Predicting work place accident types using data mining

Literature

Explaining and predicting workplace accidents using data-mining techniques , Rivas et al, 2010

“Prediction of Workplace Injuries” Sadeqi, Asgarian, Sibelia, 2019

Greedy Function Approximation: A Gradient Boosting Machine, Friedman, J. H. (February 1999)

R Packages: Mice, Random Forest, XGBoost, bnlearn

Netica Software

Jacky Ho

Teodora Lang (Saint Gobain Data Science team)

Martine Barons and Simon French (Project Supervisors)

J.Ho.5@warwick.ac.uk

1.Introduction

Saint Gobain is an international company specialising in construction materials. They are interested in limiting the number of accidents

in the workplace by assessing the risks through **data mining**. My goal is to classify the **incident** type. The incident types are defined by Saint Gobain on a scale from TF1 to TF4.

TF1 (Lost Time Injury) : Incident leading to days missed from work.

TF2 (Minor Injury) : Incident where treatment is needed to be done in an external place.

TF3 (First Aid Injury) : Incident where treatment is done internally.

TF4 (Near Miss) : Incidents where accident or injury was avoided.

2.Data

The data consists of demographical information and incident information, coming in the form of numerical and categorical data. Focus is on the smaller dataset of 9374 incidents for TF1-3s, with 16 predictors. This is narrowed down from 119772 incidents and 30 predictors, removing the TF4 incidents because of the high amounts of missing data and the ambiguous definition of the incident type.

We solve it using **MICE** (Multiple Imputation by Chained Equations) via CART. MICE imputes the missing data multiple times based on a Gibbs Sampler procedure, which predicts the missing data using full

conditionals.

Incident Type	Near Miss (TF4)	First Aid Injury (TF3)	Minor Injury (TF2)	Lost Time Injury (TF1)
Number of incidents	106061	7135	1663	584
Mean Missing Predictor Data	36.67%	20.93%	29.49%	20.21%

Fig. 1 Number of incidents by Incident type. Mean Missing Predictor Data calculated by taking an average of the missing data of each predictor.

3.Methods

Multinomial logistic regression

Assigns a probability of being classified for each class. Choose the class with the highest probability. It uses a softmax function:(Zi = Class i)

$$\boxed{\frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}}$$

XGBoost

A **gradient boosted** decision tree algorithm that is fast and has high accuracy.

Random Forests

Ensemble algorithm which takes a collection of decision trees and uses a **majority voting** rule to decide the most likely class and classifies it accordingly.

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right)$$

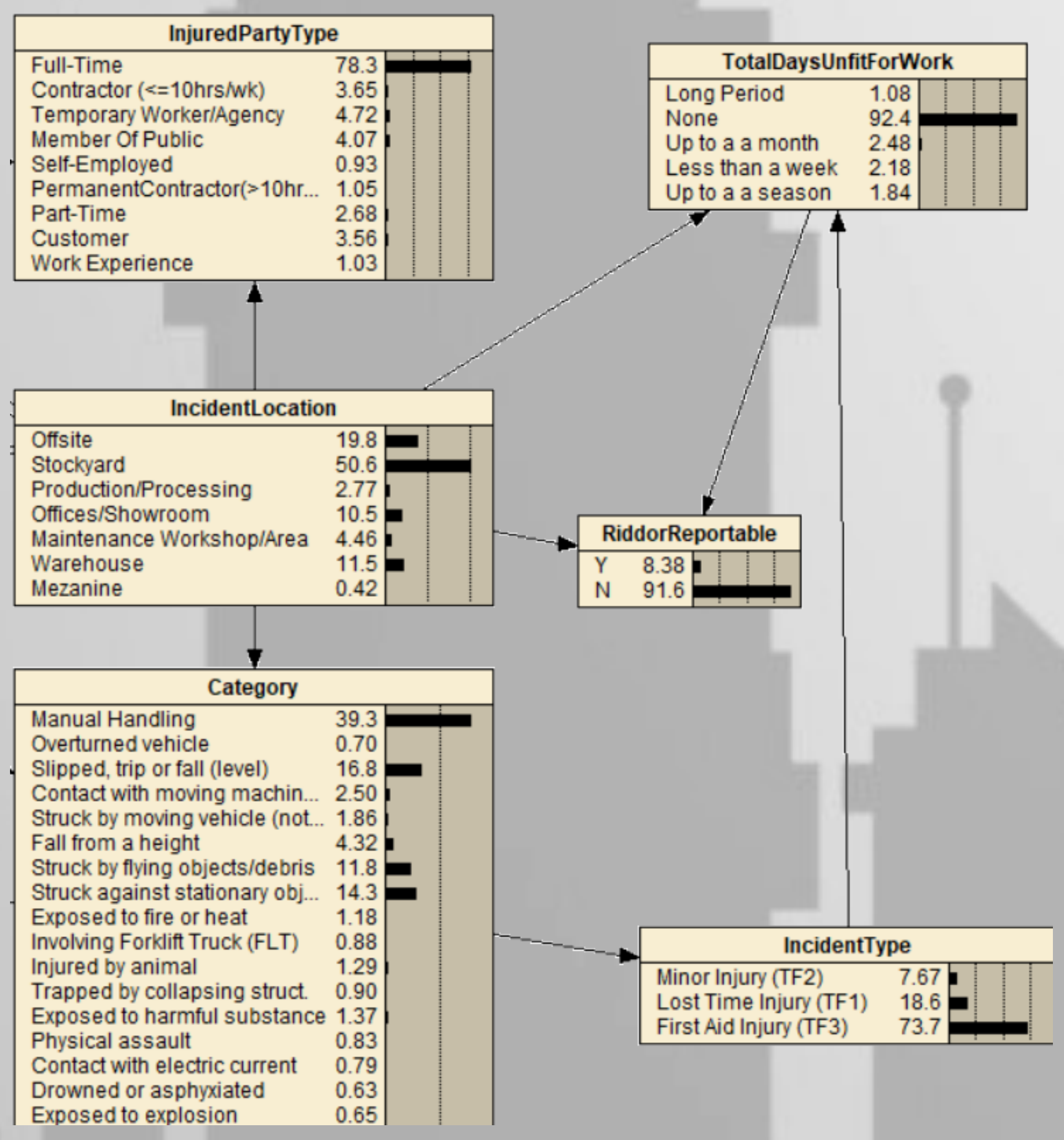
$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

Fig. 2 Gradient boost algorithm, where α is the learning rate, and h_m(x) is a learner

Bayesian Networks

The directed acrylic graph represents conditional probability statements for a set of random variables. The edges represent statistical dependence between variables (nodes). The network is easy to understand provided the domain expert has verified it plausibility of the dependencies

Conditional Probability Table



InjuredPartyType	
Full-Time	78.3
Contractor (<=10hrs/wk)	3.65
Temporary Worker/Agency	4.72
Member Of Public	4.07
Self-Employed	0.93
PermanentContractor(>10hr...	1.05
Part-Time	2.68
Customer	3.56
Work Experience	1.03

TotalDaysUnfitForWork	
Long Period	1.08
None	92.4
Up to a a month	2.48
Less than a week	2.18
Up to a a season	1.84

IncidentLocation	
Offsite	19.8
Stockyard	50.6
Production/Processing	2.77
Offices/Showroom	10.5
Maintenance Workshop/Area	4.46
Warehouse	11.5
Mezanine	0.42

RiddorReportable	
Y	8.38
N	91.6

Category	
Manual Handling	39.3
Overturned vehicle	0.70
Slipped, trip or fall (level)	16.8
Contact with moving machin...	2.50
Struck by moving vehicle (not...	1.86
Fall from a height	4.32
Struck by flying objects/debris	11.8
Struck against stationary obj...	14.3
Exposed to fire or heat	1.18
Involving Forklift Truck (FLT)	0.88
Injured by animal	1.29
Trapped by collapsing struct.	0.90
Exposed to harmful substance	1.37
Physical assault	0.83
Contact with electric current	0.79
Drowned or asphyxiated	0.63
Exposed to explosion	0.65

IncidentType	
Minor Injury (TF2)	7.67
Lost Time Injury (TF1)	18.6
First Aid Injury (TF3)	73.7

Fig.3 Selected subsection of a Bayesian network created by a Data-driven Bayesian network.

4. Current Results

Method	TF3 Accuracy	TF2 Accuracy	TF1 Accuracy
Multinomial regression	High	Low	Low
Random Forest	Medium	Low	High
XGBoost	High	Low	Low-Medium
Bayesian Network	High	High	Low

Fig.4 Employee moving heavy material.

Jewsons is the largest Saint Gobain brand.

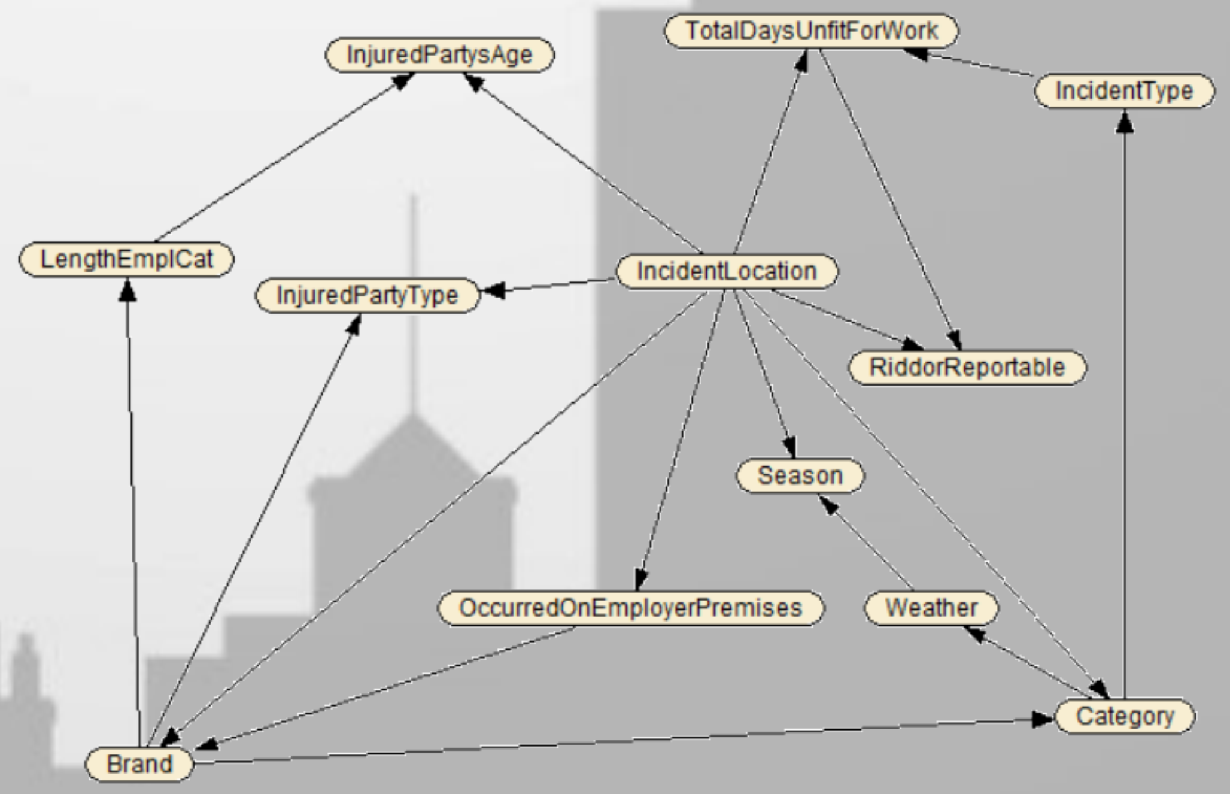


Fig.5 Data-driven Bayesian network

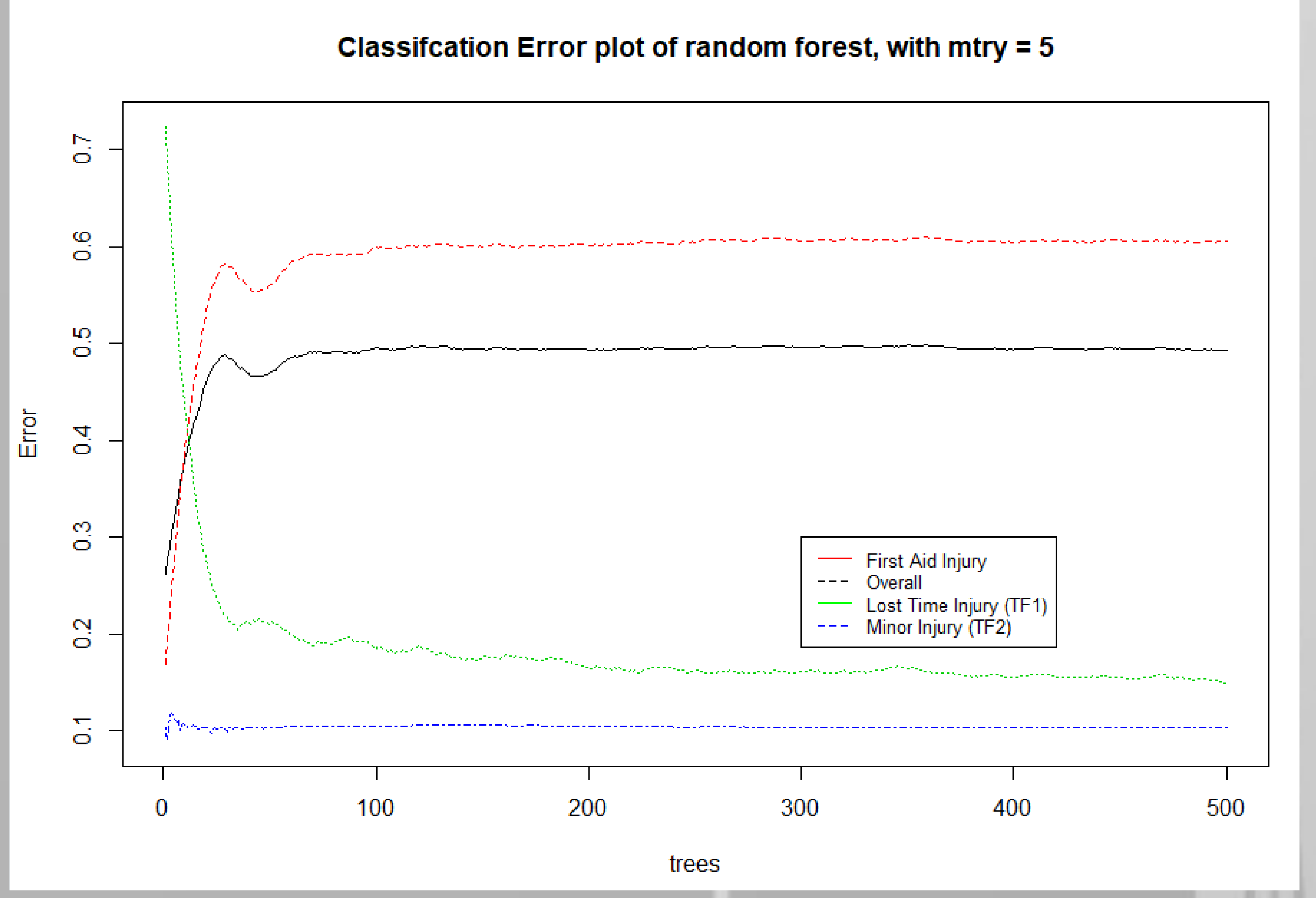


Fig. 6 Classifcation Error Plot, mtry is the number of variables randomly sampled at each split.