Labor Turnover Prediction

Fatima Abdul Rauf
BSCS Student at FCCU

Menahil Baig
BSCS Student at FCCU

Abstract—**Supervised learning algorithms are used to predict labor turnover in a large organization. The relationship between turnover and numerous factors like salary, satisfaction levels, working hours, etc, are explored in detail using data mining techniques and K Nearest Neighbors (KNN) and K-Means Clustering. The KNN model has high accuracy and precision as measured through the Confusion Matrix and Yellowbrick. A negative relationship exists between salary levels, time spent at the company, and employee turnover. A negative relationship does not exist between satisfaction levels, evaluations by the employer, and labor turnover.**

1. Introduction

Employee turnover poses significant costs and human and social capital losses to organizations making it imperative to effectively identify the antecedents of turnover. Employee turnover affects various business functions including negative impact on customer outcomes, financial performance, productivity, and sales and quality outcomes. [1] The antecedents to increased employee turnover have been recognized as low work and organizational commitment which will be further explored in this report.

With the wider applications of machine learning algorithms, HR departments have seen marked improvements in employee retention strategies. ”Supervised machine learning methods—wherein computers learn from analyses of large-scale, historical, labelled datasets—have been shown to garner insights in various fields, like biology and medical sciences, transportation, political science, as well as many other fields”. [2] Consequently, this study is will be using KMeans Clustering for model building and KNN for predicting employee turnover.

1.1. Problem Formulation

Considering the human and capital losses from employee turnover the main assumptions tend to be that increasing turnover rates are dysfunctional. This assumption has been empirically proved by various research. For example, the cost-based perspective suggests the direct and indirect costs associated with an employee exiting while the human capital perspective highlights the loss of valuable knowledge and skills that the employees developed through their experience at the company [1]. To carefully understand and reflect on the various antecedents that lead to dysfunctional turnover at any company. The hypotheses to be tested in this study are as follows.

Hypothesis 1: The relationship between turnover and satisfaction level is negative.

Hypothesis 2: The relationship between employee turnover and the last evaluation is negative.

Hypothesis 3: There is a negative relationship between time spent at the company and turnover.

Hypothesis 4: There is a positive relationship between working hours and turnover.

Hypothesis 5: There is a negative relationship between the level of salary and turnover.

2. Related Work

The study by Hancock et al. expands on the model of collective turnover by Hausknecht and Trevor by recognizing further consequences of employee turnover in different industries and different organization sizes. The study explores the relationship between employee turnover and factors like organizational performance, job level, measure of organizational performance, and size. Firstly, the study used computerized database searches from ISI Web of Science, Business Source Premier, and PsycINFO with keywords like employee turnover, turnover rate(s), organization performance, firm performance, productivity, and financial performance retention. Secondly, a meta-analysis was conducted using a manual search from various premier journals, such as the Journal of Applied Psychology and the Academy of Management Journal.

While the study by Hancock et al. focuses more on how employee turnover and organizational performance change over different organizations depending on size, industry, job levels, occupations, accidents, etc, this research mainly focuses on what factors lead to employee turnover in a large organization and predicts these antecedents.

Hancock et al. independently coded 48 studies on outcome measures and potential moderators for turnover for meta-analysis. They used three meta-analytic methods i.e. effect sizes, moderator analyses, and curvilinearity. On the other hand, machine learning models like K Nearest Neighbors (KNN) and K-Means Clustering were used to predict employee turnover in this study.

It was observed that there is a significant negative relationship between turnover and performance implying that human and social losses from collective turnover significantly outweigh the functional effects of replacing the low performers with candidates offering unique perspectives or preventing human capital stagnation. Moreover, occupation, job level, and industry negatively affected the relationship as was seen for managerial employees. The results demonstrated a stronger and more negative relationship between customer service, quality, and safety measures than labor productivity or financial performance measures. [1]

However, the limited literature on collective turnover restricts the number of available studies and impacts the strength of conclusions and depth of moderator analyses. Further research is needed to explore the effects of reduced turnover rates on organizational performance and to develop nuanced process models. The study also lacks a key differentiation between functional and dysfunctional turnover. [1]

In contrast, the study by Zhao et al. analyses various supervised learning algorithms like decision trees, random forests, etc, and their efficacy in predicting employee turnover in small-, medium-, and large-sized organizations. Where this research focuses primarily on one large-sized organization the sample in Zhao et al. involves ten organizations categorized in terms of their size.

Firstly, Decision Trees are capable of handling missing values and multiple features, along with having an intuitive interpretation. However, this method has a high model variance, and even small input changes can significantly change the tree structure limiting its prediction. Following up on Decision Trees, Random Forests were also used to enhance the basic tree design and reduce the variance while increasing the reliability by averaging the uncorrelated trees built from bootstrapped training sets and randomly selected predictors. Thirdly, Gradient-Boosting Trees were analyzed, which builds trees sequentially correcting the errors of the previous ones. Hence, reducing overfitting and improving prediction efficiency. Fourthly, Extreme Gradient Boosting optimizes the computational speed and model performance using a regularization term to reduce the overfitting effect. Naive Bayes assumes conditional independence among features while learning the joint probability distribution of inputs and computes the maximum posterior probability for predictions. Furthermore, KNN classifies new instances by a majority vote of the closest K data points, it works best with fewer features and struggles with high-dimensional data. Lastly, Linear Discriminant Analysis, Neural Networks, Support Vector Machine, and Logistic Regression were also explored. [2]

For pre-processing, Zhao et al. used median to replace missing numerical values and mode for missing categorical values. Whereas, the dataset used in this study had no missing values. Furthermore, label encoding, normalization, and standardization were performed on the dataset to prepare it for machine learning algorithms. The five evaluation metrics for the evaluation of the algorithms included "...: (1) accuracy (ACC) is defined as the percentage of the correctly classified data by the model; (2) precision (PRC) is defined as the number of true positives divided by the sum of true positives and false positives; (3) recall Employee Turnover Prediction with Machine Learning 745 (RCL) is defined as the number of true positives divided by the sum of true positives and false negatives; (4) F1 is defined as the harmonic mean of precision and recall; and (5) Receiver operating characteristic (ROC) curve is defined as a graphical plot of the tradeoff between precision and recall.." [2]. Whereas, this study only employs ROC and AUC as evaluation metrics for KNN.

The study by Zhao et al. recommends tree-based approaches for medium and large HR datasets as they tend to have the lowest data variance and a more reliable model can be built using extreme gradient boosting due to its greater predictive power and computational speed. However, this study will be using and evaluating the KNN model for predicting employee turnover for a large organization.

3. Exploratory Data Analysis

Pandas, matplotlib, numpy, seaborn, and plotly were used for the EDA. Initially, the dataset was examined to understand the features and the shape of the data. There are ten attributes in this dataset, 'satisfaction_level' ranges from 0-1 and shows the employee satisfaction, and 'last_evaluation' also ranges from 0-1 showing the evaluated performance by the employee. Then there is 'number_projects', the number of projects assigned to an employee, 'average_monthly_hours', 'time_spent_company' showing the number of years at the company, and 'work_accident' which is a binary feature, 0 for no work accident and 1 for having a work accident. Lastly, 'promotion_last_5years' is also a binary feature depicting if the employee had a promotion or not, 'Departments', 'Salary' which is a categorical feature with the values low, medium, and high, and 'left' which is the binary feature showing whether the employee left or not.

In the given dataset, we have two types of employees one who stayed and another who left the company. Around 24% of the employees left the company while approximately 76% did not.

4. Data Pre-processing

For data pre-processing scaling and label encoding are implemented as KNN uses distance algorithms sensitive to the range of features. Normalization is the scaling technique (Min-Max scaling) used to shift and rescale values to make them range from 0-1. Moreover, since distance algorithms need numerical data, the two categorical features are represented numerically through label encoding. The column for salary is changed to make the values from 'low' to 0, 'medium' to 1, and 'high' to 2 using the sklearn library with LabelEncoder.

Cluster analysis is performed to find the optimal number of clusters using the elbow method for KMeans Clustering. Cluster analysis groups together data points that are similar to one another and dissimilar to data points that are a part of other clusters. The elbow method is used to find the optimum number of clusters in this study. It plots the values of the cost function produced by the different values of k and if it increases the distortion decreases and each cluster will have fewer distinct instances which will also be closer to their respective centroids. The value of k at which improvement in distortion declines the most is called the elbow and the optimum k value for clustering. However, clustering algorithms will locate and specify clusters in data even if none are present, hence, to measure the clustering tendency the ‘Hopkins’ statistic is used.

5. Techniques

The Hopkins statistic is defined by

$$H = \frac{\sum_{i=1}^{m} u_i^d}{\sum_{i=1}^{m} u_i^d + \sum_{i=1}^{m} w_i^d}$$

where d represents the number of dimensions in the dataset. If the value of H is between 0.01-0.3 then the data is regularly spaced, if it is 0.5 the data is random, and if it is between 0.7-0.9 then the data has a high clustering tendency.

K-means clustering is performed for model building as well. It is an unsupervised learning algorithm that divides the data into clusters that share similarities and are dissimilar from instances in other clusters. After this, the data is split into a train set and a test set and the KNN model will be trained through the train set and then predict the turnover outcome in the test set.

Using K-means clustering with KNN can significantly reduce the time complexity for classification by limiting the comparisons to the points within the nearest clusters. This is beneficial since the dataset used is from a large-sized organization. Where n is the total instances, d is the dimensions in the dataset, k is the number of centroids, m is the cluster size, t is the number of iterations, the space complexity of K-means is O(nd+kd), and the time complexity is O(tnkd). While for KNN the space complexity is O(kd) and the time complexity is O(Kmd). Their combined space complexity is O(nd+kd), and the time complexity for training is O(tnkd) and for classification, the time complexity is O(kd+Kmd)

6. Experimental Evaluation

The dataset used for experimentation is a real-world dataset obtained through Kaggle. It contains 10 columns and 1499 entries, ‘Salary’ and ‘Departments’ are categorical, ‘satisfaction_level’ and ‘last_evaluation’ are continuous, ‘promotion_last_5years’, ‘work_accident’, and ‘left’ are discrete, while ‘time_spent_company’, ‘average_monthly_hours’, and ‘number_projects’ are numerical. The average satisfaction level is 0.64 while for the last evaluation by employers, it is 0.72. It is highlighted that employees tend to work for around 4 projects on average with working hours being 200 on average. However, only 2.1% of the employees have been promoted over the last five years.

For model evaluation, confusion matrix and yellowbrick are used. Confusion matrix is an NxN matrix that is used for evaluating the performance of a classification model, where N is the number of target classes giving a holistic view of the performance of the model and the errors it generates as can been seen in Figure 1. Yellowbrick is a visualization and diagnostic tool that helps with quicker model selection and will allow visualization of the AUC-ROC score. Figure 2 and 3 show the high precision and accuracy of the KNN model used in this study.

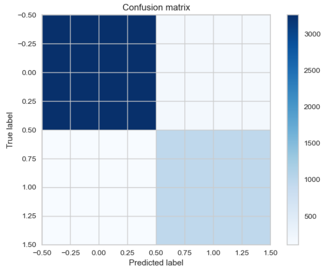

Figure 1. Confusion Matrix

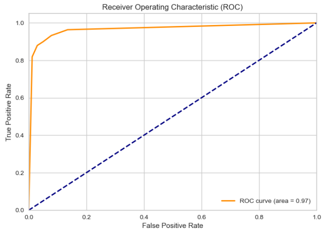

Figure 2. ROC

7. Results and Conclusion

Following Hypothesis 1, there should have been a negative relationship between satisfaction levels and turnover; however, it is negated by Figure 4. There is a significant increase in the number of those whose ’satisfaction_level’ value is between 3.5 and 4.5 and who left the company. This number exceeds the number of people who did not leave the company. Moreover, there is an increase in the number of employees who left who have satisfaction levels between 3.5-4.5 and 7-9. Subsequently, disproving hypothesis 1. Although no negative relationship was found between turnover and satisfaction levels, there might have been a few confounding variables that led to this result. Since the dataset is obtained from a secondary source, there is little clarity on how the data was collected and the opera-

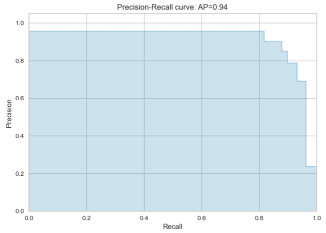

Figure 3. Precision Recall

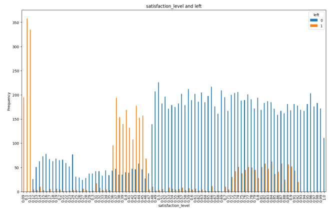

Figure 4. Satisfaction Levels and Left

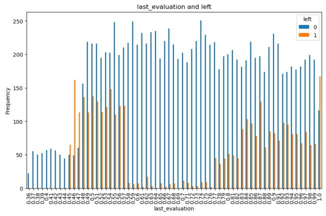

Figure 5. Last Evaluation and Left

tionalization of satisfaction level. Moreover, this assessment of satisfaction levels might not be up to date or some employees may have hidden their true feelings in the survey because of social desirability if they considered it a threat to their potential promotions or work relationships. Further research is needed to develop a holistic understanding of the relationship between satisfaction levels and employee turnover.

Hypothesis 2 is disproved as it is inferred that the relationship between employee turnover and the last evaluation is not negative. It can be seen in Figure 5 that there is a local increase between 0.46-0.6 and 0.8-1 highlighting an increase in the number of employees who had high evaluation scores and left the company. It is clarified when satisfaction levels and employer evaluation are assessed together. The employees who left can be classified into three clusters. Firstly, there is a satisfaction level of 0.4 and a last evaluation of 0.5 which indicates low organizational and work commitment. This group may need to be externally motivated for higher satisfaction and productivity. Secondly, the group with low satisfaction and high evaluations has a high turnover. This may be due to long working hours and lack of promotions which can be assessed further. Lastly, the group has high satisfaction levels and evaluations and high turnover. Further research into the reasons for their high turnover is needed. Hypothesis 3 proves there is a negative relationship between time spent at the company and turnover. It can be seen that the highest turnover rate is in the 3 years of working with the company and then the turnover decreases significantly over time.

Hypothesis 4 proves there is a positive relationship between working hours and turnover. It is inferred from Figure 6 that there is a local increase in the turnover for employees who work 125-160 and 210-290 hours on average per month. When the working hours and number of projects are analyzed together for turnover it can be concluded that the group working around 160 hours per month only has 2 projects assigned to them on average which might not be challenging or satisfying for the employees who are looking to grow hence a higher turnover in this group. Moreover, the employees working around 250 hours on average tend to have around 4 projects assigned to them showing poor work-life balance leading to higher turnover.

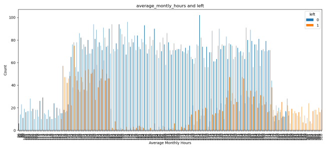

Figure 6. Average Monthly Hours and Left

Hypothesis 5 proves there is a negative relationship between the level of salary and turnover. It is inferred that the lowest turnover rate is in the highest salary group and the highest turnover rate is for the lowest salary group.

References

[1] J. I. Hancock, D. G. Allen, F. A. Bosco, K. R. McDaniel, and C. A. Pierce, “Meta-analytic review of employee turnover as a predictor of firm performance,” *Journal of Management*, vol. 39, no. 3, pp. 573–603, 2013.

[2] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, “Employee turnover prediction with machine learning: A reliable approach,” in *Intelligent Systems and Applications* (K. Arai, S. Kapoor, and R. Bhatia, eds.), (Cham), pp. 737–758, Springer International Publishing, 2019.