DOI: https://doi.org/10.24297/ijmit.v16i.9072

Predicting Tech Employee Job Satisfaction Using Machine Learning Techniques

Sumali J. Conlon[1] Lakisha L. Simmons[2] Feng Liu[3]

[1]School of Business Administration, University of Mississippi, University, MS 38677, USA

[2] BRAVE Consulting LLC, USA

[3]Department of Economics and Decision Sciences, Western Illinois University, IL 61455, USA

[1]sconlon@bus.olemiss.edu, [2]drkisha@lakishasimmons.com, [3]f-liu@wiu.edu

Abstract

High-tech industry employees are among the most talented groups of people in the workforce, and are therefore difficult to recruit and retain. We analyze employee reviews submitted by employees from five technology companies. Following the Cross-Industry Standard Process for Data Mining (CRISP-DM) and the data science life cycle process, we use machine learning techniques to analyze employees' reviews. Our goal is to predict an overall measure of whether employees are satisfied or not, using other information from the reviews, such as employer attitudes towards upper management. We also use predictive analysis to determine which features are more helpful in determining an employee's overall job satisfaction. Finally, we analyze which prediction algorithm provides the most accurate predictions. We find the percentage of true positives we correctly identify in the holdout sample is 97.4%, while the percentage of true negatives correctly identified is 72.5%.

Keywords: Employee Job Satisfaction, High-Tech Industry Employees, Machine Learning, Predictive Analysis

1. Introduction

In order for high-tech firms to be successful, they must attract high caliber and knowledgeable employees to join their teams. They also need to provide a work environment where employees feel satisfied with the company culture, their everyday work life, and their career development, among other things. The more satisfied employees are with their companies, the more willing they will be to perform to the best of their ability.

There has been a great deal of research in the human resource management area about what makes employees satisfied with their jobs. Locke (1976) defines job satisfaction as "a pleasurable or positive emotional state resulting from the appraisal of one's job or job experiences." Spector (1997) presented a list of reasons why employees are satisfied or unsatisfied with their jobs, based on factors such as Appreciation, Communication, Coworkers, Fringe benefits, Job conditions, Nature of the work, Organization, Personal growth, Policies and procedures, Promotion opportunities, Recognition, Security, and Supervision. Later, Aziri (2011) and Singh and Jain (2013) summarized employee job satisfaction factors as follows:

1. Policies of Compensation and Benefits (most important – such as rewards, salary packages compared to others in the same industries, etc.)
2. Job Security (employees can keep their jobs)
3. Working conditions (feeling of safety, comfort, and motivation)
4. Relationship with Superior Authority
5. Promotion and Career Development
6. Leadership Styles (democratic style of leadership)
7. Work Group (group dynamics, cohesiveness, and affiliation)
8. Personal Variables (such as personality, expectations, age, education, and gender differences)
9. Others (such as whether group members are outgoing (feeling like part of a family), encouragement and feedback, and use of internet and other technologies for doing job).

One of the major problems that organizations face is employee turnover. It can cost businesses a lot to find replacements. Thus, it is important to understand what causes employees to be satisfied or dissatisfied with their companies, to both prevent employee turnover and to keep current and prospective employees feeling satisfied with their work.

There are many sources of data concerning employee opinions about their companies. These include internal data collections (within the companies) and external sources such as data collections at Glassdoor.com. Datasets from open sources tend to be anonymous, which allows employees to have more freedom to share their real feelings about their places of work, without the risk of losing their jobs. As a result, true in-depth information is available, and since the reviews are typically anonymous, they often contain a voluminous amount of information. Thus, these datasets allow researchers to do a deep analysis of employee opinions, using various techniques.

Most of the current data collections from online reviews include both star ratings and open-ended opinions in a textual format. Reviews are submitted by various groups of employees, which vary by type of job (e.g., programmers or managers), status (e.g., current or former employees), rank, and anonymous versus non-anonymous, etc. Such datasets allow more in-depth analysis of what employees think about the companies they work for, which can help companies improve the recruiting and retention of these employees.

Current data analysis tools and techniques allow researchers to analyze electronic data more effectively. To structure our business predictions, we use the Cross-Industry Standard Process for Data Mining (CRISP-DM) and the data science life cycle process. CRISP-DM is a widely accepted methodology for data mining and analytics (IBM Knowledge Center). It involves five phases: 1) Business Understanding, 2) Data Understanding, 3) Data Preparation, 4) Modeling, 5) Evaluation and 6) Deployment. The data science life cycle process consists of the problem definition, ETL (extract, transform, load) & feature extraction, learning, and model deployment & development. We use machine learning techniques to analyze tech industry employees' reviews to find out whether or not employees' overall job satisfaction (as measured by star ratings) can be predicted, based on other data available in their reviews, such as star ratings of senior management, as well as written comments.

This paper complements previous research (Conlon, 2021). That previous paper looked at why data scientists consider changing jobs as a function of *environmental* variables (such as city development levels and company sizes and types). The current paper, by contrast, looks at high-tech employee job satisfaction (and so, presumably willingness to stay at their current job) in terms of *their attitudes towards various aspects of their employment situation*. This paper then examines what the major features are that contribute to accurately predicting this job satisfaction, and which algorithms perform best in predicting job satisfaction.

Using CRISP-DM, we begin with Business Understanding. That is, we define our data mining objectives and the data mining problem in terms of a set of research questions. Thus, in this research, we ask:

RQ 1: Can we use machine learning techniques to predict employee job satisfaction from employees' reviews?

RQ 2: Using machine learning techniques, which algorithms perform best in predicting employee job satisfaction?

RQ 3: Using machine learning techniques to analyze employee reviews, which features have the highest predictive power in helping the system to predict accurately?

The rest of the paper is organized as follows. First, we discuss related work in the analysis of employees' job satisfaction, text mining, sentiment analysis and machine learning. Next, we discuss the dataset used, and also does some comparisons of the data across several firms. Also, we discuss our predictive analysis methodology and presents our research findings. Then, we discuss the results from the analysis and the business implications based on the techniques and the findings. Finally, we conclude the paper and proposes future research.

2. Related Work

This research is based on two major related research areas: analysis of employee job satisfaction and the techniques used in analyzing datasets to do predictive analysis (we use machine learning techniques in this research).

2.1 Employee Job Satisfaction

There has been a great deal of research on employee job satisfaction in the literature. In general, employee job satisfaction measures various dimensions of whether or not employees are satisfied with their jobs. Locke (1976) defines job satisfaction as "a pleasurable or positive emotional state resulting from the appraisal of one's job or job experiences" (p. 1304), while Spector (1997) defines job satisfaction as “how an individual is with his or her job; whether he or she likes the job or not.” The level of job satisfaction can be evaluated at the global level (overall satisfaction) or according to the particular aspects of the job that make employees satisfied or not satisfied. Spector (1997) lists 14 common aspects of employee job satisfaction, i.e., appreciation, communication, coworkers, fringe benefits, job conditions, nature of the work, organization, personal growth, policies and procedures, promotion opportunities, recognition, security, and supervision.

Hulin and Judge (2003) argue that job satisfaction is influenced by several psychological factors including cognitive (evaluative), affective (or emotional), and behavioral factors. Cognitive job satisfaction can be evaluated using one dimension, such as benefits or supervision, or multiple dimensions if two or more facets of a job are evaluated simultaneously. Affective or emotional factors are a response to the job or to cognitive opinions about the job and reflect the degree of pleasure or happiness employees feel about their jobs. Finally, Hulin and Judge (2003) state that behavioral components of job satisfaction can be related to other key factors such as working conditions, stress level at work, etc. Other important research papers on job satisfaction can be found from Chen, 2007; Galup et al., 2008; Harden et al., 2018; Mishra, 2013; Moorman, 1993; Morris & Venkatesh, 2010; Saari & Judge, 2004; Thompson & Phua, 2012; and Vidal et al. 2007.

A great deal of research has been done analyzing what causes an employee to decide to leave an organization. This research has found that, in general, employees’ decisions to leave an organization can be influenced by several factors, including individual, organizational, and economic variables (Lee et al., 2004; Mobley, 1982; Morrell et al., 2001; Mossholder, 2005; and Mowday et al., 2001).

Employee turnover costs business a lot. Employee Benefit News (EBN) reported in August 11, 2017 that if an employee leaves a company, it will cost the employer 33% of that worker’s annual salary to hire a replacement. For example, for a media salary of $45,000 a year, the cost of finding a replacement is about $15,000 per person (Bolden-Barrett, 2017). Thus, understanding what influences employee job satisfaction can benefit firms a great deal if it can help them to prevent employee turnover.

Hinkin and Tracey (2000) lists the five major categories of employee turnover costs as:

- Predeparture (costs that are incurred once an employee has given notice),
- Recruitment (promotional materials, advertising, and recruiting sources),
- Selection (identifying the most suitable candidates - Interviewing, background and reference checks, and travel expenses),
- Orientation and Training (almost everyone requires some sort of formal or informal training),
- Productivity Loss (largest percentage of the total costs, up to 70 percent in some cases

Some important related research on employee turnover can also be found at Abraham, 1999; Hinkin & Tracey, 2000; Hinkin & Tracey, 2006; and Vidal et al., 2007.

To prevent employee turnover, organizations need to understand the causes employee dissatisfaction. Cotton and Tuttle (1986) found that the strongest predictors of voluntary turnover were age, tenure, pay, overall job satisfaction, and the employee’s perceptions of fairness. Holton et. al. (2008); Sacco & Schmitt (2005); and several other researchers found that personal or demographic variables, specifically age, gender, ethnicity, education, and marital status, were important factors in predicting voluntary employee turnover. Recently, one report, by the Work Institute, found in exit interviews that the top reasons survey respondents gave for leaving their jobs were career development (22%), work-life balance (12%), managers' behavior (11%), compensation and benefits (9%) and well-being (9%) (2017 Retention Report: Trends, Reasons & Recommendations (1)). Moreover, the Work Institute’s Retention Report studied 34,000 respondents and found that 75% of the cases of employee turnover were preventable (2017 Retention Report: Trends, Reasons & Recommendations (1)).

Thus, being able to predict whether employees are satisfied or dissatisfied with the organization should be very helpful in assisting firms to reduce employee turnover.

2.2 Machine Learning Techniques and Their Applications in Human Resource Management

Machine learning (ML) is a subarea of artificial intelligence that aims to use algorithms and statistical models to train computer systems to perform tasks such as prediction, without explicitly programming the computer for that specific task. With the rapid growth of electronic data, ML has been used in many application areas such as medical diagnosis, speech recognition, image processing, and many business applications.

In business, ML has been used widely in areas such as sales, product recommendation, dynamic pricing, marketing, finance, etc. In the human resource management area, ML has been used, for example, in predicting employee turnover (Punnoose & Ajit, 2016), analyzing employees' attrition (Alao & Adeyemo, 2013; Fallucchi, 2020; Nagadevara, 2008; and Ray & Sanyal, 2019) and analyzing and predicting employee engagement (Golestani et al., 2018). Jain et al. (2021) studies how internal marketing tools affect employees' job satisfaction of a logistic firm. For the data, several research groups use the data from IBM HR Analytics Employee Attrition & Performance available at Kaggle.com (20).

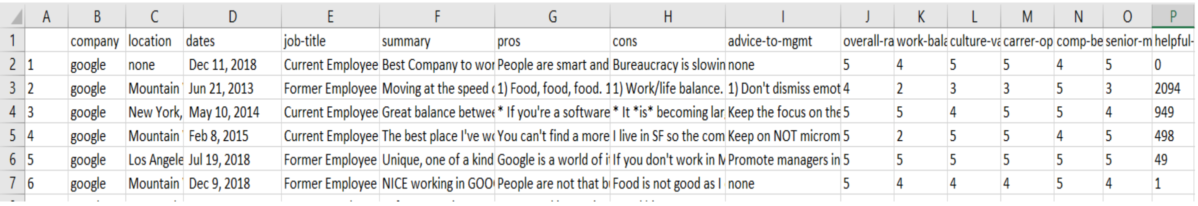
In order to analyze employee churn, Bendemra (2019) built an employee churn model to develop a strategic retention plan using Python and found that the stronger indicators of people leaving include: Monthly Income (employees with higher wages are less likely to leave), Overtime (people who work overtime are more likely to leave the company), age (25–35 are more likely to leave), Distance From Home (Employees who live further from home are more likely to leave the company), TotalWorkingYears (more experienced employees are less likely to leave, so employees who have between 5–8 years of experience should be identified as potentially having a higher-risk of leaving), YearsAtCompany (Employees who hit their two-year anniversary should be identified as potentially having a higher-risk of leaving), YearsWithCurrManager (A large number of leavers leave 6 months after they have worked for their current managers) (Bendemra, 2019).

3. Data Collection

In this section, we continue with phase two of the CRISP-DM methodology: data understanding. In this phase, we collect and describe our data collection, and assess data quality. We used online data sources for our research and analysis.

3.1 Data Source:

In this study, we used data consisting of employee reviews from technology companies. We downloaded the data from Kaggle.com (the dataset has since been removed). The data retrieved from Kaggle.com was originally scraped from the website Glassdoor.com by the original Kaggle.com collection author. The Kaggle.com page included over 67k reviews from employees of Google, Amazon, Facebook, Apple, Microsoft, and Netflix. Due to a disproportionately low number of Netflix reviews (810), we excluded Netflix from our study. Reviews are both from current and former employees, and both from anonymous employees and from employees whose identity was disclosed. An excerpt from the csv data file is shown in Figure 1.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		company	location	dates	job-title	summary	pros	cons	advice-to-mgmt	overall-ra	work-bala	culture-va	carrer-op	comp-be	senior-m	helpful-
2	1	google	none	Dec 11, 2018	Current Employee	Best Company to wor	People are smart and	Bureaucracy is slowin	none	5	4	5	5	4	5	0
3	2	google	Mountain	Jun 21, 2013	Former Employee	Moving at the speed	1) Food, food, food. 1	1) Work/life balance.	1) Don't dismiss emot	4	2	3	3	5	3	2094
4	3	google	New York,	May 10, 2014	Current Employee	Great balance betwee	* If you're a software	* It *is* becoming lar	Keep the focus on the	5	5	4	5	5	4	949
5	4	google	Mountain	Feb 8, 2015	Current Employee	The best place I've w	You can't find a more	I live in SF so the com	Keep on NOT microm	5	2	5	5	4	5	498
6	5	google	Los Angele	Jul 19, 2018	Former Employee	Unique, one of a kind	Google is a world of i	If you don't work in M	Promote managers in	5	5	5	5	5	5	49
7	6	google	Mountain	Dec 9, 2018	Former Employee	NICE working in GOO	People are not that b	Food is not good as I	none	5	4	4	4	5	4	1

Figure 1: Format of the Data

Similar to general product reviews that appear in many online sources, this data set consists of both textual and numerical information. The attributes provided by the Kaggle.com data collection page are described in Table 1

Table 1: Attribute Descriptions. NOTE: 'None' is Placed in All Cells Where No Data Value Was Found.

Column	Column Name	Description
A	Index	Index (i.e., review number)
B	Company	Company name (Amazon, Apple, Facebook, Google, and Microsoft)
C	Location	This dataset is global. As such it may include the country's name in parenthesis [i.e. "Toronto, ON (Canada)"]. However, if the location is in the USA then the dataset will only include the city and state [e.g., "Los Angeles, CA"]
D	Date Posted	Date the review was posted.
E	Job-Title	This string will also include whether the reviewer is a 'Current' or 'Former' Employee at the time of the reviews
F	Summary	A short summary of the employee review
G	Pros	Positive aspects of the job according to the employee
H	Cons	Negative aspects of the job according to the employee
I	Overall Rating	A rating scale of 1-5 where 1 is low and 5 is high.
K	Work/Life Balance Rating	A rating scale of 1-5 where 1 is low and 5 is high.
L	Culture and Values Rating	*No description given*
M	Career Opportunities Rating	*No description given*
N	Comp & Benefits Rating	*No description given*
O	Senior Management Rating	*No description given*
P	Helpful Review Count	A count of how many people found the review to be helpful
Q	Link to Review	This provides the user with a direct link to the page that contains the review. However, it is likely that this link will be outdated

Distribution of the data

The total number of reviews for each company is displayed in Figure 2.

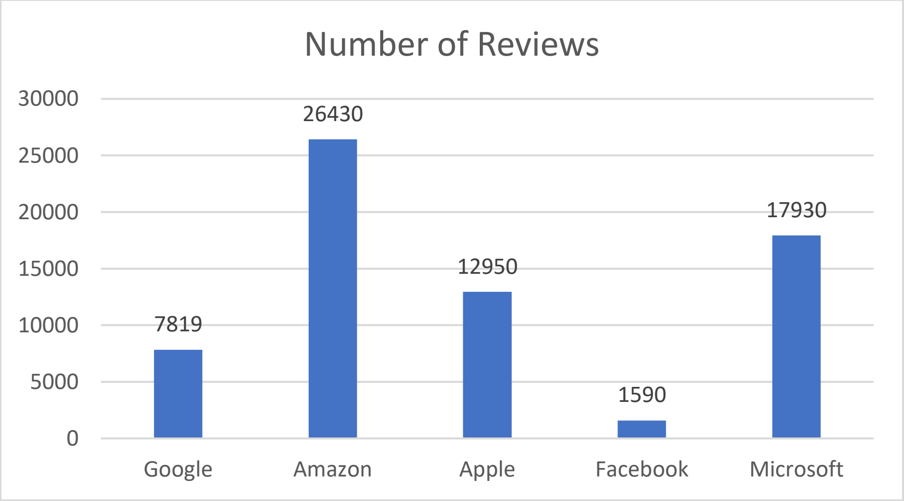


Figure 2: Number of Reviews for Each Company

A count of the overall ratings by number of stars for all companies is shown in Figure 3.

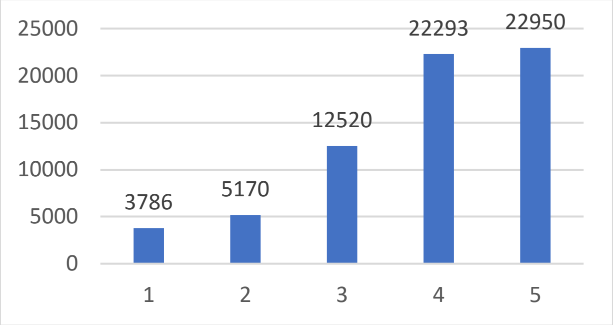


Figure 3: Number of Ratings by Stars

The star ratings are between 1 and 5 (1 is low and 5 is high). To make the predictions more accurate, we grouped the reviews into two broad categories: satisfied and dissatisfied. The reviews that contained an overall rating between 3 and 5 were classified as "satisfied" while the reviews that contained overall ratings of 1 or 2 were classified as "dissatisfied." We broke up ratings like this because we felt that only the 10%-20% most dissatisfied employees were likely to leave. However, as a robustness check, we also analyze the 123/45 classification below. Table 2 shows the proportion in percentages of the 12 vs 345 groups for each company.

Table 2: Percentage of Employees Who are Satisfied and Dissatisfied in Each Company

	Amazon		**Apple**		**Facebook**		**Google**		**Microsoft**	
	1-2 stars (%)	3-5 stars (%)	1-2 stars (%)	3-5 stars (%)	1-2 stars (%)	3-5 stars (%)	1-2 stars (%)	3-5 stars (%)	1-2 stars (%)	3-5 stars (%)
Overall Rating	20	80	10	90	6	94	5	95	11	89
Work/LifeBalance Rating	36	64	26	74	12	88	11	89	18	82
Culture and Values Rating	24	76	11	89	7	93	6	94	17	83
Career Opportunities Rating	21	79	23	77	7	93	10	90	14	86
Comp & Benefits Rating	15	85	8	92	2	98	4	96	5	95
Senior Management Rating	31	69	24	76	9	91	13	87	29	71

A graphical representation of star ratings is shown in Figure 4.

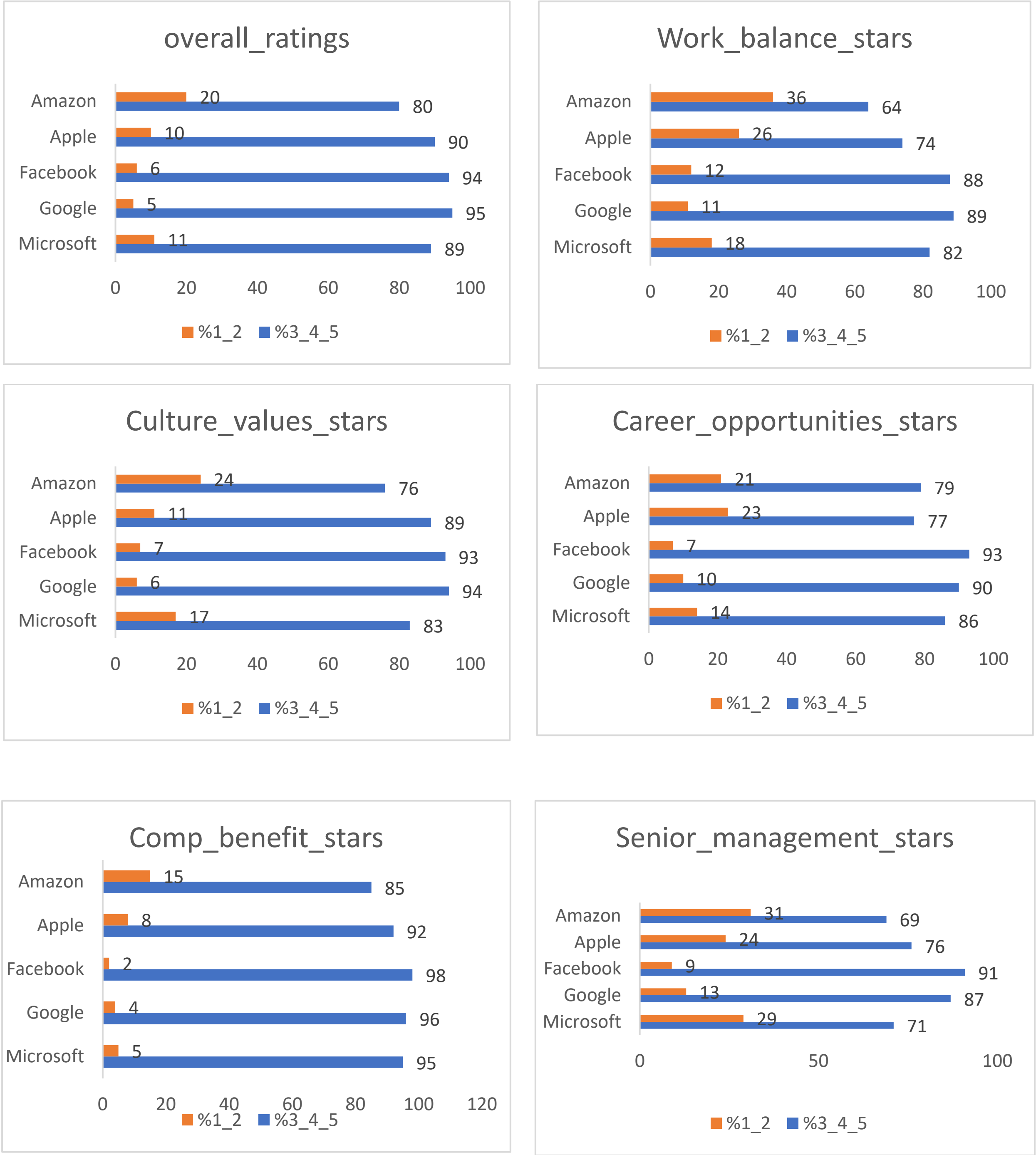


Figure 4: Graphs Showing Percentage of Reviewers That are Satisfied Versus Dissatisfied for Each Category and Company

Textual Data Analysis

There is unstructured text in the summary column, as well as in the pros and cons columns of the reviews. Text comments are important because they include an in-depth employee views about the company in question. One way to understand which terms are important in the corpus is by using a word cloud. A word cloud is a visual representation of the textual data. The level of importance and/or frequency is indicated by the size of the word. Figure 5 shows an example of a word cloud. This example shows that, for example, the term "management" is an important word that is mentioned frequently.

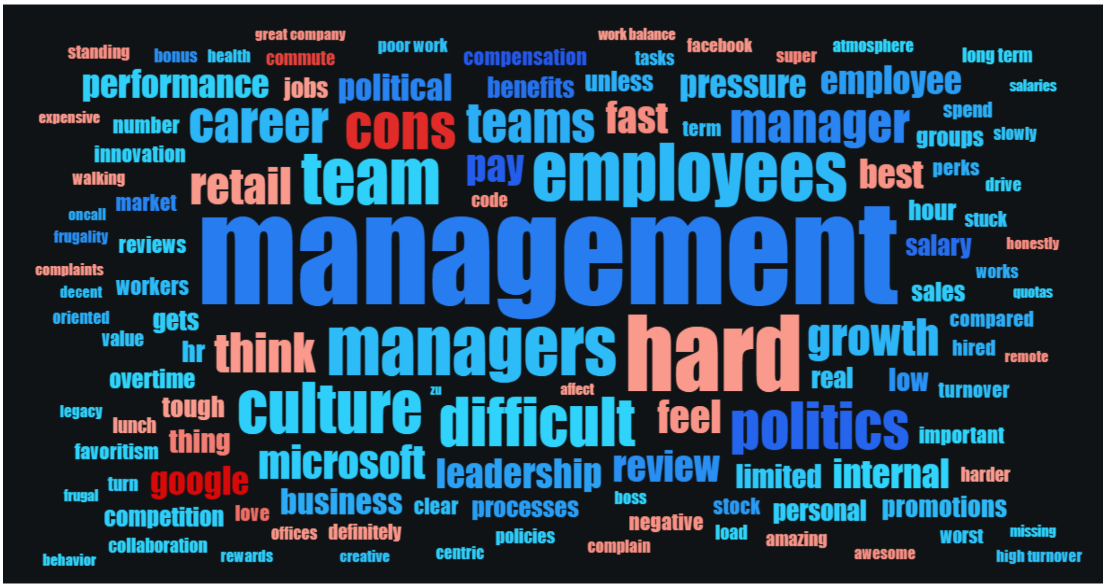


Figure 5: Word Cloud Visualization Created from the Text Portion of the Employee Reviews

Word Frequencies

Star ratings indicate the level of satisfaction but do not indicate the nature of the sentiment creating that satisfaction or dissatisfaction. To determine the specific reasons why employees feel satisfied or not, we use text analysis techniques to study the contents in the reviews. Based on the theory of information retrieval, the system first eliminates "stop words" (such as "is,", "else," "between," "the"). The number of occurrences of the remaining terms then indicates the importance of those terms in the reviews. The following table shows some examples of term occurrences in the comments from positive and negative reviews (based on the two rating categories, 1-2 being negative and 3-5 positive). Sample term occurrences from positive reviews are listed in Table 3 and from negative reviews in Table 4.

Table 3: Sampling of Term Occurrences from Positive Reviews

Positive Review Terms	Relative Frequencies	Absolute Frequencies
good	1	6476
benefits	0.899629	5826
opportunities	0.475911	3082
best	0.315473	2043
amazing	0.273935	1774
career	0.264206	1711
growth	0.257721	1669
smart people	0.245059	1587
really	0.244595	1584
opportunity	0.242279	1569
products	0.199197	1290
awesome	0.168159	1089

Table 4: Sampling of Term Occurrences from Negative Reviews

Negative Review Terms	Relative Frequencies	Absolute Frequencies
employees	0.490435	1205
job	0.455026	1118
balance	0.290598	714
life balance	0.267399	657
little	0.229141	563
amazon	0.225885	555
know	0.212047	521
poor	0.212047	521
hr	0.182743	449
retail	0.173382	426
performance	0.14652	360
worked	0.134717	331

3.2 Predictive Analysis

One of the major goals of this research is to create a system that can predict a target metric for an organization. For example, an organization may wish to predict whether an employee is satisfied with his/her company or not. We also want to examine which factors are most closely associated with overall satisfaction or dissatisfaction. The system is trained by using the data from the employee review dataset. We will use several machine learning algorithms. After the training process is complete, the system will rank the algorithms by the model that performs the best. The top performing algorithm will then be used for future predictions. The overall workflow for this process is shown in Figure 6.

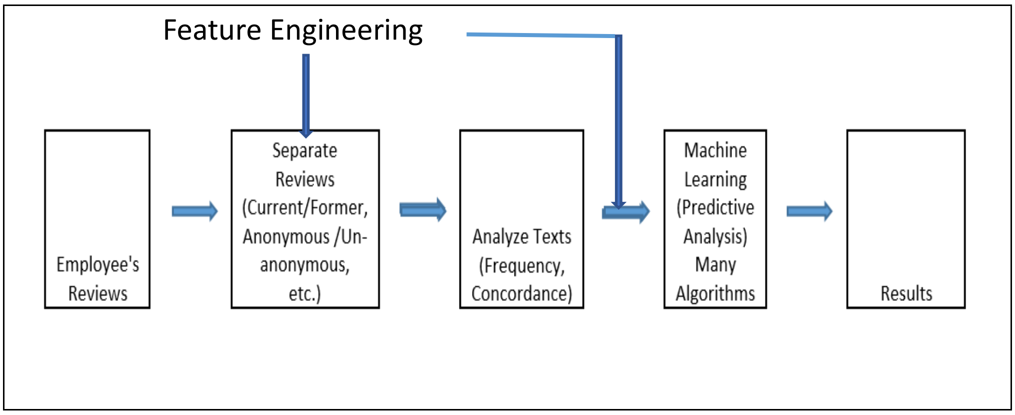

Feature Engineering
Employee's Reviews
Separate Reviews (Current/Former, Anonymous /Un-anonymous, etc.)
Analyze Texts (Frequency, Concordance)
Machine Learning (Predictive Analysis) Many Algorithms
Results

Figure 6: Workflow Diagram

There are 15 features in the employee review data set, including short summary, pros, cons, work/life balance, culture and values, career opportunities, and senior management. “Overall rating” is our target feature that we want to predict. It indicates whether employees are satisfied with their company or not. The rating values are between 1 and 5 and indicate, from low to high, the degree to which the employee is satisfied with his or her company.

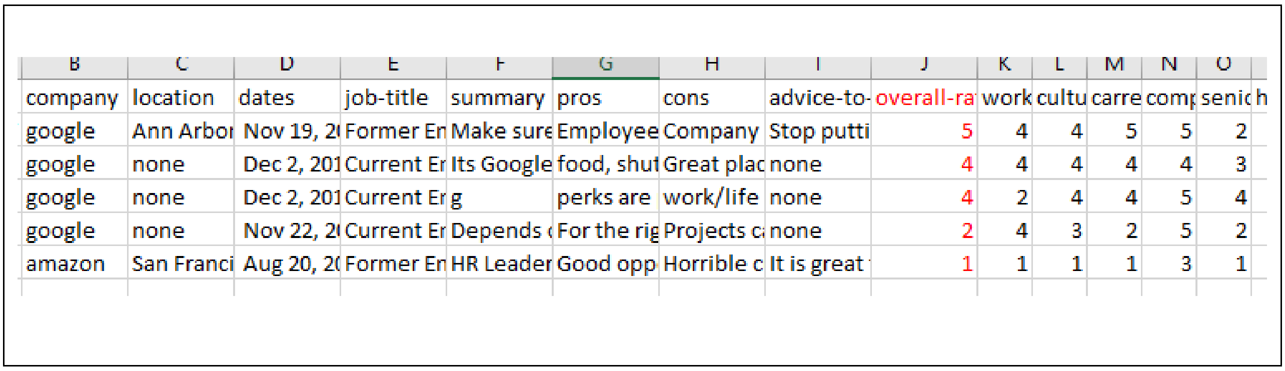
Data mining methods are categorized as either supervised (a specific target variable selected for analysis) or unsupervised (no specific target variable) (Larose and Larose, 2014). Since we know our target for prediction, we are interested in supervised learning. Supervised learning is the most common data mining method. Regression, decision trees, neural networks and support vector machines are all supervised learning algorithms.

Tools and Processes

Recall that there are five phases in CRISP-DM: 1) Business Understanding, 2) Data Understanding, 3) Data Preparation, 4) Modeling, 5) Evaluation and 6) Deployment. We are now entering into the third and fourth phases of this process. The data preparation phase is where we preprocess the data (data cleaning and preparation) and the modeling phase is where we train the model or algorithm.

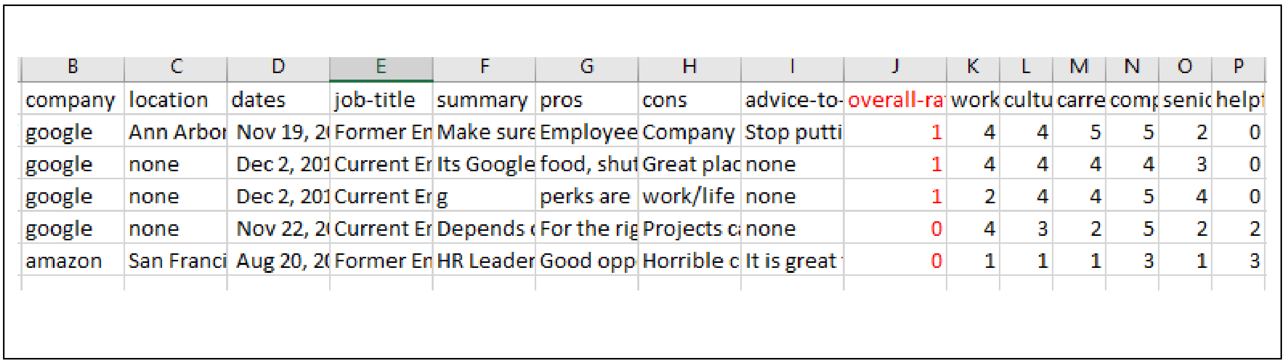
(1) **The preprocessing stage:** The data preparation stage is critical in preparing the data for modeling and analysis. We need to ensure a clean and valid dataset. For example, we need to remove the data that will not be analyzed. In this dataset, we removed the reviews from Netflix due to a low number of employee responses in comparison to the other companies. Other data cleaning and preparation tasks are based on what we want the system to predict and thus which features should be included or removed. For the data cleaning stage, we mainly use spreadsheet software to prepare our data.

For the top algorithms presented, if the prediction performance measures are not high enough, these algorithms will not predict accurately in the holdout sample. If the performance is lower than our acceptable threshold, we will make changes to the dataset such as regrouping the data for clarity, and resubmit the data to the algorithm. For example, if the target feature is “overall ratings,” which consists of 5 different values (1-5), it is more difficult to make good predictions than if we try to predict whether the employees are satisfied or not (yes or no). This is why we change the values in the target feature by grouping the reviews that have “overall ratings” 1 and 2 as “unsatisfied” and 3 through 5 as “satisfied. The data with the original and modified rankings are shown in Figure 7a and Figure 7b, respectively (compare column J in the two figures).



B	C	D	E	F	G	H	I	J	K	L	M	N	O
company	location	dates	job-title	summary	pros	cons	advice-to-	overall-ra	work	cultu	carre	comp	senic h
google	Ann Arbor	Nov 19, 2(	Former En	Make sure	Employee	Company	Stop putti	5	4	4	5	5	2
google	none	Dec 2, 201	Current Er	Its Google	food, shut	Great plac	none	4	4	4	4	4	3
google	none	Dec 2, 201	Current Er	g	perks are	work/life	none	4	2	4	4	5	4
google	none	Nov 22, 2(	Current Er	Depends (	For the rig	Projects ca	none	2	4	3	2	5	2
amazon	San Franci	Aug 20, 2(	Former En	HR Leader	Good opp	Horrible c	It is great	1	1	1	1	3	1

Figure 7a. Original Dataset with Overall Ratings between 1 and 5 (See Column J)



B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
company	location	dates	job-title	summary	pros	cons	advice-to-	overall-ra	work	cultu	carre	comp	senic	helpf
google	Ann Arbor	Nov 19, 2(	Former En	Make sure	Employee	Company	Stop putti	1	4	4	5	5	2	0
google	none	Dec 2, 201	Current Er	Its Google	food, shut	Great plac	none	1	4	4	4	4	3	0
google	none	Dec 2, 201	Current Er	g	perks are	work/life	none	1	2	4	4	5	4	0
google	none	Nov 22, 2(	Current Er	Depends (	For the rig	Projects ca	none	0	4	3	2	5	2	2
amazon	San Franci	Aug 20, 2(	Former En	HR Leader	Good opp	Horrible c	It is great	0	1	1	1	3	1	3

Figure 7b. The Modified Dataset after Changing Overall Ratings “1-2” To “0” And “3-5” To “1” (See Column J)

This modification of the data helps to improve the system performance a great deal. Intuitively, the algorithms we consider treat the number of stars as a categorial variable, rather than as an interval variable. That is, the algorithms essentially maximize precision, so a prediction of four when the true value is five is treated as if it were as bad as a prediction of one or two when the true value is five. Thus, if we do not transform our target feature, then the algorithms penalize “near misses” too much. Aggregating categories largely solves this problem. However, future work will look at algorithms which treat our target variables as interval variables, so we will not have to aggregate categories. The system performance information is discussed in Section 5.1.

(2) Training of the model: The modeling phase uses the cleaned data as the input file for the system to learn how to predict the overall "satisfied" and "dissatisfied" categories in a training data set. The resulting predictions are then evaluated on a holdout data set. Once a set of good models is obtained, they are used to determine which features are most useful in predicting satisfaction. It is also possible to use the models to predict the likely overall ratings for reviews that do not actually have an overall rating. During this stage, the machine learning tool, DataRobot, is used.

DataRobot (https://www.datarobot.com/) is a supervised automated machine learning tool implemented by several successful data scientists. It analyzes a data set for a target feature and runs the data using several algorithms. The best algorithm for predicting the target correctly with acceptable speed will be recommended for use for future predictive analysis purposes. The outputs of the other algorithms are also presented in case the user prefers using them for some purpose.

After the data set is cleaned and ready for the training phase, the data set is uploaded to DataRobot. The target feature (i.e., overall rating) is then indicated. DataRobot then uses part of the data set to train the system and identify which algorithms are most promising (i.e., can best predict the target feature). The algorithms that perform well then use more data to train until they successfully fit an acceptable amount of the training data (such as 80%). The top algorithms are then presented to the user.

4. Results and System Performance

The last phase of the CRISP-DM methodology is evaluation. Here we evaluate the performance of the algorithms. In our case, DataRobot ranked the three algorithms with the highest performance as: (1) the Nystroem Kernel SVM Classifier, (2) the eXtreme Gradient Boosted Trees Classifier with Early Stopping, and (3) the Auto-Tuned Word N-Gram Text Modeler using Token Occurrences - Cons. The best ranked algorithm for target prediction with high performance and high speed was the Nystroem Kernel SVM Classifier. The Nystroem Kernel SVM Classifier is a kernelized learning algorithm used in support vector machine classification (Chang et al., 2010).

4.1 Performance Evaluation

In general, prediction performance is measured by the AUC and log-loss measures. The AUC is the area under the (ROC) curve, where the ROC (receiver operating characteristic) curve is the curve that illustrates the tradeoff between false positives (on the horizontal axis) and true positives (on the vertical axis), for different values of the cutoff threshold. For a more demanding threshold, the system will incorrectly classify a smaller number of truly negative individuals as positive, but will also correctly identify a smaller number of truly positive individuals as positive. The ROC curve is therefore upward sloped. A value of the AUC close to one means that one can identify a large percentage of the true positives, while suffering only a small number of false positives. In this study, a "positive" indicates the individual is satisfied, overall, with the job, in the sense that their overall star rating is a 3, 4 or 5. The log loss measure, on the other hand, is simply the negative of the log likelihood function, so a small value of the log loss indicates a higher value of the likelihood function, and so, a better fit to the data.

Table 5 displays the employee job satisfaction prediction accuracy rates by algorithm, for the top three algorithms. The best performing algorithm according to both measures was the Nystroem Kernel SVM Classifier, followed by the eXtreme Gradient Boosted Trees Classifier with Early Stopping and the Auto-Tuned Word N-Gram Text Modeler using Token Occurrences – Cons. It also indicates the overall number of false negatives (FN), true negatives (TN), false positives (FP), and true positives (TP) in the holdout sample. Note, here that the percentage of true positives correctly identified in the holdout sample is 100 x [11239/(11239 + 300)] = 97.4%. Similarly, the percentage of true negatives correctly identified is 72.5%.

Table 5: Prediction Accuracy Rates by Algorithm for the 12/345 Stars Rating Classification

Algorithms (12_345)	**AUC**		**LogLoss**		**Confusion Matrix**			
	Validation	**Cross Val.**	**Validation**	**Cross Val.**	**FN**	**TN**	**FP**	**TP**
Nystroem Kernel SVM Classifier	0.9657	0.9658	0.1513	0.1510	300	1,298	491	11,239
eXtreme Gradient Boosted Trees Classifier with Early Stopping	0.9645	0.9646	0.1536	0.1533	227	1,195	594	11,312
Auto-Tuned Word N-Gram Text Modeler using Token Occurrences - Cons	0.8615	0.8595	0.2768	0.2777	268	610	1179	11,271

The DataRobot software also indicates which features were most important in each algorithm. These rankings are shown in Table 6. According to the Nystroem Kernel SVM, the top two features were number of stars for “career opportunities” and “culture values,” followed by the text categories “summary” and “cons.” The eXtreme Gradient Boosted Trees Classifier had a similar ranking, but with “culture values” and “career opportunities” switched. The Auto-Tuned Word N-Gram Text Modeler only used the “cons” feature, and so, is not included in Table 6.

Table 6: Feature Impact for each Algorithm

Algorithms (12_345)		
Nystroem Kernel SVM Classifier		
	Features	**%**
1	carrer_opportunities_stars	100
2	culture_values_stars	98.2
3	summary	62.73
4	cons	55.27
5	senior_mangemnet_stars	35.74
6	work_balance_stars	32.65
7	helpful_count	31.72
8	pros	23.92
9	comp_benefit_stars	18.97
eXtreme Gradient Boosted Trees Classifier with Early Stopping		
	Features	**%**
1	culture_values_stars	100
2	carrer_opportunities_stars	83.83
3	summary	80.66
4	cons	78.83
5	senior_mangemnet_stars	62.1
6	helpful_count	49.28
7	work_balance_stars	48.74
8	pros	39.22
9	comp_benefit_stars	20.51

4.2 Robustness

To check for the robustness of our results, we also reran our data, but with the target feature indicating that the employee was dissatisfied if the employee chose one, two or three stars and satisfied only if they chose four or five stars. Thus, more reviews were included in the dissatisfied category. The results are presented in Table 7, where again, FN, TN, FP and TP are for the holdout sample. The Nystroem Kernel SVM was again the best performing algorithm, but the Auto-Tuned Word N-Gram Text Modeler using Token Occurrences – Cons and the Auto-Tuned Word N-Gram Text Modeler using Token Occurrences – Summary algorithms were the two next best algorithms.

Table 7: Prediction Accuracy Rates by Algorithm for the 123/45 Stars Rating Classification

Algorithms (123_45)	**AUC**		**LogLoss**		**Confusion Matrix**			
	Validation	**Cross Val.**	**Validation**	**Cross Val.**	**FN**	**TN**	**FP**	**TP**
Nystroem Kernel SVM Classifier	0.9453	0.9463	0.2754	0.2728	702	3,433	857	8,336
Auto-Tuned Word N-Gram Text Modeler using token occurrences - cons	0.8105	0.8152	0.4806	0.4761	664	1,944	2,346	8,374
Auto-Tuned Word N-Gram Text Modeler using token occurrences - summary	0.4907	0.4934	0.7997	0.7965	451	1,466	2,824	8,587

Again, the Auto-Tuned Word N-Gram Text Modeler using Token Occurrences algorithms only use one feature each, i.e., the text categories “cons” and “summary,” respectively, so Table 8 only shows feature rankings for the Nystroem Kernel SVM. The rankings of the first six features are identical to those above, though the last three features are different when predicting 1 through 3 stars versus 4 or 5 stars.

Table 8: Feature Impact of Each Algorithm

Algorithms (123_45)		
Nystroem Kernel SVM Classifier		
	Features	**%**
1	carrer_opportunities_stars	100
2	culture_values_stars	96.4
3	summary	69.95
4	cons	52.96
5	senior_mangemnet_stars	51.53
6	work_balance_stars	49.76
7	comp_benefit_stars	30.7
8	pros	26.99
9	advice_to_mgmt	10.85

Note, that the features “Cons,” “Summary,” “Pros,” and “Advice to Management” are in a textual format, “Helpful count” is the number of readers that voted the comments from the reviewers helpful to them, and “stars” categories consist of numerical scores (1-5 stars).

In summary, the Nystroem Kernel SVM Classifier best predicted overall satisfied or dissatisfied employees, and the features "Career opportunities" and "Culture values," are the most important features helping the system to predict the target feature correctly, and thus have high predictive power (Burkov, 2019).

It is also worth noting that system performance was better when predicting the 12/345 breakdown of job satisfaction than when predicting the 123/45. This suggests that the three-star reviews are more appropriately classified with the four and five-star reviews than with the one and two-star reviews. For example, employees may use three stars as indicating a satisfactory work environment, like four and five stars, while they may reserve one and two stars to express extreme dissatisfaction.

5. Discussion

By using machine learning techniques to perform predictive analysis on employees' reviews, we are able to predict with high accuracy whether employees are satisfied or dissatisfied with their jobs. Thus, we are able to answer our research questions as follows:

RQ 1: Can we use machine learning techniques to predict employee job satisfaction from employees' reviews?

The AUC and Log loss as well as true versus false positives and negatives indicate that using machine learning techniques, the system is able to predict employee job satisfaction very well.

In addition to using the system evaluation using AUC and Log loss, we also evaluated the system by predicting a set of data consisting of 80 reviews that were not part of the original training set. We removed the overall satisfaction values and let the systems predict them. We compared the results created by the system with the real overall satisfaction values given by the reviewers. The system was able to predict whether those employees were satisfied with their companies or not with an accuracy rate of about 95%. Thus, we believe that the high accuracy rate is largely because the dataset is very clean and many of the features are themselves in terms of numerical scores.

Thus, we can conclude that it is possible to predict employee job satisfaction using machine learning techniques.

RQ 2: Using machine learning techniques, which algorithms perform best in predicting employee job satisfaction?

Using machine learning techniques, the system was able to predict the target feature with high accuracy rates. The most successful algorithms are: (1) the Nystroem Kernel SVM Classifier, (2) the eXtreme Gradient Boosted Trees Classifier with Early Stopping, and (3) the Auto-Tuned Word N-Gram Text Modeler using Token Occurrences – Cons.

RQ 3: Using machine learning techniques to analyze employee reviews, which features have the highest predictive power in helping the system to predict accurately?

Most features in this dataset contribute highly to the predictive analysis, since the dataset is fairly clean. However, the top two features that provided the highest prediction impact (predictive power) are "career opportunities" and "culture value." This means that, in order to predict whether an employee is satisfied with his/her company, analyzing the "career opportunities" and "culture value" provides more predictive power than other features.

One interesting thing we found in the results is that the contents of the textual "summary" feature has more impact on the prediction accuracy than the contents of the "cons" and "pros" features. We believe that the terms used in the "summary" feature must be more distinctive than those used in the other two features (the words used must contain more information).

Business Implications

In order for businesses to perform well, they need employees who are satisfied with the company. Knowing whether employees are satisfied and knowing the major indicators for satisfaction will help businesses to obtain the optimal performance from their employees. Thus, our research findings can contribute to several human resource management tasks including:

- Enhancing recruitment of strong employees. If a company knows what makes employees satisfied, it can improve the company's culture to incorporate a more satisfying workplace. As a result, the work environment can attract strong prospective employees.
- Predicting employees' retention (staying with the company) – Turnover is very costly. If employers can identify dissatisfaction in employees and rectify the situation, this may lead to greater employee retention.
- Improve the employer/employee relationship – Our study finds that "career opportunities" and "culture values" lead to more satisfied employees. More satisfied employees may be more prone to feel good about their position with the company. This positive relationship may encourage employees to perform well at work.
- Applying the algorithm to other areas of business – Machine learning tools such as DataRobot, and predictive models can be applied to other areas such as analyzing customer satisfaction to predict churn.

6. Conclusion and Future Research

In this paper, we demonstrated that it is possible to use supervised machine learning techniques to analyze and predict employee job satisfaction with high rates of accuracy. The best performed algorithm is Nystroem Kernel SVM Classifier with an accuracy rate of more than 96% according to the AUC-measure. In addition, the system can find the most important factors that provide a high impact on predicting job satisfaction.

In our study, a career opportunities and company's work culture contribute greatly to predicting employee job satisfaction.

Building on what we have found from this study, we plan to extend this research in the following areas:

- Incorporate more semantic and linguistic analyses using terms and their relations (words/phrases, lexical semantic relations, etc.) to improve system performance. E.g., how exactly does the text in the "summary" and "cons" features help to predict job satisfaction?
- Apply similar techniques to other application domains such as supply chains, health care, and security.
- Compare employee job satisfaction across similar businesses, such as Apple vs. Microsoft. From the current dataset, we find that the overall ratings for both companies are similar. However, by using semantic analysis to examine the textual comments in the employee reviews (pros, cons, and summaries) we can identify more specifically what employees think about their companies.

References

1. 2017 Retention Report: Trends, Reasons & Recommendations: http://info.workinstitute.com/retentionreport2017 (accessed on January 1, 2021).

2. Abraham, R. (1999). The Impact of Emotional Dissonance on Organizational Commitment and Intention to Turnover. *Journal of Psychology*, Vol. 133, No. 4, 441–455.

3. Alao, D. & Adeyemo, A. (2013). Analyzing employee attrition using decision tree algorithms. Comput. *Information System, Development Informatics and Allied Research Journal*. Vol. 4, No. 1, 17–28.

4. Alduayj, S. S. & Rajpoot K. (2018). Predicting Employee Attrition using Machine Learning. *International Conference on Innovations in Information Technology (IIT)*, 93-98.

5. Azari B. (2011). Job Satisfaction: A Literature Review. *Management Research and Practice,* Vol. 3, No. 4, 77-86. http://mrp.ase.ro/no34/f7.pdf.

6. Bendemra, H. (2019). Building an Employee Churn Model in Python to Develop a Strategic Retention Plan. https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d (accessed on January 1, 2021).

7. Bolden-Barrett, V. (2017). Turnover costs employers $15,000 per worker. https://www.hrdive.com/news/study-turnover-costs-employers-15000-per-worker/449142/ (accessed on January 1, 2021).

8. Burkov, A. (2019). *The Hundred-Page Machine Learning Book*, ISBN-13: 978-1999579500.

9. Chen, Ling-Hsiu (2008). "Job satisfaction among information system (IS) personnel," *Computers in Human Behavior*, Volume 24, Issue 1, 2008, Pages 105-118, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2007.01.012.

10. Conlon, S. J. (2021). Why Do Data Scientists Want to Change Jobs: Using Machine Learning Techniques to Analyze Employees' Intentions in Switching Jobs. *INTERNATIONAL JOURNAL OF MANAGEMENT & INFORMATION TECHNOLOGY*, *16*, 59–71. https://doi.org/10.24297/ijmit.v16i.9058.

11. Cotton, J. L.& Tuttle J. M. (1986). Employee turnover: A meta-analysis and review with implications for research. Academy *of management Review*, Vol. 11, No. 1, 55-70. https://doi.org/10.2307/258331.

12. Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020). "Predicting Employee Attrition Using Machine Learning Techniques." *Computers 9(4)*, 86. https://doi.org/10.3390/computers9040086.

13. Galup S, Klein G, Jiang J. (2008). The impacts of job characteristics on IS employee satisfaction: A comparison between permanent and temporary employees. *Journal of Computer Information Systems*, Vol. 48 No. 4, 58–68.

14. Golestani, A., Masli, M., Shami, N. S., Jones, J., Menon A., & Mondal J. (2018). Real-Time Prediction of Employee Engagement Using Social Media and Text Mining. *17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL, 2018, 1383-1387.

15. Harden, G., Boakye, K. G., & Ryan, S. (2018). Turnover intention of technology professionals: A social exchange theory perspective. *Journal of Computer Information Systems, 58*(4), 291–300. https://doi.org/10.1080/08874417.2016.1236356.

16. Hinkin, T. R. & Tracey, J. B. (2000). The cost of turnover: Putting a price on the learning curve. *Cornell Hotel and Restaurant Administration Quarterly, Vol.* 41, 14-21. https://hdl.handle.net/1813/71741.

17. Hinkin, T. R. & Tracey, J. B. (2006). Development and use of a web-based tool to measure the costs of employee turnover: Preliminary findings. Ithaca, NY: Cornell University School of Hotel Administration Center for Hospitality Research.

18. Holtom, B., Mitchell, T., Lee, T., & Eberly, M. (2008). Turnover and retention research: A glance at the past, a closer review of the present, and a venture into the future. *The Academy of Management Annals*, Vol. 2, No. 1, 231-274. https://doi.org/10.1080/19416520802211552.

19. Hulin, C. L., & Judge, T. A. (2003). Job attitudes. In W. C. Borman, D. R. Ligen, & R. J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology,* Hoboken, NJ: Wiley. 255-276.

20. IBM HR Analytics Employee Attrition & Performance: IBM HR Analytics Employee Attrition & Performance | Kaggle (accessed on January 1, 2021).

21. IBM Knowledge Center https://www.ibm.com/support/knowledgecenter/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html (accessed on January 1, 2021).

22. Jain D., Makkar S., Jindal L., Gupta M. (2021) Uncovering Employee Job Satisfaction Using Machine Learning: A Case Study of Om Logistics Ltd. In: Gupta D., Khanna A., Bhattacharyya S., Hassanien A., Anand S., Jaiswal A. (eds) International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, Vol 1166. Springer, Singapore. https://doi.org/10.1007/978-981-15-5148-2_33

23. Larose, D. & Larose, C. (2014). *Discovering knowledge in data: An introduction to data mining*. John Wiley and Sons, Inc.

24. Lee, T., Mitchell, T., Sablynski, C., Burton, J., & Holtom B. (2004). The effect of job embeddedness on organizational citizenship, job performance, volitional absences and voluntary turnover. *Academy of Management Journal*, Vol. 47, No. 5, 711-22. https://doi.org/10.2307/20159613.

25. Locke, E.A. (1976). The nature and causes of job satisfaction. In M.D. Dunnette (Ed.), *Handbook of industrial and organizational psychology,* Chicago: Rand McNally. Vol. 1, 1297-1343.

26. Mishra, P.K. (2013). Job Satisfaction. *Journal of Humanities and Social Science,* Vol. 14, No. 5 (Sep. - Oct. 2013), 45-54.

27. Mobley, W. H. (1982). *Employee turnover: Causes, consequences, and control,* Reading, MA: Addison Wesley.

28. Moorman, R.H. (1993). The influence of cognitive and affective based job satisfaction measures on the relationship between satisfaction and organizational citizenship behavior. Vol. 46, No. 6, 759–776. https://doi.org/10.1177/001872679304600604.

29. Morrell, K., Loan-Clarke, J., & Wilkinson, A. (2001). Unweaving leaving: The use of models in the management of employee turnover. *International Journal of Management Reviews,* Vol. 3, 219-44. https://doi.org/10.1111/1468-2370.00065.

30. Morris, M. & V. Venkatesh. "Job Characteristics and Job Satisfaction: Understanding the Role of Enterprise Resource." *MIS Q.* 34 (2010): 143-161. https://doi.org/10.2307/20721418

31. Mossholder, K., Sutton, R. P., & Henagan, S. C. (2005). A relational perspective on turnover: Examining structural, attitudinal, and behavioral predictors. *Academy of Management Journal,* Vol. 48, No. 4, 607-18. https://doi.org/10.2307/20159682.

32. Mowday, R. T., Porter, L. W., & Steers R. M. (1982). *Employee-organization linkages: The psychology of commitment, absenteeism, and turnover*, New York: Academic Press.

33. Nagadevara, V. (2008). Early Prediction of Employee Attrition in Software Companies-Application of Data Mining Techniques. Research & Practice in Human Resource Management. 16, 2020–2032. https://doi.org/10.1109/IADCC.2018.8692137.

34. Punnoose, R. & Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. *International Journal of Advanced Research in Artificial Intelligence,* Vol. 5, Issue 9. 22–26. https://doi.org/10.14569/IJARAI.2016.050904.

35. Ray, A. N. & Sanyal, J. (2019). Machine Learning Based Attrition Prediction. *2019 Global Conference for Advancement in Technology (GCAT)*, BANGALURU, India, 1-4.

36. Saari, Lise. M. & Judge, T. (2004) Employee Attitudes and Job Satisfaction. *Human Resource Management*, Winter 2004, Vol. 43, No. 4, 395–407. https://doi.org/10.1002/hrm.20032.

37. Sacco, J. M. & Schmitt, N. (2005). A dynamic multilevel model of demographic diversity and misfit effects. *Journal of Applied Psychology*, Vol. 90, No. 2, 203-231. https://doi.org/10.1037/0021-9010.90.2.203.

38. Singh, J.K. & Jain, M. (2013). A Study of Employees' Job Satisfaction and Its Impact of their performance. *Journal of Indian Research* Vol. 1, No. 4, 105-111.

39. Spector, P.E. (1997). *Job satisfaction: Application, assessment, causes and consequences. Thousand Oaks*, CA: SAGE.

40. Sumner, M. & Niederman, F. (2004). The impact of gender differences on job satisfaction, job turnover, and career experiences of information systems professionals. *Journal of Computer Information Systems*, 44: 29–39. https://doi.org/10.1145/512360.512395.

41. Thompson, E.R. & Phua F.T.T. (2012). A Brief Index of Affective Job Satisfaction. *Group & Organization Management*, Vol. 37, No. 3, 275–307. https://doi.org/10.1177/1059601111434201'

42. Tracey, J. B., & Hinkin, T. R. (2010). Contextual factors and cost profiles associated with employee turnover. In C. Enz (Ed.), *The Cornell School of Hotel Administration handbook of applied hospitality strategy,* Los Angeles, CA: SAGE. 736-753.

43. Vidal, M.E.S., Valle, R.S., & Aragón, B.M.I. (2007). Antecedents of repatriates' job satisfaction and its influence on turnover intentions: Evidence from Spanish repatriated managers. *Journal of Business Research*, Vol. 60, 1272-1281. https://doi.org/10.1016/j.jbusres.2007.05.004.