Employee Reviews' Analysis

Section 1: Motivation

1.1 Short description

My project focused on employee reviews for Google, Amazon, Facebook, Apple, Netflix and Microsoft. I aimed to find out how the employees feel of working in those top technology companies.
I chose this topic because many people like me dream to work in those companies after graduation. It's better for us to learn more about them and figure out which company fits us best.

1.2 Four questions

1) Which company has the highest average overall-rating score among those companies?
2) There are five other rating scores, which are work-balance-stars, culture-values-stars, career-opportunities-stars, company-benefit-stars and senior-management-stars. Try to find the relationship between those five rating scores and the overall-rating score.
3) What are the most frequently used words when the employees wrote pros, cons and advice for their companies?
4) Besides the overall-rating score, there are five other rating scores mentioned in question 2. We can use those five rating scores to predict the overall-rating score. Try to find the best predictive model.

Section 2: Data Source

URL: https://www.kaggle.com/petersunga/google-amazon-facebook-employee-reviews
Data format: .csv file
Important variables:
1) company: Company name (string)
2) pros, cons and advice-to-mgmt (string)
3) overall-ratings (float)
4) work-balance-stars, culture-values-stars, career- opportunities-stars, company-benefit-stars and senior-management-stars (string)
Number of Records: 67529
Time periods: 30 Jan 2008 to 10 Dec 2018

Section 3: Methods

3.1 Question 1

First, I loaded the .csv file into a Dataframe called df_reviews. Then, I extracted two columns from this Dataframe, which were column 'company' and column 'overall-ratings', and stored them in a new Dataframe called df_overall.

Fortunately, there were no missing or noisy data for this question. For data analysis, I grouped the Dataframe df_overall by company, calculated the average overall-rating score for each company and ranked their average scores. After this, I visualized the result by using histogram. I didn't encounter any challenges, since this is a basic question. But this question can give us an important overall view of the employee feelings.

3.2 Question 2

First, I extracted six columns from the df_reviews, which were overall-ratings and other five rating scores. Then, I dropped the rows where the rating scores were none.

After this, I encountered a challenge when I tried to create a heatmap to see the correlations between those variables, because the other five ratings scores were stored in string type. Thus, I used a function to convert those string type numbers to float type. Then, I successfully created a heatmap. Moreover, I conducted OLS Regression analysis to get more information of their relationships.

3.3 Question 3

First, I extracted all the pros, cons, and advice reviews for each company and stored them in different variables. After this, I dropped all the rows where the reviews were none. Then, I wanted to remove all the punctuations, non-alphabet characters and stopwords in those reviews, where I encountered the first challenge in this question. I couldn't remove them in Dataframe. Thus, I created a three-dimensional list to store all the company reviews and successfully removed those punctuations, non-alphabet characters and stopwords.

For data analysis, I used wordclouds to calculate and visualize the most frequently used words, when the employees wrote pros, cons and advice for their companies. However, after I saw the visualization results, I noticed that there are some unnecessary words, such as amazon, apple, google and company, etc. Thus, I added those words to stopwords, removed them and then created the wordcloulds of pons, cons and advice reviews for each company.

3.4 Question 4

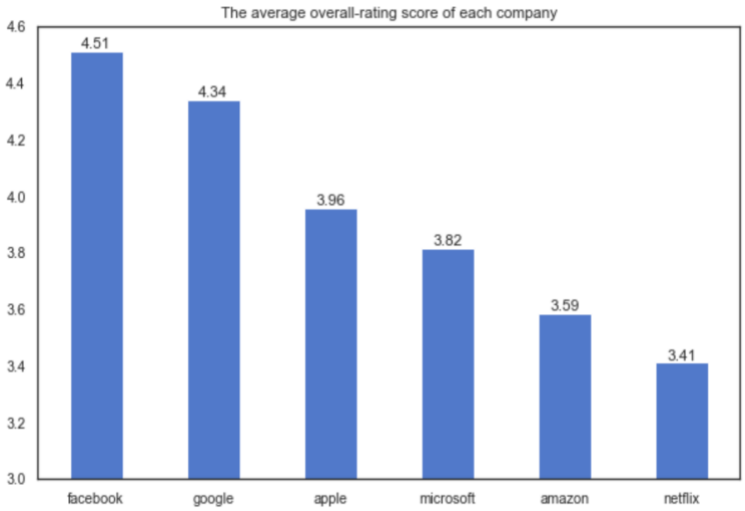
First, I extracted the overall-rating score and other five rating scores from df_reviews and stored them in a new Dataframe called df_reviews_predict. Then, I dropped all the rows where rating scores were none and converted the string type numbers in the other five rating scores to float type. Moreover, I spilt the df_reviews_predict dataset into training and testing datasets. I conducted a 70%-30% split.

Then, I needed to train my models by using the training sets and test my models by using the testing set. I used five different predictive methods to build my models, which are regression analysis, decision trees, random forests, Naïve Bayes and neural network. Moreover, for the Naïve Bayes method, I built four different models, which are GaussianNB, MultinomialNB, BernoulliNB and ComplementNB. Then, I calculated the model accuracy of each model and visualize the model accuracy results by creating a line chart.

I encountered a challenge when I calculated the model accuracy of the OLS regression model. The built-in function gave me a wrong answer. Thus, I wrote my own codes to calculate the model accuracy for this model.

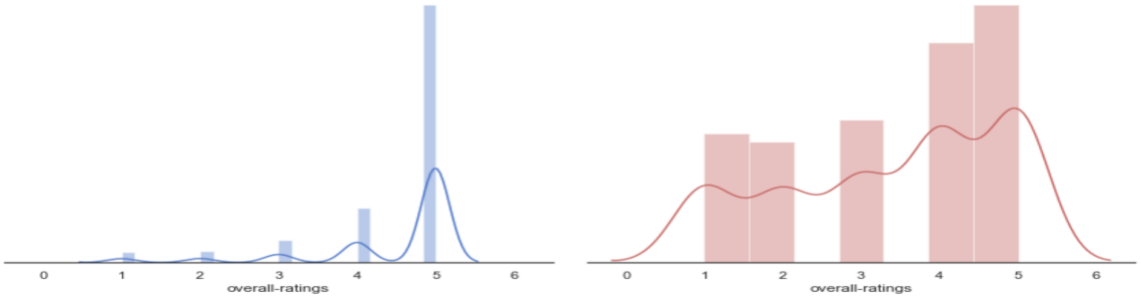
Section 4: Analysis and Results

4.1 Question 1



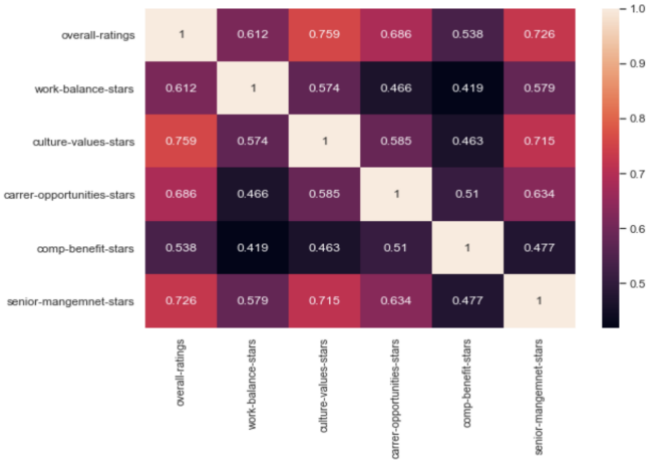
As we can see, Facebook has the highest average overall-rating score among these 6 companies, which was rated 4.51 out of 5. Moreover, Google also obtains a pretty high overall-rating score, which is 4.34. Only 2 companies were rated above 4 among these 6 companies.

However, Netflix received the lowest overall-rating score, which is 3.41 out of 5.



I also visualized the distributions of the overall-rating scores for Facebook and Netflix. As we know, Facebook obtains the highest average overall-rating score. Most of its employees rated Facebook 5 out of 5 and few of them rated Facebook 1 or 2. However, although a lot of employees of Netflix rated it 4 or 5, a large number of its employees rated it 1 or 2, which resulted in that Netflix received the lowest average overall-rating score.

4.2 Question 2



As the graph shows, the culture-values-stars has the strongest positive relationship with the overall-rating score. The senior-management-stars also obtains relatively strong relationship with the overall-rating score.

However, I'm a little bit surprised that the company-benefit-stars has the weakest positive relationship with the overall-rating score.

I run an OLS regression by taking overall-rating score as the dependent variable and other 5 rating scores as the independent variables. As we can see, the culture-values-stars variable has the largest coefficient, which is consistent with the results we got from the Heatmap above.

Moreover, the p-values of intercept and all the independent variables are 0, which means the results are statistical significant. The adjusted R-squared is 0.71, which means this model can explain around 71% all the variances.

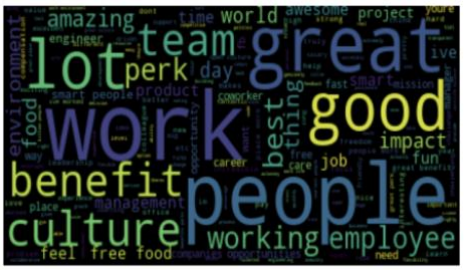
OLS Regression Results

Dep. Variable:	overall_ratings	**R-squared:**	0.713
Model:	OLS	**Adj. R-squared:**	0.713
Method:	Least Squares	**F-statistic:**	2.650e+04
Date:	Mon, 22 Apr 2019	**Prob (F-statistic):**	0.00
Time:	13:40:22	**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

4.3 Question 3

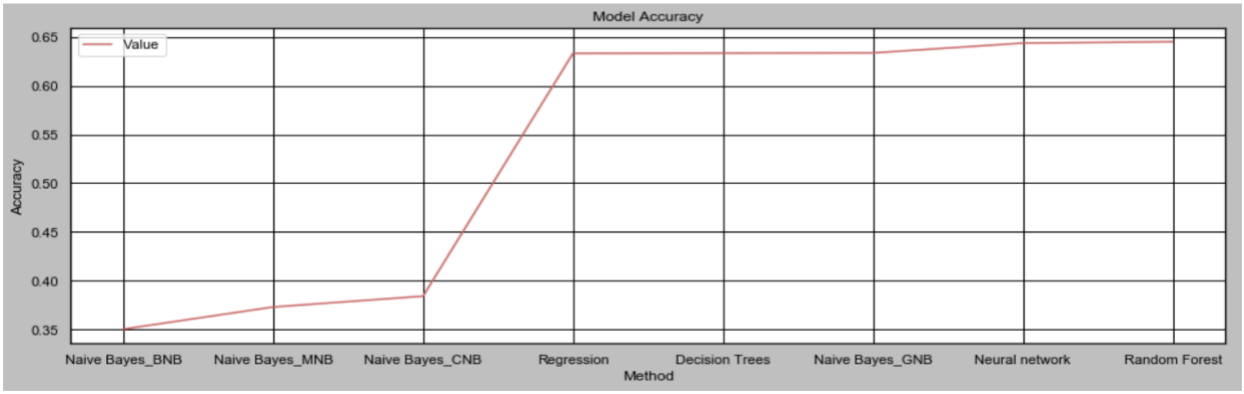


For illustration, I selected the wordcloud of the most frequently used words, when the employees of Facebook wrote pros to it.
As we can see, they used a lot of words like work, team, people, culture and benefit, etc. Thus, we can infer that they might compliment a lot about the team, people and culture in Facebook.

Moreover, I chose the worldcloud of Netflix, when the employees of Netflix wrote cons to this company. The most commonly used words are people, team, work, manager and management, etc. We can infer that the employees might complain a lot about the management issues or team work issues in Netflix.

In addition, I showed the worldcloud of Facebook, when its employees wrote advice to Facebook. They used a lot of words like employee, manager, culture and people, etc. It might because they wrote a lot of advice to their managers to ask their managers to treat them better.

4.4 Question 4



As we can see, I built eight different models to predict the overall-rating score by using other five rating scores. The three Naïve Bayes models are really not suitable for predicting the overall-rating score. However, the Random forest model and Neural network model did a pretty good job. Their model accuracies are around 0.65.