Employee Satisfaction at Tata Motors

Elaine Chua, Yuta Namba, Udita Shah, Abigail Siy

Northeastern University: Khoury College of Computer Sciences

Abstract

This study analyzes employee satisfaction at Tata Motors using a dataset of 10,752 observations. Leveraging sentiment analysis and a Random Forest Regressor model, it identifies key factors influencing job satisfaction. Results highlight the significance of work satisfaction, salary, benefits, and sentiment in predicting employee contentment. Findings offer actionable insights for Tata Motors to improve its work culture and extend implications for HR decisions in fostering inclusivity.

Introduction

Define the Problem

Tata Motors, a manufacturer headquartered in India is a subsidiary of the renowned Tata Group, one of India's largest and most reputable conglomerates. The company has gained recognition for its large lineup of vehicles. With a commitment to innovation, sustainability and delivering high quality products Tata Motors has left an imprint on the automotive landscape. Over the past year, Tata Motors has taken excellent care of their employees. Their permanent employee base has risen around 11%, and in 2016, job satisfaction rates were said to have increased by 8%. However, while they state that gender diversity is present within their organization, statistics show that only 23% of Tata Motors employees are women and the rest are men. While gender is not disclosed in the dataset, we hope that we can showcase to women that Tata Motors will provide them with an inclusive work environment.

Motivation

On our first meeting, we began discussing the diversity of each group member's cultural backgrounds, making us curious about the variations in work life across different cultures. And so, we decided to focus on a dataset pertaining to an Indian company, aligning with one of our member's cultural backgrounds. This project is significant as it can help Tata Motors and other companies to understand their employees' sentiments, leading to a more positive and productive work environment. The analysis can guide HR decisions, improve company culture, and attract top talent. Additionally, it serves as a valuable resource for data enthusiasts, NLP researchers, and AI developers interested in sentiment analysis and culture analytics.

Goal and Objectives

In this project, we will be exploring the Tata Motors Employee Reviews dataset. This dataset provides individual records on an employee's job title, location, department, job satisfaction rating, work life balance rating, skill development rating, salary and benefits rating, job security rating, career growth rating, work satisfaction rating, and their overall review of the company. We aim to understand the overall experiences of employees at Tata Motors, identify key drivers of employee satisfaction, and provide actionable recommendations to improve the work environment at Tata Motors. The goal of this project is to predict an employee's overall job satisfaction rating based on their other ratings for contributing factors and their job review.

Methodology

Data Acquisition

We obtained the data from Kaggle from a dataset titled "Tata Motors Employee Reviews". It has 10,752 observations. This dataset particularly stood out to us because not only does it have quantitative data that gives insight to an employee's experience, but it also has qualitative data about their experience in the form of a review. We thought it would be interesting to perform sentiment analysis on the reviews and include sentiment polarity scores as a feature in our model. While an employee's numerical ratings serve as reliable indicators of their overall satisfaction, we wanted to explore whether their written reviews could also be predictive of their overall satisfaction. As no dataset is perfect, we recognize that there are some limitations with the data. There are some reviews that are not in English and others that are evidently translated into English. As a result, some sentiment polarity scores may not reflect the intended meaning of the review.

Data Preparation

After loading the dataset into a pandas data frame, we began the initial data preparation and exploratory data analysis (EDA) steps. Although there was a large presence of missing values, we determined that we could appropriately handle them without skewing the distribution of the data. First, the rows that were missing more than half of their columns were removed. Next, since the missing values in the numerical rating columns accounted for less than 1% of the dataset, we performed imputation by replacing the missing values with the mode. Next, we imputed the review columns, Likes and Dislikes, with "no comment". With the remaining columns (Title, Place, Job Type, Department), we replaced the missing values with "unknown". While the missing values in those columns constituted a large percentage of the data, this approach was suitable since we chose not to use those columns as features in our model. To simplify the data, we split the Place column into two columns, City and State, and then dropped the original Place column. The majority of the data contained City values, however, a substantial portion of the corresponding State values were missing. For the purposes of our analysis, we determined the State column to be redundant and therefore dropped it. After handling all of the missing values, we processed the reviews for sentiment analysis. We removed any whitespace, emojis, punctuation, and converted everything to lowercase. Once all of the reviews were cleaned, we used the SentimentIntensityAnalyzer function from the nltk.sentiment library to compute sentiment polarity scores for each review. The polarity_scores method computes a score for each review by analyzing the words within the review. A float is returned with either a positive or negative number. A higher value in either direction indicates a more positive or negative review.

Model Selection

The final model that we selected to make our predictions was the Random Forest Regressor. We chose this model for several reasons, the first being, that our target variable was continuous as opposed to categorical. The random forest regressor also supports the use of a feature importance plot through the feature_importances_ method. We thought that incorporating this plot would assist us in making recommendations as the goal of our project is to advise Tata on areas within the company that can be enhanced to maximize employee satisfaction. Lastly, the random forest regressor excels in effectively managing datasets with a large presence of missing values. Since multiple decision trees, rather than just a couple, are constructed and utilized to make predictions, the model is less sensitive to the patterns created by missing values. This is because different trees are built using different subsets of the data.

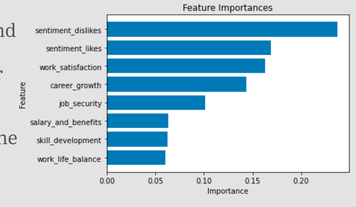
To predict the overall rating, we selected the following as features to train the model: work life balance, skill development, salary and benefits, job security, career growth, work satisfaction, and the sentiment scores for likes and dislikes. We excluded the categorical features from our model due to their imbalanced distributions, and utilizing one-hot encoding did not seem appropriate. The data was partitioned into 80% training data and 20% test data. We ensured that the results were reproducible, and the distribution of the data was preserved. In order to evaluate the performance of the model, we used the mean squared error (MSE). We also hyper tuned the model to achieve the lowest MSE. We used the n_estimators parameter to tune the model, which determines the number of decision trees that the regressor will build. A concern with our model is that we got the lowest MSE when n_estimators = 200. While increasing the number of trees generally improves the performance of the model, it also increases the risk of overfitting. The model might learn and fit too closely to the training data, making it difficult to predict on new, unseen data.

Results and Evaluation

When evaluating the performance of three regression algorithms—kNN regressor, Random Forest regressor, and Support Vector Machines (SVM)—in predicting 'overall_rating' with various features, significant findings were drawn regarding their efficacy.

The kNN regressor displayed a mean square error of 0.92. After tuning the model, its ideal parameter setting was n_neighbors set to 7, obtaining a score of 0.86. Similarly, the Random Forest regressor resulted in a mean square error of 0.88. When the model was tuned, the best score was 0.86 with its ideal parameters set with n_estimators at 200. Both models perform equally well when set to their ideal parameters, obtaining the same score of 0.86.

The feature importance scores highlight how different features affect the prediction of 'overall_rating'. 'Sentiment_dislikes' has the most significant impact, scoring over 0.2, followed closely by 'sentiment_likes' and 'work_satisfaction' at scores above 0.15. 'Career_growth' comes close to 0.15, while 'job_security' scores above 0.10. 'Salary_and_benefits', 'skill_development', and 'work_life_balance' obtained scores higher than 0.05, indicating their relatively lesser impact on predicting the overall rating.



The Support Vector Machine (SVM) model has a mean square error of 0.88. Once the model was tuned, the best 'C' value was 1 and the best 'gamma' value was 0, but the best score obtained was 1.81, which is unusually high.

The Random Forest regressor initially had a slightly lower mean square error (MSE) of 0.88 compared to the kNN regressor's 0.92 before tuning. After tuning, both models had the same ideal score of 0.86. Since the Random Forest Regressor showed a lower MSE before tuning, the Random Forest regressor is the suitable model.

The Impacts

The implementation of our model stands to significantly influence the overall work environment at Tata Motors. Our insights into the sentiments and experiences of employees offer a pathway to enhance the work environment by pinpointing strengths and weaknesses in the company's culture. Presently, our results indicate a prevalence of negative comments related to work culture. By analyzing the importance employees attribute to various features like work satisfaction, salary, benefits, job security, and so on; we can provide valuable recommendations for improvement, ultimately fostering a positive work culture and boosting employee satisfaction. The impact of our project extends beyond Tata Motors, presenting itself as a valuable resource for other companies seeking to understand their employees' sentiments. The analysis we provide can inform HR decisions, foster cultural improvement and facilitate the attraction of top talent. Our project aligns with the evolving perspective on demographic criteria in machine learning, emphasizing the practitioner's responsibility to address disparities and champion inclusivity.

Conclusion

Our analysis of employee satisfaction at Tata Motors revealed significant predictors of workplace contentment: work satisfaction, salary, benefits, and sentiment analysis. These insights provide actionable recommendations for Tata Motors to enhance its work culture and have broader implications for HR decisions aimed at fostering inclusivity. While this study represents a significant step towards understanding employee sentiments at Tata Motors, there remain opportunities for further exploration, such as delving into the nuances of cultural influences on job satisfaction and refining models for even more accurate predictions. The project's outcomes extend beyond a singular company, offering a framework for other organizations seeking to improve employee satisfaction and foster a positive work environment. Overall, this analysis not only serves as a valuable resource for Tata Motors but also contributes to the larger discourse on leveraging data-driven insights to create inclusive and satisfying work cultures.

Source: *Tata Motors Employee Reviews Dataset* by MANISH KUMAR (Link: *https://www.kaggle.com/datasets/manishkr1754/tata-motors-employee-reviews-dataset*)

December 6th, 2023