IIT MADRAS

CS4830 : Big Data Lab Project Report

Divya K Raman(EE15B085), Sahana Ramnath(EE15B109), Vaibhav Nayel(EE15B117); Team: Grobe Daten
April 22, 2019

1 PROJECT STATEMENT

In this project, we aim to predict if an employee is still working for a company or not. Various indicators like happiness level of employees over time, how they react to comments made by other employees, the comments that they make, average satisfaction level of employees belonging to the company, proportion of good and bad comments liked by the employees, etc are used to arrive at this conclusion.
This project was implemented using pyspark, kafka and spark streaming.

2 THE DATASET

For training, we are given 4 files.

The file comments_employee_mapping.csv maps each comment to its owner and also specifies the date when the comment was made.

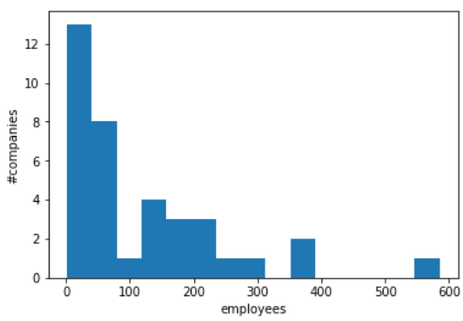
The comments_likeability.csv file has details about the reaction of employees to comments.

The file happiness_level.csv contains details about the happiness vote given by each employee on various days.

The employee_attrition.csv file maps each employee to the number of votes, last participation date and labels if the employee still exists in the company or not.

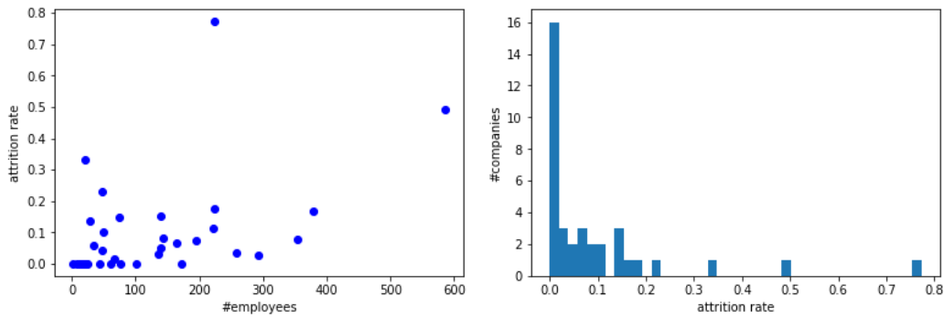
3 ANALYSIS OF THE DATASET

The dataset has 37 unique companies. Files employee_attrition.csv and happiness_level.csv contain all the companies but there are 3 companies whose employees don't like any comments and 1 company whose employees don't make any comments. This necessitates our model to be able to handle inputs which have no likes or comments. Total employees in the dataset is 4418.Therefore, the average number of employees per company is 119.405405405.
The below histogram depicts the number of employees that each company has :



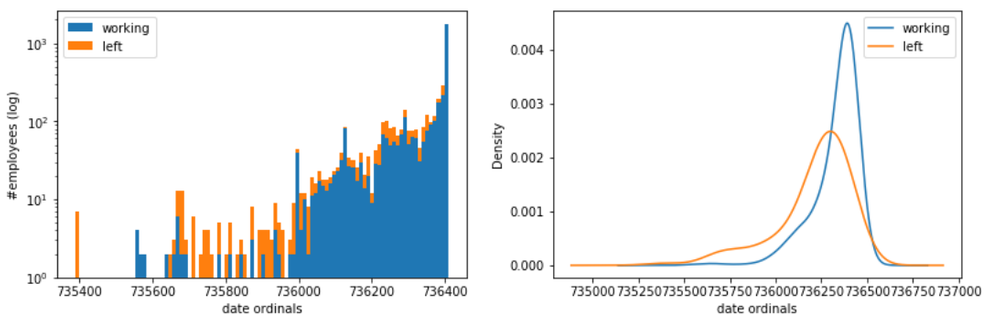
Split of employees overall, still working and those who left:

- still working: 3673
- left: 745

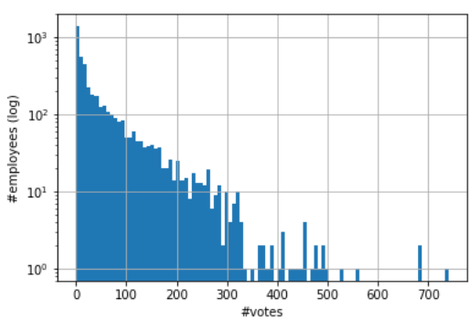


Correlation Coefficient: 0.47252133218050585
A correlation coefficient around 0.5 suggests a correlation of company size with attrition rate.

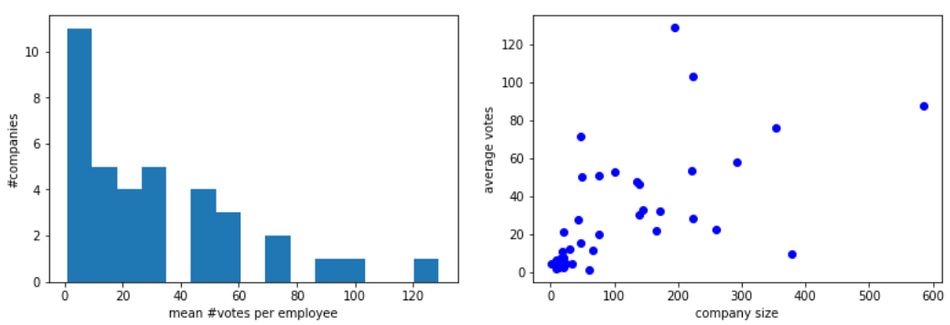
Relation between the last date of participation and likelihood of leaving :



Clearly, the further back the last date of participation is, the more likely the employee is to have left.
Number of votes made by each employee :

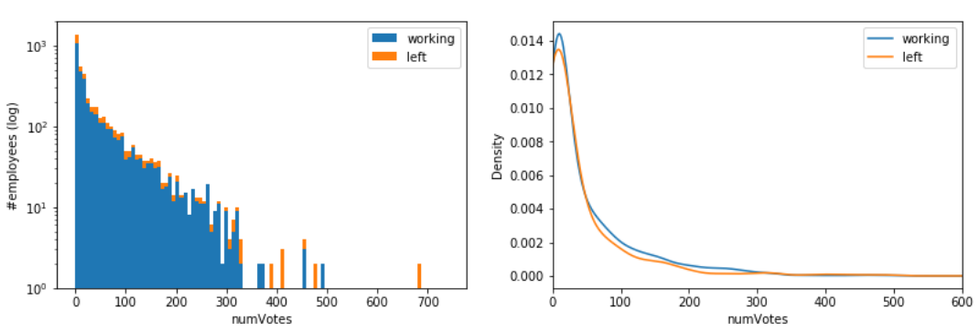


Most employees have very low vote counts.
Are the average number of votes dependent on the company? Correlation Coefficient : 0.5765644245292082

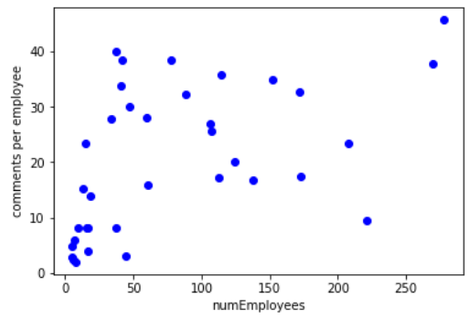


Some companies witness lower participation than other companies. This could be because employees in bigger companies are more likely to socialize or because some companies are very new and the numbers haven't had a chance to become higher.

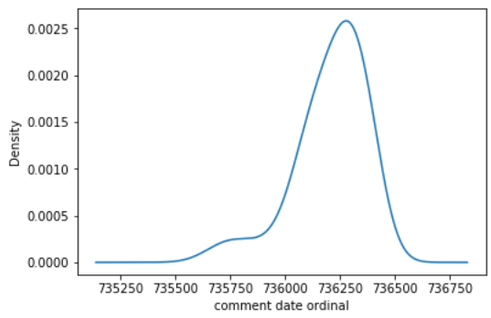
Relation between number of votes and likelihood of leaving the company:



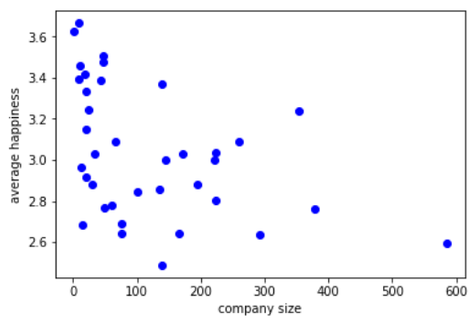
The number of votes by an employee doesn't seem to indicate anything about likelihood of leaving,i.e, unhappy employees are just as likely to vote as happy ones.
Comments per employee:



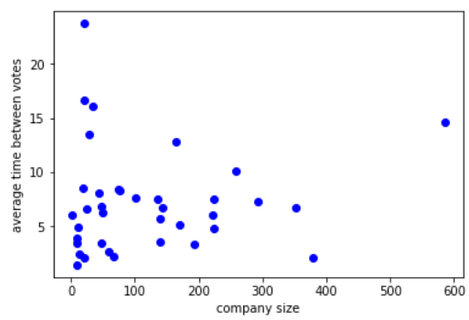
There is a clear correlation between company size and comments made per employee



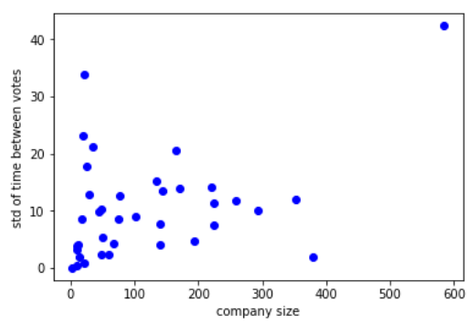
This seems to mirror the last participation date density. The number of comments seems to mirror the number of workers in the system at any time.
Variation of happiness with company : Correlation Coefficient - -0.4302256906883859



The plot shows a distinct negative correlation of company size with employee happiness.
Time between votes: Correlation Coefficient - 0.06228331496742335



Number of employees who voted only once: 458



A low correlation suggests that company size does not seem to affect the average time or variance of times between votes.

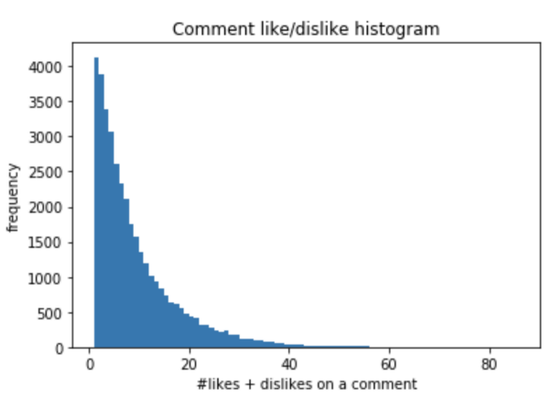
4 Preprocessing and Feature Engineering

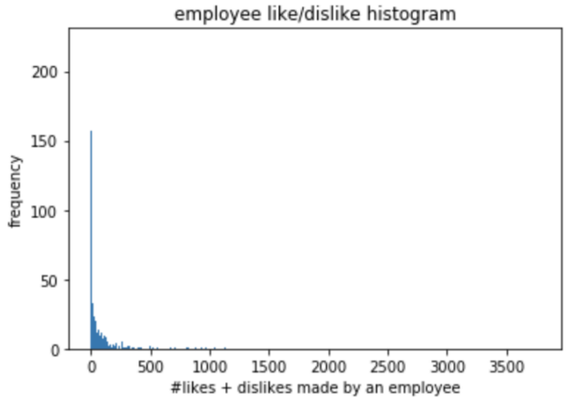
4.1 Happiness Votes

Each employee votes on a scale of 1-4 (1 is lowest, 4 is highest) as to how happy he/she is. Since the task is to predict the current employment status of the employee, we tried out two ways of using the employee's votes over multiple days. We tried exponentially weighing the votes till the present day, as well as using just the final vote(date-wise). In our experiments, we found that using the final vote gave better results, which is what we've used in our final best model.

4.2 Comments and Likes

In this case again, we've taken the last comment posted and last comment liked/disliked by the employee for our final model. We tried creating sparse comment-made and comment-liked matrices for all employees versus all comments matrices to use in our models; however, the resulting number of features was too high and was infeasible. Hence we resorted to using the latest comment made and liked/disliked.





5 Best Model

Our results were obtained using a 20% train-test split using the following features:

- Last comment made
- Last comment liked or disliked and whether like/dislike
- Last happiness vote
- Company Alias with one-hot encoding
- Last date ordinal from the attrition file

We experimented with Decision Trees, SVMs, Random Forests, Logistic Regression etc. Our final best model uses logistic regression. We obtain 84.7 percent accuracy on the train set and 83.1 percent accuracy on the test set. The code is wrtten in pyspark. We use databricks to run the code.

6 Code

We use kafka producer to send each sample of our final pre processed data(producer.py). This is received in a spark streaming session, and we use our model to generate the predicted result on this sample(sparkstream.py). These codes are run on our laptops.

7 Conclusion

In this project, multiple models were tested on the employees dataset to predict if an employee will stay in the company or not. We use kafka and spark-streaming to transfer and test on test data. We use various feature engineering and pre processing techniques to process the given data. We then test our features on various models and apply hyperparameter tuning to get the best results.