Predict who Quits

Habeebuddin Mir
Northeastern University
Toronto, ON
mir.h@northeastern.edu

Abstract

Among many drastic effects, the pandemic brought about the Great Resignation wave across the globe leading to high attrition rates faced by companies. This increases the cost borne by a company to hire, train and keep their employees. In this report, I will be using a data set of a company that wants to hire data scientists among candidates who clear their preliminary training, and leverage Machine Learning models to analyse and predict the probability of a candidate of leaving their existing company and join this new one.

1 Introduction

According to Microsoft's Work Trend Index, more than 40% of the global workforce were considering quitting their job in 2021. The COVID-19 pandemic allowed workers to rethink their careers, work conditions, and long-term goals. This coincided with the boom in Data Scientist jobs, making our particular data set very interesting to study. Not only for a student of Data Science, but this is also highly valuable data set for HR analysts to predict what kind of candidates are looking for a job switch and who can they target and optimise their recruitment process.

In this report I will be discussing how I prepossessed the data by addressing missing values and other cleaning followed by discussing which evaluation metrics to be used given the context of the data set and the model being trained on it. I'll be showing the learning curves obtained and the need for oversampling and hyper-parameter tuning to get the most optimised version of our model.

1.1 Related Works

There has been a great deal of study on employee retention in the past. Most of them have focused on surveys that showed employee satisfaction and growth. But very few have have lens on Data Scientists roles though these are the roles that are in great demand with plenty of options to move around in the industry, making it one of the most critical study to know to retain these employees.

I will be attaching the papers and work I have referred to while completing this midterm project and report in the references section.

1.2 Results

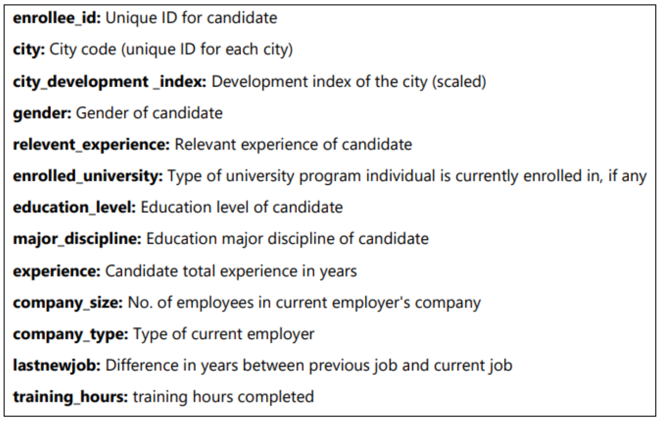
To overcome the deficiencies in the above mentioned past researches, I have trained multiple supervised machine learning models and concluded XG Boost Classifier to work the best with Test AUC Score of 77.69% after hyper-parameter tuning.

2 Problem description

A technology company wants to hire data scientists among people who complete their Data Science Boot-camp, called “training” in this study. Most of these people already have jobs but may be considering leaving their jobs to apply to this company, although that is unknown. The company wants to know which of these candidates would apply to work there after the training. This helps reduce the cost and time of research, increases the quality of training, and categorizes candidates. This data set is also designed to understand which factors might lead a person to leave their current job. Factors (predictor variables) were accumulated through candidates’ sign-up and enrollment.

There are over 10,000 observations in this data set which represents each candidate. The factors represent candidates’ demographics, education, experience, gender, previous company history, and training hours completed.

My response is a binary classification variable, target, whose classes are Y=0 if a candidate wants to remain in their current job and Y=1 if a candidate is looking for a job change in the HR researcher role as well. The goal of this study is to save the Data Science company’s resources by filtering in candidates with the highest potential and willingness for a job switch. Therefore the research questions are to identify the influential factors that output candidates who will change jobs to work as data scientists in HR research and estimate the probability that a candidate will make the switch.



enrollee_id: Unique ID for candidate

city: City code (unique ID for each city)

city_development _index: Development index of the city (scaled)

gender: Gender of candidate

relevent_experience: Relevant experience of candidate

enrolled_university: Type of university program individual is currently enrolled in, if any

education_level: Education level of candidate

major_discipline: Education major discipline of candidate

experience: Candidate total experience in years

company_size: No. of employees in current employer's company

company_type: Type of current employer

lastnewjob: Difference in years between previous job and current job

training_hours: training hours completed

Figure 1: Feature Details for the Dataset

3 Methodology

3.1 Data Tidying and Manipulation

The data-set’s categorical features like city were converted to numeric forms using Label Encoding method to make it machine readable. Some features like education level where order is important is not converted using Label Encoder as it doesn’t take into account the order. These ordinal features by making dictionaries for them and mapping them to corresponding numerical values.

The next step to clean data is to deal with missing values. I use k-nearest neighbors on features having missing data. The imbalance of data was dealt by oversamlping using Synthetic Minority Oversampling Technique (SMOTE).

3.2 Models

I used the datset to train models like Linear Regression, Decision Trees, Support Vector Machines and XG Boost Classifier. Among all of them I found Extreme Gradient (XG) Boost Classifier to perform the best, hence I chose it to make the final model presented in this report.

XGBoost is an implementation of Gradient Boosted decision trees.In this algorithm, decision trees are created in sequential form. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model.

3.3 Evaluation Metrics

During this I observed that We see that the target is skewed and thus the best metric for this binary classification problem would be Area Under the ROC Curve (AUC). We can use precision and recall too, but AUC combines these two metrics. Thus, I will be using AUC to evaluate the model that I build on this dataset.

4 Experiments

Feature Importance values show the degree to which a feature is correlated with the target feature, in a simple bi-variate relationship. It is based on an “Alternating Conditional Expectations” (ACE) score, which is used to detect non-linear relationships between individual explanatory features and the target feature. For the dataset we used, the feature importance values are shown

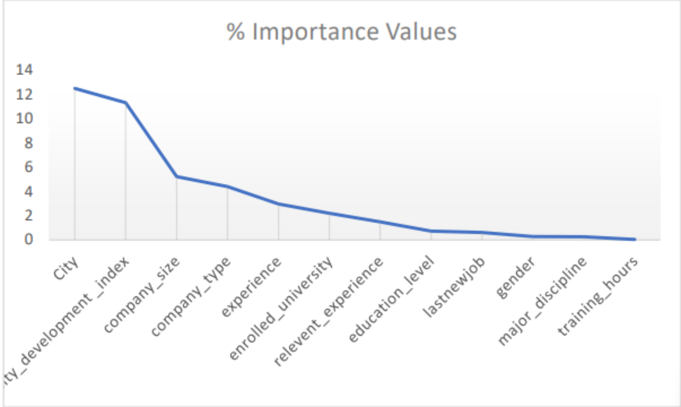


Figure 2: Bivariate Feature Importance Values for each Feature

Using the Extreme Gradient(XG) Boost Classifier on the preprocessed data and default parameters I get the following accuracy:

Train AUC Score: 0.9234348363001115
Test AUC Score: 0.7604762998949579

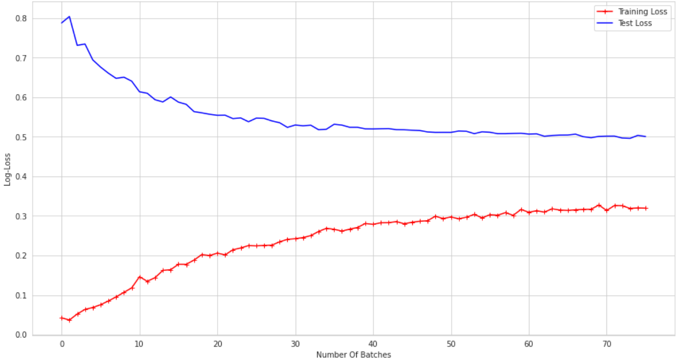
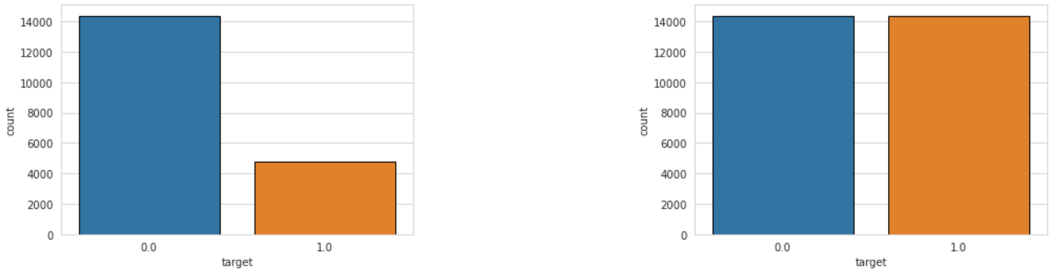


Figure 3: Learning Curve

It can be observed by the accuracy scores that the model is overfiting the data. To tackle this, we can do various things like we can increasing data set size in a balanced way and can also tune the hyper-parameters of the model.



(a) Before SMOTE

(b) After SMOTE

Figure 4: Data Balance

The learning curve shown in Figure 3 indicates we have a high variance problem as the gap is considerably high due to low training error.

Figure 3(a) shows that the data is highly unbalanced the true and false values in target class are not close to equal. This can be dealt by oversampling using Synthetic Minority Oversampling Technique (SMOTE). After SMOTE, we achieve balance as shown in Figure 4(b)

The parameters that the model has here are known as hyper-parameters, i.e. the parameters that control the training/fitting process of the model. To choose the best possible hyper-parameter values, we evaluate all the combinations and see which one improves metrics the most. I use Bayesian optimization with Gaussian process to evaluate this in the most efficient manner. Bayesian optimization algorithm need a function they can optimize. Most of the time, it's about the minimization of this function, like we minimize loss.

Now we cannot minimize the accuracy, but we can minimize it when we multiply it by -1. This way, we are minimizing the negative of accuracy, but in fact, we are maximizing accuracy. Using Bayesian optimization with Gaussian process, thus can be accomplished by using gp_minimize function from scikit-optimize (skopt) library. Doing this I got the following hyper-parameters best for my model:

Best Parameters : 'n_estimators': 1216, 'min_child_weight': 1.9330894375612764, 'gamma': 9.647431578451648, 'colsample_bytree': 0.9826077917352968

Best AUC score : -0.7708515472406663

Figure 5 below represents our new learning curve after using SMOTE and hyper-parameter tuning our model.

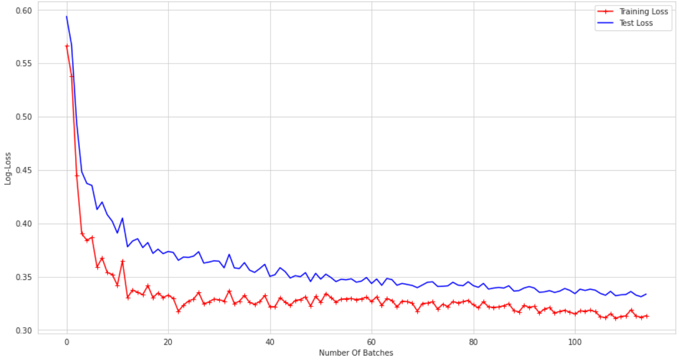


Figure 5: Learning Curve

The significant change can clearly been observed in Figure 5 where the initial high variance has been sorted.

5 Conclusion and future works

In this report, we used machine learning g techniques to analyze data about what motivates data scientists to look for new jobs. After the system was trained using the training dataset, the system was able to predict whether a data scientist is considering looking for a new job with an accuracy rate of higher than 77%. We also found that the major factors that influence this are related to the location of the individual (city and city development index),and the company the individual currently works at (size and type). From here it can be observed that individuals from underdeveloped cities switch jobs more in search for a better place to thrive.

References

[1] https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists

[2] Conlon, Sumali. (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. 10.24297/ijmit.v16i.9058.

[3] Alao, D. & Adeyemo, A. (2013). Analyzing employee attrition using decision tree algorithms. Computing, *Information Systems & Development Informatics* Vol. 4 No. 1 March, 2013. 17–28.