FINAL PROJECT

(How do we improve employee engagement?)

Mark Chen
Student ID: 5506208

1. Introduction

Employee engagement has become one of the hottest topics discussed by Human Resources professionals today. High levels of engagement promote retention of talent, foster customer loyalty and improve organizational performance and stakeholder value [1]. However, HR people have always been struggling with identifying the key factors driving employees to engage better in an organization. As a major component of employee engagement, employee satisfaction is often used as an indicator of engagement level. So this project is designed to analyse a data set containing over 67,000 employee reviews for Google, Amazon, Facebook, Apple, Microsoft, and Netflix, in order to understand what do employees value the most and try to find alternatives for engagement improvement in large high-tech companies.

This project uses CSV file, "employee_reviews.csv", downloaded from Kaggle as the dataset. I have identified several important variables such as company, location, job-title, summary and so on. Variables like company, location, job-title and summary are all Strings. Moreover, several rating scores are stored as type int such as overall-ratings, work-balance-stars and so on.

There are four questions to answer in the project. The questions are designed for leading me to analyse and visualize the data.

1. Which company has the best overall-rating score? (To find the highest mean among 6 companies)
2. What's the relationship between overall-rating score and these five stars? (Five stars: work-balance-stars, culture-value-stars, career-opportunities-stars, comp-benefit-stars and senior-management-stars.)
3. What are the most frequently mentioned Pros and Cons of each company?
4. Can we use the Five Stars (the 5 other rating score mentioned in Question b) to predict the overall rating score?

2. Related Work

The data set I choose is an open data, which has also been researched by other people in a different way. Firstly, the approaches are different. Other people focused on comparing five stars scores among different companies to find out who did better in each aspect, while I concentrated on looking for the most important factor among the five stars. Secondly, I used more tools to give a more detailed description of the data set. Moreover, I tried to create the prediction model in order to give more evidence on my assumption.

3. Process

The data set was downloaded from Kaggle. The project was originally written by Jupyter Notebook.

Question 1

First of all, I loaded the "employee_review.csv" file into DataFrame "df_reviews" as my database. Then I chose column "company" and "overall-ratings" to be a new DataFrame "df_overall". To find the highest overall-rating score among 6 companies, I grouped the "overall-ratings" data by "company" and sorted it by average "overall-ratings".

Question 2

I started with extracting all 6 columns about rating scores. To draw the relationship between overall-rating score and five stars, I had to find an effective way to take a look at correlations between these items. Then I came up with heatmap, a tool I used in one of my previous consulting projects to analyse the survey results for my client. But I didn't know how to realize heatmap by python. So I searched online and digged the official documentations to create one. However, some scores in the five stars are stored in string type. So I converted those "special" data to float type. To give an accurate description about the relationship, I also used OLS to conduct a regression analysis. Same with the heatmap, I applied the spirit from the official documentation to finish this regression analysis successfully.

Question 3

Similar to the first two questions, I firstly extracted the related columns, "pros", "cons", advice-to-mgmt". To avoid the noise from the unrelated items in the data, I used a list storing all reviews to drop all stopwords, non-alphabet words, and punctuation. To find the most frequently mentioned words in this data set, at the beginning, I came up with the idea to calculate each character's frequency and rank them by frequency value. But I found it a little annoying and confusing in writing the code. What's more, I realized it would not be an efficient way to do this work. With no clue about this challenge, I went to my friend, a data scientist, for some suggestions on the commonly used tools to solve this problem. Then I learned about

WordClouds from our discussion and I read the tutorial of generating WordClouds from DataCamp [2]. Finally I used WordClouds to show the most frequently used words in Pros and Cons of each company successfully.

Question 4

After extracting the overall-rating score and five stars to a new DataFrame "df_reviews_predict", I dropped the rows of null data and converted the strings to float. In order to find the potential predicting model, I learned the basic knowledge of machine learning. Based on the approach introduced by DataFlair website [3], I split the "df_reviews_predict" data set into training and testing data sets. I used the ratio of 80-20 for the splitting, because I learned from the internet that it's the most common practice in data science. Then I used two common methods, regression analysis and decision tree, to do the prediction job.

4. Results

Question 1

As shown by Table 1.1, Facebook has the highest average overall-rating score, which is 4.51 out of 5. Google ranks 2nd with 4.34, while Apple ranks 3rd with 3.96, Microsoft ranks 4th with 3.82 and Amazon ranks 5th with 3.59. Netflix received the lowest score among all 6 companies at 3.41 out of 5.

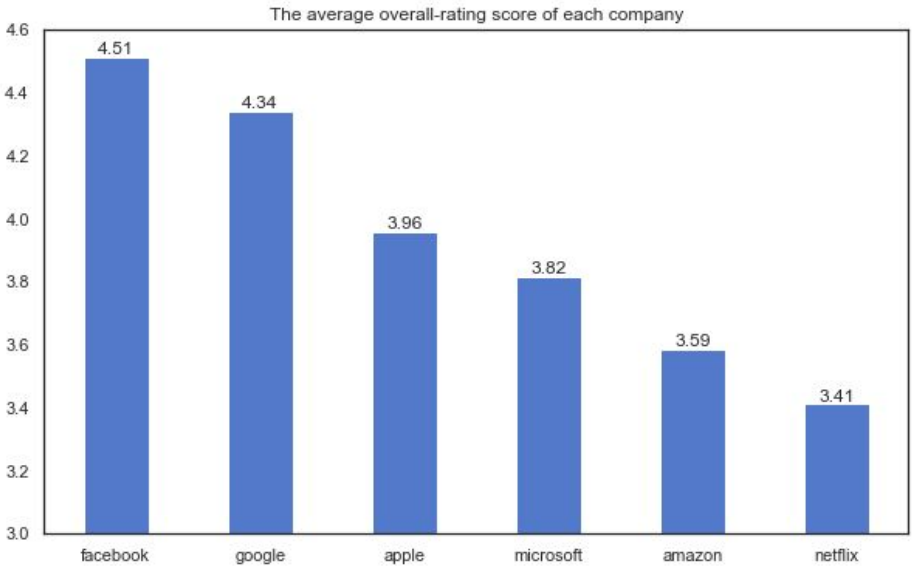

Table 1.1: The average overall-rating score of each company

Table 1.2 shows the distribution of overall-rating scores for the highest rated company Facebook and lowest rated company Netflix. The green is for Facebook, the purple is for Netflix. Then we can find the major reason why Facebook received the highest rating score is that most

employees give a full score of 5. However, for Netflix, there are a large number of employees giving the low scores of 1 and 2, which results in the low average score.

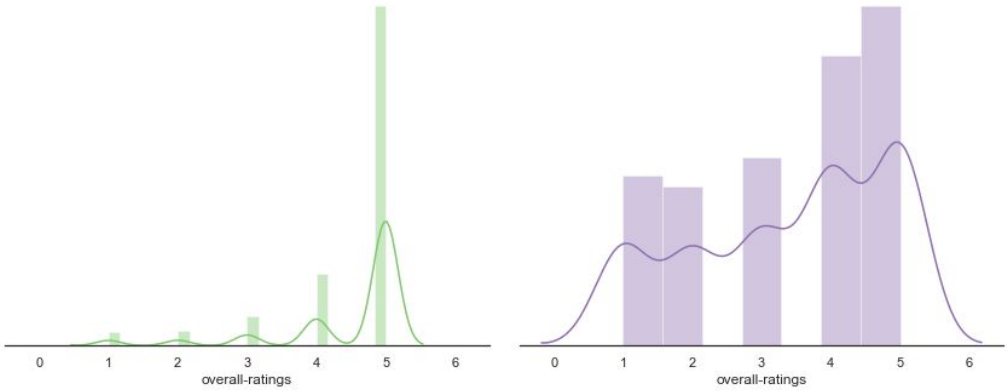

Table 1.2: Distribution of overall-rating scores for Facebook and Netflix

Question 2

From the heatmap (Table 2.1), we can see that the culture-values-stars has the strongest positive relationship with overall-ratings, while senior-management-stars has the 2nd strongest relationship with overall-ratings. The comp-benefit-stars has the weakest relationship with overall-ratings.

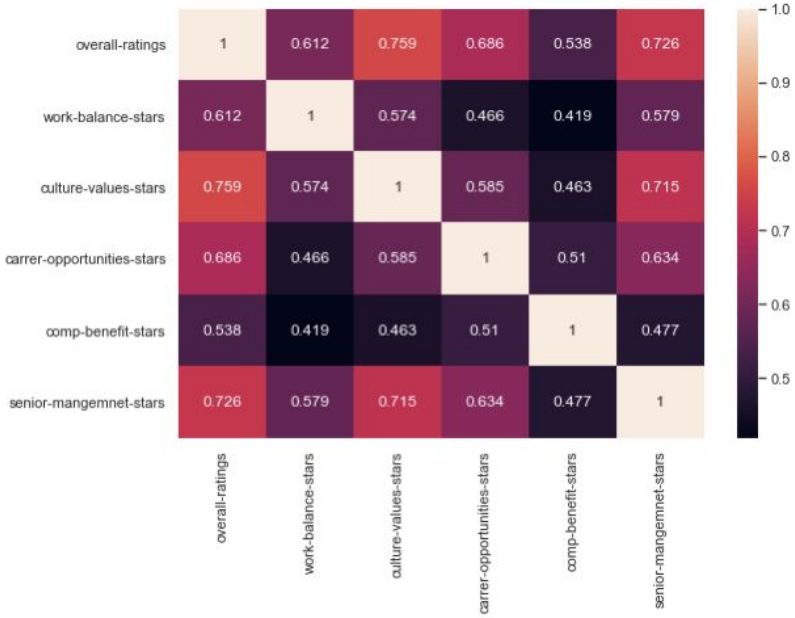

Table 2.1: Heatmap of relationship among overall-ratings and five stars

Results of culture-values-stars and senior-management-stars share a similar story with reality. Employees will be heavily affected by the company's culture and senior management. However, people also usually believe that financial incentives are the easiest way to motivate employees. The results from heatmap displays an opposite assumption, because comp-benefit-stars has the weakest relationship with overall-ratings, which means that the financial factor may play a less important role in employee engagement.

OLS Regression Results

Dep. Variable:	overall_ratings	**R-squared:**	0.713
Model:	OLS	**Adj. R-squared:**	0.713
Method:	Least Squares	**F-statistic:**	2.650e+04
Date:	Thu, 19 Dec 2019	**Prob (F-statistic):**	0.00
Time:	12:27:17	**Log-Likelihood:**	-50232.
No. Observations:	53222	**AIC:**	1.005e+05
Df Residuals:	53216	**BIC:**	1.005e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.3860	0.011	33.990	0.000	0.364	0.408
work_balance_stars	0.1298	0.003	48.922	0.000	0.125	0.135
culture_values_stars	0.3159	0.003	99.147	0.000	0.310	0.322
carrer_opportunities_stars	0.2277	0.003	75.014	0.000	0.222	0.234
comp_benefit_stars	0.1110	0.003	35.551	0.000	0.105	0.117
senior_mangemnet_stars	0.1658	0.003	50.681	0.000	0.159	0.172

Omnibus:	2701.070	**Durbin-Watson:**	1.929
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	9810.491
Skew:	-0.114	**Prob(JB):**	0.00
Kurtosis:	5.091	**Cond. No.**	36.1

Table 2.2: OLS Regression Results

The OLS Regression Results displays more details from statistics. The Adj. R-squared is 0.713, which means the model can explain about 71% of the variances. With p-value = 0, we can also say the results is statistically significant.

Question 3

Table 3.1 is the WordCloud result of the most frequently mentioned words in Facebook's employee notes about pros of the company. Then we can find that people, culture, benefit and team are the most often mentioned words. Based on this result, combining with the high overall-rating score of Facebook, we can assume that people, culture, and benefits could be the major factors affecting employee's satisfaction level.

Table 3.1: WordCloud result of Facebook (pros)

So what about the opposite example from Netflix, the company receiving the lowest rating score? Table 3.2 is the WordCloud result of Netflix's cons comments. People, team, manager, and culture are frequently mentioned when employees complain about Netflix. This result matches with the result from Facebook's pros. Both people and culture are mentioned frequently.

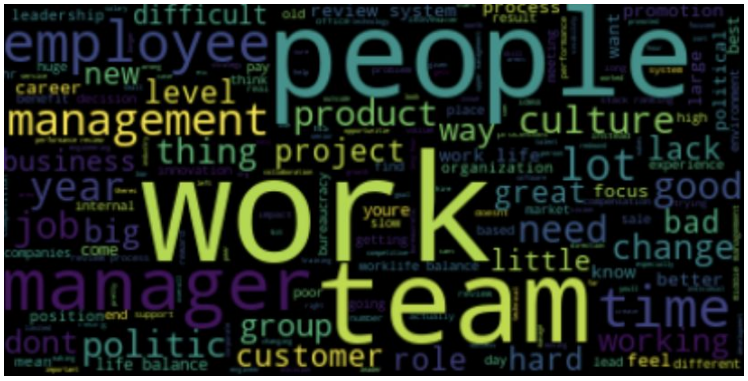

Table 3.2: WordCloud result of Netflix (Cons)

Table 3.3 shows the WordCloud result of Netflix's advice. We can easily find that people (employee), and team are frequently mentioned. Another highlighted word, "need", might indicates that employees in Netflix often want to improve something but may not receive the permission from managers. So the manager, management and change also relatively frequently appear in the advices from employees to Netflix.

Table 3.3: WordCloud result of Netflix (Advice)

Question 4

Table 4.1 shows the importances of five stars to the prediction of overall-rating score by Decision Trees. As we can see from this result, Culture-values-stars plays the most important role in predicting the overall-rating score. This provides another support to the assumption mentioned in the beginning that culture is a major factor affecting employee's engagement.

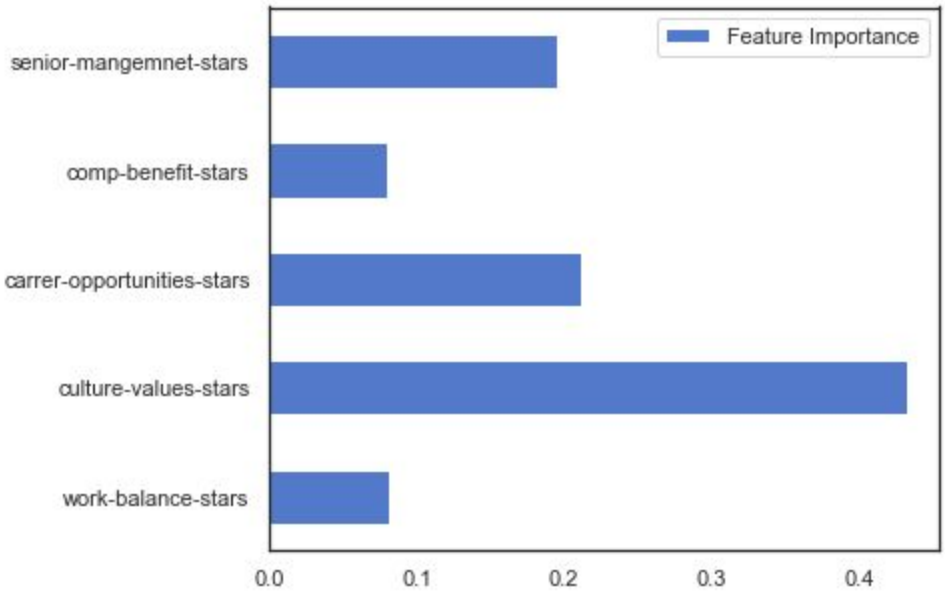

Table 4.1: Feature importance from Decision Trees

5. Suggested Alternatives

Based on the results and analysis of this project, we can make several recommendations for big high-tech companies to help improve their employee engagement level.

Strengthen culture construction within the organization. As culture is the most important factor affecting employee's overall satisfaction of the company, the organization should build up a strong company culture to motivate and retain talents.

From the result of this project, team is often mentioned by employees, which indicates that teamwork should be valued. So, culture building can start from enhancing collaboration in the company. Games are one of the best and most fun ways to build new relationships or amend old ones. The big high-tech companies could include coding competition, problem-solving games, etc. to engage tech people better.

Let managers become effective coaches. Managers also play an significant role in employee engagement. As the results suggesting above, senior management may be the most important factor after culture affecting employee engagement. Weak leaders would negatively influence the efficiency and effectiveness in an organization. To solve this problem, we need to have managers to become effective better leaders for employees.

Effective coaching can solve this problem. The goal of coaching is to solve performance problems and develop employee's capabilities. The main goals of coaching are to build a bond of trust, have a relationship to allow you to teach via expert role or process consultant [4]. So letting managers become effective coach will improve the quality of management as well as create a healthy partnership in the workplace.

According to the coaching guide from Damia Goldvarg [4], managers can start from signing the coaching agreement with coachee, then follow the steps to conduct coaching, including conversations with client, adjustment by frequent feedback, and evaluation on coaching effectiveness.

6. Conclusion

This project analyses the employee review data from 6 big high-tech companies in the US and provides deeper understanding on the relationship between different factors affecting employee engagement. Moreover, it shows the importance of culture and management in improving the engagement level in an organization.

To continue researching the engagement case after this project, we need to have more data sources, instead of using employee review data only. In addition, to understand the data from this project better, we can import more advanced tools other than WordClouds, to interpret the meaning of each whole sentence in the comment section of our data set. Because the selected frequently mentioned words may not be sufficient to show people's detailed attitude accurately, we need to understand the hidden meanings behind the whole sentence.

References

[1] Developing and Sustaining Employee Engagement. (n.d.). Retrieved from https://www.shrm.org/resourcesandtools/tools-and-samples/hr-qa/pages/whatisjobrotation.aspx.

[2] Vu, D. (2019, November 11). (Tutorial) Generate Word Clouds in Python. Retrieved from https://www.datacamp.com/community/tutorials/wordcloud-python.

[3] Team, D. F. (2019, July 10). Train and Test Set in Python Machine Learning - How to Split. Retrieved from https://data-flair.training/blogs/train-test-set-in-python-ml/.

[4] Goldvarg, D. (2018). *Professional coaching competencies: the complete guide*. Arroyo Grande, CA: Executive College Press.