Predicting Workplace Accidents using Data Mining

Jacky Ho U1614422

May 2020

Abstract

Within the workplace, accidents cause economic, social and systematic problems for companies in the construction industry . Data mining has been identified as a way to discover patterns and relationships in accident data. Saint-Gobain have an aim to achieve zero avoidable accidents. Random Forests, Bayesian Networks and XGBoost are applied with the aim of inferring characteristics of accidents. Research into the application of these techniques has been applied only on small datasets. This serves as an extension and to evaluate their performance on a larger dataset.

Acknowledgements

Martine Barons and Simon French for jointly supervising the project

Teodora Lang for introducing the project and supplying company relevant information.

Fan Zhao for jointly working on the same project title.

Contents

1 Introduction

A work accident is an unforeseen event leading to physical or mental damage to the person. Worldwide, more than 2.78 million work-related deaths occur per year, with over 374 million non-fatal accidents. [29] It is in the employers interest to seek a way to limit this. An incident is a logged event where a (potential) accident has occurred. Accidents have a consequence on the company in the form of loss of work hours , money, public image and personal trauma.

Material distribution, processing and transportation are sectors which have a risk of an accident for workers and people around them. A wide variety of jobs are employed in these areas, which leads to a wide range of causes for accidents. Saint-Gobain Building Distribution (SGBD) is an international company, founded in 1665 , which specialises in building distribution with many well-known brands such as Jewsons and British Gypsum. It employs over 100,000 people worldwide and 17000 within UK. Incidents predominantly will have occurred with skilled workers due to Saint Gobain mainly being a supplier to the industry.

Unforeseen incidents cause a lot of problems for companies in the construction sector. The company has a goal of **zero avoidable working accidents** which has led to an interest in a data driven campaign to improve safety standards. The Saint-Gobain Data Science team have provided a dataset of 119972 incidents with the aim of being able to draw conclusions for which their business can use to change their safety practices.

To reduce the number of avoidable accidents, there has been a recent rise in the use of data mining to find elements that increase the risk of accidents. This is in line with Saint-Gobain's goal of reducing the accidents in the future. [33]

The aim of data mining is the "Discovery of interesting, unexpected or valuable structures in large datasets." [9]. It uses ideas and concepts from areas such as Statistics, Big Data, machine learning and AI. We find underlying patterns in the data which would not be seen otherwise using algorithms in order to extract useful information. Popular methods include Bayesian networks and random forests. Both are flexible models which allow for non-linear relationships to be obtained from real life data which may not be appear informative using descriptive statistics.

Data mining has the benefits of

- Drawing otherwise unseen conclusions from high dimensional data that exploratory data analysis may struggle to find.
- Predictive and interpretive power
- Flexibility in being able to work with many types of data. [10]

There are also disadvantages to consider:

- The Black box nature where conclusions may lack intuition
- Being too flexible which leads to over fitting from modelling the noise in the data.
- May require a lot of data cleaning and computational time to potentially draw the same conclusions as a simple linear regression or Exploratory Data Analysis.

Saint-Gobain have two broad objectives:

1. To find characteristics leading to accidents
2. To predict where the next accident is going to happen

I will also assess which algorithm performs best in terms of individual class accuracy.

This is an investigation into what variables influence an accident and whether we can predict where the next accident will occur. The idea is to find patterns which can be incorporated into the company strategy. Integral to this will be a classification problem to classify the type of incident that occurs in terms of its severity.

I will identify predictors which explain patterns in the data and draw conclusions which Saint-Gobain can use. I would like to create a model that either assigns probabilities to an event occurring, or predicts an event of interest. Using **Bayesian networks, Random Forests and XGBoost**, I aim to assess the predictive capabilities of each algorithm and discover reasons for what makes an incident severe.

The second objective is more challenging. Predicting where the next accident will be is very difficult because safety standards in the UK are already very good in Europe. There are 0.53 fatal injuries per 100,000 employees. This is lower than France, Germany, Italy and Spain. [40], Compared to the number of hours to the number of hours the average worker works, it is very minimal. Accidents are essentially rare events which have a very small probability of occurring. Often the individual causes of these accidents are small deviations from safety norms which are really hard to predict. However there is reason to believe on a large scale, we can uncover meaningful patterns.

A target is to **identify locations with poor safety qualities** in order for the company to focus in on improving health and safety standards if they are failing. The approach I propose is to identify and rank the brands with the highest risk of accidents.

Another target be to produce a model that distinguishes between severe incidents and trivial incidents. This could be minimising TF3s and TF1s. Find characteristics which separate these two incident types.

Due to data limitations as a result of General Data Protection Regulation (GDPR), supply of confidential information to allow me to identify specific incidents is restricted.

For example I am not supplied enough information to distinguish locations to adequately locate specific branches. The information I have consists of a location code and a brand name. There are over 1000 location types, and 25 brand names. Despite attempts to identify brands with these locations, I have found it hard to identify meaningful characteristics using exploratory data analysis due to the size.

The Incident Type determines the severity of the incident, with 4 levels considered:

- A **Lost Time Injury (TF1)** corresponds to an incident where at least one full working day has been missed (Not including the day of the accident).
- A **Minor Injury (TF2)** refers to an incident which requires offsite treatment (e.g injection at a clinic), but the injured party can return straight back to work. for example, an injection or a trip to the hospital.
- A **First Aid Injury (TF3)** refers to an incident in which people report injuries that require no offsite treatment and/or lead to no days lost.

- A **Near Miss (TF4)** is an incident where incidents have occurred with potential to cause an injury. An example would be narrowly dodging a collapsing structure.

A significant distinction between the Injuries (TF1-3s) and Near-Misses (TF4s) is that TF4s are Incidents where **accident or injury was avoided** whilst TF1-3s are incidents **leading to an injury**.

Mandatory data to be inputted includes the location, incident type, whilst many other variables such as Gender, EmploymentType, and many overall safety check variables are optional unless it leads to injury.

2 Literature review

The application of data mining techniques for the purpose of improving safety standards has been applied in similar areas, such as mining and the steel industry.

Kamalesh [3] comments , that the "The safety performance of any construction activity is affected by a plethora of varied human, technical, environmental, and organizational factors. Although the apparent cause of any construction accident is generally attributed to a single trigger, very often either due to human error or due to technical fault, it is often the case that the accidents are the result of many latent factors and sub-factors contributing together to trigger an accident." This provides motivation for exploring predictive models of incidents in order to identify any underlying factors which may increase the likelihood of an accident.

Their use of Bayesian belief networks allows them to identify causal factors, construct influence diagrams and establish probability tables. They find a "strong case is made to use data mining technique together with a Bayesian belief network to predict the occurrence of accidents and injuries in a construction site." They find conversion of unstructured data into useful information that is hidden in the data to be very useful. They concede there is still much progress to develop this analytical approach in the construction industry as it lags behind the business and manufacturing industries. They presented it as a concept and hope it is usable for future researchers.

Rivas [10] explained that in the management of workplace risk, using conventional descriptive statistics fail to correctly identify cause-effect relationships and led to the inability to develop models that could predict accidents. In the context of incidents for two Spanish companies in the Mining and Construction sector, they used variable selection to draw the most important variables.

They then constructed models using Bayesian Networks and Support Vector Machines and Classification and Regression Trees. In doing so they found differences in effectiveness of Logistic Regression,a classical statistical technique and compared it to their data mining algorithms. They saw an increase in success from 72.58% for Logistic Regression to above 80%, reaching 88.71% through BayesNet-K2-8 Parents.

They found variables of greater importance to be associated with the type of work done by the worker and aspects associated with the type of employment. Variables associated with managing risk overall and that did not depend directly on the worker (Risk Assessment and Risk Management) were ranked 8th and 9th.

Their 3 most important variables were:

1. Task duration in hours
2. Company contractual status
3. Length of time doing the job

With 7 variables, most errors in the models occurred in classifying the accidents. This can be explained by the fact that accidents are less well represented in the sample than incidents. Making it more difficult for models to characterise the former. When all variables were used, the model classification errors were more evenly distributed between accidents and incidents, although the success rates fell. Implying using all variables lead to overfitting

Each of their models resulted in finding similar informative predictors.

In contrast with Saint-Gobains large dataset, Rivas' dataset is comprised of 62 completed interviews in response to an accident. The difference in sample size may mean

They went on to establish conditional probability tables which showed the conditional probabilities for each state, estimated from the data using the maximum likelihood method. They

found that when an accident occurred, the probabilities that a task lasted more than 8h or less than 4h were 0.6410 and 0.333 respectively. They concluded that an accident is less likely to happen during a task within 4-8hours because the probability of this event was lower.

They mention the advantages of Bayesian networks in the use of What-if Analysis to allow greater depth of data exploration and model workplace scenarios to be depicted. They found specific demographics of employees to have higher risk. These included employees working on part time contracts, older workers and people who are only recently employed.

They stressed that some variables had little influence on incident distribution. They found assessment of risk led to higher rates of protection. This also was correlated to having factors associated with a higher incident rate. Poor job related protection was associated with brief tasks and unspecified jobs, which had a higher accident rate.

This comes for me to question if the assessment of risk will have an impact of the type of working environment it will have. For example, risk assessment would not be as common in a office as opposed to on a construction site. Hence the risk assessment factor will not simply be a sign of better safety practice, but acknowledgement of the severity of the working environment.

Through selecting the most relevant information then identifying the best algorithms, they have discovered circumstances with have a bearing impact on accidents. This should lead to causes being clearly defined. They note that these conclusions are only valid for their companies and their time period and hence extrapolation would be approached with caution. They found their models screened out the uninformative information and used this as a recommendation for improved efficient data collection. They found CART was best used to identify circumstances associated with a greater probability of incidents. They noted that Bayesian networks have an advantage in their what if analytical capacity. They noted risk assessment appeared to have little effect and may have because better in for workers employed by subcontractors. They cite the small dataset being a weakness potentially although cite the ideal sample size cannot be pre-established because it depends on the variability of the data and the structure of the relationships between variables. They not that the improvement of the accuracy from logistic regression to other models indicates non-linear relationships between variables.

They concluded that "These studies, based on data from accident reports and interviews with workers and employers, confirm the advantages of data mining over conventional statistics in terms of the predictive function and the possibility of identifying interactions between variables with a bearing on accidents."

In the steel industry in Iran and India [37] , CHAID (Chi-squared Automatic Interaction Detection) and CART (Classifcation and Regression Trees) have been attempted with varying degrees of success.

They used the Gini index as a measure of determining best split. Pruning of the CART was seen to decrease generalisation error. A growing algorithm was used to obtain optimal parameters for accuracy. They obtained decision rules, which elicit the a probabilistic representation of the severity of the incident conditioned on the values of predictor variables. However they have not handled experimentation with missing values and did not apply it to a large dataset.

In Iran, using data from 2001 to 2014, they found 81.78% accuracy with CART and 80.73% accuracy with CHAID . [5]

The CART and CHAID algorithms were suitable for predicting the outcome of occupational accidents in a steel factory. Furthermore, safety officers can reduce the rate of accidents by using the predictions to identify factors which lead to high levels of accidents. They recommend to employ other methods of data mining such as C5, support vector machine (SVM) and Bayesian networks to predict the outcome of injuries in steel industries for future studies. [37]

Gholam states, "Safety managers can reduce the prevalence of accidents by using the predictions of CART and CHAID for detecting susceptible people"

Tixier et al [2] state "Random Forests inherits many of the advantages of trees, such as the ability to capture complex nonlinear high-order interactions among predictors, to handle highly dimensional datasets with large numbers of observations, and the robustness to outliers and to the inclusion of irrelevant predictors". They cite problems with the class imbalance in their dataset.They followed a systematic approach of optimisation of hyperparameters then applied model fitting and model evaluation. Construction injuries "do not occur in a chaotic fashion, but underlying patterns and trends do exist and can be uncovered when applied to large datasets". A suggestion is to move away from subjective decision making to using decision making backed by statistical methods. They also cite the issue where cases are instances of where there have been an accident. There is no description of the day to day processes for where there are no safety breaches or accidents occurring. This part is hard to evaluate.

3 Data

A brief summary of the predictors can be found in Figure 37 in the Appendix.

3.1 Data Source

Each incident is recorded by the responsible party (For example, the branch manager) via the system PeopleSoft, a separate company that specialised in human resources management accidents. Peoplesoft are now overseen by Oracle [38] The interface contains 4 sections to fill out. (As shown in Figure 1)

- Incident Details
- Incident People
- Injury Details
- Preventative Action

It features some inputs as a drop down or tick boxes, whilst others are optional. The optional data becomes a problem in terms of allowing records to be created with missing data. There are also issues with the definition and preciseness of the values inputted. If there is no injury, the protocol checks whether it is a dangerous occurrence or a near miss.

The dataset provided consists of **119772 Observations** with *30 variables.* We have 3 integer variables, 14 categorical variables, 11 binary variables and 2 continuous variables. 90% of the data is also TF4s. The dataset collects data predominantly from full time employees in a TF4 accident on employee premises.

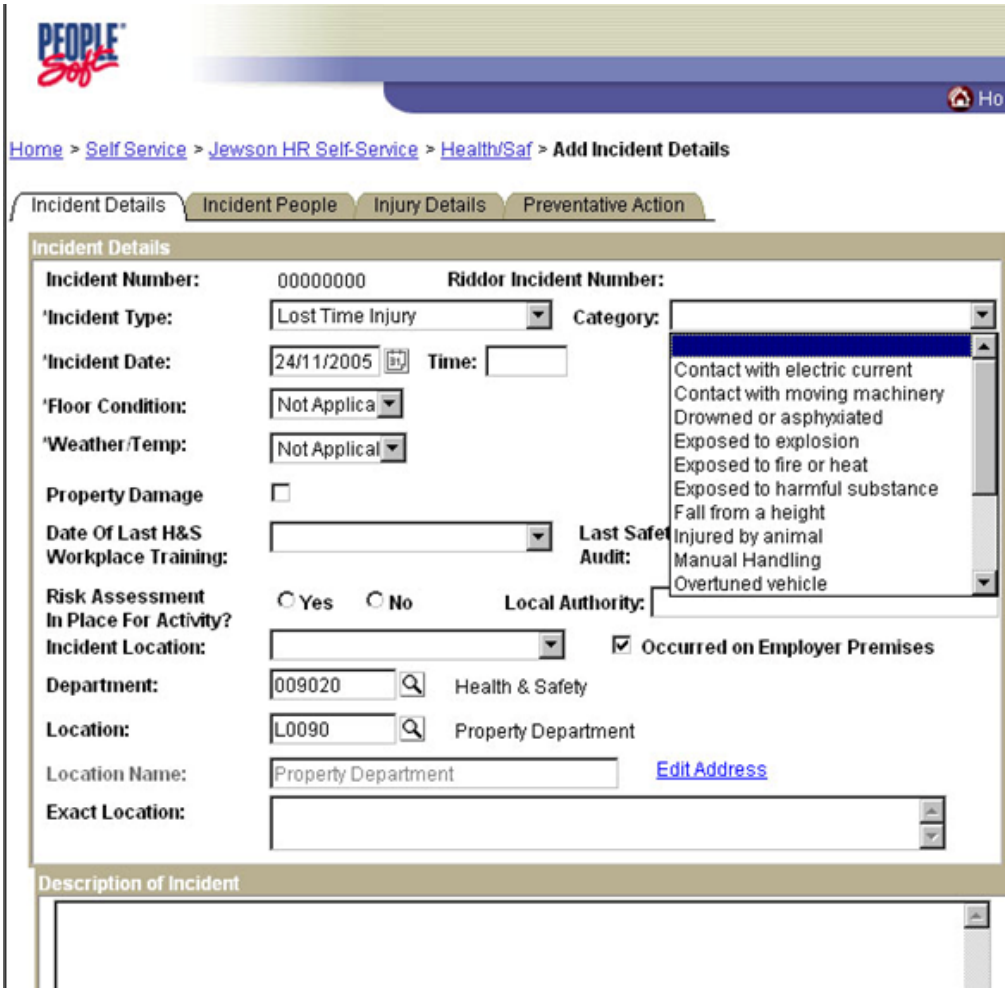

Figure 1: This is the first section of the Incident Form to fill out. The PeopleSoft Interface gives the manager an extension list of details to add. It includes additional details (More precise Geographical and personal information) I am not given due to GDPR

3.2 Description of data variables

Missing Data poses a problem because the algorithms I am using do not work well with missing values.

The imputation method I will use is Multiple Imputation by Chained Equations (MICE) is an algorithm which imputes the missing data multiple times using a specified model. This method does not work well with highly correlated variables [32]. so it is important to describe each variable and evaluate their usefulness.

1. **IncidentType** corresponds to the severity of the incident. See section 1.1 for more details.

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%

Figure 2: Number of Observations and missing data in each Incident Type. TF4. The TF1s have by far the most missing data (36.67%) and most incidents (106061). The TF3 and TF1 have the least missing data at 20%, and TF1s have the fewest incidents recorded (584)

In Figure 2, we can see TF4s is more numerous. It is 10x times more populous than TF3s and nearly 200x more populous than TF1s. There are also 3 additional categories: Dangerous Occurence (RIDDOR), Damage (TF4) and Unsafe Practice (TF5).

There are 3 additional categories which are not under consideration due to their comparatively low sample sizes and ambiguous definitions. The 3 definitions have change of use overtime. They all concern the Reporting of Injuries, Diseases and Dangerous Occurrences Regulations (RIDDOR). RIDDOR is a reportable accident. The government places duties on the employer to inform the HSE about such accidents or occurrences. This may lead to legal repercussions taken by the HSE. [11] RIDDOR Dangerous Occurrence, Unsafe Practice and Damage are all incidents not leading to time off work.

2. A **RIDDOR Dangerous Occurrence** is an incident where no personal injury occurs but had the potential to be a serious injury. Example would include a release of a harmful substance or overturned forklift. In this event, the system will advise if a RIDDOR form must be completed.

3. **Unsafe Practice** Where unsafe performance of work has been noticed and recorded. There are only 23 obserations in this category with a lot of missing data but it has been seemed unsafe enough to be recorded.

4. **Damage** Regards incidents where damage has been noticed. These are incidents which occur more often in cases where protective clothing were worn and safety checks were taken place. There is a significantly higher percentage of the following predictors taking the following values: Mechanical equipment being used that the dataset average (55%) and authorisation to operate has been given 99.9% of the time that it is relevant and is relevant 40% of the time. Protective clothing is worn 57% of the time. A significant proportion of the incidents in the dataset which have property damage have been labelled as "Damage". (793 out of 2613)

   The main disadvantage of these three classes is that they are not recorded frequently, which them having 130, 23 and 4130 points respectively. This is compounded by the different periods that have been sampled from, as seen in Figure 4.

5. **IncidentDate** gives information as to whether specific days of the calendar are injury prone. I have made modifications to this to represent it in terms of months or seasons in order to find more patterns specific periods of the year and I do not want to measure the changes over the years in my modelling. .

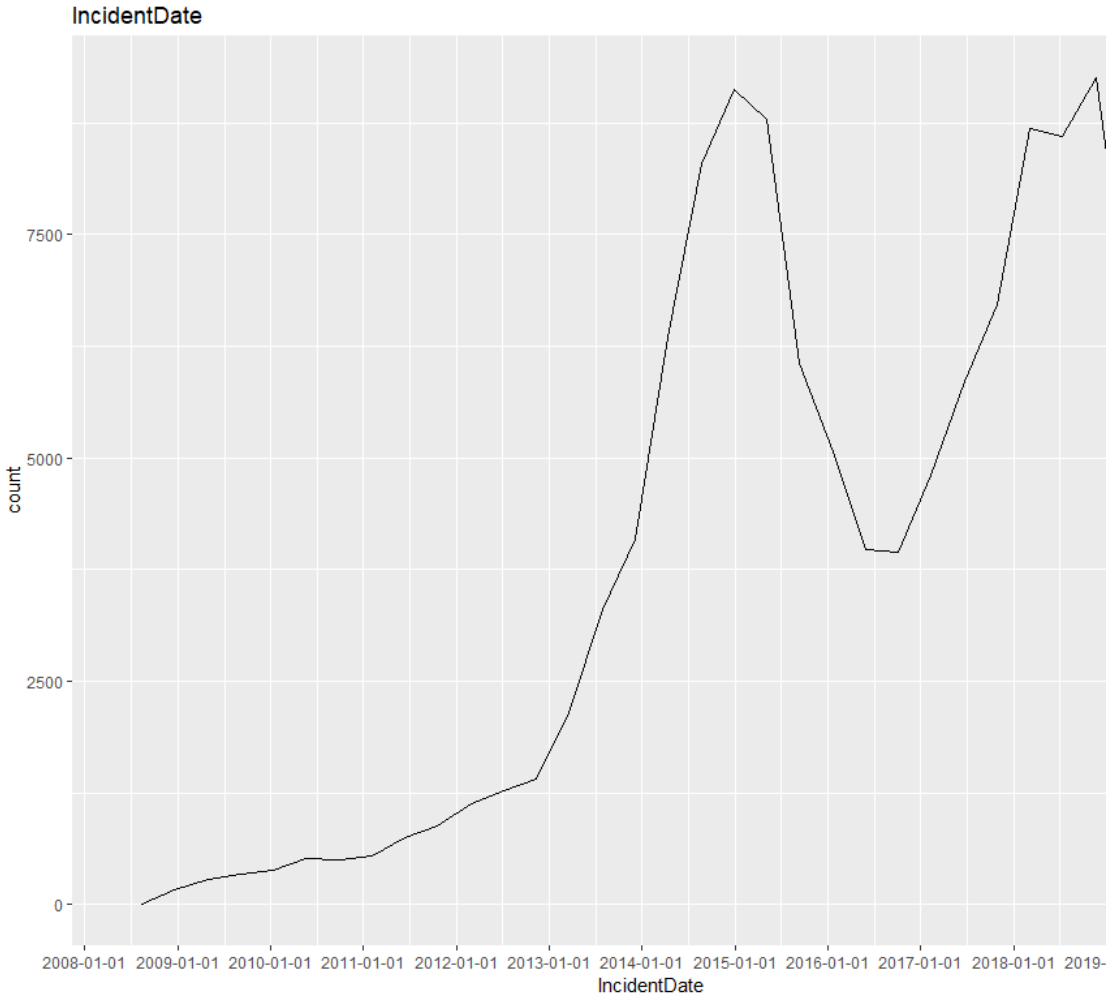

Figure 3: Number of incident dates on a monthly basis. The amount of incidents recorded begins to rise significantly after 2013.It begins to drop down (To a level still significant higher than pre-2013) before rising again

6. **IncidentTime** gives information on the incident time, given in intervals of 5 minutes.The idea is to identify whether certain times of the day are more prone to accidents, for example when commuting to work begins or at the end of the day. Accidents have a Gaussian-shape distribution based on the time. So it suggests that the most incidents happen around 11am and goes down earlier and later in the day. Unfortunately it is hard to determine whether this is due to more people working or more dangerous work being done around this time. Data relating to this information was not available from Saint-Gobain and would still require me to make several assumptions.

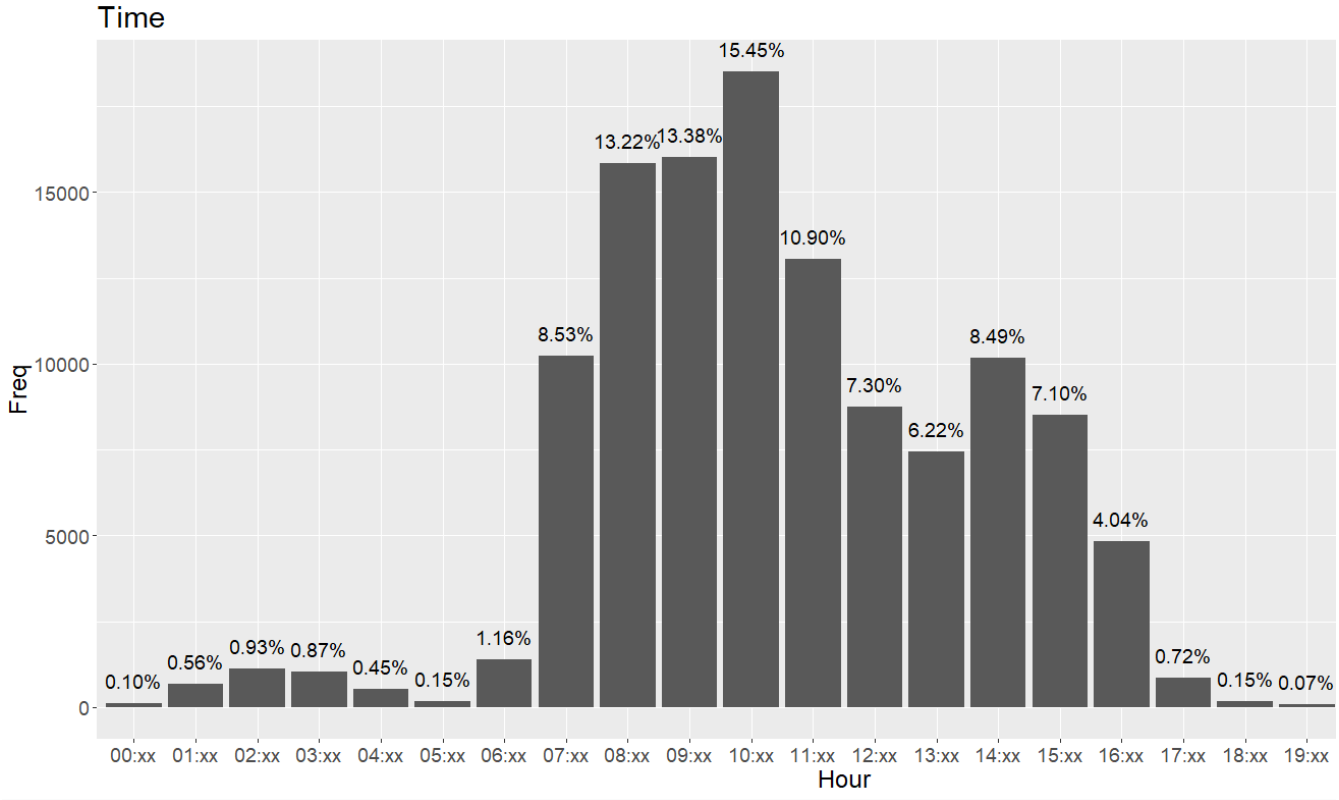

Figure 4: Number of incidents over time. The hour notation, 08:xx, means any time in the 8th hour of the day. The majority of the incident times roughly correspond to the average working day, where most people work between 07:xx to 17:xx. Over 0% of incidents occur during 8:xx and 10:xx.

7. **Category** will be a description of the events of the incident. This may include accidents involving manual handling, falls from height, slips from level ground, or struck by flying object. There is also a distinction between forklift truck accidents and other moving vehicles

   The most common categories of accidents do not appear to be too dependent on the Incident Type, because the same most common categories appear high in the table in Figure 5. The most common categories are Manual Handling, Slips and Falls from Height and being struck by a flying object.

	TF1	TF2	TF3	TF4	TF5	RIDDOR
1st	Manual Handling 29%	Manual handling 41%	Manual Handing 45%	35% Slips	Struck by flying object 23%	17% Struck flying object
2nd	Slips, Trips 27%	24% slips, trips	17% struck by stationary object	17% struck flying object	Struck by stationary object 16%	Fall from height 14%
3rd	Falls from height 14%	11% struck by stationary object	Slip 16%	14% manual handling	14% Fall from height	Slip 13%
4th	Struck by flying object 12%	10% struck y flying object	Struck by moving vehicle 13%	Fall from height 9%	13% Involving FLT	Contact with moving machinery 10%
5th	Contact Moving Machinery 7%	Fall from height... moving vehicle/machinery	Minor	6-7% Slip	9% Manual Handling	Struck by stationary object 9%

Figure 5: Table of the Categories separated by Incident Type. Manual handling and slips are the leading cause of injuries. Minor, the 5th most common category is to denote that the 5th most common category is less than 3% and is not significantly more common than the 6th-9th categories.

8. **Floor Conditions** describes whether the floor was wet or dry. There is an option for normal, which is presumed to mean dry.

9. **Weather** is a qualitative statement on the weather, being a combination of the temperature and an indicator of wet or dry conditions.

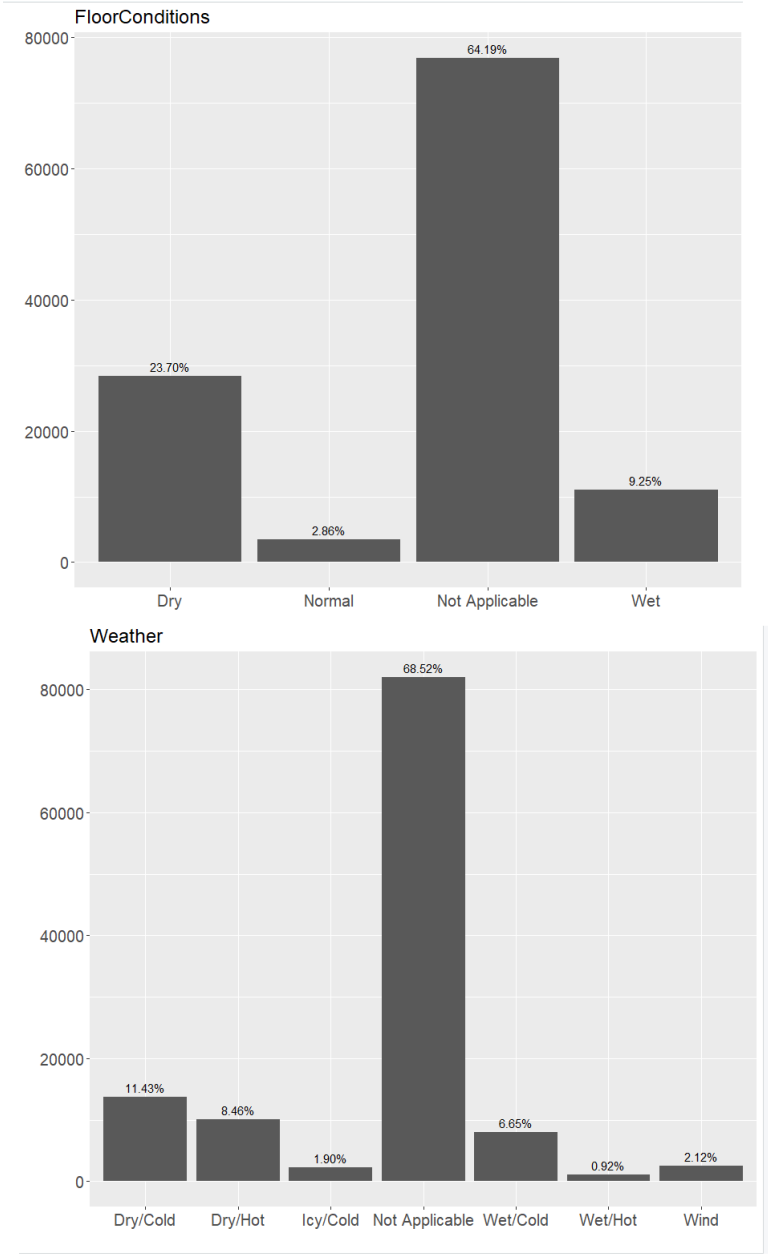

Figure 6: Floor and weather bar plots. Percentages of each category are given. Not Applicable means the detail was deemed not relevant. For example, it could be argued that whether or not it is sunny or rainy should not have an impact on a accident inside a building and it offers a clear distinction from being a missing value.

10. **Incident Location** indicates the type of location of the accident e.g warehouse. This is in hope to find which locations are more prone to accidents.

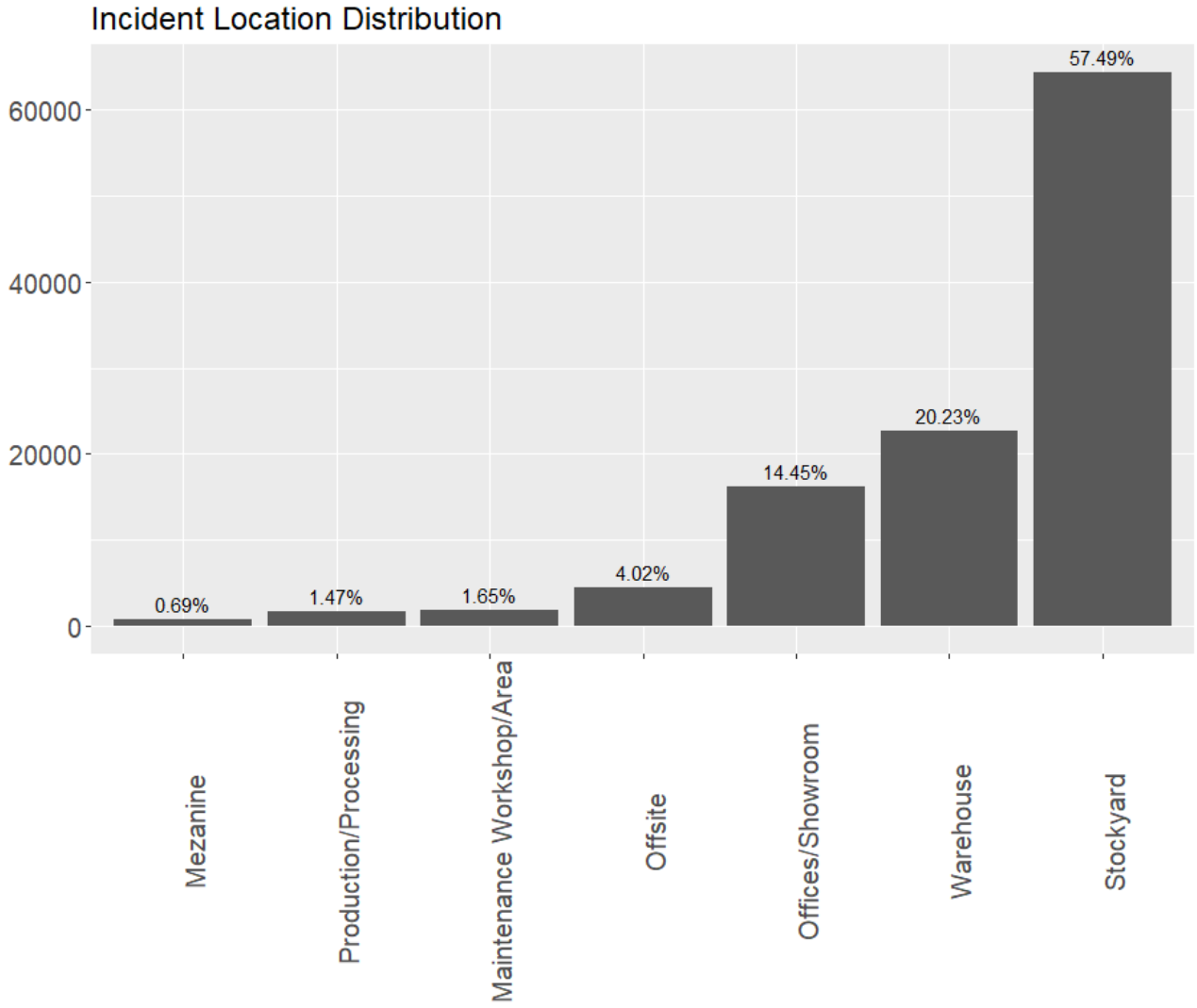

Figure 7: Incident Location. Over 90% of incidents occur in the 3 Incident Locations Stockyard, Warehouse and Offices/Showroom.

The following 5 variables are binary variables with varying degrees of of missing data. It is important that the variables are used appropriately in the given context. For example, the levels of mechanical handling will be different in different work locations.

Each business unit shall ensure suitable and sufficient arrangements are in place to ensure the control and selection these variables are managed. This shall be achieved by ensuring suitable and sufficient assessment, procurement, training in use, issue, wearing of, maintenance, repair and replacement where appropriate. These factors shall be checked on a regular basis to ensure that it is within the guidance lines. Employees shall be trained in the understanding of this concept and be informed of their responsibilities in practicing these guidelines and reporting any problems.

11. **Risk Assessment** indicates whether risk assessment has been taken place at all.

    The purpose of a risk assessment is to assist in determining what measures should be taken to comply with Saint-Gobain's duties under the relevant Saint-Gobain or national and local statutory provisions

    Each business unit shall ensure suitable and sufficient arrangements are in place so that risk assessments are reviewed at regular intervals or immediately if:

    - There is any reason to suppose that the original assessment is no longer valid
    - Any of the circumstances of the work environment have changed significantly.

    Each business unit shall ensure suitable and sufficient arrangements are in place to review risk assessments at regular intervals, the time taken between review being dependent on the nature of the risks and the degree of change likely in the work activity. As a result it is only asked whether a risk assessment has been taken at all. [7]

12. **Occurred On Employer Premises** checks whether incidents are more likely to occur offsite or onsite. This is important as if it is on their premises, they are responsible for making sure the premises are well maintained in order to identify dangerous areas and steps are taken to reduce the risk in the area, removing the risk entirely if possible.

13. **Was Protective Clothing Worn** is considered to make sure suitable personal protective equipment is managed.

14. **Was Mechanical Handling Equipment Used** is important as Saint Gobain claims "Unsafe manual handling leads to more accidents than anything else." [8] Thus Mechanical Handling Equipment can reduce the amount of accidents, through reducing the manual workload. Manual Handling is the second most common category of incidents with 18804 incidents. Thus an indication that Mechanical Hadling Equipment was used in the appropriate context can be a good sign of good safety practice.

15. **RIDDOR Reportable** Indicates whether the incident was required to be reported to Health & Safety Executive (HSE)

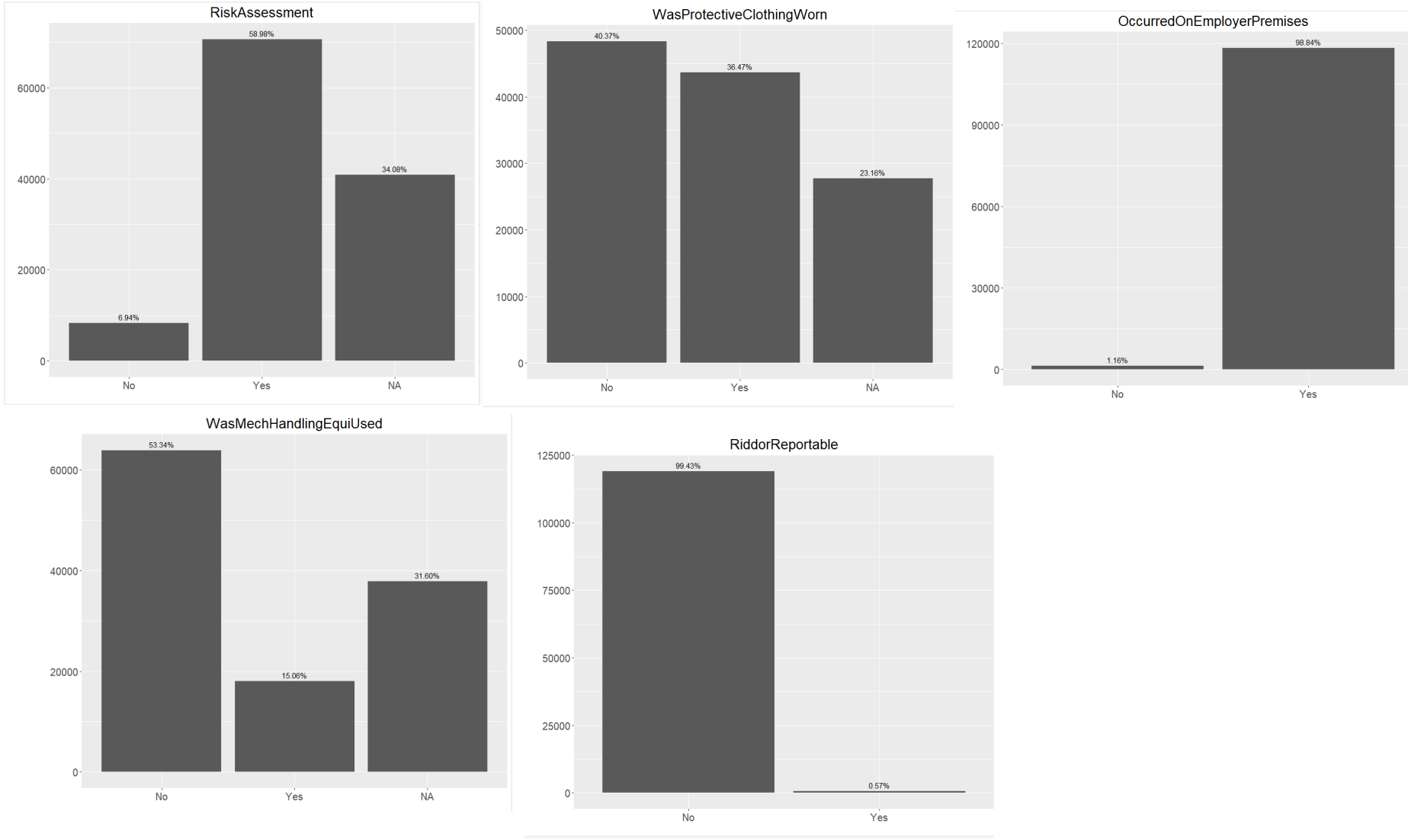

Figure 8: 5 Binary barplots with an indication of how much missing data there is. The 3 variables relating to management of safety standards have a significant amount of missing data.

16. **Total Days Unfit For Work** indicates how many working days have been missed as a result of the incident.

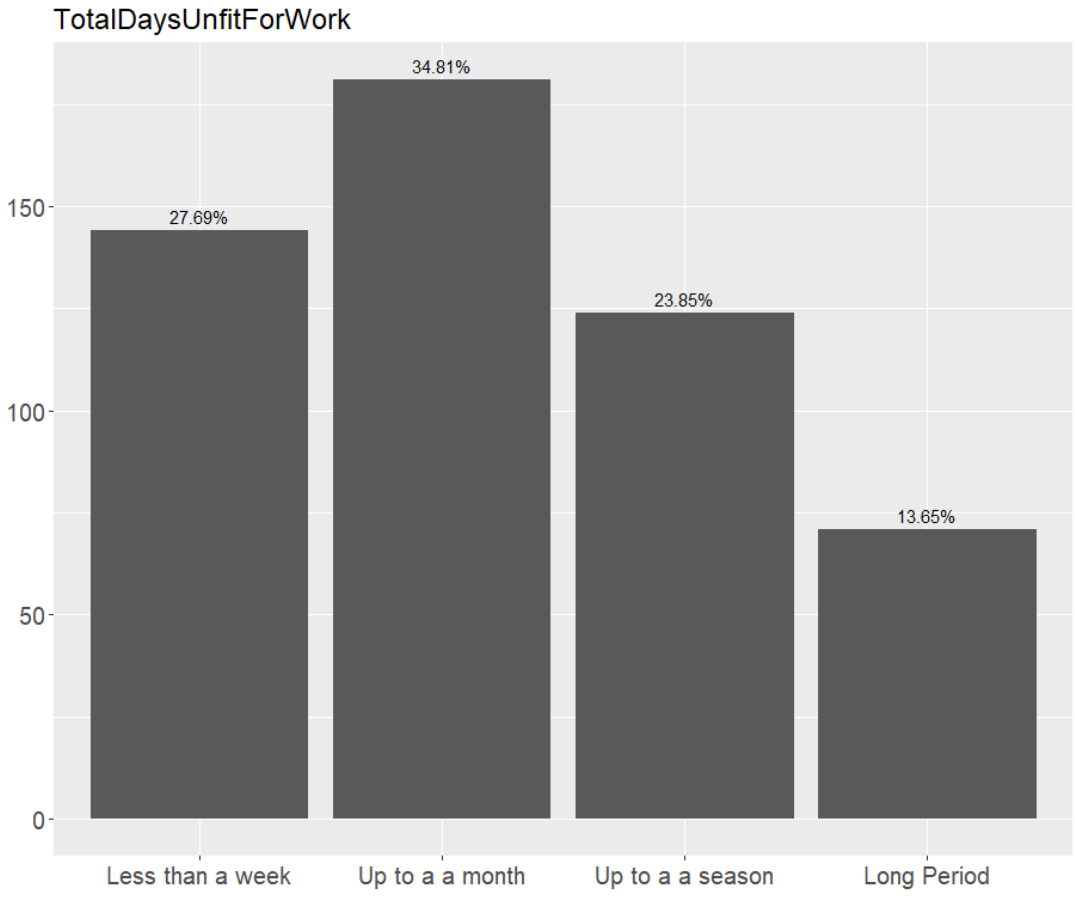

Figure 9: Total Days Unfit for Work for those **who have missed time off work**. There are many outliers which obscure the distribution, so a discrete representation has been chosen. The Time ranges for Total Days Unfit For Work represent respectively: 1-6 Days; 7-30 Days; 31-90 Days; 90 days or more.

17. **Injured Partys Age** indicates the age of the injured party. In figure 10, the age involved in the most incidents is around 50 year old. It falls down sharply after. People in the age 30 to 45 seems to have the same incident rate. The tails on both sides may be attributed to less workers being of this age.

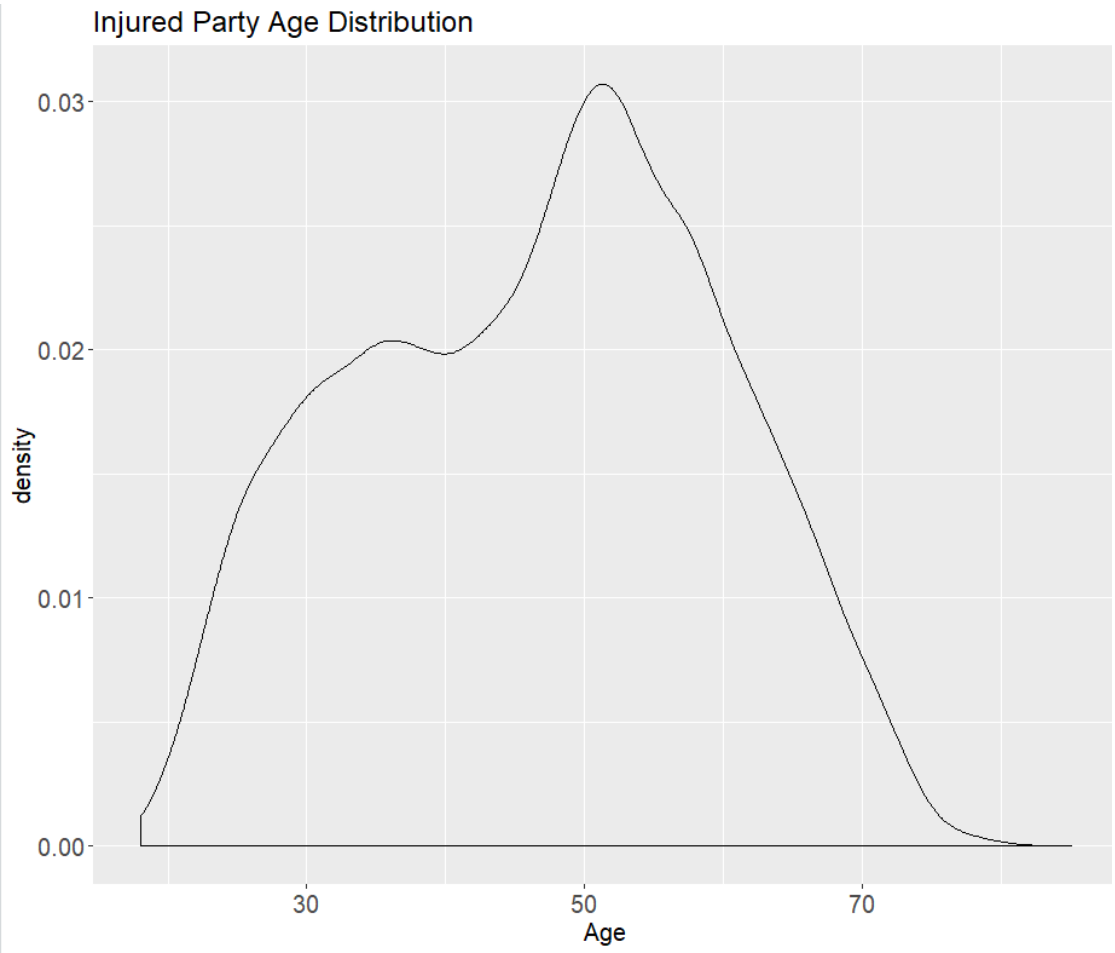

Figure 10: Injured Partys Age distribution in terms of a density plot taken over the whole dataset. The higher the density, the more likely an person of specific age was in an incident that was recorded in the dataset.

18. **Length Empl Cat** indicates how long they have been employed for.

19. **Brand** given to indicate the brand of Saint-Gobain. We observe Jewsons is by far the biggest brand by a considerable amount. As a result any pattern of accidents in the Jewsons Brand will obscure patterns in the other brands.

	Brand	Frequency
1	Jewson	84312
2	Graham	6556
3	Minster Insulation and Drylining	5623
4	International Timber	1605
5	Pasquill	1497
6	J.P. Corry Group Ltd	1085
7	Normans	929
8	International Decorative Surfaces	859
9	Gibbs and Dandy Group	778
10	George Boyd	750
11	Calders & Grandidge	646
12	Frazer	553
13	Ceramic Tile Distributors	546
14	Jewson Civils	527
15	Priority Plumbing	190
16	Chadwicks	141
17	Bassetts	125
18	IDS Independant	114
19	Ideal Bathrooms Ltd	83
20	Neville Lumb	74
21	World's End Tiles	21
22	Ideal Bathrooms Independant	6
23	Tile Depot	6
24	Blackpool Power Tools	3
25	PDM Ltd	2

Figure 11: Saint Gobain consists 25 brands, with Jewsons being by far the largest brand. This has resulted in it having a higher number of total incidents. Differences in working cultures and data collection may be reasons for differences in frequency

The following 4 variables dropped for having over 90% missing data data:

20. **Date HS Work Place Training** is the last recorded work place training. This is given as a bin. This training is "to ensure that employees have the necessary training, education and skills to reduce the level of risk that they are exposed to. This will also enable employees to identify potential losses and assist in the creation of a zero harm culture"

21. **Last Safety Audit** tells us how long ago safety audits were performed.

22. **Property Damage** tells us whether property damage was sustained as a result of the incident. In the dataset, 55 out of 2614 incidents, which had property damage, led to an injury. This combined with the high amounts of missing data makes it not very informative.

23. **Auth To Operate** indicates whether people have authorisation to handle the equipment.In cases where there are Injuries, "AuthToOperate" was put down as yes for the majority. This suggests the majority of incidents are perhaps more accidental and despite Authorisation to operate, incidents still occur. Having over 90% missing data is a challenge because

24. **Location** is a code given to represent the location. It is unclear what the location code exactly relates to due to the restircitons of General Data Protection Rights (GDPR). There are 1181 unique values making it hard to use without being able to capture any numerical relationships between the incident types and location. It is possible to identify certain locations with high amounts of severe incidents. As such I will not be able to incorporate it into my modelling.

    The following three variables give very similar information.

25. **EmploymentType** Assesses whether being employed by the company or being a non employer has a significant effect.

26. **InjuredPartyType** gives information on their relation to the comapny i.e whether they are full-time, part-time, contractors or members of the public.

27. **EmplType** goes into the type of work that the person involved in the incident does. It gives an indication of the intensity of their day to day work life and it is expected that they are at different levels of risk to accidents. For example, a HGB/LGV Driver will be at a higher risk than someone working in internal office sales. To avoid co-linearity between predictor variables I will omit the variable EmploymentType as it is very similar to the InjuredPartyType. The EmploymentType consists of two values: a Non-employer and a employer. There are only 569 non-employers recorded with 119173 employers recorded. As such for modelling purposes this will not be as useful.

    Additional variables (Which are not included in the modelling but may provide informative context) are included in the Appendix in section 7. These variables have been found to have minimal relevance based on the EDA and modelling.

It is not clear in many categories with missing values, whether the missing value is due to the variable not being relevant to the location or incident or rather because it is genuinely information that was not available. For example, in the use of mechanical handling equipment, there is a lot of unknowns in Office (As many managers who input the information may view this variable as not mandatory to fill in), but a higher proportion of "No" than in production processing.

The definitions of the variables have slight inaccuracies. For example, there are Lost Time Injuries (TF1) with 0 days lost from work. It could be assumed this to be human error as the managers who entering the information in the PeopleSafe system may interpret the classifications wrong or input the wrong amount of days. The clear rise in incidents after 2013 means there is a potential time effect with the data. However that is not something I will consider in my modelling. There has been changes in data collection over the ten year period as Saint-Gobain has reviewed its data collection practices and its safety regulations. [8]

3.3 Comparing TF4 to TF1-TF3

In order to simplify the classification problem I will limit the predictions to TF1s to TF3s.

The dataset has different characteristics, depending on whether it is a TF4 or a TF1-TF3. The distinct difference comes in the form of missing data, incident type definition and changes in data collection. In the TF4s, 7 variables have no data. This presents a big issue in the idea of MNAR, which means we cannot impute for missing data. This leads us to question the validity of combining the data from TF1-TF3s to TF4s.

Due to the system of reporting, the missing data in the TF4s are often caused by the information for many variables being optional to report. This creates a data imbalance where there is 36.67% Missing data in the TF4s as opposed to 20% in the TF1s and TF3s. The boxplots in Figure 12 show that the majority of TF1s-TF2s occur earlier in the data collection period. However, TF3s, TF4s seem to be collecting more often in the future. This hints towards a data driven initiative.

"Damage" and "Unsafe Practice" are categories that have been recently introduced. It seems RIDDOR Dangerous Occurrence has gone down overtime.

This is because the incident types Damage, Dangerous Occurrences and Unsafe Practice are more reflective of the time period in which they were collected and TF4s have issues with high missing data and ambiguous definitions. Furthermore removal of TF4s will reduce the effect of imbalanced data. The data is now 76% TF3s, 18% TF2s and 6%TF1s, compared to being 90% TF4s, and the other 5 Incident types being a comparatively small proportion of the rest of the data.

The definition for each type has been said to be subjective in the sense that two individuals recording an incident may interpret the same incident as different incident types. Thus there is inaccuracy in the strict classification. It may be possible that the classification of Minor Injuries (TF2) has decreased and a preference of classifying as either TF1 or TF3 was made. A possible change in the definition may have occurred due to a change of priorities in locating significant injuries.

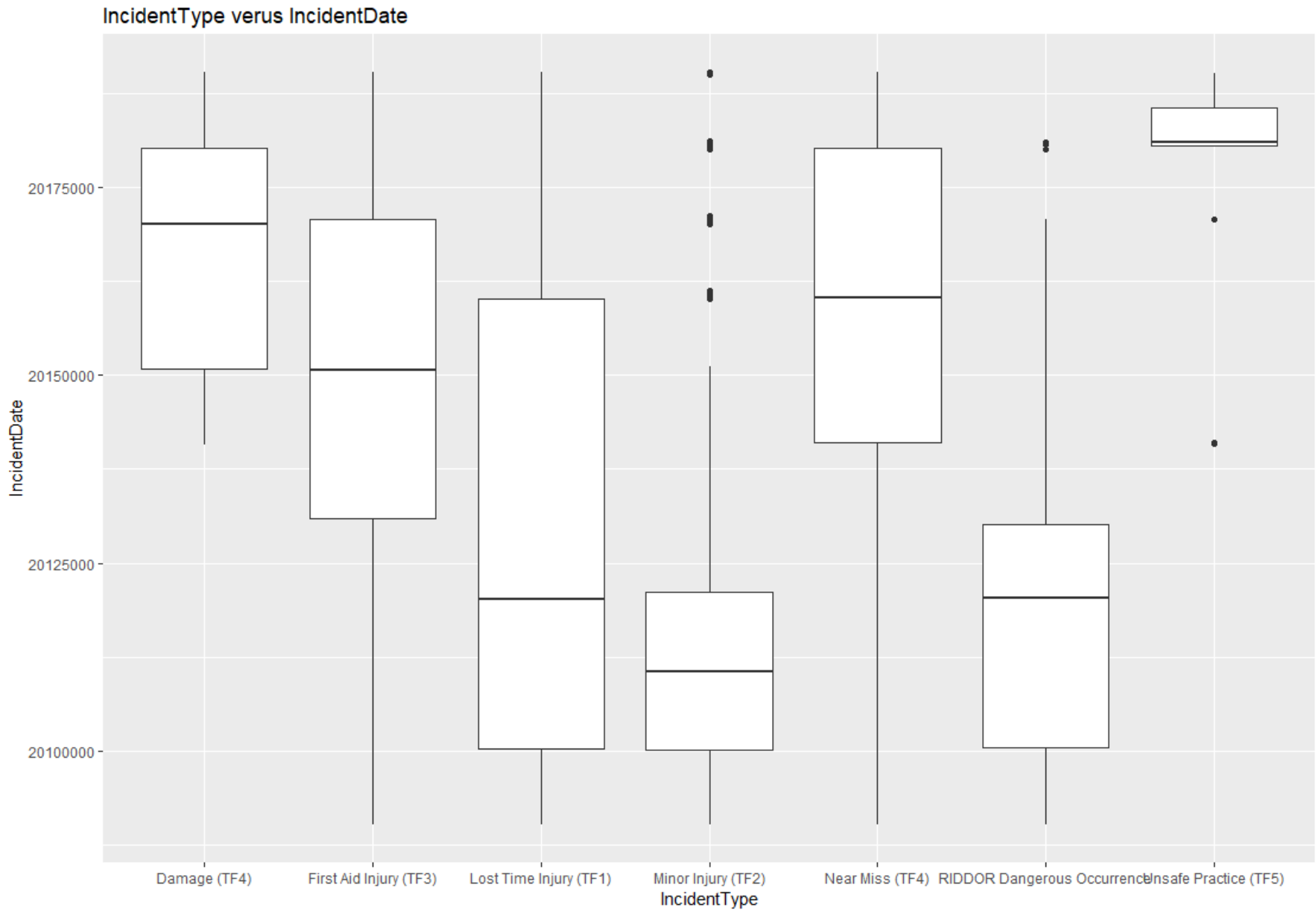

Figure 12: Boxplot showing how the collection of Incident Type has changed over time. The black dots indicate outliers, which are significantly greater than the standard deviation in each group.

From a computational perspective, it will be less intensive to compute data for a dataset 10 times smaller, as methods such as Random Forest which perform multiple bootstraps which is computationally expensive. The definition of being a near-miss (TF4) means it is plausible that a near-miss could easily be a TF1 or TF3 in more unfortunate circumstances so it may cause problems in classification.

TF1-TF3s

PropertyDamage	0.973757200768082	*Location*	0.0414977597610412
AuthToOperate	0.917324514614892	*WasProtectiveClothingWorn*	0.0381907403456369
LastSafetyAudit	0.610945167484532	*IncidentNumber*	0
EmplType	0.591956475357372	*IncidentType*	0
ru11ind	0.584062299978664	*IncidentDate*	0
DateHSWorkPlaceTraining	0.580861958608918	*IncidentTime*	0
LengthEmplCat	0.57008747599744	*Category*	0
Gender	0.512161297205035	*FloorConditions*	0
Ethnicity	0.429165777682953	*Weather*	0
Country	0.170044804779176	*OccurredOnEmployerPremises*	0
Brand	0.170044804779176	*EmploymentType*	0
RiskAssessment	0.149562620012801	*InjuredPartyType*	0
IncidentLocation	0.134947727757627	*TotalDaysUnfitForWork*	0
WasMechHandlingEquiUsed	0.0935566460422445	*RiddorReportable*	0
InjuredPartysAge	0.0514188180072541	*Day*	0

TF4s

DateHSWorkPlaceTraining	1	*Brand*	0.100972283782382
LastSafetyAudit	1	*IncidentLocation*	0.061345354067843
InjuredPartyType	1	*Location*	0.00452663642622054
Gender	1	*IncidentNumber*	0
LengthEmplCat	1	*IncidentType*	0
InjuredPartysAge	1	*IncidentDate*	0
EmplType	0.999981139014891	*IncidentTime*	0
PropertyDamage	0.971463329529701	*Category*	0
AuthToOperate	0.883307085129056	*FloorConditions*	0
ru11ind	0.539612783975707	*Weather*	0
Ethnicity	0.388951234923	*OccurredOnEmployerPremises*	0
RiskAssessment	0.3603108290346	*EmploymentType*	0
WasMechHandlingEquiUsed	0.340223879893247	*TotalDaysUnfitForWork*	0
WasProtectiveClothingWorn	0.252133648940484	*RiddorReportable*	0
Country	0.100972283782382	*Day*	0

Figure 13: Proprtion of Missing data of TF4s compared to TF1-3s. 1 indicates all data is missing, and 0.061 would indicate 6.1% of the data is missing. 6 Variables with no data in the TF4s

3.4 Data Collection Changes over time

There was a fatal injury in 2012 in a Jewsons branch which caused a change in the data collection methods in 2013. [12].

"With regards to health, in 2013, Saint-Gobain adopted a Health policy that is in continuity with the actions already undertaken by the Group. It establishes the guidelines of its action for protecting the health of its employees, its customers and users of its products, as well as for residents adjacent to its sites. All the Group's sites throughout the world have to implement it, in accordance with their local regulations and in addition to the health and industrial hygiene standards and tools already in place."

As a result ,starting in 2013, we see a change in data collection. The question that needs to be addressed is whether we can consider the data before and after 2013 as different datasets, or can we eliminate the factor of time. [8]

From my exploratory data analysis, I have only noticed a few differences between the data before and after 2013:

To do: Add percentages

- A large increase in data collection from 8009 Incidents to 111733 Incidents.
- Proportion of TF1s have gone down (11% compared to 6%) whilst the proportion of TF3s has gone up from 45% to 90%.
- Proportion of TF2s has gone down from 16.0% to 4.7%
- A significant increase in TF4s being reported from 61.2% to 90.5%
- The Introduction of recording incidents within a warehouse
- Average age of individuals involved in incidents has gone down from 43.38 to 41.57.
- The number of RIDDOR Reportable incidents has decreased increased from 277 to 402 however the number of non-RIDDOR Reportable incidents has increased from 7732 to 111331.
- The variables "LengthEmplCat" and "Gender" and "EmplType" have not been used pre-2013.
- A wider list of options for InjuredPartyType
- More onsite incidents are reported after 2013

Saint Gobain is a company that has annual reviews of safety regulations and hence it may have an effect on any changes in accident numbers

Dangerous Occurrences, Unsafe Practice and Damage all are collected predominantly in different yearly periods. This may be representative of the changing focus of Saint-Gobain's strategy. For example, Damage (TF4) and Unsafe Practice appear to be newly introduced Incident Types. TF3 and TF4s are collected mainly in latter years whilst TF1s and TF2s have a median Incident Date that is earlier in the time period. There is evidence to suggest that TF1s are going down over time.

In Figure 15, we can see the data collection changes over time.

3.5 Transforming features

Discussion with the decision maker is required to get a better understanding with the raw data. The Saint-Gobain Data Science lead, Teodora, was very helpful in providing context and was in contact over the project.

Transformation of features is appropriate in order to extract more meaningful patterns, or make it usable for specific algorithms. Clarification was required on whether "Not Applicable" and "Normal" held the same meaning as NA or not. Furthermore small typos in the data were required to be inspected and modified appropriately. For modelling it may be better to convert variables into a form where the information is easier to interpret or process. We will convert incident date into seasons. This is to group the data into a more flexible form. I have not looked at yearly changes as part of my modelling to simplify the approach.

One-hot encoding is encoding a categorical variable with a value of 1 or 0, which indicates their presence of absence (1 indicating the presence of that variable). It will often lead to sparse matrices. XGBoost, an algorithm I am using, requires these sparse representation.

I will explore the option of converting "TotalDaysUnfitForWork" and "InjuredPartysAge" into "bins" in order to have a fully discrete dataset. The choice of the width of the bin will require the consideration of a tradeoff between having a bin that is very informative of the age (small width) and having a bin that has a good sample size (large width). The choice of a small bins may make it more accurate for prediction, but not generalise well as a large bin. [34]

There has been relatively low amount of research into this question. As such, I have opted for bin widths that are commonly used for that category.

4 Variable Selection and Missing Value Analysis

4.1 Variable Selection

Feature selection is the process of selecting a subset of relevant features or attributes as dependent attributes in a predictive model, thus leading to reduction of model overfitting and improved prediction accuracy. There is room to explore how models have the ability to identify important variables. Variable selection can be done through modelling and exploratory data analysis

There is an argument for us to remove some of the variables for the following reasons:

1. Lack of informative properties
2. Highly correlated with other predictors (Colinearity)
3. For the goal of parsimony we may want to remove predictors to make the model simpler. Parsimony will come into conflict with potentially removing useful info, so we may want to do statistical tests through modelling to show irrelevancy. We can also use exploratory data analysis to get a feel for how the variables affect the severity of the incident type.
4. To provide a recommendation to Saint-Gobain to remove this variable if it takes a lot of time to collect or ask for improved data collection methods
5. Heavily missing data
6. Inconsistencies in collection or methodology

Firstly, Incident Number and Incident Date are highly correlated, as the incident number is an automatic number assigned according to when it was inputted. The relationship is not completely linear only due to cases where inputting of the details occurs a period later after the incident. For example, this may be due to the accidents occurring over the weekend, and the person inputting on the following day does not input each accident in their chronological order. Furthermore, 103 incidents were omitted based on there not being an incident number.

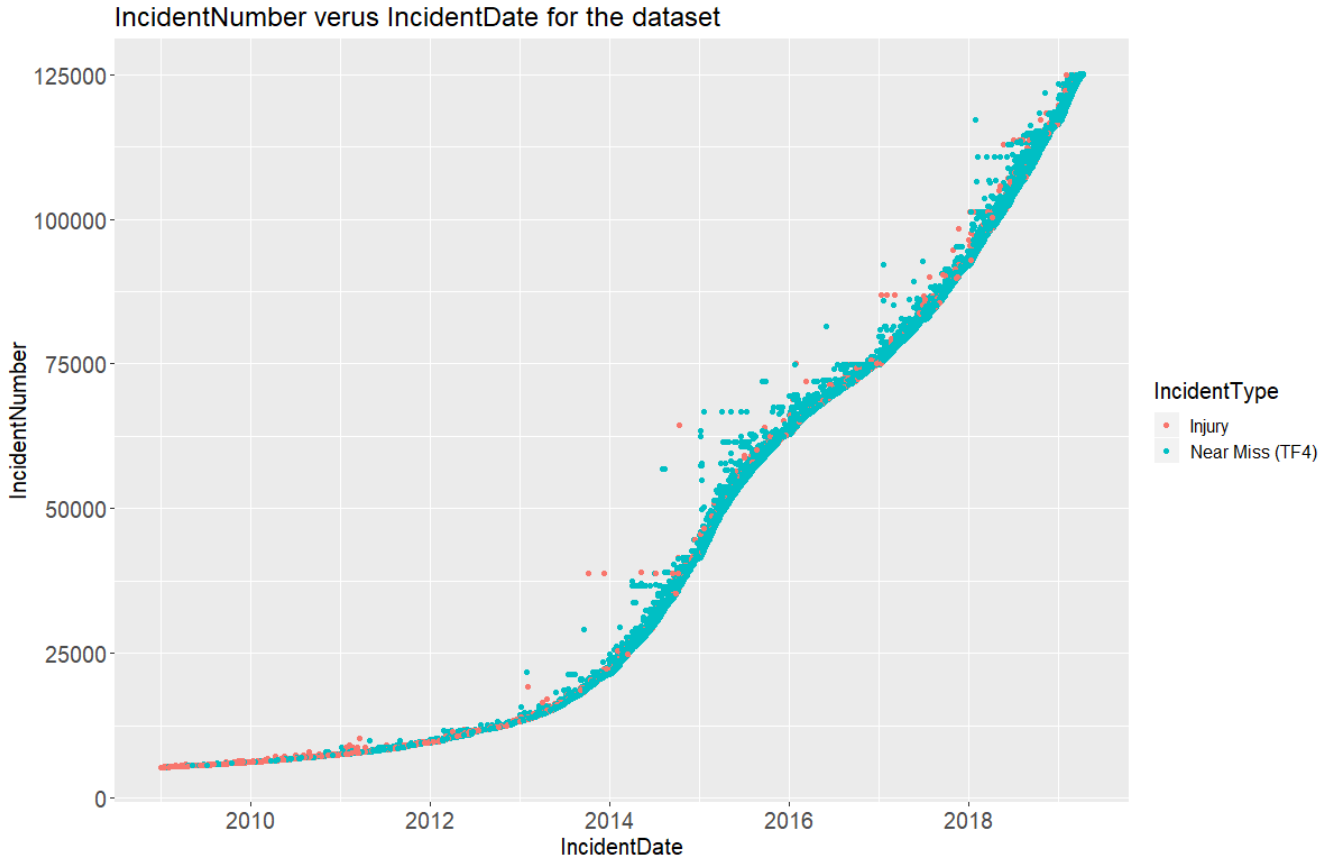

Figure 14: The Incident Number is logged with each incident chronologically. The points outside of the curve are the points which have been logged at a much later date. (I.e have a larger incident number than the majority around the same incident date. We can observe the time lag here. The maximum seems to be around a year. Typically it is a few days but sometimes weeks or month. After 2013, many incidents are logged many months or years after the accident occurred. This may be due to further investigation or administrative issues. It appears on average that incidents leading to injuries are logged later on based on the plot. The longest delay is just over a year.

4.2 Missing Value Analysis

Due to systematic and human error, The dataset has a high level of missing data. In the literature in morphometrics, 5% missing data resulted in 50% of the samples being affected. [30] However, Yanjun Qi states that good results can be obtained even with high amounts of missing data, if there are variables in the dataset which can provide similar information. [28]

It seems there is conflicting literature on how much missing data is too much and it may need to be taken on a case by case basis. It is important that the missing value imputation is accurate (in terms of producing data that converges to a distribution).

LastSafetyAudit	97
EmplType	96.8
DateHSWorkPlaceTraining	96.7
LengthEmplCat	96.6
PropertyDamage	96.4
Gender	96.2
InjuredPartysAge	92.6
InjuredPartyType	92.2
AuthToOperate	87.7
ru11ind	54.4
Ethnicity	39.1
RiskAssessment	34.1
WasMechHandlingEquiUsed	31.6
WasProtectiveClothingWorn	23.2
Country	10.6
Brand	10.6
IncidentLocation	6.5
Location	0.8
IncidentNumber	0
IncidentType	0
IncidentDate	0
IncidentTime	0
Category	0
FloorConditions	0
Weather	0
OccurredOnEmployerPremises	0
EmploymentType	0
TotalDaysUnfitForWork	0
RiddorReportable	0
Day	0

Figure 15: Percentage of missing data in the TF3s

The algorithms I will use do not have inbuilt capabilities to deal with the missing data hence I will need a dataset with no missing values. I will describe the types of missing data and give an explanation as to why data may be missing. Lastly, I will explain how I will use MICE to impute the missing values.

Missing data comes under 3 main categories:

1. Missing Not at Random (MNAR): The missingness is explained by a reason that we either do not know or through patterns in the data. This type is not workable with MICE.

2. Missing at Random: (MAR): The missingness is completely coincidental and can be solved by prediction based on the other data.

3. Missing Completely at Random (MCAR): Whether or not the person has missing data is completely unrelated to the other information in the data.

Many details are optional depending on whether an accident led to an injury, meaning a lot of missing data is missing not at random. Furthermore, if injured parties are not employees, personal information seemed to not be necessary. Other reasons such as the variables not being relevant to accident may mean managers choose to not fill in the information.

For MICE, MAR and MCAR are not an issue and can be imputed with good accuracy. Due to large amount of missing data in TF4s (see Figure 15) are MNAR, it is a good reason for removal and focusing on TF1-TF3s.

4.2.1 MICE algorithm

Multiple Imputation by Chained Equations (MICE) is an algorithm which imputes the missing data multiple times, using a specified model. This is based on a Gibbs sampler, where the next value of the (previously) missing data is updated based on the current imputed data. With an appropriate model MICE can impute missing data for all types of data. Initially I have done multiple imputation on the smaller TF1-TF3 dataset. [32]

It is important for us to check this either mend it via manual or multiple imputation or remove unusable data. It is also important to check whether the missing values in a variable will correspond to extra information or valuable missing data, as that will guide my approach. This is an interpretation of whether the information is MNAR or MAR/MCAR. I can check this by seeing whether there are correlated variables in the data, and seeing how the model changes under different conditioning. (CART).

The chained equation process can be broken down into four general steps [15]:

1. Values are imputed for every missing value. These represent placeholders. Popular options include the mean.

2. For one variable x, the placeholders are set back to missing

3. Observed values from variable x are regressed as the response, on the other variables in the imputation model. The regression model would require the same assumptions as when modelling on a full dataset with no missing values. The model used is a Classification and Regression Tree.

4. Missing values in variable x are then predicted from the model . When other variables are being imputed for their missing data, any variable used as a predictor which contains previously imputed points, will take into consideration those previously missing imputed points for forming the model.

5. Steps 2-4 are then repeated for each variable with missing data. One iteration would involve imputation for each variable,

6. Step 2 to 4 are repeated for many iterations. The imputations for each variable is updated at each iteration. after a certain amount of iterations, we obtain an imputed dataset. The aim is to achieve convergence of the coefficients, which means the order in which the variables are imputted does not affect the values of the coefficients.

4.2.2 Classification and Regression Trees (CART)

The model chosen for mice is **Classification and Regression Trees CART.** This is because it works sufficiently well for regression and classification. Main benefit it runs fast for large datasets, as it has no assumptions on the structure of the data. Missing values do not effect the building of the tree, as they use surrogates which are similar data points, modelled on the existing data, to make a prediction on the missing values. [35]

Classification and regression trees (CART) was originally described by Breiman in 1984 [24] and is a popular way to identify structure and relationships in high dimensional data. As the name suggests, it uses classification trees on discrete outcomes and regression trees on continuous outcomes. A iterative algorithm minimises the within-group variance through partitioning of individuals. [16] [13]

4.2.3 Mice Analysis

There a few important assumptions of mice.

1. We need the variables to fit with the chosen imputation method (For example, using Linear Regression in MICE would require a continuous response).

2. The method needs to be able to take numerical and categorical predictors.

3. We also cannot impute data that has too many missing data or else there will be problems or it won't converge or impute accurately.

   After fulfilling these conditions, we need to assess whether the missing data is truly accurate, and does not have glaring affects on the data. One idea to measure this is to assess the convergence. If the imputation produces similar results for each variable every time, it is on the right path.

   We want convergence to a distribution, not a point, as we are looking for an accurate representation of the whole dataset. Thus, each chain should be stationary. Therefore the posterior densities from each chain, after burn in,should look the same. The variation between values within one chain should similar to the variation between values in different chains. More formally, we would expect the variation between values in one chain (within chain sum of squares) to be similar to the variation between values in different chains (between chain sum of squares).

The classical diagnostic for MICE would be to check trace plots of the mean and variance against the iteration number. On convergence, the different streams should be freely intermingled with one another, without showing any definite trends. Convergence is diagnosed when the variance between different sequences is no larger than the variance within each individual sequence. Inspection of the streams may reveal particular problems of the imputation model. A pathological case of non-convergence occurs with the plots in Figure 16

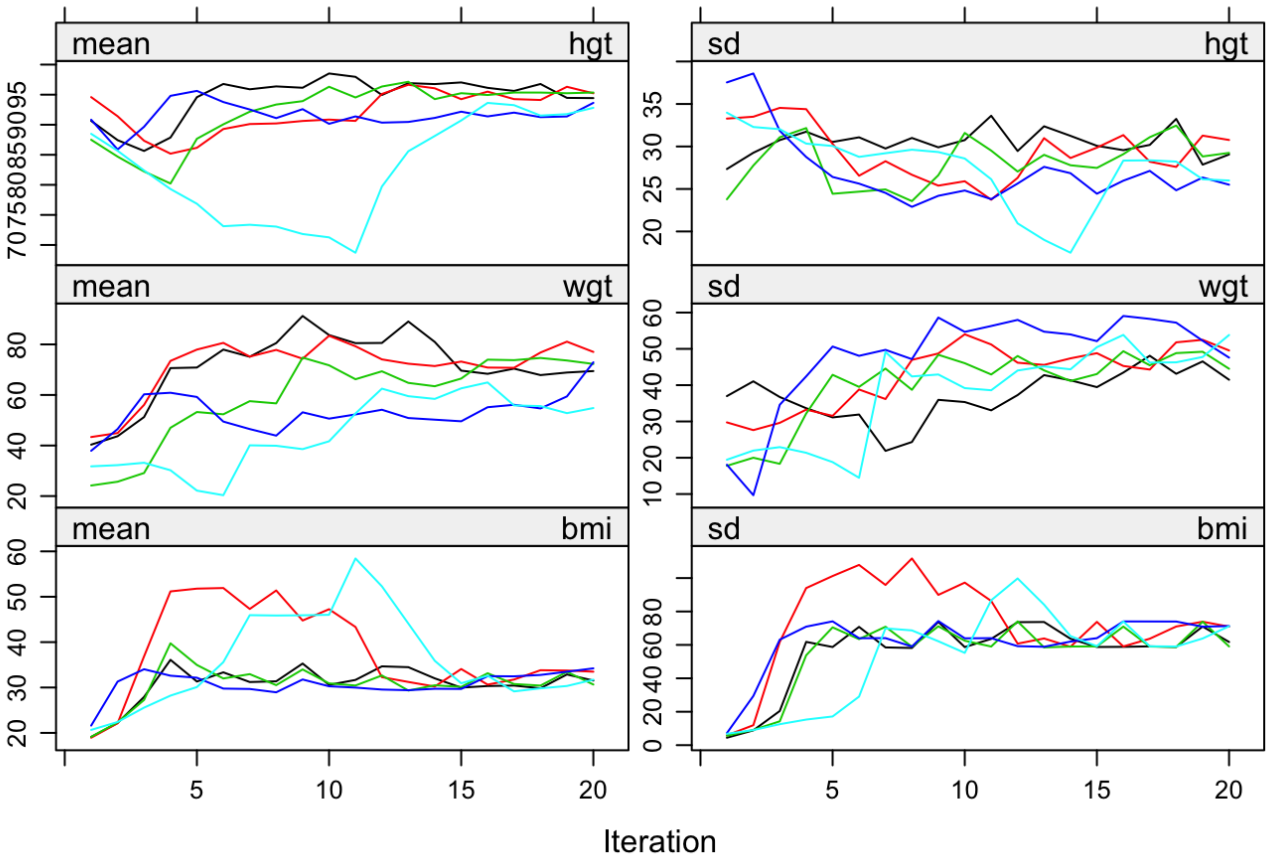

Figure 16: Non-convergence of the MICE algorithm caused by feedback of bmi into hgt and wgt. Each colored line represented a run of the MICE algorithm, plotted against its average value. The non-convergence is characterised by the runs not overlapping and mixing. Formally, convergence is found when the variance between different sequences is no larger than the variance within each individual sequence.

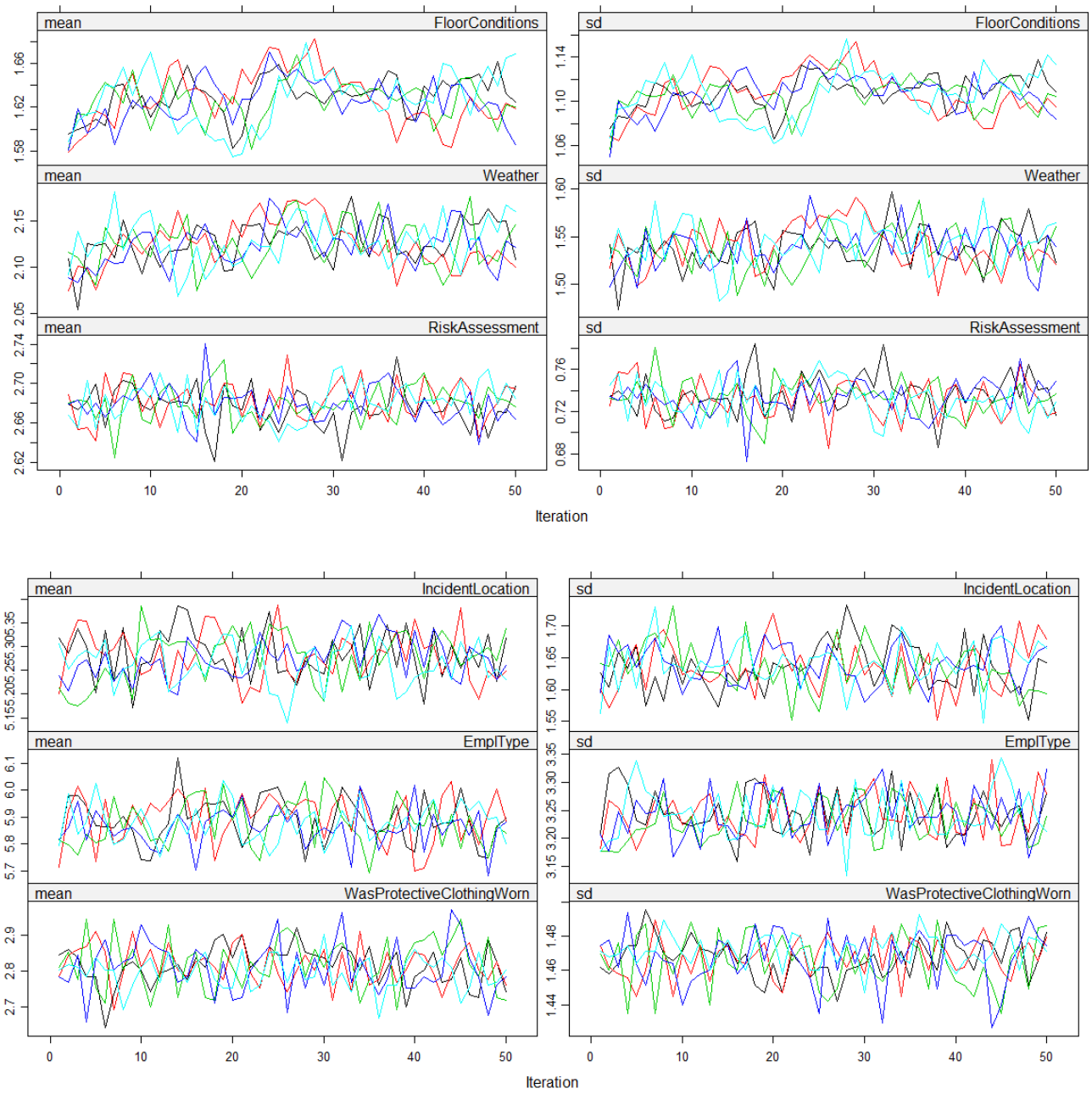

Figure 17: The Missing data is computed on the dataset containing only TF1s, TF2s and TF3s. We have also removed 13 variables which were previously deemed non-informative or have had too much missing data. Diagnostics of the MICE can be shown in the following plots. Now applying MICE to the dataset of TF1-3s, using CART we obtain an imputed dataset. These convergence plots show 5 runs of the chain. The trace plots are the average values of each variables In the following plots we see that there is relatively good mixing.

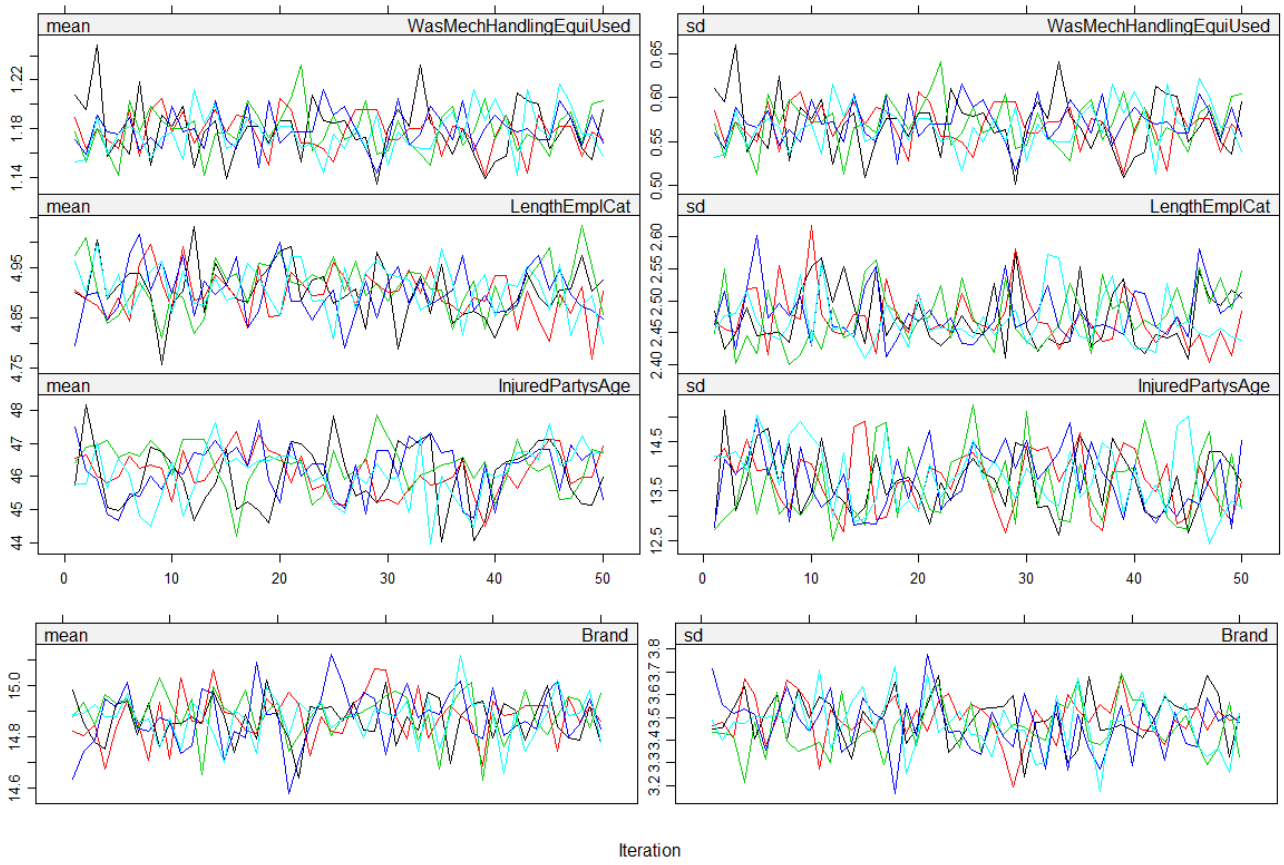

Figure 18: However the majority of the missing data is categorical, so looking at the diagnostic plots should be taken with caution. The means and averages are assuming that the categories are ordered in levels. A few of the variables example being brand do not have a clear ordering.

In summary I have taken the steps:

1. Removed 30 observations with no information
2. Removed variables with over 90% missing data
3. Removed variables which are correlated with other variables or appear very irrelevant based on EDA.(Based on filtering by Incident Type and looking for any meaningful differences)
4. Fixed missing data by MICE with CART.
5. Assessed the completion of the missing data.

5 Modelling

To summarise our completed dataset, we have 9374 datapoints to work from, with 16 predictor variables and aim to predict the IncidentType (TF1s, TF2s and TF3s)

We have filled in missing data for:

- Brand
- InjuredPartysAge
- LengthEmplCat
- WasMechHandlingEquiUsed
- WasProtectiveClothingWorn
- EmplType
- IncidentLocation
- RiskAssesment
- Weather
- FloorConditions.

There will be implementation of various modelling techniques with two purposes. **Accuracy** and **Interpretability**

Bayesian Networks will have an emphasis on interpretability whilst **Random Forest and XGBoost** will have more focus on accuracy.

Integral to my modelling will be evaluation of confusion matrices.

A confusion matrix is a table which compares predicted values with the true values. The quality of the predictions can thus be interpreted through different angles.

1. **Accuracy** is the $\frac{\text{Total number of correctly predicted cases}}{\text{Number of observations}}$
2. **Precision**, also known as the recall, is the $\frac{\text{Number of true positives}}{\text{Number of positive predictions}}$
3. **Sensitivity** is the $\frac{\text{Number of true positives}}{\text{Number of actual positives}}$

5.1 Random Forest

A Random Forest is an ensemble algorithm which creates multiple decision trees (CART) and then combines the output generated by each of the decision trees. Each tree outputs a class prediction, and the class with the most votes becomes the models prediction. (Majority Voting). The idea is that via this ensemble method, the trees protect each other from their individual errors, as although many trees may output the wrong classification, the majority vote should be the accurate classification.

Fundamental to the splitting of the tree is the Gini Index.

5.1.1 Gini Index

The Gini Index to determine which attribute to split on during the tree learning phase. The Gini index measures the level of impurity or inequality,which refers to a variables impact in the splitting the samples assigned to a node which was based on a split at its parent. The more homogeneous a node is, the smaller the Gini value is. The gini value of two child nodes is less than the parent node if a specific feature is used to determine the split. A feature's Gini importance in a single tree is then defined as the sum of the Gini index reduction (from parent to child) where this specific feature was used in the splitting. The overall feature importance in the Random Forest is thus defined as the sum or the average of its importance value among all trees in the forest.

Gini Index :

$$i(t) = 1 - \sum_{i=1}^{c-1} p(i|t)^2$$

Change in Gini Index :

$$\Delta i(t) = i(t) - \frac{N_L}{N} i(t_L) - \frac{N_R}{N} i(t_R)$$

Where c = Number of child nodes

t = training data Where p(i) is the probability of an object being classified to a particular class.

N_L being the number of samples reaching the left node

t_L being the node on the left split

t_R, andN_R are defined similarly. [37]

5.1.2 Decision Trees and Random Forest

A **Decision Tree** is a classification model which works on the concept of Gini index at every node. The decision tree will classify data points at each of the nodes, by partitioning the data and finding the variable which achieves the most separation and check for the change in Gini index at each node. It will then classify at the node where change in Gini index. Because decision trees are very sensitive to their training data, small changes in training result in different tree structures. Decision trees are very biased models.

A decision tree considers every possible feature, picking the one that produces the most separation between the two nodes. One issue is that the trees may be very correlated, is not as usual for model prediction, because if they are correlated, each tree will be using similar information which will lead to the trees predicting the same output. If this prediction is correct then it can lead to high training accuracy however if it is not accurate on test data it can lead to the model perofrming very poorly. (High generalisation errror)

However each tree in a random forest can pick only from a randomly sampled subset of features and data points. A low correlation between models is important for the majority voting mechanism to produce accurate predictions, or else the trees will tell the same information. The Random Forest uses **bagging** (bootstrap aggregation), in the form of its majority voting mechanism.

We take bootstrap samples, where we randomly sample from the dataset used for modelling with replacement, and form a tree. We do n of these bootstraps and train individual trees based on each bootstrap sample. This collectively forms a random forest. Then we output the bagged output (Majority vote)

$$f(X) = \sum_{i=1}^{n} f_i(\hat{X}_i)$$

Where f_i is the i-th tree, formed using the i-th bootstrap ($\hat{X}_i$

There is a potential downside where if the majority voting outputs the wrong classification. we may find that a random forest may actually make the error worse in which case a decision tree may be more suitable.

The performance is tested on an out of bag sample. The out of bag sample is the set of data points which were not used in the training of respective decision tree. This allows the samples to be tested on new data and reduces the need for cross validation and a test and training data split. The cutoff will be used to factor in the imbalance in the data. It aims to balance the population sizes, because the majority voting may have a preference to favour a class with more samples. A vector of length equal to number of classes. The 'winning' class for an observation is the one with the maximum ratio of proportion of votes to cutoff. The suggest option by the documentation [6] is $frac1k$, where k is the number of classes (i.e. the majority vote wins)

5.1.3 Hyper Parameters

To generalise best on new data, there are tuning parameters.

The **number of trees to grow** should be large enough to ensure every input row gets predicted at least a few times. But at some point the extra trees may start overfitting and finding no or minimal improvements. We can decide this parameter via error plots and seeing when the errors converge.

The **Mtry** is the number of variables randomly sampled as candidates at each split. We want to find the mtry that leads to the lowest class error. Higher mtry allows more expression in the prediction whilst lower may lead to less correlated trees.

These two variables will also determine how costly the predictions will be to produce.

To classify IncidentType using a Random Forest, we will work on the following criteria:

- The individual class error (Precision and Sensitivity)
- The total error
- What variables seem important
- The number of trees to grow
- The number of variables randomly sampled as candidates at each split (Mtry).

MeanDecreaseGini uses the Gini Index, a metric used in decision trees.
Recall, Gini Index = $i(t) = 1 - \sum_{i=1}^{c-1} p(i|t)^2$
If $i(.)$ is Gini index, then $I(.)$ is called Mean Decrease Gini function.

$$I(X_k) = \frac{1}{M} \sum_{m} \sum_{t} \frac{N_t}{N} \Delta i(t)$$

M is the number of total number of trees int he Random Forest Model
N_t is the number of data points used for node t, based on variable X_k
N is the number of data points in total [39]

5.1.4 Variable Importance

It is said that the random forest is viewed as a black box because of its sheer size. [36] However we can still interpret the importance of each variable to the model.

It is important to know which variables are important. Reducing the number of uninformative variables that a tree can sample from will reduce the number of trees producing bad outputs. This should improve the overall accuracy. There are several importance measures to consider.

- MeanGiniDecrease
- Mean Min Depth
- AccuracyDecrease
- Number of Nodes
- Number of Times A Root

The **Gini decrease** is the decrease in Gini in the absence of this variable. At every split one of the Mtry variables is used to form the split and there is a resulting decrease in the Gini. The sum of all decreases in the forest due to a given variable is normalised by the number of trees.

The **Mean Decrease in Gini** is a measure of variable importance for estimating a target variable (X_k. It is the mean of a variables total decrease in Gini impurity, weighted by the proportion of samples reaching that node in each individual decision tree in the random forest.

Mean Decrease in Gini measures the predictive power qualitatively but it does not have a coefficient like interpretation. Having twice the decrease in Gini does not mean the variable is twice as important. [28]

The **Mean Min Depth** is the average minimum depth before the variable is chosen at a node. A lower Mean Min Depth will indicate a higher variable importance. [1]

$$\text{Mean Min Depth} = \frac{1}{B^*}\sum_{b=1}^{B^*} min(depth(T_b^i)), i = 1 : p$$

$depth(T_b^i)$ is the depth of the i-th variable in the b-th random forest tree

The **Accuracy decrease** is the decrease in accuracy in the absence of this variable. For each tree using the observations not selected, randomly permute the values for the j-th variable. We then evaluate each tree and see if the error rate improved and by how much.

Number Of Nodes is the number of average number of nodes it is chosen to split, more splits in discrete categories

Average number of nodes for variable i is =

$$\frac{1}{N}\sum_{n=1}^{N} N(T_n^i), i = 1 : p$$

Where $N(T_b^i)$ is the number of nodes where the i-th variable is used to split the b-th random forest tree.

Number of Times A Root is the number of times the variable appears as the root node of a tree.

5.2 Random Forest Results

I find the TotalDaysUnfitForWork to be the most important feature, however it would not be appropriate to include it for predicting incident type because we would not have access to the days missed before the incident occurs. Due to the out of bag values, cross validation is not required to estimate errors.

To optimise the mtry, a search was performed, until the out of bag error estimate did not improve by 0.1%. (mtry initialised at 2)

Based on the tables in Figure 20, setting mtry to 5 and randomly sampling 5 variables at each split seems best Higher mtry do not gain good results. This may be linked with the idea that the first 5-6 variables are a lot more important than the other ones. Thus setting mtry as 5 gives us a relatively good chance of each tree finding at least one of the variables, leading to a good prediction.

We find similar performance between a full discrete dataset and keeping Total Days Unfit For Work and InjuredPartysAge continuous.

ntree = 500, mtry = 2	test data			
Accuracy	TF3	TF2	TF1	Overall
RandomForest no cutoff	99.83%	89.21%	99.64%	81.54%
RandomForest with cutoff	88.78%	73.68%	97.20%	78.16%
Discrete RandomForest	87.17%	74.88%	95.48%	76.77%
Discrete RandomForest with removed variables	93.70%	88.88%	95.48%	73.96%
Randomforest with removed variables	93.53%	87.68%	95.48%	79.33%

Figure 19: Table of Random Forest results showing the effect of having a cutoff, comparison with a discretised dataset and removal of some variables (TotalDaysUnfitForWork)

mtry	TF3	TF2	TF1	Overall
2	87.17%	25.12%	95.48%	76.77%
3	72.03%	50.48%	95.48%	70.25%
4	66.67%	57.25%	95.48%	66.92%
5	63.59%	60.63%	95.48%	65.17%
8	60.95%	61.35%	95.48%	63.26%

Figure 20: Table Of Random Forest Results comparing the different mtry done on test data.

The algorithm classifies TF1 and TF3s with high sensitivity (87.79% and 80.77%) and precision (64.32% and 4.84%), but has low specificity (59.90%) and precision (28.77%) on TF2s. As we increase the number of variables sampled at each split (mtry), we find that the TF3 accuracy goes down from 87.17% whilst the accuracy of TF2s goes up from 25.12% to 1.35%. The TF1 accuracy does not get affected by change in mtry by more than 0.1%.

mtry = 2 ntree = 500	TF3	TF2	TF1
Precision Rate	84.01%	5.80%	69.81%
Recall	87.51%	11.32%	95.48%
mtry = 5 ntree = 500	TF3	TF2	TF1
Precision Rate	87.79%	28.77%	80.77%
Recall	64.32%	59.90%	94.84%

Figure 21: Table of Random Forest results looking at the precision and recall.

Variable Importance Plots were performed on a Random Forest with 200 trees and 5 variables randomly sampled at each split.

The most important variables for predicting Incident Type are

- RIDDOR Reportable
- Weather
- Injured Partys Age
- LengthEmplCat
- Incident Location.

From the Variance Importance plots, Injured Partys Age had the greatest Gini Decrease. RiddorReportable had a high accuracy decrease. Riddor Reportable was very often used as a root node. This is because it is a very good indicator of a TF1 incident as the nature of the definition indicates an incident severe enough to be reported to HSE. Likewise, Incident Location, Category and Floor conditions were often root nodes. The variable importance plot between mean min depth and times a root were often very linked. There is a clear relationship between these two factors.

Variables which did seem to struggle were variables relating to the safety procedures. This is interesting because you would expect poor safety practice would be a reason for incidents however perhaps the safety standard is at a suitable level, and only rare incidents get flagged up regardless of safety practice.

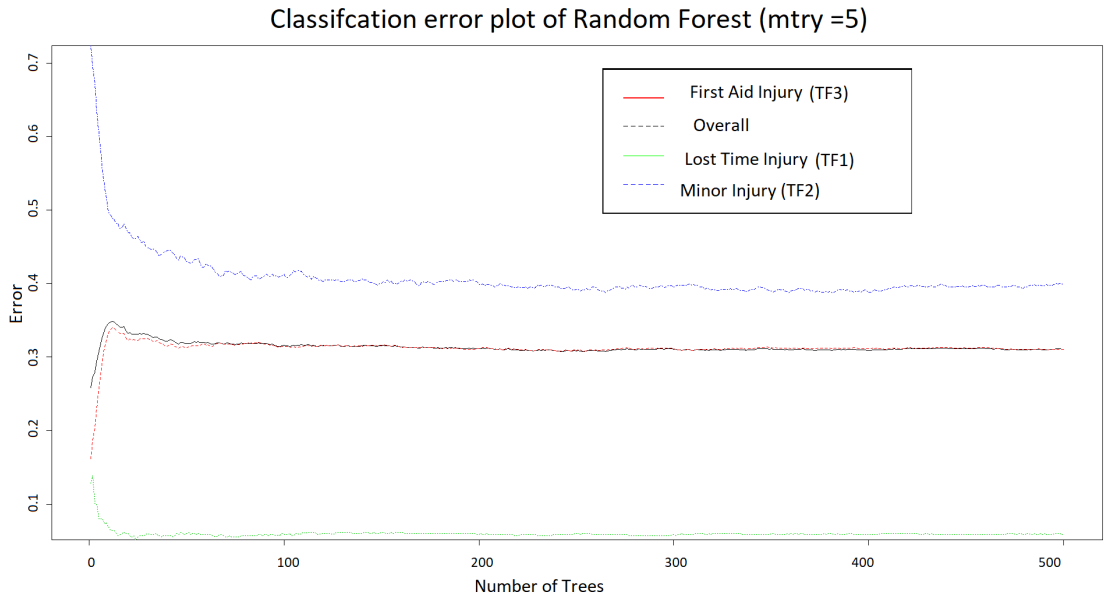

Figure 22: The Overall Error rate is characterised by the First Aid Injury (TF3) Rate as it is the largest class. The errors start to converge after around 100 trees. The Lost Time Injury finds the lowest error, whilst the error increases in TF3s and decreases in TF2 as the number of trees increases.

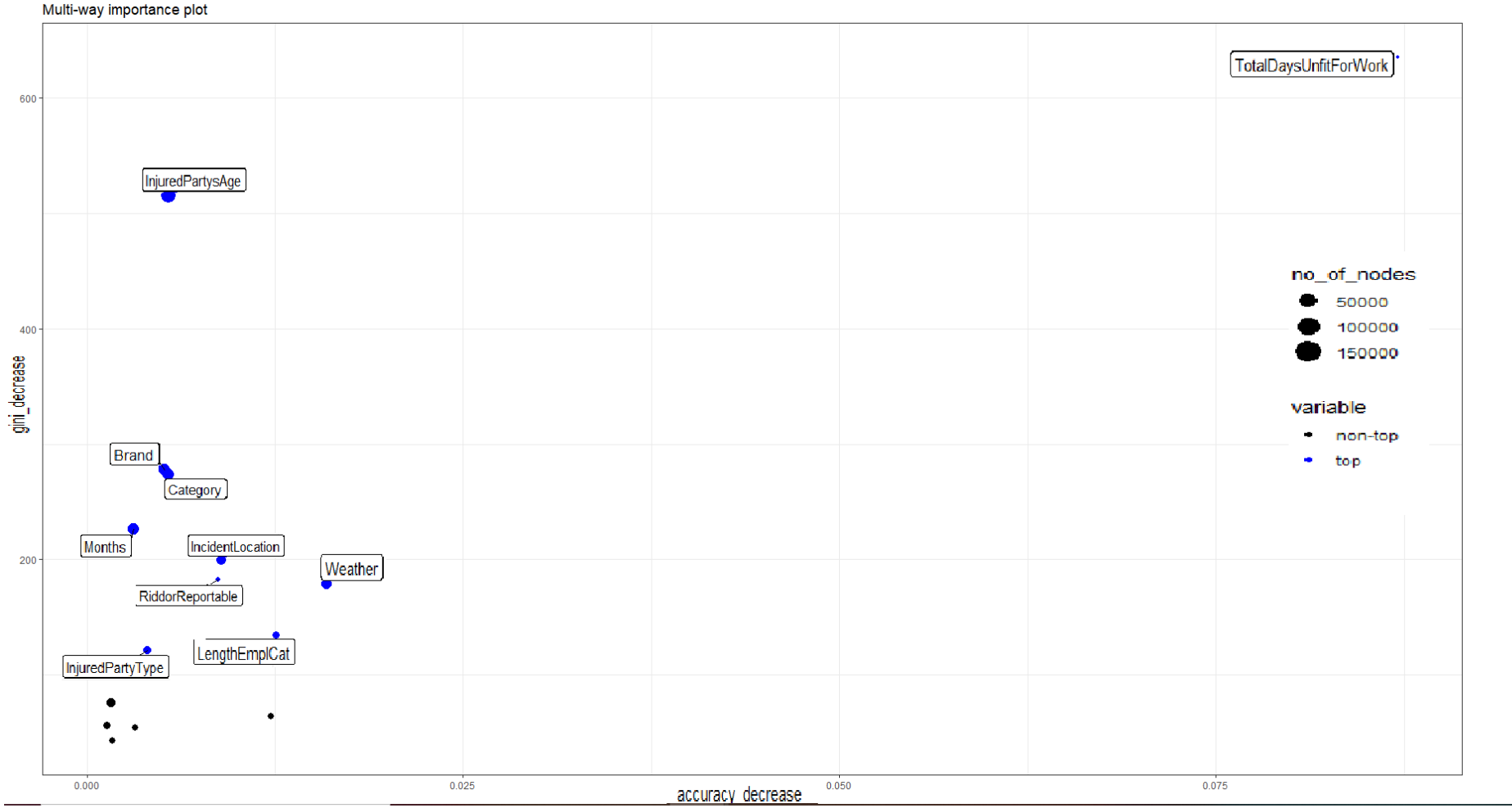

Figure 23: Variable Importance Plot of Full dataset comparing the accuracy decrease with the gini decrease. The 10 most important variables are highlighted blue.

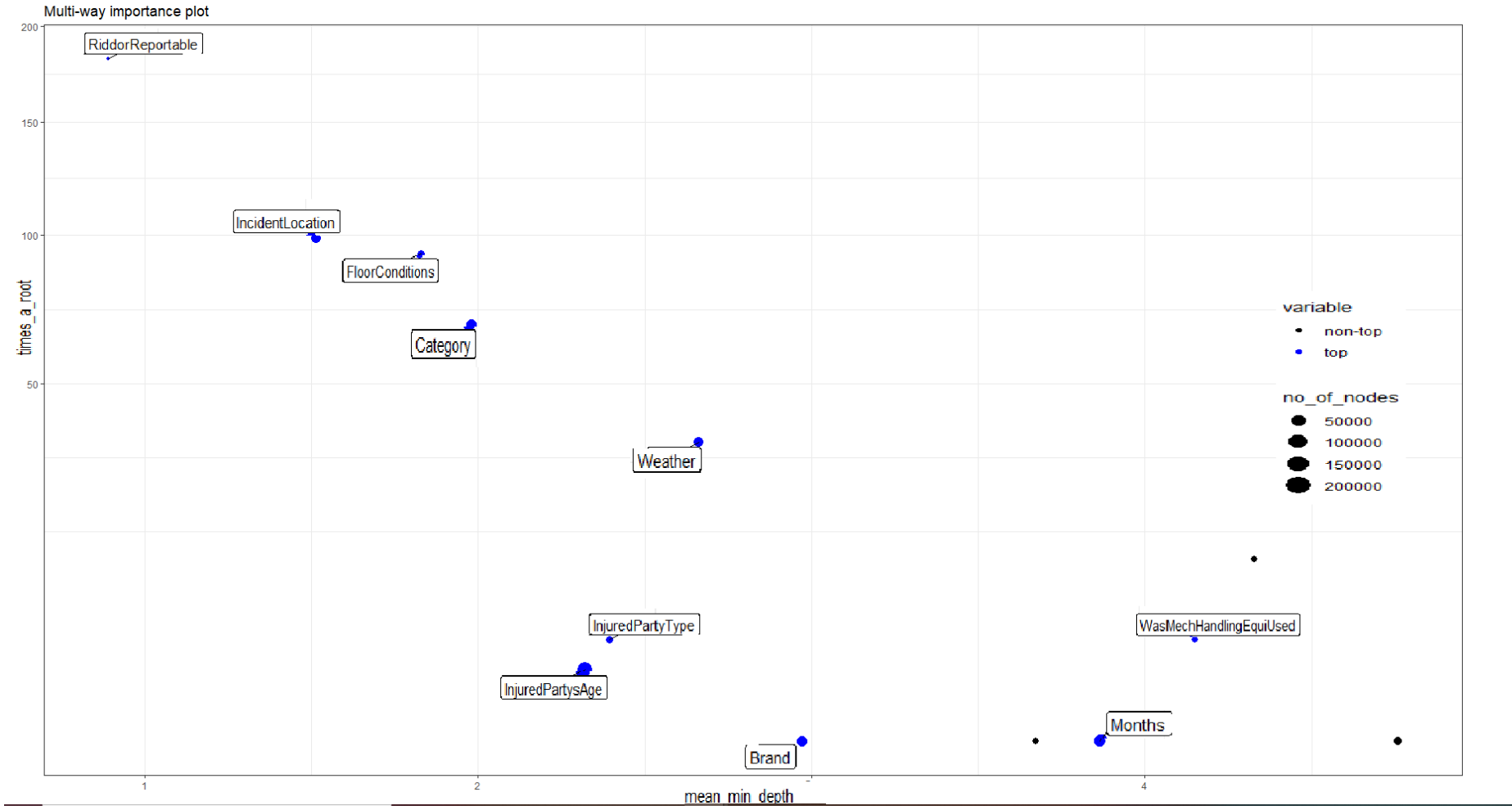

Figure 24: Variable Importance Plot of dataset without dominant variables 2 comparing the mean min depth with the number of times each variable appears as a root.

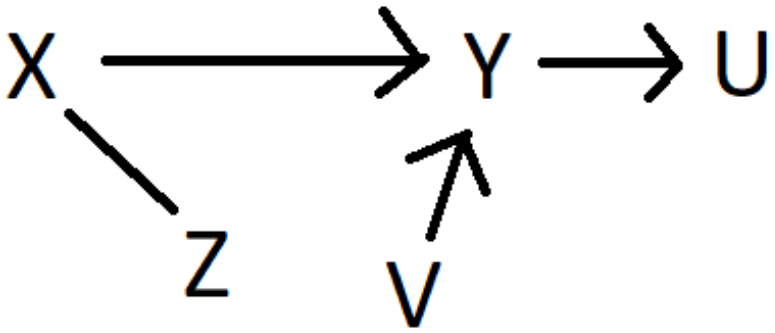

Figure 25: Toy example of a Bayesian Network of the variables X,Y,U, Z,V,

5.3 Bayesian Networks

5.3.1 Bayesian Network Terminlogy

Definition 1 Node,

A circle connected to an arc is a **node**. It represents a random variable

Definition 2 Arc

An uninterrupted line connecting two nodes is an **arc** and represented the conditional dependency between two random variables.

Definition 3 Path A **path** is a directional arc connecting two nodes.

Definition 4 Parent

A node x with a path to another node y is considered the **parent** of node y.

Definition 5 V-structure

A **v-structure** is obtained if a node has un-married parents.

Definition 6 Ancestor

If there exists a path from A to B, then A is an **ancestor** of B.

Definition 7 Moralisation In a Bayesian network, the **moralised** graph is the graph where every unmarried parents of the same are connected.

Definition 8 Skeleton

A Skeleton is a DAG where all directional arcs are replaced with non directional arcs.

Theorem 1 d-separation theorem

in the skeleton of the moralised ancestral graph, a variable is conditionally independent of other variables that are non parents if graphical separation is achieved.

Definition 9 Directed Acrylic Graph where arcs represent conditional dependencies between variables which are represented by distinct nodes is called a **Bayesian Network**. This model can represent discrete and continuous data.

In the toy example in Figure 25, the nodes are the letters X,Y,U,Z,U. The arcs connect the variables. X is a parent of Y and is an ancestor of U. We find that using d-separation theorem, U is conditionally independent of X given Y.

A Bayesian Network can be used for descriptive and predictive statistics. The Node-Arc structure provides information on independence/dependence relationships between each variable and its relation to the response. Bayesian models also have an advantage of automatically updating when new evidence is included. A distinguishing feature of the technique is its capacity for identifying possible relationships not only between covariables and response but also between covariables themselves.

Data with a lot of predictors, many being categorical, can become complicated and lack interpretability, The main benefit of the Bayesian network is its ability to create simple interpretations by implementing an arc strength threshold. The algorithm will use d-separation theorem which allows us to conclude variables are irrelevant to other variables given another variable. This allows us to obtain an easily interpretable model which will eliminate issues of co-linearity through irrelevance statements and allows me to assign conditional probabilities of events happening. This should allow us to obtain simple conclusions which will be relevant for a business such as Saint-Gobain to implement. I am particualrly interested in **IncidentType, Incident Location, RIDDOR Reportable and Category** particularly, as they have been flagged to be important predictors in the Random Forest.

In the Bayesian Networks, I want to get a visualisation of how the variables interact by using data driven methods.

I have explored two methods:

1. **Hill Climbing**, score based algorithm to create my Bayesian Network using the R package bnlearn.

2. Using **Tree Augmented Naive Bayes** (TAN) to create a Bayesian Network. The software Netica gives me an environment to implement this.

5.3.2 Hill climb

The Hill Climb algorithm learns a Bayesian network from the Data provided. It determines the best network by using a the **hill climb algorithm**. This scored based algorithm, which is a discrete version of the gradient descent algorithm which that explores the space of the Directed Acrylic Graphs by single arc addition, removal and reversals; with random restarts to avoid local optima.The algorithm starts with an initial network and determines a graph that improves the network by including, eliminating or inverting an arc in the graph. The process is repeated until there is no neighbouring arc that improves the BIC score of the current network. [27]

$$BIC = \ln(p)k - 2\ln(\hat{L})))$$

Where $\hat{L}$ is the maximised value of the likelihood function of the model
n is the sample size
p is the number of parameters. [10]

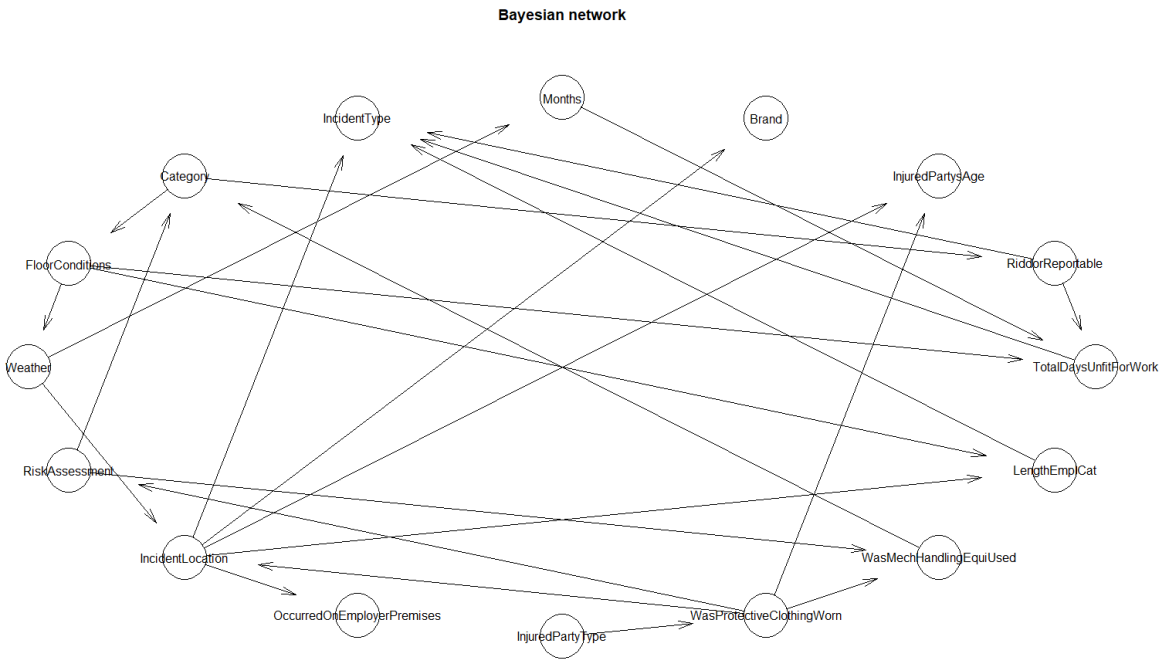

Figure 26: Bayesian network created by hill climb algorithm. This is the Bayesian network with all the predictors. It is hard to draw conclusions from this because there are many relationships. Any arc in the Bayesian network which points towards a variable not of interest (whom has no arcs of interest) will be considered for deletion.

With this network, we want to achieve d-separation theorem meaning elimination of cycles and v-structures. The Bayesian network produced by this hill climbing function has contains many weak arcs, leading to poor interpretability and cycles. Having too many variables can limit the interpretability so we want to find irrelevant variables.

One criteria to simplify would be to remove variables irrelevant of two away from the variable incident type and do not lead to it. This would consist of removing WasMechHandlingEquiUsed, LengthEmplCat, Weather, Seasons and InjuredPartyType

From the Hill Climb model of the dataset, we can remove weather and season as floor conditions tells us enough because weather and season are independent of the Incident type given the Floor Condiitons.

We can reduce the number by implementing a threshold on the arc strength.

We measure the strength of the probabilistic relationships expressed by the arcs of a Bayesian network, and use model averaging to build a network containing only the significant arcs

I can use a boot strength function to eliminate weaker arcs to order to achieve d-separation theorem. This in turn should eliminate cycles. Boot.strength estimates the strength of each arc as its empirical frequency over a set of networks learned from bootstrap samples. It calculates the strength of the arc (how often we observe x to y) and the strength of direction (how often we observe x to y when we observe an arc at all between x to y.) Boot.strength will also compute the threshold that will be used to decide whether an arc is strong enough to be included in the network network. The default value is the threshold attribute of the strength argument.

The boot strength of each arc is estimated as its empirical frequency over a set of networks learned from bootstrap samples. It computes the probability of each arc (modulo its direction) and the probabilities of each arc's directions conditional on the arc being present in the graph (in either direction).

If arc strengths have been computed using bootstrap, any strength coefficient (which is the relative frequency of the arc in the networks learned from the bootstrap replicates) greater or equal than the threshold is considered significant and is included in the averaged network. The threshold is determined by approximating the ideal asymptotic empirical CDF $F_{\hat{p}(.)}$ with its finite sample estimate $F_{\hat{p}(.)}$ CDF. This involves a choice of norm, with L_1 norm being the default.

Note that although arcs are individually identified as significant or non significant, they are not identified independently of each other as $\hat{t}$ is a function of the whole set of configurations

As such the identification of significant arcs can be thought of either as a L_1 approximation of the form

$$\hat{t} = \arg\min_{t \in [0,1]} L_1(t; \hat{p}(.))$$

t is a measure of the fraction of non-significant arc, where E_0 is the set of significant arcs:

$$e_i \in E_0 \iff \hat{p}_(i) > F^{-1}_{\hat{p}(.)}(t)$$

$$F^{-1}_{\hat{p}(.)}(t) = \inf_{x \in \mathbb{R}} \{F_{\hat{p}(.)}(x) \geq t\}$$

[20]

Which states an arc is significant if the probability of this arc being in the network is greater than the quantile function

The threshold argument is used to determine which arcs are supported strongly enough by the data to be deemed significant. Any arc strength which has been computed by the network with a strength coefficient (which is the change of the network score caused by the removal of the

arc) lower than the threshold is considered significant. In this case the default value of threshold is 0.

The idea is as follows:

1. Resample the data using bootstrap;
2. Learn a separate network from each bootstrap sample;
3. Check how often each possible arc appears in the networks;
4. Construct a consensus network with the arcs that appear more often (More Significant). [14]

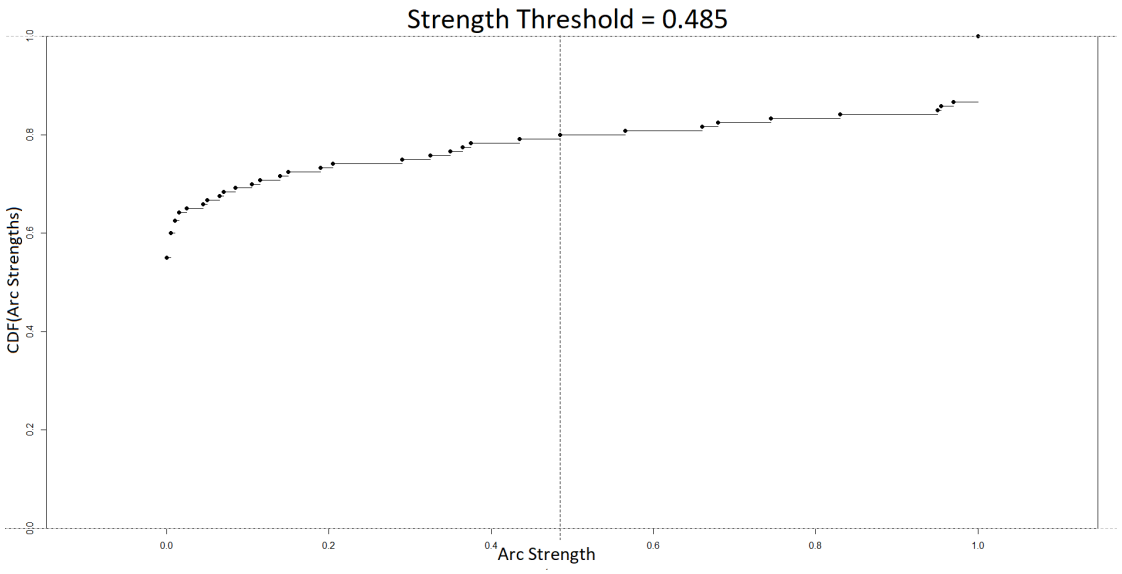

Figure 27: Strength plot with threshold. showing the cumulative cdf of the arc strength. The majority of the arcs have a low strength so only the strong arcs are plotted.

5.3.3 Bayesian Network Results

The strength plot is a cumulative density plot of the arc strengths. There are 32 unique values for the arc strength. Our algorithm picks the arcs with strength greater than 0.485, resulting in the following Bayesian network.

Removal of variables **LengthEmplCat,WasMechHandlingUsed and RiskAssessment** follows as they do not have a directed arc to a variable of interest.

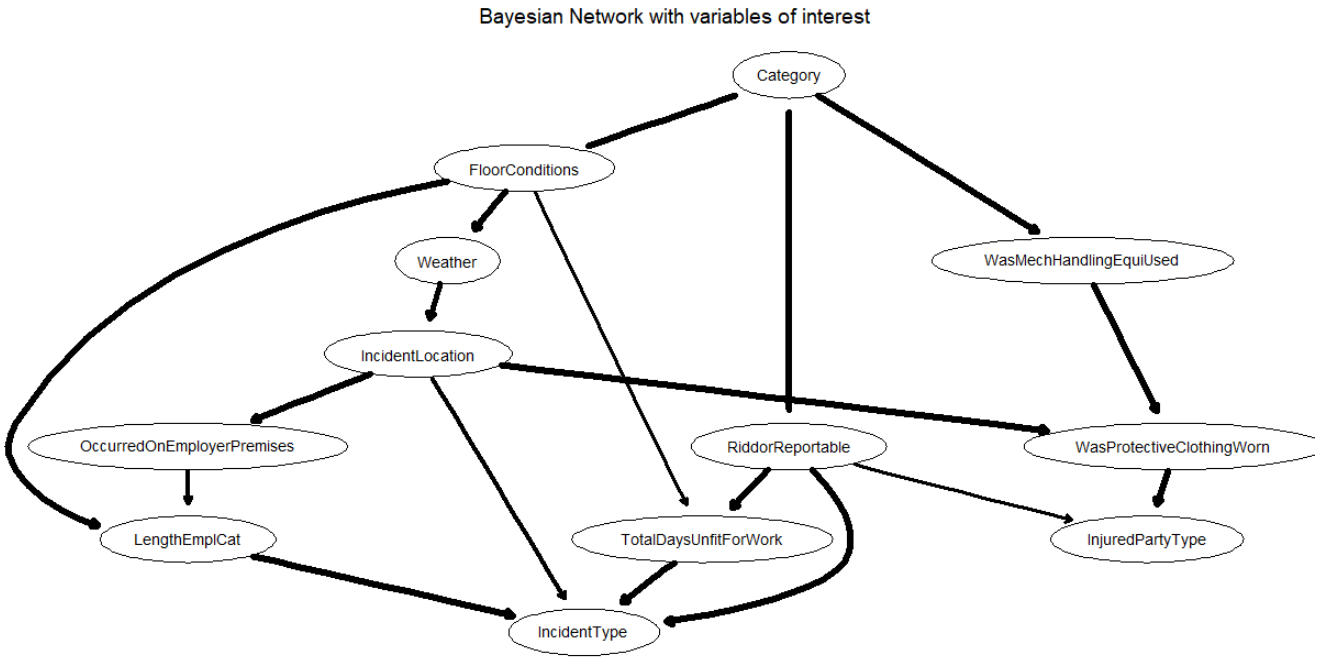

Figure 28: Bayesian network containing the variables of interest for the modelling. The thickness of the directed arcs correspond to how strong the relationship is.

Using Bayesian cross validation, we obtain the following results. It predict well on TF3s and TF1s, but struggles with TF2s. We use the Hill Climb algorithm and use the predicted weighted likelihood loss.

Variables linked directly to IncidentType are TotalDaysUnfitForWork, IncidentLocation, LengthEmplCat and Riddor Reportable.

5.3.4 Tree Augmented Naive Bayes and What If analysis

Naïve Bayes is the assumption that the presence of a particular feature in a class is unrelated to the presence of any other feature. Features independently contribute to the probability of a class.

"Tree Augmented Naive Bayes is a semi-naive Bayesian Learning method. It relaxes the naive Bayes attribute independence assumption by employing a tree structure, in which each variable only depends on the response and one other variable. A maximum weighted spanning tree that maximizes the likelihood of the training data is used to perform classification." This is a reduction in the bias and increases variance. [21]

Netica is a program created by the company Norsys to build Bayesian Networks and Influence Diagrams. It has a user-friendly interface for forming and manipuating networks. It can build networks from scratch or learn networks using pre-existing data. Netica requires a dataset consisting only of discrete variables to form a Bayesian Network created by using Tree Augmented Naive Bayes (TAN). Thus, TotalDaysunfitForWork and InjuredPartysAge are converted into categorical variables. This corresponds to changing from individual numbers to a range of numbers. (I.e 15 days unfit for work would go into "Up to a month ") Likewise for Age, where the categories are listed in a range of 10 years.

5.3.5 Sensitivity to Findings analysis

In the Netica software, there is avenue to explore how sensitive a response is to changes in the predictors. This gives us a sense of how important variables are for determining the incident type.

The **Mutual information** is a measure of cross-entropy between the target node and one of its parents (In the TAN setting, every variable is a parent of the target node). [22]

$$\sum_{q}\sum_{f} P(q,f)\frac{log(P(q,f))}{P(q)P(f}$$

q is the target node

f is a parent node.

The **Variance of node belief** $Var(q) \in [0,1]$ is the expected change squared of beliefs of the target node taken over all its states due to a finding at the parent node. [23]

$$Var(q) = \sum_{f}\sum_{q} P(q,f)[P(q|f) - P(q)]^2$$

Sensitivity of 'IncidentType' to a finding at another node:

Node	Mutual Info	Percent	Variance of Beliefs
IncidentType	0.99260	100	0.2184740
TotalDaysUnfitForWork	0.25409	25.6	0.0425096
RiddorReportable	0.12821	12.9	0.0167446
Category	0.03509	3.54	0.0051527
IncidentLocation	0.03487	3.51	0.0060029
Weather	0.02200	2.22	0.0042905
InjuredPartyType	0.01383	1.39	0.0022438
InjuredPartysAge	0.01316	1.33	0.0024548
WasMechHandlingEquiUsed	0.01240	1.25	0.0019308
Brand	0.01079	1.09	0.0014340
OccurredOnEmployerPremis	0.01078	1.09	0.0018193
LengthEmplCat	0.00966	0.973	0.0018495
FloorConditions	0.00603	0.607	0.0012302
WasProtectiveClothingWor	0.00297	0.299	0.0001833
Months	0.00084	0.0851	0.0001297

Figure 29: TotalDaysUnfitForwOrk and RiddorReportable are the most important variables. Category, Incident Location and Weather seem significant.

5.3.6 What If Analysis

The Netica software gives us the ability to condition on events occurring, then seeing how the response changes, calculating the probabilities by maximum likelihood estimation.

Using mechanical equipment in wet conditions has a higher risk of severe injuries

Increased chance of Lost Time Injuries (TF1) due to being self-employed, on work experience or contractors. On the other hand, members of public, part-time/customers have more First Aid Injurys (TF3). A few brands appear to have a higher risk of severe injuries. For the dataset, there is as 6.23% chance of a TF1 Injury. For many individual brands it can be found the chance is higher than 10%. One example is the Minister Insulation and Drylining, which has the 3rd highest overall number of incidents, has a TF1 rate of 11%. It also has a higher percentage of incidents leading to significant time off work. This may indicate the brand has poorer safety standards.

Severe categories include accidents involving moving vehicles, slips from heights have 20-30% of incidents being Lost Time Injury.

If something occurs off employer premises, we find an increase from 17.3% to 22.9% TF1s and decrease in TF3 injuries from 77% to 22.9% to 17.8%.

5.4 XGBoost

XGBoost is a gradient boosting framework with stochastic gradient boosted decision tree algorithms developed in 2014 by Tianqi Chen [18] has been said to have good performance over many Data Science competitions hosted on the Data Science website Kaggle. Tianqi notes that "Among the 29 challenge winning solutions published at Kaggle's blog during 2015, 17 solutions used XGBoost".

XGBoost uses the concept of boosting to improve the model. Boosting uses a sequence of weak learners to iteratively improve the model. They are considered "weak" because they may only be marginally better than guessing. The iterative feature identifies areas in which the previous iteration performed poorly and improves on it. The improvement is implemented via **gradient boosted trees**.

The XGBoost framework uses parallel computing, which as the name suggests, runs many operations simultaneously (scalability) to improve model exploration time. "The scalability in all scenarios means it is ten times faster than existing popular solutions". [18]

It deals with the presence of categorical data by creating a sparse matrix with one-hot encoding.

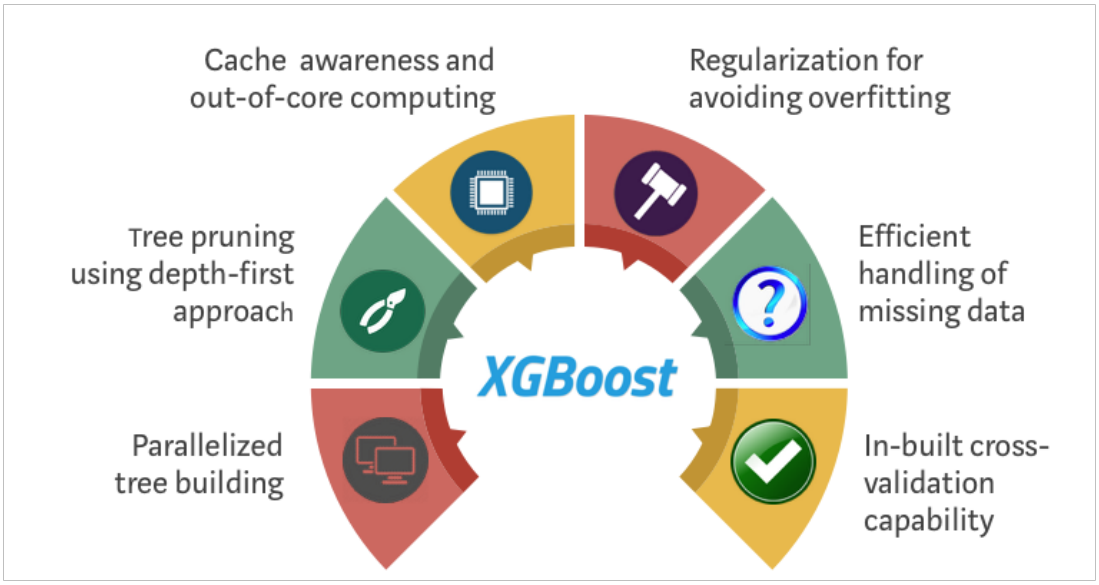

Figure 30: Explanation of XGBoost benefits [17]

5.4.1 Gradient Boosted Decision Tree

For a given dataset with n samples and m features

$$D = (xi, yi)(x_i \in \mathbb{R}^m, y_i \in \mathbb{R}),$$

a tree ensemble model uses k additive functions to predict the output.

$$y_i = \sum_{k}^{K} f_k(x_i), f_k \in F$$

F is the space of classification trees.

$$F = \{f(x) = w_{q(x)} | q : R^m \to T, w. \in \mathbb{R}^T\}$$

q represents the structure of each tree that maps an example to the corresponding leaf index.
T is the number of leaves in the tree.
Each classification tree f_k has an associated tree structure q and leaf weights $w_q(x)$
We will use the decision rules in the trees (given by q) to classify it into the leaves and calculate the final prediction by summing up the score in the corresponding leaves (given by w).

5.4.2 Objective function

To learn the set of functions used in the model, we minimise:

$$l(\phi) = \sum_{i} L(\hat{y}_i, y_i) + \sum_{k} \Omega(f_k)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda |||w||^2$$

$L(\hat{y}_i, y_i)$ is a loss function between the predicted values and observed values.
$\sum_k \Omega(f_k)$ is a regularisation term consisting of:
w, the summation of the score in the corresponding leaves
T, the number of leaves in each tree.
Common choices for loss functions may include logistic loss, squared error and the Softmax function.
Let y_i^t be the prediction of i^{th} instance at t^{th} iteration. We will need to add the tree f_t that minimises

$$L^t = \sum_{i=1}^{t} l(y_i^t, (y_i^{t-1} + f_t(x_i)) + \frac{1}{2}\lambda ||w||$$

This means we greedily add the tree, f_t that minimises the objective function.
We may find this objective function becomes too complicated, so Friedman has suggested that second-order approximations can be used to quickly optimize the objective in the general setting. [31]

5.4.3 Hyperparameters to tune

There are a lot of hyperparameters involved. XGBoost hyperparameters can be divided into three categories (as suggested by its authors):

1. **General Parameters**: Controls the booster type in the model which eventually drives overall functioning. There is no parameter I am interested in changing.

2. **Booster Parameters**: Controls the performance of the selected booster. Many of these parameters may need to be tuned by cross validation.

3. **Learning Task Parameters**: Sets and evaluates the learning process of the booster from the given data

   multi:softmax is multi-class classification technique which uses the softmax function. It returns predicted class labels.

   We will explore the following evaluation functions:

   - AUC Area under the Receiver Operating Characteristic (ROC) curve
   - Multinomial classification error rate (merror) $\frac{No.wrongcases}{No.allcases}$

I will focus on the tree booster parameters to tune [26].

5.4.4 Tree Booster Parameters

- **nrounds** controls the maximum number of iterations of the tree. For classification, it is similar to the number of trees to grow.

- **eta** controls the rate at which the model learns from the data. After every round, it shrinks the feature weights to reach the best optimum. eta and nrounds are complimentary, so a lower eta will need to be supported by an increase in nrounds to keep computation speeds reasonable. Typically, it lies between 0.01 to 0.3

- **gamma** is a regularisation parameter to reduce overfitting. This is to improve performance on test data. The optimal value of gamma depends on the dataset and other parameter values. Higher the value, higher the regularization. gamma = 5 has been recommended by Tianqi. [18] The Higher the gamma, lower the difference in train and test cross validation (Generalisation error).

  The following are regularisation parameters used to control the complexity of the trees.

- **Max Depth** controls the depth of the tree. The larger the depth, more complex the model; higher chances of overfitting. Larger datasets require deep trees to learn the rules from data.

- **Min Child Weight** blocks the potential feature interactions to prevent overfitting as it requires value sufficiently big enough. If the leaf node has a minimum sum of instance weight (calculated by using the hessian for each node) lower than the the Min Child Weight, the tree splitting stops.

- **Colsample bytree** controls the number of variables supplied to a tree. Typically, its values lie between (0.5,0.9). This feature is a reminiscent of the mtry hyperparamater in the Random Forest [19]

5.4.5 XGBoost Results

The issue I have encountered with XGBoost is that we have a lot of categorical features which are encoded via one hot encoding. this will decrease the significance of the high number of categorical variables in my dataset. Despite this, the model performs relatively well compared to Random Forests and Bayesian Networks.

To select the best model, judging it by minimisation of the overall classification error may not be ideal, as this would favour algorithms that predict TF1s the best potentially as the class are imbalanced. As such we may want to look at how well it classifies the classes individually.

The model performs well on test data in terms of accuracy without cross validation. Without performing any cross validation, it achieves 88.75% accuracy on TF1s with 92% on TF3s. It performs poorly on TF2s, achieving 17.66% in which it mainly assigns incorrect classifcations to TF3s.

In this case, we find the prediction ranges from 98.86% to 92%. This range is because I have found TotalDaysUnfitForwork and RIDDOR reportable to be the most dominant predictors.

There are 22 incorrectly classified incidents.

1773 correct TF3s, 134 correct TF1s, and 24 incidents incorrectly predicted as TF3s

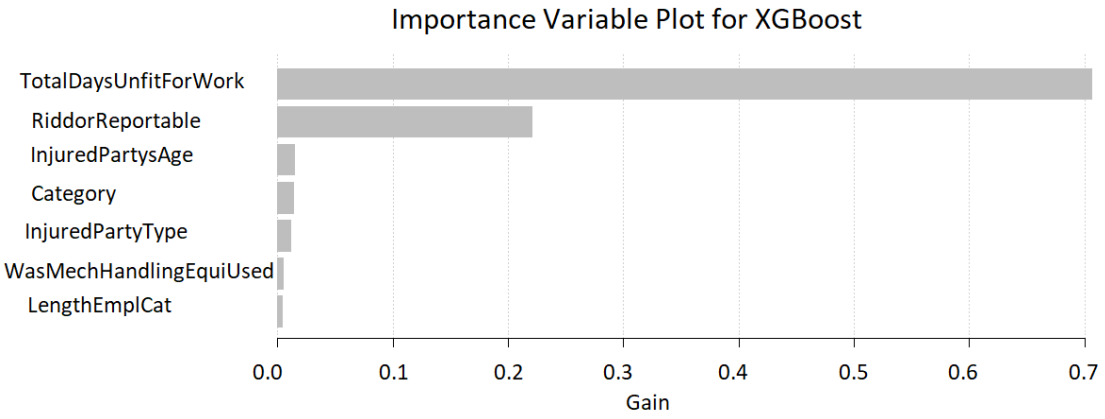

Figure 31: The evaluation of variable importance is done by the Gain. "The Gain implies the relative contribution of the corresponding feature to the model calculated by taking each feature's contribution for each tree in the model. A higher value of this metric when compared to another feature implies it is more important for generating a prediction." [4]

We have an accuracy of 95.85% and 80 incorrect classifcations, 62 misclassified as TF3s 18 misclassified as TF1s with 95 correct TF1s, 1754 correct TF3s.

I find that in the absence of RIDDOR Reportable and TotaDaysunfitForWork, it struggles to find any prediction in the TF3s, assigning all to TF1s. So in this algorithm it seems to depend on these two key variables.

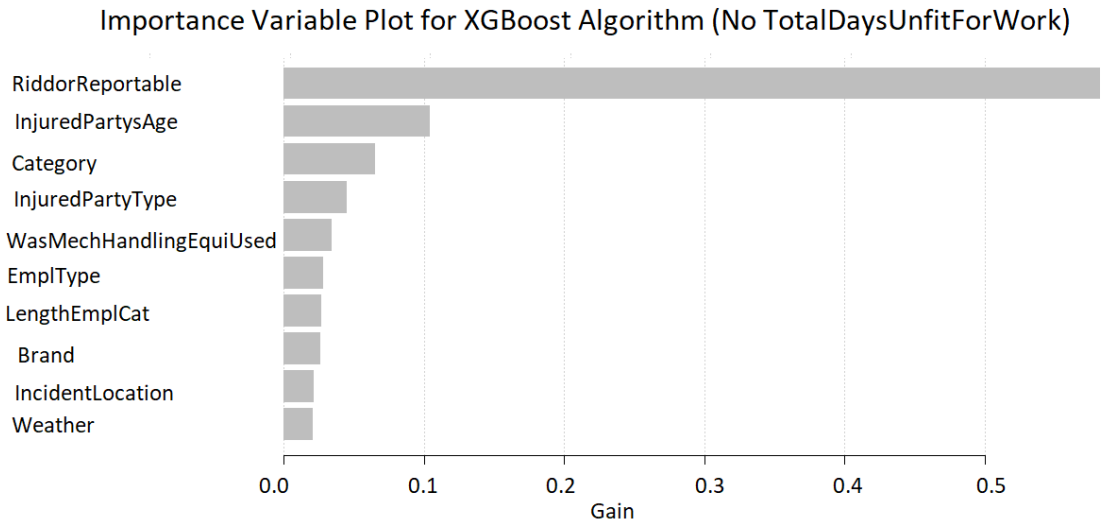

Figure 32: Variable Importance plot without TotalDaysUnfitForWork. We find the most important variables to be RIDDOR Reportable, Age and Category.

There is an issue with the cross validation. When we apply cross validation with the hopes of minimising the test error, it finds the minimum test error at the point where every observation is classed as TF1 or TF3. The cross validation specifies the nrounds which has the lowest test error. A high gamma seems to reduce the generalisation error, but it will end up reducing the overall test accuracy.

One option I have to to eliminate the TF2s and try to focus on a binary classification.

In this case, we have a binary classification problem between TF1s and TF3s. The data has 7714 incidents and 7131 TF3s and 583 TF1s. It will do binary logistic regression, outputting out a probability for a binary classification. The loss function is the log loss,

$$-\frac{1}{N}\sum_{i=1}^{N} y_i log(p(y_i)) + (1-y_i) log(1-p(y_i))$$

Where y_i is the label, with $p(y_i)$ being the probability of y_i.

5.5 Full Time Equivalent (FTE)

The Full Time Equivalent (FTE) is the amount of hours an average worker works at each Saint Gobain Branch. Upon request, Additional information was supplied from Teodora, the Data science lead at saint Gobain, to gain some information about the locations. The Turnover represents the annual revenue for each location in a year. This value will be uniquely representative for each location branch.

Using the turnover from each location, I applied a linear transformation it to find the approximate FTE hours worked per week. This my approach to numerically convert each location with an interpretable statistic, as the FTE is linked to the location they work at.

$$FTE = 3.2 + 0.000008 * Turnover$$

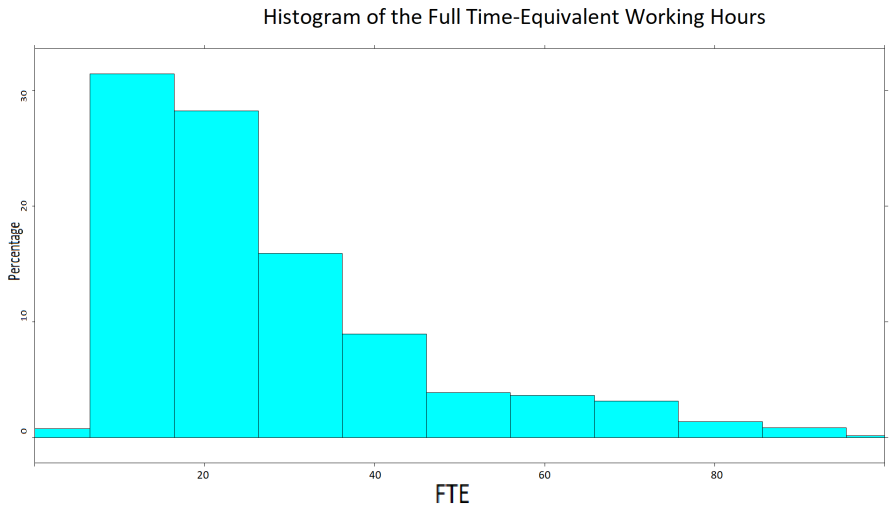

Figure 33: Histogram of the Full-Time Equivalent hours. This plot has been truncated at 100 FTE, as less than 1% of workers worked over 100 hours per week.

If people are working an amount of hours that is in the upper quantile (top 25% most hours), there are 78% more Lost Time Injuries. There are more instances of damage by 10% and more there are 670 TF2s in the upper half compared to 420. There are many near misses in all categories.

MICE was used to fill in the missing data for the TF1-3s. This dataset consists of roughly 6000 observations as a lot of the locations were missing.

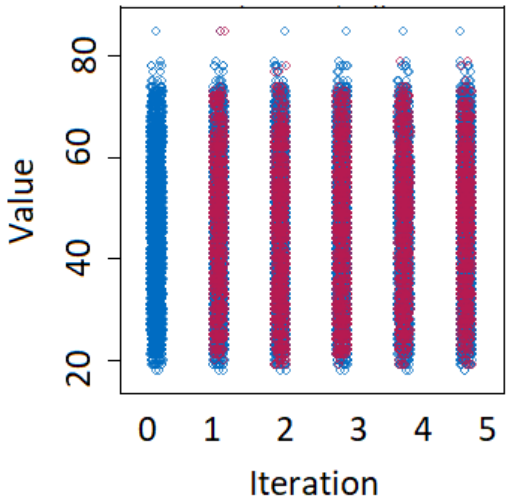

Figure 34: Stripplot showing how the original data compares to the imputted data. We see that it looks very similar.

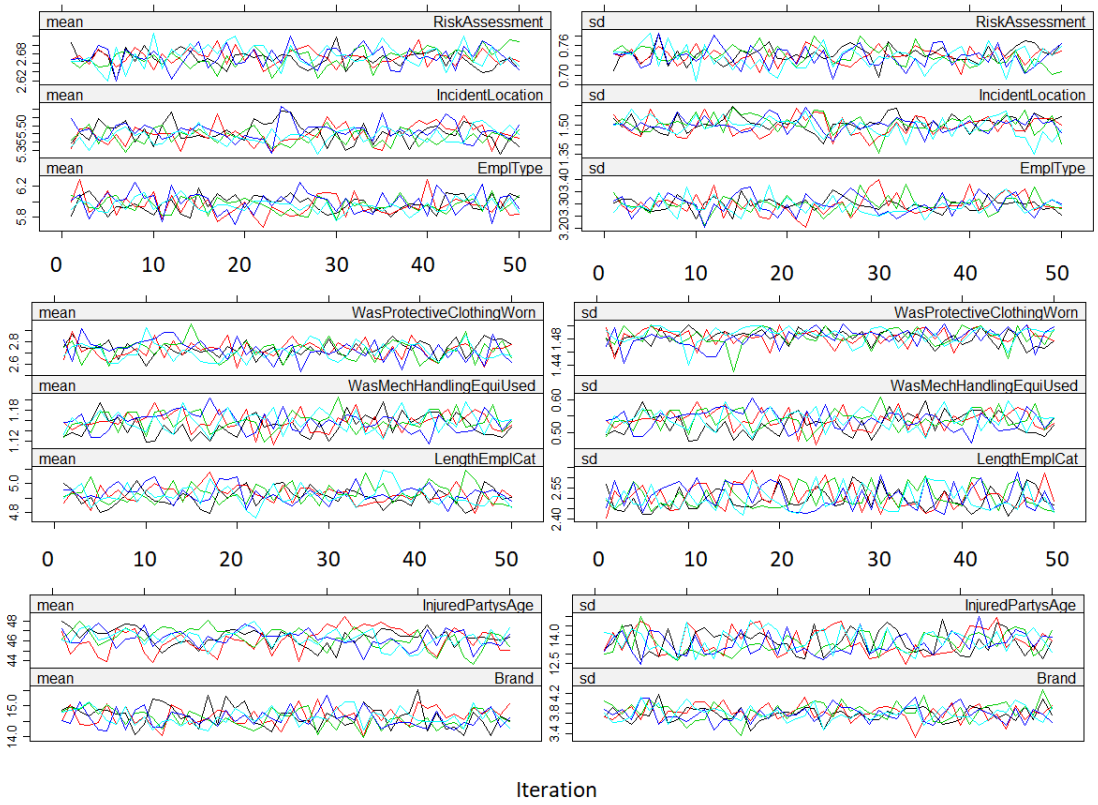

Figure 35: FTEConvergence, Convergence plots show no abnormal results of the Multiple Imputation

Using a Random Forest, we get convergence of error around 100 ntrees. With these two new variables it seems that more incorrect predictions for TF3 are going to TF2s. However TF1 accuracy remains very high. (94%) Increasing mtry reduces TF2 errror but increases Tf3 error.

Similar to Random Forest models previously explored, the optimal mtry is 5. Based on the variable importance plots, FTE is an important variables, ranking in the top 5 most important variables in the importance plots.

6 Conclusion

Management of safety and prevention of accidents is a very important aspect not just for Saint-Gobain but for all companies. In this project we have

- Looked at exploratory data analysis for the dataset
- Analysed missing values and imputed to form a complete dataset to model the TF1s-TF13s
- Removed uninformative variables
- Created models in terms of Bayesian Networks, XGBoost and Random Forest
- Important variables seem to be ones in the related to "Personal and Situational Details" as opposed to direct safety precautions the company applies.

The employment of Bayesian Networks offers an intuitive way of explaining relationships with the predictors and the incident type. An avenue explored by what if analysis is the identification of brands with higher levels of severe incidents. In terms of flexibility of interpretability the what-if analysis is very useful. Results show that TotalDaysUnfitForWork, RIDDORReportable, Category, Incident Location and Weather are the most important.

Important variables seem to be ones in the related to "Personal and Situational Details" as opposed to direct safety precautions the company applies. Working for long hours will increase the severity of incidents, however generally basic safety standards are followed, for example wearing protective equipment and having health and safety training. A suggestion would be to focus on what causes workers to begin to lose concentration. This could be in terms of working too long or lack of experience or their age.

Random Forest performs well due to its high precision on TF1s (87.79%) and TF3s (80.77%) with some access to variable explanation Bayesian network offers flexibility in What If Analysis to see how decisions affect the likelihood of accidents

XGBoost is an interesting exploration into a package with many hyper parameters. However in practice there is difficulty getting it to work and it seems to be very biased, struggling to generalise well.

Incident Location, Category are important predictors for Incident Type which is a similar conclusion to Rivas' study, it is found that variables associated with management of risk overall did not seem to be important. There is a link between working long hours and increased severity of accidents.

6.1 Recommendations

Efficient data collection has been identified as a place for improvement by Saint Gobain. In order to maximise the potential usefulness of the data there are many recommendations I can give.

Firstly, a lot of the data does not have an appropriate context attached them. For example, we may have 10 times more people working in one location compared to another and the different in incident type sizes may purely be reflective of this. The idea of counting how many people work on each location is a suggestion. It would only be an approximate estimate as it falters when you factor that many workers work between locations (e.g lorry drivers) and it may not be appropriate for offsite locations.

Secondly, there may not be a clear enough distinction between the Minor Injuries and First Aid Injuries. The algorithms report less than 30% accuracy for predictions on Minor Injuries. Attempts at a binary classification between TF1s and TF3s using XGBoost has proven successful with a 92% accuracy.

Thirdly, this problem represents a minority class prediction, where being able to predict and identify Lost Time Injuries is very important. All the algorithms naturally will have a high overall accuracy rate, simply with a target of predicting the majority class well. One option to tackle this is to resample the data. There is room for more research on this area through the use of SMOTEBoost

Finally, keeping up high safety standards (By having regular risk assessments and adequate training and by ensuring safety procedures are followed) are important. However accidents still occur despite these guidelines being followed. A bigger contributing factor is the conditions that workers work in. Working long hours and working in unsafe physical conditions (In terms of Incident Location, Weather and Category) are key contributors to accidents for which Saint Gobain can address.

7 Appendix

Main libraries used:

- mice
- XGBoost
- randomForest
- bnlearn

Information about the packages and the subsequent functions can be found in their CRAN Documentation.

Netica Software can be downloaded from their website. A tutorial can be found on their website.

norsys.com/netica.html

7.1 Additional Data description

1. **Incident Number** is a number associated with each incident, which is chronologically allocated and so it is strongly correlated with incident date.
2. The **Day** of the week to see whether some days are more prone. Incident count across days is relatively the same, except for Sunday where not many incidents occur.

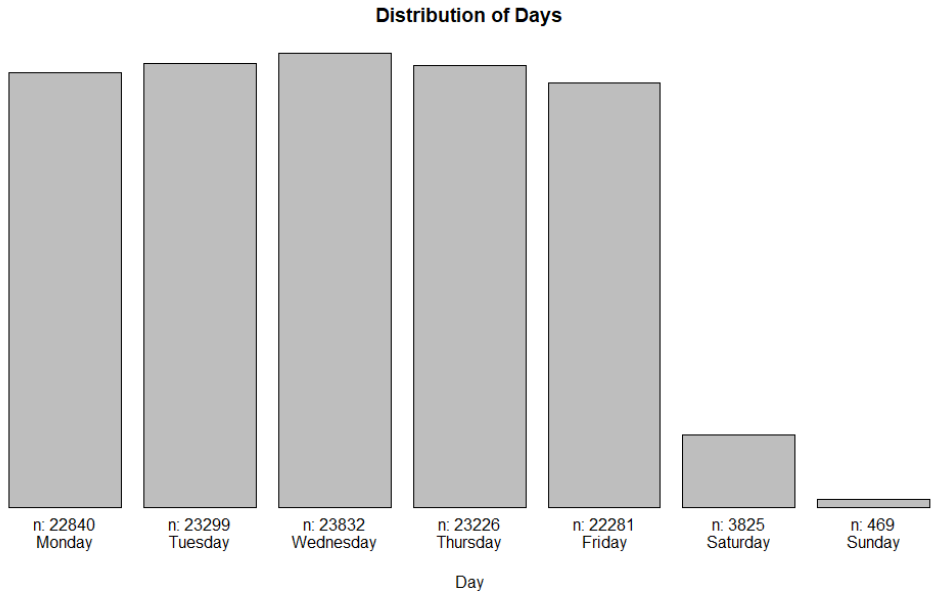

Figure 36: There is very small difference between the number of accidents in the different days, although it gets marginally bigger in the middle of the week. There are fewer accidents at the weekend, presumably as less people are working.

3. **Gender** assesses whether the gender has an impact of what type of accident occurs. Based on EDA it does not have informative features.

4. **Ethnicity** is used to see whether it has an impact

5. **Country** given in form of England, Scotland, Wales, Northern Island. It is with hope that it may be possible for working cultures in different countries to be different leading to different incident risks, but the information was not informative based on EDA.

   Spatial data in the form of Country, ru11ind and Location unfortunately do not prove as useful based on EDA plots. There is a systematic fault in ru11ind where it is hard to compare when each country has a different rural index. Furthermore it has a high amount of missing data.

Variables	DataType	Type	Description	Examples
IncidentNumber	Numeric	*Integer*	To distinguish between distinct incidents	5258..... 125133
IncidentType	Text	*Ordinal*	To distinguish between incident types	Lost Time Injury (TF1), Minor Injury (TF2), First Aid Injury (TF3), Near Miss (TF4)
IncidentDate	date	*Date*	To identify whether certain days / weeks are more prone to accidents (ie. Mondays, or days following a Bank Holiday)	20190405
IncidentTime	text	*Time*	To identify whether certain times of the day are more prone to accidents	9:00, 15:30
Category	Text	*Categorical*	To distinguish between duggerent types of accidents, could also be an outcome variable	Manual Handling Overturned vehicle Slipped, trip or fall (level) Struck against stationary object Struck by flying objects/debris Fall from a height Struck by moving vehicle (not FLT) Contact with moving machinery Exposed to fire or heat Injured by animal Exposed to harmful substance Contact with electric current Exposed to explosion Trapped by collapsing struct. Physical assault Involving Forklift Truck (FLT) Drowned or asphyxiated
DateHSWorkPlaceTraining	Text	*Categorical*	To identify the impact on incidents	<2 Months, Between 2 Months and 1 Year, Between 1 Year & 2 Years, > 2 Years
LastSafetyAudit	Text	*Categorical*	To identify the impact on incidents	<6 Months, Between 6 Months and 1 Year, Between 1 Year and 2 Years, Between 2 Years and 3 Years, >3 Years or None
FloorConditions	Text	*Categorical*	To identify the impact on incidents	Dry, Wet, Not Applicable, Normal
Weather	Text	*Categorical*	To identify the impact on incidents	Dry/Hot, Wet/Cold, Dry/Cold, Not Applicable, Icy/Cold, Wet/Hot, Wind
PropertyDamage	Text	*Binary*	To identify the impact on incidents	Y/N
RiskAssessment	Text	*Binary*	To identify the impact on incidents	Y/N
IncidentLocation	Text	*Categorical*	To idenitfy which locations are more prone to incidents	Offsite Stockyard Offices/Showroom Production/Processing Maintenance Workshop/Area Warehouse MezanineOffsite, Stockyard, Offices/Showroom
OccurredOnEmployerPremises	Text	*Binary*	To understand whether incidents are more likely to occure on-site / off-site	Y/N
Location	text	*Categorical*	To understand the geographical distribution of incidents	L9663, L0490....

Figure 37: Description of the variables, with given example and motivation

Variables	DataType	Type	Description	Examples
Employmenttype	text	*Binary*	To identify whether employment type (via training / culture etc) has an impact on incidents	Employer, Non Employer
InjuredPartyType	Text	*Categorical*	To identify whether employment type (via training / culture etc) has an impact on incidents	Contractor(<=10hrs/week), Customer, Full-Time, Member of Public, Part-Time, PermanentContractor(>10hrs/week), Self-Employed, Temporary Worker/Agency, Work Experience
EmplType	Text	*Categorical*	To link employment types to incidents, and factors likely leading to accidents	HGV/LGV Driver Machinery Operative Management Staff Field Based Employee/Sales Internal Sales Office Staff Counter Sales Staff Yard/Warehouse Operative MHE/Forklift Truck Operator Representative
WasProtectiveClothingWorn	Text	*Binary*	To identify the impact on incidents	Y/N
WasMechHandlingEquiUsed	Text	*Binary*	To identify the impact on incidents	Y/N
AuthToOperate	Text	*Binary*	To identify the impact on incidents	Y/N
Gender	Text	*Binary*	To identify the impact on incidents	Y/N
LengthEmplCat	Text	*Categorical*	To identify the impact on incidents	10-20 Years, 5-10 Years, 6-12 Months, 20+ Years, 2-3 Years, 1-2 Years, 3-6 Months, 3-5 Years, 0-3 Months
TotalDaysUnfitForWork	numeric	*Continuous*	This can serve as an other outcome variable (ie. factors ABC will likely lead to X days off	0-1000+
RiddorReportable	text	*Binary*	To use as a proxy for serious incidents	Y/N
InjuredPartysAge	numeric	*Continuous*	To identify the impact on incidents	0 (Unknown), 70+
Ethnicity	text	*Categorical*	To identify the impact on incidents	Labelled A-M
Dav	text	*Categorical*	To identify whether certain days are more	Monday. Tuesday...
Brand	Text	*Categorical*	To identify spatial relationships	Ceramic Tile Distributors PDM Ltd Jewson George Boyd Jewson Civils Graham Minster Insulation and Drylining Gibbs and Dandy Group Bassetts J.P. Corry Group Ltd International Decorative Surfaces IDS Independant Pasquill Frazer Neville Lumb Priority Plumbing International Timber Blackpool Power Tools Chadwicks Ideal Bathrooms Independant Tile Depot Normans Ideal Bathrooms Ltd Calders & Grandidge World's End Tiles
ru11ind	Text	*Ordinal/Categorical*	To identify spatial relationships. Rural index tells us whether the location is more urban or rural. England and Scotland have two distinct systems. (For example 1 or A1 would describe a metropolitan area whilst Z9 or 8 would be a hamlet.	1 2 4 5 6 8 A1 B1 C1 C2 D1 E1 E2 F1 F2 Z9

Figure 38: A continuation of this. A list of every Saint Gobain brand is given. Some Saint Gobain brands may have been recently purchased.

Rural index is described as follows:

ru11ind is a Rural Index which classifies the postcode of the location. It is scaled from 1 to 9 for Scotland and is scaled from A1 to Z9 for England and Wales. 1 and A1 indicating more dense areas. More information can be found in the appendix. Information for Northern Ireland and the Republic of Ireland were not covered by this classification.

A1 = urban major conurbation: OA falls within a built-up area with a population of 10,000 or more and is assigned to the 'major conurbation' settlement category. The wider surrounding area is less sparsely populated;

B1 = urban minor conurbation: OA falls within a built-up area with a population of 10,000 or more and is assigned to the 'minor conurbation' settlement category. The wider surrounding area is less sparsely populated;

C1 = urban city and town: OA falls within a built-up area with a population of 10,000 or more and is assigned to the 'city and town' settlement category. The wider surrounding area is less sparsely populated;

C2 = urban city and town in a sparse setting: OA falls within a built-up area with a population of 10,000 or more and is assigned to the 'city and town' settlement category. The wider surrounding area is sparsely populated;

D1 = rural town and fringe: OA is assigned to the 'town and fringe' settlement category. The wider surrounding area is less sparsely populated

D2 = rural town and fringe in a sparse setting: OA is assigned to the 'town and fringe' settlement category. The wider surrounding area is sparsely populated;

E1 = rural village: OA is assigned to the 'village' settlement category. The wider surrounding area is less sparsely populated;

E2 = rural village in a sparse setting: OA is assigned to the 'village' settlement category. The wider surrounding area is sparsely populated;

F1 = rural hamlet and isolated dwellings: OA is assigned to the 'hamlet and isolated dwelling' settlement category. The wider surrounding area is less sparsely populated;

F2 = rural hamlet and isolated dwellings in a sparse setting: OA is assigned to the 'hamlet and isolated dwelling' settlement category. The wider surrounding area is sparsely populated.

The rural-urban classification in Scotland is consistent with the Scottish Executive's core definition of rurality that defines settlements of 3,000 or less people to be rural. It also classifies areas as remote based on drive times from settlements of 10,000 or more people. This definition is unchanged from the 2001 Census: 1 = Large Urban Area: Settlement of over 125,000 people;

2 = Other Urban Area: Settlement of 10,000 to 125,000 people;

3 = Accessible Small Town: Settlement of 3,000 to 10,000 people, within 30 minutes' drive of a

settlement of 10,000 or more; 4 = Remote Small Town: Settlement of 3,000 to 10,000 people, with a drive time of 30 to 60 minutes to a settlement of 10,000 or more;

5 = Very Remote Small Town: Settlement of 3,000 to 10,000 people, with a drive time of over 60 minutes to a settlement of 10,000 or more;

6 = Accessible Rural: Settlement of less than 3,000 people, within 30 minutes' drive of a settlement of 10,000 or more;

7 = Remote Rural: Settlement of less than 3,000 people, with a drive time of 30 to 60 minutes to a settlement of 10,000 or more;

8 = Very Remote Rural: Settlement of less than 3,000 people, with a drive time of over 60 minutes to a settlement of 10,000 or more.

7.2 Code

In the attached zip we have:

- Original Data
- Dataset of mice (With a discrete version for use in Netica)
- Dataset with FTEs
- Dataset with FTEs with missing data filled in.
- Sample code
- Random Forest Results
- Examples of Netica Bayesian Networks

References

[1] Teresa Brunsdon, Topics In Data Science Lecture Notes University of Warwick, 2020, https://warwick.ac.uk/fac/sci/statistics/currentstudents/modules/st3/st343/

[2] Tixier, Antoine and Hallowell, Matthew and Rajagopalan, Balaji and Bowman, Dean, 2016, Application of machine learning to construction injury prediction, volume = 69, journal = Automation in Construction, http://civil.colorado.edu/ balajir/CVEN6833/const-data/application$_o f_m l_t o_i njury_p rediction_t ixier_e t_a l_2$016.*pdf*

[3] Predictive Models from Accident Reports, Kamalesh Panthi and Syed Munawwar Ahmed and Greenville Nc, 2015

[4] The Multiple faces of 'Feature importance' in XGBoost 2019 https://towardsdatascience.com/be-careful-when-interpreting-your-features-importance-in-xgboost-6e16132588e7

[5] Shirali, Gholam and Noroozi, Moloud and Malehi, Amal, 2018, Predicting the outcome of occupational accidents by CART and CHAID methods at a steel factory in Iran, Journal of Public Health Research doi = doi.org/10.4081/jphr.2018.1361

[6] Breiman and Cutler's Random Forests for Classification and Regression 2018-03-22 https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

[7] Saint Gobain UK Ireland, Our Vision For EHS 2025 2016, https://www.saint-gobain.co.uk/sites/saint-gobain.co.uk/files/2018-11/EHS

[8] 2018 Registration Document, Saint Gobain Pages 68-69 https://www.saint-gobain.com/sites/sgcom.master/files/ddr2018$_v a.pdf$

[9] David.J.Hand, *Principles of Data Mining.* Imperial College London, United Kingdom, 2012, volume 30 pages 621-622, www.https://www.deepdyve.com/lp/springer-journal/principles-of-data-mining-IfNvcz0X0A?key=springer

[10] T.Rivas et al. *Explaining and predicting workplace accidents using data-mining techniques.* University of Vigo, Spain, 2011,www.elsevier.com/locate/ress

[11] HSE *Reporting of Injuries, Diseases and Dangerous Occurrences Regulations.* United Kingdom, 2013, https://www.hse.gov.uk/riddor/, accessed: 17.03.2020

[12] BBC *Forklift truck accident at Jewsons in Banbury kills man.* United Kingdom, 2012, https://www.bbc.co.uk/news/uk-england-oxfordshire-16996825?fbclid=IwAR30vZuZvPJxCUgnOaJmSZvs4jzhjs1OGwl8f5fbPg0te2TfggfCFbmt1Bq/, accessed: 17.03.2020

[13] Andrea S.Foulkes *Classifcation and Regression Trees. In: Applied Statistical Genetics with R.* University of Massachusetts,USA, 2009, DOI : *doi.org/10.1007/978-0-387-89554-3*$_6$

[14] Marco Scutari, *bnlearn - an R package for Bayesian network learning and inference.* Switzerland, 2014, DOI = https://www.bnlearn.com/documentation",

[15] Melissa Azur, *Multiple imputation by chained equations: what is it and how does it work?.* Maryland, USA, 2010, DOI : //www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/",

[16] Roger Lewis, *An Introduction to Classification and Regression Tree (CART) Analysis.* California, USA, 2000, DOI : https://www.researchgate.net/publication/240719582$_{A}n_{I}ntroduction_{t}o_{C}lassification_{a}nd_{R}egression_{T}ree_{C}ART_{A}na$

[17] Megha sharma, Niharika Priyadarshini, *XGBoost using python.* 2019, DOI = //http://dataanalyticsedge.com/2019/11/23/xgboost-using-python/

[18] Tianqi Chen and Carlos Guestrin, *AXGBoost: A Scalable Tree Boosting System", 22nd SIGKDD Conference on Knowledge Discovery and Data Mining.* University of Washington, USA, 2016, DOI : https://arxiv.org/pdf/1603.02754.pdf

[19] Tianqi Chen et al, *Understand your dataset with Xgboost.* DOI : https://cran.r-project.org/web/packages/xgboost/vignettes/discoverYourData.html

[20] Marco Scutaria, Radhakrishnan Nagarajanb, *Identifying Significant Edges in Graphical Models of Molecular Networks.* University College London, United Kingdom, 2018, DOI : https://arxiv.org/pdf/1104.0896.pdff

[21] Fei ZhengGeoffrey I. Webb, *Tree Augmented Naive Bayes. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning..* Boston, USA, 2011, DOI : https://doi.org/10.1007/978-0-387-30164-8$_{8}$50

[22] Fengzhan Tian et al, *Learning Bayesian Networks Based on a Mutual Information Scoring Function and EMI Method* Beijing Jiaotong University,, China, 2007, DOI : https://doi.org/10.1007/978-3-540-72393-6$_{5}$0

[23] Norsys, *Sensitivity Equations* Beijing Jiaotong University,, China, 2007, DOI = https://www.norsys.com/WebHelp/NETICA/X$_{S}ensitivity_{E}quations.html$

[24] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. CRC Pres 84.

[25] W. N. Venables and B. D. Ripley Springer (mid 2002) Modern Applied Statistics with SFourth edition 2002

[26] Aarshay Jain, Analytics Vidhya, Complete Guide to Parameter Tuning in XGBoost with codes in Python, 2016 https://tinyurl.com/zv52wj8

[27] Dimitris Margaritis, Learning Bayesian Network Model Structure from Data, Carnegie Mellon University, Pittsburgh, USA,2003,

https://www.cs.cmu.edu/ dmarg/Papers/PhD-Thesis-Margaritis.pdf

[28] Yanjun Qi, Random Forest for Bioinformatics, Springer, Boston, MA, 2012 https://doi.org/10.1007/978-1-4419-9326-7$_{1}$1

[29] International Labour Organisation, Safety and health at work, 2020 http://www.ilo.org/global/topics/safety-and-health-at-work

[30] Julien Clavel, Gildas Merceron, Gilles Escarguel, Missing Data Estimation in Morphometrics: How Much is Too Much?, Systematic Biology, Volume 63, Issue 2, March 2014, Pages 203–218, https://doi.org/10.1093/sysbio/syt100

[31] Stef van Buuren, Multivariate Imputation by Chained Equations logistic, "Package "mice" February 21, 2020

[32] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. Annals of Statistics, 28(2):337–407, 2000. https://cran.r-project.org/web/packages/mice/mice.pdf

[33] @phdthesisphdthesis, author = Tiwari, Prayag, year = 2017, month = 06, pages = , title = Accident Analysis by using Data Mining Techniques, doi = 10.13140/RG.2.2.20091.41766/1

[34] M. P. Wand (1997) Data-Based Choice of Histogram Bin Width, The American Statistician, 51:1, 59-64

https://doi.org/10.1080/00031305.1997.10473591

[35] Salford Systems, Using Surrogates to Improve Datasets with Missing Values https://www.salford-systems.com/resources/webinars-tutorials/tips-and-tricks/using-surrogates-to-improve-datasets-with-missing-values Accessed: 04/05/2020

[36] Zhang, H., Wang, M. (2009). Search for the smallest random forest. Statistics and its interface, 2(3), 381. https://doi.org/10.4310/sii.2009.v2.n3.a11

[37] author = Sarkar, Sobhan and Patel, Atul and Madaan, Sarthak and Maiti, Jhareswar, year = 2016, month = 12, pages = , title = Prediction of Occupational Accidents Using Decision Tree Approach, doi = 10.1109/INDICON.2016.7838969

[38] Oracle Customer Success—Saint-Gobain, https://www.oracle.com/customers/saint-gobain-mktg-cld-story-1.html

[39] Arthur Charpentier 'Variable Importance Plot' and Variable Selection June 17 2015 https://www.r-bloggers.com/variable-importance-plot-and-variable-selection/

[40] Health and Safety Executive, 2019, Health and Safety statistics in the United Kingdom, 2019 https://www.hse.gov.uk/statistics/european/european-comparisons.pdf