

National College of Ireland

Project Submission Sheet – 2018/2019

School of Computing

Student Name: Sankalp Ram Saoji

Student ID: X17154171

Programme: MSc. Data Analytics **Year:** Jan. 2018

Module: Analytical CRM

Lecturer: Vikas Sahni

Submission Due Date: 22/07/2018

Project Title: Absenteeism at work

Word Count: 1539

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:

Date:

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. **Please do not bind projects or place in covers unless specifically requested.**
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	



Analytical CRM Project

Topic: Absenteeism at work

By:

Sankalp Ram Saoji

X17154171

MSc. In Data analytics

School of Computing

National college of Ireland

Table of Contents

Dataset Background.....	5
Data pre-processing and assumptions.....	5
Research and investigation into suitable techniques	6
Specification of hypothesis	8
Implementation	9
Conclusion.....	13
References	13

Dataset Background

The dataset selected for this project is about absenteeism at work. The dataset was downloaded from UCI Machine Learning repository. The dataset contains several attributes which are related to absenteeism at work. The dataset refers to a company in Brazil. For the implementation of this project, we have selected some of the variables such as age, work load average per day, social smoker and drinker, number of child, distance from residence to work place, disciplinary failure, education level. The dependent variable for our dataset is Absenteeism, the absenteeism variable contains binomial categorical values less absent and more absent.

Data pre-processing and assumptions

The dependent variable was encoded in categorical and dichotomous form for successful execution of machine learning models. Assumption made for the dataset are as follows. The independent variable age was in numerical we have encoded it into categorical as people ranging from age 0 to 30 are considered as Young, people ranging from 31 to 45 are encoded as mid age people and above 45 are considered as senior employees. Social drinker and smoker were converted to no and yes as it was 0 and 1 in the dataset. Education level were encoded to 1 has high school, 2 as graduate, 3 as post graduate and 4 were labelled as doctorate according to the meta data of the dataset. Thus, the target variable absenteeism is predicted as less absent or more absent. Machine learning is used to find the interdependence between the target variable and the predictors.

Research and investigation into suitable techniques

The relationship between work, health, age, education, social activities can be demonstrated using absenteeism. Chronic diseases can lead to more number of absenteeism at work. Quality of education can also lead to absenteeism at work as on different education levels people have different patterns towards commitment of work. Increase in age can sometimes cause serious health issues which ultimately leads to absenteeism at work. Absenteeism is not only related to health care there are many other causes which lead to absenteeism examples: family and personal problems, distance from residence to work, working conditions, relative help, children concerns, disciplinary failures. (Rhodes and Steers (1990), Ose (2005), Michie and Williams (2003), Drakopoulos and Grimani (2011)). Absenteeism is worth investing because absenteeism directly impacts to company's workforce which leads to various types of losses. Losses can be in the terms of cost, time, status.

There are different patterns of absenteeism for every individual as everyone has their own problems and they leave accordingly. Someone might be injured or face some serious health issues may take more leave in recent days but was regular for the whole year. Someone might be absent in past and now is very regular at work so we can assume that they have recovered well from the reason of absence.

In a survey by Garcia and Malo (2014) founded that in European countries the average absenteeism rate for a healthy individual is 0.5 to 0.8 days per month and for disabled people it ranges from 0.9 to 2.8 per month (Mullen and Rennane, 2017).

The absenteeism rate at work for individual who belong to the age group of 50 and above averagely takes eleven days off from work per year by Xu, Jensen (2012) (Mullen and Rennane, 2017).

The literature on different health conditions also determines absenteeism. Sharma and lynch (2013) concluded that smoker have high absenteeism rate that is 8 days per year compared to 6 average days per year for non-smokers. The conclusion was made based on survey carried out on employees of a large company (Mullen and Rennane, 2017).

Schultz and Edington (2007) investigated that the higher the body weight and less physical activity can also lead to higher number of absenteeism at work (Mullen and Rennane, 2017). Healthcare is one of the most important attribute for absenteeism at work, but the question arises does the introduction of health insurance has an impact on absenteeism or not? XU and Jenser (2017) founded that health insurance did not have any impact on absenteeism at work. Absenteeism at work is directly correlated to health status (Mullen and Rennane, 2017).

Susser and Zieberth (2010) in a study suggested that women are more likely to be absent at work compared to men (Mullen and Rennane, 2017).

Researchers such as George and Jones (2002) concludes that job satisfaction has primary impact on absenteeism at work. This means that the more dissatisfied an individual is there are more chances to be absent at work. Robbins et. Al (2003) defines job satisfaction as “*The difference the rewards employee receives and the reward they believe they should receive*” (Drakopoulos and Grimani, 2011).

Lau et. Al (2003) researched that age has negative correlation with absenteeism at work. This indicates that the younger people are more absent than the senior employees. The reason for this negative correlation might be because of the higher responsibilities of senior people at work (Drakopoulos and Grimani, 2011).

Ordinary Least Squares regression and Tobit model was used for analysis of the above relationships and interdependencies on absenteeism at work with age, gender, education, health by author Drakopoulos and Grimani, (2011).

The author Cohen and Golan (2007) used general linear regression model and concluded that absenteeism in past is directly correlated to absenteeism in future. Age factor is not strongly correlated to absenteeism as younger people take more leave than older people. Marital status does not have an impact on absenteeism, it had very low correlation with absenteeism at work. Job satisfaction had high correlation with respect to absenteeism at work. Health is one of the importance factors for absenteeism at work, rather other factors contribute more than health issues in absenteeism at work (Cohen and Golan, 2007).

Specification of hypothesis

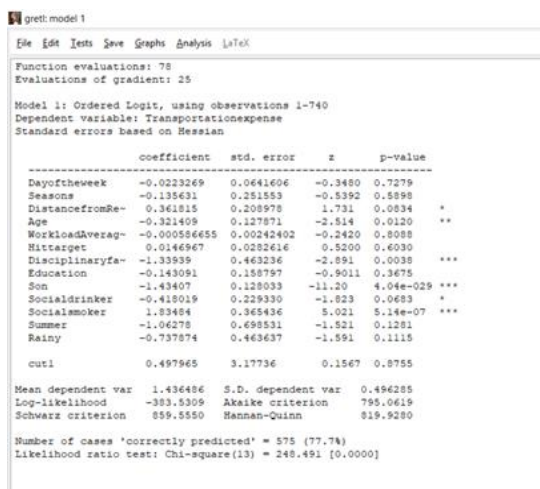
Researchers have conducted various analysis on absenteeism at work. With the use of literature review we came across some combinations which can be relatable for absenteeism at work. The question arises that, does age and workload in combination affect absenteeism? Which age group has maximum probability of absenteeism with respect to workload average per day, number children, social smoker.

H1: “Does high work load average per day and age has a direct correlation with absenteeism at work”.

H2: “With the pressure of high workload do smoking and having children affect’s middle-aged people for more absenteeism”.

Implementation

The implementation of this project is conducted using two main machine learning and statistic software known as RapidMiner and Gretl. We have already stated about the data preprocessing required for the successful execution of machine learning models. We are using logit function and logistic regression for successful implementation of this project. Logistic regression is used for predicting the output result of dependent variable that are more absent and less absent using our independent variables. The logit function in Gretl is used to check which are the most important attributes affecting the dependent variable. We have found that age, social smoker, having children and disciplinary failures are most important variables in our predictors.



gretl: model 1

File Edit Tests Save Graphs Analysis LaTeX

Function evaluations: 78
Evaluations of gradient: 25

Model 1: Ordered Logit, using observations 1-740
Dependent variable: Transportationexpense
Standard errors based on Hessian

	coefficient	std. error	z	p-value
Dayoftheweek	-0.0223269	0.0641606	-0.3480	0.7279
Seasons	-0.135431	0.251553	-0.5392	0.5898
DistancefromRe-	0.361815	0.208978	1.731	0.0834 *
Age	-0.321409	0.127871	-2.514	0.0120 **
WorkloadAverag-	-0.000586655	0.00242402	-0.2420	0.8088
Hittarget	0.0146967	0.0282616	0.5200	0.6030
Disciplinaryfa-	-1.33939	0.463236	-2.891	0.0038 ***
Education	-0.143091	0.158757	-0.9011	0.3675
Son	-1.43407	0.128033	-11.20	4.04e-029 ***
Socialdrinker	-0.418019	0.229330	-1.823	0.0693 *
Socialsmoker	1.83484	0.365436	5.021	5.14e-07 ***
Summer	-1.04278	0.498531	-1.921	0.1281
Rainy	-0.737874	0.463637	-1.591	0.1115
cut1	0.497965	3.17736	0.1567	0.8755
Mean dependent var	1.436486	S.D. dependent var	0.496285	
Log-likelihood	-383.5309	Akaike criterion	795.0619	
Schwarz criterion	859.5550	Hannan-Quinn	819.9280	

Number of cases 'correctly predicted' = 575 (77.7%)
Likelihood ratio test: Chi-square(13) = 248.491 [0.0000]

Fig.1 Most important variables.

Successful implementation of logistic regression in RapidMiner for age and work load average per days as predictors we found the prediction accuracy of 64.19%. The result obtained also contains classification error, precision, recall, F-measure, sensitivity, specificity.

Accuracy	64.19%
Classification error	35.81%
Precision	64.54%
Recall	96.81%
F-measure	77.45%
Sensitivity	96.81%
Specificity	7.41%

Table 1. Logistic regression confusion matrix parameters.

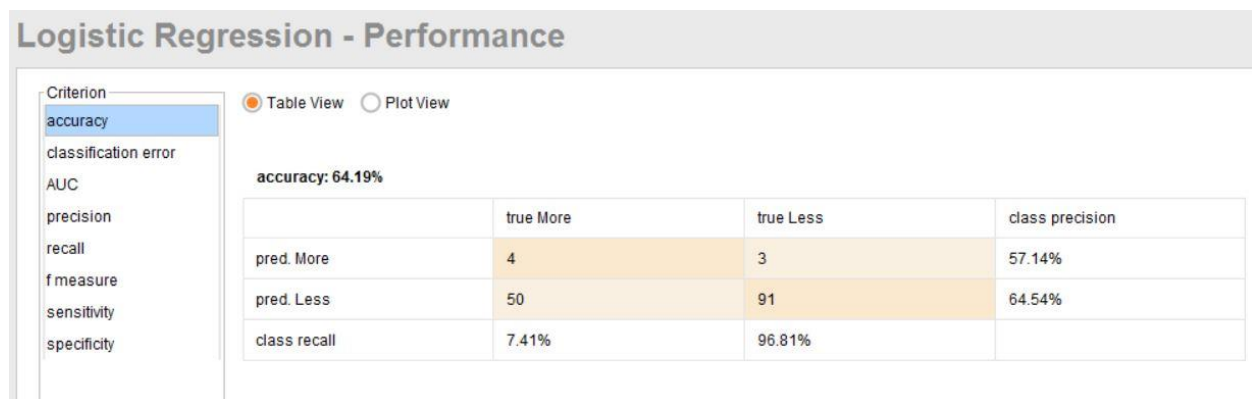


Fig.2 Accuracy obtained for age and workload using logistic regression.

The results obtained by using logistic regression for our hypothesis (H1) concluded that age is not directly correlated to absenteeism at work. Workload average per day is

independently strongly correlated to absenteeism. The project predicted that the middle-aged people with high workload have 53% probability of been in more absent class, compared to senior people have 39% probability and young generation have 42% probability to be classified in the more absent class.

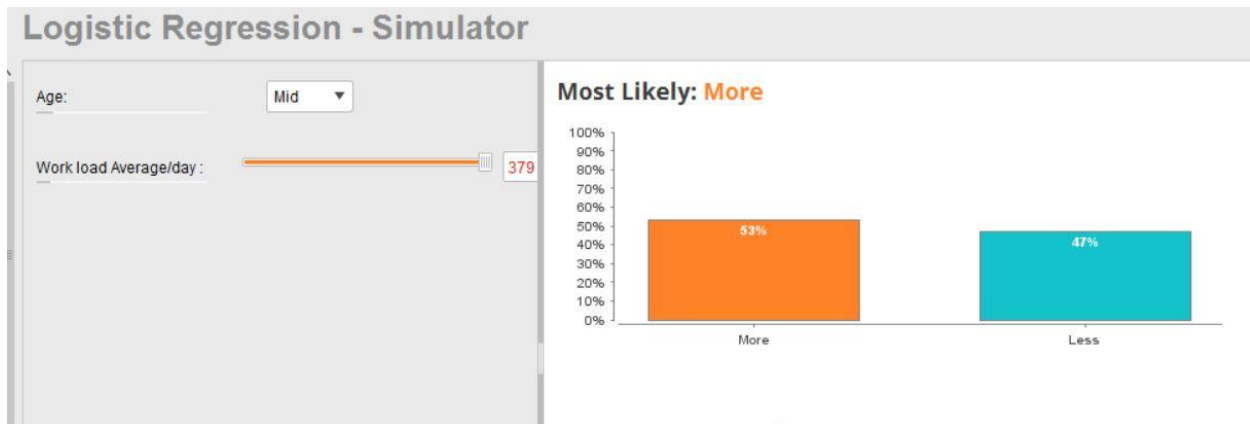


Fig.3 Probability for middle aged people with high workload.

There are many machine learning algorithms which can be used for prediction and implementation of this project some of them are naïve Bayes, Decision tree and Generalized linear model. The generalized linear models were used by other researcher's for predicting absenteeism at work. We have implemented using logistic regression, the advantages of logistic regression are as follows.

- There is no requirement of normally distributed independent variables.
- Logistic regression is robust with categorical values.
- Logistic regression works well with binary dependent variable.
- There is no assumption for homogeneity of variances.

Thus, we got the result that middle-aged people are more likely to be absent at work when workload is high. So, our second hypothesis (H2) is that social smoker and having more

number of kids has a positive or negative correlation with middle-aged people. The output obtained using the same logistic regression model is that yes, people who smoker their chances of being absent is more and number of children also has a strong positive correlation with absenteeism. Hence, both smoking and having children are directly correlated to absenteeism individually, as well as in combination it increases probability of middle aged people for being more absent.

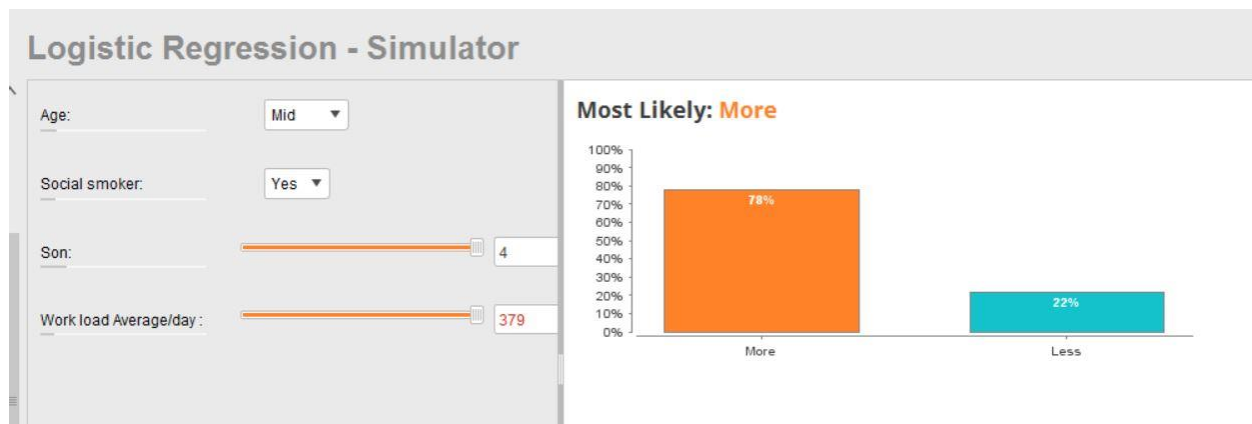


Fig.4 Probability for middle aged people who are smokers plus have more children.

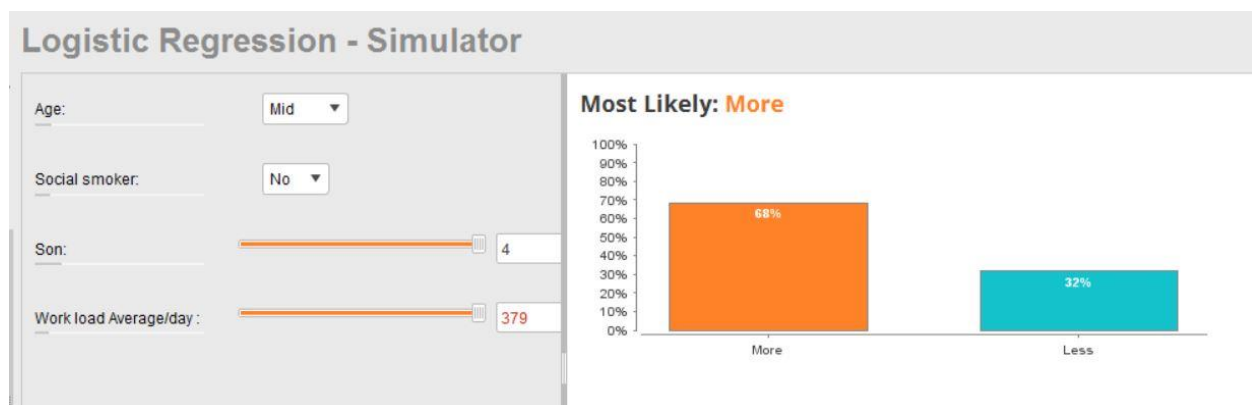


Fig.5 Probability for middle aged people who non-smokers but have more children.

Conclusion

Thus, we have successfully implemented our hypothesis (H1) using logistic regression model and we found that age is not directly correlated to absenteeism at work rather workload individually has a strong correlation with absenteeism. For hypothesis (H2) using the same logistic regression model, we conclude that smoking and having number of children in combination strongly influences absenteeism at work.

References

- Cohen, A. and Golan, R., 2007. Predicting absenteeism and turnover intentions by past absenteeism and work attitudes: An empirical examination of female employees in long term nursing care facilities. *Career Development International*, 12(5), pp.416-432.
- Drakopoulos, S.A. and Grimani, A., 2011. The relationship between absence from work and job satisfaction: Greece and UK comparisons.
- Garcia-Serrano, Carlos and Miguel Á. Malo. 2014. "How disability affects absenteeism: An empirical analysis for six European countries." *International Labour Review*, 153(3): 455-471.
- George, J.M., & Jones, G.R. (2002). *Organisational behaviour*, (3rd ed.). New Jersey: Prentice Hall.
- Lau, V.C.S., Au, W.T. & Ho, J.M.C. (2003). A qualitative and quantitative review of antecedents of counterproductive behaviour in organizations, *Journal of Business and Psychology*, 18(1):73-98.
- Michie, S. and Williams, S. (2003). Reducing work-related psychological ill health and sickness absence: a systematic literature review, *Occupational and Environment Medicine*, 60: 3–9.
- Mullen, K.J. and Rennane, S., 2017. *Worker Absenteeism and Employment Outcomes: A Literature Review* (No. odrc17-20). National Bureau of Economic Research.

- Ose, S.O. (2005). Working conditions, compensation and absenteeism, *Journal of Health Economics*, 24: 161–188.
- Rhodes, S.R., & Steers, R.M. (1990). *Managing Employee Absenteeism*, Addison: Wesley Publishing Company.
- Robbins, S., Odendaal, A. & Roodt, G. (2003). *Organizational behaviour – Global and Southern African perspectives*, South Africa: Pearson Education.
- Schultz, Alyssa B. and Dee W. Edington. 2007. “Employee Health and Presenteeism: A Systematic Review.” *Journal of Occupational Rehabilitation*, 17:547–579.
- Sherman. Bruce W. and Wendy D. Lynch. 2013. “The Relationship Between Smoking and Health Care, Workers’ Compensation, and Productivity Costs for a Large Employer.” *Journal of Occupational and Environmental Medicine*, 2013, 55(8):879-884.
- Susser, Philip and Nicolas Ziebarth. 2016. “Profiling the U.S. Sick Leave Landscape: Presenteeism among Females.” *Health Services Research*, 51(6): 2305-2317.
- Xu, Xiao and Gail A. Jensen. 2012. “Does health insurance reduce illness-related worker absenteeism?” *Applied Economics*, 44, 4591–4603.