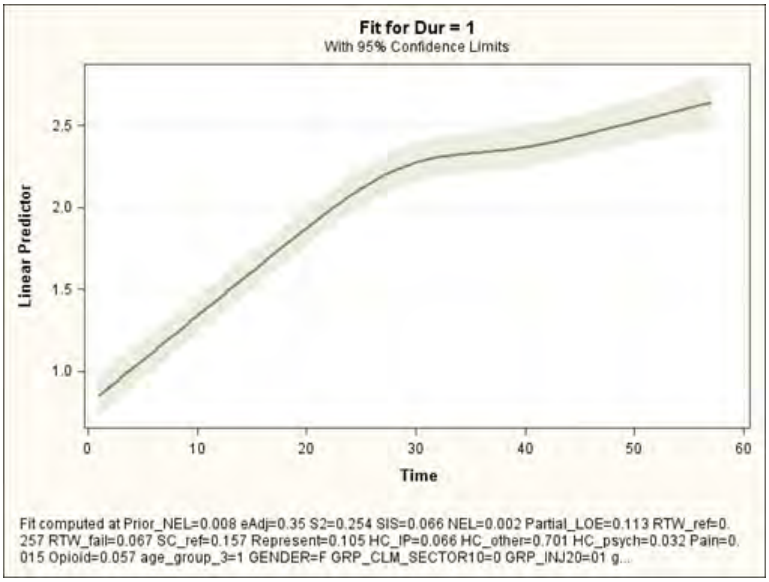


Figure 1. Plot of spline for time variable

Factor	Estimate	95% Confidence Limits	
Partial_LOE at Time=4	**0.657**	0.636	0.678
Partial_LOE at Time=8	**0.551**	0.538	0.564
Partial_LOE at Time=13	**0.443**	0.428	0.458
Partial_LOE at Time=17	**0.384**	0.367	0.4
Partial_LOE at Time=22	**0.356**	0.339	0.374
Partial_LOE at Time=26	**0.361**	0.334	0.389
Partial_LOE at Time=34	**0.345**	0.318	0.374
Partial_LOE at Time=42	**0.336**	0.301	0.375
Partial_LOE at Time=52	**0.392**	0.338	0.456

Table 3. Hazard ratios with confidence limits for time-dependent Partial LOE variable at different time points (weeks of claim duration)

Conventional modelling with the LOGISTIC procedure allows us to provide very detailed information on the effect of various factors on the modelled outcome (very good for explanatory modelling). In recent years, Machine Learning methods, including Random Forests (James, 2014), started to gain popularity, especially when emphasis of the modelling is accurate prediction, and there is no particular need for the explanatory component. For comparative purposes we applied random forest model to our expanded discrete time data set to estimate the outcome.

Classification and regression trees work by recursive partitioning of the data into groups ("nodes") that are increasingly homogeneous with respect to some kind of a criterion. Usually, mean squared error is used for regression trees, and Entropy or the Gini index is used for classification trees. Random Forest takes predictions from many classification or regression trees and combines them to construct more accurate predictions through the following algorithm:

- Many random samples are drawn from the original data set. Observations in the original data set that are not in a particular random sample are said to be "out-of-bag" (OOB) for that sample.
- To each random sample a classification or regression tree is fitted without any pruning. Predictors for each tree are randomly chosen.
- The fitted tree is used to make predictions for all the observations that are out-of-bag for the sample the tree is fitted to.
- For a given observation, the predictions from the trees on all of the samples for which the observation was out-of-bag are combined.

Classification Trees and Random Forests take into account all of the necessary interactions, the lack of which in many cases results in worse predictive power for conventional regressions.

SAS Enterprise Miner® high-performance procedure HPFOREST was used for RF; however, actual implementation was done using SAS® coding in SAS Enterprise Guide®. It should be noted that PROC HPFOREST could be called from the programming interface of SAS Enterprise Guide only if SAS Enterprise Miner® is also installed on the same SAS Server.

SAS code showing an example of discrete time survival analysis with estimation using Machine Learning with Random Forest is shown below. We use a number of INPUT statements to specify the variables that we want to include for modelling (one for interval variables, and one for nominal variables). We also specify our target (variable "Dur"), and state that this variable is binary. SAVE statement allows us to save the random forest model into a binary file for future scoring of (new) data using the HP4SCORE

procedure. We save a number of tables from RF modelling output for future reference using ODS OUTPUT statement:

```
proc hpforest data=dur_surv seed=12345 maxtrees=200
          alpha=0.05 vars_to_try=15;
     input Time Acc_Age Wage Prior_NEL eAdj S2 SIS NEL
          Partial_LOE RTW_ref RTW_fail SC_ref Represent HC_other
          HC_Psych Pain Opioid HC_IP / level=interval;
     input Gender GRP_CLM_SECTOR10 GRP_INJ20 GRP_INJSTICK
          GRP_FIRMSIZE FLANGUAGE NOC1 source1 event1
          Prior_claims / level=nominal;
     target Dur / level=binary;
     save file = "\\srvscudd2\PM
DEV2\Projects\Claim_risk_scoring\dur_surv_model_RF.bin";
     performance details;
     ods output fitstatistics = crs.RF_fit
          VariableImportance = crs.RF_VarImportance
          ModelInfo = crs.RF_ModelInfo;
run;
```

Random Forest has a number of parameters that can be tuned to improve the model accuracy. In this paper, we will show an example of tuning one of the most important parameters using graphical analysis: number of variables to try ("VARS_TO_TRY"). "VARS_TO_TRY=m | ALL" syntax specifies the number of input variables to consider splitting on in a node. *m* ranges from 1 to the number of input variables, *v*. The default value of *m* is $\sqrt{v}$; however, we can run a number of models trying different values for *m* and choosing the best model using "out-of-bag" (OOB) prediction error and/or misclassification rate. The HPFOREST procedure computes the average square error (ASE) measure of prediction error. For a binary or nominal target, PROC HPFOREST also computes the misclassification rate and the log-loss. Figure 2 shows OOB prediction error and misclassification rate for random forests with a different number of "variables to try" (5, 7, 9, 11, 13, or 15). Probably due to a discrete time survival analysis set-up of our (expanded) dataset, the OOB misclassification rate does not seem to be very informative. Based on the OOB prediction error, we can see that the model with 15 variables to try achieves the best performance.

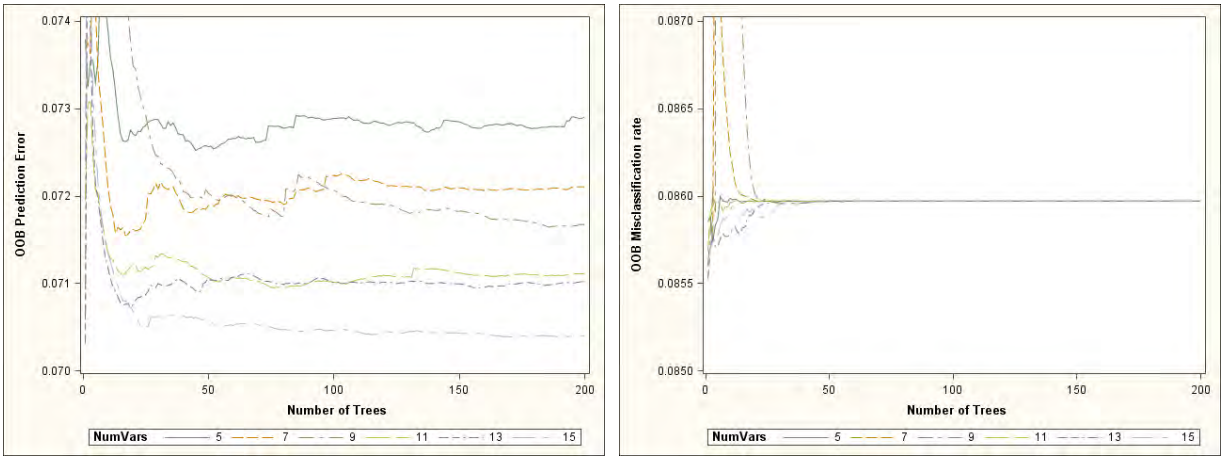


Figure 2. Out-of-bag prediction error and misclassification rate for Random Forests with a different number of "variables to try"

Figure 3 shows the final model (vars_to_try = 15) OOB vs Training (Full data) ASE Prediction error and Misclassification rate.

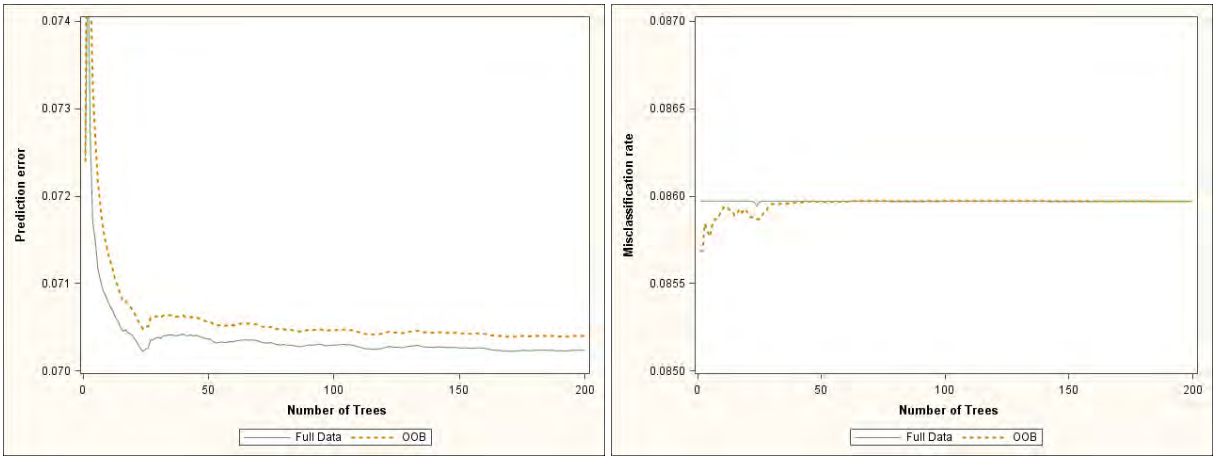


Figure 3. Final model (vars_to_try = 15), OOB vs Training (Full data) ASE Prediction error and Misclassification rate

Variable importance from the Random Forest final model is shown in Table 4. This table provides information on the number of times the variable was used to split a node, as well as Gini, Margin, Gini Out-of-Bag (OOB), and Margin Out-of-Bag metrics. As can be seen, the Time variable is the most important variable (based on Gini metric), which warrants a survival analysis framework approach to this data and suggests that the hazards may be not constant over time. Type of injury is the second most important predictor, followed by partial return to work on modified duties. In Figure 4, we also plotted the logit of Random Forest Prediction versus Time (holding all other variables at their corresponding means or the same reference levels as in the logistic regression) to compare it to Figure 1 from logistic regression with regard to the estimated baseline hazard. The two plots are not exactly the same, but both suggest that the effect of Time is clearly not linear.

Variable	NRules	Gini	Margin	GiniOOB	MarginOOB
Time	9154	0.00835	0.01671	0.04321	0.06701
GRP_INJ20	4565	0.00275	0.0055	0.00501	0.00328
Partial_LOE	1399	0.0008	0.00159	0.09813	0.10708
SC_ref	1287	0.00047	0.00093	0.04025	0.0438
event1	1910	0.00027	0.00053	0.01481	0.01141
grp_injstick	1538	0.00022	0.00044	0.02222	0.0122
ACC_AGE	2678	0.00022	0.00043	-0.0011	-0.0002
RTW_ref	1433	0.00017	0.00034	0.00989	-0.01299
HC_other	1297	9.9E-05	0.0002	-0.2244	-0.2139
GRP_CLM_SECTOR10	1143	9.5E-05	0.00019	0.00069	0.00091
NOC1	1455	9.3E-05	0.00019	0.00564	0.00249
WAGE	1752	7.9E-05	0.00016	0.0006	0.00117
source1	1054	7.2E-05	0.00014	-0.0017	-0.00249
eAdj	656	6.7E-05	0.00013	-0.0048	-0.00502
HC_IP	820	6.2E-05	0.00012	0.00223	0.00298
GRP_FIRMSIZE	1240	4.2E-05	8.4E-05	0.00008	0.00011
SIS	391	4.2E-05	8.4E-05	-0.0025	-0.00339
Represent	517	3.3E-05	6.6E-05	0.00083	0.00142
GENDER	686	1.3E-05	2.6E-05	0	0.00001
Opioid	216	1.3E-05	2.7E-05	0.00025	0.00039
S2	661	8E-06	1.7E-05	0	0.00003
RTW_fail	284	6E-06	1.1E-05	-0.0007	0.0004
HC_psych	116	4E-06	8E-06	0.00002	0.00029
Prior_claims	45	1E-06	1E-06	0	0
Pain	26	0	1E-06	0	0.00002
NEL	0	0	0	0	0
Prior_NEL	0	0	0	0	0
FLANGUAGE	0	0	0	0	0

Table 4. Variable Importance from Random Forest.

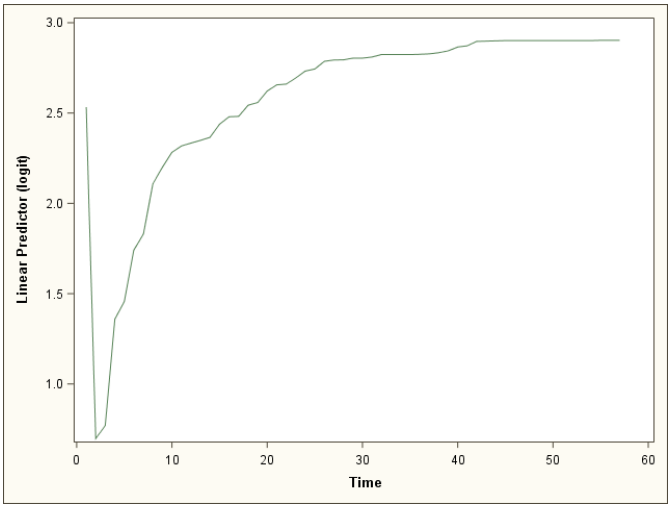


Figure 4. Logit of the Random Forest Prediction versus Time

Once we have our discrete time survival analysis model estimated using these two methods (logistic regression and random forest), we can score (new) data and calculate the survival probability. Below is an example of the SAS code:

```
*Score Logistic;
proc plm restore=crs.Dur_surv_model;
      show effects parameters;
      score data=dur_surv_expand out=dur_surv_score predicted;
run;

*Score Random Forest;
proc hp4score data=dur_surv_expand;
      id _ALL_;
      score file= "\\srvscudd2\PM
DEV2\Projects\Claim_risk_scoring\dur_surv_model_RF.bin"
      out=dur_surv_score(rename=(P_Dur1=Prob));
      performance details;
run;
```

```
      *Calculate survival probability;
      data dur_surv_score;
            set dur_surv_score;
            by clmno;
            retain Prev_Surv_prob;
      *     Prob = exp(Predicted) / (1 + exp(Predicted)); *Comment out
for RF;
            if first.clmno then Prev_Surv_prob = 1;
            Surv_prob = Prev_Surv_prob * Prob; *(1 - Prob) if modelled
Dur=0;
            output;
            Prev_Surv_prob = Surv_prob;
            drop Prev_Surv_prob;
      run;
```

Please note that we need to use a "Prob = exp(Predicted)/(1+exp(Predicted))" statement for data scored by PLM procedure (it produces a linear score (on a logit scale), and we need to convert it back to the hazard). For scoring of data using HP4SCORE procedure, this statement has to be commented out (not needed).

In order to calculate the survival probability, we keep in mind that survival function at time t_i can be written in terms of the hazard at all prior times $t_1, \ldots, t_{i-1}$, as

$$S_i = (1 - h_1)\,(1 - h_2)\ \ldots\ (1 - h_{i-1})$$

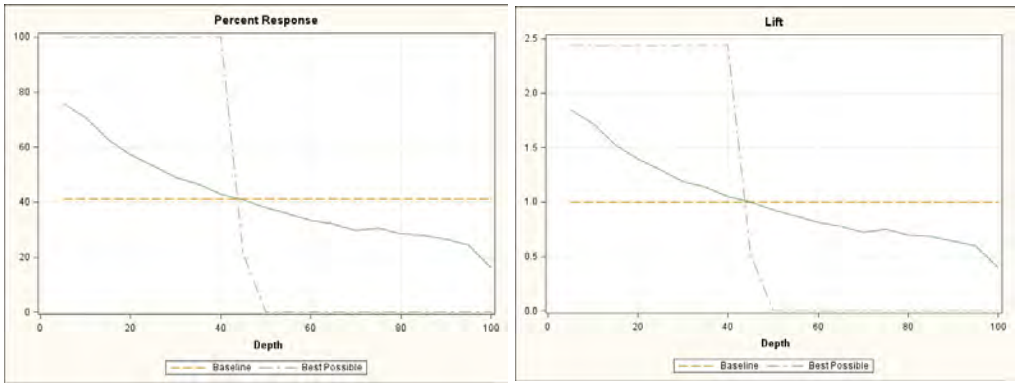
In other words, this result states that in order to survive to time t_i one must first survive t_1, then one must survive t_2 given that one survived t_1, and so on, finally surviving t_{i-1} given survival up to that point. (Rodríguez, 2017). We implement this calculation using DATA step with BY and RETAIN statements as shown in the SAS code above. Please note that we are using in the formula (Prob) ("Prob" is a variable for estimated hazard) and not (1-Prob) since we are estimating Dur = 1 and not Dur = 0 in our particular data set up.

RESULTS

Time-specific percent response and lift charts, accuracy and sensitivity statistics were used to evaluate the predictive power of the models. By time-specific we mean that the risk scoring is done for claims that survived to a certain time period (risk week, in our terminology), and we estimate a risk of being on LOE benefits in the next month. Time-specific slicing is possible due to our survival analysis framework approach to modelling.

Figure 5 and Figure 6 show percent response and lift charts for risk weeks 8 and 12 correspondingly. As can be seen, the RF model achieves better performance for the riskiest claim buckets in early stages of life-cycle duration. As the claims mature, the two estimation methods (RF and logistic) become more and more similar in their predictive power (Figure 7 and Figure 8 for risk weeks 28 and 52 correspondingly). Probability of staying on benefits in the next month for claims that managed to survive long is very high, and the model becomes less and less discriminative at later stages of the claim life cycle. Looking at Percent Response graphs, we can see that for claims that survived to risk week 8, only 40% on average remain on benefits after one month (orange horizontal dotted line), while for claims that survived to risk week 52, almost 80% remain on benefits one month later. For the riskiest bucket of claims, the lift is around 2 for claims that survived to risk week 8, and only around 1.25 for claims that survived to risk week 52.

Logistic regression with splines and interactions



Random Forest Machine Learning

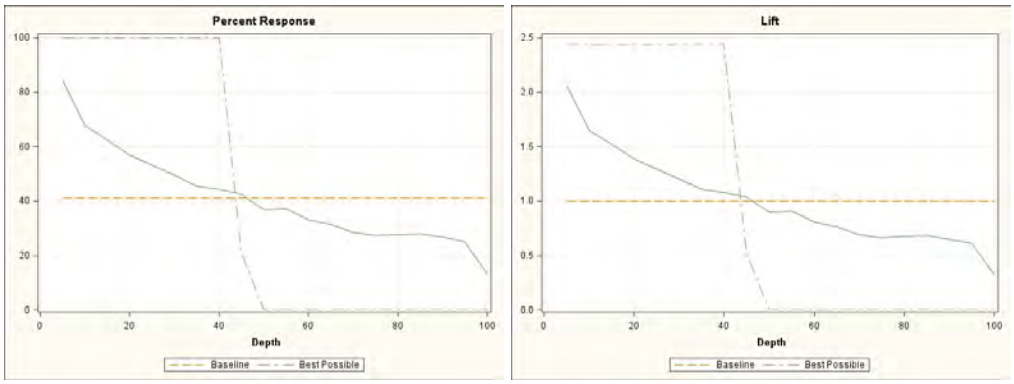
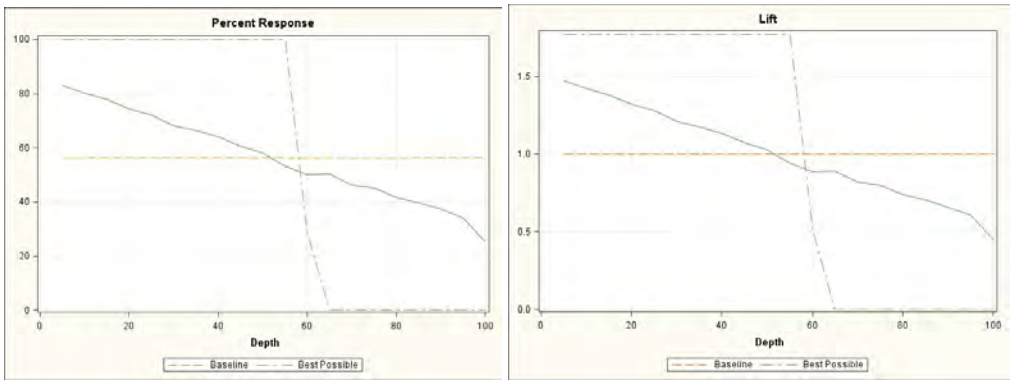


Figure 5. Percent Response and Lift charts: risk week 8

Logistic regression with splines and interactions



Random Forest Machine Learning

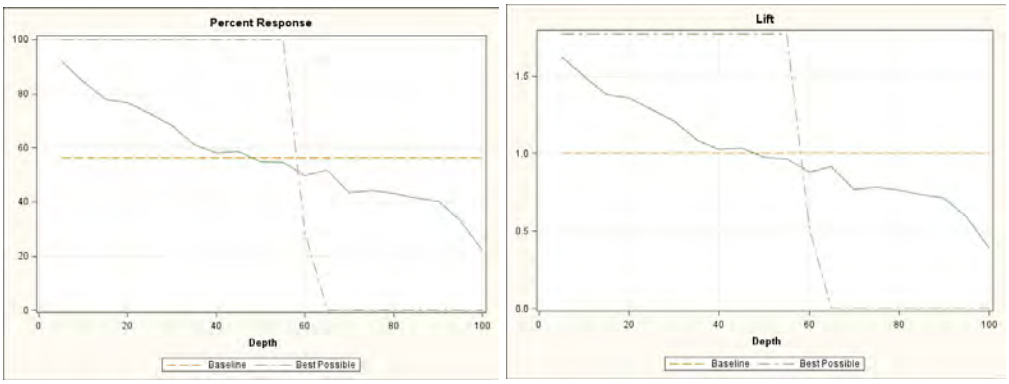
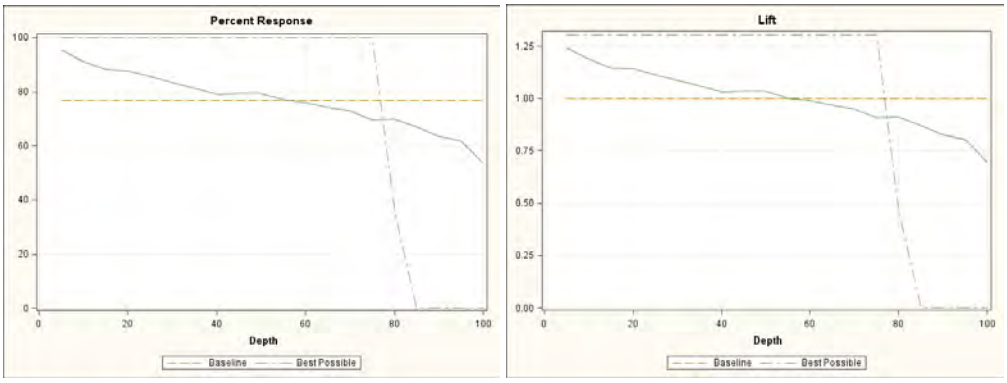


Figure 6. Percent Response and Lift charts: risk week 12

Logistic regression with splines and interactions



Random Forest Machine Learning

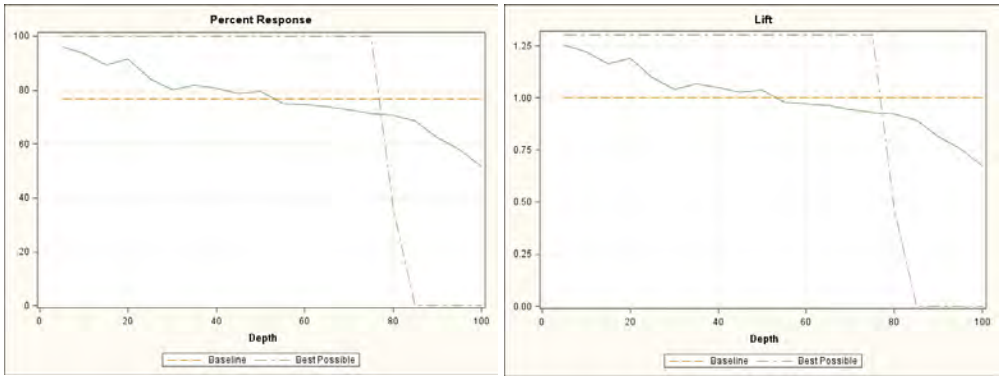
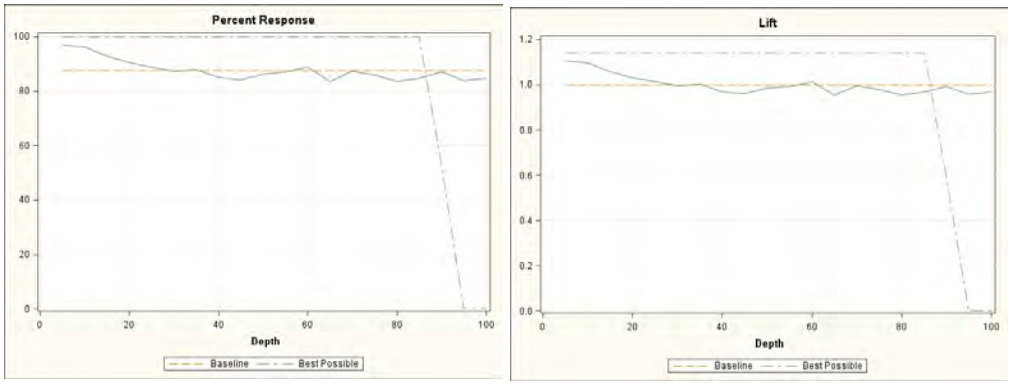


Figure 7. Percent Response and Lift charts: risk week 28

Logistic regression with splines and interactions



Random Forest Machine Learning

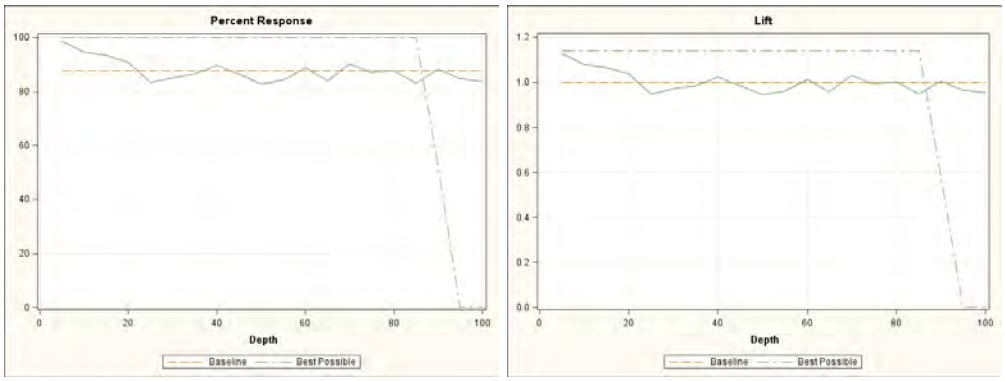


Figure 8. Percent Response and Lift charts: risk week 52

Time-specific calculated sensitivity and accuracy are presented in Table 5. The table also shows percent on benefits in the next month for claims that survived up to that time point (risk week), as well as arbitrarily chosen model cut-offs for survival probability to label risky claims. In many cases the model performance could be optimized if cutoffs corresponded to the underlying prevalence of an event of interest (in our case, percent on benefits). However, we modified the cut-offs to meet capacity requirements (i.e., how many claims could be physically followed up given available resources). In any case, the cut-offs are the same for both estimation methods (Random Forest and logistic regression), and the models could be directly compared. As we can see, the Random Forest achieves slightly better predictive power than logistic regression in early stages of the claims life cycle, and the performance is almost identical for long-surviving claims.

Risk week	Random Forest Machine Learning		Logistic with splines and interactions		Percent on benefits in one month	Existing Cut-offs, Top
	Sensitivity	Accuracy	Sensitivity	Accuracy		
8	**56.5%**	**65.3%**	**55.7%**	**64.7%**	**41.1%**	**40.0%**
12	**52.4%**	**62.7%**	**52.0%**	**62.2%**	**56.4%**	**40.0%**
16	51.9%	62.4%	50.7%	60.9%	61.5%	40.0%
20	49.7%	59.6%	49.0%	58.7%	67.0%	40.0%
24	47.1%	55.7%	46.8%	55.3%	72.6%	40.0%
28	**65.5%**	**63.8%**	**65.3%**	**63.5%**	**76.8%**	**60.0%**
32	63.9%	62.2%	64.0%	62.4%	79.7%	60.0%
36	62.6%	60.5%	63.0%	61.2%	81.8%	60.0%
40	62.1%	60.1%	62.6%	60.9%	83.2%	60.0%
44	61.6%	59.8%	62.0%	60.5%	85.4%	60.0%
48	80.6%	73.2%	80.7%	73.4%	87.1%	80.0%
52	**80.6%**	**73.7%**	**80.6%**	**73.6%**	**87.6%**	**80.0%**

Table 5. Sensitivity and Accuracy

The following formulas are used to calculate sensitivity and accuracy of the model at different time points (risk weeks):

$$\text{Sensitivity} = TP / (TP + FN)$$

$$\text{Accuracy} = (TP + TN) / (P + N)$$

Where TP – true positive, TN – true negative, FP – false positive, FN – false negative, P – positive, N – negative.

In order to do the validation of the modelling approaches, we partitioned our data into the training data set (60%) and validation data set (40%) using cluster sampling (cluster = claim) to ensure that the whole claim with all its time observations, and not individual records, is being sampled. We re-trained both logistic regression and Random Forest models only on the training data set, and we scored the hold-out validation data set. Table 6 shows sensitivity and specificity on the hold-out validation data set, and as can be seen, the results are very similar to our full sample results shown in Table 5. Once again, the Random Forest achieves slightly better predictive power than logistic regression in early stages of the

claims life cycle, and the performance is almost identical for long-surviving claims on the hold-out validation data set.

Risk_wk	Random Forest Machine Learning		Logistic with splines and interactions	
	Sensitivity	Accuracy	Sensitivity	Accuracy
8	**56.9%**	**65.6%**	**55.5%**	**64.4%**
12	**52.8%**	**63.1%**	**51.8%**	**62.1%**
16	52.4%	62.9%	50.5%	60.6%
20	50.5%	60.7%	49.2%	58.9%
24	47.7%	56.6%	47.0%	55.6%
28	**66.3%**	**64.9%**	**65.8%**	**64.2%**
32	64.4%	63.0%	64.0%	62.4%
36	62.8%	61.1%	62.7%	60.9%
40	62.0%	59.9%	62.0%	59.9%
44	61.5%	59.5%	61.7%	59.9%
48	80.2%	72.5%	80.5%	73.1%
52	**80.5%**	**73.2%**	**80.6%**	**73.3%**

Table 6. Sensitivity and Accuracy on the hold-out validation data set

CONCLUSION

This paper presents a proof-of-concept for using Survival Analysis and Machine Learning with Random Forest for claim risk scoring purposes.

Both estimation methods (conventional logistic regression and Random Forest) show very good goodness-of-fit across all time points (weeks of claim duration); however, the models at longer durations become progressively less and less useful. Claims with longer and longer durations have very low propensity to close in the next time period. All of such claims are effectively “very risky,” and should probably be subject to intensive management/interventions irrespective of any model.

Machine Learning with Random Forest estimation is very similar in predictive power to a sophisticated “conventional” logistic regression with splines and interactions. However, RF achieves better prediction power for the riskiest claims in early stages of the claim life-cycle, so it may warrant a switch to RF as a primary tool for claim risk scoring for this particular data. Since Random Forest focuses on prediction and not explanation, it provides fewer benefits for understanding the impact of various factors on duration outcomes. We still need conventional modelling to understand the exact impact of individual factors for operational improvement initiatives. Machine Learning with Random Forest was implemented in the Claim Risk scoring project as a viable (and superior) alternative to conventional modelling.

REFERENCES

Allison, P. D. 2010. *Survival Analysis Using SAS®: A Practical Guide, Second Edition*. Cary, NC: SAS Institute Inc.

James, G., Witten, D., Hastie, T., Tibshirani R. 2014. *An Introduction to Statistical Learning: with Applications in R*. Springer Publishing Company, Incorporated.

Rodríguez G. 2017. *Discrete Time Models.* Princeton University. http://data.princeton.edu/wws509/notes/c7s6.html (accessed December 20, 2017).

ACKNOWLEDGMENTS

The authors would like to thank Frank Ferriola, Charles Schwab & Co., and Lorne Rothman, SAS Canada, for their thoughtful comments and peer review of the draft paper.

CONTACT INFORMATION

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

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