P8106_MidtermProject_mb4757 Report

Minjie Bao

3/29/2021

Introduction

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 collected from the information that the candidates provided. There are 19158 rows and 14 columns in the raw data set. 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 for the courses and categorization of candidates.

We want to understand that what factors lead a person to leave their current jobs. We are going to predict the probability of weather a candidate will look for a new job or will work for the company, and interpreting affected factors on employee decisions.

Data preparation/cleaning

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, which can substitute to `city` variable, and `enrollee_id` variable is not useful for model prediction. I keep all the other variables: city_development_index, training_hours, gender, relevent_experience, enrolled_university, education_level, major_discipline, experience, company_size, company_type and last_new_job as predictors. For the outcome variable `target`, I recode the values "1" and "0" 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.

To dealing with the missing data, the whole dataset 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. Figure 1 is a visualization plot of missing data.

Exploratory analysis/Visualization

From the Gender Distribution table and Education Level by Gender plot in Figure 2, 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 feature plot in Figure 3, we can see the distribution of the only two continuous variable in the data set `city_development_index` and `training_hours`.

Models

Since we are going to predict binary response, I choose 6 models for classification: GLM (logistic regression), GLMN (penalized logistic regression), LDA, QDA, GAM and MARS. Each model has different assumptions. For GLM logistic regression model, it assumes independence of errors, linearity in the logit for continuous variables, absence of multicollinearity, and lack of strongly influential outliers. 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. For the Mars model, I choose tuning parameter: degree = 1:3 and nprune = 8:15 because I have 11 predictors. For the confusion matrix I choose 0.5 as the cutoff point.

I use confusion matrix to compare the training and testing performance. From the confusion matrix output by using test data: The accuracy is 0.7716, which means the overall fraction of correct prediction is 0.7716 with 95% CI between 0.758 and 0.7848. The NIR (No Information Rate) is 0.7507, which is the fraction of "Yes" class in both predicted and trained dataset. The p-value is 0.001374 < 0.05, which means we reject the null hypothesis and conclude that accuracy > no information rate. The kappa value is 0.2467, which is the agreement between the predictive value and the true value. A kappa value of 1 represents perfect agreement, while a value of 0 represents no agreement. The sensitivity is 0.25759, measures the proportion of actual positives that are correctly identified TP/(TP+FN). The specificity is 0.94228, measures the proportion of actual negative that are correctly identified TN/(FP+TN).

The confusion matrix results by using training data: The accuracy is 0.7644. The NIR (No Information Rate) is 0.7506. The p-value is 3.855e-05 < 0.05. The kappa value is 0.221. The sensitivity is 0.23993. The specificity is 0.93864. Comparing the results with testing data, we can see that the accuracy of confusion matrix by using training data is higher than test data, NIRs are similar, p value, kappa value, sensitivity and specificity of training data are smaller than test data.

From the importance plot in MARS (Figure 4), 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.

Comparing the six models' ROC curves (Figure 6) and their AUC values: AUC for GLM = 0.736, AUC for GLMN = 0.735, AUC for LDA = 0.734, AUC for QDA = 0.732, AUC for GLMN = 0.747, and AUC for MARS = 0.786. We can see that MARS model has the largest AUC = 0.786. 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.

Limitation

This data set contains too many categorical variables and there are only two continuous variables:`city_development_index` and `training_hours`. This dataset doesn't contain enough essential factors such like age, offered salary, accommodation, and the company profile. As we know that the employee satisfaction index is a key for making such decision. This dataset can include more essential factors to help the company create a more accurate model.

For models, the major limitation of GLM Logistic Regression model is the assumption of linearity between the dependent variable and the independent variables. Also, Logistic Regression requires average or no multicollinearity between independent variables. Logistic regression attempts to predict binary outcomes based on a set of independent variables, but logit models are vulnerable to overconfidence. That is, the models can appear to have more predictive power than they actually do as a result of sampling bias. This will cause overfitting problem. A disadvantage of QDA is that it cannot be used as a dimensionality reduction technique. The limitation of GAM and MARS model is slower to train the model.

Conclusions

In conclusion, MARS model is the best model with AUC = 0.786. The important variables: city_development_index, relevent_experience, education_level, company_size, enrolled_university, and last_new_job will lead a person to leave their current jobs. The city_development_index is the most important variable that will affect people to look for a new job or work for the Data Science Company. We can see a clear relationship between `city_development_index` and `target` from the partial dependence plot in Figure 6. The people in the city with city_development_index > 6.2 are more likely to change their jobs. As the city_development_index increases, the predicted `target` value more approach to 1. Although there seems to be a small decrease of job changes in city_development_index between 0.91 and 0.93, the total plot trend is increasing. This indicates that people in a high development city tend to look for a new job actively, which makes sense since high development cities always have more opportunities than low development cities.

Figure 1:

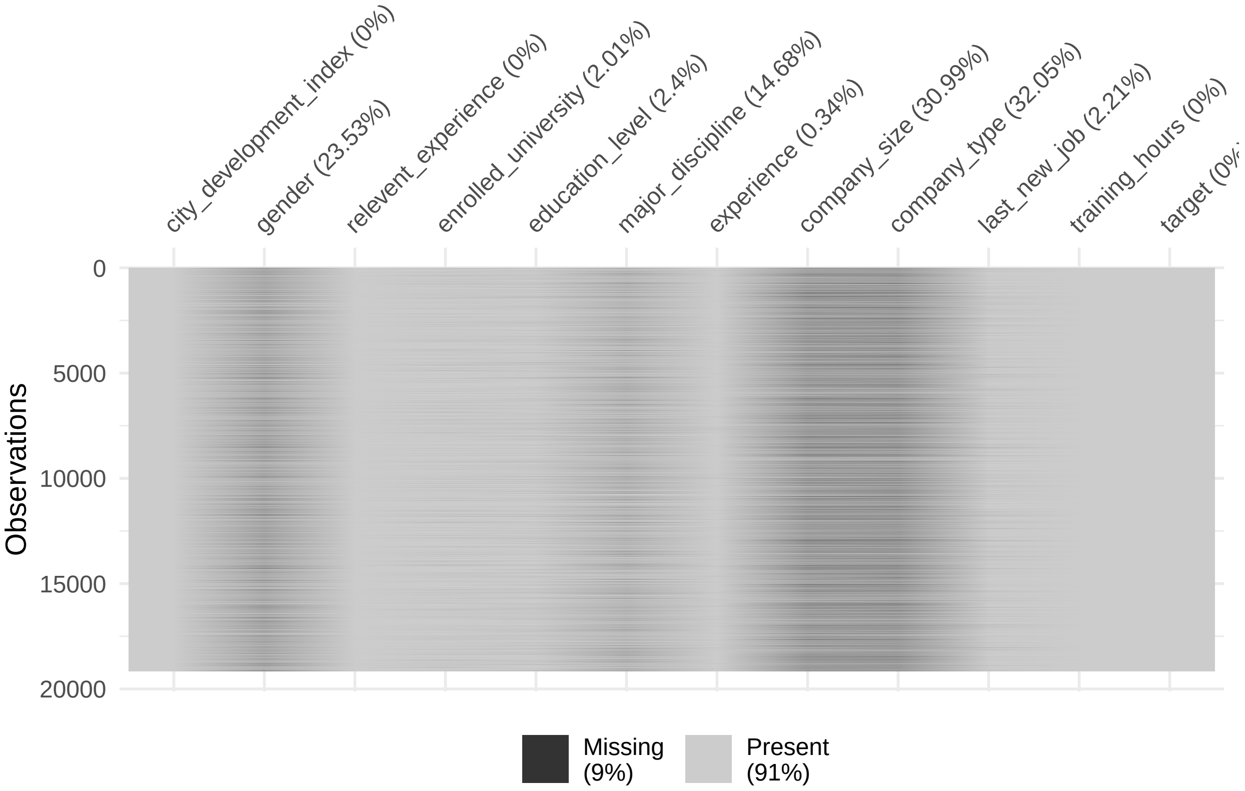


Figure 2:

gender	total	Percent
Male	14188	93%
Female	985	6%
Other	154	1%

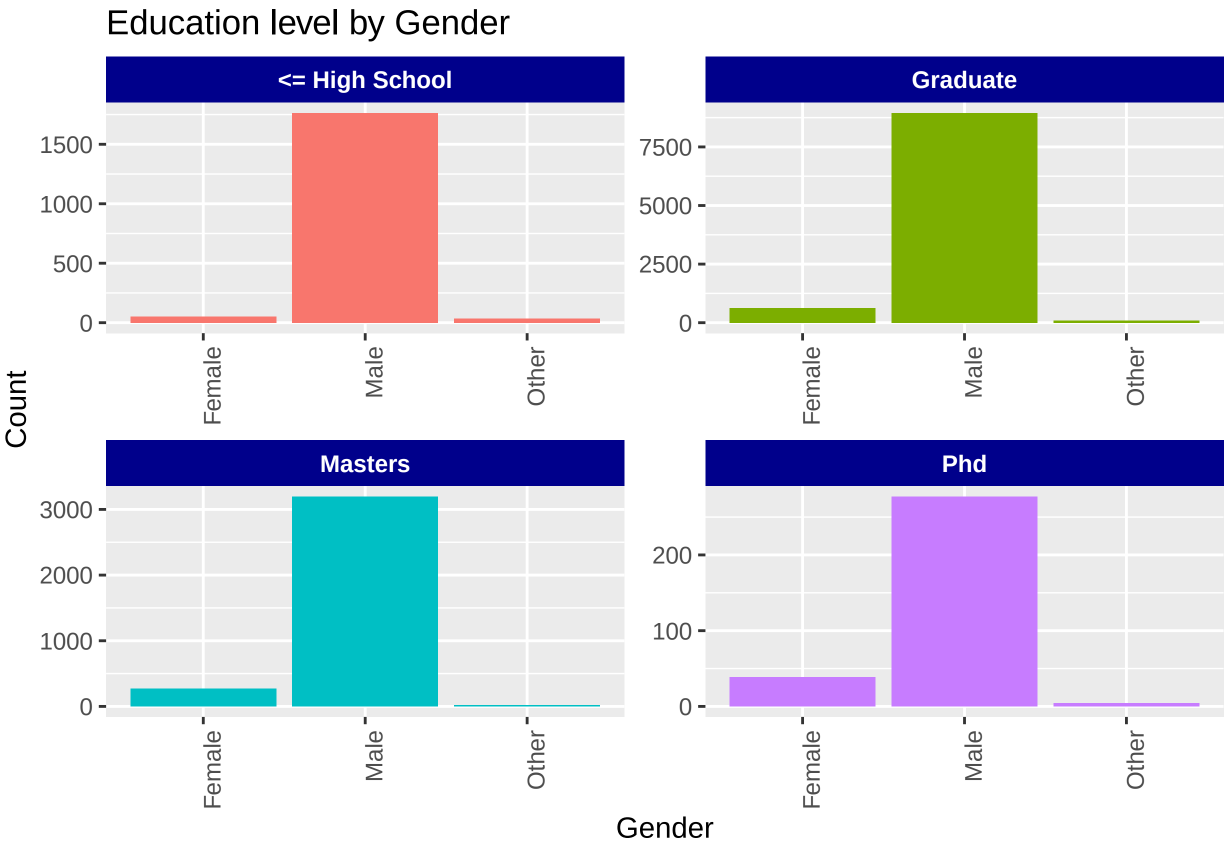


Figure 3:

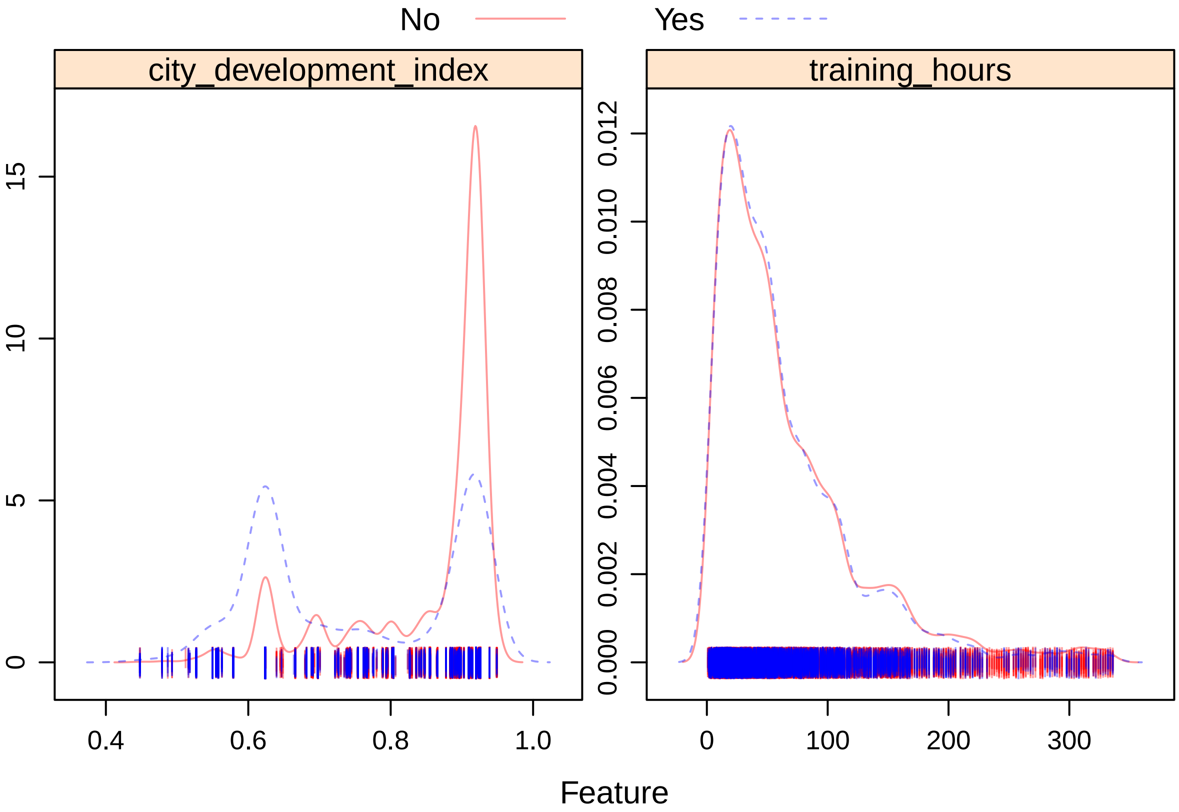

No
Yes
city_development_index
training_hours
0
5
10
15
0.4
0.6
0.8
1.0
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0
100
200
300
Feature

Figure 4:

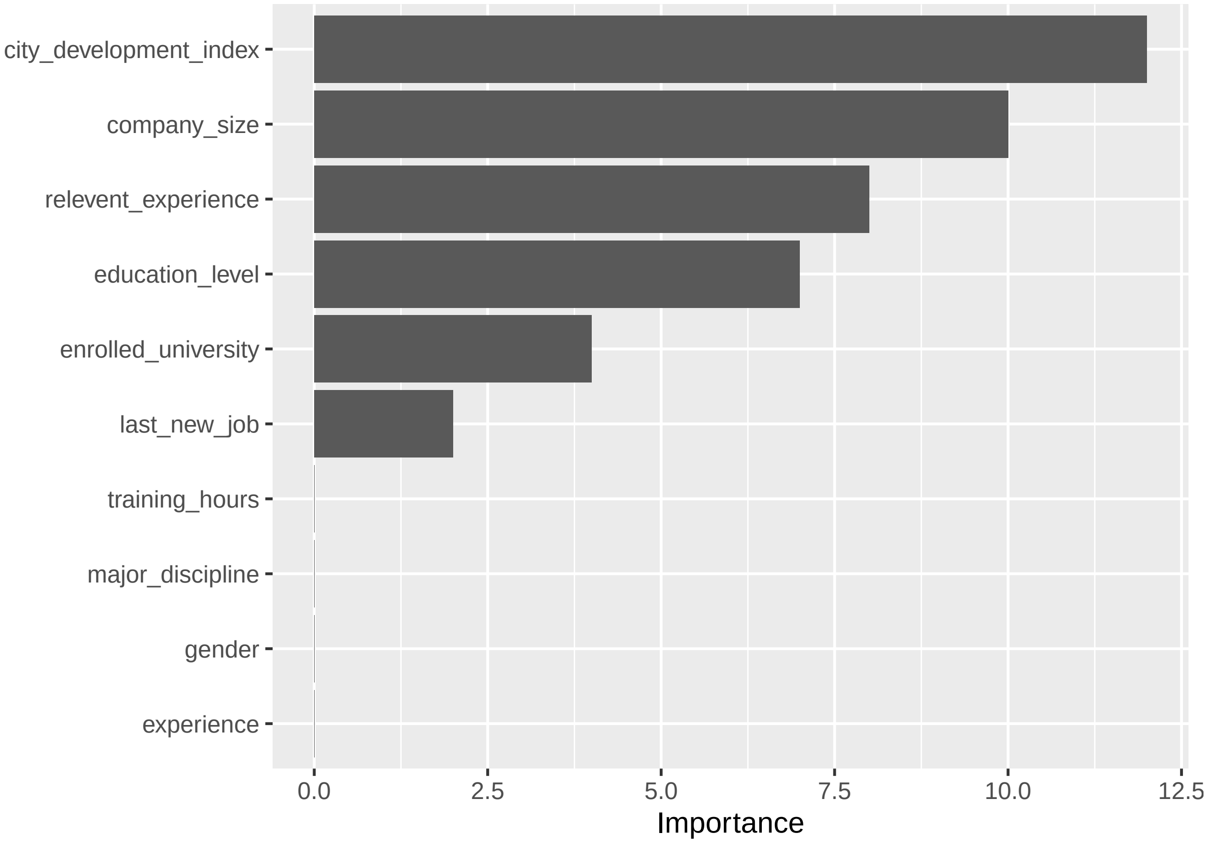

city_development_index
company_size
relevent_experience
education_level
enrolled_university
last_new_job
training_hours
major_discipline
gender
experience
0.0
2.5
5.0
7.5
10.0
12.5
Importance

Figure 5:

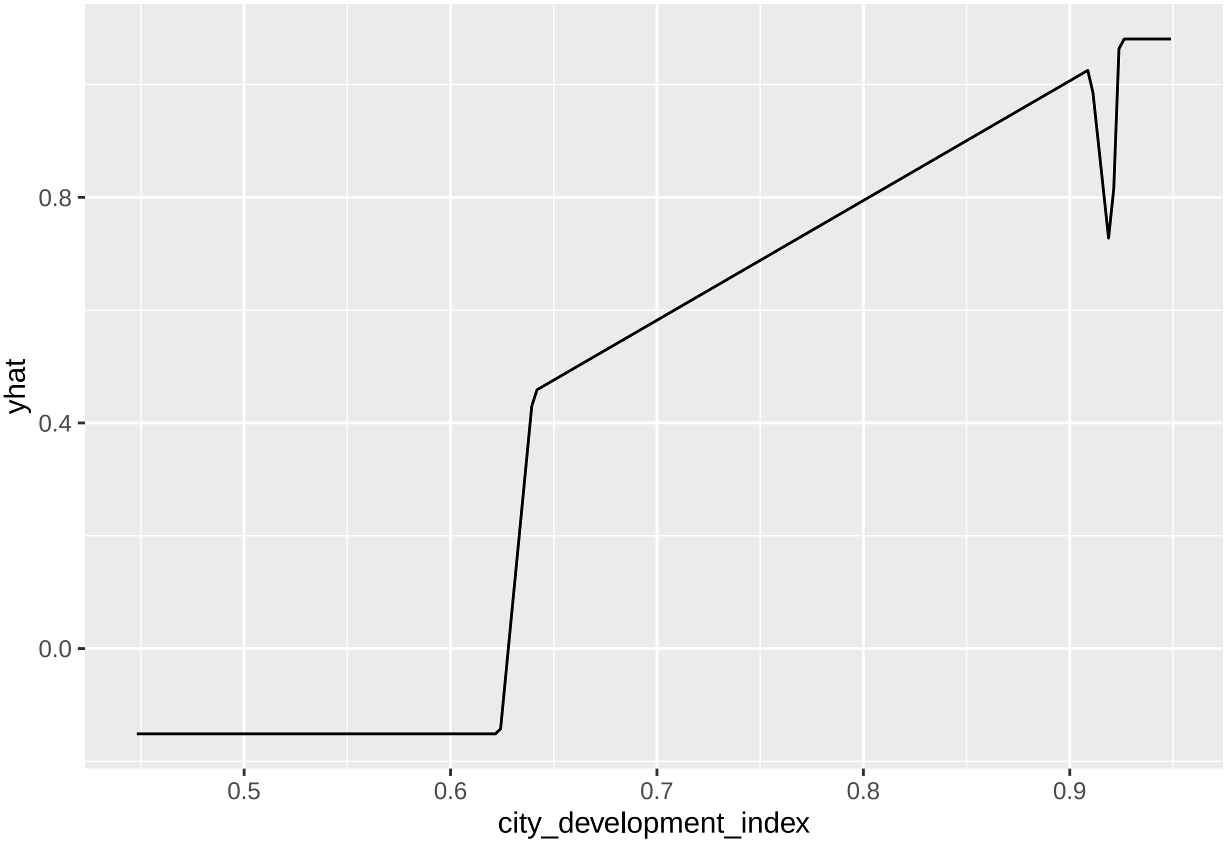
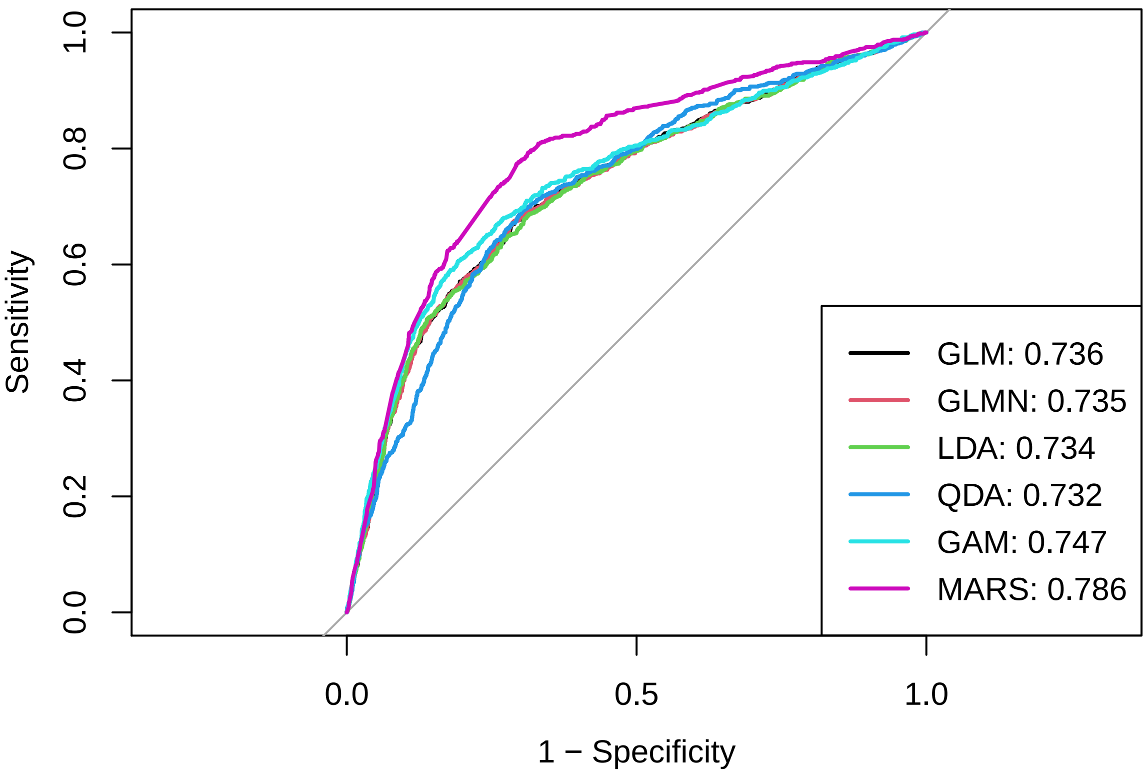


Figure 6:



Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1 − Specificity
GLM: 0.736
GLMN: 0.735
LDA: 0.734
QDA: 0.732
GAM: 0.747
MARS: 0.786