Team 1 (“Destroyers”) Project Report
Using Human Resources Records to Recommend Employment: a Classification Problem

Jesus Vazquez, Leo Li, Ying Zhang, Tian Zhao

BIOS 735, Spring 2021
May 8, 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.

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 prior 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.

In this project, we use a data set 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 12 feature variables and 19158 observations. 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 higher missing percentages than others, and the missing percentages for the variables in the data set range from 2% to 30%.

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 available features. By using the statistical techniques in the course, we will practice the steps to build a statistical model and estimate parameters through maximum likelihood estimation (MLE). While for the machine learning model, a generic random forest (FR) and support vector machine (SVM) model will be built to compare the accuracy of predictions. The second aim is to evaluate the performance of different classifiers through the area under the curve (AUC) of the receiver operator characteristics (ROC) curve, as well as the 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.

The rest of the report is organized as follows: Section 2 provides descriptions of the data preprocessing steps, the rationales of the proposed statistical and machine learning models, metrics for performance comparison, and functions included in the R package; Section 3 provides a summary of key outcomes/variables and the main findings of the project; Section 4 describes the conclusions based on the results and the limitations of the study.

2 Methods

In this project, we use logistic regression, ridge logistic regression, unweighted-SVM, weighted-SVM, and unweighted-RF to predict the probability of a candidate to look for a new job or will work for the company. To evaluate the performance of the five methods, we use the AUC of the ROC curve, as well as the sensitivity and specificity of the predicted values as metrics. In addition, we use complete-case analysis to handle the missing data in our training and test sets, 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. MCAR is usually very strong in practice and for this reason, we will consider multiple imputation (MI), which is also valid under the missing-at-random (MAR) assumption, to handle the missing data only for the logistic regression and ridge-logistic regression [7]. MI is only common in estimation problems when using parametric methods, this is why we will only be conducting MI for the logistic and ridge logistic regression. We then compare the impact of complete-case analysis and MI to handling missing data in terms of model performance for the parametric models. Class imbalance in the outcome variable can become problematic in classification problems. Due to the class imbalance of our outcome, we decide to test the performance of the unweighted SVM and RF by training the models using the complete-case training set, and by down-sampling the complete-case training set. To compare the performance of various methods using variations of the training set, we decide to only consider the complete cases of the test set. This would allow us to make a direct comparison between methods given that they will all be using the same test set.

2.1 Data Pre-processing Steps

Before conducting the analysis, we pre-process the data set as follows:

1. We start by converting the categorical variables to dummy variable notations using reference cell coding;

2. We then split the data set into a training and test set with a ratio of 70% to 30%;

3. To prepare the data sets for complete-case analysis, we only keep the complete observations in the training and test data sets;

4. To handle the missing data issue, we consider the method of multiple imputation. To prepare the data sets for multiple imputation, we use the *mice* package in R to conduct the multiple imputation. Specifically, the *mice* package imputes the missing data through the fully conditional specification (FCS) method, which specifies a conditional imputation model for each missing variable conditional on the other variables and assumes that a joint distribution exists that corresponds to this set of conditional distributions. The training data sets are obtained by imputing 5 data sets, in which separate analyses will be conducted on each one, and the final parameter estimates will be combined and adjusted by Rubin Formula. The test set is, again, formed by only keeping the complete observations in the set portion of the original data set.

5. Since the data set is very imbalanced, we will consider the approach of down-sampling. To down-sample the training data set, we use the R function *downSample*(). This function will keep all observations whose outcome is the minority class and will use random sampling to select

an equal amount of observations of the majority class. This will allow for our training data set to have a class-imbalance ratio between the outcome categories of 1:1. The data set used to down-sample will be the complete-case training data set from step 3.

To sum up, we have the following data sets ready from the data pre-processing step:

1. *train_cc.csv*: Training data set for complete-case analysis;
2. *train_mi_1.csv, train_mi_2.csv, train_mi_3.csv, train_mi_4.csv, train_mi_5.csv*: Five training data sets for multiple imputation;
3. *train_cc_downsample.csv*: Training data set for down-sampling analysis.
4. *test.csv*: Test data set for all types of analysis.;

2.2 Statistical and Machine Learning Models

2.2.1 Logistic Regression

To address the classification problem, one simple way is to use logistic regression. The logistic regression model arises from the desire to model the posterior probabilities of a binary outcome via a linear function of the available features, while at the same time ensuring that they sum up to one and remain in $[0, 1]$ through a logit transformation [9].

In the logistic regression model, we assume that,

$$y_i|x_i \sim Bernoulli(p_i),$$

and that,

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

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

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

We could solve for β by maximizing the log-likelihood function. To achieve this, we use IRLS method. The algorithm goes as follows:

```
Algorithm 1: Iteratively reweighted least squares
Result: $\beta_m$
initialization: $\beta_0 = 0$,$m = 0$;
while $\beta_m$ not converged do
    Compute $\mathbf{p}$ by setting its elements to:
        $$p(x_i;\beta) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}} i = 1, 2, ..., N \qquad (1)$$
    Compute the diagonal matrix $\mathbf{W}$. The ith diagonal element is:
        $$p(x_i;\beta)(1-p(x_i;\beta)) \qquad (2)$$
    $\beta_{m+1} \leftarrow \beta_m + (\mathbf{X^T W X})^{-1}\mathbf{X^T}(\mathbf{y}-\mathbf{p})$;
    $m \leftarrow m+1$;
end
```

To improve computation efficiency, we could compute the $N \times p$ matrix $\tilde{\mathbf{X}} = \mathbf{WX}$, which is:

$$\begin{pmatrix} p(x_1;\beta)(1-p(x_1;\beta))x_1^T \\ p(x_2;\beta)(1-p(x_2;\beta))x_2^T \\ ... \\ p(x_N;\beta)(1-p(x_N;\beta))x_N^T \end{pmatrix} \tag{3}$$

In this way, we could avoid doing operations on $N \times N$ matrix W, and instead, doing multiplication on smaller objects.

2.2.2 Ridge Logistic Regression

One high-dimensional extension of logistic regression is ridge logistic regression, which is motivated by the multicollinearity among the feature space. In particular, when there are many correlated variables in a logistic regression model, their coefficients can become poorly determined and exhibit high variance. A wildly large positive coefficient on one variable can be canceled by a similarly large negative coefficient on its correlated cousin. By utilizing the ridge logistic regression, this problem can be alleviated [9].

Specifically, the ridge logistic regression is the penalized version of logistic regression with L_2 loss, where the following log-likelihood function is maximized [4]:

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

Similar to logistic regression, we use IRLS algorithm to compute the estimates to the regression coefficients:

```
Algorithm 2: Iteratively reweighted least squares (Ridge Logistic Regression)
Result: $\beta_m$
initialization: $\beta_m = 0, m = 0$;
while $\beta_m$ not converged do
    Compute p by setting its elements to:
        $$p(x_i;\beta) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}} i = 1,2,...,N \tag{4}$$
    Compute the diagonal matrix W. The ith diagonal element is:
        $$p(x_i;\beta)(1-p(x_i;\beta)) \tag{5}$$
    $\beta_{m+1} \leftarrow \beta_m + (\mathbf{X^T W X} + \lambda\mathbf{I_p})^{\mathbf{-1}}(\mathbf{X^T}(\mathbf{y}-\mathbf{p}) - \lambda\beta_m)$;
    $m \leftarrow m+1$
end
```

Similarly, we can compute $\tilde{\mathbf{X}} = \mathbf{WX}$ to replace $\mathbf{WX}$ in the above iteration.

2.2.3 Support Vector Machine

Support Vector Machine (SVM) is a classification approach that utilizes the geometry of the data, in which we can formulate 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. The optimal model was picked given the highest Cohen-Kappa statistic based on a 5-Fold Cross Validation Procedure and a C-grid ranging between 1 and 3000. The C-grid extends to large values due to the high level of class imbalance in the data.

2.2.4 Weighted Support Vector Machine

Weighed-SVM is an extension of SVM in which the weights are used to account for the class imbalance of the outcome. Again, we can formulate the optimization problem as a convex minimization problem with two linear constraints [10]:

$$\min \frac{1}{2}\|\beta\|^2 + C_1 \sum_{i=1}^{n_1} \xi_i + C_2 \sum_{j=1}^{n_2} \xi_j,$$

subject to $y_i(\beta_0 + x_i^T\beta) \geq (1-\xi_i)$, where $\xi_i \geq 0$, for $i = 1, \ldots, n_1 + n_2$ where n_1 and n_2 are the number of observations in each class. C_1 and C_2 are the cost parameters pertaining to the class membership. We will fix C_1, our weight, to the inverse of the proportion of observations in category 1, while searching for the optimal C_2 based on a search grid between 1 and 10, as well as the inverse of the proportion of observations in class 2. Again, we will use 5-fold cross validation and will pick the model with the highest Cohen-Kappa.

2.2.5 RF

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$\{\hat{C}_b(x)\}_1^B$. When using model forest one must tune the number of ensemble trees and the number of variables randomly sampled as candidates at each node split. For this project, we fixed the total number of trees to 500. Additionally, we searched through a grid between 1 and the total number of covariates to find the optimal number of variables to be randomly sampled. The optimal model was then picked according to the highest Cohen-Kappa statistic based on a 5-fold cross validation procedure.

2.3 Metrics for Performance Comparison

To evaluate the performance of the classifiers that we consider, we use the AUC of the ROC curve, as well as the sensitivity and specificity of the predicted values as metrics.

Sensitivity (or true positive rate) is defined as the proportion of the cases that are correctly predicted as positive among all the positive samples. Specificity (or true negative rate) is defined as the proportion of the cases that are correctly predicted as negative among all the negative samples.

A ROC curve is a commonly used summary for assessing the trade-off between sensitivity and specificity, and it is a plot of the sensitivity versus 1-specificity as we vary the parameters of a classification rule. The AUC of the ROC curve is a commonly used method to evaluate the overall performance of classification rules, in which a greater AUC implies a better classification method.

Note that the AUC, sensitivity, and specificity are all constructed and computed based on the test data set, using the estimates resulted from fitting the model on the training data set.

2.4 R functions

In this project, we develop an R package called *ridgeLR*, which includes the following functions:

1. *logistic_regression(y, X)*:
   - **Description:** Fit the logistic regression model.
   - **Arguments:** (1) y: an n by 1 matrix containing the outcome variable; (2) X: an n by p design matrix.
   - **Outputs:** Maximum likelihood estimates of the regression coefficients.
   - Implemented using Rcpp.
2. *ridge_logistic_regression_beta (y, X, lambda)*:
   - **Description:** Fit the ridge logistic regression model for a given tuning parameter.
   - **Arguments:** (1) y: an n by 1 matrix containing the outcome variable; (2) X: an n by p design matrix; (3) lambda: a numeric value of the tuning parameter.
   - **Outputs:** Maximum likelihood estimates of the regression coefficients.
   - Implemented using Rcpp.
3. *cv_ridge_logistic_regression (y, X, lambdas, fold)*:
   - **Description:** Find the optimal value for the tuning parameter of ridge logistic regression by k-folds cross validation.
   - **Arguments:** (1) y: an n by 1 matrix containing the outcome variable; (2) X: an n by p design matrix; (3) lambdas: a sequence of values of tuning parameters; (4) fold: number of folds for cross validation.
   - **Outputs:** The optimal tuning parameter for ridge logistic regression by grid search.
   - Implemented using Rcpp.

The R package have man pages for all the functions that are exported, and the R package has passed the following package tests:

- **"Correct dimension":** test whether the dimension of output regression coefficient estimates from *logistic_regression(y, X)* is equal to the number of columns of the input design matrix.
- **"Not full rank case":** test whether the *logistic_regression(y, X)* function will give a warning message when the input design matrix is less than a full rank.
- **"LR parameter estimates":** test whether the *logistic_regression(y, X)* function will give the same parameter estimate as the *glm()* function in *glmnet* package.
- **"LR vs. ridge LR":** for the same input data set, test whether the parameter estimates from *logistic_regression(y, X)* will have a larger absolute values than the parameter estimates from *ridge_logistic_regression_beta(y, X, lambda)*, when lambda is chosen to be 2.
- **"ridge LR tuning":** for the same input data set, test whether the parameter estimates from *ridge_logistic_regression_beta(y, X, 1)* will have a larger absolute values than the parameter estimates from *ridge_logistic_regression_beta(y, X, 2)*.

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	Development index of the city (scaled)
gender	Gender of candidate
relevant_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	0 - Not looking for job change, 1 - Looking for a job change

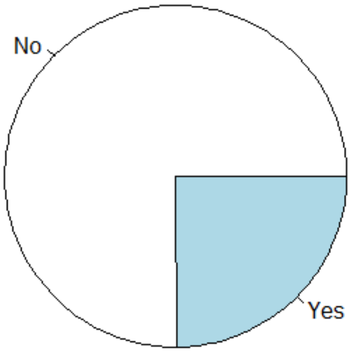
3 Results

3.1 Key Outcomes/Variables

There are 12 feature variables in the data set, which include some characteristics of the cities of the companies, such as the development index, as well as some other demographic and educational variables of the employees, such as gender, experience, and education levels. The details of the variables in the data set are provided in Table 1. The outcome variable of interest is the “target” variable, which is an indicator of whether a candidate would like to look for a job change.

Another feature of the data set is that it is highly imbalanced. As illustrated in Figure 1, only one-fourth of the candidates in the data set would like to look for a job change. This imbalanced nature of the data set adds some further challenges to the analysis, and we will use various approaches to improve the performances of the machine learning methods, such as down-sampling and incorporating weights into the analysis.

Figure 1: Pie Chart of the Target Variable in the Data Set



3.2 Main Findings

In Table 2, the logistic regression and ridge logistic regression give almost the same AUC regardless of using complete-case analysis or MI. Figure 2 also shows the same ROC curve for the logistic regression and ridge logistic regression, which is due to the small value of the penalized parameter (λ) tuned by the 5-folds cross validation. Moreover, we also find out the AUC of logistic regression using complete-case analysis as training data set is just slightly better than using imputed training data sets. As a result, we could conclude the prediction performance of the logistic regression is not affected significantly by using complete-case analysis or MI. For multiple imputation, we have also tried fitting the model by imputing 10 data sets, and the model performance is very similar to that based on 5 imputed data sets, so we proceed with using 5 imputed data sets for all MI-based analyses.

In Table 3, the RF and SVM using the unweighted method, equalling to using the complete cases as the training set, give the greatest AUC compared to the other models. Then, the SVM with weights give similar but slightly lower AUC compared to the unweighted SVM. In order to solve the problem of imbalance sampling, we also try to fit RF and SVM with the down-sampling method. The RF still gives comparable AUC with the unweighted RF. However, the SVM with the down-sampling method has the least AUC among all machine learning methods. The differences of the AUC among the SVM with the down-sampling method and other models are also shown in Figure 3. In the left bottom of the ROC curve, when the cut-off point for the positive outcome is very close to one, SVM with down-sampling has the lowest sensitivity under fixed specificity among all models. While with the decreasing of the cut-off points and increasing of the value of x-axis, the sensitivity of SVM with down-sampling method increases even faster than the other models after the x-axis over 0.4. Thus, the performance of the SVM with down-sampling is more affected by the choice of cut-off points than other models.

We choose the top 3 important variables by the variable importance for machine learning methods and absolute values of estimated coefficients for logistic regression and ridge logistic regression. The important variables are summarized in Table 4. All the models choose the city development index as the most important variable for prediction. Next, the logistic regression model and SVM chooses the employee's experience variables about the job while the RF choose the training hour in the session as the second important variable. Last but not the least, the third most important variable is still focused on the relevant experience for logistic regression but is the education-related variable for the other machine learning methods. In conclusion, the top 3 variables chosen by logistic regression and machine learning methods are quite similar, which means they abstract the similar variable information for prediction.

Table 2: AUC of the logistic regression and ridge logistic regression using complete cases and imputed cases

Model	LR[a]	RLR[b]	LR with imputed cases	RLR with imputed cases
AUC	0.763	0.763	0.761	0.761
tuned λ	-	0.12	-	(0.06-0.16)

a:LR- Logistic regression/ with complete cases, b:RLR- Ridge logistic regression/ with complete cases

Table 3: AUC of the machine learning methods

Model	RF(unweighted)	SVM(unweighted)	RF(downsampling)	SVM (downsampling)	SVM (weighted)
AUC	0.759	0.759	0.756	0.743	0.754

Figure 2: ROC curve of logistic regression and ridge logistic regression

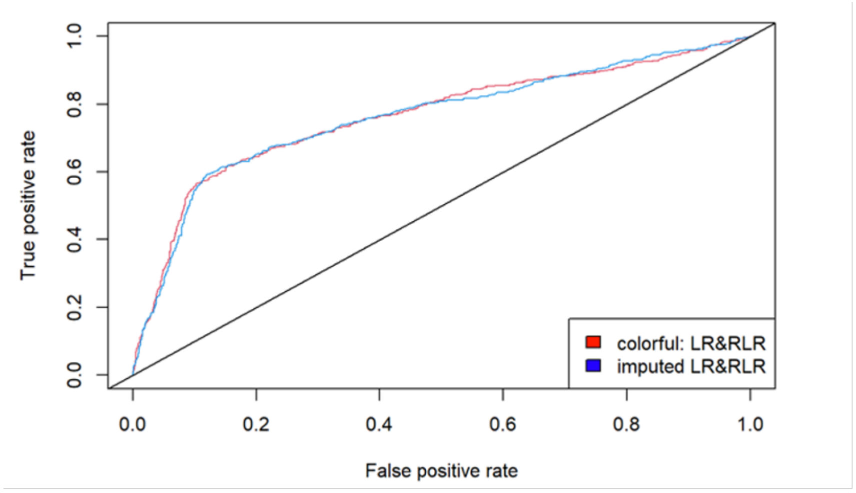
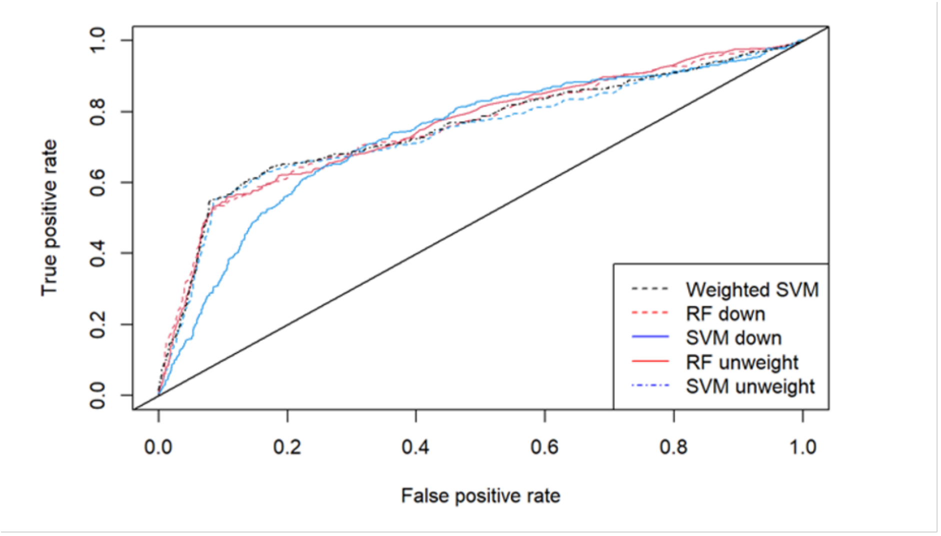


Table 4: Ranked the Top 3 most important variables by different models

Model	Top 1	Top 2	Top 3
Logistic regression (LR)	City development index	Experience =2	Experience =4
multiple imputation (imputed LR)	City development index	Experience =2	Relevant experience
RF unweighted	City development index	Training hours	Education level = 2
SVM unweighted	City development index	Relevant experience	Enrolled University=2
SVM with weights	City development index	Experience =2	Enrolled University = 2
RF with down sampling	City development index	Training hours	Experience =2
SVM with down sampling	City development index	Experience =2	Enrolled University = 2

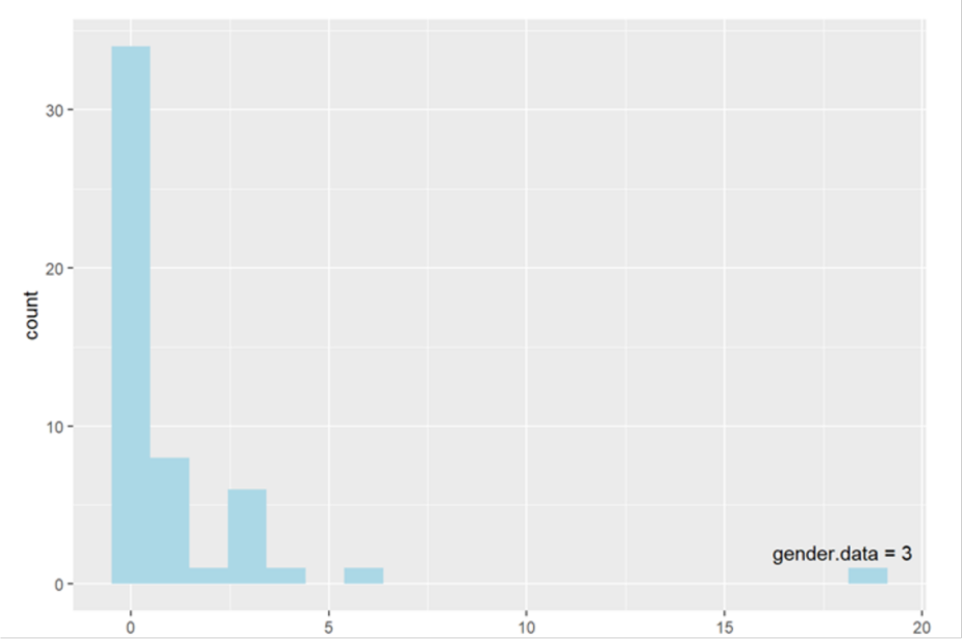
Figure 3: ROC curve of machine learning models



3.3 Sensitivity Analysis

In the sensitivity analysis, we aim to discover whether the estimated coefficient estimation of logistic regression is influenced by using different methods to handle the missing data (i.e., using complete-case analysis or MI). The histogram in Figure 4 shows the distribution of the absolute scaled differences for estimated coefficients (β's) by logistic regression using the complete data and imputed data. We find out that most, if not all, of the parameter estimates, are very close, with the only exception of the "gender.data = 3" (non-binary), which has a significant difference in estimates of using complete-case analysis and MI.

Figure 4: Scaled differences of coefficients between complete cases and imputed cases of LR



4 Conclusions

Based on the results of the prediction performances for various methods that we consider, we conclude that all the statistical and machine learning methods have very similar performances as measured by AUC of ROC curve, sensitivity, and specificity, although the logistic regression has a slightly better performance than all the other methods. For parametric models that we consider, the estimated regression coefficients for logistic regression are very similar to those of ridge logistic regression and the tuning parameter for ridge logistic regression is very close to 0, indicating that there is no obvious evidence of collinearity. In addition, unweighted SVM and weighted SVM have very similar performance, but the down-sampling SVM has a much lower AUC. The unweighted SVM has a much longer computation time than the down-sampling SVM. Regarding the performance of RF, we find that the performance of unweighted RF is similar to the down-sampling RF, but the computation time for unweighted RF is much longer.

Our results also shed some light on the data generation process of the HR data that we use. Since logistic regression has the best performance among all the methods that we consider, one could say that the assumptions made by the logistic regression (i.e., the logit transformation of the predicted probabilities and the linear relationship between the outcome and main effects of the covariates) seem to be reasonable for prediction. If the data set involves highly non-linear relationships or interactions among the covariates, then we would expect that the machine learning methods (i.e., SVM and RF) would have better performances. Therefore, in terms of the statistical and machine learning models, we recommend logistics regression over the other more complicated methods, because it has comparable performance as other methods and it is also easier to interpret.

According to the logistic regression model, the city development index and the experience level of the candidate are the two most important variables in predicting whether a candidate wants to look for new employment. Our general recommendation to the HRs is that people in more developed cities appear to be less likely to look for new employment, and people with around 2 years of experience appear to be less likely to look for new employment, compare to those people with more or fewer years of experience.

In the sensitivity analysis, we compare the parameter estimates from logistic regression and ridge logistic regression using complete case analysis with their counterparts using MI. We can see that most of the parameter estimates for complete-case analysis are similar to those for multiple imputation, indicating that the assumption of the MCAR missing mechanism may be reasonable. The only coefficient which demonstrates some extent of difference is 'gender.data = 3', which is the non-binary gender category, and the difference may be caused by the relatively small sample size in the non-binary gender category in the complete case data set.

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.

8 R.L. Schaefer , L.D. Roi R.A. Wolfe (1984) A ridge logistic estimator, Communications in Statistics - Theory and Methods, 13:1, 99-113.

9 Friedman, J., Hastie, T., Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.

10 Yang, Xulei and Song, Qing and Wang, Yue (2007). "A weighted support vector machine for data classification". International Journal of Pattern Recognition and Artificial Intelligence.