Absenteeism prediction from timesheet data: A case study

Peter Zupančič
1A Internet d.o.o.
Naselje nuklearne elektrarne 2
Krško, Slovenia
peter.zupancic91@gmail.com

Biljana Mileva Boshkoska
Faculty of Information Studies in
Novo mesto, Ljubljanska cesta 31a,
Novo mesto, Slovenia
Jožef Stefan Institute, Jamova cesta
39, Ljubljana, Slovenia
biljana.mileva@fis.unm.si

Panče Panov
Jožef Stefan Institute and
Jožef Stefan International
Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia
pance.panov@ijs.si

ABSTRACT

Absenteeism, or employee absence from work, is a perpetual problem for all businesses, given the necessity to replace an absent worker to avoid a loss of revenue. In this paper, we focus on the task of predicting worker's absence based on historical timesheet data. The data are obtained from MojeUre, a system for tracking and recording working hours, which includes timesheet profiles of employees from different companies in Slovenia. More specifically, based on historical data for one year, we want to predict, under (which) certain conditions, if an employee will be absent from work and for how long (e.g., a week, a month). In this respect, we compare the performance of different predictive modeling methods by defining the prediction task as a binary classification task and as a regression task. Furthermore, in the case of one week ahead prediction, we test if we can improve the predictions by using additional aggregate descriptive attributes, together with the timesheet profiles.

KEYWORDS

Absenteeism at work, absence prediction, predictive modeling, timesheet data, human resource management

1 INTRODUCTION

Companies strive to have better predictive accuracy in their day to day operations, with the main goal of improving the productivity of the human resources (HR) department and hence obtaining higher profits and lower HR expenditures. They obtain information and insight from the large collections of human resource management (HRM) data that each employer owns, to support day to day operations and decision making, as well as, to comply to the national and international legislation.

The new era of HR executives is moving from settling on receptive choices exclusively taking into account reports and dashboards towards connecting business information and human asset information to foresee future results which will bring changes. Having such data enables them to detect patterns and trends, anticipate events and spot anomalies, forecast using what-if simulations and learn of changes in employee behaviour so that employee can take actions that lead to desired business outcomes. The purpose of HRM is measuring employee performance and engagement, studying workforce collaboration patterns, analyzing employee churn and turnover and modelling employee lifetime value [1].

Information society '20, October 5–9, 2020, Ljubljana, Slovenia

In this paper, we address the task of absenteeism prediction from time sheets data. More specifically, based on data that we get from MojeUre time attendance register system, we want to build a predictive model to predict if or for how many days an employee would be absent. In this case, we are considering one-week-ahead prediction from workers profiles and one year historical time sheets data. To predict if an employee will be absent in a given week, we employee the task of binary classification, which can be addressed by using a large number of binary classification methods. On the other hand, to predict the number of days an employee would be absent in a given week, we employee regression, which can be addressed by using regression methods. Furthermore, we observe and discuss how adding of aggregate attributes influences the prediction power if used together with the timesheet profiles.

2 DATA

In this section, we present the MojeUre system and then describe the structure of the raw data, as well as the process of data cleaning. Then we present the structure of the dataset, used for learning the predictive and the aggregate attributes, we constructed in order to test if they would improve the predictive power of the predictive models.

2.1 MojeUre system

The MojeUre system (https://mojeure.si) was developed to support the process of planning workers schedules, as well as for recording work attendance and absenteeism. In addition to the easy recording of the working hours of employees by a company, the system also provides access to each employee's own working hours, vacation control, sick leave, travel orders, etc. The system can be accessed using the web or by using a mobile application.

The entry of working hours is done either through a web application or a mobile application. In the case the company also wants to invests into a working time registrar, this can be done through the registrar where the employee has a personalized card for clock-in or clock-out (for example usage of break, such as a lunch break, a private break, etc.). The system allows different types of registered hours to be entered in the system in a single day.

All data used in the paper was obtained from the electronic system for recording working hours. There are currently more than 150 different companies that use the system for registering workers attendance. The basic function of the system is to record the arrivals and departures of an employee at work and to record the various types of employee absence, such as sick leave and vacation leave. In addition, the system covers other absences such as paternity leave, maternity leave, part-time leave, study leave, student leave, etc.

In this paper, we use data from the MojeUre system for the year 2019 and we have timesheet attendance data for all 52 weeks. The data instances are composed of three types of attributes: (1) attributes describing workers profiles (See Table 1), (2) attributes describing timesheets absence profiles of each worker (See Table 2), and (3) attributes that are aggregates from timesheets profiles constructed using domain knowledge (more details about the attributes is provided in Section 2.2). The timesheets attributes composing the absence profile of each worker are calculated based on the logged presence and absence logging data aggregated on the week level.. The entire dataset for the whole year consists of 232 different attributes and 2363 employees which are defined as each row.

Table 1: Workers profile attributes

Attribute name	Type	Description
EmployeeID	numeric	Unique employee identifier.
WorkHour	numeric	Data indicating how many hours per day an employee is employed by contract.
CompanyType	nominal	Company type by specific categories.
EmploymentYears	numeric	Describes how many years the person has been employed by the current company.
JobType	nominal	Describes type of job (e.g. permanent, part-time).
Region	nominal	The region in which the employee's company is located.

Table 2: Timesheet absence profile attributes

Attribute name	Type	Description
WeekWNYTotal	numeric	The number of all absences in a given week, including the sum of sick leave and (vacation) leave.
WeekWNY VacationLeave	numeric	The number of absences with type vacation leave in a given week.
WeekWNY SickLeave	nominal	The number of absences with type sick leave in a given week.
WeekWNY Absence	nominal	Value tells if employee was absent at least 1 day in whole week.

2.2 Data prepossessing and feature engineering

Feature Engineering is an art (Shekhar A, 2018) and involves the process of using domain knowledge to create features with the goal to increase the predictive power of machine learning algorithms. In this section, we describe the newly constructed attributes using domain knowledge. Furthermore, we present the process of data cleaning. Before cleaning, the original dataset contains 2087 instances of individual employees. The engineered aggregate attributes using domain knowledge from timesheets profiles are presented in Table 3.

Table 3: Attributes representing the workers profiles

Attribute name	Type	Description
VacationLeave TotalDays	numeric	Total days of vacation leave for all weeks, which are defined in the timesheets data used for the descriptive attribute space.
SickLeave TotalDays	numeric	Total days of sick leave for all weeks, which are defined in the timesheets data used for the descriptive attribute space.
ShortTerm VacationLeave3	numeric	A count of how many times an employee was at vacation leave for at least 3 days per week.
LongTerm VacationLeave5	numeric	A count of how many times an employee was on vacation leave for at last 5 days per week.
ShortTerm SickLeave3	numeric	A count of how many times an employee was on sick leave for at least 3 days.
LongTerm SickLeave5	numeric	A count of how many times an employee was on sick leave for at least 5 days.
WinterVacation LeaveAbsence	numeric	The number of vacation leave days that were used in winter.
SpringVacation LeaveAbsence	numeric	The number of vacation leave days that were used in spring.
SummerVacation LeaveAbsence	numeric	The number of vacation leave days that were used in summer.
AutumnVacation LeaveAbsence	numeric	The number of vacation leave days that were used in autumn.
WinterSickLeave Absence	numeric	The number of sick leave days that were used in winter.
SpringSick LeaveAbsence	numeric	The number of sick leave days that were used in spring.
SummerSick LeaveAbsence	numeric	The number of sick leave days that were used in summer.
AutumnSick LeaveAbsence	numeric	The number of sick leave days that were used in autumn.
WinterVacation LeaveHoliday	numeric	The number of vacation leave days that were used in winter during school holidays.
SpringVacation LeaveHoliday	numeric	The number of vacation leave days that were used in spring during school spring holidays.
SummerVacation LeaveHoliday	numeric	The number of vacation leave days that were used in summer during school summer holidays.
AutumnVacation LeaveHoliday	numeric	The number of vacation leave days that were used in autumn during school holidays.

The period we are considering in our analysis is one year, that is composed of 52 weeks. For construction of the aggregate attributes, we have defined our seasons by weeks, defined as follows: (1) the winter season is defined from week 51 in the previous year to week 12 in the New year; (2) the spring season is defined from week 13 to week 25; (3) the summer season is defined from week 26 week to week 39; and (4) the autumn season is defined from week 40 week to week 49.

In addition, we also defined the school holidays by weeks, which are defined as follows: (1) the winter holidays are defined from week 7 to 8; (2) the spring holidays are defined from week 18 to 19; (3) the summer holidays are defined from week 26 to week 35; and (4) the autumn holidays are defined from week 44 to week 45.

After we cleaned up the initial dataset, we obtained a smaller number of dataset instances. This resulted in a dataset with 961 distinct rows or more precisely different employees. The main control statement for the data cleaning was a test if an employee has less than one VacationLeaveTotalDays in the defined period. This would mean that: (1) an employee that fulfills this condition doesn't work any more in company; or (2) the company doesn't use recording system anymore; or (3) the employee is student and for students the vacation leave days are not recorded as they are usually paid per working hour only.

The most of employees in the dataset are working in company type called "Izobraževanje, prevajanje, kultura, šport" (Education, translation services, culture, sports). In addition, most of the employees are coming from the region "Osrednjeslovenska" (Central Slovenia region). The largest number of absence vacation leave or holiday leave was in week 52, which is the last week in year 2019 which is expected.

Descriptive attributes		**Target attribute**
Worker profile	Timesheet absence binary profile 1-(K-1) week	Week K Absence

(a) Without aggregate attributes

Descriptive attributes			**Target attribute**
Worker profile	Timesheet absence binary profile 1-(K-1) week	Timesheet absence aggregates 1-(K-1) week	Week K Absence

(b) With aggregate attributes

Figure 1: The structure of the data instances used for learning predictive models

3 DATA ANALYSIS SCENARIOS AND EXPERIMENTS

Research question. In general, in this paper we want to perform one-week ahead prediction of employee absence, using worker profile data, historical timesheet data aggregated on a week level, as well as aggregated attributes described in the previous section. We explore the task of predicting employee absence both as a binary classification task and as a regression task. In the experiments, we want to test if and how the aggregates attributes influence the predictive power of the built models both for the case of binary classification and regression.

Tasks. In the binary classification task, we want only to predict if an employee will be absent in a given week. For this case, we use the boolean attribute *WeekWNYAbsence* as a target attribute (WNY is the identifier of the target week). In the regression task, we want to predict the number of absence days. For this case, we use one of the following numeric attributes as targets *WeekWNYTotal* (for predicting the total number of absence days), *WeekWNYVacationLeave* (for predicting the number of vacation leave days), or *WeekWNYSickLeave* (for predicting the number of sick leave days).

Construction of the experimental datasets For the purpose of analysis, we construct two types of datasets: (1) the first type contain worker profile and timesheet absence profiles as descriptive attributes (see Figure 1a); and (2) the second type includes also timesheets absence aggregates (see Figure 1b).

In order to perform analysis, we need to properly construct the datasets used for learning predicting models. For example, if we want to predict workers absence for week 15, we use historical timesheets data from week 1-14 together with the aggregates calculated on this period as descriptive attributes.

We decided to split the year consisting of 52 weeks in four quarters (Q1: W1-W13, Q2: W14-W26, Q3:W27-W39, Q4:W40-W52), each containing 13 weeks. The absence data for the first 12 weeks were used as historical timesheet profiles, out of which the aggregate attributes were calculated. The absence of the 13^{th} week was used a target attribute. For each quarter, we constructed two different variants of datasets, one containing the aggregate attributes and the other without the aggregate attributes. This procedure was done for both tasks: binary classification and regression.

Experimental setup. For our paper, we used Weka as main software [2] to execute predictive modelling experiments. WEKA is an open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that one can develop machine learning techniques and apply them to real-world data mining problems. In the experiments, for all methods we used the default method settings from Weka mining software. The evaluation method used was 10 fold cross-validation.

Methods. Here, we used different predictive methods implemented in the WEKA software with different settings. For the regression task, we compare the performance of the following methods Linear regression (LR), M5P (both regression and model trees)[3], RandomForest (RF) [4] with M5P trees as base learners, Bagg (Bag) [5] having M5P trees as base learners, IBK (nearest neighbour classifier with different number of neighbours) [6] and SMOreg (support vector regression) [7].

For binary prediction, we compare the performance of the following methods: jRIP (decision rules) J48 (decision trees) RandomForest (RF), Bagging (Bagg) having J48 trees as base learners, RandomSubSpace (RS) [8] having J48 trees as base learners, SMO (support vector machines) [9], and IBK (nearest neighbour classifier with different number of neighbours).

Evaluation measures. To answer our research question for the case of regression, we use several measures for regression analysis, such as: Mean Absolute Error (MAE), Root mean squared error (RMSE), and Correlation coefficient (CC).

For the case of classification, we use several measures for classification analysis, such as: the percentage of correctly classified instances (classification accuracy), precision, and recall.

Table 4: Predictive performance results. The bold value denotes the highest value when we compare datasets with (A) or without (NA) added aggregate attributes. The gray cells denote the best performing method for each dataset.

(a) Performance results for the regression task - RMSE measure (less is better)

Dataset	LR	MP5	M5P-R	RF	Bagg	IBK(K=1)	IBK(K=3)	IBK(K=7)	SMOreg
Q1-A	0.789	0.692	0.775	**0.688**	**0.64**	0.804	**0.687**	0.734	0.681
Q1-NA	**0.723**	**0.674**	**0.767**	0.729	0.647	**0.798**	0.693	**0.724**	**0.659**
Q2-A	1.692	**1.369**	1.422	**1.412**	1.438	1.894	**1.476**	**1.382**	1.617
Q2-NA	**1.44**	1.382	**1.396**	1.457	**1.379**	**1.752**	1.506	1.425	**1.497**
Q3-A	0.942	**0.919**	0.976	0.999	0.935	1.409	1.074	**1.015**	**0.963**
Q3-NA	**0.911**	0.929	**0.956**	**0.968**	**0.927**	**1.223**	**1.046**	1.017	0.969
Q4-A	**0.977**	**0.947**	**0.961**	**0.923**	**0.922**	1.222	**1.029**	1.005	**0.984**
Q4-NA	0.992	0.985	0.976	1.024	0.975	**1.186**	1.066	**0.999**	1.007

(b) Performance results for the classification task - Accuracy in% (more is better)

Dataset	JRip	j48	RF	Bagg	RS	SMO	IBK(K=1)	IBK(K=3)	IBK(K=7)
Q1-A	**87.429**	**90.810**	90.357	**90.833**	89.881	**92.762**	87.452	**91.810**	90.810
Q1-NA	**87.429**	**90.810**	**90.381**	89.857	**90.357**	90.833	**89.429**	**91.810**	**90.833**
Q2-A	63.645	**68.879**	**65.751**	65.419	**66.736**	69.200	58.153	**64.347**	**68.842**
Q2-NA	**66.466**	68.177	67.118	**66.441**	66.429	**66.773**	**65.049**	62.291	67.463
Q3-A	**84.429**	**84.404**	**83.288**	83.061	84.409	**86.677**	77.182	82.616	**85.333**
Q3-NA	83.737	83.520	82.379	**83.737**	**84.864**	86.449	**81.263**	**85.101**	84.879
Q4-A	**71.130**	67.277	**72.150**	**70.460**	**70.305**	**70.452**	**69.627**	**70.644**	**70.302**
Q4-NA	70.455	**68.266**	66.774	67.441	69.791	69.466	66.093	67.610	68.960

4 RESULTS AND DISCUSSION

Regression task[1]**.** In Table 4a, we present the results for RMSE measure. It indicates how close the observed data points are to the model's predicted values, and lower values indicate better fit. From the results, we can observe that in general Bagging of M5P trees obtains the best performance. Predicting absence in week 13 from Q1 is generally better without using aggregate attributes. We have similar behaviour for predicting absence in week 26 (Q2) and week 39 (Q3). Predicting absence for the last week in the year from Q4 is generally better done using additional aggregate attributes. If we consider MAE, the best performing method is SMOreg, and for Q1, Q2 better results are obtained without the use of aggregate attributes, opposite to the Q3 and Q4. Finally, if we consider CC the best performing method is Bagging, and for Q1 and Q4 better results are obtained without using aggregate attributes, opposite to Q2 and Q3.

Classification task[2]**.** In Table 4b, we present the results for accuracy. From the results, we can observe that in general SMO obtains the best performance. For Q1, we obtain better results if we do not include aggregate attributes. For Q2, Q3 and Q4 the best results are obtained by using the additional aggregate attributes. If we consider precision the best performing methods are SMO and JRip, while for recall the best performing method is IBK using 7 nearest neighbours.

5 CONCLUSION AND FUTURE WORK

The main goal of the paper was to test if adding additional timesheet aggregate attributes can influence the predictive power in the case of one-week ahead absenteeism prediction from timesheet data. The research was performed on data from year 2019, collected by the MojeUre work attendance register system. We used various predictive modelling methods formulating the prediction task as regression (predicting the number of absent days in a week) and classification (predicting if an employee will be absent in a given week). To see the difference in performance, we performed experiments on datasets constructed on different quarters of the year. The best prediction method in the case of regression is Bagging and in general we could say that predictions are slightly better if we don't use aggregate attributes. The best method in the case of classification is SMO. Again almost same results with using or not using external aggregate attributes.

In future work, we plan to perform selective analysis of absenteeism using the same data based on different criteria, such as seasonality, closeness to holidays (before, after), critical weeks for certain professions etc. In addition, we plan to perform regional analysis and workers domain analysis which is based on company type. Moreover, more insight into absence patterns will be available after collecting several years of attendance data for each employee. Finally, we plan to compare the different granularity of prediction (day - based vs. week - based vs. half a month based vs. month based analysis).

[1]Complete results for regression are presented at the following URL https://tinyurl.com/yyp85vfr

[2]Complete results for classification are presented at the following URL https://tinyurl.com/y6o6h6d8

ACKNOWLEDGMENTS

We thank the company 1A Internet d.o.o., which provided us access to the data which were used in our research. Panče Panov is supported by the Slovenian Research Agency grant J2-9230.

REFERENCES

[1] Malisetty, S., Archana, R. V., & Kumari, K. V. (2017). *Predictive analytics in HR management.*, Indian Journal of Public Health Research & Development, 8(3), 115-120.
[2] Witten, I. H., & Frank, E. (2002). *Data mining: practical machine learning tools and techniques with Java implementations.*, Acm Sigmod Record, 31(1), 76-77.
[3] Ross J. Quinlan. *Learning with Continuous Classes. In: 5th Australian Joint Conference on Artificial Intelligence.*, Singapore, 343-348, 1992.
[4] Leo Breiman (2001). *Random Forests.*, Machine Learning. 45(1):5-32.
[5] Leo Breiman (1996). *Bagging predictors.*, Machine Learning. 24(2):123-140.
[6] D. Aha, D. Kibler (1991). *Instance-based learning algorithms.*, Machine Learning. 6:37-66.
[7] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy. *Improvements to the SMO Algorithm for SVM Regression.*, In: IEEE Transactions on Neural Networks, 1999.
[8] Tin Kam Ho (1998) *The Random Subspace Method for Constructing Decision Forests.*, IEEE Transactions on Pattern Analysis and Machine Intelligence. 20(8):832-844. URL http://citeseer.ist.psu.edu/ho98random.html.
[9] J. Platt. *Fast Training of Support Vector Machines using Sequential Minimal Optimization.*, In B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning, 1998.