Loan funding: a study covering the probability of credit defaults

The Lending Club platform is an American company that finances loans from individuals to individuals. Founded in 2006, by May 2020 it had financed up to $59 billion in loans. As an alternative to bank financing but also to financial investment, this platform allows both the project realizations and a low-risk investment for the people granting the loans. However, to ensure this low risk, the allocation of loans must be optimized to avoid credit default. We will focus on this issue. Using two classification models, we will propose a segmentation of loan applicants in order to help investors in the selection of their financing. We will first present modifications to the data so that it is usable, robust and relevant. We will then use a logistic regression model, the main interest of which is to inform on the probabilistic impact of each variable on loan repayment. We will also proceed to a classification of individuals using a random forest classifier. Finally, we will present the Receiver Operating Characteristic (ROC) and calculate the Area Under the Curve (AUC) which will be our main metric.

1. DATA MANAGEMENT

The dataframe “Lending Club” is initially composed of 151 variables and 2260701 observations. There are 113 float variables and 38 objects. We will, in this first part, begin by presenting the data management we made in order to have a clear and clean database for the models we will do later in this analysis.

As the subject specifies it, there is information that we are not supposed to know regarding the time where the loan credit was granted. Thereby, we look at each variable one by one in order to know if the information provided is whether or not relevant. We then verify that the database does not have any duplicates.

In order to deal with our database, we proceed by eliminating variables that contains too many modalities such as the id or the applicant’s profession title. Indeed, since a huge amount of

specific information were given, we decided to eliminate those variables since they don't provide any interesting evidence.

In the meantime, since our database contains missing values, we fixed a deletion threshold. As a result, variables which had more than 80% missing values were deleted. For the other part of the sample, there were interesting variables and not all of them are necessary missing values since some client may don't have the answer on a specific request. Therefore, variables that had a lot of missing values but at the same time, were under our deletion threshold, we computed a specific value of "–10" since it was important to not create biases. In this case, the missing values was an information. On the contrary, for variables which had a few missing values, NaN were a lack of information. We observed that some features' missing values appeared for individuals having low salaries. Thus, we computed the minima value for them. For other missing values, having no information on their possible meaning, and wishing to avoid any unbalanced observations, we used the mean value of the feature.

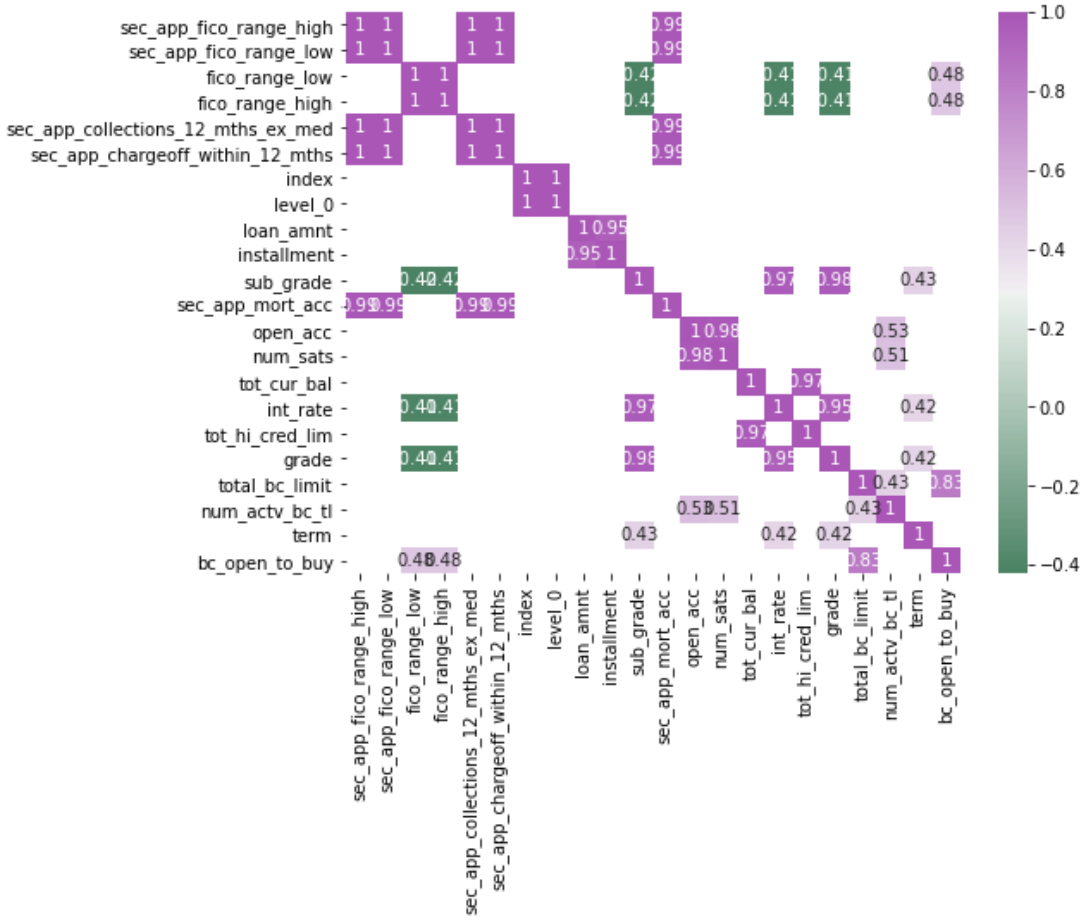
Now that our database is clean, another step is necessary before creating our models. Thus, we proceed to encode qualitative variables. For our target, the loan status, we implemented a mapping thanks to the replace function. On the other hand, for the others categorical variables, we used the target encoding.

In our database we have several categorical variables. However, in the rest of our study we will need a database composed only of numerical variables. We will therefore encode the categorical variables so that they become numerical. Thereby, we use the target encoding smoothing method. We will perform a Laplace smoothing which allows to smooth the categorical data. We chose this method because we have a lot of variables, we have a target and the smoothing will avoid the problem of overfitting that occurs when the modalities of the categorical variables are not well represented.

We must therefore choose the "weight", m, we wish to attribute to the mean of our target. It must be strictly greater than zero. We will consider that the value of m will be too high if the values that will be replaced in the categorical variables are close to the mean of our target. Taking these conditions into account, we have chosen a value of 10 for m.

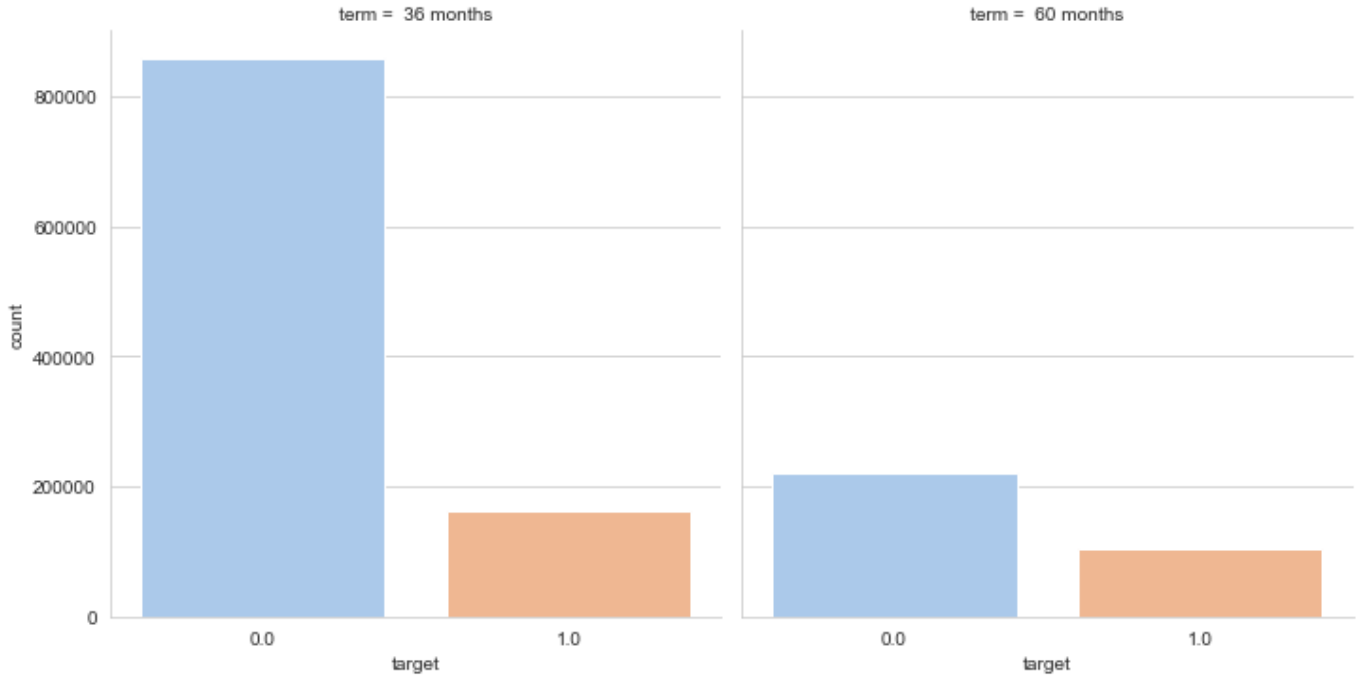
With the encoded data set, we will be able to do the VIF which is the next tools that we will use.

About our variable's selection, we decided to estimate multicollinearity between variables and delete the ones that have a strong correlation. Indeed, a strong multicollinearity in generalized model could distort interpretation of coefficients due to the inflation of these coefficients' variance. Thereby, we import variance_inflation_factor from statmodels and check correlations clusters thanks to a graph. It appears that some of variables relative to fico_range and open_account measure the same phenomenon. Thus, we deleted these and create a 'fico_range_mean' instead of each fico_range variables. Finally, the VIF enabled to delete 12 variables that were highly correlated.

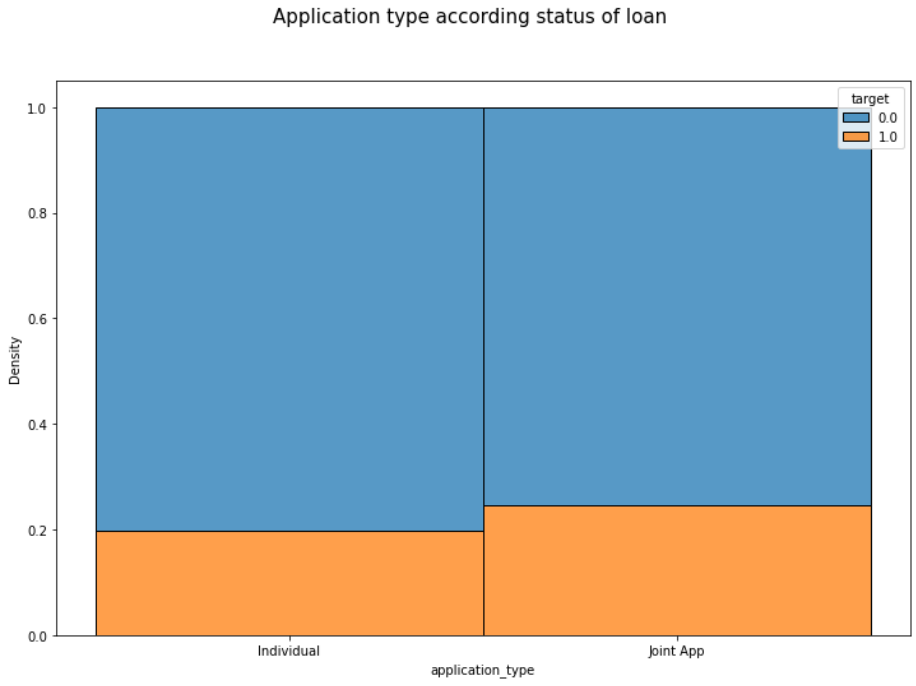


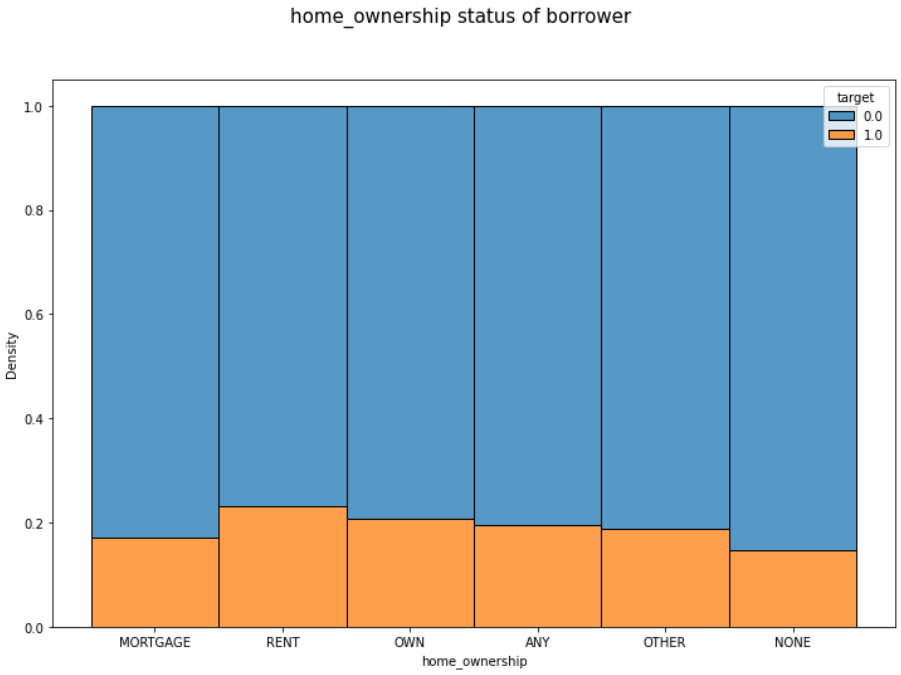
In order to describe in the most efficient way our dataframe, we make graph on variables that seem important to us. After some researches about the database Lending club, we represent interest rate, home status, application type, amount and term of the loan, according to the status of the loan.

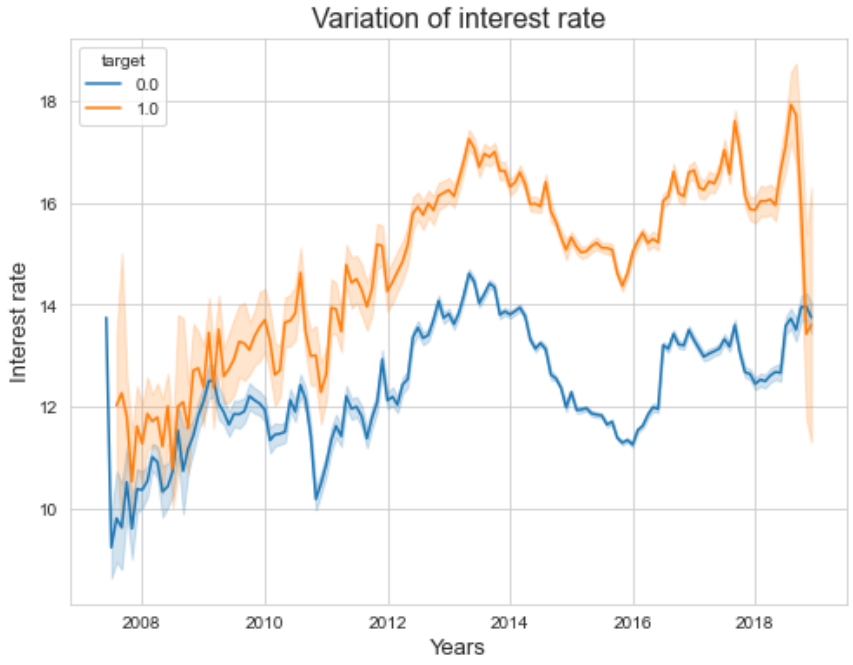
At this time, we have already selected only two types of status loan that are “fully paid” and “charge off": the target variable. Thus, we obtain a database composed of 20% people categorized in “charge off” and 80% of people in “fully paid”.



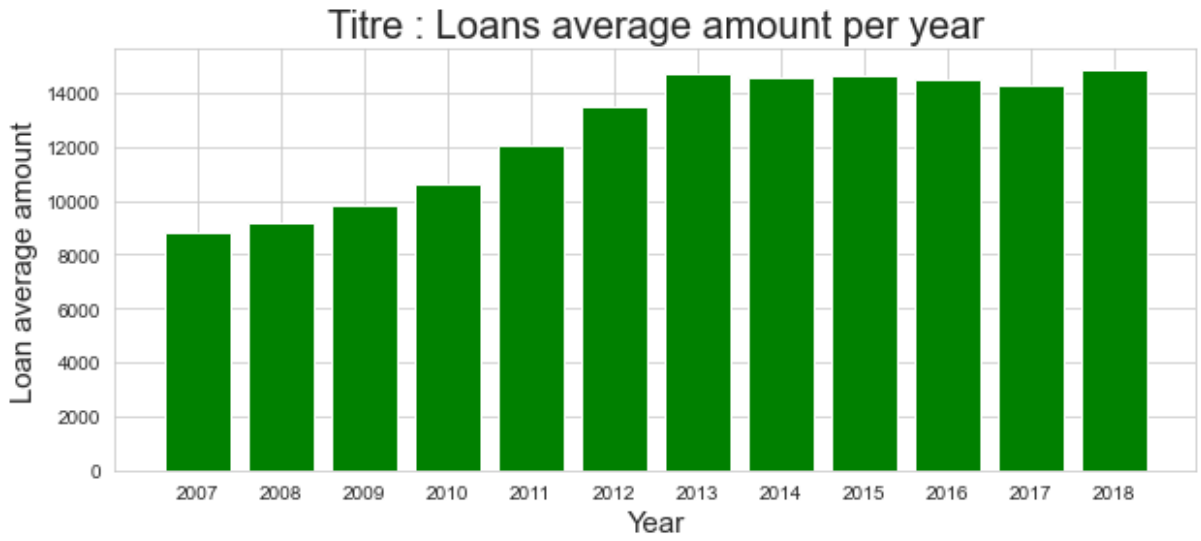
First analyzes reveal that the proportion of charge off is more important in 60-month term contract than 36-month term compared to fully paid. Besides, in term of density they tend to take the loan with a second applicant and most of them are rented. Moreover, the interest rate is significantly higher for person in charge off since 2010.







Actually, the mean interest rate 12% for person in fully paid case and 15% for person in charge off. Nevertheless, the mean amount of loan is stable for all borrowers since 2013 after a large increase. We can notice these persons have an annual income lower than the other one and a loan amount higher than these.



2. VARIABLE SELECTION

As said before, we chose to do a Random Forest Classifier and a Logistic Regression. In order to do this and have a robust model, we split our database in 3 parts.

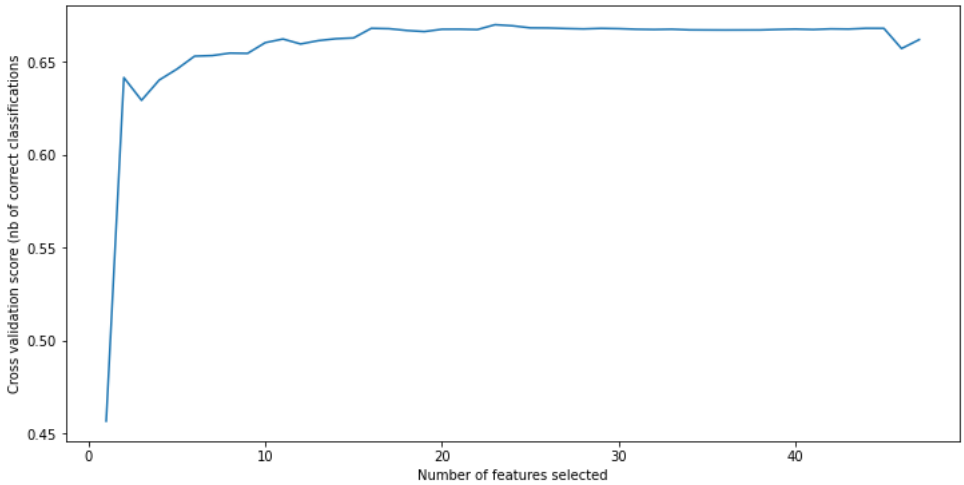
In this part, we start by separating the dataset into two parts according to the date 12/31/2016: the first part is for training and validation and the second part is for the test. Then, we separated the first part for training (80%) and the second part for validation (20%).

After the data management phase, we have 55 variables. Wanting to have a robust and efficient model, it is better to select some other features to delete. In order to choose the features to keep we use a Recursive Feature Elimination with Cross-Validation (RFECV) and a Recursive Feature Elimination (RFE). This method allows us to select the best variables to describe our target, in order to categorize individuals depending on their probability of default.

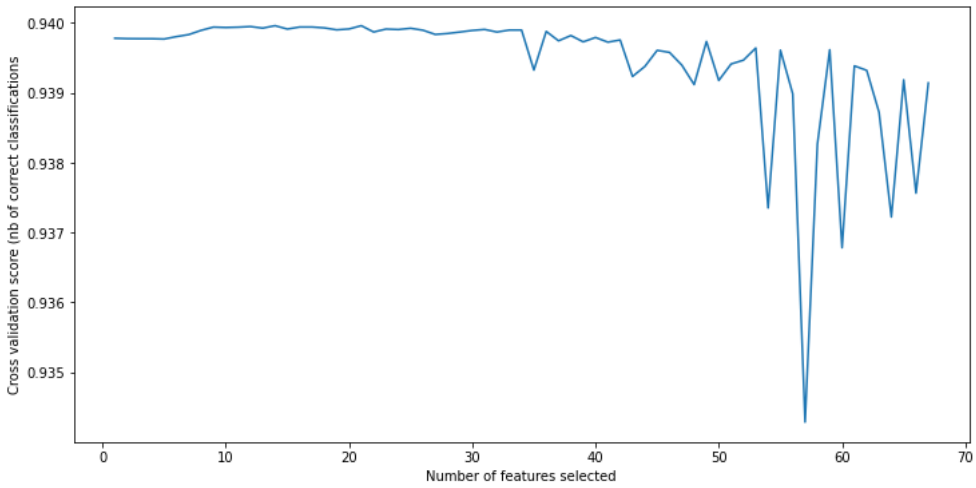
In its operation, the RFECV select the best variables depending on an estimator and scoring type (we use the accuracy). We do this for both Random Forest Classifier and the Logistic Regression and obtain respectably a number of 14 and 28 features to select.

The advantage of the RFECV is that it allows to have an accuracy curve depending on the number of features selected. We see on the curve below that for a logistic regression, we could have a higher accuracy using 42 variables. But we can have a quite similar accuracy using

only 28 variables. As a model is faster with less variables, we select the first 28 features. The RFE will then tell us what are the main 28 variables that explain the target, using a logistic regression.



Here, we selected 14 features because we see that it's the number of features giving the best accuracy score.



Having the number of features we want to keep in both models, we can do the RFE that will allow us to know these variables. This will select the best features for the model based on the estimator we selected. We will also fit the RFE on our train dataset for both models.

We finally have 28 variables for the Logistic Regression : 'term', 'int_rate', 'grade', 'emp_length', 'home_ownership','verification_status','purpose','addr_state','dti','delinq_2yrs','open_acc','pu

b_rec','total_acc','initial_list_status','application_type','open_acc_6m','open_act_il','open_rv_12m','all_util','acc_open_past_24mths','chargeoff_within_12_mths','mort_acc','num_actv_rev_tl','num_tl_120dpd_2m','num_tl_op_past_12m','pub_rec_bankruptcies','hardship_plan_binary', and 'fico_range_mean'.

And 14 variables for the Random Forest: 'term', 'int_rate', 'grade', 'emp_length', 'annual_inc', 'dti','collection_recovery_fee','tot_cur_bal','open_rv_24m','max_bal_bc','acc_open_past_24mths','avg_cur_bal','num_tl_op_past_12m', and 'hardship_plan_binary'

3. Modeling: Logit Regression and Random Forest

A. Methodology

As said before, we have three datasets: a train, a validation and a test dataset. First, the train part allows to train the model in order to obtain a good parameterization and a good fit.

Then, during the validation part, we apply our model to the valid dataset. The objective of this second step is to know if our model can be generalized. The validation set makes it possible to obtain a better evaluation by adjusting overfitting. Accuracy is the metric used to analyze overfitting and underfitting. We analyze this metric through its probability and a representation of the validation curves. Accuracy is the sum of true positives and true negatives over the sum of all estimates. It is therefore the probability that the model gives the correct estimates. The validation curves will allow us to know which number of iterations to choose to have a generalizable model (without overfitting and underfitting). There is underfitting when the accuracies of the train and valid dataset is low. In this case, we increase the number of iterations in our train part for a better training. There is overfitting when the accuracy of the train is greater than the accuracy of the validation.
When analyzing the validation curves, it is still necessary to keep a small difference between the accuracy of the train part and the accuracy of the valid part to avoid overfitting. With overfitting, the model is not generalizable. In this case, we lower the number of iterations in our train part. Thus, the model is less trained. Finally, the model is efficient when the accuracies are close.

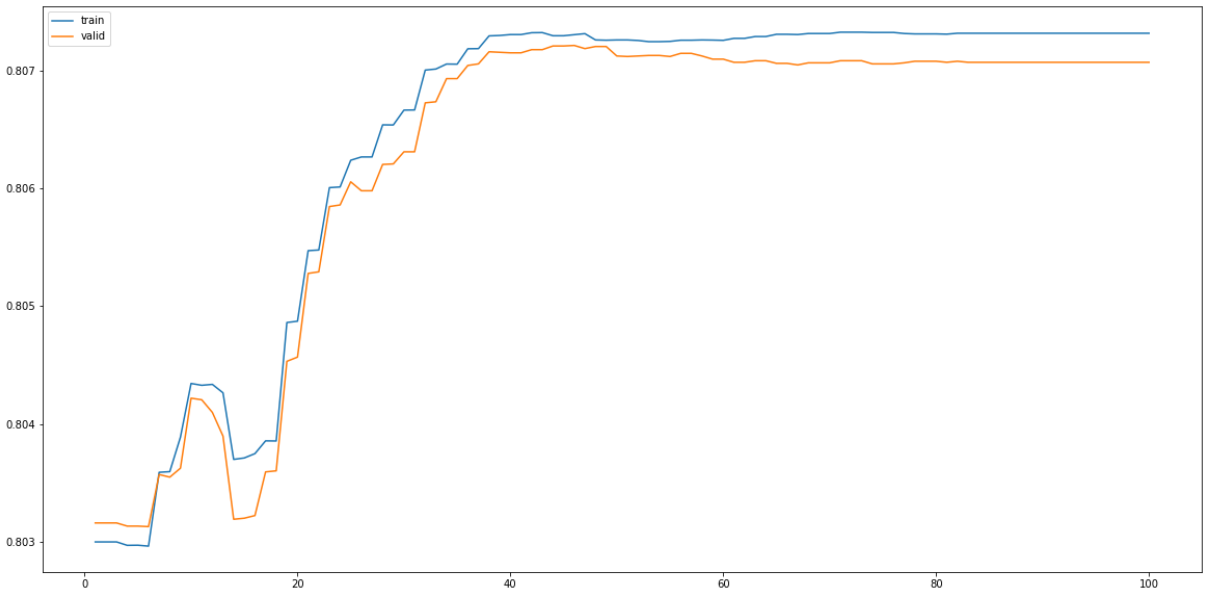
The third step called testing is used to provide a final model fit and estimate the performances of our model. The dependent variable is predicted on the test dataset. The performance evaluation is done by the ROC and the AUC metric. The ROC is a graphical representation of the relationship between the true positive rate (sensitivity) and the false positive rate (specificity) for different thresholds. This curve makes it possible to know for which threshold we obtain the lowest false positive rate and the highest true positive rate. Thus, choosing this threshold enables to reduce the cost and therefore to be more profitable. The AUC metric is a score that calculates the quality of precision of our model regardless of all thresholds. The closer the AUC is to 1, the more accurate the model is.

B. Logistic regression

Logistic regression is a statistical model that use a logistic function to model a binary variable. We give the value 1 to the persons who are in default of credit and 0 otherwise.

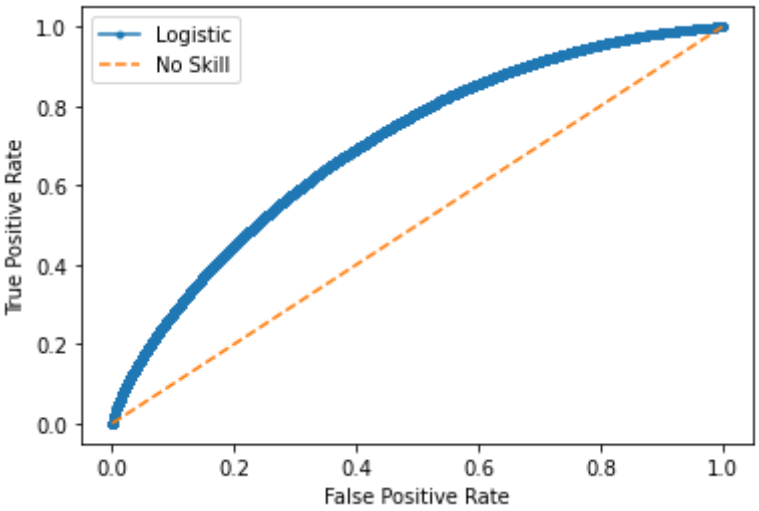
First step: We apply the logistic model to the train part of our model to train it. We use the newton-cg solver which is an algorithm that uses gradient descent to minimize the cost function. Regulation by « l2 » also makes it possible to minimize the cost function. The number of iterations (max_iter) is the parameter that allows us to regulate underfitting and overfitting.

Second step: We apply the model to the valid dataset and calculate the accuracies and the validation curves. According to the analysis of the validation curves, we must set the maximum iteration at 65.



When we set max_iter to 65, the accuracies of the train and validation parts are respectively 0.8073 and 0.8070. As these accuracies are important, there is no underfitting. Also, they are close but not equal which means there is no overfitting. Thus, the model is generalizable.

Third step: We predict the risk of credit default on the test part. AUC is 0.7. It is relatively close to 1 which means that our model is quite efficient.



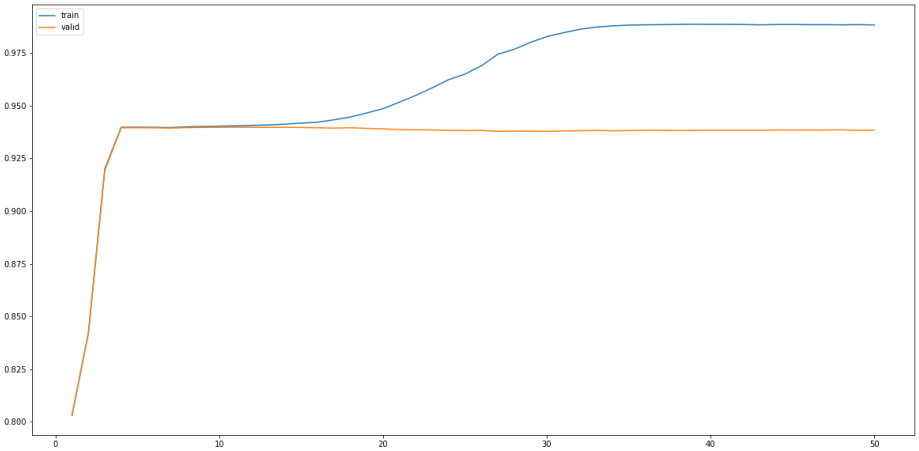
The ROC curve indicates that people with a probability of default greater than 0.7 are classified as insolvent borrowers. This threshold makes it possible to reduce costs and to be the most profitable.

C. Random Forest

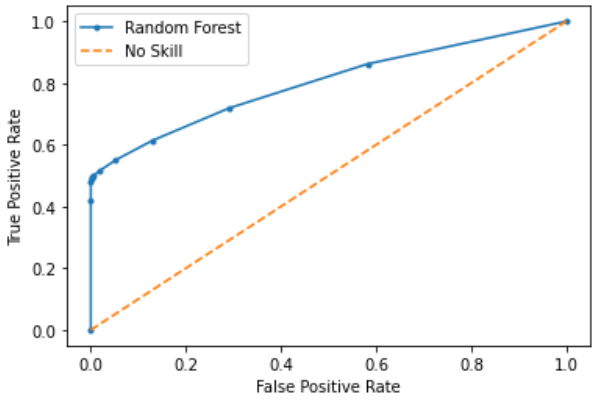
The Random Forest is a Machin Learning algorithm that classifies individuals according to their characteristics. It works by building a multitude of decision trees that take the class out. We give the value 1 to the persons who are in default of credit and 0 otherwise.

First step: We apply the random forest model to the train part of the model to train it. We set the number of trees in the forest at 10 to have a good performance without taking too much time. The depth of the tree (max_depth) is the parameter that allows us to regulate underfitting and overfitting.

Second step: We apply our model to the valid dataset and calculate the accuracies and the validation curves. According to the analysis of the validation curves, we must set the maximum depth at 35.



When we set max_depth to 35, the accuracies of the train and validation parts are respectively 0.9883 and 0.9381. As these accuracies are important, there is no underfitting. Also, they are close but not equal which means there is no overfitting. Thus, the model is generalizable.
Third step: We predict the risk of credit default on the test part. AUC score is 0.8. It is close to 1 which means that the model is efficient.



The ROC indicates that people with a probability of default greater than 0.7 are classified as insolvent borrowers. This threshold makes it possible to reduce costs and to be the most profitable.

4. Conclusion

Finally, we analyzed the lending club loan applicants according to their probability of default. Starting by cleaning our database, selecting the variables of interest and available at the time of the loan, we also proceeded to change the variables and manage the missing data. In order to select the most relevant variables for our analysis, we first performed a correlation analysis, then an RFECV and RFE which allowed us to select the most relevant variables for our two models. Finally, we trained our models on the train dataset, validated by avoiding overfitting in order to ensure the generalization of the models thanks to the validation dataset, and finally applied our models to the test dataset. The two models have an ROC of 0.7 which means that people with a score above 0.7 will not get credit. This reduces the risk for investors and thus increases profits. Finally, both models have accuracies greater than 0.8 which shows a good fit of the model to the data. It is important to note that the Random Forest is more efficient because it has half the number of variables of the logistic regression, is faster, and has higher accuracies. We therefore recommend the use of this classification model.