Prediction of Workplace Injuries

Mehdi Sadeqi*, Azin Asgarian and Ariel Sibilia

Cority Inc in collaboration with Georgian Partners Inc

Toronto, Ontario, Canada

Mehdi.Sadeqi@cority.com

Problem Description

We collected employees' safety-related information from different organizations during years 2016-2017. We treat the learning problem as a binary classification task. Using the data collected during 2016, the objective is to predict whether an employee was injured or not in 2017. Although the collected datasets differ in size and distribution, they are all highly imbalanced (1-7% injury cases).

In all datasets, the employee records are represented by 38 engineered features that capture two main groups of information: general employee information (*e.g.* age), and event-based information. Event-based information are either associated with the employee (*e.g.* number of absences) or with the employee's site (*e.g.* the risk assessments scores). In this work, we use XGBoost as our base predictive model.

Handling Imbalanced Data

We used four methods that combine ensemble-based supervised learning algorithms with re-sampling methods (to rebalance the class distribution in each bag of the bagging or in each iteration of training weak learners of the boosting):

1. UnderBagging: Random under-sampling combined with bagging
2. SMOTEBagging: SMOTE or over-sampling combined with bagging
3. RUSBoost: Random under-sampling combined with AdaBoost.M2
4. SMOTEBoost: SMOTE or over-sampling combined with AdaBoost.M2

Transfer Learning

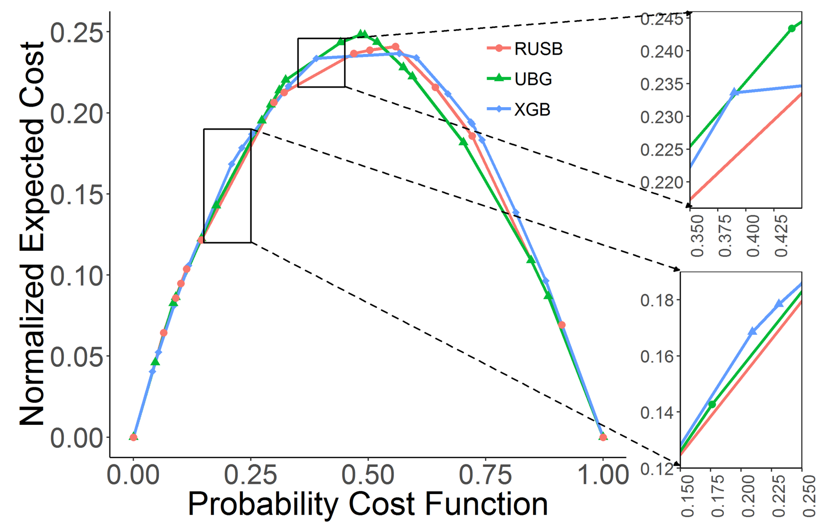
We then employed instance-based transfer learning method to control overall relative importance between source and target samples and to transfer knowledge learned from one organizations (source domain) to a new organization (target domain).

Baseline Models Source ($\mathcal{A}_S$), target ($\mathcal{A}_T$), union of source and target ($\mathcal{A}_{S\cup T}$), all the weights set to 1 ($\mathcal{A}_1$), and evaluate source sample weights assuming Gaussian distribution for target and source samples ($\mathcal{A}_G$)

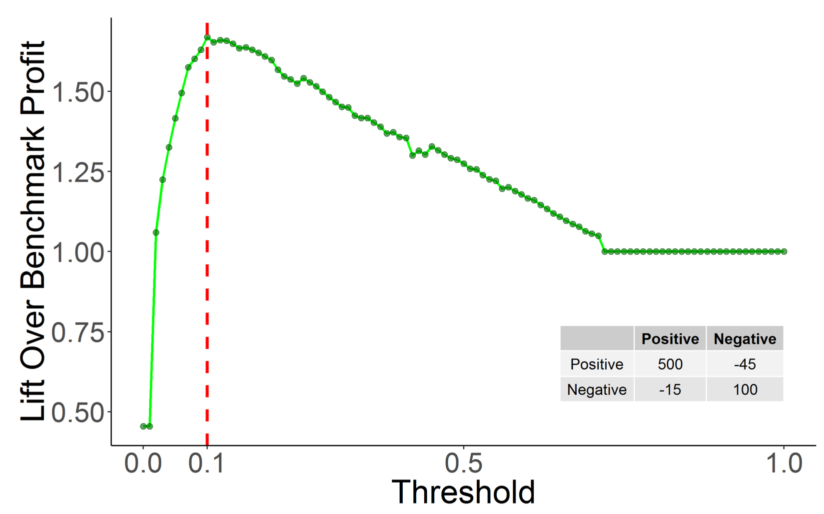
Hybrid Weights Combine similarity of source samples to target samples (measured by training a logistic regression binary classifier) with relevance of source samples to the target task (by using the distance of sample x to the decision boundary of an XGBoost binary classifier trained on all source and target samples), i.e., $w_x = w_{domain_x} + w_{task_x}$

Ensemble-Based Resampling Methods Results

Cost-Curve Evaluation We used these curves to evaluate and compare classifiers in deployment conditions of two important and usually unknown or time-varying factors: class distributions and misclassification costs. RUSBoost and UnderBagging showed a better performance than the XGBoost model in handling class imbalance.



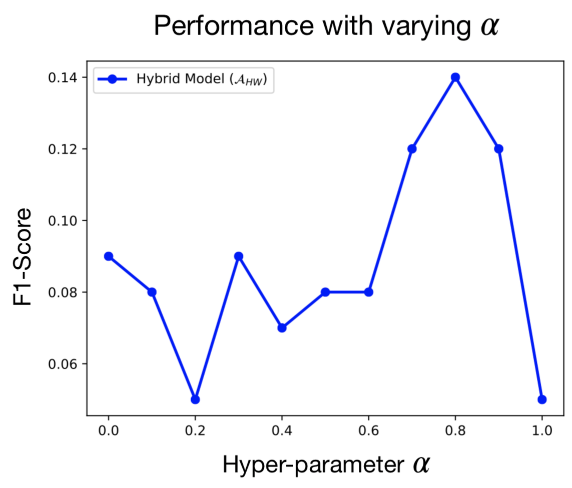
Cost-Sensitive Learning In the presence of a cost matrix, we can find the optimum threshold that maximizes the corresponding cost function. In our XGBoost model, this threshold will be another hyper-parameter that should be optimized inside a cross-validation pipeline.



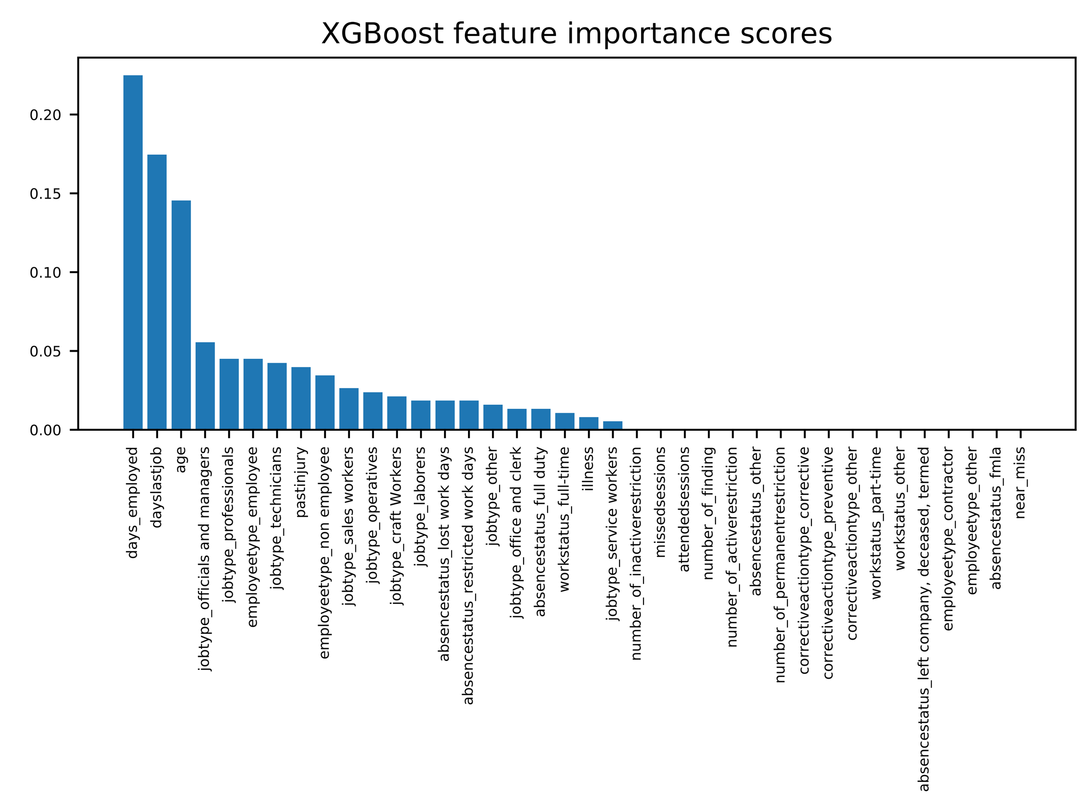
Instance-Based Transfer Learning Results

Method	Precision	Recall	F_1-score	AUCPR
$\mathcal{A}_T$	0.07	0.06	0.06	0.0375
$\mathcal{A}_S$	0.04	0.18	0.07	0.0405
$\mathcal{A}_{S\cup T}$	0.13	0.06	0.08	0.0478
$\mathcal{A}_1$	0.06	0.12	0.08	0.0456
$\mathcal{A}_G$	0.07	0.16	0.10	0.0532
$\mathcal{A}_{HW}$	0.11	0.12	**0.12**	**0.0542**

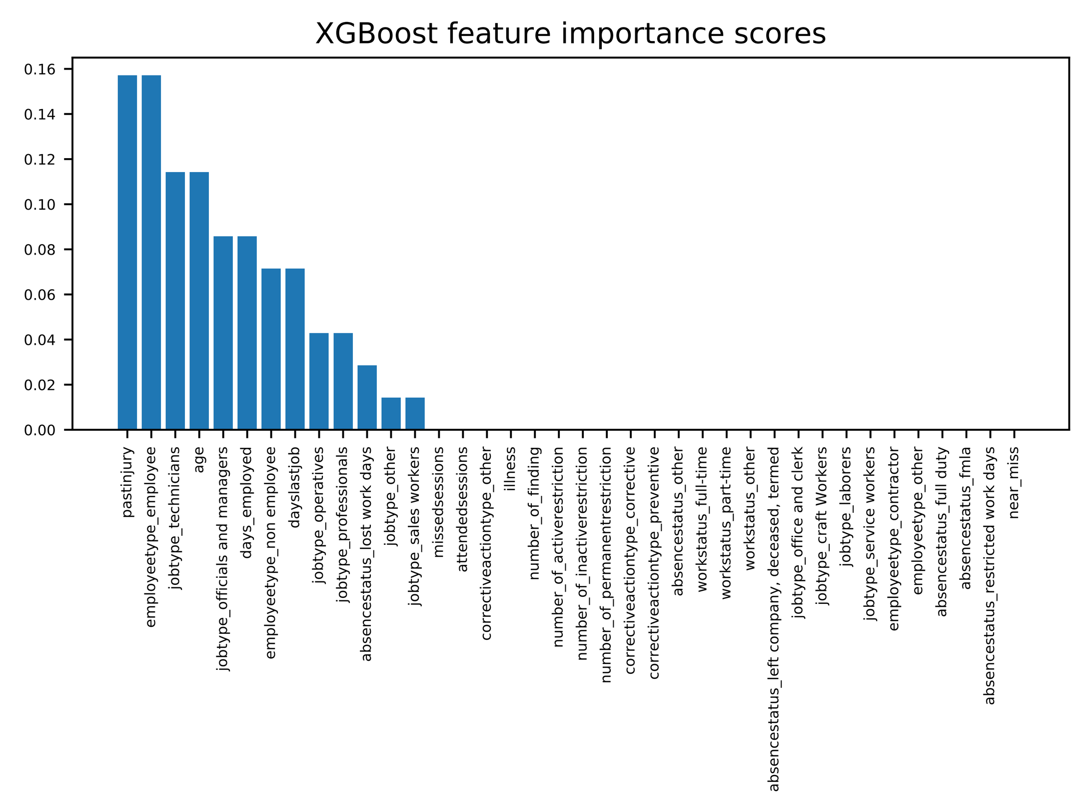
We used a total of 58,271 samples (12,225 and 46,046 from target and source organizations training sets, respectively) and evaluated the models on 3,057 samples from target test set. $\mathcal{A}_{HW}$ considerably improved model performance. The best performance is achieved with $\alpha = 0.7$.



Features importance scores for model $\mathcal{A}_T$

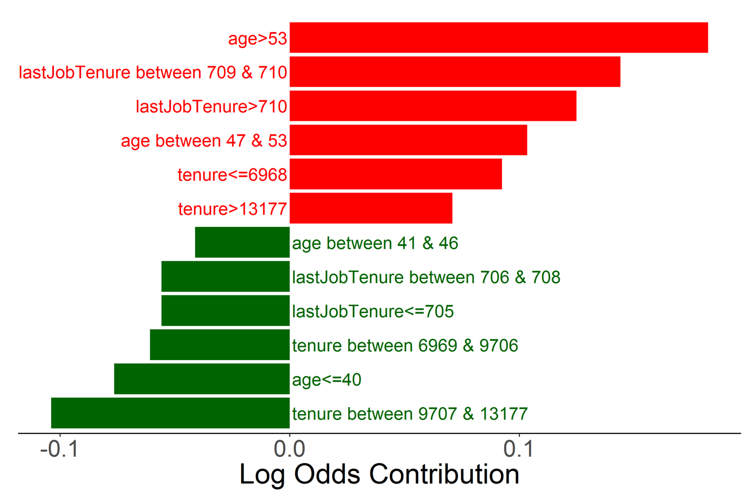


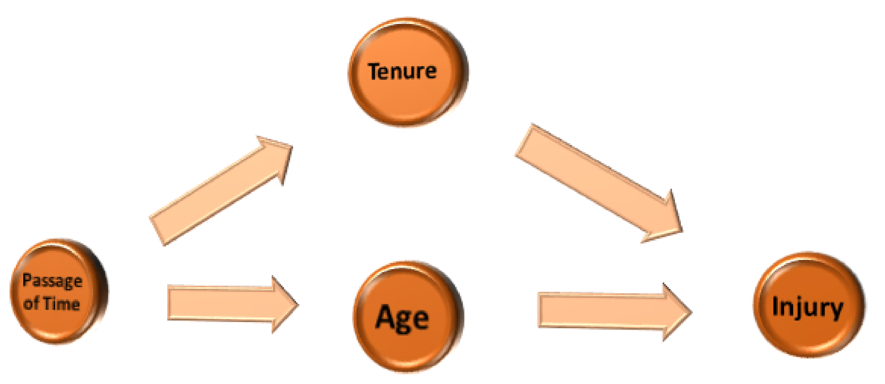
Features importance scores for model $\mathcal{A}_{HW}$



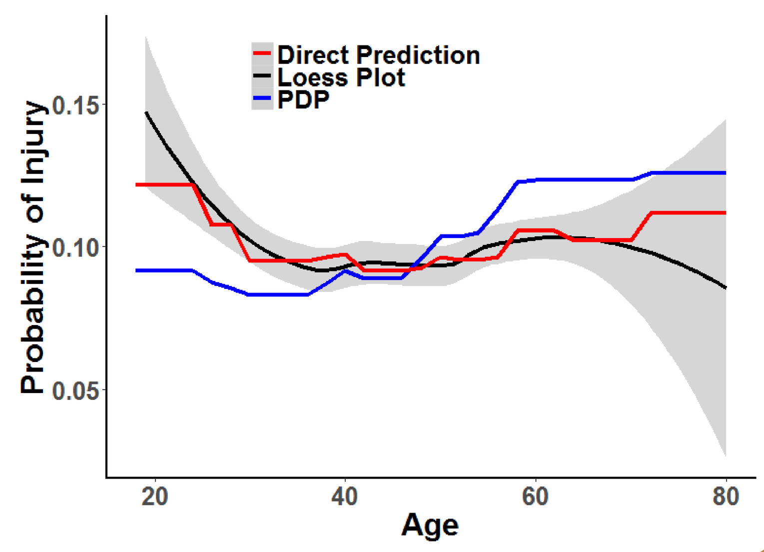
Actionable Insights

Measuring Association First, we find the average log-odds contribution of each feature to each sample. Next, for each discrete value of each feature (continuous variables are binned and then treated as categorical), we average the sample-based contribution over all samples with matching values.





Causal Relationship Partial dependence plots (PDPs) show the average relationship between two (or more) variables over a population by marginalizing over the distribution of all other variables. Partial dependence calculation that averages over a set of variables is equivalent to controlling for those variables using Pearl's back-door adjustment formula.



Conclusions and Future Research

We investigated the problem of injury risk prediction in a supervised learning framework.

Model Creation and Improvement To improve on our baseline XGBoost model with highly imbalanced data, we employed Ensemble-Based Resampling methods and Transfer Learning.

Model Interpretability and Causal Inference We used average log-odds contribution of each feature to measure associations and Partial Dependence Plots along with Back-Door Path Criterion to determine the causal effect of a feature on the risk of injury.

Causal Inference and Observational Data In general, measuring causal effect of a given variable in an observational study is a challenging task and will be our future research direction.