Assignment 2 - Data Modelling & Presentation

Absenteeism At Work

Reuben Rajeev - s3717497
Jeremy Quintana - s3719476

Contact Details:	Reuben - s3717497@student.rmit.edu.au Jeremy - s3719476@student.rmit.edu.au
Affiliations:	Royal Melbourne Institute of Technology
Date:	26/05/2019

Table of Contents

Abstract

Utilizing data sets donated through UCL, a data set was chosen to complete a data science classification task. Using the absenteeism at work data set collected from a courier company, the data contained entries of absences and the reasons for these absences and possible factors that could lead to the absences. Using this data set, the task was to create a model that would be able to predict the reason for the absences using the collected attributes. The traditional data science process was utilized, hence the data set was properly prepared then explored to view any relations between attributes. Two models were then used against the data set, to enable predictions on future data and analyse patterns. Upon analysing the data, the data showed little or no relationships with each other and upon attempting to train the model to predict the reasons for absence, the model also had trouble attempting to find relations. It would be recommended that if future tasks attempting to predict the reasons for a worker's absence was conducted, more data be collected on the previous history of the absentee. Additionally, it would be recommended that further safety and assistance be placed for the workers participating in heavy load due to the large amount of absences being for physiotherapy and issues with connective tissue and the skeletal system

Introduction

The data set, absenteeism at work was collected from a courier company within Brazil between July 2007 and July 2010. Courier companies are tasked with the delivery of items to destinations within a large area. An example of a well-known courier company would be FedEx whom deliver a multitude of items to areas around the world. Workers within these courier companies have varying roles although the main workforce requires the use of large manual labour to move around packages.

Employee health and safety is an increasing issue and will continue to be an issue until research is performed to identify the underlying cause and solve it. In addition, improving the health and safety for employees increases the productivity of the courier company due to less absences occurring, and more work being done.

It is known that people working high manual labour jobs, are more likely to contract issues with their health overtime. And as a result, research will be conducted concerning the absences within the company to investigate whether the courier company holds and liability for the absences but also how they can improve safety of their employees and hence productivity and profit through the reduction of these absences.

This report will follow the data science process to analyse the data set. The process will then be discussed and evaluated upon, to outline issues and alterations performed on the data set to improve final results. These results will then be analysed and evaluated elaborating on the significant details of the findings. A model will also be created, to display that there is a relationship between the data obtained and the reasons for a worker's absence. Recommendations will then be created based on the findings that were discovered through the data exploration. The model will also shed light on the specific attributes, regarding an employee absence that a company will want to avoid improving the productivity of their workers.

The data being used will mainly focus on the reason of absence. These reasons are absences attested by the International Code of Disease (ICD) and CID which encompasses all other absences that do not fall under the category of illness, disease or reason for an inability to function optimally. Accompanying the reason of absence attribute, there is information regarding the employee details at the time of absence, and some information regarding their personal lifestyle.

Methodology

The analysis of this data follows the same guidelines as a typical data science approach. Although, extra data preparation was performed within data modelling to improve the model effectiveness as well as during the data exploration stage to improve overall readability of the created visualisations.

Using the absenteeism at work as the data set to complete a classification task. Research was conducted into the domain, and the particular attributes within the data set. This research was to ensure all data within the set was interpreted correctly upon analysis and preparation. In addition, the research would enable feature engineering, to improve the resultant model's ability to predict.

Looking into the accompanying attribute information file and the previously completed research, the data was to be prepared to ensure that no errors were contained within the data set. With the attribute information document conveying that all data was numerical, a simple type check within Python was conducted to ensure that this was the case. As all data was numerical, checks for whitespace, capital letter mismatches and typo were not conducted. Sanity checks were performed, using the bounds provided within the attribute information documentation and using logical estimates for any unknown bounds. Missing value checks were also conducted for each attribute ensuring that all observations could draw relevant relations to their corresponding existing values.

When performing sanity checks, errors were found within the month attribute which led to the discovery of the errors within the season attribute. The month attribute contained values of zero which contradicted the logical estimate used for the sanity bounds. To initially solve this the observations were grouped into seasonal attribute values, with the median of each group being imputed to the corresponding observations with months in error containing the same seasonal value. It was then identified that the seasonal values did not line up to the relevant months of the year, hence the median month value for all observations was imputed into the errors and the season attribute data was recalculated with the relevant months.

When exploring data, a number of graphs were created to analyse the significance of the attribute in the overall data set and hence their relationships with other attributes were predicted and graphed. A series of presumed most significant individual attributes contained within the data set was graphed and analysed to view their relevance and importance in the overall data set. Another set of graphs were created using presumed significant relations between attributes of the data set. Although, the main relations should contain a relation with the reason of absence, other relations were formed to classify employee attributes that may cause more absences.

For singular attributes in data exploration a series of bar charts, density plots and pie charts were used. This is mainly to display the overall frequency that the values are used within the data set. Meanwhile for graphing relations, scatterplots were used with some being hexbins to emphasise density, and boxplots some of which being ordered for those with a large magnitude of boxes.

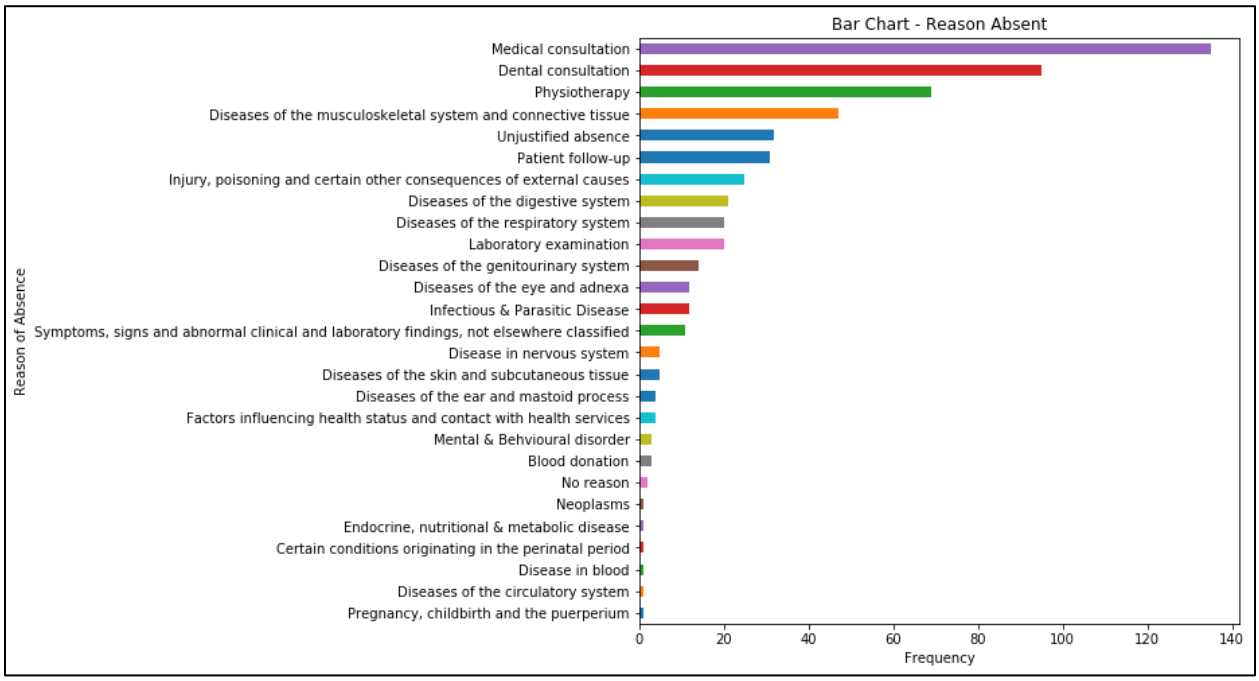
Treating this as a classification task, two models were created and trained to attempt to predict the reason an employee would be absent, but also identify a clear relation between the reason absent attribute and the remaining attributes within the data set. The two models used were K Nearest Neighbour and Decision Tree Classifier, with altered parameters to ensure that the models are trained to achieve the best accuracy based on the data set. To train and test the models, three splits into train and test data was performed on the data set, then these splits was used for both models. And the resultant accuracy for each model and split combination was then collected and compared to choose the overall best model based on accuracy.

Feature engineering was also performed to maximise the accuracy of the model. This was done through removing attributes to be placed within the model to reduce the complexity and noise that the model could ignore. Editing existing features to become more readable for the model. And creating interaction features by combing two pre-existing features so that the model can identify a clear relationship with the reason of absence. Additionally, to assist with the creation of the model, outliers were removed, although not before data exploration as these outliers could still display valuable information in graphs

Results

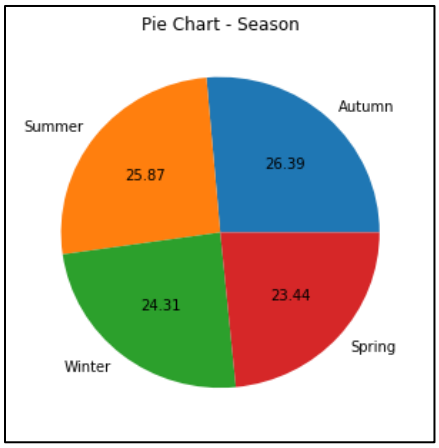
Graph of Proposed Significant Individual Attributes

Figure 1.



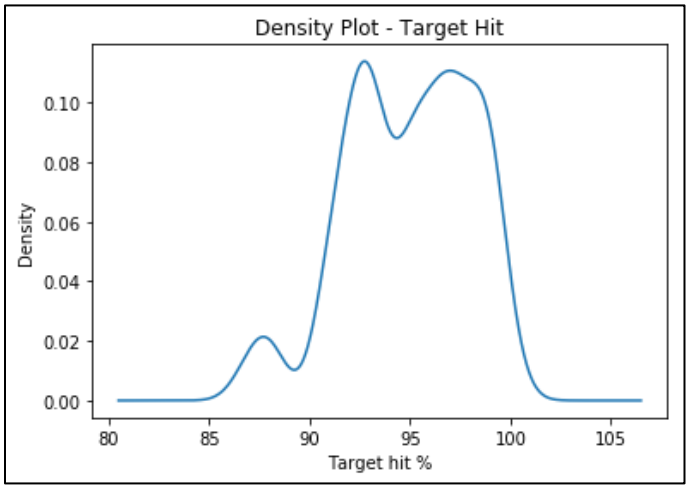
Ordered bar chart displaying the frequency of which absent reasons were most frequent to least

Figure 2.



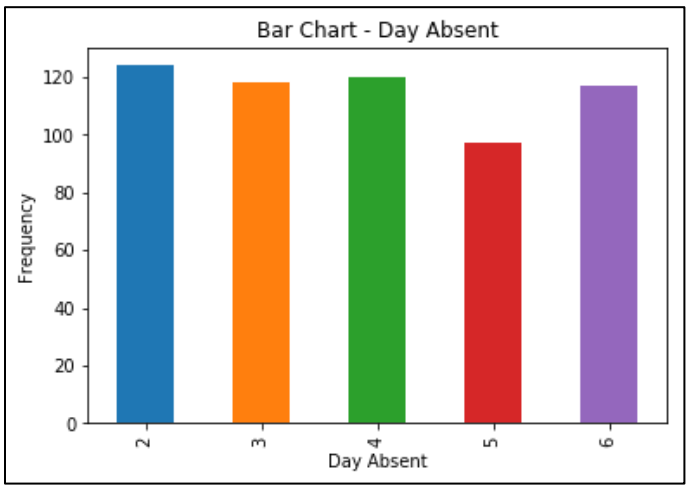
Pie chart displaying the frequency of the absences in each season

Figure 3.



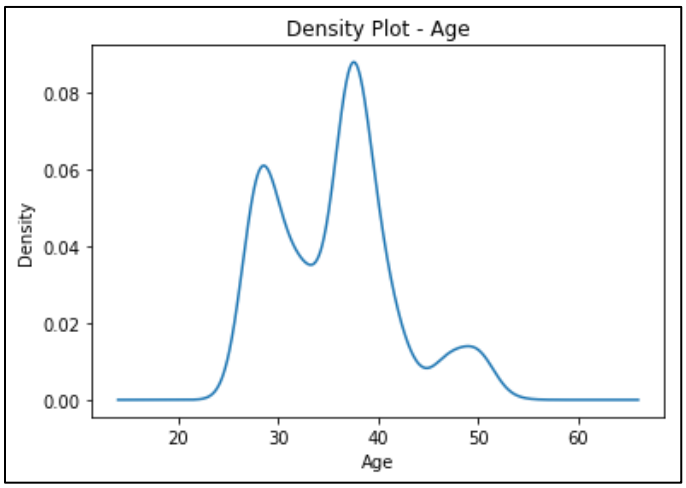
Density plot displaying the absences and the target workload hit during that absence

Figure 4.



Bar chart displaying the frequency of absences for each day

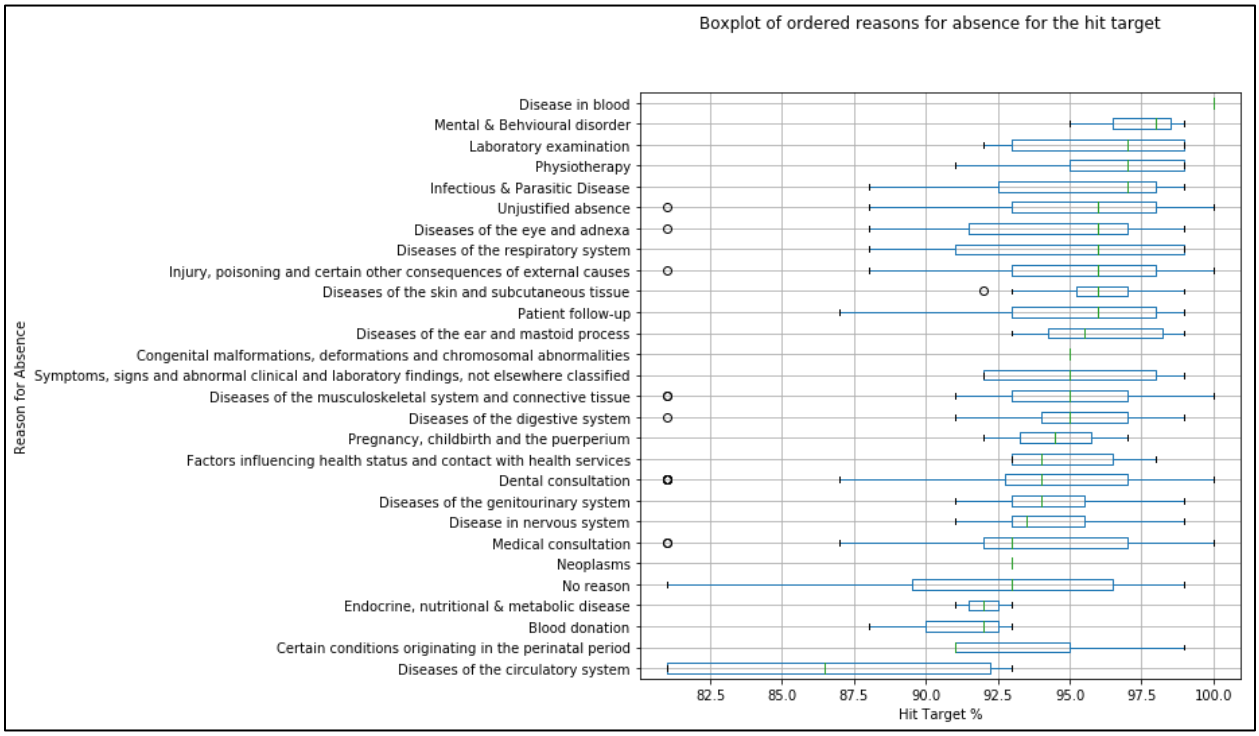
Figure 5.



Density plot displaying the most frequent age to be absent

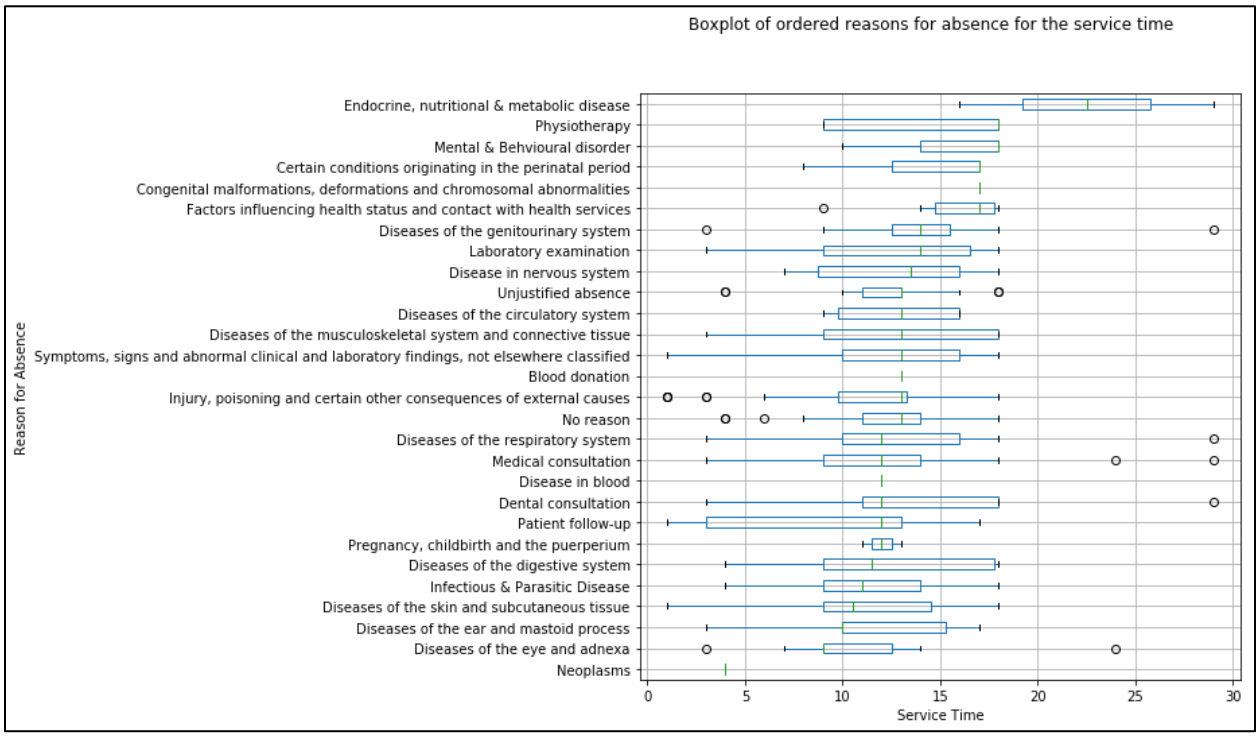
Graphs of Proposed Significant Attribute Relations

Figure 6.



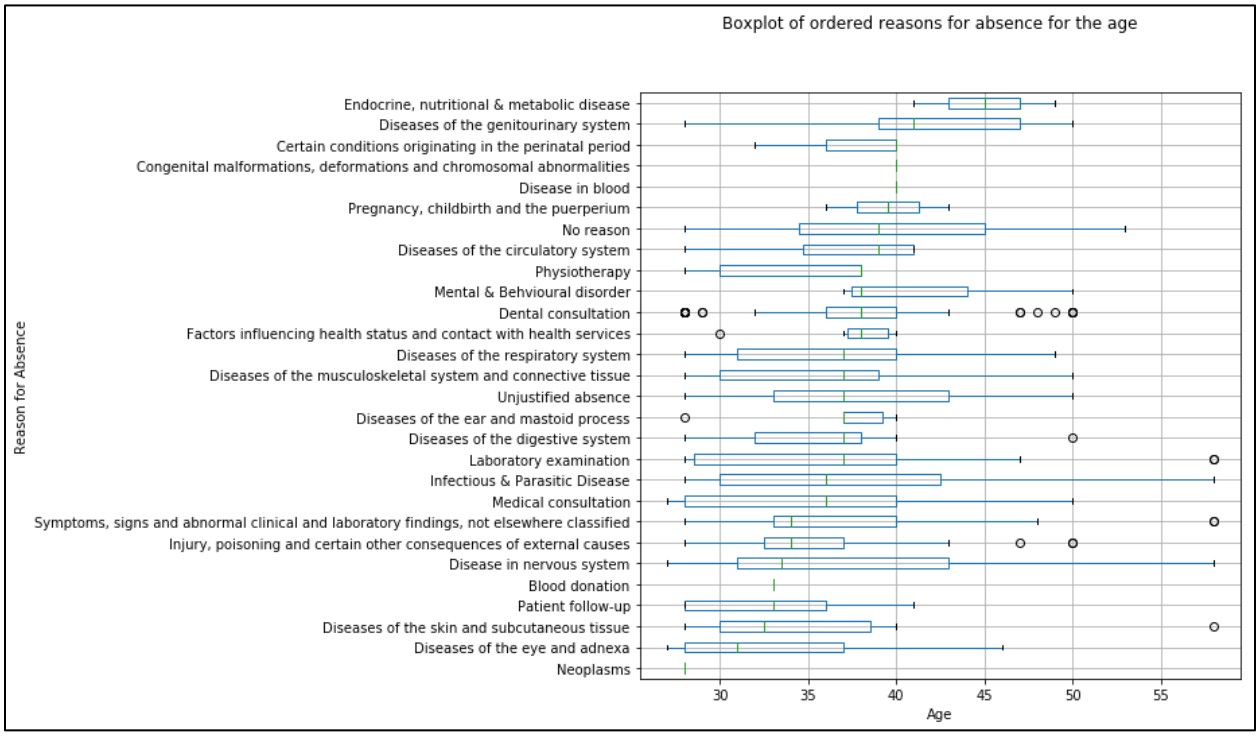
Boxplot of hit target grouped by reasons absent, showing the circulatory issues lowers the hit target workload

Figure 7.



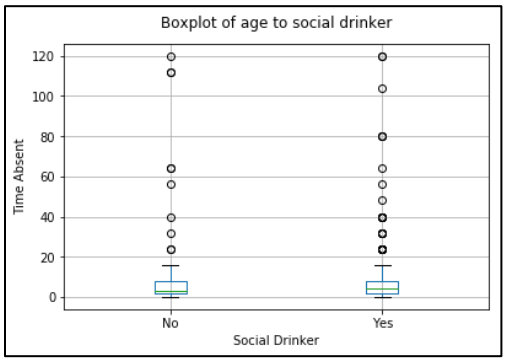
Boxplot of service time grouped by reasons absent, showing that service time is increased for those with metabolic, nutritional or endocrine diseases

Figure 8.



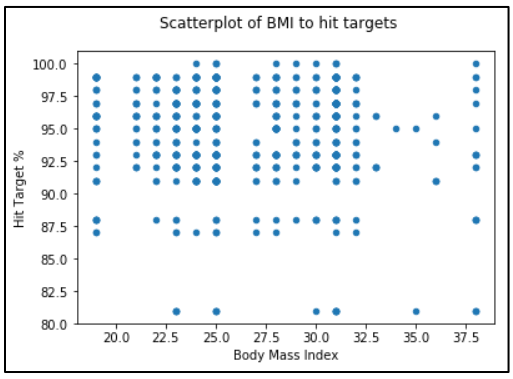
Boxplot of age grouped by reasons absent age showing that age doesn't contribute much the reasons of absence

Figure 9.



Boxplot of time absent grouped by social drinker, showing there is little relation between the two

Figure 10.



Scatterplot of BMI to hit target workload, showing there is little relation between the two

Model Results

Figure 11.

K Nearest Neighbour – 50/50 Split	Pre-Feature Engineering	Post-Feature Engineering
Precision	0.33	0.33
Recall	0.36	0.40
F1 Score	0.32	0.32
Classification Error Rate	63.54%	60.42%

Figure 12.

K Nearest Neighbour	50% train, 50% test	60% train, 40% test	80% train, 20% test
Precision	0.33	0.34	0.40
Recall	0.40	0.41	0.47
F1 Score	0.32	0.34	0.39
Classification Error Rate	60.42%	58.87%	52.59%

Figure 13.

Decision Tree Classifier	50% train, 50% test	60% train, 40% test	80% train, 20% test
Precision	0.46	0.50	0.54
Recall	0.44	0.47	0.50
F1 Score	0.43	0.48	0.51
Classification Error Rate	56.25%	53.81%	50.00%

Discussion

Through the initial exploration of the data, a single column was analysed to identify the significance of the data towards the entire data set.

Figure 1 was used to identify the most frequent reasons for absences within the company, with the expectation being that there would be a considerable amount of absences due to skeletal and tissue issues. By using this ordered bar graph, the figure was able to display the highest to lowest reasons of absence, with medical consultations being the most frequent. Although the hypothesis proved true as the highest reason for absence that was classified as an ICD was issues with the skeletal system and tissues. This pushes the idea that the courier company holds partial liability to these absences due to excess unassisted manual labour being performed.

Figure 2 was used to identify if the season of year affects the frequency of absences at a given time within the company, and it was expected that winter season would have an increase of absences due to there being more sicknesses at that time. By using a pie chart to display the frequency of the absences per season, the seasons become much easier to compare. Although as the frequency between each season is similar, percentages are displayed for a more accurate reading. As the winter frequency of absences is the highest, the hypothesis is proven true, although due to the low disparity between the seasons the hypothesis cannot be thoroughly verified.

Further analysing whether a given time affects the amount of absences, Figure 4 was used to prove the hypothesis that Mondays tend to contain higher amount absences due the high dislike for attending work after weekends. By using a bar chart, it was also able to easily identify the highest workday with absences which was Monday. This

proved the hypothesis with Mondays having higher absences which may be due to the dislike of working after weekends. Although, alike with the previously mentioned seasonal hypothesis, the disparity is too small to thoroughly verify the hypothesis.

It is to be expected that the frequency of the hit target workloads would be considerably lower than 100%, and this hypothesis is analysed through Figure 3. Using a density chart grants the capability to display continuous values of the hit target workloads and their densities. The graph whilst displaying a peak at 80% hit target also displayed other peaks, conveying that for different given absent reasons the hit target workloads can vary.

Regarding age, it is hypothesised that employees of higher age will occupy most of the observations due to those of higher age being more prone to diseases and illnesses. Within figure 5 a density plot is used to plot the densities throughout all the values of age forming three peaks at age 25, 38, 53. This does not thoroughly support the hypothesis as there are peaks at age 25 and 38, and the highest peak being 38. Although this could be due to there being a lower number of employees within the company at older ages. Additionally, the peak at age 25 could be due to absences resulting from social drinking and of the like.

A series of graphs were then created for data exploration between attributes, attempting to support hypothesis and discover significant relationships between attributes that may affect the individual attribute counterparts.

Ordered boxplots grouped by reasons of absence were created to identify if any reason directly influenced another attribute. It is hypothesised that all ICD categorised absences would hit not complete more of the target workload as compared to absent reasons categorised into CID due to ICD absences affecting the overall health of the employee. Figure 6 creates this boxplot against the percentage the employee hit to the workload displaying the mean of the percentage accomplished workload for each absence reason with absences due to diseases with circulatory systems hitting a percentage. This graph does not thoroughly support the hypothesis as not all ICD absences are lower than CID absences although, it does display that circulatory diseases affects what an employee can accomplish while other reasons do not create such a large disadvantage when completing target workloads.

The following figure 7 also uses a boxplot grouped by reason of absence and placed against the service time of a task given to an employee. It was hypothesised that the service time of a given task would be greater with all ICD related absences due to ICD absences affecting the overall efficiency of the employee hence slowing their productivity. Although, the graph displays that only absences due to metabolic, nutritional or endocrine diseases delay the overall service time of a task. This could be due to those with nutritional diseases typically having diabetes, and using previous studies, the previous study explains that diabetes lowers overall productivity within the workplace through more absences and a slower rate of productivity which is evident within the created graph.

Figure 8 uses another boxplot grouped by reason of absence and is then placed against the age of the employee absent. This will display if a particular reason of absence only occurs with specific age groups within the company, with the prediction that all physical ICD diseases would have a mean towards higher age groups of over 45 years old. The graph of figure 8 does not display such a large disparity between the means between the absent reasons with the two highest means being with metabolic disease and genitourinary disease which have been proven to occur more towards older age groups. This though does not completely support the hypothesis as not all physical ICD absent reasons were not towards an older age group.

The following boxplot of figure 9 does not use an ordered box plot as the possible grouping is binary and is used to identify if the time an employee is absent is affected if they are a social drinker. It would be hypothesised that being a social drinker would increase the time absent due to the body having a decreased ability to be able to cope with diseases. Observing the graph, there was little disparity between the mean time absent of an employee

whether they were social drinkers or not, although even by a small margin, social drinkers did have a longer time absent than non-drinkers.

Figure 10 uses a scatterplot of body mass index against the percentage of the target workload completed to identify any pattern or clusters within the two attributes. It would be expected that those of a higher body mass index would not be less likely to target workload due the various diseases that may accompany those with higher body mass indexes. The graph displayed a seemingly random combination of percentage hit targets and body mass indexes. This finding contradicted the overall hypothesis, although could support findings in figure 6 as only circulatory diseases affected percentage hit targets.

Models were created as part of the data modelling process to predict the reason of absence but also identify if there is a relation for the reason to the from features within the data set. The models used K Nearest Neighbour and Decision Tree Classifier was used to classify a given group and and hence predict the reason of absence if the attributes placed the observation within this group of classifications.

Using a 50/50 train test split, the K Nearest Neighbour model was used as an initial test to predict the overall accuracy possible to be maintained by the model. Using figure 11, the model had a classification error rate of 63.54% which was far too high for use of an accurate relation hence feature engineering was used which lowered the classification error rate to 60.42%. As shown both precision and F1 score maintained the same for post and pre feature engineering, although the recall was higher meaning the feature engineering was successful as the relevant instances that were predicted increased.

All K nearest neighbour models in all three splits displayed a reasonable but inaccurate result in predicting the reason of absence. Even after feature engineering and editing the value of the K parameter the lowest classification error rate was shown in figure 12 as 52.59%. Each recall, precision and F1 score increased as the classification error rate decreased, meaning that the higher the train to test split, the higher the overall score accuracy of the model. With such a low accuracy for predicting the reason of absence, using the presumably optimal K value to cluster the data, with feature engineered data, and altered parameters to accommodate the data being used to train the model, it may be concluded that the reason of absence is too randomised to be able to accurately predict with the current attributes of data. This possible lack of relationship may not be true though as relations between attributes were observed from the data exploration stage in figure 6 as an example, although this was a single attribute affecting a single reason of absence. If more data on new attributes were collected, this model may display a given relationship.

Similarly, the Decision Tree Classifier model was trained three times for the same train test split ratios to predict the reason of absence. The accuracy of these models is presented in figure 13 with the best overall version being with an 80/20 train test split, which follows the K nearest neighbour rule that was observed, were the larger the train data, the greater the accuracy. For this best version of this model, the classification error rate was 50% which was extremely high considering previous models. Precision was relatively high with a 0.54 value although, the recall was lower with a value of 0.50 meaning that the decision tree was not able to identify relevant instances over all instances.

Overall using the best iteration of both Decision Tree Classifier and K Nearest Neighbour models, the decision tree classifier was the best model for this data set. Comparing all results to determine the accuracy of the K Nearest Neighbour to the Decision Tree Classifier: precision is 0.40 to 0.54, recall is 0.47 to 0.50, F1 score is 0.39 to 0.51, classification error rate is 52.59% to 50%. Decision tree being fairly simple, means that the calculation of relevant instances is not far better than K nearest neighbour although all other scores are higher. This could be due to the decision tree not being affected at all by outliers, and despite removing outliers before modelling there could still

be outliers missed within the data set. As observed through data exploration majority of the data seemed random, which hinders the k nearest neighbour model despite using distance as a weight parameter. It could also be due to Decision Tree being a non-parametric method as opposed to K nearest neighbour, meaning that the wrong k value was used with the model.

Conclusion & Recommendations

This analysis of data collected from a courier company in Brazil was used to identify possible reasons for absenteeism within the workplace and how it can be reduced. For this particular company being a courier company, it was predicted that the company may be liable to some of the absences which can reflect on the health and safety issues within the company particularly due to the heavy manual labour required of the employees within the company.

Using the data set, the hypothesis that the courier company may be liable for some of the absences due to the manual labour can be supported though not thoroughly proven. This is detailed through the data exploration stages, which displayed in figure 1 that the one of the highest ICD categorised reason for absence was due to physiotherapy and issues with the skeletal system and connective tissue. Considering the high manual labour required within the company it supports the evidence that the one of the highest absence reasons is due to issues regarding the physical movement of the body.

Overall recommendation for the company would be to improve assistance and provide more effective tools to assist in manual labour for workers. With improved tools and training for workers, regarding proper technique when moving and lifting packages there would be a reduced number of skeletal system and connective tissue issues. This reduced number of issues for this particular issue results in a reduced number of appointments dedicated to physiotherapy, which would drastically reduce the number of absences due to the reduction two large reasons that employees are absent. In addition to this reduction of absences, from a business perspective a reduction of absences and disease within the company, results in more productivity and hence more profit.

Direct relations were also discovered that displays how disease directly affects productivity and hence profit of the company. Using figure 6, it displays how metabolic, endocrine and nutritional disease can affect the service time of a given task overall delaying the process meaning the package delivery is delayed. And using figure 7 displays how employees suffering from circulatory issues struggle to reach the target goal workload for a given day. Overall diseases are decreasing the efficiency and productivity of any company hence should be looked into, with employees being specialised and assisted to work around or cure their disease.

It would be recommended that each worker is analysed for diseases, and assigned to tasks where, their disease affects their overall productivity the least. Additionally, purely from business view, future employees should be tested for metabolic, nutritional and endocrine disease along with circulatory diseases, as these two diseases seem to affect the courier company in terms of overall productivity.

Regarding the models created and the model chosen to be most effective to predict the reason of absence. The accuracy of the best model, which was the decision tree classifier, was fairly low and inaccurate at a 50% classification error rate. With this accuracy, it shows that there is a relationship between the reason of absence and the other attributes. Although, this reason may be only partial meaning that a given feature from the data would not directly alter the model output reason of absence. Although it may also be that the data contained too much noise that was not filtered out enough to grant a more accurate result. Overall this data can be used to

attempt to predict the reason of absence, although is not recommended due to its accuracy, and it would be recommended that it be further refined with more data attributes for future studies.

With the inaccurate model but graphs still displaying relations, questions have risen regarding what data attributes are needed to accurately predict the reasons absence. This would be a particularly hard task to complete in the future due to the seemingly large amount of reasons an employee may be absent, with each reason needing to be recorded. Although if found, it will be able to answer accurately and improve company efficiency completely as certain attributes regarding a given employee can be tracked and prevented to ensure they optimise their time working within the company.

To conclude, the company does hold partial liability for the absences within the company and should improve their employee safety for manual labour to reduce absences and increase profit. Workers with specific diseases should also be monitored and given special assistance and placement in the workplace, to maximise their capabilities within the company. The model though does not display much relations with a 50% error rate, although it can be used it is not recommended and should be refined through the collection of more isolated attributes as it has been proven that a single reason for absence but not group can hold direct relation to a another single attribute for the data. If done, the courier company could monitor the employee behaviour to cater to their predicted diseases and hence improve the overall productivity and profit of the company. Although this model and the findings does not necessarily need to be restricted to a courier company but all companies to predict employee absences

References

- Andrea Martiniano (1), Ricardo Pinto Ferreira (2), and Renato Jose Sassi (3). (2018). *Absenteeism at work Data Set* . Available: http://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work. Last accessed 02/06/2019.
- Unknown. (2018). *Feature Engineering.* Available: https://elitedatascience.com/feature-engineering. Last accessed 02/06/2019.
- Chris Albon. (2017). *Create a Column Based on a Conditional in pandas.* Available: https://chrisalbon.com/python/data_wrangling/pandas_create_column_using_conditional/. Last accessed 02/06/2019.
- Natasha Sharma. (2018). *Ways to Detect and Remove the Outliers.*Available: https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba. Last accessed 02/06/2019.
- Imran. (2015). *How to choose the value of K in knn algorithm.*Available: https://discuss.analyticsvidhya.com/t/how-to-choose-the-value-of-k-in-knn-algorithm/2606. Last accessed 02/06/2019.
- Shovalt. (2017). *UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted samples.* Available: https://stackoverflow.com/questions/43162506/undefinedmetricwarning-f-score-is-ill-defined-and-being-set-to-0-0-in-labels-wi. Last accessed 02/06/2019.
- Prashant Gupta. (2017). *Cross-Validation in Machine Learning.*Available: https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f. Last accessed 02/06/2019.
- Center for Health Services Research. (2005). *The impact of diabetes on employment and work productivity..* Available: https://www.ncbi.nlm.nih.gov/pubmed/16249536. Last accessed 02/06/2019.