Human Resource Analyzing and Predicting Employee Turnover

Prepared By

Wissam Mohammed

9931067600.ups@htu.edu.jo

Data Science - The National ICT Up skilling program

Prepared for

HTU Al-Hussain Technical University

September 10 2021

Abstract

In this Report, we will establish exploratory data analysis and model prediction on Kaggle Dataset Employee Turnover (Attrition). To understand and discover the reasons behind why employees takes decision to resign from work and which department have higher tendency to quit and leave his work. The records and features will be analyzed through Statistical techniques. Firstly, Pearson-r correlation matrix Algorithm, it will help us analyze relationship between numerical features in the records.

In Analysis. We will study single variable analysis which also known as univariate analysis, analyzing one feature separately [3] Next, Bivariate analysis, to analyze linear-nonlinear relationship. Next, multivariate analysis to understand the spread & distribution of variable.

Machine learning models prefer data to be normally distributed before model evaluation. If some features are not normally distributed we will apply log transform to force some feature to be somehow normally distributed. Machine learning likes to be fed with clean and normalized data, for better accuracy and performance. Next, we will apply supervised machine learning classification model algorithm. The first part apply of supervised machine learning classification models will be used for evaluation and testing. The three algorithms are Decision tree classifier random forest classifier and XGBoost classifier. All three algorithm are categorized as supervised machine learning models. The last part will be used for prediction accuracy, for employee turnover. A classification confusion matrix report will printed, to explain the evaluation and predicted accuracy. The confusion matrix will summarize the performance of a classification algorithm [7]. We will also use The (AUC) .Area under the curve is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve [8]. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

Contents

1. Introduction

Human Resource analytics is also known People analytics, which is a data driven approach to managing people at work .A common definition of employee turnover is the loss of talent in workforce overtime, this includes any employee departure, including resignations, termination, and retirements [1].Employee turnover may result in high cost for company, which may leads into affecting company's hiring and retention decision. Other names for employee turnover is employee churn or Attrition [1].

1.1 Objectives

Our main aim in the Human resource analytics project is to analyze reasons behind resign of an employee and predict the turnover of an employee.

1.2 Contribution

The main contribution of the following report:

- We apply Exploratory Data Analysis to understand certain behavior & what is affecting employee turnover
- Supervised Machine learning classifiers Evaluations and performance
- We apply supervised machine learning classifiers to predict the accuracy of employee turnover

2. Data

Human resource dataset is created by Kaggle community. Its fictional dataset was made to stimulate real life data in modern companies in Human resource section[2], since HR data is considered confidential it's impossible to provide these information to the public or for data science communities. Kaggle data science community decided to create human resource dataset similar to real HR data in modern companies, in order to test machine learning models on it and also to apply exploratory data analysis.

Our human resource dataset consist of 10 features x 14999 Records

2.1 Features Description

Columns	Description
Satisfaction_level	Employee satisfaction level
Last_evaluation	Employee last evaluation score %
Number_project	Number of project assigned to employee
Average_monthly_hours	Monthly hours of employee
Time_spend_company	Years spend in company
Work_accident	Employee work mistake
Churn	Turnover , 1 for Left & 0 for Stay
Promotion_last_5years	Promotions upgrade-level
Department	Sections(IT,Sales,Support,Accountant ,etc)
Salary	0 Low , 1 medium , 2 High

2.2 Datatypes description

A. Quantitative Measurement scale

Quantitative data	Measurement Scale
Satisfaction Level	Continuous Ratio
Last_evaluation	Continuous Ratio
Number projects	Discrete
Average monthly hours	Discrete
Time spend company	Discrete
Work Accident	Discrete
Churn	Discrete Binary
Promotion Last 5 years	Discrete

B. Qualitative Measurement scale

Qualitative data	Measurement Scale
Department	Nominal
Salary	Nominal Ordinal

3. Methods

3.1 Three Algorithms models on full data.

We performed three model for evaluation, starting by decision tree. Decision tree are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. [4]. Second model is random forest .Random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting [5].Finally, XGBoost model. It provides a parallel tree boosting it allows for the optimization of arbitrary differentiable loss functions [6]. In each stage n-class regression trees are fit on the negative gradient of the binomial or multinomial deviance loss function. Binary classification is a special case where only a single regression tree is induced [6]

3.2 Metrics used on the Three algorithm Models

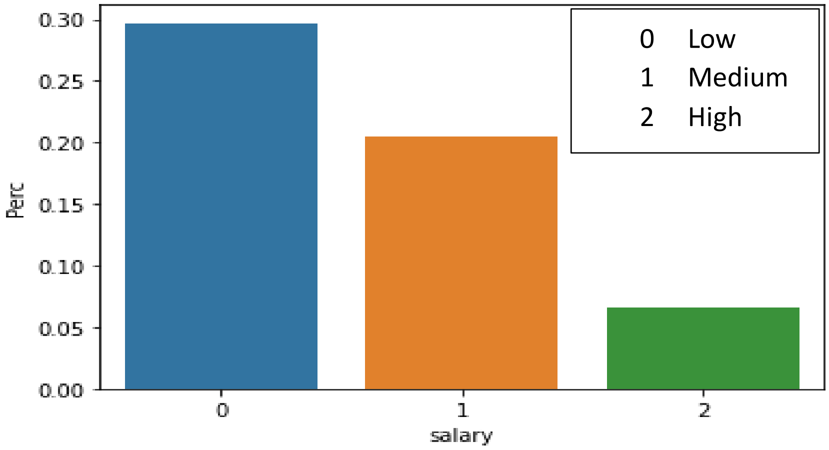
Cross validation technique with 10-k-fold .to make sure to provide the best result for all three models. Since our data is imbalanced, we had to add balanced method on class-Weight. Our Random state method is adjusted to 42, to avoid providing prediction repetition on one sample test. The test size for all three algorithm is 0.2 = 20%.Min sample leaf is 150 and the max depth is 8 for all three model.

3.3 Exploratory Data Analysis

During the analysis we found that **75.19** %(11428) stayed did not turnover and **23.80**% did turnover (3571)

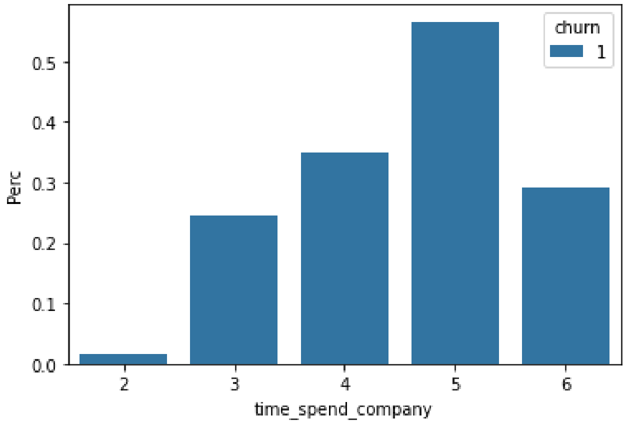
We plot the salary feature in bar graph and group it by turnover of employee Low salary which associated with number 0 in our dataset the explanation means that employee with low salary are more likely to leave the company (turnover)

3.3.1 Salary and turnover



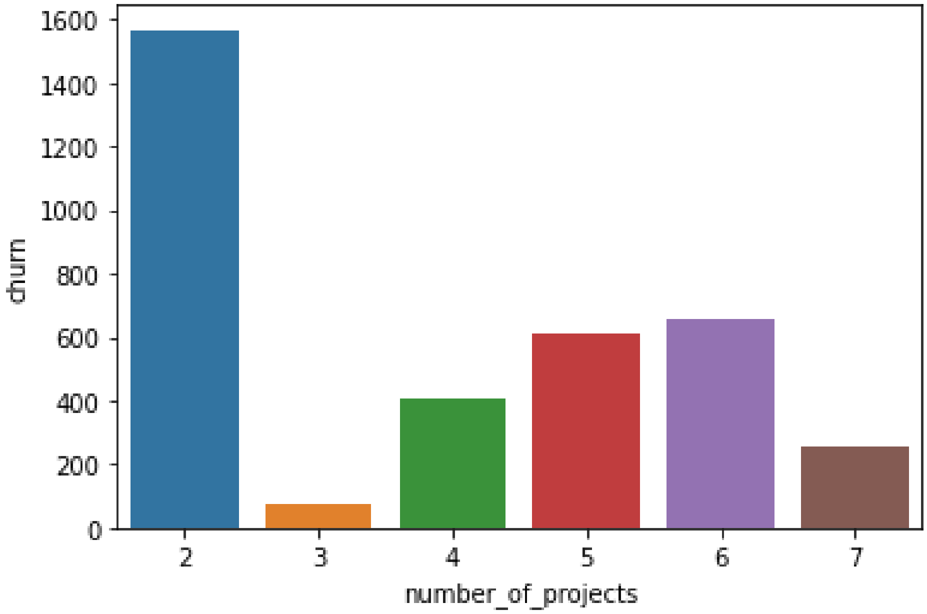
3.3.2 Time spend at company and turnover (churn)

There is 58% chance to leave the company after 5 years of service



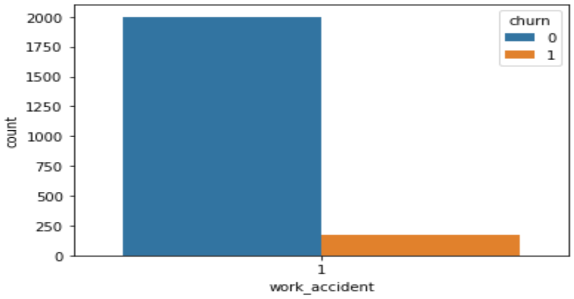
3.3.3 Number of project and turnover (churn)

Almost 1600 left the company after being assigned the second task for them (project)



3.3.4 Work Accident and turnover (churn)

Almost 2000 employees involved in work accident. But only less than 200 left the company



4. Result and Discussion

4.1 Model Evaluation Results

Model	Cross validation Prediction Score	Area under the curve Score
Decision Tree classifier	95.03	94.07
Random Forest classifier	95.70	93.97
XGBoost classifier	96.29	96.23

4.2 Model Evaluation Discussion

All models performed well. But at the end the XGBoost had the upper hand when it comes to performance and accuracy. The accuracy score with cross validation XGBoost scored 96.29.It was 1.26% percent different from Decision tree and 0.59 different from random forest classifier, it is still high accuracy. We can implement Decision tree and Random forest in our studies since the accuracy is high, but when it comes to Area under the curve score we will have to rely on XGBoost Classifier. To conclude, the best model in our evaluation used in the HR data is XGBoost Classifier Algorithm.

4.3 Confusion Matrix Result

XGBoost Classifier Confusion matrix	Positive	Negative
Positive	True Positive (2731)	False Positive (132)
Negative	False Negative (26)	True Negative (871)

4.4 Confusion matrix discussion

True Positive (TP) = 2731; meaning 2731 positive class data points were correctly classified by the model

True Negative (TN) = 871; meaning 871 negative class data points were correctly classified by the model

False Positive (FP) = 132; meaning 132 negative class data points were incorrectly classified as belonging to the positive class by the model

False Negative (FN) = 26; meaning 26 positive class data points were incorrectly classified as belonging to the negative class by the model

5. Conclusion

The accuracy of our model XGBoost 96.29.This can be explained as that all feature effect the employee turnover such as salary and number of project assigned And year spend & work accident but only 23.80 out of 75.19 decided to resign .During the analysis our main focus was on: Number of project assigned and salary, years spent, work accident, it was highly correlated with employee turnover.so the XGBoost score suggest that most likely employee turnover is due low salary and number of projects, the number of project had huge effect on turnover is either due to failure or lack of knowledge.

6. Future Work

We still considered our work is limited. Since the percentage of turnover of employee is 23% we can somehow say only temporary that most employee resign “quit” is likely based on their own personal behavior. So we have to look deeper into our analysis. Feature extraction algorithm such as RFE Model will be applied in the next analysis, it help us determine which feature is most likely has huge impact on turnover unlike correlation matrix, which explain in general the correlation, might not be accurate, in some cases.

6. References

[1] What Is Employee Turnover (and Why It Matters) | Workest (zenefits.com)

[2] Human Resource | Kaggle

[3] https://homeweb.csulb.edu/~msaintg/ppa696/696uni.htm

[4] 1.10. Decision Trees — scikit-learn 0.24.2 documentation

[5] sklearn.ensemble.RandomForestClassifier — scikit-learn 0.24.2 documentation

[6] Getting Started with XGBoost in scikit-learn | by Corey Wade | Towards Data Science

[7] sklearn.metrics.confusion_matrix — scikit-learn 0.24.2 documentation

[8] AUC-ROC Curve in Machine Learning Clearly Explained - Analytics Vidhya