Team 38

Jack Lusk

Sasakorn Phanitsombat

Introduction

Employee attrition is a significant challenge faced by organizations today. High turnover rates not only impact productivity and team dynamics, but also result in substantial costs associated with hiring, training, and onboarding new employees. This project aims to leverage IBM's employee data to build a predictive model that can help identify the factors influencing employee longevity and provide insight for improving employee retention strategies.

Dataset and data preparation

IBM HR Analytics Employee Attrition & Performance | Kaggle
https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

The project will utilize a dataset with 1470 rows of employee data from IBM. The team performed a cursory data validation for blanks or invalid data and found the data has been processed. The variables BusinessTravel, Department, education, and EducationalField were changed to categorical variables. Also, after looking at the four salary variables: DailyRate, HourlyRate, MonthlyIncome, and MonthlyRate the team found discrepancies and we are using MonthlyIncome for the models.

The next step is to choose the significant indicator variables. This can be done with R lm function and finding variables with strong correlation.

Planned Approach and Methodology

There are at least two models that we will try to use to perform a prediction of whether an employee based on certain attributes are likely to stay with the organization.

1. Cluster model: we can cluster employees based on different attributes e.g. job satisfaction, years at the company, performance rating, salary, percent of raise, etc. We will be separating the data into similar groups based on the shared attributes. We will then need to choose the number of clusters and balance the trade-off between simplicity and accuracy. This trade-off needs to be made as we should ensure our model is relatively straightforward to understand to the non-analytics decision makers in the organization while we should ensure the model accuracy is high by taking into account the most important patterns and data nuances. In terms of data transformation, we need to perform standard scaling which converts our variables in a way that the mean becomes zero and the standard deviation becomes 1. Besides, min-max scaling may be required since it helps fit our dataset range to a specified scale without impacting the distribution shape.

2. Regression model: Depending on our final variables to be used in the model, we will sort out a formula that represents the relationship between the variables in the dataset e.g. whether it's a linear relationship or something else. Meanwhile, a logistic regression model can be considered as we would like to predict the probability that an employee will likely stay based on different attributes. In terms of data transformation, we may require a log transformation. It helps reduce the skewness of the data and makes the relationship between variables clearer. Box-Cox transformation may be required if the data is not normally distributed and this will help improve the accuracy of the predictions for linear regression models.

In terms of the dataset used, we will split our datasets into three; train, validation, and test. The proportion of the training dataset is 80% and the validation dataset is 10% while the test dataset is 20%. The models will use a training dataset to build the model. The validation dataset will be used to give an evaluation of the model fitness in the training dataset and this is to help us fine-tune hyperparameters. In terms of model comparison, the test dataset will then be the last to help compare the models that we develop to understand which model yields the optimal recall and accuracy.

To go deeper into fine-tuning hyperparameters, we can use any of the following methods: Bayesian optimization, random search, and grid search. For example, Bayesian optimization takes the model as an optimization model, and we can discover the good parameters combination after a couple of iterations.

Literature Survey

1. S. Kaur and R. Vijay, "Job Satisfaction – A Major Factor Behind Attrition or Retention in Retail Industry", *Imperial Journal of Interdisciplinary Research*, vol. 2, no. 8, 2016.

Job Satisfaction – A Major Factor Behind Attrition or Retention in Retail Industry | Semantic Scholar

The authors focus on the elements of positive work environment, developing, and retaining talent in retail environment. It is helpful to get another industry other than IBM.

2. G. K. P. V. Vijaya Saradhi, "Employee churn prediction", *Expert Systems with Applications*, vol. 38, no. 3, pp. 1999-2006, 2011.

Employee churn prediction - ScienceDirect

This paper looks at customer churn and uses elements for improving loss of customers for loss of employees. I find this an interesting approach that should be further investigated.

3. D. A. B. A. Alao, "Analyzing employee attrition using decision tree algorithms", *Computing Information Systems Development Informatics and Allied Research Journal*, no. 4, 2013.

ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS | D. | Computing, Information Systems, Development Informatics and Allied Research Journal (iiste.org)

The dataset is 109 observations for workers at a Nigeria institution. Waikato Environment for Knowledge Analysis (WEKA) and See5 were used to generate a decision tree model. They mainly used demographics information.

Conclusion

The team addresses the challenge of employee attrition and proposes a predictive modeling approach using IBM's employee data. The dataset, consisting of 1470 rows, has been validated and prepared for analysis. The report outlines two models that will be employed: a cluster model to group employees based on shared attributes and a regression model to predict the likelihood of an employee staying based on various factors. The dataset will be split into training, validation, and test sets, with the test set used for model comparison. We suggest exploring hyperparameter fine-tuning methods such as Bayesian optimization, random search, and grid search. The team also reviewed three journals related to attrition. The team will further explore looking at datasets and analysis for customer churn as many factors may be similar.