[1]

Predicting Glassdoor Workplace Factor Ratings from Reviewer Profession

Group 4: David Olsen (19714, DAOL@itu.dk), Gustav Jandrup (19750, GUSJ@itu.dk), Klara Jensen (19784, KLPJ@itu.dk)
Course code: KSDAMIN1KU

***Abstract*— Glassdoor is a popular platform for employees to leave anonymous reviews and ratings of their employer and can be helpful for job seekers and others who are considering working for a particular company. Glassdoor ratings may have a significant impact on a company's reputation and success. This study explores whether a Glassdoor reviewer's profession can predict whether a rating is high or low using a large (>100k) dataset collected from Kaggle and converted to binary with One-Hot Encoding. We use an unsupervised learning and a supervised learning approach. The unsupervised approach combines *k*-Modes Clustering with Frequent Pattern Mining while the Supervised learning approach picks and tunes a Logistic Regression Classifier as the best supervised model. We find reviewers with job titles belonging to the category "Service Professional" can accurately predict an overall low satisfaction with most rating parameters, while 'Technology Professional' can accurately predict an overall low satisfaction. Finally, we find that 'Business Professional' from the Consulting industry can predict high satisfaction with career opportunities and low satisfaction with work life balance. The results raise some questions towards an unfortunate tendency for service professionals to be unsatisfied with their workplace and how organisastions may consider if part of their workforce are systemically dissatisfied.**

***Keyword*—association mining, classification, clustering, company culture, data mining, data preprocessing, employee reviews, Glassdoor, machine learning.**

I. INTRODUCTION

I

n today's business environment, online reviews and ratings can have a significant impact on a company's reputation and success, as they provide a platform for customers and employees to share their experiences and opinions about the company. An example of such a platform is Glassdoor, which is a popular platform for employees to leave anonymous reviews and ratings of their employer, and can be helpful for job seekers and others who are considering working for a particular company [1][2]. A positive work environment has several benefits, e.g., productivity, improvement of employee competencies and collaboration. Thus, improving the factors that result in a positive work environment can lead to these benefits, while improving access to talent [3][4]. In this study, using unsupervised and supervised machine learning methods, we seek to explore the following research questions: a) How does a reviewer's job title on Glassdoor affect their evaluation of a company, and what is the best fitted model to predict such evaluation? b) Is it possible to accurately predict job satisfaction based on broad job categories?

II. METHODOLOGY

A. Data Collection, Preparation and Preprocessing

A Glassdoor review dataset collected from Kaggle was used for this study. The original source of data is Glassdoor, but the dataset was made available by a user named David Gauthier [5]. The dataset contains over 800,000 reviews for a variety of companies, as well as information about the reviewers' job titles, location, tenure, and review date. It includes several ratings provided by the reviewers based on a scale from 1 to 5 where 1 represents "Very Dissatisfied" and 5 represents "Very Satisfied" [6]. These ratings are 'Overall Rating', 'Work Life Balance', 'Culture and Values', 'Diversity and Inclusion', 'Career Opportunities', 'Compensation and Benefits' and 'Senior Management', and enables the reviewer to provide more detailed feedback on specific aspects of their work experience.

We are interested in understanding the relationship between a reviewer's job title and their evaluation of a company. To do this, we excluded reviews with no job title and those that were made by anonymous employees as the scope of our model is to use job titles as features as the scope of our model is to use job titles as features. Also, the distribution of ratings was skewed towards higher ratings (see code for descriptive statistics). Specifically, the number of reviews with ratings of 5 was 232,255, with 4 it was 278,277, with 3 it was 194,267, with 2 it was 74,809, and with 1 it was 58,958. Since the majority of the reviews have given an overall rating of 4 or 5, we decided to transform all ratings to binary, where 0 represents ratings of 1 to 3, and 1 represents ratings of 4 and 5. This was done as an imbalanced dataset can be a problem in machine learning because the model may become biased towards the more prevalent class, leading to poorer performance on the less prevalent class [7].

We further prepared the dataset by splitting the 'current' column into a column with data on the reviewer's current employee status and a column with their tenure. The current employee status column was converted into binary form, and the tenure column's text data was first replaced with integers, and later, encoded with Panda's get_dummies function [8]. The column with ratings on diversity and inclusion was dropped, as it held 702,500 NaN values, corresponding to around 85 percent of the entire dataset. Lastly, to avoid having a potentially infinite number of dimensions in our features, we decided to only include the twenty most frequent job titles in the dataset which represented 145,063 reviews.

As there are potentially close to potentially close to infinite job titles, we decided to create yet another dataset that instead of the twenty most frequent job titles held four different job type labels. These labels are 'job_Service Professional',

[1].

'job_Management Professional', 'job_Business Professional' and 'job_Technology Professional' and are derived from the twenty most frequent job titles in the former dataset. This was done to test the predictive power of labeled data and to avoid building an algorithm that relies on ever-increasing dimensionality as new job titles are introduced or an algorithm that only can predict the twenty most frequent job titles in the dataset. To ensure that the labels had a high accuracy, the modeling and evaluation process was run twice: one time with the job titles dataset as input variable and one time with the labels dataset.

B. Unsupervised Learning Methods

We applied unsupervised machine learning with the purpose of first detecting any clusters of reviews based on the combination of high/low reviews across the variables 'Work Life Balance', 'Culture and Values', 'Career Opportunities', 'Compensation and Benefits' and 'Senior Management'. The intention was to detect patterns of high/low job satisfaction across multiple parameters. We then applied frequent pattern mining to detect frequency of association rules of antecedent itemsets containing job titles or labels and consequent itemsets containing a specific cluster [9].

1) k-Modes Clustering

To detect clusters of reviews, we used Vos' [10] implementation of *k*-Modes for clustering the reviews. *k*-Modes is a distance-based partitioning method that, given a set of *n* objects, constructs *k* partitions of the data using an iterative relocation approach. *k*-Modes uses initialization methods to define the centroid of each cluster and iteratively partitions the dataset and selects the mode of each cluster until the centroids converge, and the mode does not change over the next iteration [9].

Since the dataset consists of binary variables and the means cannot be calculated meaningfully, we used *k*-Modes as an alternative to *k*-Means. To initialize our centroids, we used Cao et al's [11] density-based initialization method. Cao et al's [11] density-based initialization for *k*-Modes is an extension of the standard *k*-Modes algorithm that attempts to select more representative initial centroids by taking into account the density of the data points in the feature space. Specifically, it uses a density-based clustering algorithm to identify high-density regions in the feature space, and then selects the centroids from these high-density regions. The idea is that these centroids are more likely to be representative of the data points in the cluster and will therefore produce better clusters [11]. The sum of squares error, also known as 'inertia', was applied as an objective function for measuring within-cluster variation [9]. We used the elbow-curve method by plotting inertia for each *k* in a range to pick the best *k* value [12].

2) Frequent Pattern Mining

To detect association rules between job titles and clusters and labels and clusters, we used the mlextend frequent pattern library implementation of apriori and association rules algorithms [13]. The apriori algorithm detects frequent item sets above a defined 'support' threshold measuring the relative frequency of the item set. The association rules algorithm detects patterns between itemsets above a defined 'confidence' threshold measuring the conditional probability of the antecedent and consequent itemsets given the support of the antecedent itemset. We also used the lift metric to filter negatively correlated or independent consequents and antecedents [9].

As a result of having a large dataset with more than 100,000 rows, we set a low support threshold of just 0.5 percent for the apriori algorithm. This was done as a result of multiple iterations showing frequent patterns between the rating parameters, possibly as a result of the overall probability distribution where high ratings in one parameter tends to lead to high ratings in other parameters and vice versa. For the association rules, we defined a confidence threshold of 60 percent, assuming that a confidence score close to 50 percent or below would either be too random or improbable. We set a lift threshold of 1.2 to ensure the association rules we found had a positive correlation, e.g, if job_title 'Cashier' as part of an antecedent item set would be correlated with consequent itemsets containing cluster 0. Finally, we applied apriori and frequent pattern mining to a dataset consisting of just the job titles and clusters, excluding all other variables.

C. Supervised Learning Methods

1) Selecting the Best Model

A supervised machine learning algorithm was created to predict ratings on job titles or labels. To find the best performing algorithm, we introduced a collection of models to test both the job titles dataset and the labels dataset. For both datasets, each model was trained to predict the variables of 'Work Life Balance', 'Culture and Values', 'Career Opportunities', 'Compensation and Benefits' and 'Senior Management', and each model was evaluated by the aggregate score of the output variables.

The model collection held a naive classification model, linear, nonlinear and an ensemble model. For the naive classification model, we chose the Prior model that predicts the prior probability for each class, or simply put, the probability of the class occurring in the training data. This model was trained to get a benchmark on which the other models could be evaluated [9]. For linear algorithms, we trained a Logistic Regression (LR), Linear Discriminant Analysis (LDA), and Bernoulli Naive Bayes (BNB) model. For nonlinear algorithms, we trained a Decision Tree (DTC) and *k*-Nearest Neighbors (KNN), and finally, for ensemble algorithms, we trained a Random Forest (RF) classifier [9].

2) Cross-Validation of Accuracy Scores

To evaluate the models, we splitted the datasets into a training set containing 70 percent of the dataset, and the remaining 30 percent into a test set. The model collection was only trained on the training set, and the final logistic regression model was trained on the training set and tested on the test set. To calculate the accuracy score of each model, we deployed the cross-validation method, 10-*k* Fold. Cross-validation is a resampling method used to evaluate a

model's performance by dividing the training data into a set of folds (*k)*, training the model on different combinations of these folds (*k-1)*, and evaluating the model's accuracy score based on the remaining fold [9].

3) Evaluation Framework

To evaluate the different models, we decided on three evaluation criteria: a) accuracy of model, b) the model's tendency to overfit, and c) model interpretability. In evaluating the collection of models, the aim was to choose the model with the highest accuracy that is least prone to overfitting and can be interpreted. The accuracy score is a metric that evaluates the number of correct predictions made. This is done by taking the number of correct predictions made divided by the number of total predictions [14]. The accuracy score should not be used as a metric for imbalanced classification problems, but after converting the dataset to binary form, the datasets are more balanced.

To reduce overfitting, we considered the bias-variance trade-off [15]. The bias-variance trade-off refers to the notion that it is generally easy to train a model with high bias and low variance or high variance and low bias, leading to a trade-off between both options. Thus, for our final choice of model we wanted to avoid picking a model with tendencies to overfit as a result of having a relatively large amount of features. Parametric models tend to have more bias while non-parametric models tend to have higher variance. Thus, some models, such as DTC or KNN might be prone to overfitting on the training data. Parametric classification models, such as LR, have a number of assumptions about the structure of the decision boundary and the underlying distribution of the data. These assumptions allow parametric models to be simpler and faster to train but may also make them less flexible and less accurate on datasets that do not fit the assumptions of the model. As a result, parametric models tend to be more interpretable and have more bias [15]. On the other hand, non-parametric classification models, such as KNN and RF, do not make any assumptions about the distribution of the data. This lack of assumptions allows non-parametric models to be more flexible and potentially more accurate, but at the cost of computational performance [15]. Non-parametric models tend to have higher variance and may be more prone to overfitting on the training data.

Interpretability is about if humans can determine how the algorithm has arrived at a given prediction. Differently put, it is about how well the algorithm can be understood and explained by humans. Linear models and DTC are defined as interpretable models [16].

4) Logistic Regression Classification and Hyperparameter Tuning

The chosen model was a LR. Despite its name, logistic regression is a classification model used to predict the probability of a target variable of binary form [17][18]. The logistic regression model uses the S-shaped standard logistic function, as illustrated in formula A, to model the relationship between the predictor variables and target variable.

In a logistic regression model, the coefficients β_0 and β_1 are used to represent the relationship between the predictor variables and the target variable. The coefficients are estimated during the model training process, and they are used to make predictions on new data. For each predictor variable in the model, there is a corresponding coefficient that represents the strength and direction of the relationship between the predictor and the response. A positive coefficient indicates that an increase in the predictor is associated with an increase in the probability of the response, while a negative coefficient indicates that an increase in the predictor is associated with a decrease in the probability of the response. The coefficients are typically found using maximum likelihood estimation, a method that finds the coefficients that maximize the likelihood of the observed data according to the model. To find the best fitted model, the maximum likelihood is calculated for each fitted model [17][18].

Fig. 1. Standard Logistic Function [15, p. 134]

To improve the final model's performance, we applied the Scikit-Learn function, RandomizedSearchCV [19]. Random search is an optimization method that randomly combines different hyperparameters within a search space. A search space is a defined range of possible values for the hyperparameters, and randomly sampling points within this range allows for the exploration of different combinations of these values [20].

III. RESULTS

A. k-Modes Clustering and Frequent Pattern Mining

Applying *k*-Modes to each rating feature and plotting the sum of squares error where $k = 1,...,9$ produces the elbow curve seen in figure 2. The curve shows a sharp decline from $k = 1$ to $k = 2$ and a diminishing decline from $k = 3$ to $k = 9$ indicating that the "elbow" is around 3.

As seen in the attached Jupyter Notebook file (code file), cluster 0 primarily consists of reviews in which all ratings are 3 or below with only relatively few being above 3. Cluster 0 is the largest cluster measured absolutely, as it contains approximately 80k rows, which is more than 50 percent of the total dataset. Cluster 1 primarily consists of reviews in which all ratings are more than 3 and relatively few at 3 or below. This cluster is the second largest cluster, covering approximately 50,000 rows. Finally, cluster 2 consists of reviews that rated 'Work Life Balance', 'Senior Management' and 'Compensation and Benefits' at 3 or below with 'Career Opportunities' and 'Culture and Values' being rated above 3.

As seen in the code file, applying the apriori algorithm and association rules algorithm to the entire dataset, that includes the clusters, demonstrates some patterns regarding job satisfaction.

The five most frequent job titles included in the antecedent itemset to cluster 0 as a consequent are 'Crew Member', 'Cashier', 'Sales Assistant', 'Manager' and 'Consultant'. The five most frequent job titles included in the antecedent itemset

to cluster 1 as a consequent are ‘Software Engineer’, ‘Consultant’, ‘Manager’ and ‘Senior Software Engineer’. The five most frequent job titles included in the antecedent itemset to cluster 2 as a consequent are ‘Senior Associate’, ‘Consultant’, ‘Senior Consultant’, ‘Manager’ and ‘Associate’. The results indicate that employees working in customer-facing service jobs as ‘Crew Members’ or ‘Cashiers' tend to predominantly rate between 1 and 3 while employees working in ‘Software Engineering’ tend to predominantly rate from 4 to 5. Finally, consulting-related job titles are distributed relatively evenly across cluster 0 and 1 while cluster 2 almost exclusively consists of consulting type job titles, which indicates that these reviewers tend to rate ‘Career Opportunities’ and ‘Culture and Values’ highly while being mostly dissatisfied with all other parameters.

The result of running apriori and association rules on the dataset exclusively with job titles and clusters returns three rules as illustrated in figure 3. All association rules have low support values, which was expected due to the low antecedent support. However, confidence scores are relatively high (close to 70 percent or higher) and lift values are above threshold.

It is not possible to discern clear patterns from association rules describing the labeled dataset. This is because the antecedent itemsets containing each job label are relatively evenly spread out, except for the itemsets predicting cluster 2 in which ‘job_Business Professional’ are more frequent in each antecedent compared with other categories.

As illustrated in figure 4, running apriori and association rules algorithm using the dataset consisting exclusively of the job labels and the clusters, detects a stronger association rule

Fig. 3. Association Rules Predicting Cluster 0

than running the algorithms with all job titles. ‘job_Service professional’ predicts cluster 0 with a relatively high support value of 15 percent and confidence score at 70 percent, this is a significant increase in support compared with the three association rules discovered without labels where the highest support value was 4 percent, the support for the labeled antecedent are 15 percent.

Arguably, from applying clustering and association rules there are no clear and substantive patterns predicting high rating, but there is a clear pattern of ‘job_Service Professional’ being associated with ratings at 3 or below.

B. Evaluation of Supervised Models

Figure 5 illustrates a boxplot of the model collection with each model’s accuracy score on the dataset with labels and the dataset with job titles. To select the final model, we evaluated the models on three criteria: accuracy, tendency to overfit and interpretability. The Prior for both datasets had a median of approximately 0.54. First, we excluded the KNN model as it had the lowest accuracy score compared to the other models.

Also, the KNN model performed worse than the Prior, which was used as the benchmark. Then, as the accuracy scores between the remaining models were close, we excluded based on tendency to overfit. Here, the RF model and DTC model model were excluded because of their non-parametric model type. The remaining models were LR. LDA and BNB. From these three, we selected the LR model, as it enables one to see probability coefficients which can provide insights into feature importance later on.

Fig. 4. Association Rule Predicting Cluster 0

C. Tuned Logistic Regression Classifier

1) Hyperparameter Tuning

We applied RandomizedSearchCV to search for hyperparameters for the LR [18]. For the RandomizedSearch parameter grid the following hyperparameters were included to allow for a broad range of configurations. For both datasets, we included solver, penalty type, and C, which is the degree of regularization. As the datasets contained several ratings, we ran the code with each individual rating as the target variable.

As seen in the code file, the results showed that the Stochastic Average Gradient Descent (SAG) method performs best on four out of the five ratings on the unlabeled dataset, with Limited Memory-BFGS (LBFGS) being chosen as the solving method for ‘Career Opportunities’ as the only outlier.

In contrast, when applied to the labeled dataset, LBFGS performed best three out of five times with similar accuracy scores to the SAG method. LBFGS as a solving method function well for large datasets but as it stores memory as part of its optimization method it might be slower. SAG performs similarly to LBFGS on most ratings, leaving us with two

Fig. 5. Boxplots of accuracy scores on each model in the model collection.

solving methods that are similarly suited for large datasets. SAG was chosen as it is more efficient on memory usage and allows for larger scalability for future model tuning with more features without a significant penalty in accuracy [21].

The penalty hyperparameter on the two datasets applied the L2 (ridge) regularization and none. The L2 penalty lowers the weights, and thus, reduces the complexity of the dataset and that allows for a more generalizable model while reducing potential overfit and was thus included in the final model.

The C values that were tested in the RandomSearchCV all had low values, with four out of five values being between 0.2 and 0.0 for the unlabeled dataset and with a similar pattern in the labeled dataset. This suggested that the datasets predict class labels reliably, and do not require a high C value for accurate predictions. The final model applied the following hyperparameters C = 0.3, solver = 'sag', and penalty = 'l2'.

2) Final Logistic Regression Model

Running the final tuned model provided the test and training accuracy results, which can be seen in table I. The test and training accuracy was almost the same on both datasets, which may indicate the models have not been overfit, but has a reasonable balance between bias and variance [15]. The models tended to predict 'career_opp' best but accuracy scores were similar overall for each y variable, which might indicate that our chosen hyperparameters can be used in general for Glassdoor reviews. Finally, the labeled dataset predicted almost with same accuracy as the dataset with job titles. This may indicate that it is possible to label Glassdoor reviewer's job titles meaningfully to avoid a potentially infinitely high dimensionality, but also that it is possible to gain some insights for each individual job title.

TABLE I

ACCURACY OF JOB TITLES DATASET

	Training Accuracy	Test Accuracy
career_opp	0,63	0,63
work_life_balance	0,60	0,60
senior_mgmt	0,58	0,58
comp_benefits	0,56	0,56
culture_values	0,61	0,61

ACCURACY OF LABELED DATASET

	Training Accuracy	Test Accuracy
career_opp	0,63	0,63
work_life_balance	0,59	0,59
senior_mgmt	0,57	0,57
comp_benefits	0,55	0,55
culture_values	0,61	0,60

The 'career_opp' rating tended to be predicted accurately / similarly (>60% accuracy) across all job titles except for reviewers who work as a project manager. Looking at the labeled data, a similar pattern was seen, but the accuracy score was highest for 'job_Business Professionals' and 'job_Service professionals'. 'work_life_balance' was predicted most accurately for 'Senior Associate' (69%), 'Senior Software Engineer' (71%) and 'Software Engineer' (74%). For the labeled data 'job_Technology Professional' (73%) tends to be the most accurate label for predicting work life balance satisfaction. Accuracy in predicting 'senior_mgmt' satisfaction tends to be in the lower end (<60%) compared with the accuracy of the other target variables but was most accurately predicted by observations where the job title was "Crew member' (69%), 'Customer Service Representative' (64%), 'Sales Assistant' (63%) or 'Cashier' (63%). Similarly, satisfaction with Senior Management was best predicted by observations labeled 'Service professional' (64%). Accuracy of predicting Compensation and Benefits were lower compared with the other y variables overall (<55%), but was predicted well by observations that contained job titles 'Cashier' (68%), 'Crew Member' (76%), 'Sales assistant' (63%) and 'Software Engineer' (64%). Similarly, the accuracy scores were highest for observations with labels 'Service Professional' (64%). 'culture_values' was predicted with a relatively high accuracy (above 60%) across most job titles. They were predicted particularly well for 'Software engineer' (68%), 'Senior Software Engineer' (66%), 'Senior Manager' (66%) and 'Crew member' (64%). Similarly, the highest accuracy score for 'culture_values' across the labeled data was found in the observations that contained the label 'job_Technology Professional' (67%).

For the dataset containing job titles, the most important features tended to have negative coefficients when predicting reviews overall (<-0,60). When predicting 'career_opp', both the 'Crew Member' and 'Sales Assistant' variables had a negative coefficient below -0.7. Similarly, 'job_Service_Professional' was the only labeled variable to have a coefficient below -0.7. When predicting 'work_life_balance' ratings 'Senior Associate' had a relatively large negative coefficient of -0,77. However, the rest of the variables with negative coefficients did not have nearly as large negative coefficients as 'Senior Associate'. Overall, thereweres just one large negative coefficient for 'senior_mgmtt', which was 'Crew Member'. This can also be seen on the labeled data, in which there were no coefficients higher than -0,7 when predicting 'senior_mgmt'. When predicting 'comp_benefits', most job title variables had small coefficients except for 'Cashier' and 'Crew Member' that had large negative coefficients as respectively -0,7 and -1,08. Similarly, the largest negative coefficient, when training the model on the labeled dataset was 'job_Service Professional' (-0,52). The pattern was the same for 'culture_values' in which 'Cashier' (-0,57) and 'Crew Member (-0,79) as well as 'Service Professionals' (-0,46) had the largest negative coefficients while the rest of the negative coefficients were small (>-0,20).

IV. DISCUSSION

A. Implications of Results

To evaluate the models, we calculated the class priors for

the two datasets, and for both datasets it was 0,54. From the results, it seemed that the trained models all had higher accuracy scores than the benchmark, which suggests that they are performing better than a baseline model. Though this is the case, it is worth considering if a model with a 0,60 accuracy score should be deployed in real life.

The results might also reflect other, general findings. We found that there were two large clusters: cluster 0 and cluster 1. Cluster 0 includes reviews that are mostly dissatisfied across all categories, where cluster 1 includes reviews that are mostly satisfied across all categories.

Findings from the association rule mining suggests a strong relationship between clutter 0 and 'Service Professionals'. This means that what we define as service professionals tend to review their workplace negatively. Unfortunately, this finding validates the assumption that the work environment is not supportive of service professionals. Similarly, the LR model found that job titles included in the service professional label tend to have the largest negative coefficients across all categories. The tendency for service professionals to review negatively, are further substantiated by our supervised methods in which most of the largest negative coefficients for predicting each rating outcome belonged to service professionals.

While no clear pattern or strong association rules amongst other than service professionals, there were more association rules found for cluster 1 with Software engineers including in the antecedent itemset compared with other job titles. This is similar to the strong positive coefficients found with logistic regression, especially across 'Work Life Balance' and 'Compensation and Benefits' ratings.

'Consultants', 'Senior Consultant' and 'Senior Associate' belonging to 'Business Professional' seem to have coefficients that match the association rules found for cluster 2 where 'Career Opportunities' is rated high, but 'Work Life Balance' is rated low. In further research it might be interesting to create some more granular labels such as 'Consulting Professional', as these professions may be quite different from a workplace review perspective than 'Project Manager'.

Overall, the results may indicate that a) working as a service professional can predict overall dissatisfaction with the workplace, b) working in consulting indicates a high satisfaction with career opportunities but at the expense of work life balance while c) software engineering somewhat reliably predicts an overall high satisfaction, but especially a high satisfaction with work life balance and compensation. A factor that may have implications for the resulting clusters of the dataset is the relatively low amounts of reviews from jobs of the 'Service Professional' label, as the majority of reviews was found in the Business and Technology Professional labels in the labeled dataset, and similarly with the job titles associated with the same labels.

B. Implications of Practice

It can be helpful for organizations to understand the factors that influence how their employees evaluate their job performance and the overall work environment. This can help them identify areas for improvement and make changes that are more likely to be well received by their employees. Understanding the relationship between reviewer's jobs and their evaluation might be particularly useful, as it can provide insights into the specific aspects of the job that are most important to employees and how well the organization is meeting their needs. Furthermore, the relationship between a reviewer's job and their evaluation of their work environment can have societal implications in a number of ways. For example, if certain job roles consistently receive low ratings in certain areas such as service professionals, it could be a sign that the work environment is not supportive of employees in those roles or that there are broader systemic issues at play. This could have implications for the overall well-being and satisfaction of employees in those roles, as well as for the quality and effectiveness of the work that they do. Additionally, if the evaluations of certain job roles consistently differ from those of other job roles, it could be a sign of unequal treatment or discrimination within the organization, which could have broader societal implications in terms of fairness and equity. Thus, deploying the model to gain insights into the organization's work environment can help organizations create a more supportive and inclusive work environment that benefits both their employees and society as a whole.

V. Conclusion

This project has sought to understand if reviewers' professions on Glassdoor affect their Workplace Factor ratings. Our results indicate moderately effective predictions for each individual rating. The project applied the accuracy metric as scoring based on our final model. The parameters and hyperparameters chosen for the supervised model followed three evaluative criteria, namely, accuracy tendency to overfit, and model interpretability. The model showed that the labels 'job_Service_Proffesional' and 'job_Technology_Proffesional' had LR coefficients that showed a relative dissatisfaction in the ratings left by service professionals concerning ratings on work benefits and high ratings left on work life balance by the technology professional group. Finally, we find that 'Business Professional' which is primarily formed by reviewers working within Consulting firms, can predict high satisfaction with career opportunities and low satisfaction with work life balance. Increasing the amount of reviews or adding more job titles and labels could improve the models predictive capacity further as the current accuracy scores does not allow for a high confidence in predicting individual job title ratings reliably. Applying data mining methods allowed to establish relationships between individual job titles based on similarity. and further research should follow so as to increase the interpretability of our results

References

[1] N. El-Rayes, M. Fang, M. Smith, and S. M. Taylor, "Predicting employee attrition using tree-based models," International Journal of

Organizational Analysis, vol. 28, no. 6, pp. 1273–1291, Mar. 2020, doi: 10.1108/ijoa-10-2019-1903.

[2] B. Sainju, C. Hartwell, and J. Edwards, "Job satisfaction and employee turnover determinants in Fortune 50 companies: Insights from employee reviews from Indeed.com," Decision Support Systems, vol. 148, p. 113582, May 2021, doi: 10.1016/j.dss.2021.113582.

[3] V. Chang, Y. Mou, Q. A. Xu, and Y. Xu, "Job satisfaction and turnover decision of employees in the Internet sector in the US," Enterprise Information Systems, Oct. 2022, doi: 10.1080/17517575.2022.2130013.

[4] J. L. West, K. M. Knippenberg, and J. M. Sitkin, "The Influence of Employee Online Ratings on Hiring Decisions," Organizational Behavior and Human Decision Processes, vol. 47, Art. no. 2, 2015, doi: 10.1016/j.obhdp.2015.02.004.

[5] Kaggle, "Glassdoor Job Reviews," https://www.kaggle.com/datasets/davidgauthier/glassdoor-job-reviews. https://www.kaggle.com/datasets/davidgauthier/glassdoor-job-reviews (accessed Jan. 06, 2023).

[6] Glassdoor, "Glassdoor Help Center," help.glassdoor.com, Mar. 03, 2022. https://help.glassdoor.com/s/article/Ratings-on-Glassdoor?language=en_US

[7] J. Brownlee, "A Gentle Introduction to Imbalanced Classification," Machine Learning Mastery, Dec. 22, 2019. https://machinelearningmastery.com/what-is-imbalanced-classification/ (accessed Jan. 07, 2023).

[8] Pandas, "pandas.get_dummies — pandas 1.2.4 documentation," pandas.pydata.org. https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html (accessed Jan. 08, 2023).

[9] J. Han and M. Kamber, Data Mining : Concepts and Techniques, 3rd ed. Haryana, India ; Burlington, Ma: Elsevier, 2012.

[10] N. J. de Vos, "kmodes: Python implementations of the k-modes and k-prototypes clustering algorithms for clustering categorical data.," PyPI, Sep. 06, 2022. https://pypi.org/project/kmodes (accessed Jan. 06, 2023).

[11] F. Cao, J. Liang, and L. Bai, "A new initialization method for categorical data clustering," Expert Systems with Applications, vol. 36, no. 7, pp. 10223–10228, Sep. 2009, doi: 10.1016/j.eswa.2009.01.060.

[12] Andriy Burkov, The Hundred-Page Machine Learning Book. Quebec, Canada: Andriy Burkov, 2019.

[13] mlxtend, "Mlxtend.frequent patterns - mlxtend," rasbt.github.io, 2022. http://rasbt.github.io/mlxtend/api_subpackages/mlxtend.frequent_patterns/

[14] SciKit, "Metrics and scoring: quantifying the quality of predictions — scikit-learn 0.23.2 documentation," scikit-learn.org. https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score (accessed Jan. 08, 2023).

[15] G. James, D. Witten, T. Hastie, and R. Tibshirani, INTRODUCTION TO STATISTICAL LEARNING : with applications in r. S.L.: Springer-Verlag New York, 2021.

[16] P. R. Nagbøl, O. Müller, and O. Krancher, "Designing a Risk Assessment Tool for Artificial Intelligence Systems," The Next Wave of Sociotechnical Design, pp. 328–339, 2021, doi: 10.1007/978-3-030-82405-1_32.

[17] Carnegie Mellon University, "Logistic Regression Advanced Methods for Data Analysis (36-402/36-608) 1 Classification," 2014. Accessed: Jan. 08, 2023. [Online]. Available: https://www.stat.cmu.edu/~ryantibs/advmethods/notes/logreg.pdf

[18] Scikit-Learn, "RandomizedSearchCV," Scikit-learn.org, 2019. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html (accessed Jan. 08, 2023).

[19] G. Chinazzo, "Investigating the indoor environmental quality of different workplaces through web-scraping and text-mining of Glassdoor reviews," Building Research & Information, pp. 1–19, Apr. 2021, doi: 10.1080/09613218.2021.1908879.

[20] S. Gottipati, K. J. Shim, and S. Sahoo, "Glassdoor Job Description Analytics – Analyzing Data Science Professional Roles and Skills," 2021 IEEE Global Engineering Education Conference (EDUCON), Apr. 2021, doi: 10.1109/educon46332.2021.9453931.

[21] SciKit-Learn, "sklearn.linear_model.LogisticRegression — scikit-learn 0.21.2 documentation," Scikit-learn.org, 2014. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed Jan. 08, 2023).

[22] J. Brownlee, "Hyperparameter Optimization With Random Search and Grid Search," Machine Learning Mastery, Sep. 13, 2020. https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/

Appendix I

Writing Activities	
Section	**Subsection**
I. Introduction	
II. Methodology	A. Data Collection, Preparation and Preprocessing
	B. Unsupervised Learning Methods
	B.1. *k*-Modes Clustering
	B.2. Frequent Pattern Mining
	C. Supervised Learning Methods
	C.1. Selecting the Best Model
	C.2. Cross-Validation of Accuracy Scores
	C.3. Evaluation Framework
	C.4. Logistic Regression Classification and Hyperparameter Tuning
III. Results	A. *k*-Modes Clustering and Frequent Pattern Mining
	B. Evaluation of Supervised Models
	C. Tuned Logistic Regression Classifier
	C.1. Hyperparameter Tuning
	C.2. Final Logistic Regression Model
IV. Discussion	A. Implication of Results
	B. Implication of Practice
V. Conclusion	

Appendix II

Accuracy

Final_df with Job Titles

job	career_opp	work_life_balance
job_title_Analyst	0,59	0,53
job_title_Assistant Manager	0,61	0,61
job_title_Associate	0,67	0,62
job_title_Business Analyst	0,61	0,52
job_title_Cashier	0,62	0,56
job_title_Consultant	0,67	0,63
job_title_Crew Member	0,65	0,54
job_title_Customer Assistant	0,62	0,52
job_title_Customer Service Representative	0,61	0,56
job_title_Director	0,60	0,59
job_title_Manager	0,62	0,58
job_title_Project Manager	0,54	0,57
job_title_Sales Assistant	0,68	0,49
job_title_Sales Associate	0,62	0,55
job_title_Senior Associate	0,70	0,69
job_title_Senior Consultant	0,67	0,65
job_title_Senior Manager	0,65	0,55
job_title_Senior Software Engineer	0,57	0,71
job_title_Software Engineer	0,63	0,74
job_title_Vice President	0,57	0,54

Final_df with Job Labels

job	career_opp	work_life_balance
job_Business Professional	0,64	0,59
job_Management Professional	0,61	0,56
job_Service Professional	0,64	0,53
job_Technology Professional	0,61	0,73

Coefficients

Title: Coefficients of data with Job Titles

Feature	career_opp	work_
job_title_Analyst	0,16	
job_title_Assistant Manager	0,31	
job_title_Associate	0,48	
job_title_Business Analyst	0,22	
job_title_Cashier	-0,62	
job_title_Consultant	0,61	
job_title_Crew Member	-0,76	
job_title_Customer Assistant	-0,67	
job_title_Customer Service Representative	-0,60	
job_title_Director	0,19	
job_title_Manager	0,29	
job_title_Project Manager	-0,12	
job_title_Sales Assistant	-0,87	
job_title_Sales Associate	-0,48	
job_title_Senior Associate	0,63	
job_title_Senior Consultant	0,52	
job_title_Senior Manager	0,34	
job_title_Senior Software Engineer	0,11	
job_title_Software Engineer	0,32	
job_title_Vice President	-0,08	
tenure_0	-0,02	
tenure_1	-0,17	
tenure_10	-0,26	
tenure_3	-0,12	
tenure_5	-0,09	
tenure_8	-0,14	
current_0	-0,20	
current_1	0,20	

Title: Coefficients of Labeled Dataset

feature	career_opp	work_life_
tenure_0	-0,02	
tenure_1	-0,15	
tenure_10	-0,34	
tenure_3	-0,09	
tenure_5	-0,10	
tenure_8	-0,19	
current_0	-0,20	
current_1	0,20	
job_Business Professional	0,35	
job_Management Professional	0,20	
job_Service Professional	-0,74	
job_Technology Professional	0,20	