report

Minjie Bao

3/27/2021

Introduction

Describe your data set. Provide proper motivation for your work.

A data science company wants to hire data scientists among people who successfully pass some courses provided by the Company. Many people sign up for their training. The data is from the information that candidates provided. There are 19158 rows and 14 columns in the raw data set. This data set contains too many categorical variables and there are only two continuous variables:`city_development_index` and `training_hours`.

Our motivation for this project is to help the Company know which of these candidates really want to work for the company or will look for a new employment. This project can help the Company to reduce the cost and time as well as the quality of training or planning the courses and categorization of candidates.

What questions are you trying to answer?

The questions we want to answer is what factors lead a person to leave their current jobs. We are going to predict the probability of a candidate looking for a new job or will work for the company, and interpreting affected factors on employee decisions.

How did you prepare and clean the data?

First, I recode all the character type variables to categorical variables in different levels, and then convert them to numeric. I delete the variables `city` and `enrollee_id` since we already has `city_development_index` variable and `enrollee_id` is not useful for model prediction. I keep all the other variables as predictors. I recode the "1" and "0" values in the outcome variable `target` as "Yes" and "No", and then converted `target` as factor.

After cleaning the raw data, I split the whole dataset as 80% trainData and 20% testData. The trainData has 15,327 rows and 12 variables, and the testData has 3,831 rows and 12 variables.

Finally, the data has a lot of missing values (9%), especially in `gender`, `company_size`, `company_type`, and `major_discipline`. These four predictors' compete_rates are $< 90\%$. The variables are missing at random (MAR), which means the missingness depends only on the observed data. Therefore, I consider to use imputation method. I choose median imputation to impute the missing values in the trainData and testData separately. Since all the missing data are categorical variables, median imputation seems better than knn and bag imputation. Because knnImpute and bagImpute return digital values for the missing data, which is not appropriate for our categorical data.

visulization

From the Gender Distribution plot and Education Level by Gender plot, we can see that there are too many males(93%) in the training data set, which indicates that gender is a biased variable, and it is not a good predictor. However, I am still going to keep this variable in the prediction model.

From the density plot, we can see the distribution of the only two continuous variable in the data set `city_development_index` and `training_hours`.

What predictor variables did you include?

I include city_development_index, training_hours, gender, relevent_experience, enrolled_university, education_level, major_discipline, experience, company_size, company_type and last_new_job as predictor variables.

What technique did you use? What assumptions, if any, are being made by using this technique?

Since we are going to predict binary response, I choose 6 models for classification: GLM, GLMN, LDA, QDA, GAM and Mars. GAM and MARS models assume nonlinear relationship between the dependent variable and covariates. LDA assumes equality of covariances among the predictor variables X across each all levels of Y, and this assumption is relaxed with the QDA model.

If there were tuning parameters, how did you pick their values?

Finding the tuning parameter is to find the optimal value. We can use bestTune in the model to figure out the tuning parameters.

Discuss the training/test performance if you have a test data set.

Which variables play important roles in predicting the response?

From the importance plot in MARS, we can see that city_development_index is the most important variable. Other variables like relevent_experience, education_level, company_size, enrolled_university, and last_new_job also play important roles in predicting the response.

What are the limitations of the models you used (if there are any)? Are the models flexible enough to capture the underlying truth?

The limitations of GAM and MARS are that they cannot include interaction terms, so that we don't know if there's interaction between variables. Also, these two models are slower to train.

Conclusions

What were your findings? Are they what you expect? What insights into the data can you make?

After comparing the six models and their AUC values, we can see that MARS model has the largest AUC = 0.779. This indicates that MARS model has a better performance than other models. All these models'

AUC are close to 0.7, which means there is a 70% chance that the model will be able to distinguish between positive class and negative class. From the ROC curves, we can also see that MARS (purple ROC curve) is more efficient than other models since the ROC curve is closer to the upper left corner.