Employer Features Generation on Claim Risk

Diane Zhu / MBAN, Schulich School of Business, July 24th, 2018

wsib ONTARIO

Table of Content

1. Project Scope

wsib ONTARIO

Background

5,000,000+
WORKERS COVERED

290,000+
EMPLOYERS COVERED

- WSIB, as one of the largest insurance organizations in North America, is a government agency that, guided by legislation, provides workplace compensation and services.
- Entirely funded by employer premiums and administered as an independent trust agency.

- Under the Strategy and Analytics cluster, the Corporate Business Information and Analytics (CBIA) division supports the organization and drives better outcomes for workers and employers by transforming data into business relevant solutions that inform evidence based decision making.

Diane Zhu

Program Associate - Data Science, Advanced Analytics
Corporate Business Information & Analytics Division
Strategy and Analytics Cluster

wsib ONTARIO

Strategic Goals & Priorities

Promote health and safety in Ontario workplaces

- Embracing the role as a key health and safety system partner by using advanced analytics to help our system partners better assess system performance

Achieve better return-to-work and recovery outcomes and administer benefits fairly

- Enhance adjudication processes and initiate a customer-centric operation model, as well as productivity and responsiveness

Deliver service excellence, quality and care through innovation

- Invest in processes and technologies to determine which solutions are most likely to succeed, where there are bottlenecks in service delivery and how to better respond to trends

Implement data governance and value-add analytics through the Data Governance Program, **predictive models** and building the WSIB's **data science capacity**.

wsib ONTARIO

Business Needs

Supporting the goals and priorities…

- Enhance decision-making in operations through understanding the risk associated with a newly registered claim and the factors influencing the risk.
- Case managers will better see the underlying needs and costs of the claim, and decide whether extra efforts and intervention are necessary.

Predictive Models

We already have a model *(Claim Risk Scoring)* that scores the adjusted risk of a claim. We want to further examine the risk of a newly registered claim by looking at:

How each of the external entities, such as employers and health care service providers, add to the risk of a newly registered claim?

Particularly, I will be looking at **employers' influence.**

Definitions

- **Registered claims** are claims for injuries, illnesses or fatalities reported to the WSIB in the year and includes all allowed, denied, abandoned and pending claims.
- The **risk of a claim** in this case is chosen is to be measured by its likelihood to continue to be on benefits (meaning that WSIB will compensate the injured worker) after a duration of 3 months.
- An **employer** is only considered if they are under WSIB coverage.

wsib ONTARIO

Starting Point: The Existing Claim Risk Scoring

Claim Risk Scoring (Work by my colleague: Yuriy Chechulin)

- Claim risk scoring can allow claims to be identified at most risk of prolonged duration. Early identification of such claims may help targeting these claims with interventions and tailored claim management initiatives to improve duration and health outcomes.

Discrete-time Survival Analysis Framework (logistic regression with splines and interactions)

- Each claim survival history is broken down into a set of discrete time units that are treated as distinct observations. We have an “expanded” data set where each claim has as many records as there are “alive” time points, until this claim is off benefits.
- Estimated whether an event did or did not occur in each time unit using logistic regression model.

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

- Scores the risk – the likelihood of a claim being on or off benefits after a duration of 3 months using time, injury features, worker characteristics, and firm size.
- **For 3 months: Sensitivity = 52%; Specificity = 62%**
- This means for all the historical claims, we have their expected risk that does not take specific external entities into account.

Published Paper: https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/2521-2018.pdf

wsib ONTARIO

2. Project Objective & Deliverables

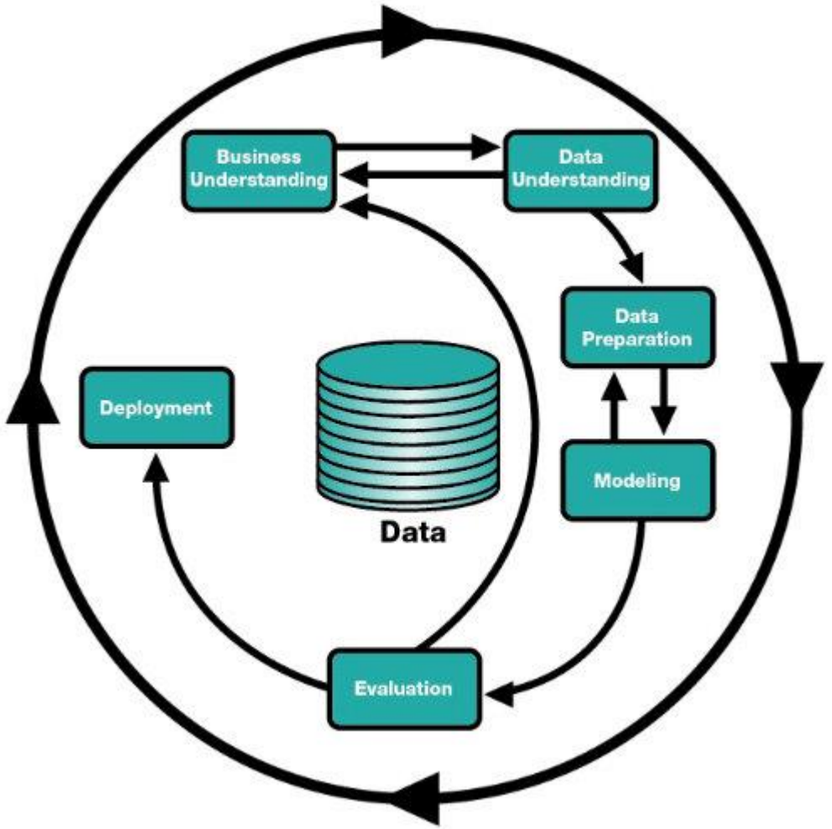
wsib ONTARIO

Project Objective

Building on the existing claim risking scoring model produced by Yuriy, the objective of this project is to develop a model that predicts whether the expected risk of a newly registered claim will be increased, decreased, or unchanged based on its employer features.

Deliverables:

- CRISP-DM methodology/process will be used as a structure approach to track this project (see the chart on the right)
- Predictive modeling techniques will be used on employer features and the claim outcome; different types of models will be built, compared and tested for prediction.
- The future deployment and application of the model will be discussed.



3. Business & Data Understanding & Preparation

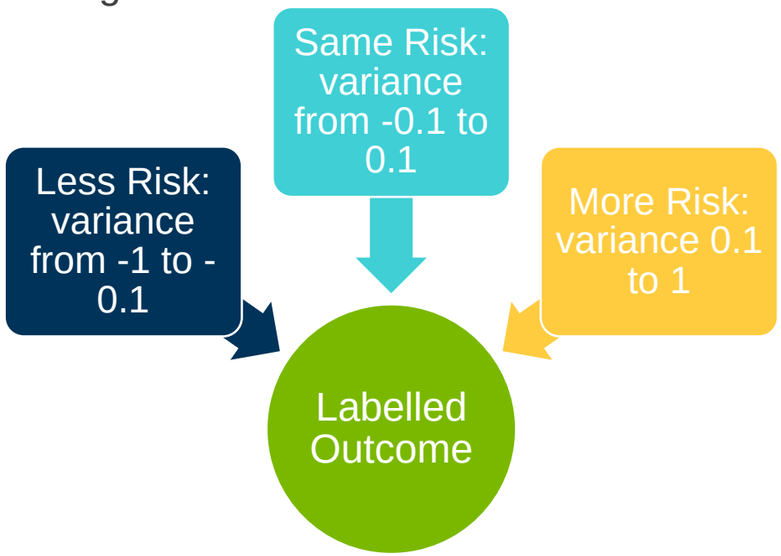
wsib ONTARIO

Business Understanding

As mentioned in Project Scope, the business need is to allow case managers to understand the underlying risk of a newly registered claim, and decide whether intervention with the employers is necessary, which will increase the effectiveness of case management and achieve better return to work and recovery outcomes.

A **multi-classification** analysis will fulfill the need the most:

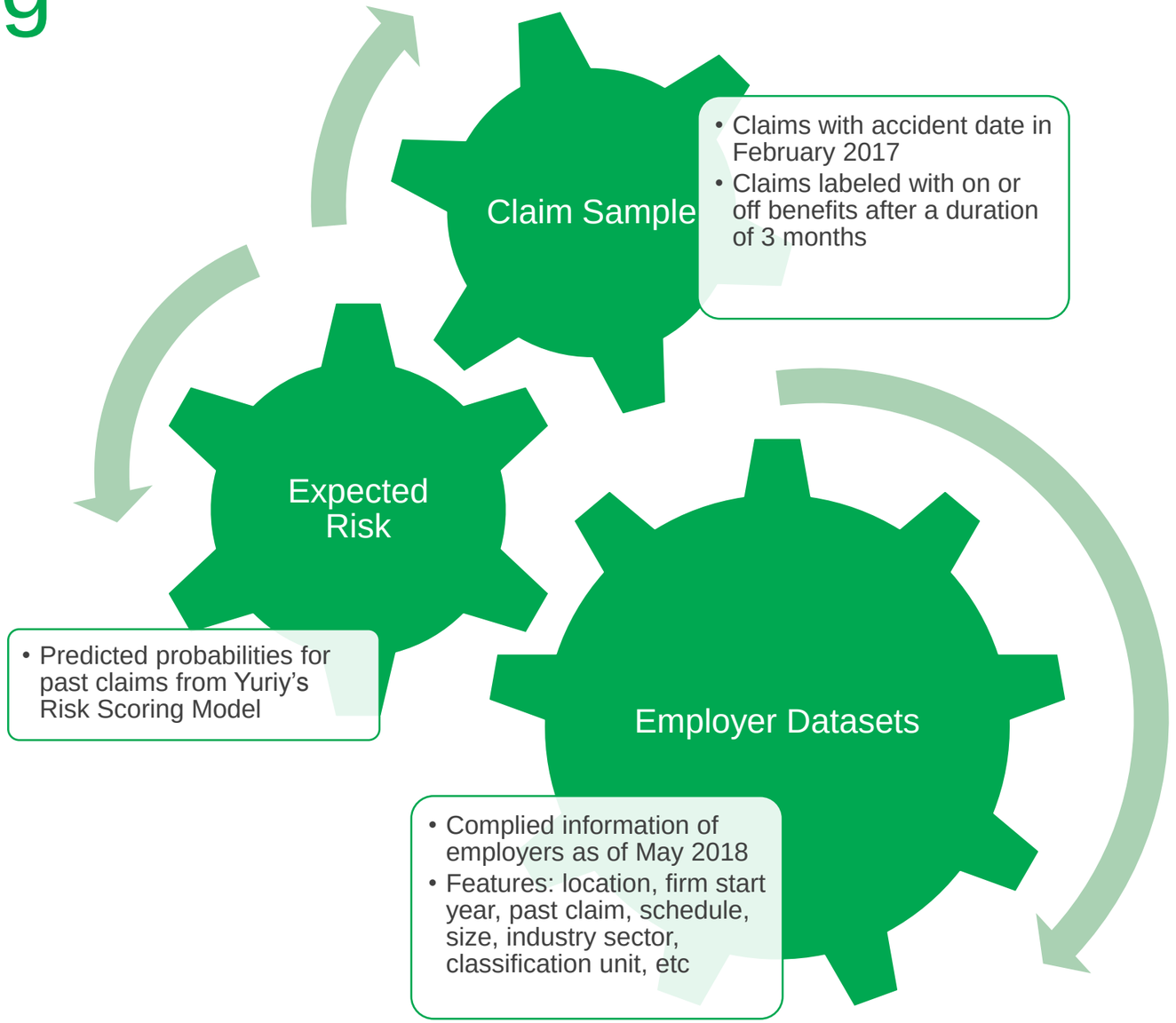
- The variance between the expected risk (probability of being on or off benefits after 3 months) predicted by Yuriy's model and the actual outcomes (1 being on and 0 being off) will be calculated:
  - The variance will have a range from -1 to 1: negative means the expected risk is higher than actual, and positive means the expected risk is lower.
- If the employer features can explain the variance, they can be used to predict and adjust the risk of a claim.
- The variance will be divided into 3 classes, and each class will correspond to a risk decile for easier understanding and decision making for case managers.



Data Understanding

Challenges

- Some important employer features including industry sector and firm size may not be used as they were already used in Yuriy's model
- Two of the employer features are categorical variables with more than 500 levels: city and classification unit



Data Preparation (using R)

Merge Datasets into Final Dataset

- Merged all the datasets into one dataset and removed ID variables; final dataset has 35323 observations and 13 variables

Label Target Variable

- Calculated variance and assigned 3 classes:
  risk$RISK1<-cut(risk$RISK, c(-1, 0.1,0.1,1),labels=1:3)

Categorical Variables Treatment

- Thresholding for Random Forest
- One-Hot Encoding (dummy coding) for XGBoost

Data Splitting

- 70% for training
- 30% for testing

wsib ONTARIO

Final Dataset

12 Predictor Variables:

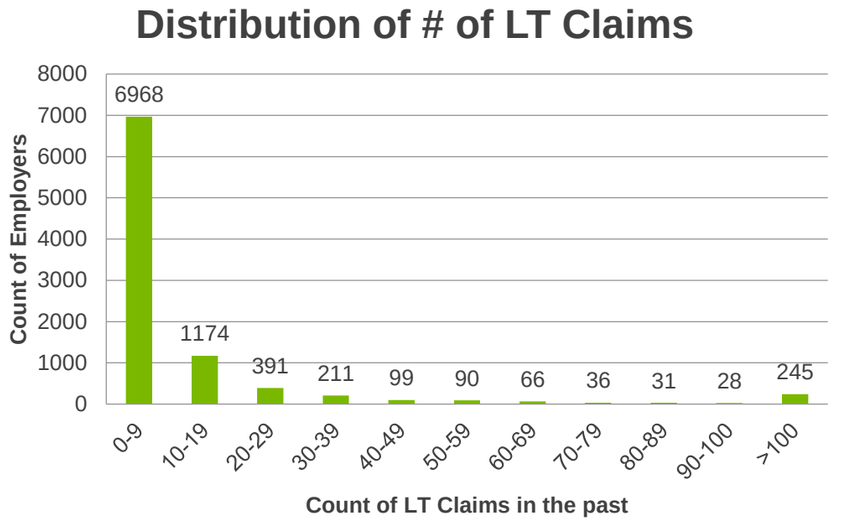
Numeric Variables

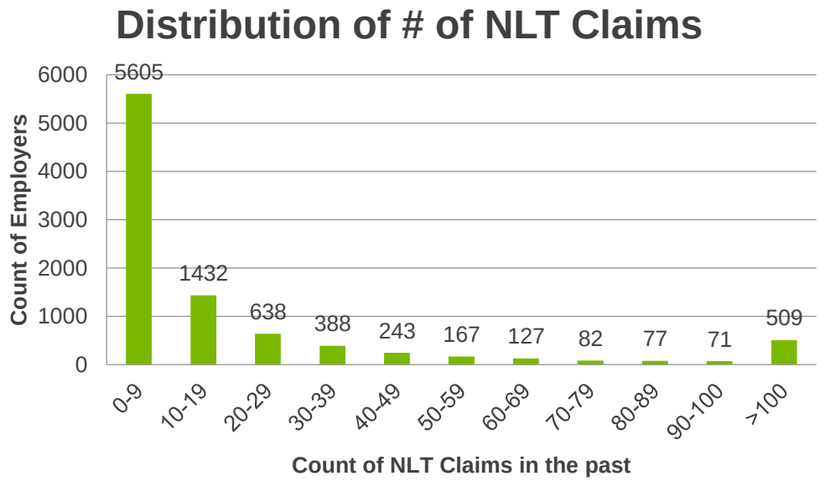
- # of Lost Time Claims in the past
  - When a worker suffers a work-related injury/disease which results in being off work and loss of wages/earnings, or a permanent disability/impairment
- # of Non-Lost Time Claims
  - When no time is lost from work but health care is needed

- *Looking at the distributions, employers tend have fewer LT claims than NLT claims, and most employers have very small amount of claims (<20)*

- # of Fatality Claims
  - 2% of the sample employers have had fatalities

- # of Claims that were on Benefits for 3 Months
  - 72% of the sample employers have had 0 claim on benefits for 3 months
- # of Claims that were on Benefits for 6 Months
  - 81% of the sample employers have had 0 claim on benefits for 6 months – risk tends to decrease overtime
- Allowance Rate
  - Percentage of claims where entitlement to benefits was authorized
  - The average allowance rate for the sample employers is 75%



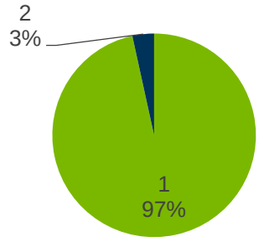


Final Dataset

12 Predictor Variables:

Categorical Variables

- Schedule
  - Schedule 1: employers for which the WSIB is liable to pay benefit compensation for workers' claims
  - Schedule 2: employers that self-insure the provisions of benefits



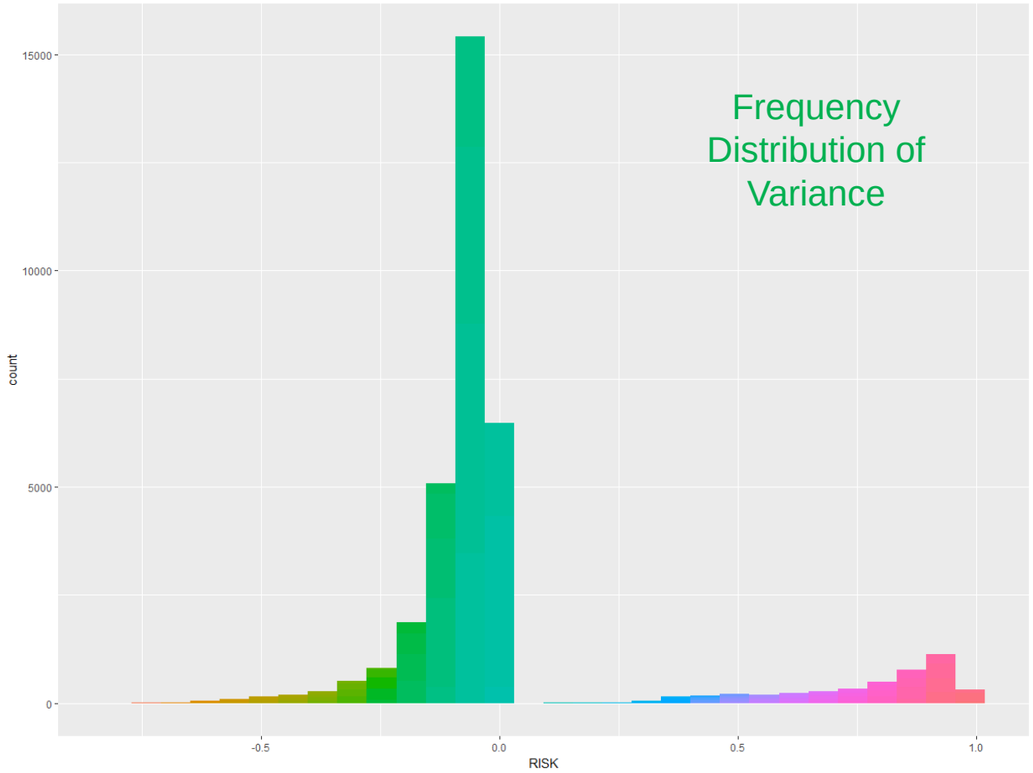
- Classification Unit
  - Used by the WSIB to classify the business activities of employers
  - 640 levels

- Organization City: 911 cities
- Organization Province: 37 provinces
- Organization Country: Canada and US

- Organization Start Year: 1937 to 2017

Target Variable with imbalanced classification:

- Class 1 (Less Risk): 8,137
- Class 2 (Same Risk): 22,837
- Class 3 (More Risk): 4,349



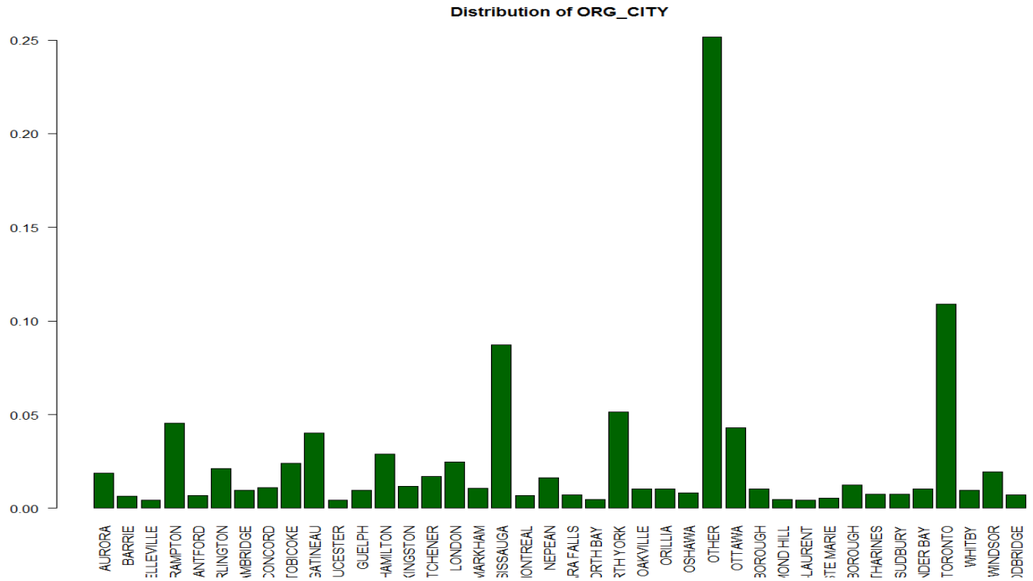
Categorical Variable Treatments

Thresholding Based on Frequency:

- The function “group_category” from library “Data Explorer” will group the sparse categories for a discrete feature based on a given threshold
  - group_category(data, feature, threshold, measure, update = FALSE, category_name = "OTHER“, …)
- For example, “ORG_CITY” had 911 levels but a random forest model cannot process a variable with > 50 levels
  - After thresholding, cities with a frequency in the bottom 25% are grouped to “other”; now we have 40 levels

Dummy Encoding:

- The function “sparse.model.matrix” from library “matrix” constructs a sparse model or “design” matrix
  - sparse.model.matrix(object, data = environment(object), contrasts.arg = NULL, ...)
  - Does dummy coding for categorical variables (transform each level into a binary variable) and removes the original variables
- XGBoost only accepts numeric features
  - After dummy coding, I had 1595 predictor variables



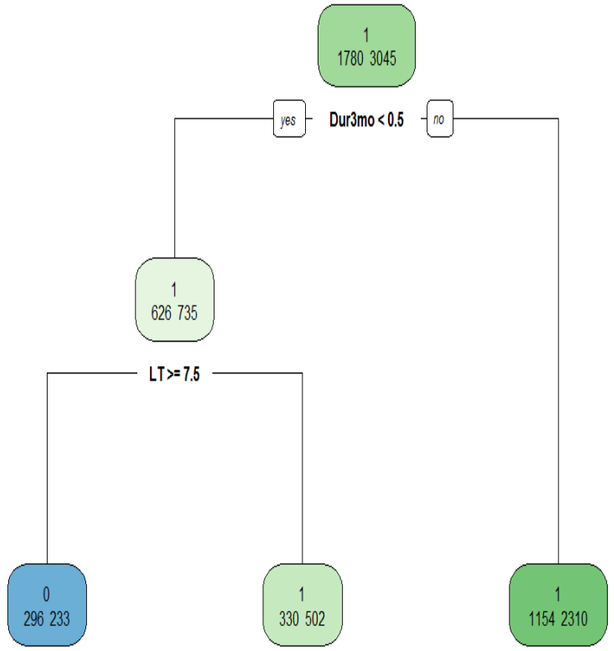
wsib ONTARIO

4. Modeling & Evaluation

wsib ONTARIO

Random Forest

Example of A Decision Tree:



- RF is a powerful **ensembling** algorithm - a type of supervised learning technique where multiple models are trained on a dataset and their individual outputs are combined by some rule to derive the final output.
- RF draws bootstrap samples, builds multiple **decision trees** and merges them together to get a more accurate and stable prediction.

- RF creates random samples and trains each sample independently to remove bias - instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features.
- RF reduces overfitting
- RF can be modelled for both regression and classification, and for categorical variables.

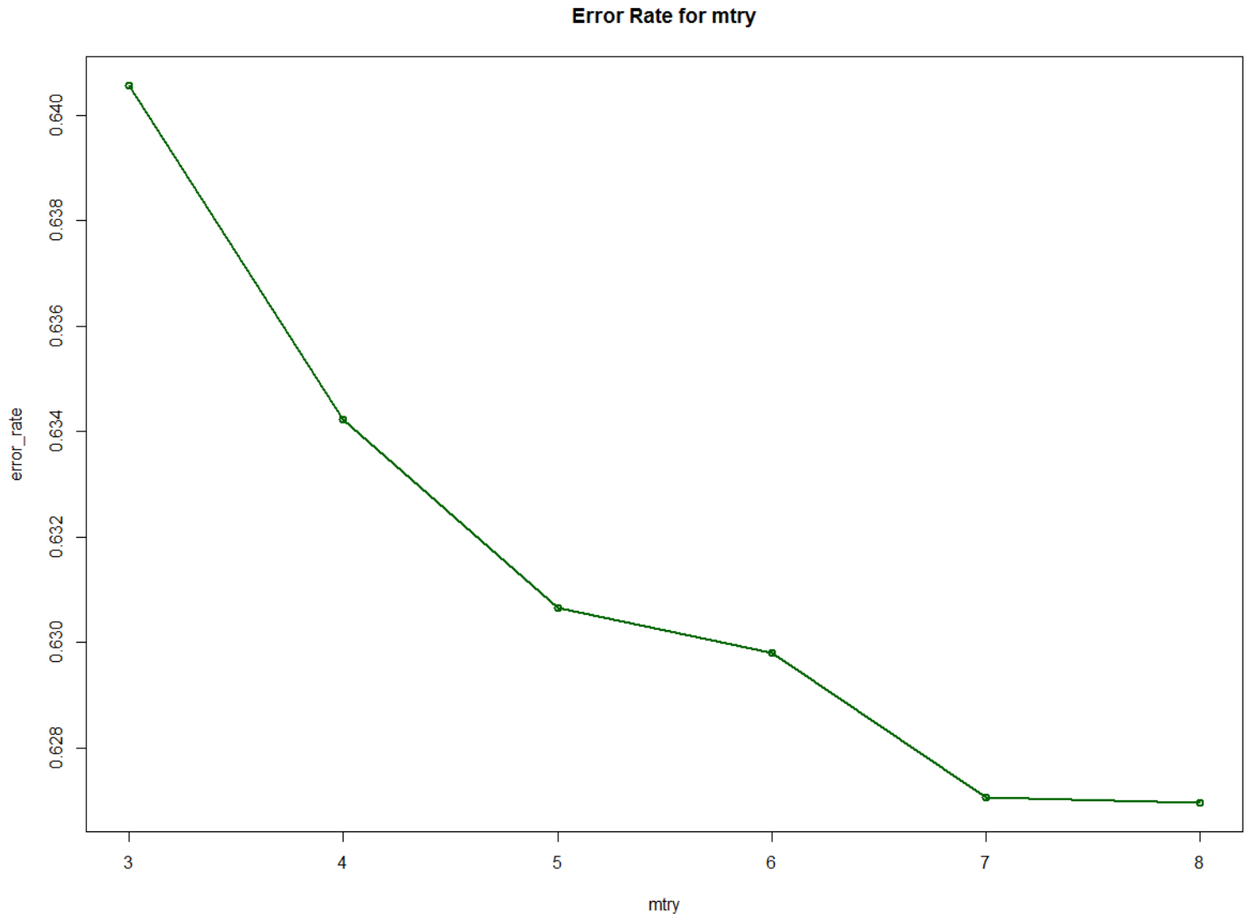
- A decision tree involves recursively partitioning the data into subsets that contain instances with homogenous values.
- Each internal node is labeled with an input feature and each leaf of the tree is labeled with a class.
- The goal is to find the attribute that returns the highest information gain (the purest branches) at each split.

Disadvantages:

- The larger the number of trees, the slower the model
- It cannot take categorical variables that have more than 50 levels

wsib ONTARIO

Random Forest - Modeling



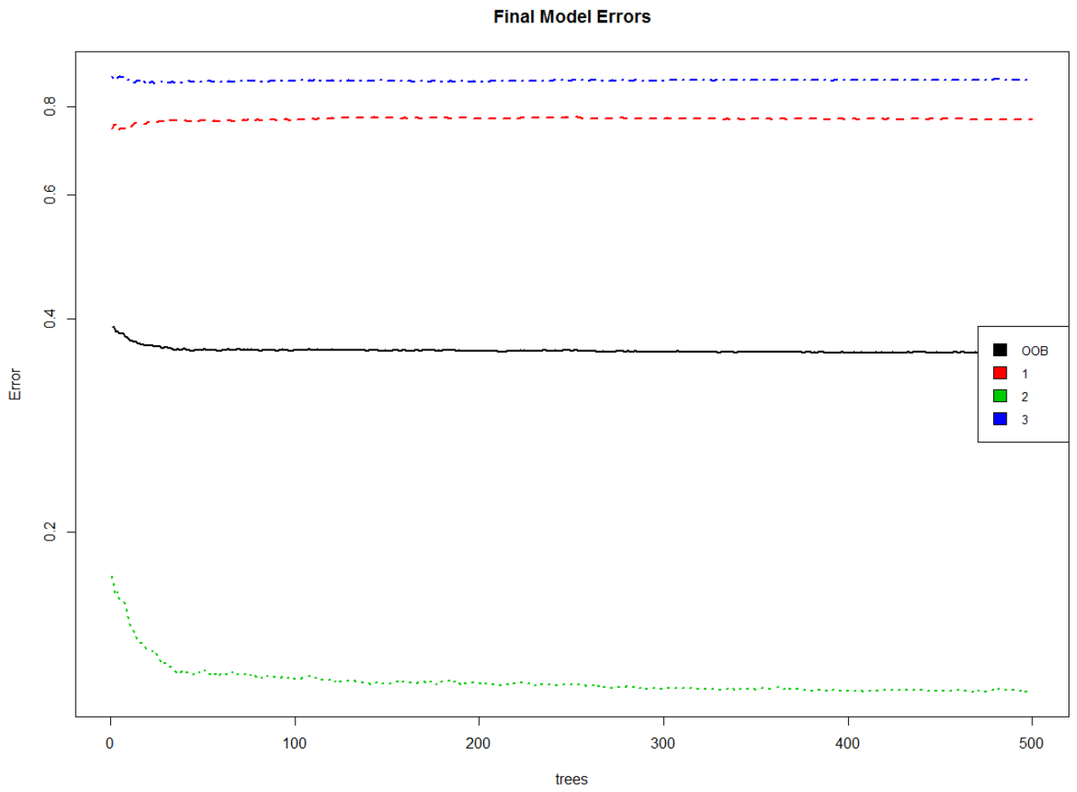
Parameter Highlights

- ntree: Number of trees to grow; a large number will ensure that every input row gets predicted at least a few times.
- mtry: Number of variables randomly sampled as candidates at each split.
- Used a “For” loop to identify the right mtry for model (very time-consuming to generate and plot).
- Mtry=8 produces the smallest error.

wsib ONTARIO

Random Forest - Modeling

```
#Build the model and predict
rf<-randomForest(RISK1 ~ . , data = train, nTree=500, mtry=8,
do.trace=100)
pred <-predict(rf,testx)
pred_prob = predict(rf, testx, type = "prob")
```



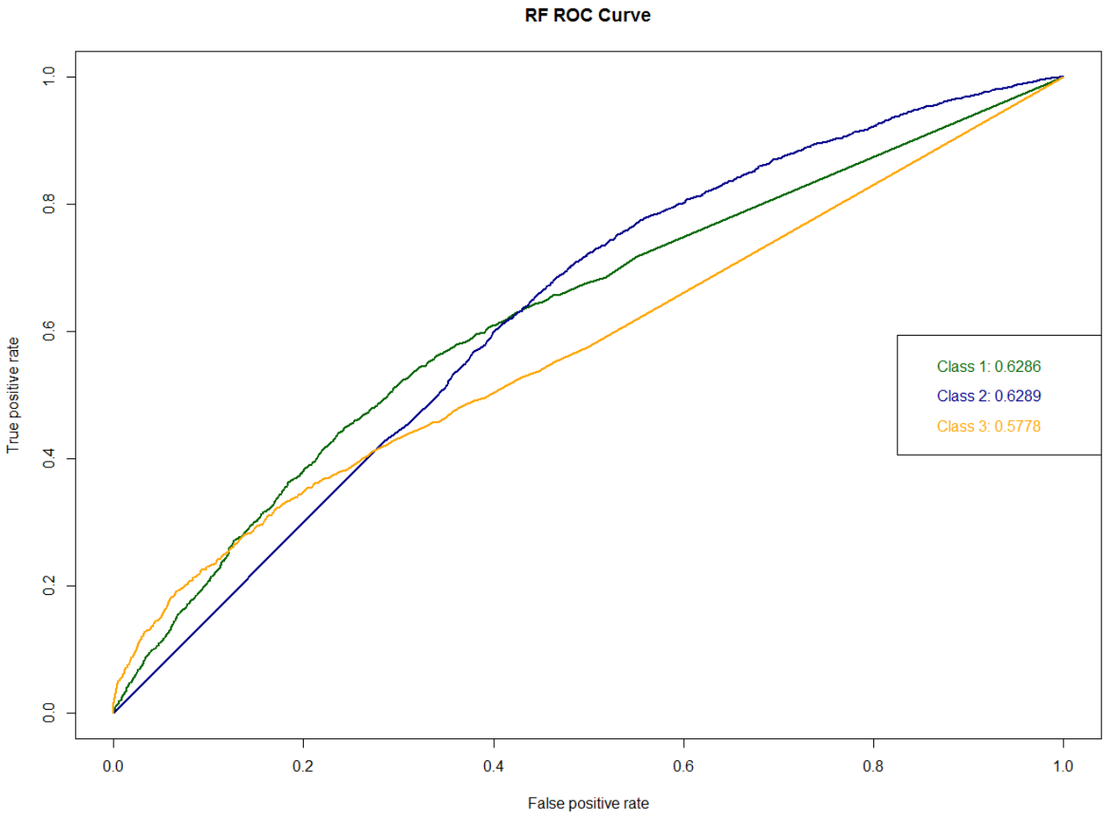
Out of Bag (OOB) Error

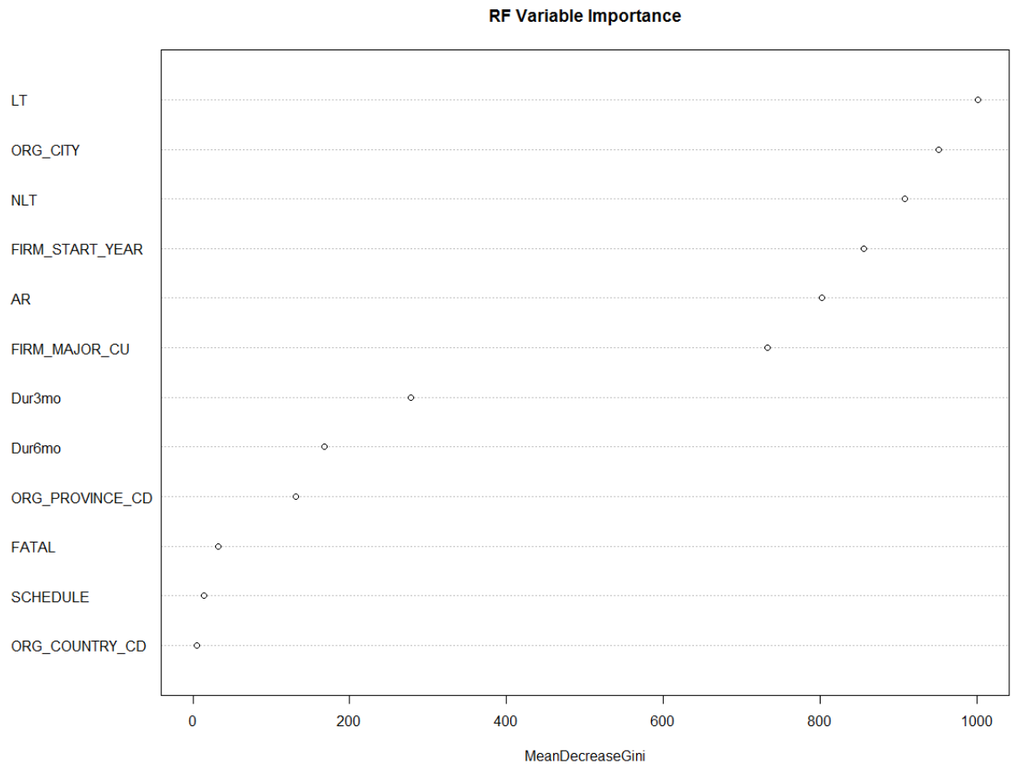
- Since each tree is constructed using a different bootstrap sample from the original data, about 1/3 of the cases are left out of the bootstrap sample.
- These cases for each tree is passed through that tree and the outputs are aggregated to produce the percentage error.
- Does not require further cross validation.

- Best number of trees is 500.

- Final model's OOB estimate: 35.89%
  - Class 1: 76.72%
  - Class 2: 11.95%
  - Class 3: 87.20%

Random Forest - Results





Interpretation	• In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (1-Specificity) for different cut-off points of a parameter. • For each class, the AUC is higher than the baseline, with Class 2 having the best performance. • In term of variable importance, *# of Lost Time Claims in the past, organization city, and the # of Non-Lost Time Claims* contribute the most to the model.

Extreme Gradient Boost

XGBoost	• One of the hottest libraries in supervised machine learning; designed for **superior speed and performance** • Supports various objective functions, including regression, classification, and ranking • XGBoost shines when we have lots of training data where the features are numeric or a mixture of numeric and categorical fields
Algorithm	• An implementation of **gradient boosting decision tree algorithm** • XGBoost build trees one at a time, where each new tree helps to correct errors made by previously trained tree. With each tree added, the model becomes even more expressive
Limitations	• Only takes numeric variables; categorical variables must be converted to numeric • More prone to overfitting than Random Forest; however, we can tune the parameters to adjust the model

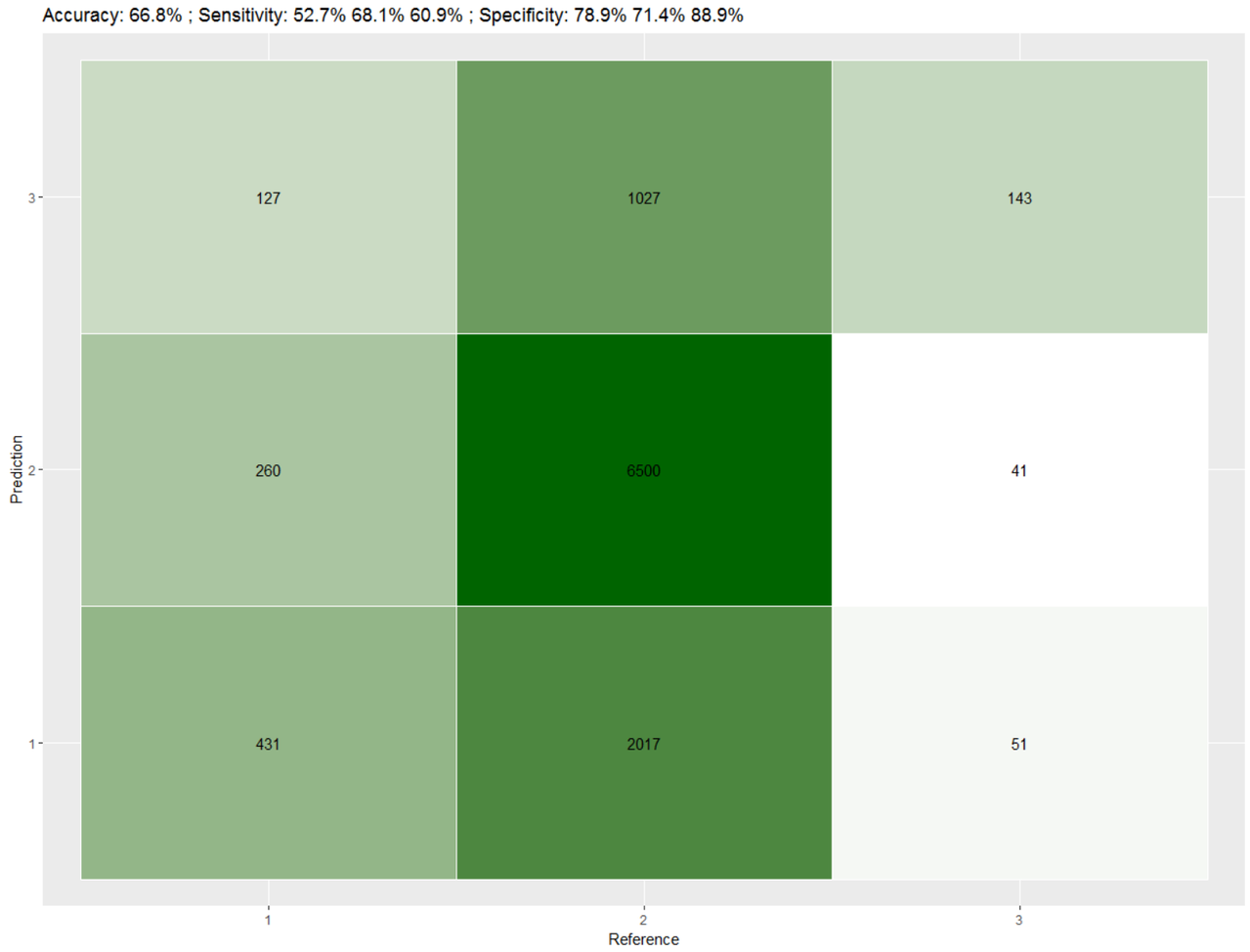
XGBoost - Modeling

```
#Final codes

params <- list(booster = "gbtree", objective = "multi:softprob",
num_class = 3, eval_metric="mlogloss", max_depth=4, etc = 0.1,
min_child_weight=8)

xgbcv <- xgb.cv(params = params, data = train_matrix, nfold = 5,
nrounds = 200, stratified = TRUE, print_every_n=20,
prediction=TRUE)


#Build the final model to predict

model <- xgb.train(params = params, data = train_matrix,
nrounds = 100)

pred <- predict(model, newdata = sparse.matrix.test)
```

Parameter Highlights

- Booster type: “gbtree” for classification; “gblinear” for regression
- The objective of “multi:softprob” leads to probabilistic classification and the evaluation metric “mlogloss” is the multi-class logloss function which should be minimized
- “Max_depth” controls the complexity of the model; the default is 6
- “Min_child_weight” blocks the potential feature interactions to prevent overfitting
- The “nrounds” parameter tells XGBoost how many trees to grow, which is tied to the learning rate “eta” which shrinks the feature weights to reach the best optimum
- The cross-validation function “xgb.cv” performs k-fold validation and is used for tuning the parameters

XGBoost - Results

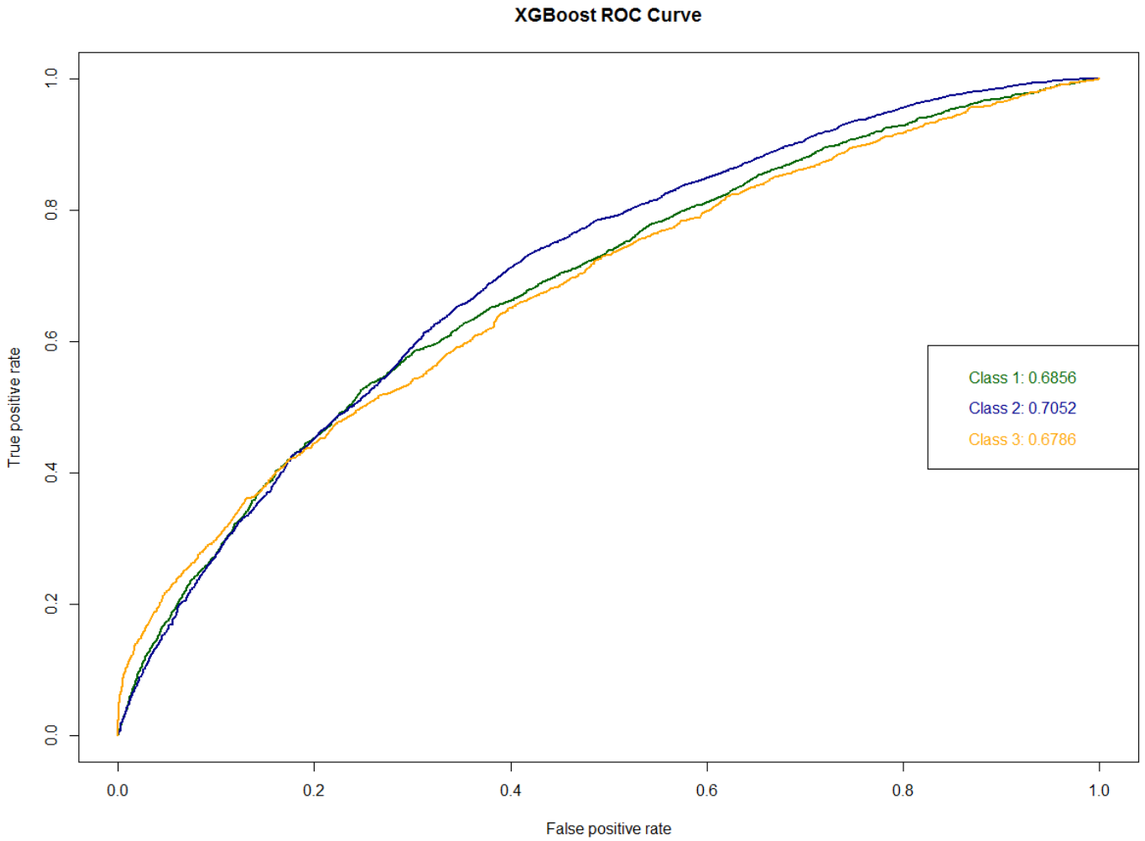


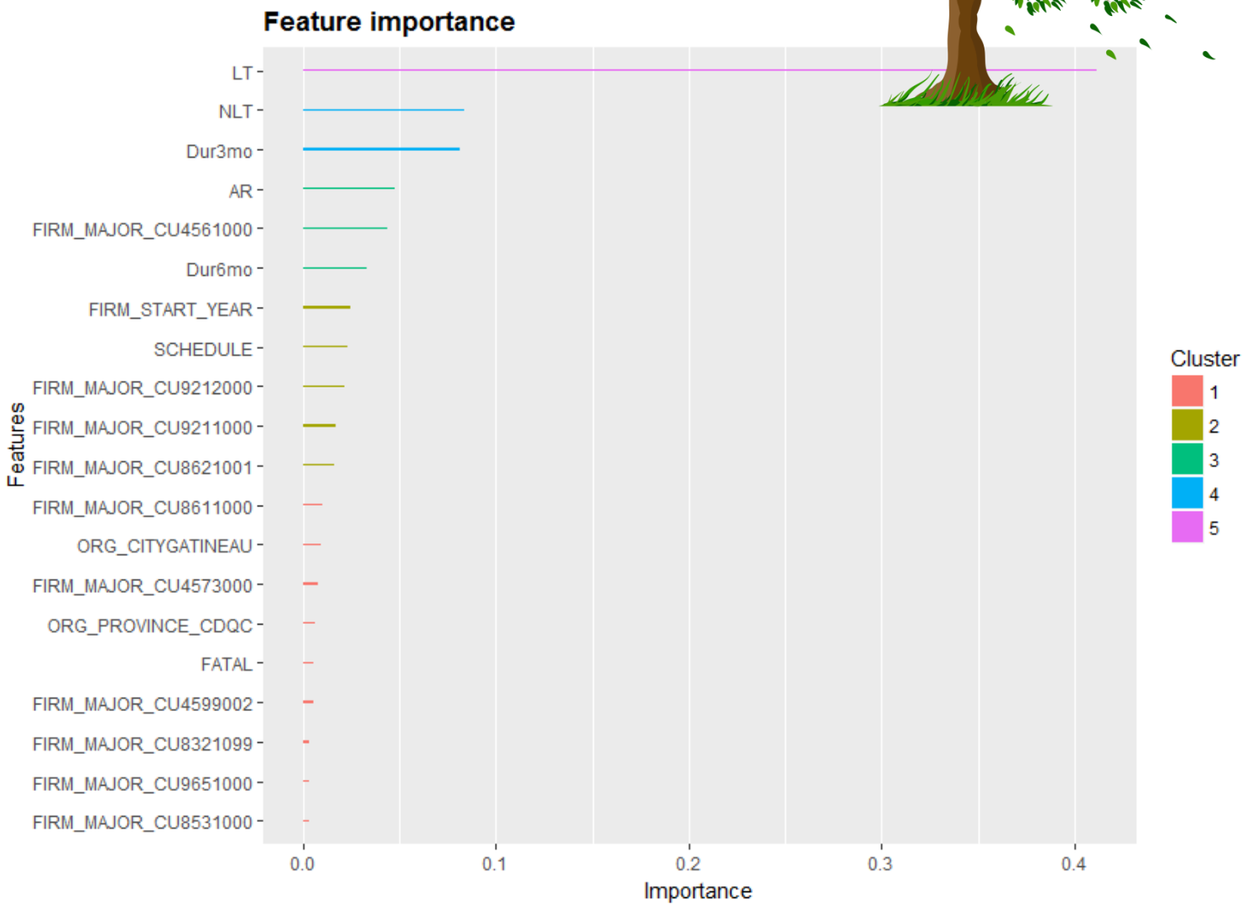
Library(ggplot)

Interpretation

- Sensitivity measures the proportion of actual positives that are correctly identified.
- The model is better at correctly classifying Class 2 (Same Risk) as positive.
- Using the employer features as predictor variables, the model predicts the variance in expected risk against actual outcome with an overall accuracy of 66.8% - reasonable in real world and helpful in adjusting the risk of a claim.

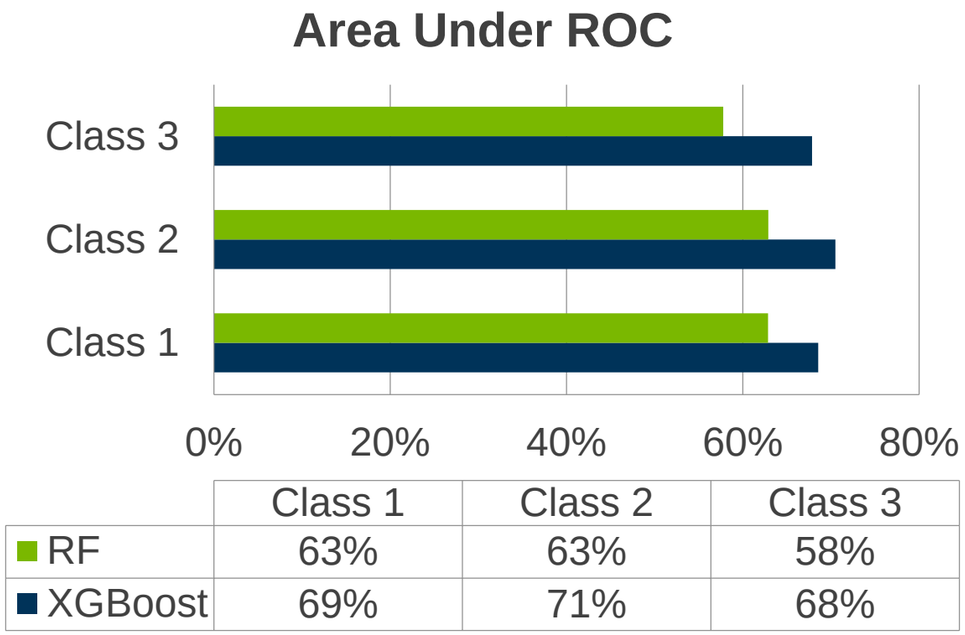
XGBoost - Results



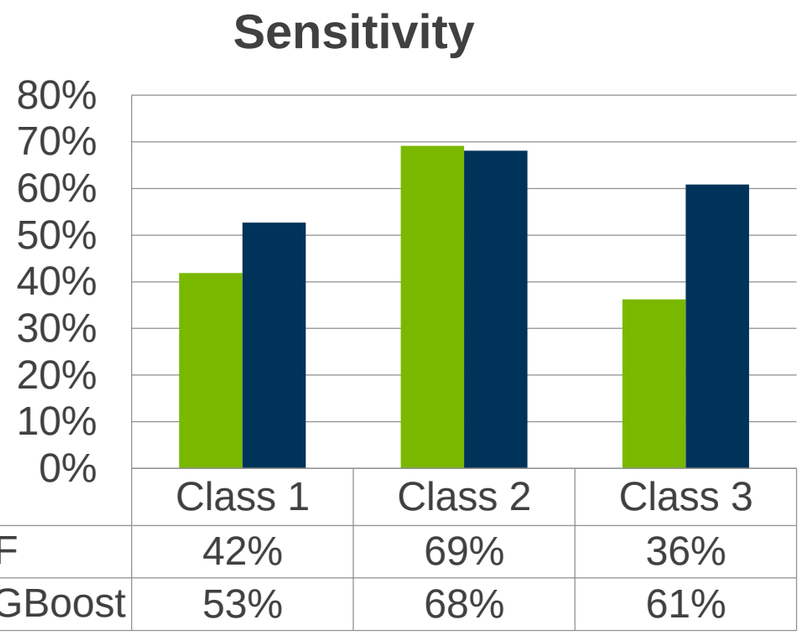


Interpretation	• For each class, the AUC is higher than the baseline, with Class 2 having the best performance (similar to RF model). • Features *# of Lost Time Claims, # of Non-Lost Time Claims*, and *# of Claims that were on Benefits for 3 Months* are the most important in classifying – these features should be given high regard.

Model Comparisons



	Class 1	Class 2	Class 3
RF	63%	63%	58%
XGBoost	69%	71%	68%



	Class 1	Class 2	Class 3
RF	42%	69%	36%
XGBoost	53%	68%	61%

Comments

- Better at predicting Class 2 – the expected risk of claim does not need to be adjusted by employer features.
- Class 1 and Class 3 have an AUC much higher than random classifier, which will serve as a reference in determining whether the risk of a registered claim should be increased or decreased due to employer features.
- XGBoost performs better than Random Forest both in successfully distinguishing positive and negative cases, and in correctly predicting positive cases for each class.

wsib ONTARIO

5. Future Steps

wsib ONTARIO

Considerations

Since we only used claims from February 2017, we will **increase the sample size** and **balance the target classes** to include more employers and claims, reducing bias and enhancing accuracy.

We will also look for more unique employer features to be added into the models, given that we have only used 13 variables (some of which are quite correlated- e.g. city and province).

Deployment: We will test the models on new data for 3 months before official deployment. The models will then be integrated with the reporting tool. The deployment now will not be in report, but directly into the guidewire system.

Monitoring: The models will be monitored systematically for performance, and we will receive feedback from case managers. Models will be modified and new models will be used if necessary.

The project is still in progress…

wsib ONTARIO

Discussion

wsib ONTARIO

Appendix

Link to My R Codes: https://gist.github.com/zhu701/63e680a95a92dffbac80a03f3630a183

References:

- Library DataExplorer: https://cran.r-project.org/web/packages/DataExplorer/DataExplorer.pdf
- Random Forest: https://www.r-bloggers.com/how-to-implement-random-forests-in-r/
- XGBoost:
  - https://rpubs.com/mharris/multiclass_xgboost
  - https://datascienceplus.com/extreme-gradient-boosting-with-r/
  - https://cran.r-project.org/web/packages/xgboost/vignettes/discoverYourData.html
  - https://rpubs.com/hariteja/xgb
- R Resources

wsib ONTARIO