ST443 Group Project:
Applications of Machine Learning Techniques on Real Word Data

Department of Statistics
The London School of Economics and Political Science

Group Members:
14876 (Contribution - 33.33%)
11245 (Contribution - 33.33%)
10722 (Contribution - 33.33%)

December, 2020

Contents

1 Applications of Machine Learning Techniques on the IBM Employee Attrition Data

1.1 Introduction

1.1.1 Data Description & Objective

There are many reasons that cause an employee to leave a company, however, from the company's perspective, having a high employee attrition rate would always make the company suffer from a high human resource cost and an unstable employee management system. In this project, we use the *IBM HR Analytic Employee Attrition & Performance* data set that originally has 1470 observations and 35 features. We are interested in predicting the attrition status of an employee and exploring the factors that would mostly affect a person's decision of whether to leave a company. The objective of our project is to apply different machine learning methods on this classification problem and find the best one in terms of having a good balance between high prediction accuracy and reasonable model complexity.

1.1.2 Methods & Performance Measurement

12 classification methods are used in this project. In addition, 10-fold cross validation is used when performing each model, AUC (under ROC) and sensitivity (i.e. the true positive rate, also known as recall) are considered when comparing performances of these models. Sensitivity is preferred since companies may pay more attention on the number of people that are correctly classified.

1.2 Data Preprocessing

1.2.1 Data Cleaning

The data set is clean in terms of having no missing values. However, there are more cleaning procedures we conducted in order to obtain consistent and meaningful analysis. We first drop the four variables (e.g. Employee Count) that are invariant over time and irrelevant to our analysis. We then change the class of ordinal variables (e.g. Job Satisfaction) that recorded as numerical values from numeric to factor. We also group several indicators that measure different aspects of job satisfaction into the "Total Satisfaction" variable. Lastly, we standardise all numerical variables and convert all factors to dummy variables.

1.2.2 Exploratory Data Analysis

We find that people work as Laboratory Technicians and Sale Executives contribute 26.2% and 24.1% of the attrition at IBM, respectively. In addition, employees with a bachelor degree contribute most of the attrition at 41.8%, however, the attrition rate of experts who holds a PhD degree is only 2.1%. It is worth noticing that people with a bachelor and at the age around 30 have the most incentive to leave the company, they form over 50% of the attrition.

Overall, the data shows that employees who have more work overtime experience and more business travel experience tend to have low total job satisfaction. These people also are more incentive to leave the company.

1.2.3 Imbalanced Data Problem

This data set is imbalance as 83.9% of the respondents still stayed at the company and only 16.1% left the company at the time that the survey was conducted. To mitigate this issue, two approaches are considered:

1. We choose to use the SMOTE (Synthetic Minority Oversampling Technique) to generate "synthetic" data from the minority class. However, it may lead to over-fitting problem if k-fold cross validation (CV) is applied after conducting this technique. Thus, we decide to apply SMOTE during the re-sampling procedure, this is allowed by specifying the 'sampling' argument in the trainControl function from the caret package.

2. As an alternative, we also apply the stratified 10-fold CV which ensures that the distribution of the attrition variable in each fold is the same as the distribution in the origin data set.

After conducting the two approaches mentioned above for each classification method, we find that the stratified 10-fold CV performs much better than the SMOTE especially in terms of sensitivity (see Appendix for the comparison). One possible reason is that the SMOTE may generate more noise rather than meaningful data within the high-dimension setting of the data set. Therefore, we decide to use the stratified 10-fold CV for model comparison and further analysis.

1.3 Experiments & Results

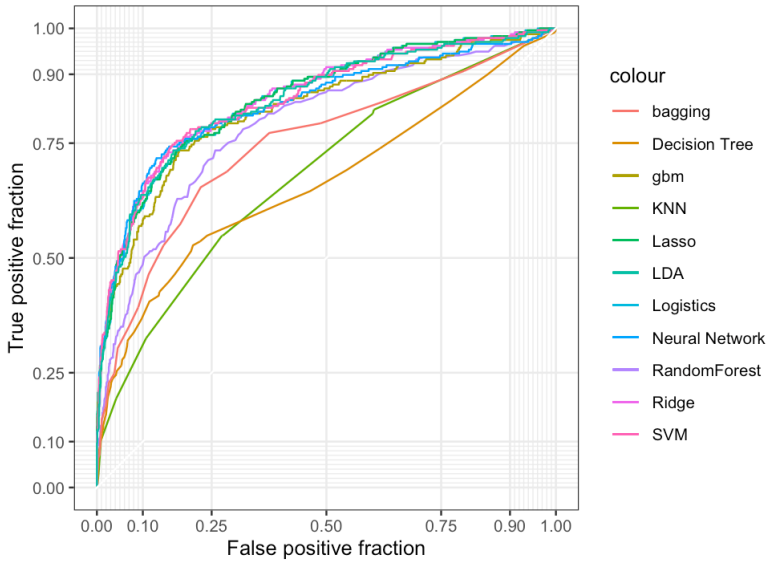

Figure 1: ROC Curve

	AUC	Sensitivity
Ridge	0.852	0.977
Logistic	0.851	0.961
Lasso	0.851	0.962
SVM	0.848	0.968
NeuralNetwork	0.844	0.959
LDA	0.844	0.967
gbm	0.824	0.989
AdaBoost	0.821	0.987
RandomForest	0.801	0.999
Bagging	0.780	0.964
KNN	0.692	0.989
DecisionTree	0.671	0.974

Figure 2: Model Performance

From the above table and graph, we observe that the logistic regression, SVM, LDA and Neural Networks have the highest AUC and sensitivity while KNN (K=5) and decision tree provide the worst performance.

1.3.1 Reviews of the Best-performed Approaches

Logistic regression and LDA are used to predict the probability that a person would resign, where logistic regression tries to maximize its log-likelihood function, and LDA tries to maximize the discriminant function, which is proportional to its posterior probability.

SVM is a non-parametric classification method that can provide a non-linear decision boundary. This method uses kernels to enlarge the original feature space where the enlarged space is linearly separable. We also conduct the neural network algorithm for this binary classification problem since it is one of the most popular supervised learning methods that is widely used in many machine learning fields. Inspired by the working mechanism of human/animal brains, this method consists of forward-propagation and back-propagation, which helps the computer "understand" the input data better and provide more accurate results if sufficient data is provided.

We also apply both gradients boosting and adaptive boosting (AdaBoost) procedures, the gradient boosting technique sequentially fits the decision tree to the residuals whereas AdaBoost fits the tree to the misclassified observations by giving them heavier weights.

1.3.2 Best Model Selection & Tuning

We consider the logistic regression as the best method since it not only has the greatest performance but also it is less complex compared with other models. To further improve the performance of the logistic model in the sense of avoiding overfitting and reducing the variance, shrinkage methods including both of the Ridge and

Lasso procedures are conducted. The optimal value of the tuning parameter lambda for the Ridge and Lasso is chosen at 0.005 and 0 by performing 10-fold CV, respectively. From figure 2, even though Ridge performs slightly better than the logistic model, in order to have a more interpretable model, we choose the original logistic regression as the final model.

Below, we show the first six most important factors that may affect one's attrition status. Combining with the estimated coefficients of these variables, we conclude that people with long working time, low total satisfaction, high number of previously worked company, high frequency of business travel, long distance from home and less stock option are more likely to leave the company, which is consistent with our EDA result.

	estimated_coefficient <dbl>	Importance <dbl>
OverTimeYes	2.0971717	10.213799
TotalSatisfaction	-0.8735792	8.836311
NumCompaniesWorked	0.5135197	5.056781
BusinessTravelTravel_Frequently	2.1123731	4.899038
DistanceFromHome	0.4057232	4.434587
StockOptionLevel1	-1.2599239	4.017813

Figure 3: Importance of Factors

1.4 Conclusion

In this project, we aim to predict employee's attrition at IBM, the logistic regression model is chosen as the best one since it has the strongest prediction power, a good interpretability and a reasonable model complexity. Factors such as overtime working, job satisfaction and frequency of business travel are found to be closely related to the person's decision of leaving the company. Therefore, a company that suffers from high attrition rate may need to develop various strategies that provide more benefits to its employees to increase employees' loyalty to the company.

2 Applications of Machine Learning Techniques on the Concrete Compressive Strength Data Set

2.1 Introduction

2.1.1 Data Description & Objective

Concrete is one of the most important construction engineering materials used in almost every structure of civil engineering, such as road, bridge, and water conservancy. It is mainly used to bear load or resist any forces, thus, the compressive strength is considered to be the most crucial property for evaluating the quality of concrete. In this project, we aim to predict the compressive strength of concrete based on the *Concrete Compressive Strength Data Set* given by the UCI Machine Learning Repository. We are also interested in finding the factors that are most related to the concrete compressive strength. This dataset consists of 1030 observations and 9 columns, including the dependent variable CC.Strength and the other 8 variables such as cement, fly ash and water that are measured in kg in a m^3 mixture.

2.1.2 Methods Performance Measurement

We aim to find the best model that has both high prediction power and reasonable model complexity. Twelve regression methods are applied to this regression problem. 10-fold cross-validation (10-fold CV) is used throughout the procedure of evaluating the performance of each model since it provides a much more concrete estimation of the prediction error than the validation set approach. MAE (Mean absolute error), RMSE (Root Mean Squared Error), and R-squared are the three evaluation metrics that we use in this project.

2.2 Exploratory Data Analysis & Data Standardisation

There are 9 numeric variables in the dataset. To explore how each variable is distributed, we plot the histograms of them and find that variables such as blast-furnace slag, fly ash and super plasticizer are noncontinuous and right-skewed with more than 80% centered around 0. Additionally, since the age (days) variable only has 14 discrete values, we divide the data of the age variable into 4 groups and found that about 66% of the observations fall into the age interval $(0, 100]$.

We further plot the scatter plots between CC.Strength, age group, and other variables. We find that the compressive strength increases with age, cement, and superplasticizer, decreases with the amount of water used. We also use the corrplot() function to see the actual correlation between variables(see Appendix). From the plot, we can tell that most of the explanatory variables are weakly correlated, with the correlation < 0.5. However, the variable water and superplasticizer are, to some extent, strongly correlated, with the correlation of -0.66, which makes sense since the superplasticizer is commonly used as a good substitution of water in construction.

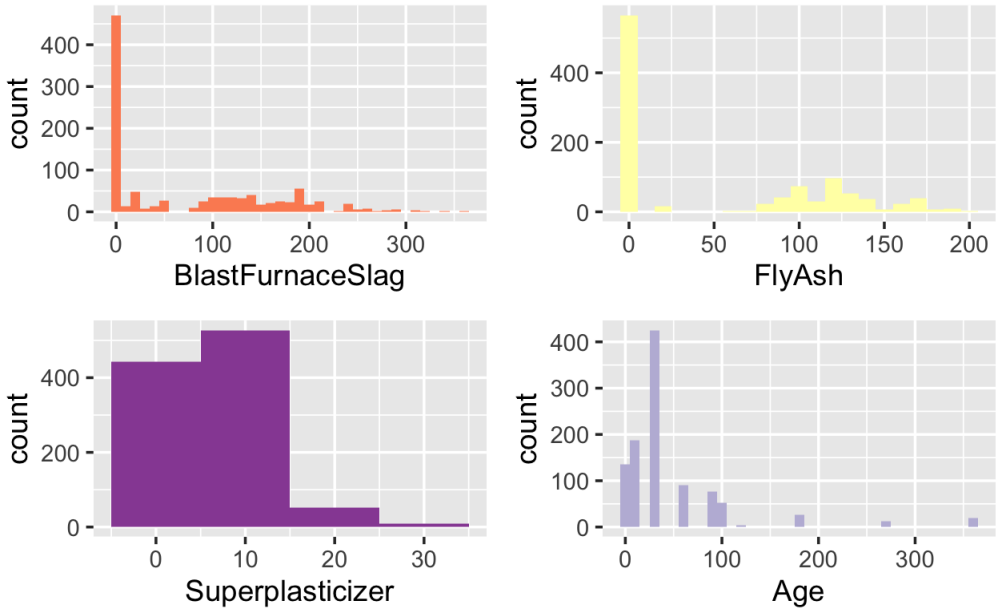

Figure 4: Histograms of variables

2.3 Experiments & Results

2.3.1 Summary of Results

Model	MAE	RMSE	Rsquared
GBM	2.629	3.854	0.946
RandomForest	3.247	4.587	0.928
Bagging	3.768	5.096	0.907
GAM	5.423	6.944	0.829
KNN	6.781	8.994	0.717
Lasso	8.291	10.413	0.616
Linear	8.290	10.435	0.610
ElasticNet	8.302	10.454	0.611
Ridge	8.470	10.554	0.601
SVM	8.203	10.786	0.597
DecisionTree	10.084	12.497	0.447
NeuralNetwork	34.818	38.609	NA

Figure 5: Model Performance

For the linear regression-based methods, we first conduct the simple linear regression then consider adding penalty terms (lasso, ridge, and elastic net) to improve the out-of-sample performance. For the tree-based nonlinear regression models, we apply the decision tree, bagging, random forest, and gradient boosting machines (GBM). Other methods, including Generalized Additive Model (GAM), KNN (with K=5), SVM, and Neural Network are also considered.

From the table, we conclude that GBM has the best performance in terms of having the lowest MAE and RMSE and the highest R-squared. Furthermore, the linear regression-based methods have very similar prediction power, and they are much less predictive than the nonlinear methods except for the simple decision tree.

2.3.2 Reviews of the Tree-Based Approaches

The GBM is an additive model that fits small trees to the residuals and sequentially add the tree to the previously fitted one to minimize the loss function and update the residuals. Bagging applies the bootstrap approach such that it fits a tree on each bootstrapped dataset and average all the trees to reduce the high variance of the single tree method. A similar approach is applied by random forest, but it only selects a random subset of all the predictors as split candidates when fitting the tree.

From the results, we observe that the performance of the tree-based models is significantly better than the family of linear regression models. Given the previous EDA, we know there exists relatively strong colinearity between certain variables; one possible reason is that the tree-based models can handle the colinearity better than the linear regression models. In addition, since there are certain variables in the dataset clustered at 0 but highly right-skewed, which is harmful to the prediction power of the linear regression model due to the high-leverage point effect; however, the tree-based models would not be affected by this issue as it has a much more flexible structure in general.

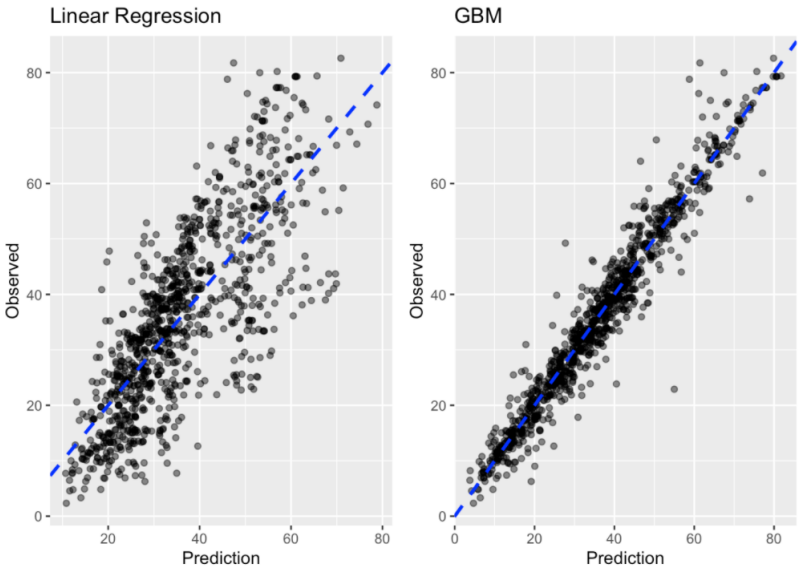

Figure 6: LR vs. GBM

2.4 Best Model Selection

Even though GBM is less interpretable than linear regression, based on the excellent performance of the GBM, there is no doubt that we would choose it as our final choice. Additionally, we are able to find the relative influence that measures how much each covariate contributes to the reduction of the sum of squared error loss function. We can see that age and cement are the two most important features that determine the compressive strength of concrete. This result makes sense since the correlation plot suggests that the correlation between cement and compressive strength is the highest among all the explanatory variables.

var <chr>	rel.inf <dbl>
Age	32.321207
Cement	30.007730
Superplasticizer	9.471336
Water	9.366036
BlastFurnaceSlag	9.007580
FineAggregate	4.822232
CoarseAggregate	3.537793
FlyAsh	1.466085

Figure 7: Relative Importance

2.5 Conclusion

In this project, we apply different regression machine learning methods to predict the concrete compressive strength. We observe that the tree-based methods, especially the Gradient Boosting Machine, have the best performance. Furthermore, the age (days) of the concrete and the use of cement (kg/m^3) are the most important factors in evaluating the concrete compressive strength.

3 Appendix

Project 1:

	AUC	Sensitivity
SVM	0.850	0.835
Logistic	0.847	0.830
Lasso	0.847	0.839
LDA	0.846	0.818
Ridge	0.845	0.832
NeuralNetwork	0.839	0.848
gbm	0.835	0.919
AdaBoost	0.824	0.900
RandomForest	0.797	0.959
Bagging	0.769	0.886
DecisionTree	0.713	0.861
KNN	0.706	0.635

Figure 8: Model Performance with SMOTE CV

	AUC	Sensitivity
Ridge	0.852	0.977
Logistic	0.851	0.961
Lasso	0.851	0.962
SVM	0.848	0.968
NeuralNetwork	0.844	0.959
LDA	0.844	0.967
gbm	0.824	0.989
AdaBoost	0.821	0.987
RandomForest	0.801	0.999
Bagging	0.780	0.964
KNN	0.692	0.989
DecisionTree	0.671	0.974

Figure 9: Model Performance with stratified CV

Project 2:

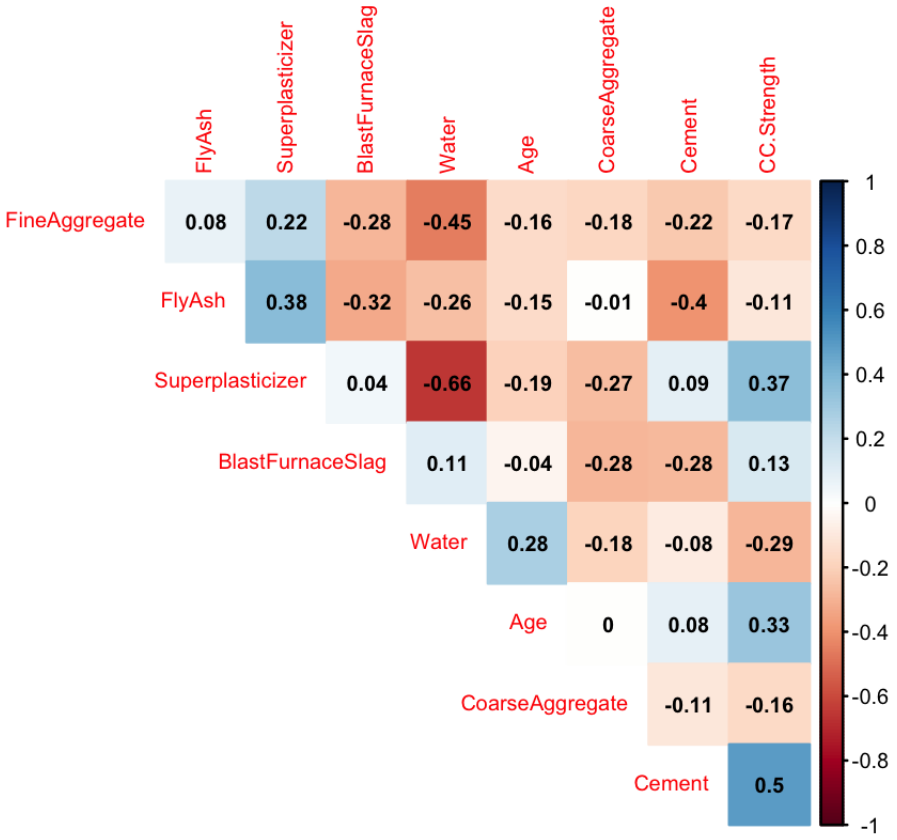

Figure 10: Correlation Plot