Assignment 1: Predicting Absenteeism

Workforce Analytics

Group 10
Adrien Talbot 01817798
Callum Fenn Macalister 01748909
Ilias Mylonas 01770605
Konstantinos Paganopoulos 01769789
Marios Zoulias 01766825
Paolo Cristini 01800434

05/05/2020

Contents

1. Exploratory Data Analysis (EDA)

The exploratory data analysis provided a better understanding of the employees' votes, interactions and absenteeisms from 2017 to 2019. It was observed that, whilst the total number of votes increased from 2017 to 2019, the average votes of the employees using the happy force platform decreased because the proportion of employees voting negatively (vote of 1 or 2) rose. The available data sample for absenteeism is very small since the recorded periods are just for the month of June, August and October 2018; overall, the number of absences are roughly around 15 to 25 on average per month. Comparing data from HappyForce platforms between absent and non-absent employees, it is not possible to clearly identify differences in "social" behaviour as both groups heavily participated in the voting system rather than comment/feedback feature. Surprisingly, among non-absent employees, a small proportion of comments have been classified as "Criticism" while, in this category, no comment at all is present in the absent sample.

2. Predicting Absenteeism using Classification

Given the insights provided by the EDA, it was decided that the most efficient approach to predicting absenteeism, for the next T-day period, was based on the Happyforce usage of the last T-day period, where $T_{optimal}$ is equal to 15. This required an algorithm that takes as inputs the current datasets and returns one final dataset with employee Ids, their Happyforce usage features for all different 15-day periods and a final indicator whether they were absent next period or not (6 in total as we have absenteeism data for months).

Several machine learning algorithms were evaluated for performance on this data, namely: kNN, Logistic regression, Decision Trees, Random Forest, SVM and Naïve Bayes. To deal with the problem of imbalance in the number of absent and non-absent records, the random oversample method was used. Since the key criterion of the company was to predict absenteeism, the true positive rate (sensitivity) needs to be high, with the trade-off of a reduction in the false positive rate and eventually accuracy score. In terms of efficiency, the best 'Area under Curve' (ROC) was given by the kNN algorithm.

3. Employee Categorisation using Clustering

To categorise employees, a K-means clustering algorithm was implemented over the five numerical features used in the classification as well as emotional stability and the categorical variable 'Absent_next_period'.

Since K-means is sensitive to starting order, the input data was shuffled multiple times and the K-value yielding the highest average cluster stability was taken; the Elbow method and Silhouette Score were also considered. As K-means is a heuristic finding local maxima, the starting centroid position was also tuned. The optimum number of clusters was found to be four, with a 99.8% stability.

The clusters can be summarised as follows:

Cluster 0: Absentees

- All have been absent
- Smallest proportion of the company, approximately 5%
- More participative than the average employee and seemingly liked more by colleagues, as judged by their below average dislikes from others

Since the absentees appear to be positively engaged with the company, it is likely that these employees are absent for genuine physical health reasons. It is recommended that the company conducts further investigation cross-checking the 5% with industry standard absenteeism statistics.

Cluster 1: Highly Active Employees

- Never absent
- Approximately 30% of the company
- Highly participative in the surveys, all scoring well above average

The employees in this cluster represent the ideal or model employees as they frequently engage with the platform whilst never being absent. Further analysis could seek to combine KPI or performance metric data to build a model of the true ideal employee.

Cluster 2: Standard Employee

- Never absent
- Approximately 25% of the company
- Engage with surveys at an average level

Despite only representing a quarter of the company, this is the expected average employee. They are not of concern to the company regarding absenteeism.

Cluster 3: Non-Survey Participants

- Never absent
- Majority of the company (~42%)
- Rarely participate in surveys hence hard to track or assess

It is highly likely that the data, and hence clusters, are skewed by the large proportion of non-respondees. This is something the company may wish to address e.g. by offering an incentive scheme, if more accurate models are to be built in the future.

4. Actions to Mitigate Absenteeism

The company's initial goal should be to better understand the reasons behind absenteeism and, hence, mitigate circumstances which cause it to arise. Productivity of the workforce is more important than just presence, and the company should aim to keep employees motivated, even if they require several days off work.

4.1 Actions with Available Data from Happyforce

Using the classifier to identify likely candidates for absenteeism, the company should seek to discuss with these employees any problems or complaints that they may have - hopefully preventing any avoidable absenteeism.

4.2 Actions with more Data from Happyforce

The company could also ask Happyforce developers to provide additional information. Features such as the exact wording of the comments may allow specific problems to be categorised using keyword analysis. The addition of more categories for comment type would also be helpful; categories such as “OTHER” are difficult to conduct analysis with.

4.3 Actions from the HR Department

HR should seek to provide a more detailed absence report where sickness and accidents are split so that the company can distinguish between genuine and fake absences from sickness. Subject to the privacy policy of the company, details of personal characteristics for each employee such as age, seniority in the firm and performance reviews would likely improve the predictive modelling of absenteeism. Personality tests for the employees, such as the Big Five Personality Traits test, could also help the firm understand better each employee’s character.