Data Science/Theoretical Insights

The number of absentees predicted by the Support Vector Machine and Random Forest ML algorithms are brittle.

SVR leads to a higher R^2 score (0.530) than Random Forest (0.294) - R^2 ranges from 0 to 1, with 1 being the best outcome. Cross-validation suggest that SVR is the better model of the two, with c53% of the variance in the dependent variable (number of absentees per day) is predictable from the independent variables e.g. average score, comment length/count, number of dis/likes, feedback type, dis/agree etc.

Testing the model using the test dataset (the most recent 25% of data that I did not use in training/validation) to assess brittleness was revealing. Both models led to negative test R^2 scores: SVR (-0.055) & Random Forest (-0.119). This implies that the mean of the test data supplies a better fit to the outcomes than do the fitted model values. In short, our machine learning models do a poor job are predicting the number of absentees per day i.e. ML models are brittle.

Why the poor predictions? One plausible reason is that we have used the default hyper-parameters for the ML models and have not taken the time to fine-tune these parameters to see if we can improve the predictions. This is a potential future avenue of investigation.

But this is fine, through cluster analysis we can still segment the data into different personas.

Actionable Insights

Given the data science insights, In segmenting the employees, we can create 5 personas for our actionable insights –

0: Happy employee - high vote rating, minimal absences apart from common sickness, but little participation in commenting/voting

1: Voter employee - like the "Happy" employee, but is an active voter

2: Commenter employee - like the "Happy" employee, but is an active commenter

3: Absent employee - often sick employee, minimal participation, and low vote rating

4: Vocal employee - like the "Happy" employee, but is both an active voter & commenter

These personas will inform us on what we can analyze the issue to be. From SVM & Random Forest mode of analysis, we note that it is difficult to predict the number of absenteeism per day. However, RFE suggests that there may be some drivers that are fitting with "common-sense". E.g.

- The vote score re. "How are you today?" is particularly revealing. The organization could check these scores across different teams, and deep-dive into root-cause when the score dip / are low e.g. has there been a change in leadership / management of these teams?

- Actively watch the comments / feedback e.g. has there been a drop on by number of comments - what are the reasons? E.g.
  - It could be it because employees do not see management take actions in response to feedback posted?
    - *Management* should read the comments (CONGRATULATION, CRITICISM, OTHER, SUGGESTION) to set up the reasons behind this, and publicly show there is measurable with respect to employee suggestions.

Finally, clustering suggests that employees interact with the Data Collection app (deployed by the client) differently - some prefer to vote, others prefer to comment. However, the often-absent employee has low participation rate on the app, and when they do take part, have a low rating. This can be used as a "red flag" early-warning indicator on which employees are particularly at a low-ebb and managers may use this to intervene / signpost the employee to more support e.g. Cognitive Behavior Therapy, family care, DEI programs to check for undue biases etc.