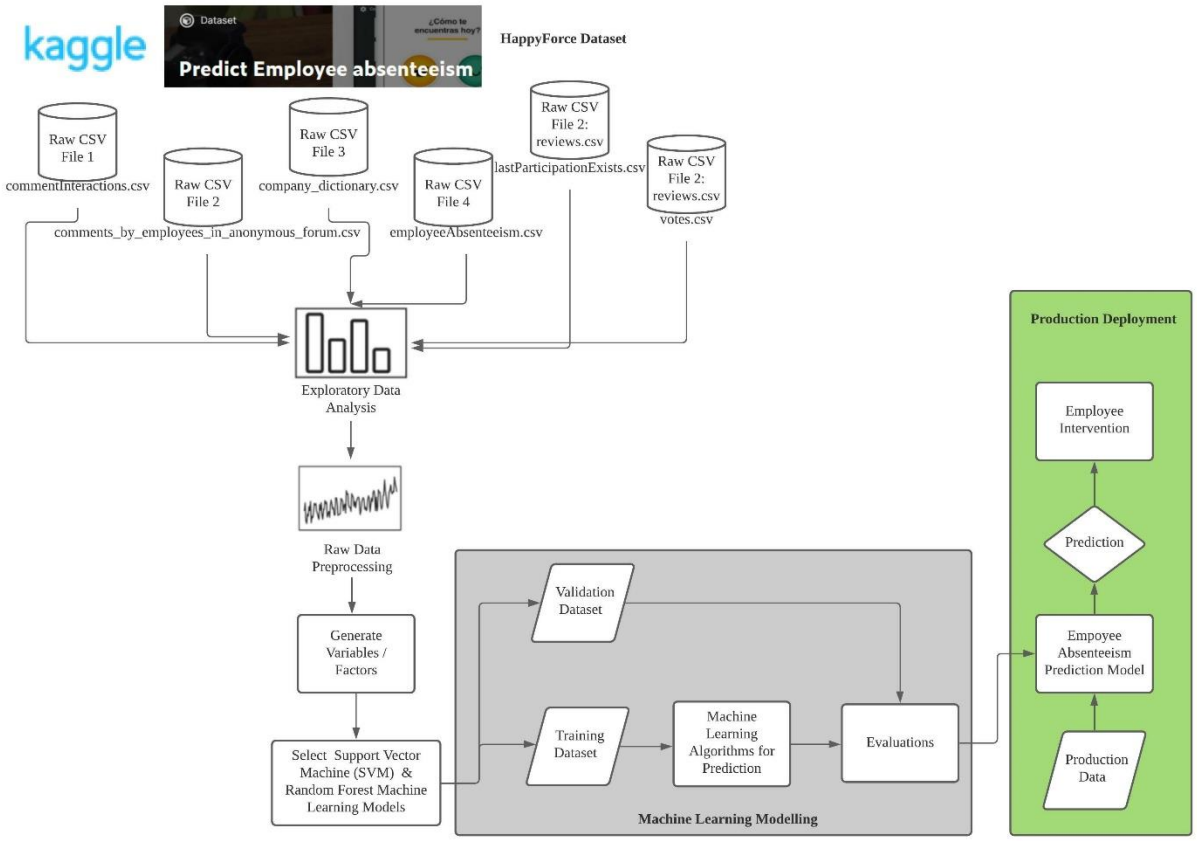
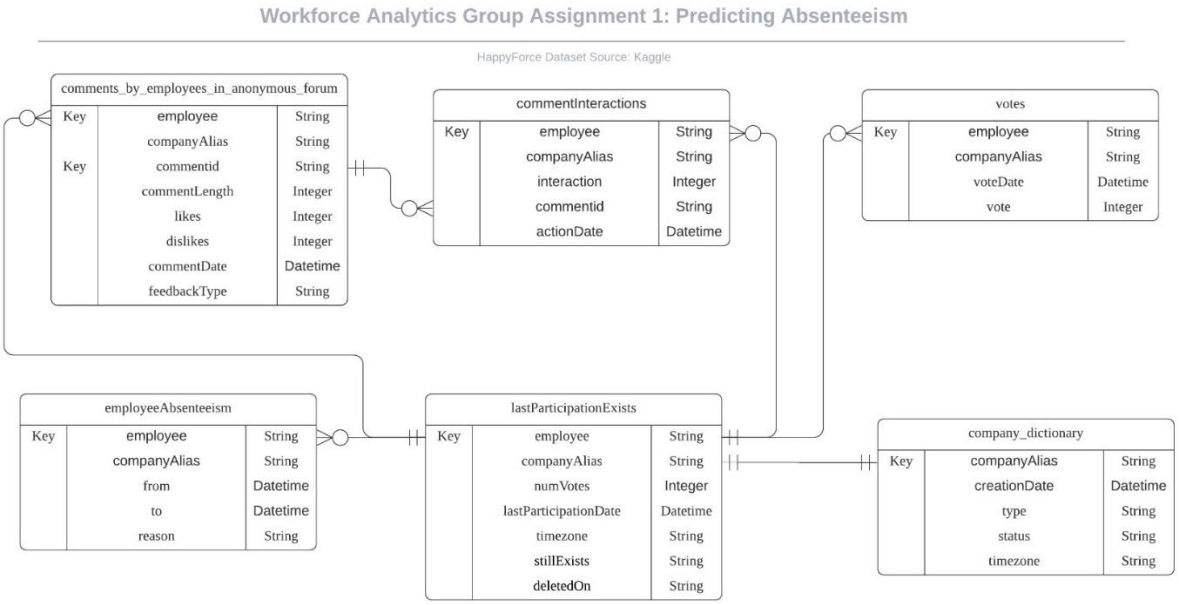
Introduction & Approach

The HappyForce dataset consists of employee related data for an organization. The overall approach adopted and summary of csv data files are shown below

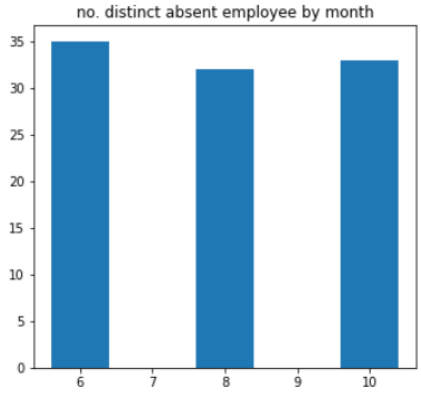


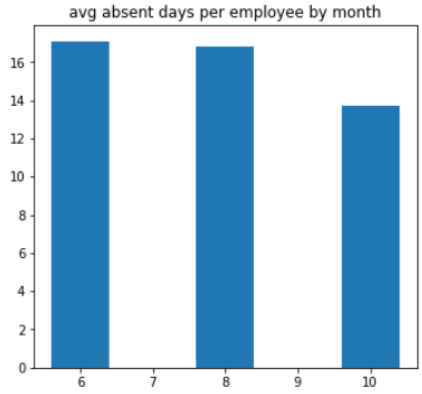


Summary of EDA

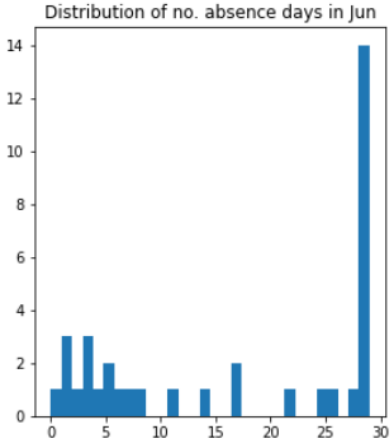
We noted seven key insights:

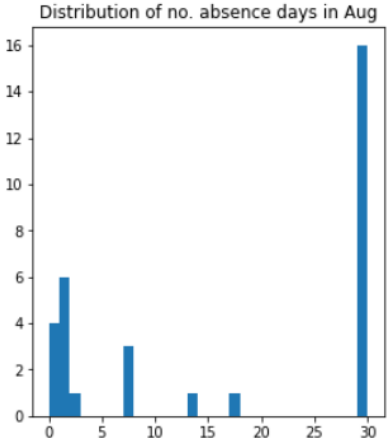
1. All absenteeism records took place in 2018, specifically only 3 months (Jun, Aug and Oct) while other datasets had data from May 2017 to Feb 2019.
2. There were 106 absent records, arising from only 62 out of 480 employees.
3. More than 30 employees were absent in each of those months, with an average of at least 15 days of absent per employee.

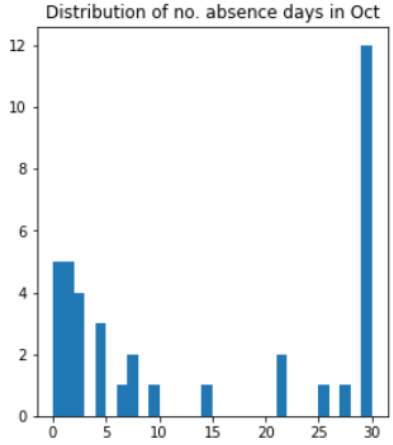




4. A significant number of employees were missing for almost the entire month.







5. Most absenteeism days were attributed to 'Common sickness or accident not related to job'

month	Common sickness or accident not related to the job	Long term sick leave	Non job related sickness	Workplace accident
6	551.0	29.0	29.0	6.0
8	471.0	60.0	NaN	7.0
10	409.0	57.0	NaN	55.0

6. 'Information' and 'Other' are the primary feedback types recorded and also generated a high number of dislikes.

feedbackType	employee	companyAlias	commentId	commentLength	likes	dislikes	commentDate
CONGRATULATION	115	1	325	200	76	22	299
CRITICISM	15	1	16	16	6	5	16
INFORMATION	158	1	1302	497	82	47	1111
OTHER	260	1	3188	540	65	43	2097
SUGGESTION	88	1	241	190	54	33	232

Absenteeism Machine Learning ("ML")

We use ML to predict absenteeism using two absenteeism definitions:

1. Scenario 1: we want to predict employees who will be absent for more than 2 days. We assume 2 days absenteeism has minimum impact on the business.
2. Scenario 2: we want to predict employees that will have more than 20 days of absenteeism. Long term absenteeism can lead to decreased productivity and resource management issues (Forbes, 2013). Pre-emptive actions need to be taken to ensure business continuity. 20 days appear to be a reasonable cut-off.

As a large proportion of absent employees were absent for a long duration where no activities during the absence period could be captured, and also due to the limitation that only 3 months absenteeism data is available, we decided to use employees' behavior data in last month to predict if the employees will be absent in the upcoming month for at least the numbers of days we specified in the scenarios above.

i. Features

We have used the following features:

1. Number of votes
2. Mean score of votes
3. Relative mean score (defined as employees mean score of votes divide by average mean score for the company)
4. Total dislikes
5. Total likes
6. Total dislikes by feedback types
7. Total likes by feedback types
8. Likeability (defined as "total likes" divide by sum of "total likes' and "total dislikes')
9. Total number of feedbacks
10. Number of feedbacks by types
11. Locations of employees

ii. SVM and Random Forest comparison

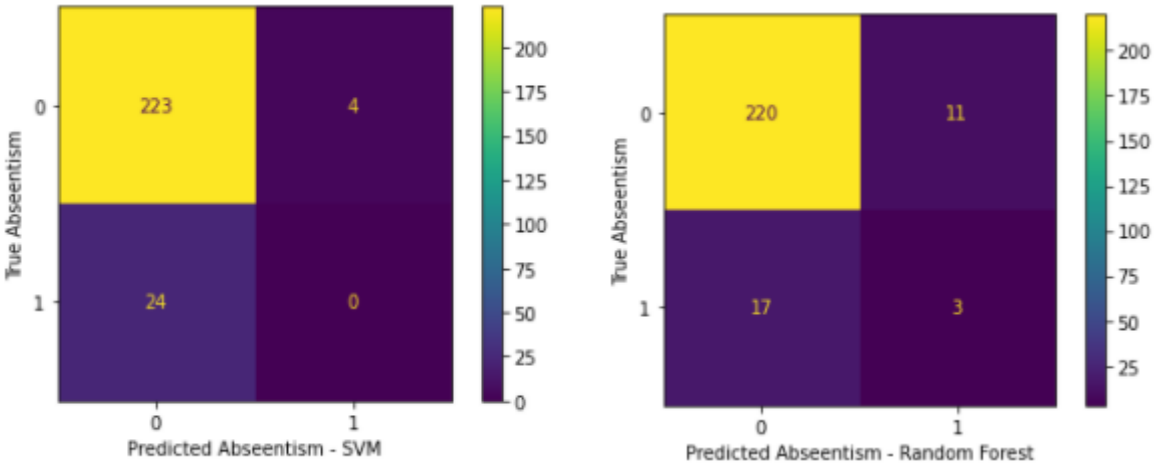
We ran SVM and Random Forest (10-fold cross validation) models on scenario 1.

The dependent variable is extremely imbalanced with limited positive cases and most ML techniques will ignore which will lead to poor performance on the minority class, despite it being most critical to successfully predict the minority class, i.e., absentees. We have used Random Oversampling and SMOTE (Synthetic Minority Oversampling Technique) to address this challenge.

	Proportion of Non-absentee	Proportion of Absentee
Scenario 1 (>2 days)	**0.925**	**0.075**
Scenario 2a (<= 20 days)	**0.946**	**0.054**
Scenario 2b (> 20 days)	**0.952**	**0.048**

Due to this imbalance issue, overall model accuracy is not an ideal metric. Random Forest is superior in terms of recall and precision compared to SVM, i.e., it predicts a higher number of absenteeism correctly, this is despite tree-based model being more prone to over-fitting. To improve the performance of SVM would require more hyperparameter tuning.

	Support Vector Machine	Random Forest
Accuracy	0.92	0.89
Recall	0.00	0.15
Precision	0.00	0.21

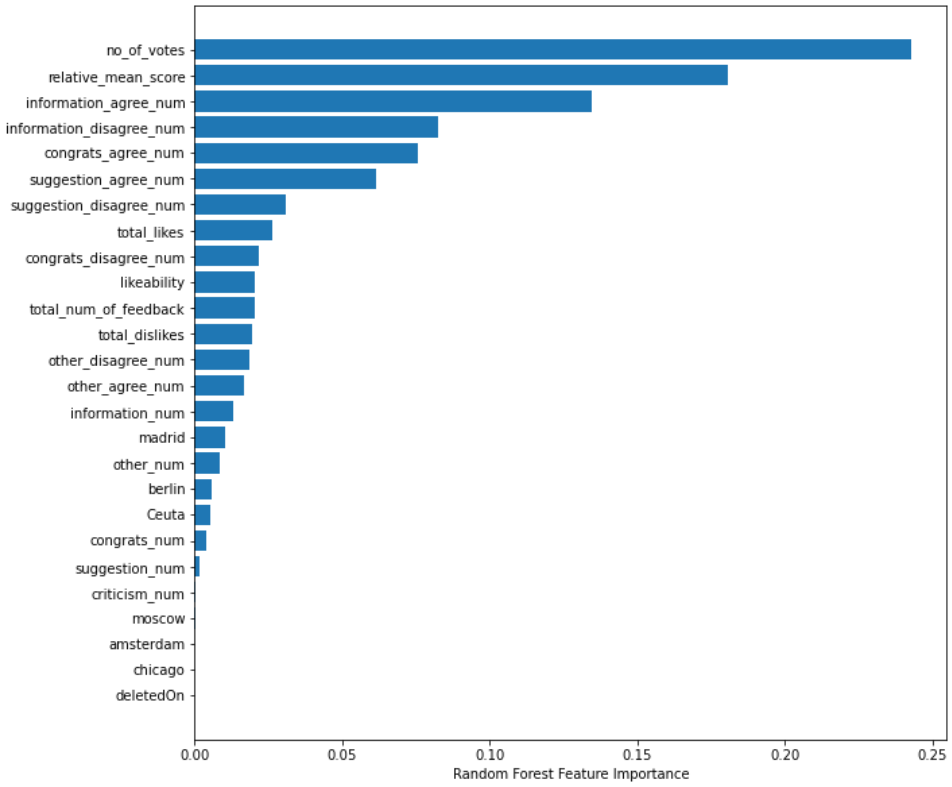


iii. Random Forest model

We decided to adopt Random Forest as the primary model to apply on both scenario 1 and 2. The results are shown below and 2 insights are noted:

- The model performed better when we predict more than 2 days of absenteeism and this could be due to the slightly bigger absenteeism dataset.
- The most predictive features are those related to votes and 'information' feedback type.

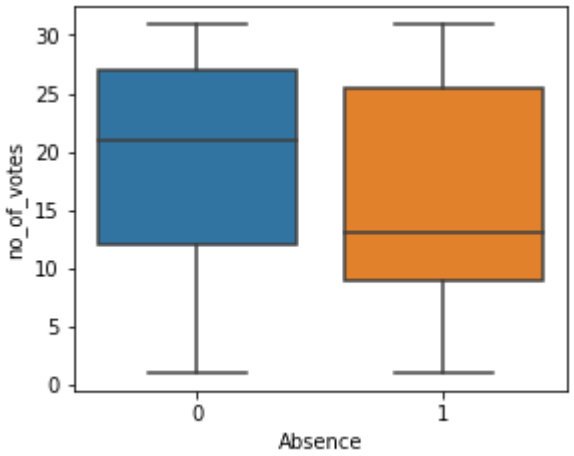
	Scenario 1 (>2 days)	Scenario 2a (<= 20 days)	Scenario 2b (> 20 days)
Accuracy	0.93	0.95	0.95
Recall	0.10	0.16	0.04
Precision	0.48	0.42	0.20
Most predictive features	1. Number of votes 2. Relative mean score of votes 3. No. of agree to 'Information' type feedback 4. No. of disagree to 'Information' type feedback 5. No. of agree to 'suggestion' type feedback	1. Number of votes 2. Relative mean score of votes 3. No. of agree to 'Information' type feedback 4. No. of agree to 'suggestion' type feedback 5. No. of agree to 'congrats' type feedback	1. Number of votes 2. Relative mean score of votes 3. No. of agree to 'Information' type feedback 4. No. of disagree to 'Information' type feedback 5. No. of agree to 'suggestion' type feedback



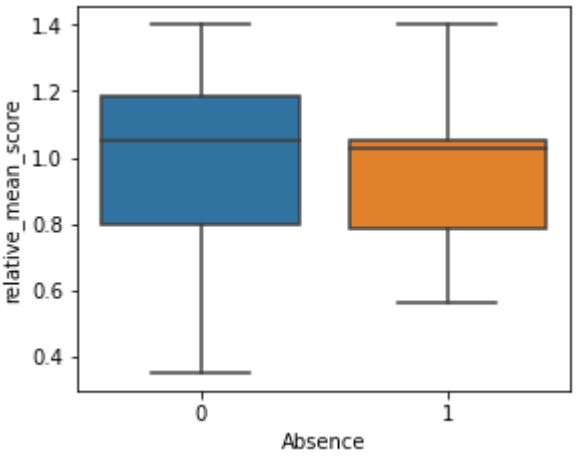
iv. Employee's categories

We built employees categories for using the above 4 most predictive features using scenario 1.

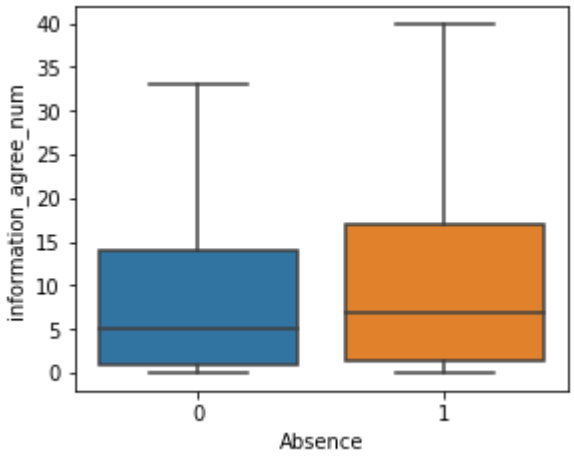
No_of_votes: Employees who tend to be absent generally vote less as the medium of 'absent' (~13) is significantly lower than that of 'present' (~21).



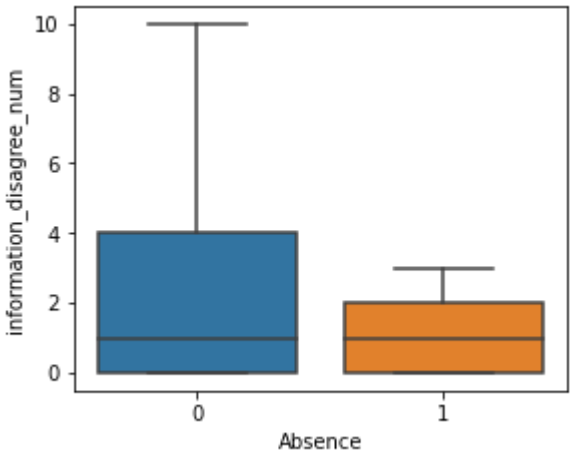
Relative_mean_score: Absent employees are almost certain to be less happy than the general population. However, ~50% of the non-absent employees also had below-average happiness score.



information_agree_num:



information_disagree_num:



Limitations

We recognized 3 limitations that could represent areas of improving the accuracy of the ML:

1. Limited absenteeism data provided as we had only 3 months of records compared to multi-year records for other features.
2. Data for absenteeism is imbalanced which depreciates the accuracy of ML models. If we then impose rebalancing, we sacrifice the data quantity, again resulting in less robust models.
3. We do not have any visibility to the actual comments made despite recognizing these comments can be highly predictive.

Recommendations to reduce absenteeism

Studies have shown that causes of employee absenteeism could include but not limited to burnout, stress and low morale, disengagement, depression, illness, injuries and family circumstances (Forbes, 2013).

Based on our analysis, participation on Happyforce to the question "How are you today?" is an important indicator and should be monitored closely. Employees are likely to be less engaged and hence more likely to be absent when they stop/have reduced participation on Happyforce.

According to HappyForce, information comments are words of encouragement. We noted that absent employees tend to agree more and disagree less with “information” type feedback. This implies that absent employees may need more encouragement (HappyForce, 2021).

Based on our analysis, the following are criteria we have set to determine whether an employee is of higher risk for being absent in the following month:

- Number of votes is fewer than 13 in the past month
- Relative mean score of votes is less than 1 in the past month
- Number of agree to ‘information’ type feedback is more than 15 in the past month

The following interventions could be adopted:

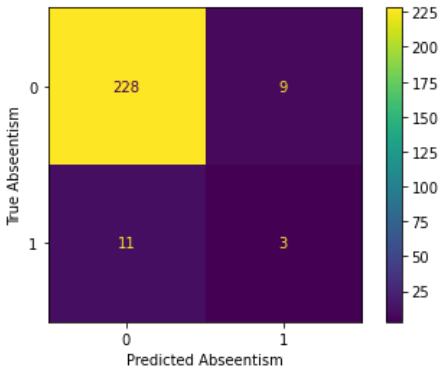
1. High-risk employees
    a. Semi-Automatic intervention: HappyForce system can be configured to send alerts to HR & the supervisor. They could take active intervention steps to engage the employees.
    b. Full-Automatic intervention: HappyForce & HR system could interact to automatically send these employees some rewards to motivate them like gifts or inspirational material.
2. Employees in general who exhibit positive trends, for a defined period of time, could be recognized or unlock further rewards.

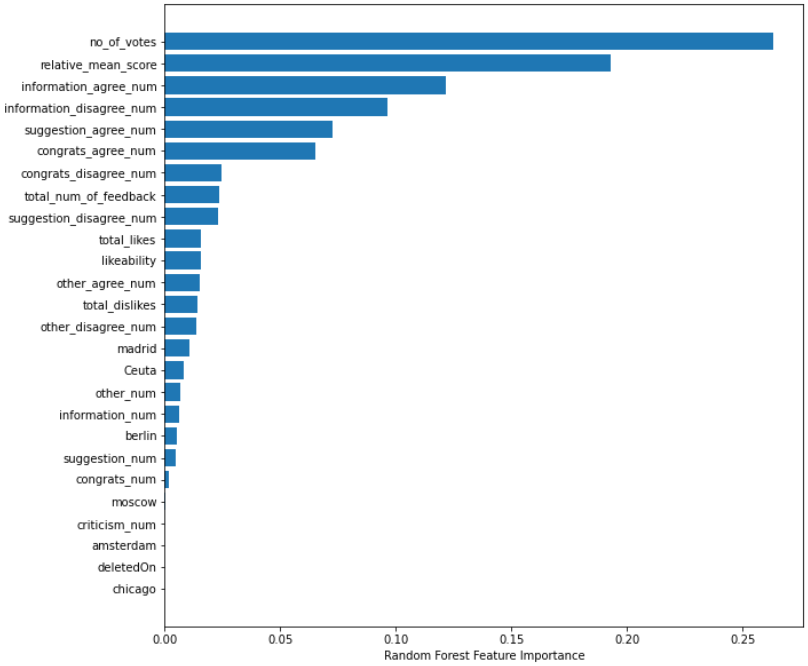
References:

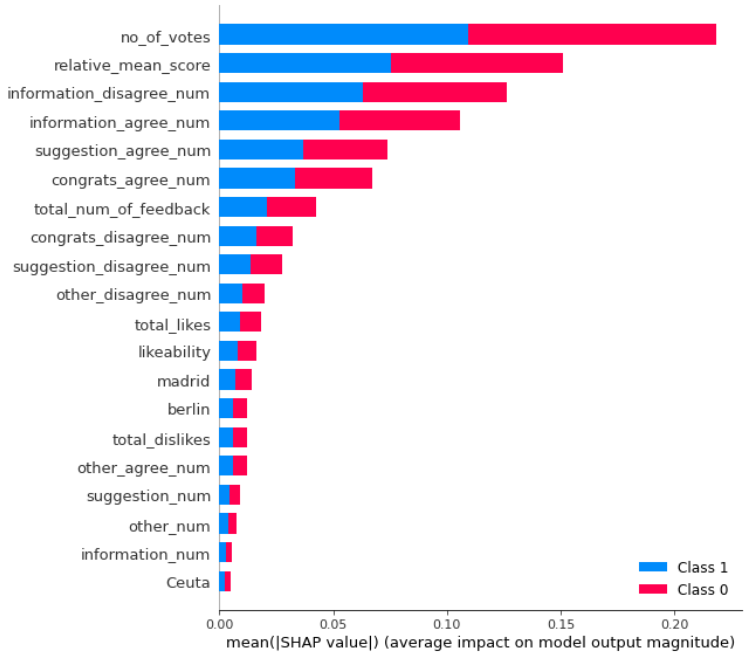
1. Forbes Investopedia, 2013, *The Causes And Costs Of Absenteeism In The Workplace*, viewed 01 November 2021, <https://www.forbes.com/sites/investopedia/2013/07/10/the-causes-and-costs-of-absenteeism-in-the-workplace/?sh=4f2400403eb6>.
2. HappyForce, 2021, viewed 10 November 2021, https://myhappyforce.com/en/

Appendix – selected outputs from Random Forest

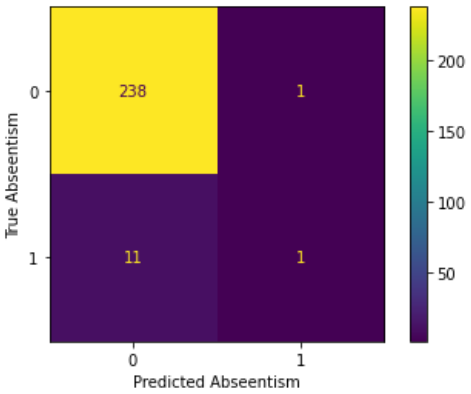
- More than 2 Days of Absenteeism

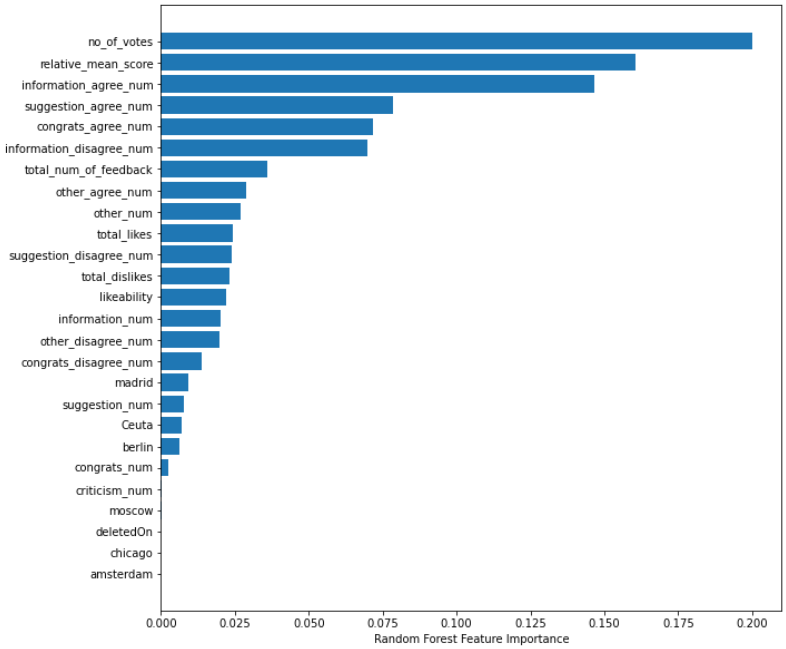


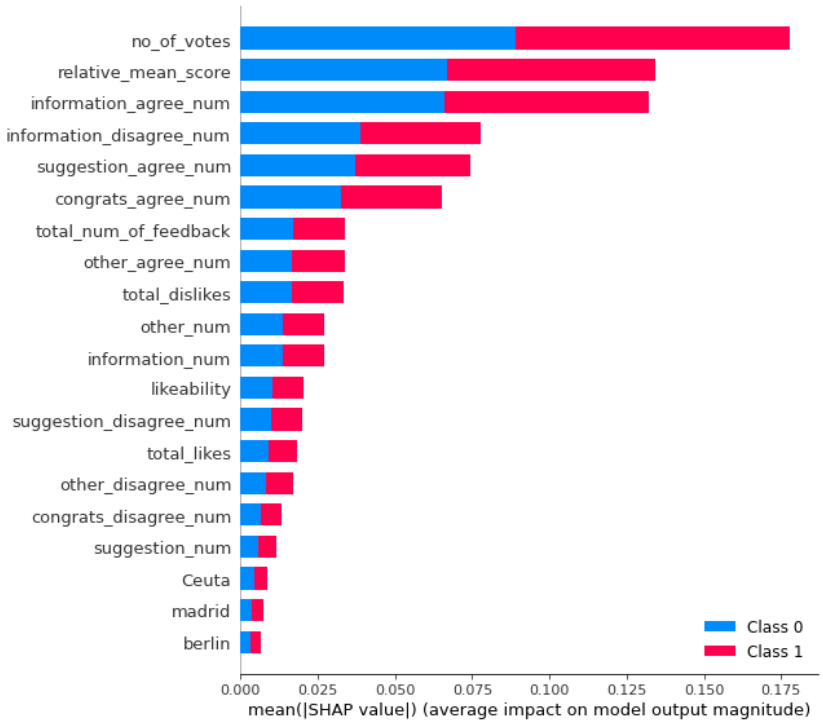




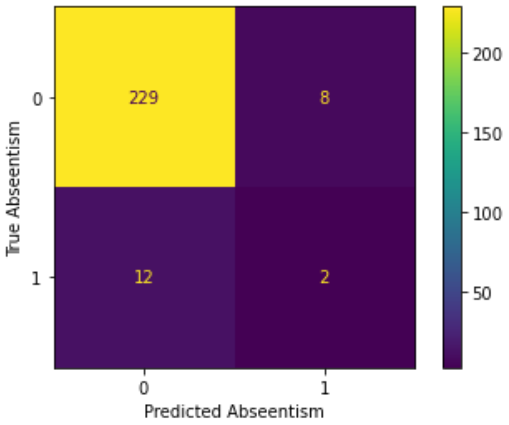
- Less than 20 days of Absenteeism

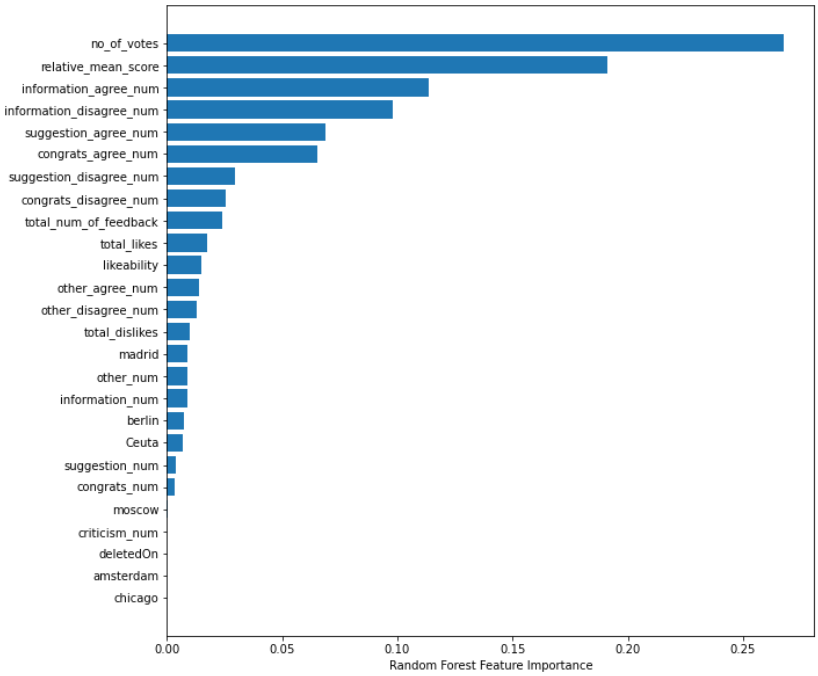

238
1
11
1
0
1
True Abseentism
Predicted Abseentism
200
150
100
50

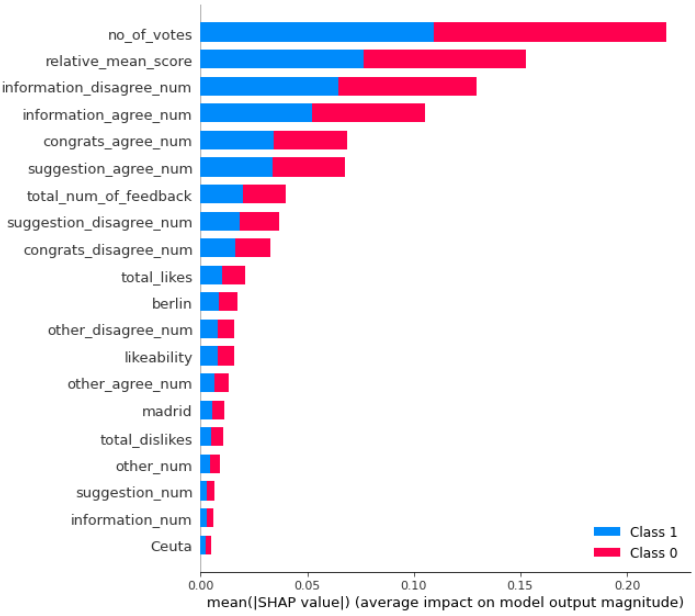

no_of_votes
relative_mean_score
information_agree_num
suggestion_agree_num
congrats_agree_num
information_disagree_num
total_num_of_feedback
other_agree_num
other_num
total_likes
suggestion_disagree_num
total_dislikes
likeability
information_num
other_disagree_num
congrats_disagree_num
madrid
suggestion_num
Ceuta
berlin
congrats_num
criticism_num
moscow
deletedOn
chicago
amsterdam
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Random Forest Feature Importance


no_of_votes
relative_mean_score
information_agree_num
information_disagree_num
suggestion_agree_num
congrats_agree_num
total_num_of_feedback
other_agree_num
total_dislikes
other_num
information_num
likeability
suggestion_disagree_num
total_likes
other_disagree_num
congrats_disagree_num
suggestion_num
Ceuta
madrid
berlin
Class 0
Class 1
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
mean(|SHAP value|) (average impact on model output magnitude)

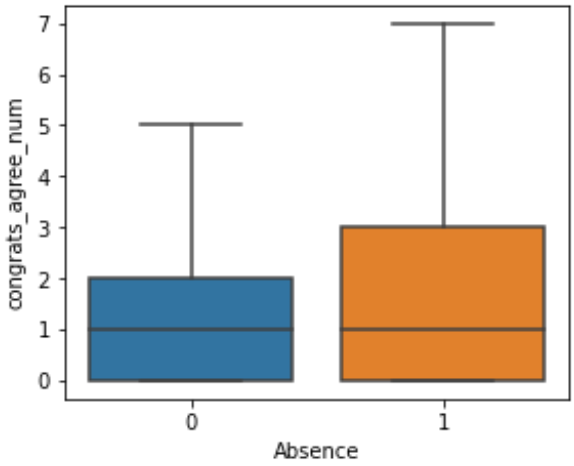
- More than 20 days of Absenteeism



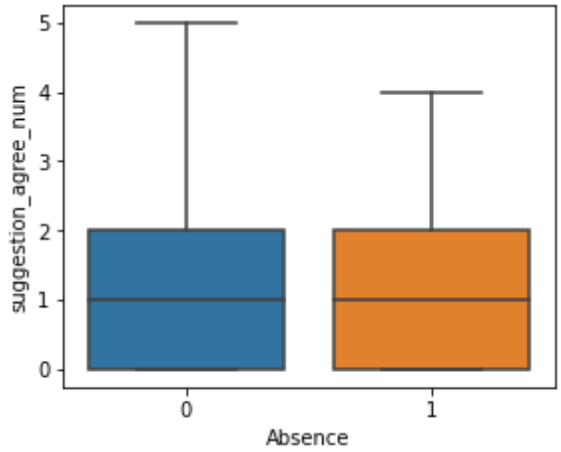




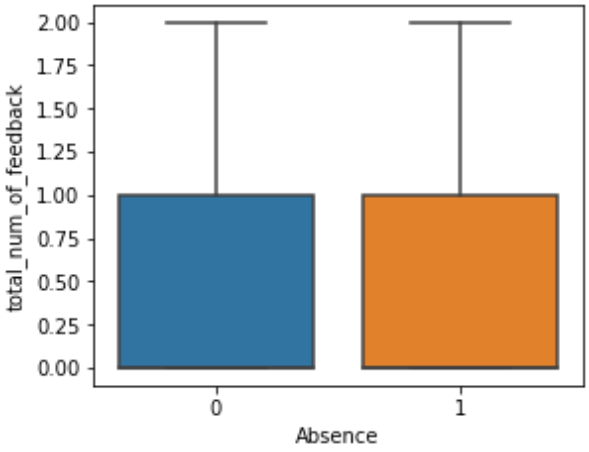
Congrats_agree_no:

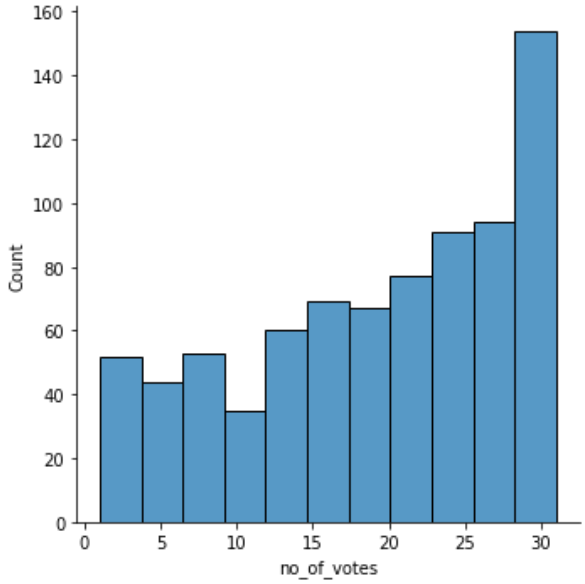
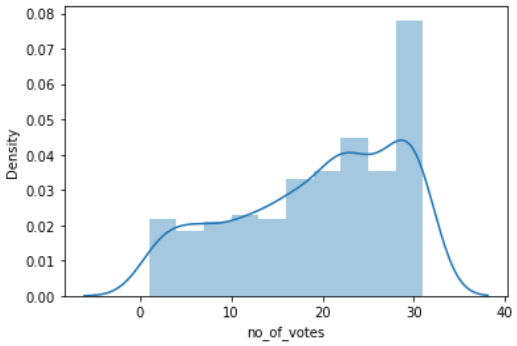
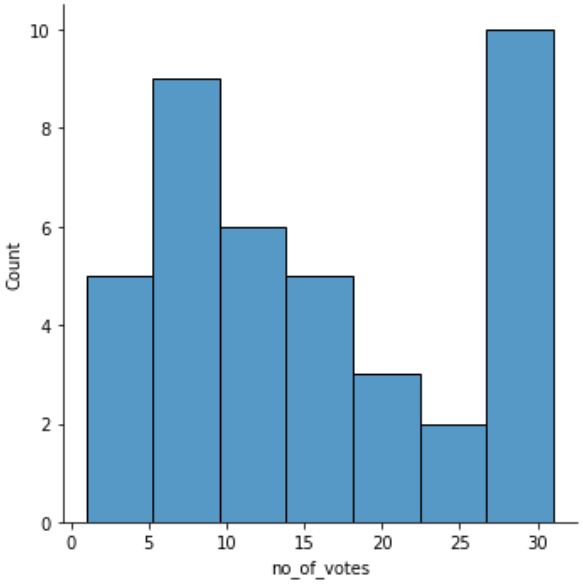
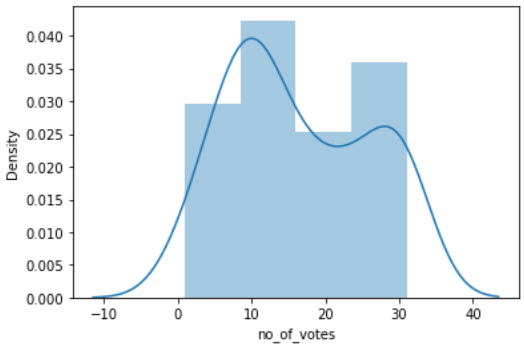
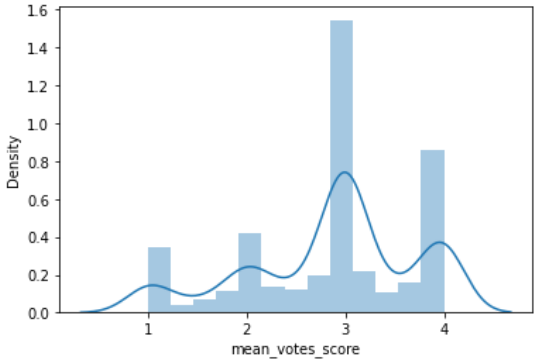
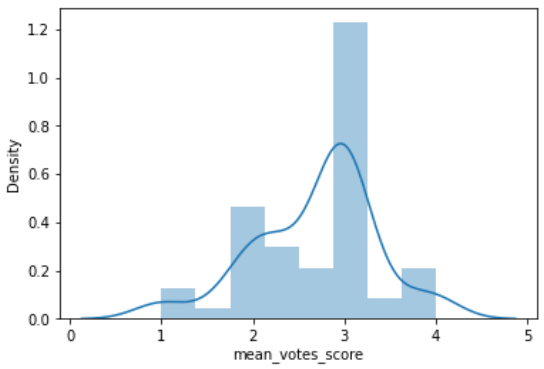


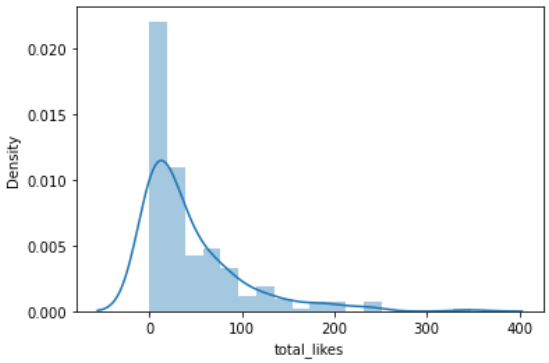
suggestion_agree_num:



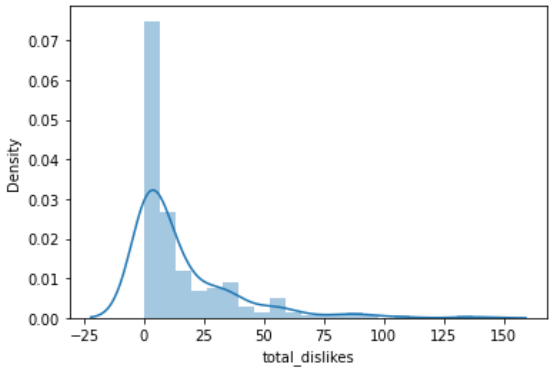
total_num_of_feedback

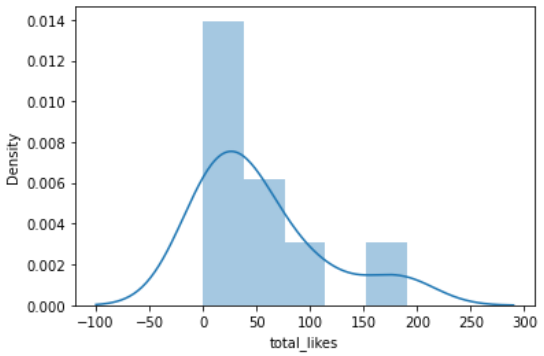


Present	Absent for June, August and October 2018
Number of votes   	Number of votes  
Mean Vote Score  	Mean Vote Score 
Total Likes	Total Likes



Total Dislikes





Total Dislikes