Statistical & Machine Learning Individual Project

MSc in Big Data Analytics for Business
Elina KELLY

Contents

Introduction

Machine learning is frequently used for making predictions that are performed through algorithms that are trained with data. It can be divided into 2 main groups: the unsupervised learning method, which means that the data we train only contains features, and the other approach is the supervised learning, so the training set is labeled, in other words, the target 'variable' is already known. The latter option will be used in this application, for a classification problem with a qualitative binary response. The data set used is named "Credit Card Default" and includes features that describe a client's profile and the target variable that indicates whether a default will occur the next month. The goal will be to build 5 machine learning algorithms after polishing the data, then apply the models and compare the results thanks to their accuracy in order to find the optimal approach.

I/ Preprocessing

I.a. Missing Values

Out of the 20 000 rows and 25 columns in the Credit Card Default data set, 4448 values were missing. It was important to treat them in order to produce more effective models and to minimize bias in later steps. The missing categorical values were treated differently as the numerical ones.

Missing Categorical Values

After defining all the variables in the data set that contained categorical values, the missing values were filled with the class that occurred the most frequently in each specific column. Therefore, these missing values were imputed based on the median.

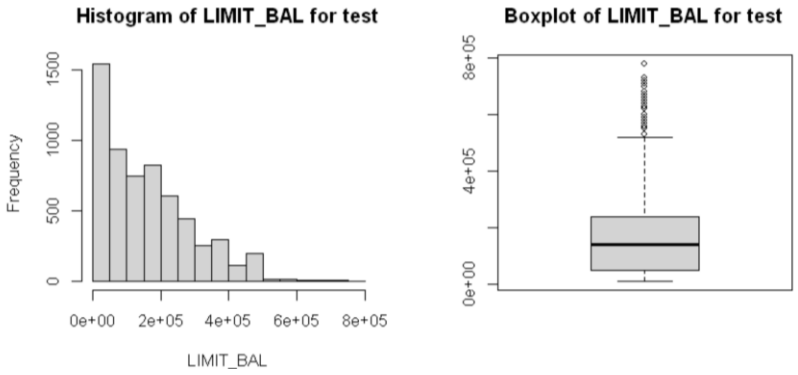
Missing Numerical Values

Handling missing numerical values required a splitting process. By splitting the data set into a training set (70%) and a test set (the remaining 30%), the missing values in each column in the training set was replaced by the mean of their corresponding column in the training set.
Whereas the missing values in the test set were also replaced with the mean of the columns from the training set.

In addition, the numerical value age was transformed into a categorical variable by defining 10-year ranges. For instance, if a person is 24 years old, in other words, if they are in their 20s, they would be placed in the category 20 in the new variable 'age_group'.

I.b. Outliers

To verify the regularity of the numerical variables, it was useful to plot their distribution.
The example below shows the variable 'LIMIT_BAL' in the testing set and the corresponding frequency for each value. Visually, we can say that the data presents outliers after ~5e+05, the maximum value of the variable equals 780 000 and indicates that it is 5 times the amount of the mean of the variable (equal to 168 402) in the testing set.

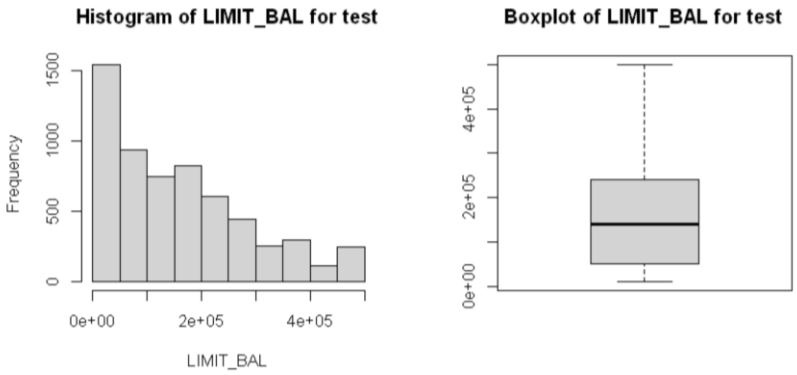


A method named winsorization was used to fix the issue of outliers. Winsorisation involves replacing the largest or smallest value of the data set with observations closest to them.
In the example above with the variable 'LIMIT_BAL' in the test set, winsorization helped to replace the largest value with values closest to them. This process limited outliers by replacing those largest extreme values with less extreme ones, as seen in the graphs below.

More precisely, this step involved determining a threshold, the defined threshold for both the training and test set was the sum of the 3rd quartile (75% of data is placed below) of the training set and a constant 1.5 multiplied by the interquartile range of that variable in the training set (so Q3 + 1.5 * IQR and Q1 - 1.5 * IQR for values in lower extreme values).
Then the next step was to replace all the values above this benchmark with the values defined in the benchmark.

Before winsorization … After winsorization …

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.			Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
training	10000	50000	140000	165963	230000	1000000	⟵	⟶	10000	50000	140000	165278	230000	500000
testing	10000	50000	140000	168402	240000	780000	⟵	⟶	10000	50000	140000	167568	240000	500000



This process was applied to most of the numerical variables in the training and testing set as many of them presented outliers.

I.c. Feature Selection

Feature selection is an important step before applying a model as it improves model interpretability and increases the model's accuracy. In this application, we use the direct approach which reduces the number of features by removing the irrelevant ones that don't affect the target variable much or those that are highly correlated and contain the same information.
Then in the next step, the models will be fitted on the reduced selection of variables.

Forward Stepwise Selection

The forward selection approach starts off with an empty model with no features, then adds them in one by one by selecting the features that have the smallest RSS and the highest R^2.
Advantage: Forward selection starts by considering small size models, so reduces the chances of collinearity among features.
Disadvantage: This method has a path dependency, the first uniquely selected feature must also be selected in the size 2 model, and in the size 3 model, etc.

Backward Stepwise Selection

As opposed to the forward selection model, the backward selection begins with all the features, then reduces the number by removing the least significant features one by one.
Advantage: This approach considers more combinations of features at first than the forward selection, and the forwards selection may not be able to reach because of its limitations.
Disadvantage: There must be a greater number of samples than number of features.

Best Subset Selection

The best subset selection model considers all the possible models with first only 1 feature, then 2, then 3, etc. until all features are in the model. Then it selects the best model of size 1 (model containing only 1 feature), size 2 (model containing a combination of 2 variable), size 3, etc. Finally, it chooses the best overall model, again, with a low RSS and high R^2.

Advantage: It considers every possible combination between each feature, and has a higher chance of finding the best possible model.
Disadvantage: The number of models with different sizes will increase exponentially with the number of features available in the data set.

Comparison

By evaluating the different feature selection models (table below on the left), we find that the one with the best performance corresponds to the Best Subset Selection with the features presented in the list on the right below.

Model	Evaluation
<chr>	<dbl>
Forward Selection adjr2	0.1505473
Forward Selection BIC	0.1505800
Forward Selection CV	0.1505651
Backward Selection adjr2	0.1505461
Backward Selection BIC	0.1508329
Best Subset Selection	0.1504855

BSS Train Variables	
	<dbl>
(Intercept)	3.376896e-01
EDUCATION	-1.816416e-02
MARRIAGE	-1.760113e-02
PAY_0	9.135788e-02
PAY_2	1.971939e-02
PAY_3	1.334314e-02
PAY_5	1.533207e-02
BILL_AMT1	-8.817982e-07
BILL_AMT3	4.690423e-07
PAY_AMT1	-5.861775e-06
PAY_AMT2	-4.862437e-06
PAY_AMT3	-3.261560e-06
PAY_AMT6	-3.152566e-06
age_group	1.034231e-03

II/ Modeling

II.a. Logistic Regression

Logistic regression measures the probability that a response variable belongs to a specific category. In the credit default data set, the model measures the probability that the customer will default next month with the qualitative binary response ranging between 0 and 1 for the variable 'default.payment.next.month'.

The logistic regression model is defined by the function:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

Where $X = (X_1,\ X_2,\ \ldots,\ X_p)$ represents p predictors. This function measures the probability of a customer having a default next month and provides outputs only between 0 and 1 and is always represented graphically by a S-shaped curve.
The maximum likelihood method is used to fit the model on the training set and estimate the unknown regression coefficients $\beta_0,\ \beta_1,\ \ldots,\ \beta_p$.

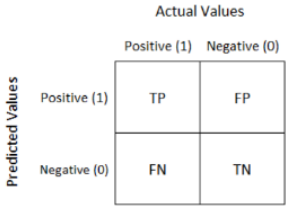
$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} \left(1 - p(x_{i'})\right)$$

By maximizing this likelihood function to fit the logistic regression, we maximize the probability or likelihood of obtaining the same outputs as the actual data with those coefficients. In other words, the higher the likelihood, the closer the outputs are to the actual categories.
As we are dealing with a logistic regression, this function assumes a Binomial probability distribution, meaning that it measures the probability of the number of successful outcomes, that is why the maximum likelihood function has a similar form to the Bernouilli distribution:

$$P(Y_i|x_i;\theta) = h(x_i)^{y_i}(1-h(x_i))^{1-y_i}$$

In this application, the logistic regression was fit on the training set, then predicted on the test set with the previously selected variables. One of the ways to verify the effectiveness of a model is by building a confusion matrix:

```
pred_class1    0    1
       Down 4550  987
       Up    143  320
```



https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

This matrix indicates that 4550 values were predicted as having default next month and was in fact the case, and 320 were predicted as non-defaulters and it was the actual case. The errors below indicate that for error type 1, the model had predicted non-defaulter as defaulters for 987 of the values, and that it predicted non defaulters, but it was false for 143 values.
The accuracy is then deducted by calculating the ratio of the correct predictions over the total predictions.

Advantages: Simple and highly interpretable.
Disadvantages: Logistic regression is limited, in the sense that is can only handle binary classification problems properly.

… Cross validation

The logistic regression performed slightly better on the training set with the previously selected features demonstrating an accuracy of 0.8116 versus 0.8113. Applying the k-fold cross validation measures the logistic regression model's performance on different portions of the data set. To do this, it first split the data into 10 folds in this context, trained the model on a portion by leaving a small one out, then testing it on that one. The cross validation repeated these steps until each portion had been in the testing portion. At each step, the model had been calculating the average prediction error, therefore, at the end, it computed the average of the collection of these errors, giving a final accuracy of 0.806 for the model after being applied on the unseen testing set.

```
Generalized Linear Model

20000 samples
   13 predictor
    2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 18000, 18000, 18001, 18000, 17999, 18001, ...
Resampling results:

  Accuracy   Kappa
  0.8059004  0.2524231
```

II.b. Linear Discriminant Analysis

LDA is an alternative approach to the logistic regression, they essentially differ in the fitting step. LDA makes predictions separately to calculate the probability that an input belongs to each of the individual output categories. The output category corresponds to the category with the highest probability, that is measured with the Bayes Theorem:

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

Where π_k (also called prior probability) represents the probability that an observation belongs to a class in the training set. Considering this application involves a binary classification problem, π_k would equal 0.5 as the probability would be split into 2. The $f_k(x)$ corresponds to the density function of X for an observation that belongs to a certain class in the training set. In other words, LDA measures the distribution of the features in each category before applying the Bayes theorem to transform them into estimates to find the probability of the output category given X.

Advantages: Not only is LDA well suited for binary classification problems but it is also adapted to multi-class classification problems.

Disadvantages: The model is less efficient when the size of the sample is small and less stable when the classes are well-separated.

II.c. Quadratic Discriminant Analysis

QDA is an extension to the LDA. The 2 classifiers differ in the QDA's 2nd assumption, the first is that the observations correspond to a Gaussian distribution, and the second is that each category k has a covariance matrix Σ_k with $X \sim N(\mu_k, \Sigma_k)$.
In this application, the QDA performs worse than the LDA, with respective accuracies of 0.764 and 0.812. These results suggest that when fitting, the QDA was too flexible and that the decision boundaries are rather linear in this training set.

Advantages: QDA is a more flexible model with higher variance compared to LDA.
Disadvantages: DQA will not preform well if there are many observations in the training set due to the variance of the classifier.

II.d. Decision Tree

A classification tree starts at the top with a node based on a condition or threshold, then it splits into edges also called the branches extending to another node and so on until it reaches the leaves at the end which correspond to the final decisions. The tree is built through recursive binary splitting which takes place based on a criterion: the classification error rate. However, this is measure is not always enough, other common methods include the Gini index, and the entropy where for both, if the value is close to 0, it can be interpreted as the node containing observations from a unique category.

Advantages: There is no need to dummy encode the variables before applying a decision tree and the process and results are easily interpretable.
Disadvantages: One of the downsides is that the performance, the accuracy may be low and also no change can be made to the training set as decision trees are very sensitive.

II.e. Random Forest

Random forest is a combination of uncorrelated decision trees that each predict a category for a given input. The most occurring category in the random forest then becomes the final prediction for that input, this method is known as the wisdom of crowds.

If each tree inside a random forest selects strong predictors each time, the correlation between them would be high. Random forests avoid this kind of correlation between the decision trees, by not allowing the trees to select the strongest predictors that lead to the most separations as top splits, but rather by considering a more random subset, and a set of new predictors is selected at each split. Another particularity about random forests is that each tree randomly selects observations from the data set. This is because a change in the training set usually significantly affects the decision trees.

Advantages: Random Forest has the power to decorrelate the trees.
Disadvantages: Reduced interpretability compared to decision trees and lower prediction accuracy compared to other tree-based methods (e.g. gradient boosting tree).

III/ Comparison

With the table below, we can determine that the decision tree was the model that performed the best with the highest accuracy of 0.82.

Every model was first applied with all the features, then the process was repeated with the previously selected features with the best subset selection. All the models, except for the random forest performed better with the dimension reduction.

	Model	Evaluation
	<chr>	<dbl>
7	Decision Tree (selected vars)	0.8153674
8	Decision Tree (all vars)	0.8153674
3	LDA (selected vars)	0.8120000
1	Logistic regression (selected vars)	0.8116667
4	LDA (all)	0.8113333
2	Logistic regression (all vars)	0.8100000
10	Random Forest (all vars)	0.8098333
9	Random Forest (selected vars)	0.8065000
5	QDA (selected vars)	0.7640000
6	QDA (all vars)	0.7238333

After comparing the table above with the accuracy of all the models, the same was applied without the winsorization in the preprocessing step to remove all the outliers. The table below shows that some models performed better without the winsorization, such as LDA and Logistic Regression. For the LDA, it can be explained by the fact that it already includes a dimension reduction technique. We can also observe that the decision tree results do not change whether the outliers are removed or not.

	Model	Evaluation
	<chr>	<dbl>
4	Decision Tree (all vars)	0.8153674
2	LDA (all)	0.8128333
1	Logistic regression (all vars)	0.8121667
5	Random Forest (all vars)	0.5000000
3	QDA (all vars)	0.4283333

0.8120000

0.8100000

Conclusion

Predicting whether a customer will default or not the next month is extremely useful for banks in deciding whether to grant credit to a customer or not.
The goal of this application was to define the model that would perform the best to find out which category a customer belongs to, whether they would be in the default (1) or non-default (0) category.

First, the data needed to be cleaned by handling all the missing values by replacing them, removing outliers with winsorization that later proved to be not as useful for some predictive algorithms (namely LDA, logistic regression, decision tree), and selecting the best features to fit the model on.

After applying 5 different models of logistic regression, such as LDA, QDA, decision tree and random forest, and by analyzing and comparing them it was clear that the decision tree with the variables selected from the best subset selection had the highest accuracy out of the 5 models. This means that the decision tree model is best able to predict whether a customer will default next month.

References

Book: An Introduction to Statistical Learning
by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Random forests:
- https://towardsdatascience.com/understanding-random-forest-58381e0602d2#:~:text=The%20random%20forest%20is%20a,that%20of%20any%20individual%20tree

LDA:
- https://www.analyticsvidhya.com/blog/2021/08/a-brief-introduction-to-linear-discriminant-analysis/

Decision trees:
- https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

K-fold cross-validation:
- http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-in-r/
- https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85

Linear discriminant analysis:
- https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/