Team 1 ("Destroyers") Project Proposal

Using Human Resources Records to Recommend Employment: a Classicification Problem

Jesus Vazquez, Leo Li, Ying Zhang, Tian Zhao

BIOS 735, Spring 2021
March 31, 2021

1 Introduction

Recruiting the right employees is key to business growth and success, which makes staffing an essential business operation. While recruiting companies like LinkedIn and Indeed have helped the business expand their grid of search for new employees at competitive prices, staffing can be expensive. According to the 2019 Training Business Report, a survey study of 240 U.S.-based corporations and educational institutions with 100 or more employees, reported about 83 billion dollars were spent in 2019 in the US alone for training-related expenditures [1]. At the individual level, on average companies spent around $1,300 per trainee in 2019. The large expenditures related to recruitment and training raise the importance for companies to keep employee-retention high. Exploring employee factors such as socio-demographic characteristics or tenure and experience in their respective field of study may help human resourced (HR) managers achieve higher employee-retention rates and subsequently benefit long-term company health and development.

2 Data Description

A company which is active in big data and data science wants to hire data scientists among people who successfully pass some courses which are conducted by the company. The HR representatives in the company want to know which candidates really want to work for the company after training. Compounding the issue is some candidates only take these courses in order to look for new employment in other companies. An a priori understanding of the attributes which may predict which candidate is looking for a job change could help reduce the cost and time for planning these courses. There are 13 demographic and educational features provided in the data set, and the details of the variables in the data set are provided in Table 1.

This data set is offered by Kaggle, a subsidiary of Google Inc. that has publicly available data sets for machine-learning and data-science practitioners. The data set contains 19158 observations, and we will be splitting it into a training and test set with a ratio of 70% to 30%. The data set also has several important characteristics. First, the data set is very imbalanced for some variables, which means the classification model might have poor prediction accuracy due to those variables with skewed distribution. Second, there are different types of features, including binary, categorical, and continuous variables, which adds modeling and analytical complexity. Third, there are many missing values in the data set. Some variables tend to have more missingness than others, and the missing percentages for the variables in the data set range from 2% to 30%.

Table 1: Description of the Variables in the Data Set

Feature Name	Details
enrollee_id	Unique ID for candidate
city	City code
city_development_index	Developement index of the city (scaled)
gender	Gender of candidate
relevent_experience	Relevant experience of candidate
enrolled_university	Type of University course enrolled 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
last_new_job	Difference in years between previous job and current job
training_hours	Training hours completed
target	Not looking for job change, 1 – Looking for a job change

3 Aims

The primary aim of the project is to build statistical and machine learning models to predict whether a candidate wants to look for new employment based on features described in table 1. By using the statistical techniques in the course, we will practice the steps to build a statistical model and estimate parameters through MLE. While for the machine learning model, a generic random forest and SVM model will be built to compare the accuracy of predictions. The second aim is to explain the sensitivity of results based on varying model types and features in prediction through the area under the curve (AUC), sensitivity, and specificity of the predicted values. The third aim is to practice our skills in building R packages to solve real-world data problems.

4 Methods

In order to predict the probability of a candidate to look for a new job or will work for the company, we consider the following four methods:

- **Logistic Regression:** In logistic regression model, we assume that,

$$y_i|x_i \sim Binomial(m_i, p_i),$$

and that,

$$logit(p_i) = \log(\frac{p_i}{1-p_i}) = x_i^T\beta,$$

for $i = 1, \ldots, n$ [3]. Then the log-likelihood function for solving β can be expressed as,

$$l_n(\beta) = \sum_{i=1}^{n}\{y_i x_i^T\beta - m_i \log(1 + \exp(x_i^T\beta))\}$$

- **Ridge Logistic Regression:** Ridge logistic regression is the penalized version of logistic regression with L_2 loss, where the following log-likelihood function is maximized [4]:

$$l_n^{(ridge)}(\beta) = \sum_{i=1}^{n}\{y_i x_i^T\beta - m_i \log(1 + \exp(x_i^T\beta))\} - \lambda\sum_{j=1}^{p}\beta_j^2.$$

- **Support Vector Machine (SVM):** In SVM for classification, we can fomulate the optimization problem as the following convex minimization problem with linear constraints [5]:

$$\min \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i,$$

subject to $y_i(\beta_0 + x_i^T\beta) \geq (1-\xi_i)$, where $\xi_i \geq 0$, for $i = 1, \ldots, n$, and C is a given cost parameter.

- **Random Forest:** The random forest builds the decision rules by including multiple decision trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree [6]. Specifically, the algorithm of random forest for classification is summarized as follows: for $b = 1, \ldots, B$,

  (a). Draw a bootstrap sample $Z_{1:n}^{*(b)}$ of size n from the training data.

  (b). Grow a classification tree to $Z_{1:n}^{*(b)}$ by the following steps until the minimum node size, k_{min}, is reached:

  1. Randomly select n out of p variables in the design matrix.
  2. grow a classification tree using only the selected variables.

  Then output the ensemble of trees. In order to make a prediction at a new point x, let $\hat{C}_b(x)$ be the class prediction of the b^{th} random-forest tree. Then $\hat{C}_{rf}^B(x) =$ majority vote$\{\hat{C}_b(x)\}_1^B$.

For the four methods mentioned above, we will use complete-case analysis to handle the missing data in our data set, i.e., only those observations with complete data in all variables are used for analysis. However, the complete-case analysis would only be valid under the missing-completely-at-random (MCAR) assumption, which is usually very strong in practice, so that we would also like to consider multiple imputation, which is also valid under the missing-at-random (MAR) assumption, to handle the missing data for logistic regression and ridge logistic regression [7]. We will then compare the impact of different approaches to handling missing data (complete-case analysis and multiple imputation) in terms of model performance.

5 References

1 Training Magazine. "2019 Training Industry Report." "https://trainingmag.com/sites/default/files/2019_industry_report.pdf". Accessed March 11, 2021.

2 HR Analytics: Job Change of Data Scientists, Predict who will move to a new job "https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists"

3 McCullagh, P. Nelder, J. (1989), Generalized Linear Models, Second Edition , Chapman Hall .

4 Hoerl, Arthur E., and Robert W. Kennard. "Ridge regression: Biased estimation for nonorthogonal problems." Technometrics 12.1 (1970): 55-67.

5 Cortes, C. Vapnik, V., 1995. Support-vector networks. Machine learning, 20(3), pp.273–297.

6 Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

7 Little, R. J., Rubin, D. B. (2019). Statistical analysis with missing data (Vol. 793). John Wiley Sons.