Prediction of Workplace Injuries

Mehdi Sadeqi [1] **Azin Asgarian** [2] **Ariel Sibilia** [1]

Abstract

Workplace injuries result in substantial human and financial losses. As reported by the International Labour Organization (ILO), there are more than 374 million work-related injuries reported every year. In this study, we investigate the problem of injury risk prediction and prevention in a work environment. While injuries represent a significant number across all organizations, they are rare events within a single organization. Hence, collecting a sufficiently large dataset from a single organization is extremely difficult. In addition, the collected datasets are often highly imbalanced which increases the problem difficulty. Finally, risk predictions need to provide additional context for injuries to be prevented. We propose and evaluate the following for a complete solution: 1) several ensemble-based resampling methods to address the class imbalance issues, 2) a novel transfer learning approach to transfer the knowledge across organizations, and 3) various techniques to uncover the association and causal effect of different variables on injury risk, while controlling for relevant confounding factors.

1. Introduction

Workplace injuries can affect workers' lives and can cause substantial economic burden to employees, employers, and more generally to society (ILO, 2018; Sarkar et al., 2016). There are more than 374 million work-related injuries reported every year, resulting in more than 2.3 million deaths annually (ILO, 2018). The yearly cost to the global economy from work-related injuries alone is a staggering $3 trillion, estimated by ILO.

Predicting injuries and providing actionable insights on factors associated with injuries are critical for improving workplace safety. Recent research has focused on this problem in sports (Naglah et al., 2018; Rossi et al., 2018), construction (Tixier et al., 2016; Poh et al., 2018), and various workplace settings (Sánchez et al., 2011; Rivas et al., 2011; Sarkar et al., 2016). Despite introducing many interesting frameworks, these studies do not address some of the main challenges such as lack of labeled data and class imbalance issues. In addition, previous works do not investigate the causal relationships between different variables and injury incidents.

We propose a framework that employs ensemble-based resampling methods and a novel transfer learning approach to address class imbalance and data availability issues. We apply a method to predictive features of injuries to highlight their direct causal effect and we utilize a visualization technique that provides interpretability. We demonstrate the utility of our framework through experiments performed on real-world datasets. More specifically, we show that ensemble-based resampling and transfer learning techniques can increase the F_1-score by 100% and area under precision recall curve by 44%, when compared to a model trained on a single organization dataset.

In the remainder of this paper, we first provide a brief overview of the problem and our machine learning framework in Section 2.1. We present the employed ensemble-based resampling techniques, our instance-based transfer learning method, and the approaches used to provide actionable insights in Sections 2.2, 2.3, and 2.4, respectively. Section 3 describes our results and Section 4 covers conclusions and future work.

2. Injury Prediction as Supervised Learning

2.1. Data and Problem Description

To conduct this study, 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. The collected datasets differ in size and distribution, however, 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 infor-

[1]Cority Inc, Toronto, Ontario, Canada [2]Georgian Partners Inc, Toronto, Ontario, Canada. Correspondence to: Mehdi Sadeqi <Mehdi.Sadeqi@cority.com>.

Appearing at the International Conference on Machine Learning AI for Social Good Workshop, Long Beach, United States, 2019.

mation are either associated with the employee (*e.g.* number of absences) or with the employee's site (*e.g.* the risk assessments scores).[1] In this work, we use XGBoost (Friedman, 2001b) as our base predictive model.

2.2. Imbalanced Data

To address the problem of highly imbalanced data, several approaches are proposed in the literature. Among the most common ones are over-sampling and under-sampling methods (Chawla, 2003), neighbor-based techniques (Wilson, 1972; Tomek, 1976), Synthetic Minority Over-sampling TEchnique (SMOTE) (Chawla et al., 2002), adjusting class weights, boosting techniques, and anomaly detection methods. From these solutions, we are particularly interested in four methods that combine ensemble-based supervised learning algorithms with resampling methods (*UnderBagging*, *SMOTEBagging*, *RUSBoost*, and *SMOTEBoost*). We give a brief overview of these methods below.

UnderBagging and SMOTEBagging methods try to rebalance the class distribution in each bag of the bagging algorithms. UnderBagging uses random under-sampling while SMOTEBagging uses SMOTE or over-sampling to achieve this goal (Galar et al., 2012). Alternatively, RUSBoost (Seiffert et al., 2010) and SMOTEBoost (Chawla et al., 2003) combine AdaBoost.M2 (Freund & Schapire, 1997) boosting algorithms with resampling methods to address the class imbalance issues. Similar to UnderBagging and SMOTEBagging, in each iteration of training weak learners, these two algorithms respectively use random under-sampling (to reduce majority instances) and SMOTE (to increase minority instances). Moreover, these four approaches have the advantage of having very few number of hyper-parameters. We provide a comparison of these methods in Section 3.1.

2.3. Transfer Learning

To handle the data unavailability issues for a new organization (target domain), we leverage the knowledge learned from other organizations (source domain) by employing an instance-based transfer learning method. Given a set of target training samples $\{(x_i, y_i)|i \in \{1, 2, ..., N_T\}\}$ and a loss function $\mathcal{L}(.)$ the goal of supervised learning is to find model $\mathcal{A}^*$ that minimizes the expected error, *i.e.*, $\mathcal{A}^* = \arg\min_{\mathcal{A}\in\mathbb{A}} \mathbb{E}_{x\sim P_T}\big[\mathcal{L}(\mathcal{A}(x), y)\big]$. Here x is an arbitrary sample and P_T is the probability distribution of target samples. Following the idea of importance sampling (Liu, 2008; Asgarian et al., 2018) for transferring the knowledge from source domain (S) to target domain (T), we can express the expected error as $\alpha\mathbb{E}_{x\sim P_T}\Big[\epsilon(x)\Big] + (1-\alpha)\mathbb{E}_{x\sim P_S}[\epsilon(x)\frac{P_T(x)}{P_S(x)}]$. Here $\epsilon(x)$ shows the error for each sample and α is a hyper-parameter that controls the overall relative importance between source and target samples. Source sample weights $\{w_{x_j} = \frac{P_T(x_j)}{P_S(x_j)} \mid j \in \{1, ..., N_S\}\}$ play a major role in instance-based transfer learning methods, as they control the individual effect of source samples (Asgarian et al., 2017). We describe different weighting approaches including our five baselines models and our proposed weighting strategy in the following.

Baselines: Models $\mathcal{A}_S$, $\mathcal{A}_T$, and $\mathcal{A}_{S\cup T}$ trained respectively on source, target and the union of source and target, serve as minimum baselines that a transfer learning method must outperform. Our fourth baseline model is an instance-weighted model ($\mathcal{A}_\mathbf{1}$) with all the weights set to $\mathbf{1}$ (*i.e.*, $\mathbf{W}_S = \mathbb{1}$). This is similar to $\mathcal{A}_{S\cup T}$, except in this model we use α to determine the relative overall importance between source and target samples. Our last baseline model ($\mathcal{A}_G$), assumes Gaussian distributions for target and source samples to evaluate the source sample weights w_{x_j}.

Hybrid Weights: Previous methods evaluate the source sample weights solely based on their similarity to the target domain. We argue that it is also important to measure the relevance of source samples to the target task. Hence, we define weights $w_x = w_{domain_x} + w_{task_x}$, where w_{domain_x} measures the similarity of an arbitrary source sample x to the target domain, while w_{task_x} measures the importance of sample x in the target task.

For evaluating w_{domain_x}, unlike the previous methods that employ generative approaches to estimate P_T and P_S, we directly approximate weights $w_{domain_x} = \frac{P_T(x)}{P_S(x)}$ with a discriminative classifier. More specifically, using $\{(x, l_x) \,|x \in S \cup T$, and $l_x = 1$ if $x \in S$, $l_x = 0$ otherwise$\}$, we train a binary classifier (e.g., logistic regression (LR)) to differentiate source and target samples. Next, we use the learned weights of this classifier (w_{lr} and c_{lr}) to estimate source sample weights $w_{domain_x} = \frac{P_T(x)}{P_S(x)} \approx \frac{1}{\exp(x^T w_{lr} + c_{lr})}$.

To compute w_{task_x}, we train an instance of our predictive model (XGBoost) using all samples from source and target with their corresponding injury labels. We then define w_{task_x} to be the uncertainty of this model about sample x. We define the uncertainty of a model about sample x to be the distance of x to the decision boundary. Note that this value could be negative (thus subtracting from w_x) when the decision is incorrect, or positive (thus adding to w_x) when the decision is correct. We denote this model as $\mathcal{A}_{HW}$.

2.4. Actionable Insights

To visualize the relationship between injuries and its predictors, we explore one method to show the straightforward association, and one to find causal relationship.[2]

[1] The names of the organizations are masked due to confidentiality reasons.

[2] Please note that in order to infer meaningful and accurate causal relationships using this approach, we need adequately accu-

To measure the association between features and the target variable, we average the contribution of each feature's possible values to the log-odds ratio of all samples which match that value. We bin every continuous variable, treating them as categorical, so that such matching is possible for all variables. We use the *xgboost_explainer* package in Python, which is inspired by (Foster), to find the average log-odds contribution of each feature to each sample. Next, for each discrete value of each feature, we average the sample-based contribution over all samples with matching values. This gives us a visualization of both the average impact and direction of each variable as seen in Figure 1. This is a unique approach for visualizing the association between safety-related variables and injuries and it also gives an interpretation of the model. However, it must be noted that each effect size measured here is an average over the population, not a deterministic effect.

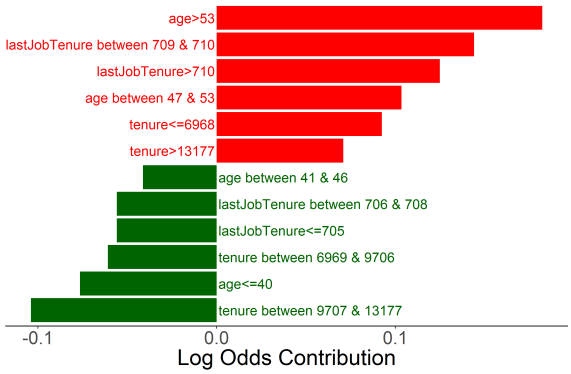


Figure 1: Log-odds contribution of binned continuous variables for a hypothetical example company.

Partial dependence plots (PDPs) (Friedman, 2001a) are an especially interesting and fairly natural way of adding interpretability to more complex models. PDPs show the average relationship between two variables over a population by marginalizing over the distribution of all other variables. For a trained model, this is approximated by summing over the training data, where, unlike the marginalized variables, the variables to be plotted are held constant (Friedman, 2001a). PDPs may also yield a causal interpretation of the effect of a variable on injuries (Zhao & Hastie, 2017). Zhao & Hastie showed that a partial dependence calculation that averages over a set of variables is equivalent to controlling for those variables using Pearl's back-door adjustment formula (Pearl, 1993). For example, there are instances where the relationship between an input variable and injuries is reversed when doing a partial dependence calculation over another variable. In (Pearl, 2014), Pearl shows that instances of the Simpson's paradox can be properly explained when using the back-door criterion to adjust for a variable.

rate models. The ensemble-based resampling methods and transfer learning are an attempt in this direction.

In our example, we observe the effect of age on probability of injuries after adjusting for tenure. In our causal hypothesis, both age and tenure have a causal impact on injuries. However, neither are considered to directly cause each other. Nevertheless, they are strongly correlated, so we connect each of their exogenous variables in a causal graph. Intuitively, this can be described as both age and tenure being caused by the "passage of time". In our causal path from age to injury there is one back-door path, which is blocked by tenure. Therefore, tenure satisfies the back-door criterion from age to injury. In Figure 2, we plot probability of injury versus age generated in three different ways. The direct prediction and loess (locally estimated scatterplot smoothing) plot both estimate the direct association between age and probability of injury. The partial dependence plot shows the same association, after adjusting for tenure.

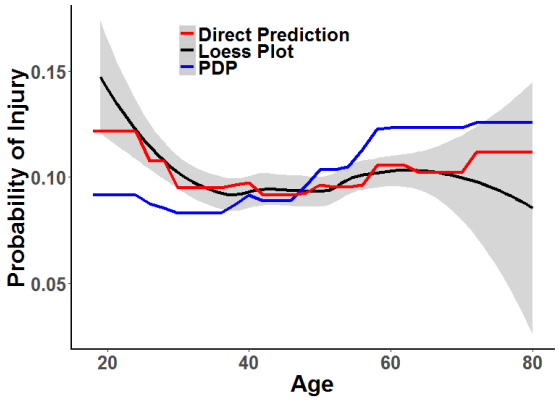


Figure 2: The loss curve and model prediction show the direct association between age and injury. The PDP curve shows the same relationship when controlling for tenure.

3. Experiments

3.1. Imbalanced Data

For comparing the ensemble-based resampling methods to our benchmark XGBoost model, we adopted an effective visualization technique called cost-curves (Drummond & Holte, 2006). These curves allow us to evaluate a classifier in deployment conditions of two important factors–*class distributions* and *misclassification costs*–which are usually unknown or varying with time. Using cost-curves, we can visualize a classifier's performance for the whole range of these unknown factors. Receiver operating characteristic (ROC) Curves are point/line dual with cost-curves and they convey the same information implicitly, but they are not visually as informative. Hence, we used cost-curves in addition to our other evaluation metrics. These curves also helps to find the conditions for which a classifier shows better performance compared to other classifiers and particularly to trivial classifiers (Drummond & Holte, 2006).

In all datasets, RUSBoost and UnderBagging showed a bet-

ter performance than the XGBoost model in handling class imbalance (Figure 3). SMOTEBagging and SMOTEBoost, however, showed a lower performance compared to the XGBoost. To avoid a cluttered plot, they are not shown in Figure 3. That being said, this is a data-dependent behavior and one should test each of these algorithms to see which one best matches the data.

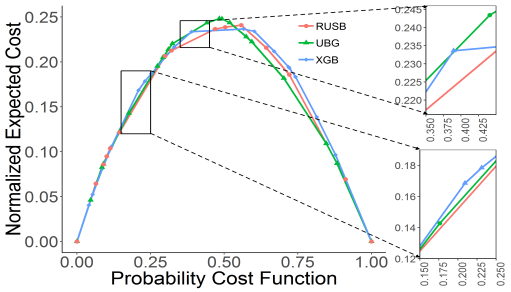


Figure 3: Cost-curves of RUSBoost and UnderBagging versus XGBoost.

The cost-curve performance comparison and model selection is only needed if misclassification costs are unknown in advance or deployment class distribution is different from the test data class distribution. If this is not the case, we can find the optimum threshold that maximizes a profit function defined by a profit matrix. This threshold will be another hyper-parameter that should be optimized inside a cross-validation pipeline. In our XGBoost model, we used the profit function as the evaluation metric for *early stopping* for 100 different thresholds. For simplicity, here we kept all other XGBoost parameters fixed and optimized only for threshold. Figure 4 shows the ratio of model profit to a benchmark profit as a function of threshold values. The profit matrix and the optimum threshold value 0.1 are shown in this figure.

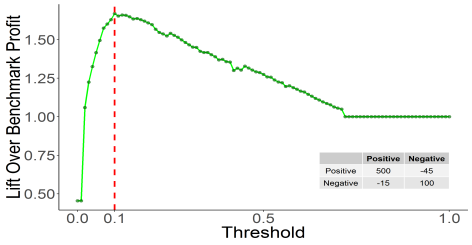


Figure 4: Optimum threshold for a given profit matrix.

3.2. Transfer Learning

In our transfer learning framework, we considered Organization-1's dataset as the target domain and Organization-2's dataset as the source domain. For training, we used 58,271 samples from target (12,225) and source (46,046) training sets, and evaluated the models on 3,057 samples from target test set. Since the datasets were highly imbalanced (1-7% injury cases), we used precision, recall, F_1-score (macro), and area under the precision-recall curve (AUCPR) as our four evaluation metrics. Results of our quantitative evaluation is shown in Table 1.

In Table 1, we see that model $\mathcal{A}_T$ has a poor performance with F_1-score equal to 0.06 and AUCPR of 0.0375. This is possibly due to data sparsity issue and lack of expressiveness of the model. On the other hand, $\mathcal{A}_S$ has a higher F_1-score and AUCPR, but the precision is diminished. Also model $\mathcal{A}_{S\cup T}$ performs better in terms of F_1-score and AUCPR compared to both models $\mathcal{A}_T$ and $\mathcal{A}_S$. The best result is obtained with our model $\mathcal{A}_{HW}$, which increases the F_1-score and AUCPR considerably.

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
$\mathcal{A}_\mathbf{1}$	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**

Table 1: Performance of different methods on Company-1's data.

Figure 5 shows the AUCPR obtained with model $\mathcal{A}_{HW}$ as a function of hyper-parameter α. We see that the best performance is achieved with α equal to 0.7. However, increasing or decreasing α results in lower AUCPR, as it enhances the influence of target or source samples respectively.

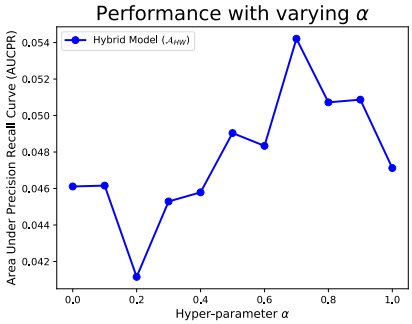


Figure 5: Effect of α on performance.

4. Conclusions And Future Work

In this paper, we investigate the problem of injury risk prediction in a supervised learning framework. To improve the performance in presence of highly imbalanced data, we employ ensemble-based resampling techniques. To address the lack of labeled data, we propose an instance-based transfer learning method. Additionally, we provide actionable insights to prevent injuries and show the effectiveness of our framework experimentally. In our future work, we focus further on discovering the causal relationships from observational data as it is a key element in injury prevention.

Acknowledgment

We would like to acknowledge Parinaz Sobhani, Madalin Mihailescu, Chang Liu, and Diego Huang for their invaluable assistance and insightful comments on the initial draft of this work. We are specially grateful to our management team at Cority and Georgian Partners including Stan Marsden, David Vuong, Madalin Mihailescu, and Ji Chao Zhang for encouraging and supporting our research activities, without which we were unable to complete this work. Finally, we would like to thank Babak Taati for his direction and guidance on this study.

References

Safety and health at work, 2018. URL http://www.ilo.org/global/topics/safety-and-health-at-work/lang--en/index.htm.

Asgarian, A., Ashraf, A. B., Fleet, D., and Taati, B. Subspace selection to suppress confounding source domain information in aam transfer learning. In *IEEE IJCB*, 2017.

Asgarian, A., Sobhani, P., Zhang, J. C., Mihailescu, M., Sibilia, A., Ashraf, A. B., and Taati, B. A hybrid instance-based transfer learning method. *arXiv:1812.01063*, 2018.

Chawla, N. V. C4. 5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In *Proceedings of the ICML*, 2003.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002.

Chawla, N. V., Lazarevic, A., Hall, L. O., and Bowyer, K. W. Smoteboost: Improving prediction of the minority class in boosting. In *In Proceedings of the Principles of Knowledge Discovery in Databases (PKDD)*, 2003.

Drummond, C. and Holte, R. C. Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 2006.

Foster, D. *xgboostExplainer: XGBoost Model Explainer*. R package version 0.1.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1997.

Friedman, J. H. Greedy Function Approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5): 1189–1232, 2001a.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 2001b.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *Trans. Sys. Man Cyber Part C*, 2012.

Liu, J. S. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

Naglah, A., Khalifa, F., Mahmoud, A., Ghazal, M., Jones, P., Murray, T., Elmaghraby, A. S., and El-baz, A. Athlete-customized injury prediction using training load statistical records and machine learning. In *IEEE ISSPIT*, 2018.

Pearl, J. *Statist. Sci.*, 1993. doi: 10.1214/ss/1177010894.

Pearl, J. Comment: Understanding simpsons paradox. *The American Statistician*, 2014.

Poh, C. Q., Ubeynarayana, C. U., and Goh, Y. M. Safety leading indicators for construction sites: A machine learning approach. *Automation in Construction*, 2018.

Rivas, T., Paz, M., Martín, J., Matías, J. M., García, J., and Taboada, J. Explaining and predicting workplace accidents using data-mining techniques. *Reliability Engineering & System Safety*, 2011.

Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernàndez, J., and Medina, D. Effective injury forecasting in soccer with gps training data and machine learning. *PloS one*, 13(7):e0201264, 2018.

Sánchez, A. S., Fernández, P. R., Lasheras, F. S., de Cos Juez, F. J., and Nieto, P. G. Prediction of work-related accidents according to working conditions using support vector machines. *Applied Mathematics and Computation*, 2011.

Sarkar, S., Patel, A., Madaan, S., and Maiti, J. Prediction of occupational accidents using decision tree approach. In *IEEE Annual India Conference*, 2016.

Seiffert, C., Khoshgoftaar, T. M., Hulse, J. V., and Napolitano, A. Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Trans. Systems, Man, and Cybernetics, Part A*, 2010.

Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B., and Bowman, D. Application of machine learning to construction injury prediction. *Automation in construction*, 2016.

Tomek, I. Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, 1976.

Wilson, D. L. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, 1972.

Zhao, Q. and Hastie, T. J. Causal interpretations of black-box models. In *Journal of Business & Economic Statistics*, 2017.

Supplementary Material

1. Transfer Learning

Given a loss function $\mathcal{L}(.)$ and a set of target training samples $\{(x_i, y_i)|i \in \{1, 2, ..., N_T\}\}$, the goal of supervised learning is to find model $\mathcal{A}^*$ that minimizes the expected error, *i.e.*, $\mathcal{A}^* = \arg\min_{\mathcal{A}\in\mathbb{A}} \mathbb{E}_{x\sim P_T}\big[\mathcal{L}(\mathcal{A}(x), y)\big]$. Here x is an arbitrary sample and P_T is the probability distribution of target samples. We follow the idea of importance sampling (Liu, 2008; Asgarian et al., 2018) to transfer the knowledge from source domain (S) to target domain (T). Assuming an infinite number of training samples, we can express the expected error based on the samples from target and source domains as follows (Haase et al., 2014).

$$\begin{aligned}\mathbb{E}_{x\sim P_T}\big[\underbrace{\mathcal{L}(\mathcal{A}(x), y)}_{\epsilon(x)}\big] &= \int \epsilon(x) P_T(x) dx \\ &= \int \epsilon(x) \underbrace{\left[\alpha + (1-\alpha)\frac{P_S(x)}{P_S(x)}\right]}_{=1} P_T(x) dx \\ &= \alpha \mathbb{E}_{x\sim P_T}\Big[\epsilon(x)\Big] + (1-\alpha)\mathbb{E}_{x\sim P_S}\left[\epsilon(x)\underbrace{\frac{P_T(x)}{P_S(x)}}_{w_x}\right]\end{aligned} \tag{1}$$

Here $\epsilon(x)$ shows the error for each sample and α is a hyper-parameter that controls the overall relative importance between source and target samples. Source sample weights $\{w_{x_j} = \frac{P_T(x_j)}{P_S(x_j)} \mid j \in \{1, ..., N_S\}\}$ have a major role in instance-based transfer learning methods, as they control the individual effect of source samples. Considering the case where a finite number of source (N_S) and target (N_T) training samples are available, we can replace the expected values and other terms in Equation 1 with their respective counterparts.

$$\Theta^* = \arg\min_{\Theta}\left(\frac{\alpha}{N_T}\sum_{i=1}^{N_T}\epsilon(x_i, \Theta) + \frac{1-\alpha}{N_S}\sum_{j=1}^{N_S}\epsilon(x_j, \Theta) w_{x_j}\right) \tag{2}$$

Models $\mathcal{A}_S$, $\mathcal{A}_T$, and $\mathcal{A}_{S\cup T}$ trained respectively on source, target and the union of source and target, serve as minimum baselines that a transfer learning method must outperform. Our fourth baseline model is an instance-weighted model ($\mathcal{A}_1$) with all the weights set to $\mathbf{1}$ (*i.e.*, $\mathbf{W}_S = \mathbb{1}$). This is similar to $\mathcal{A}_{S\cup T}$, except in this model we use α to determine the relative overall importance between source and target samples.

Our last baseline model ($\mathcal{A}_G$), assumes Gaussian distributions for target and source samples, which leads to $w_{x_j} = \frac{P_T(x_j)}{P_S(x_j)} = \frac{\mathcal{N}(x_j;\mu_T,\Sigma_T)}{\mathcal{N}(x_j;\mu_S,\Sigma_S)}$. Here μ_T and μ_S show the mean and Σ_T and Σ_S show the covariance matrices for target and source distributions respectively. We call this model with $\mathcal{A}_G$.

Hybrid Weights: Previous methods evaluate the source sample weights solely based on their similarity to the target domain. We argue that it is also important to measure the relevance of source samples to the target task. Hence, we define weights $w_x = w_{domain_x} + w_{task_x}$, where w_{domain_x} measures the similarity of an arbitrary source sample x to the target domain, while w_{task_x} measures the importance of sample x in the target task.

For evaluating w_{domain_x}, unlike the previous methods that employ generative approaches to estimate P_T and P_S, we directly approximate weights $w_{domain_x} = \frac{P_T(x)}{P_S(x)}$ with a discriminative classifier. More specifically, using $\{(x, l_x) \,|x \in S \cup T$, and $l_x = 1$ if $x \in S$, $l_x = 0$ otherwise$\}$, we train a binary classifier (e.g., logistic regression (LR)) to differentiate source and target samples. Next, we use the learned weights of this classifier (w_{lr} and c_{lr}) to estimate source sample weights $w_{domain_x} = \frac{P_T(x)}{P_S(x)} \approx \frac{1}{\exp(x^T w_{lr} + c_{lr})}$.

To compute w_{task_x}, we train an instance of our predictive model (XGBoost) using all samples from source and target with their corresponding injury labels. We then define w_{task_x} to be the uncertainty of this model about sample x. We define the uncertainty of a model about sample x to be the distance of x to the decision boundary. Note that this value could be negative (thus subtracting from w_x) when the decision is incorrect, or positive (thus adding to w_x) when the decision is correct. We denote this model as $\mathcal{A}_{HW}$.

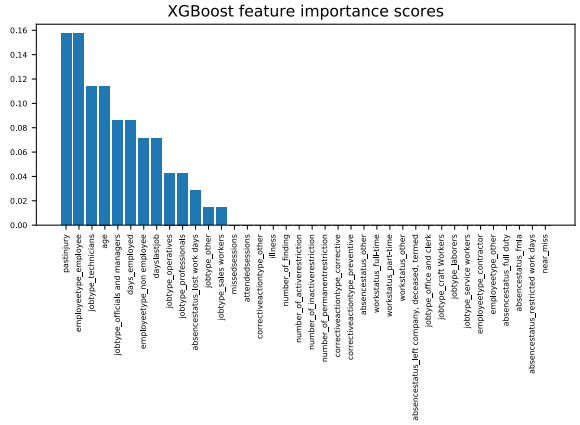
2. Additional Results

Table 1 shows the performance of various methods on positive (injury) class and negative class. As it can be seen, the performance of all methods is very high on negative class and relatively very low on positive class, which could be explained by the imbalance issues in the dataset. The trend is that almost all instance-based methods perform better compared to the baseline models $\mathcal{A}_T$, $\mathcal{A}_S$, and $\mathcal{A}_{S\cup T}$, highlighting the effect of transfer learning. Among instance-based transfer learning methods, our model $\mathcal{A}_{HW}$ has the best performance in terms of F_1-score and AUCPR while having a high F_1-score on the negative class.

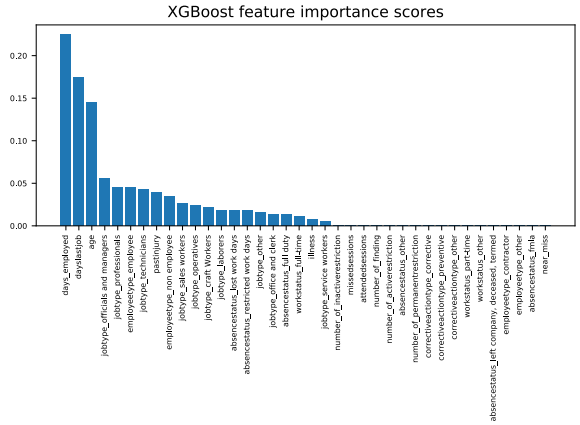
Figure 1 shows the feature importance scores obtained with target-only model ($\mathcal{A}_T$) and our model ($\mathcal{A}_{HW}$). We can see that model $\mathcal{A}_T$ has a very low importance score for 25 features from the 38 engineered features and mainly considers 13 features. Among these 25 features, we can see features like the number of illnesses (`illness`), if employee are technician or laborers (`jobtype_technician` and `jobtype_laborer`), if employees are working full-time (`workstatus_full-time`). However, our model $\mathcal{A}_{HW}$ utilizes 23 features which includes $\mathcal{A}_T$'s 13 features plus 10 more features including the number of illnesses, if employee are technician or laborers, and if they are working full-time.

Methods	Positive Class				Negative Class		
	Precision	Recall	F_1-score	AUCPR	Precision	Recall	F_1-score
$\mathcal{A}_T$	0.07	0.06	0.06	0.0375	0.98	0.99	0.99
$\mathcal{A}_S$	0.04	0.18	0.07	0.0405	0.99	0.93	0.96
$\mathcal{A}_{S\cup T}$	0.13	0.06	0.08	0.0478	0.98	0.99	0.99
$\mathcal{A}_{\mathbf{1}}$	0.06	0.12	0.08	0.0456	0.99	0.97	0.98
$\mathcal{A}_G$	0.07	0.16	0.10	0.0532	0.99	0.97	0.98
$\mathcal{A}_{HW}$	0.11	0.12	0.12	0.0542	0.99	0.98	0.98

Table 1. Per class performance of different methods on Company-1's data.



(a) Features importance scores for model $\mathcal{A}_T$.



(b) Features importance scores for model $\mathcal{A}_{HW}$.

Figure 1. Illustration of feature importance scores obtained with target-only model ($\mathcal{A}_T$) and our hybrid model ($\mathcal{A}_{HW}$). Company names are masked for confidentiality reasons.

References

Asgarian, A., Sobhani, P., Zhang, J. C., Mihailescu, M., Sibilia, A., Ashraf, A. B., and Taati, B. A hybrid instance-based transfer learning method. *arXiv:1812.01063*, 2018.

Haase, D., Rodner, E., and Denzler, J. Instance-weighted transfer learning of active appearance models. In *2014 IEEE CVPR*, 2014.

Liu, J. S. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.