A Study on the Attributes of Companies, and their Effect on the Overall Glassdoor Company Rating

Erin M. Swan-Siegel

Northwest Missouri State University, Maryville MO 64468, USA
S201660@nwmissouri.edu

Abstract. As consumers, our lives are filled with reviews that aid us in deciding what products and services to purchase; five gold stars helping us make even the most minute of decisions. When we see similar items boasting similar ratings, it would be reasonable to conclude that the items are comparable in some way - sharing common characteristics. Applying it to employers present on the job-search platform, Glassdoor, this hypothesis is explored and tested. A strategic web-scraper was built to gather company attributes found on each employer's profile, accessed from a specific starting page. Once completed, supplemented, and cleaned, 3,829 records consisting of seven primary attributes were analyzed in an effort to determine the validity of the supposition. Training the data in Random Forest, Decision Trees, Gradient Boosting, Lasso Regression, and K-Nearest Neighbor classification models. After Hyperparameter tuning, the Random Forest model performed the best with an average error of 0.42. The maximum accuracy of the model is only 14.3%. This is either due to the smaller sample size, outliers within the data, or the complexity of the resulting Company, compared to the attribute data recorded on Glassdoor.

Keywords: business · employee satisfaction · company culture · company rating

1 Introduction

Every day, consumers are bombarded with products and businesses touting the product's utility or the benefits of their particular service through qualitative or quantitative summary statements; declarations such as "9 out of 10 doctors recommend", "4.5 stars with over twenty-thousand reviews!", and "Voted the best in the Midwest". We consumers rely on these ratings and reviews to help us decide on everything from where to eat, what product to buy, where to spend our vacation, and even where to work.

But what special mixture of characteristics contribute to a higher rating, when determined by both past and present employees - the longevity of the company; whether the company answers to shareholders or their cause; the size

of the employee base, perhaps? Furthermore, can one predict a company's rating based on key characteristics?

A Glassdoor company profile is created when the first self-reported past or present employee submits a review on the platform. Companies can create a free profile, allowing them to maintain certain aspects of the associated information card.[7] A company's overall rating is a calculated average of all review submissions.

Limitations to this study include the fact that a company's Glassdoor rating is not based on *every* past or present employee - just the subset of self-reported employees who decided to participate. These reviews can also be skewed / affected by employees who are new to the company, leave the company under stressful circumstances, or have recently been denied a promotion. Similarly, company data gathered is initially populated by a self-identifying employee and may or may not have been viewed or updated by an official company representative, the result of which could be incorrect attribute values.[7] In summary, the context and attributes of those contributing to the rating cannot be accounted for or controlled, and must therefore be taken with the common-sense understanding that many consumers do for other location, service, or product review page.

1.1 Associated Literature

Recruiting and retaining top talent is one of the major drives of any business that aims to be successful. Said another way, companies with the ability to keep their employees satisfied have a greater chance of performing well. This satisfaction - like many human experiences - is determined by a multitude of factors including the ability to advance, work-life balance, and the company's overall culture and values.[6],[5]

A company's ability to hit the common criteria for happy employees can be facilitated by higher employee base or increased revenues; for others, it can be a reason for hampering it.

1.2 Goals of this Study

The aim of this study is to predict a company's rating when a data Machine Learning model is provided four key characteristics.

1.3 Methodology

In order to answer the questions posed by this study, company data was scraped from the job platform, Glassdoor (GitHub Project Repository) The resulting data frame was further populated with company data scraped by other individuals analyzing Data Science job descriptions on the same platform.

Four key company features - Global Size, Company Ownership Type, Year Founded, and Estimated Yearly Revenue - were independent variables identified

as those potentially contributing to the company's rating. Records missing one or multiple of the key attributes were removed as one part of the data cleaning process.

2 Data

Data for this study was intended to be gathered solely by the author's web-scraper, built together with Software Engineer, Daniel Swan. Using code from a Kaggle contributor as a base. The resulting scraper extracts company data from an employer card found on a particular Glassdoor site, clicks into each company's Glassdoor employer page and extracts additional data. Unfortunately for this study, platforms are continuing to build better protections over their systems and their data, and the web scraper built for this study - while it does work as expected - encountered scraping prevention measures.

Web scraping prevention exists for many reasons, including blocking excess traffic that slows down the site for non-bot users and for protecting a company's intellectual property[8]. Options to bypass these security measures are accessible, but required significant re-work of the author's script or a research partnership with the Glassdoor platform itself. With a similar output resource available, company data published to Kaggle in 2020 was used to supplement the web-scraped data.

2.1 Cleaning

Data in the tenth field of the web-scraped output file was parsed with the Text-to-Columns feature of MicroSoft Excel, creating eight additional features using the delimiter specified in the web-scraper. Before the generated and the procured data were joined, the Kaggle data set was further processed for this study by removing unnecessary fields related to the job listings, cleaning the rating from the company name, and using the "Remove Duplicates" function to reduce the records to only those with unique attribute values. Using the approximately 8000 company name values, a V-lookup process against a copy of the original Kaggle data allowed for the first respective Job Description to become a substitute for the "Company Description" attribute.

Further cleaning was accomplished through a free Visual Studio plug-in, Data Wrangler (DW), and Jupyter lab. Opening a tabular data file with DW allows a user to modify their data as they would in a sheets program like MicroSoft Excel, while providing the scripting needed to make the same changes with Python code, including find and replace, computations, and string extractions. Prior to its removal from the data set, the original values for the field "Headquarters" were split into City and State (or Country) through the use of DW, and the two records whose State value was "61" was substituted with the correct value of "NY". The application was further used to substitute numeric values for category values found in non-numeric fields - Company Ownership Type, Est Yearly

Revenue, Global Size, Headquarters (State/Country), and Industry. The distinct values from each field were ordered alphabetically or by size, and assigned a unique numeric value in Excel. These values then replaced the string values, as detailed in the following subsection. The new field "Company Age" was created by subtracting the value found in "Year Founded" from the current year (2023). The resulting script was applied to the data set through Jupyter Labs.

During the Data Exploration phase, the attributes "Headquarters" and "Industry" proved out to have nearly no correlation (r = ¡ 0.05) with the test variable, Company Rating, and were removed from the resulting set of attributes utilizing Python. While the supplemental Kaggle data set did not contain any NULL/missing values, most fields contained some "-1" and "Unknown" values. Using a python file, records containing one or more default values were not considered for further analysis, including the Company Age value "2024", which resulted from a previous cleaning step. The final data set contained 3,829 clean records.

The values found in non-numeric fields were replaced, utilizing the unique numeric values assigned (described in the previous passage and detailed below). This step allowed the attribute data to be compared in more meaningful ways during Data Exploration. The Python notebook used in this phase is available at `https://github.com/progswan2022/scraping-glassdoor-selenium`.

Resulting Data Landscape ATTRIBUTES

Attribute	**Description**
Company Name	Name of the company reviewed
Global Size	Number of employees across all of the companys' locations, as reported by current or past employees who leave a review on the platform
Company Rating	The average rating, reported on a scale of 1 to 5, as determined by current and past employees who leave a review on the platform
Company Ownership Type	Indicates the company's designation as a privately-held, publicly-held, non-profit, educational institute, etc
Company Age	The year the company was reported to have been founded, subtracted from the current year (2023)
Est Yearly Revenue	Estimated yearly revenue of the company, as reported by current or past employees who submit a review on the platform

VALUE CONVERSION FOR NON-NUMERIC FIELDS

• Global Size

Global Size	**Order number**
1 to 50 employees	1
51 to 200 employees	2
201 to 500 employees	3
501 to 1000 employees	4
1001 to 5000 employees	5
5001 to 10000 employees	6
10000+ Employees	7

- Company Ownership Type

Company Ownership Type	**Category Number**
College / University	1
Company - Private	2
Company - Public	3
Contract	4
Franchise	5
Government	6
Hospital	7
Nonprofit Organization	8
Other Organization	9
Private Practice / Firm	10
School / School District	11
Self-employed	12

- Est Yearly Revenue

Est Revenue	**Order Number**
Less than $1 million (USD)	1
$1 to $5 million (USD)	2
$5 to $10 million (USD)	3
$10 to $25 million (USD)	4
$25 to $50 million (USD)	5
$50 to $100 million (USD)	6
$100 to $500 million (USD)	7
$500 million to $1 billion (USD)	8
$1 to $2 billion (USD)	9
$2 to $5 billion (USD)	10
$5 to $10 billion (USD)	11
$10+ billion (USD)	12

2.2 Exploration

Following the guidelines and advice of YouTuber Ken Jee[10], in conjuncture with the Visual Studio plugin - Data Wrangler - the Company Rating data set was passed through a series of tests / analysis. Conclusions drawn during the Data Exploration phase resulted in additional data cleaning steps be performed (documented in previous subsection) including the conversion of values in the categorical fields and the removal of the fields "Headquarters" and "Industry".

The Data Exploration piece began with a simple field correlation matrix, which showed a nearly zero relationship between Company Rating and Industry, and Company Rating and Headquarters (State/Country). A moderate negative relationship between Company Rating and Global Size, as well as Company Rating and Est Yearly Revenue, were noted. Strong correlations were observed between revenue, size, and age. This allows the conclusion that the older the company, the longer they have had to grow in profitability and employee resources. (See image: Correlation Matrix; Heat Map)

An iterative approach to displaying charts through the Seaborn library was created to define and build a series of bar plots. The script is included below. Through the visualizations, it can be seen that the dependent variable, Company Rating, as a fairly normal distribution, centered on a 3.63 average rating, across all companies represented. Companies reported as Privately or Publicly owned make up 3200 of the 3829 companies (approx 84%). An exponential distribution was observed for Company Age, with a mean of 48 years and a max of 355. (See image: Company Age Distribution Plot)

```
Python code for the definition and creation of Bar Plots:
    # Define the fields to be graphed
    df_col = df2[['Company Rating','Global Size','Est Yearly
    Revenue','Company Ownership Type','Company Age']]
    # Create a for-loop
    for i in df_col.columns:
        col_num = df_col[i].value_counts()
        # Include a graph title
        print("Visualization for %s: Total = %d" % (i,
        len(col_num)))
        # Define the values of x- and y-
        chart = sns.barplot(x=col_num.index, y=col_num, data=df)
        # Rotate the x-value labels for easier reading
        chart.set_xticklabels(chart.get_xticklabels(), rotation=90)
        # Display the graph
        plt.show()
```

Box plots quickly provide lots of information about the data being plotted, including the mean, the 25% and 75% quartiles, if the data is tightly or loosely grouped, and if there are outlier values that may need to be removed.[2] Three of the Company attributes had data values that can be considered outliers - existing above or below the whisker ranges. The companies who ages are above

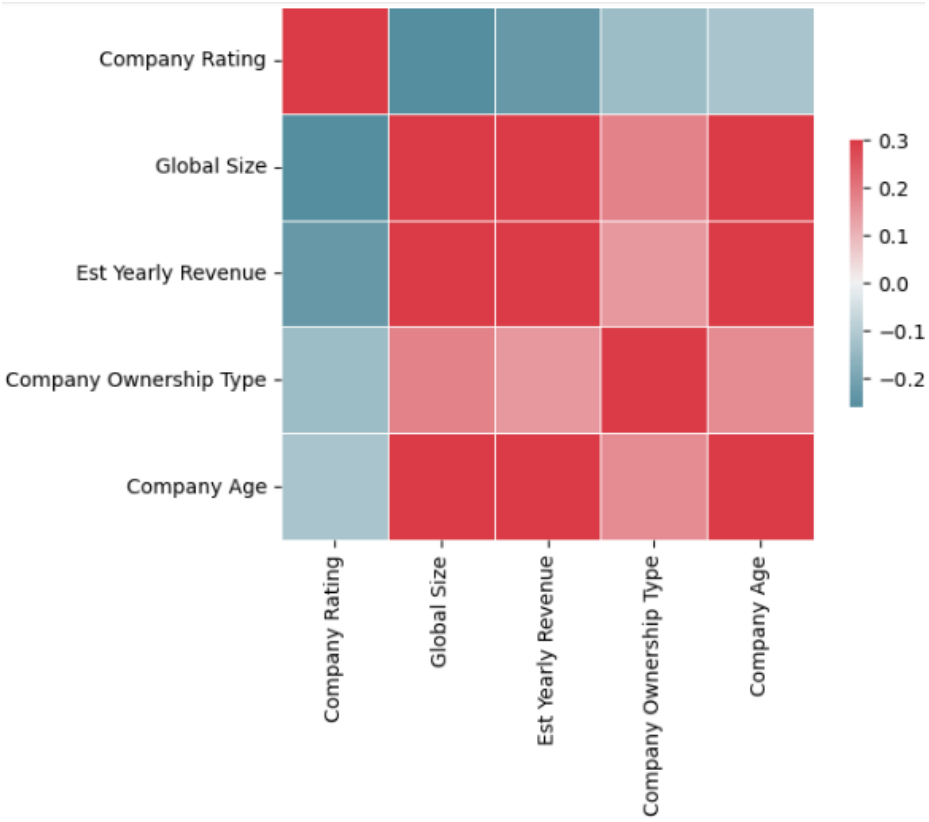

Fig. 1. Correlation Matrix; Heat Map

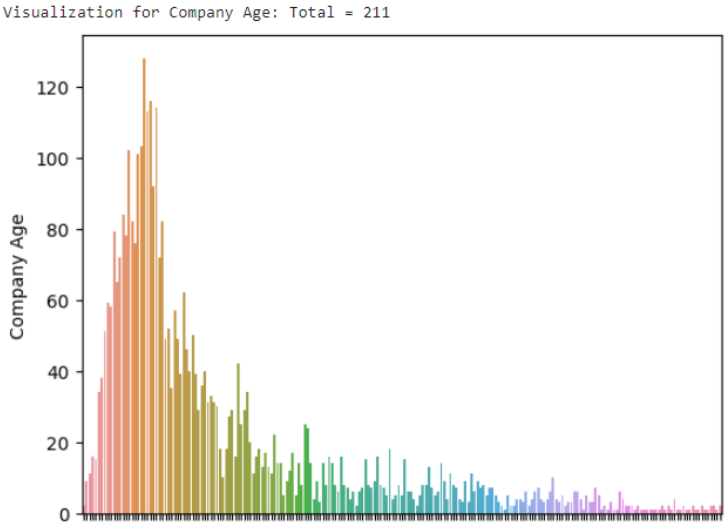

Fig. 2. Exponential Distribution of Company Age

approx 135 years, are of an Ownership Type other than Private or Public, or have a Company Rating lower than about 2.3, fall outside of the lower and upper data ranges.

3 Model Selection

The data for this project is categorical and has a dependent variable attempting to be predicted. Consequently, the use of Supervised Machine Learning is the most appropriate for this study. Supervised learning is the process of teaching a machine to predict a value, based on an input of one or more attributes. When considering Supervised Learning models, there is a wide array to choose from – Neural Networks, Naïve Bayes, Linear Regression, Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbor, and Random Forest.[9]

The data set for this project is categorical / characteristic data, lending itself better to Random Forest (including Decision Tree), Logistic Regression (including Lasso Regression and Gradient Boosting), and K-Nearest Neighbor models.

3.1 Building

In order to convert the data into values that can be utilized for numeric modeling, it first had to be converted to dummy data. This was something previously attempted using a system of manual conversion by the author.[10] The use of the function "get_dummies" allows the values of categorical data to be broken into a series of indicators. The data was then broken into a test and training

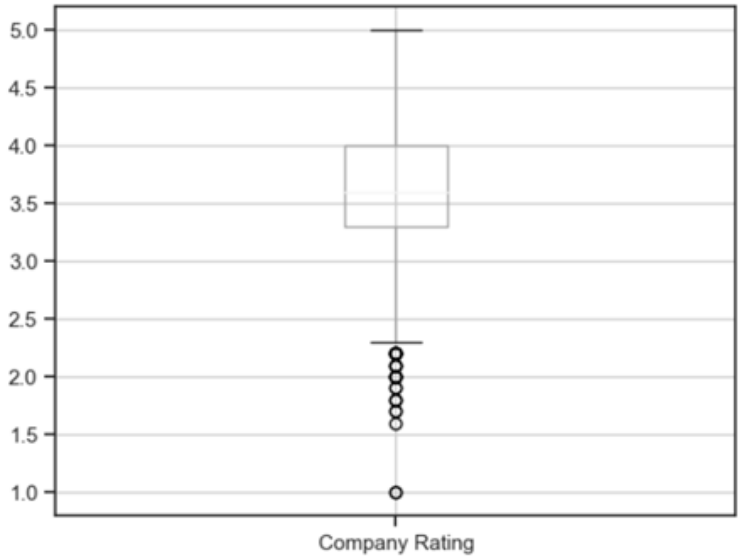

Fig. 3. Box-and-Whisker Plot of Company Rating shows outliers less than 2.3

set (20% and 80%, respectively). The training and testing data sets were run through several models individually, then SKLearn's GridSearchCV module was used to identify the best parameters (Hyperparameter Tuning). The performance of combinations of multiple models were also explored.

There are a number of statistical measures which help to inform the accuracy of a model by analyzing the variations between the predicted values and the original values. The most recognizable being the measure of linear correlation, the R-squared value. One analysis technique is to calculate the average of the absolute error of the predicted values (Mean Absolute Error). Further model assessment is achieved by squaring each absolute variation (Mean Squared Error), then taking the square-root of that value (Root Mean Squared Error). Using the GridSearchCV module, the best parameters for the models were determined.

Random Forest Best Parameters:

```
max_depth: 10, min_samples_split: 10, n_estimators: 300
Random Forest Best MAE: 0.4243
Random Forest Best MSE: 0.2830
Random Forest Best RMSE: 0.5320
Random Forest Best R-squared: 0.1426
```

Gradient Boosting Best Parameters:

```
learning_rate: 0.01, max_depth: 3, n_estimators: 300
Gradient Boosting Best MAE: 0.4221
Gradient Boosting Best MSE: 0.2892
```

```
Gradient Boosting Best RMSE: 0.5378
Gradient Boosting Best R-squared: 0.1239
```

Decision Tree Best Parameters:

```
max_depth: 10, min_samples_leaf: 2, min_samples_split: 10
Decision Tree Best MAE: 0.4400
Decision Tree Best MSE: 0.3066
Decision Tree Best RMSE: 0.5538
Decision Tree Best R-squared: 0.0711
```

K-Nearest Neighbors Best Parameters:

```
algorithm: auto, n_neighbors: 7, weights: uniform
K-Nearest Neighbors Best MAE: 0.4432
K-Nearest Neighbors Best MSE: 0.3016
K-Nearest Neighbors Best RMSE: 0.5492
K-Nearest Neighbors Best R-squared: 0.0864
```

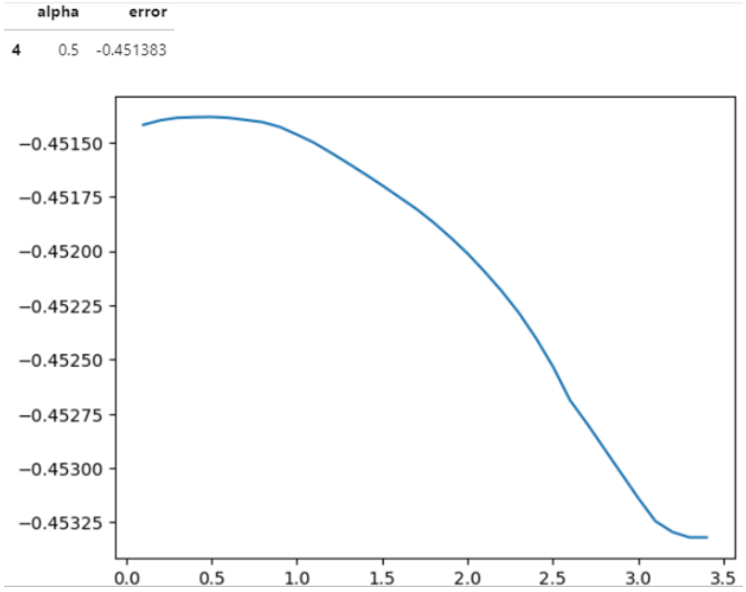

Fig. 4. Graph of the Lasso Regression Model with Hyperparameter Tuning

4 Results

The following readouts are the statistical results of the models with their respective optimal parameter values.

Random Forest Results

```
Training R-squared: 0.3111, Testing R-squared: 0.1426
Training MAE: 0.3777, Testing MAE: 0.4145
Training MSE: 0.2276, Testing MSE: 0.2830
Training RMSE: 0.4771, Testing RMSE: 0.5320
```

Gradient Boosting

```
Training R-squared: 0.1623, Testing R-squared: 0.1239
Training MAE: 0.4149, Testing MAE: 0.4184
Training MSE: 0.2768, Testing MSE: 0.2892
Training RMSE: 0.5261, Testing RMSE: 0.5378
```

Decision Tree

```
Training R-squared: 0.2556, Testing R-squared: 0.0711
Training MAE: 0.3872, Testing MAE: 0.4288
Training MSE: 0.2459, Testing MSE: 0.3066
Training RMSE: 0.4959, Testing RMSE: 0.5538
```

K-Nearest Neighbor

```
Training R-squared: 0.2792, Testing R-squared: 0.0864
Training MAE: 0.3822, Testing MAE: 0.4256
Training MSE: 0.2382, Testing MSE: 0.3016
Training RMSE: 0.4880, Testing RMSE: 0.5492
```

Lasso

```
Training R-squared: 0.0145, Testing R-squared: 0.0137
Training MAE: 0.4512, Testing MAE: 0.4425
Training MSE: 0.3256, Testing MSE: 0.3256
Training RMSE: 0.5706, Testing RMSE: 0.5706
```

For linear models, the Coefficient of Determination (R-squared) value represents the percentage of the dependent variable can be explained by the model. The Mean Absolute Error (MAE) is a great way to assess the overall error between the model and the actual values, but could mask pockets of extreme deviation. The Mean Squared Error (MSE) provides another perspective by exacerbating errors – the higher the value, the more significant or numerous the outliers. If the MAE is higher than the MSE, it suggests that the model has errors that have accumulate due to "randomness, variability, or imprecision in the data." [1] The Root Mean Squared Error (RMSE) effectively calculates the standard deviation of the errors. [4]

5 Conclusion

Based on the models tested, Random Forest best parameter fit had the most success with a 14.3% determination and an average absolute error of 0.42 Company Rating points.

5.1 Discussion

Many of the models performed similarly; there isn't a single model that strongly fits the data set, which speaks to the variability in the strength of the attribute relationship to the Company Rating. The larger number of degrees of freedom created a weaker model and more difficulty in estimating the coefficients of the algorithm. [3] Different approaches and techniques that could be taken are discussed in Future Exploration.

5.2 Summary

Utilizing largely supplemental company data from the job-search platform, Glassdoor, company attributes - including global employee size and annual estimated revenue - were evaluated for their relationship to the associated company's rating. The hyperparameter tuning of categorical supervised learning models led to a Random Forest model performing with 14.3% accuracy.

5.3 Limitations

There were three major restrictions encountered during the course of this study. The most hindering and unexpected was the anti-scraping measure encountered during data collection of Glassdoor Company data, which consequently required the use of supplement data. The additive data available was acquired as a part of analysis involving job listings and not all desired attributes - such as number of reviews and the company description, were present. The absence of expected data reduced the available attributes that could have potentially strengthened the statistical models, and removed the ability to perform a clean language analysis of company descriptions with relation to the associated rating. Another limitation was the sample size. Using stagnant supplemental data - with no ability to gather additional data - left only 3,829 clean data records. More data may have led to stronger models. An additional hindrance were the inconsistent categorical attribute values. Continuous data or normalized interval counts for attributes - such as global employee size - would have greatly improved the ability to create better statistical models.

5.4 Future Exploration

Alterations to the data source, the ability to by-pass Glassdoor's anti-scraping measures, and narrowing the scope of the analysis are possible future steps. Glassdoor company data includes attributes with categorical values that are not of a consistent interval, making analysis that much more difficult. Finding a company rating data source with attribute data that is more suited to analysis would potentially lead to more promising results. Implementing a technique which would allow for more data to be gathered from the site, without disrupting the data scraping, would provide a larger data set. Based on the box plots and distribution graphs created during Data Exploration, narrowing the scope of the study to only companies that are privately-owned with a rating greater than 2.5 would likely yield cleaner outcomes and more meaningful results.

References

1. AI, L., Community: What is the difference between mean squared error and mean absolute error in machine learning metrics?, https://www.linkedin.com/advice/0/what-difference-between-mean-squared-error-tz1mc
2. Catalogue, T.D.V.: Box and whisker plots, https://datavizcatalogue.com/methods/box_plot.html
3. Community, S.E.: Random forest underfitting, https://stats.stackexchange.com/questions/151556/random-forest-underfitting
4. Frost, J.: Root mean squared error, https://statisticsbyjim.com/regression/root-mean-square-error-rmse/
5. Glassdoor: Glassdoor research: Satisfaction drivers remain intact, https://www.glassdoor.com/research/satisfaction-drivers-remain-intact
6. Glassdoor: Glassdoor research: Satisfied workers stay, https://www.glassdoor.com/research/satisfied-workers-stay
7. Glassdoor: How did my company get on glassdoor?, https://stage-help.glassdoor.com
8. GPT, C.: Why websites prevent data scraping, https://chat.openai.com/
9. IBM: What is supervised learning?, https://www.ibm.com/topics/supervised-learning
10. Jee, K.: Data science project from scratch, https://www.youtube.com/@KenJee_ds