Prediction of consumer credit risk

Marie-Laure Charpignon
mcharpig@stanford.edu

Enguerrand Horel
ehorel@stanford.edu

Flora Tixier
ftixier@stanford.edu

Abstract

Because of the increasing number of companies or startups created in the field of microcredit and peer to peer lending, we tried through this project to build an efficient tool to peer to peer lending managers, so that they can easily and accurately assess the default risk of their clients. Precisely, the main purpose of this project is to predict if a consumer will experience a serious delinquency (90 days or worse) during the next two years (thus it is a classification problem). The dataset consists of roughly 100,000 consumers characterized by 10 variables. Two of the models we implemented present a very good predictive power (AUC around 0.85): they are obtained by combining trees, bootstrap and gradient boosting techniques.

1 Introduction

Credit and default risks have been in the forefront of financial news since the subprime mortgage crisis that began in 2008. Indeed, people realized that one of the main causes of that crisis was that loans were granted to people whose risk profile was too high. That is why, in order to restore trust in the finance system and to prevent this from happening again, banks and other credit companies have recently tried to develop new models to assess the credit risk of individuals even more accurately. Besides, the financialization of our economies implies that more and more stakeholders are involved, however it can still be very difficult for some people - either because of their banking history or of their atypical situations - to get a loan. This imbalance has led to the development of new alternatives to the bank system. The number of peer to peer lending websites, MicroFinance Institutions (MFI) and companies that back their development, is currently growing quickly, and the quite recent stock market listing of
LendingClub is adding evidence of that. It is precisely in that dynamic that this project fits, its main goal is to predict if a consumer will experience a serious delinquency (90 days or worse) during the next two years. The data, the methods and the models used will be presented in sections two and three, then the results will be interpreted and discussed in section four.

2 Data

2.1 Presentation of the data

The data used in this project comes from the competition "Give me some credit" launched on the website Kaggle. It consists of 120,269 consumers, each characterized by the following 10 variables:

- age of the borrower;
- number of dependents in family;
- monthly income;
- monthly expenditures divided by monthly gross income;
- total balance on credit cards divided by the sum of credit limits;
- number of open loans and lines of credit;
- number of mortgage and real estate loans;
- number of times the borrower has been 30-59 days past due but no worse in the last

two years;
- number of times the borrower has been 60-89 days past due but no worse in the last two years;
- number of times the borrower has been 90 days or more past due.

They are all continuous variables and the dependent variable is if a person experienced 90 days past due delinquency or worse in the last two years (1 if yes and 0 if not).

2.2 Processing

When we looked initially at the data, we thought that they certainly should not all be relevant. For instance the age of the borrower does not seem so important, and the last three variables look redundant. That is why we decided it could be interesting to try to select the most useful variables. To do so, we tested the significance of each of the variables using linear and logistic regressions. They both revealed that the variable "balance on credit cards divided by sum of credit limits" was not really significant. However omitting it did not improve the final results, so we decided to keep it. To go further in that analysis, we also did a PCA of our data. It highlighted that a certain combination of the three variables "number of times the borrower has been some days past due in the last two years" was the first principal component, and that two other combinations of the same variables were the last two components whose associated variances were the lowest. It confirmed the intuition that these three variables could be redundant if they were not considered in the right proportion. We tried to keep only the first eight principal components but again it did not improve the results so we used the original data.

In order to normalized the range of this dataset, we decided to scale all the data. We also realized that this dataset was very unbalanced, the proportion of positive outputs (consumers who had a default) was only 6%. As we wanted to predict if a person would experience a delinquency, we thought it could increase the predictive power of our models to train them on a dataset where the proportion of positive outputs was higher. In this context we increased this proportion to 30% in the training set. This was done by randomly selecting the positive outputs to add in the training set. This improvement has allowed us to obtain much more precise results.

3 Methods

3.1 Models

Classification trees are appropriate for this problem, as they successively determine decision criteria based on subsets of the initial variables. It corresponds to an intuitive representation of the consumers, each one being associated with a cluster linked to its credit profile.

We chose to use four different models:

- Logistic regression as it is a very classic model for this type of problems;
- Classification and Regression Trees (CART): we read in the literature that trees were particularly efficient in classification;
- Random Forests: this model averages multiple deep decision trees trained on different parts of the training set (this aims at reducing the variance);
- Gradient Boosting Trees (GBT): gradient boosting algorithm improves the accuracy of a predictive function through incremental minimisation of the error term. After the initial tree is grown, each tree in the series is fitted with the purpose of reducing the error. A tree at step m partitions the input space into J disjoint regions $R_{1m}, \ldots, R_{jm}$. The output is then

$$h_m(x) = \sum_{j=1}^{J} b_{jm} \mathbb{1}(x \in R_{jm})$$

where b_{jm} is the value predicted in the region R_{jm}. The update rule of the model is

$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$$

where L is a loss function (the MSE for instance). Thus,

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J} \gamma_{jm} \mathbb{1}(x \in R_{jm})$$

3.2 Methods

To assess and compare the precision of our models, we realized that we could not use the classic error measure (number of wrong prediction over the total number of predictions) as the models implemented tend to underestimate the proportion of positive outputs which is already very low in the dataset we worked on. We prefer to use the two following metrics: AUC and F1 score, as they are complementary and both adapted to binary classification. The AUC is the Area Under Curve of the true positive rate versus the false positive rate and F1-score is the harmonic mean between precision (proportions of positive and negative results that are true positive and true negative) and recall (true positive rate). These two metrics are between 0 and 1 and the bigger they are, the better the associated model is.

The results presented in the next section are calculated as an average over thirty iterations of the models. At each iteration the dataset is randomly split into two subsets: a training and a testing set. The proportion of positive outputs is increased in the training set and then the trained models are tested on the unbalanced testing set.

4 Results

4.1 Presentation of the results

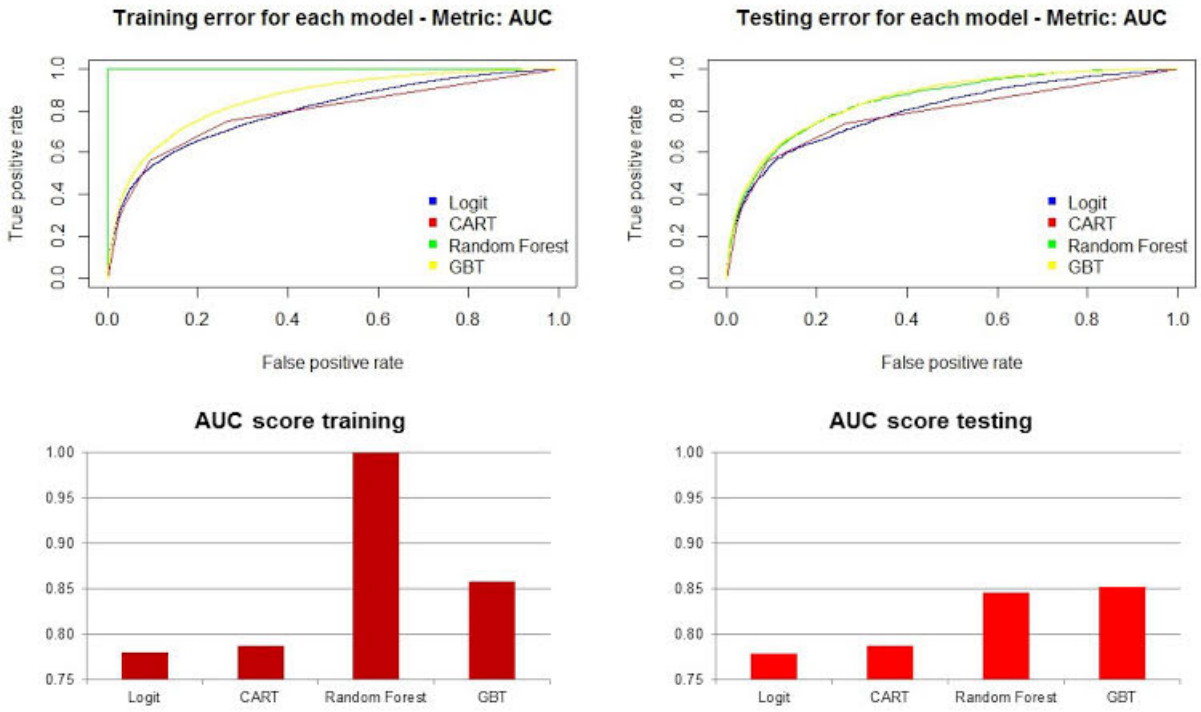

Figure 1: Training and testing error with the AUC metric

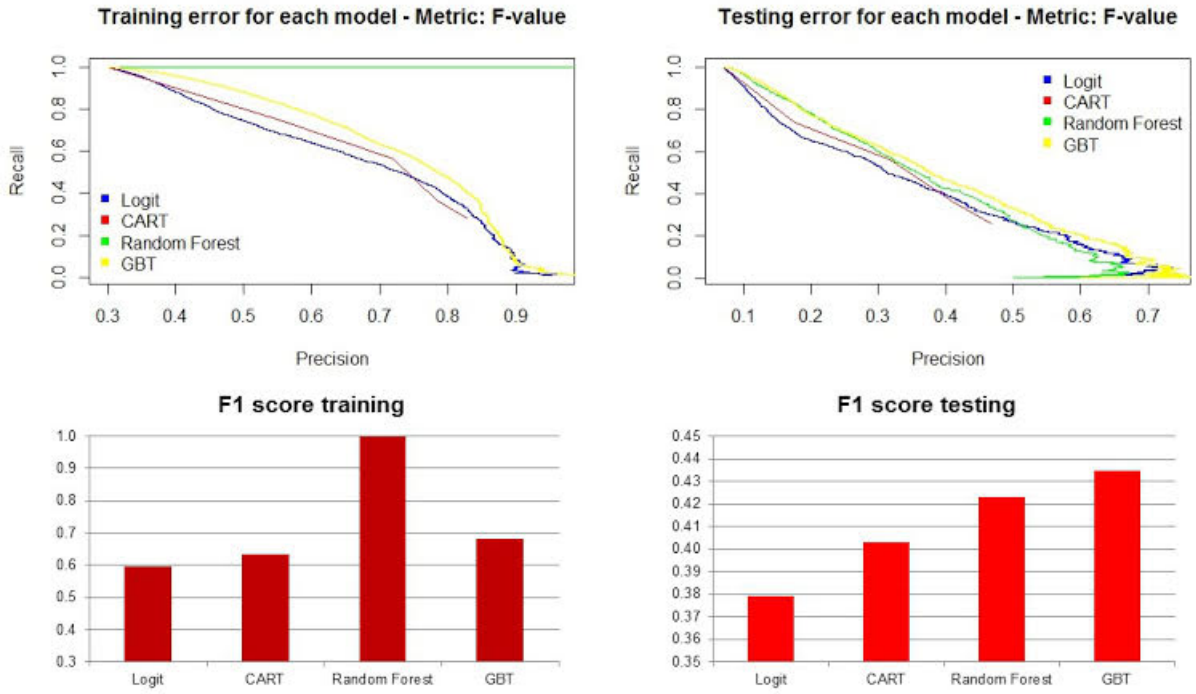

Figure 2: Training and testing error with the F_1 metric

Comparison references

As explained in the previous section, we decided to implement a Logit model in order to have some reference to which we could compare the results from the other three models, both in training and testing. Indeed, Logit is known to be one of the most appropriate algorithms for classification problems.

Comments

Looking at the testing and training results for the AUC metric, we can clearly state that two distinct groups of models appear: Logit and CART constitute the first one; the more sophisticated tree models - Random Forest and GBT - form the second one. We also notice that the performance is quite similar for testing and training, using this metric. Unlike AUC, F1-score introduces a bigger gap between training and testing values. Moreover, it has a more gradual evolution. Yet, it also indicates that GBT is the best model.

4.2 Discussion - interpretation

Our two best models are successful - with an AUC around 0.85 - in predicting if a consumer will experience a serious delinquency in the next two years. Our results are very satisfying compared with those of the best competitors of the Kaggle competition from which we collected our data. When it comes to testing, our models are efficient, for two major reasons: the first one is that the structure of trees is adapted to classification problems; and the second one is that they are sophisticated, compared with the basic CART, as they involve statistical and machine learning techniques such as bootstrap or Gradient Boosting. The only aspect that surprised us a lot was the fact that Random Forest highly overfits: it is astonishing because it is not what is expected from this model. By construction, it is indeed supposed to have a lower variance than CART. There is only one determining parameter for this model (the number of trees) and the same result has been obtained for dif-

ferent values of it: the problem may come from our database, and one possible improvement could be to test with others.

5 Conclusion

By combining trees and gradient boosting technique (GBT model), we have implemented a model which presents two principal features. First, its predictive power is very accurate. With an AUC of 0.86, GBT beats the other models we used and especially Logit (which was our reference). Second, its small variance makes it reliable: unlike Random Forest, its training and testing errors are on the same scale which means that it does not tend to overfit.

6 Future

To go on with this project, we thought about some ideas for improvement. We used a database of ten variables for this study. It could be interesting to try to add new variables (associated with some characteristics of the loan for instance) and see if it improves the predictive performances of the models. For example, LendingClub is using more than 100 variables to predict the default risk. Besides, according to the literature, neural networks offer very good performance for credit scoring problems. Thus, comparing its predictive power with the one of our models could allow us to put our results into more perspective.

In order to create a practical and useful application from this study, we could develop a credit risk management tool for peer to peer lending companies. This tool could provide for instance the ideal interest rate for a loan in order to minimize its risk. A peer to peer lending company connects borrowers and lenders, the latter being investors looking for certain returns and risk ratios based on their risk profile. Predictions of credit risk of individuals could also be used to create portfolios of loans in order to diversify their risk and to help investors reaching their specific return over risk target.

References

[1] Brown I., Mues C., *An experimental comparison of classification algorithms for imbalanced credit scoring data sets*, (Expert Systems with Applicaions #39 , 3446-3453, 2012.)

[2] Chawla N. V., *Data Mining for Imbalanced Datasets : an Overview* (Springer, 853-867, 2005.)

[3] Galindo J., Tamayo P., *Credit Risk Assment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications* (Computational Economics #15, 107-143, 2000.)

[4] Hand D. J., Henley W. E., *Statistical Classification Methods in Consumer Credit Scoring : a Review* (Journal of the R. Statist. Soc. #160 , 523-541 , 1997.)

[5] Khandani A. E., Kim A. J., Lo A. W., *Consumer credit-risk models via machine-learning algorithms.* (Journal of Banking & Finance #34 , 2767-2787, 2010.)

[6] Thomas L. C., *A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers*, (International Journal of Forecasting #16 , 149-172, 2000.)