APPLIED MACHINE LEARNING PROJECT

CREDIT RISK AND EXPLAINABLE AI

X

by:

[redacted]

Julien DOS-REIS

Professor: [redacted]

Subject: Applied Machine Learning

Semester 2

31 May 2023

1. Goal

The aim of the project is to demonstrate the interest or not of an XAI approach in the context of credit scoring. To realize it, we build a machine learning algorithm to predict the risk of default proxied in the dataset by the loan status. The XAI algorithm is then applied on the machine learning algorithm to identify the important features in the data. The ultimate goal is to be able to make sense of the different information that we will obtain and assess the importance of the XAI algorithm.

2. Choice of the dataset

To build the machine learning algorithm, three dataset from Kaggle were provided. We based the choice of the dataset on the number of datapoints, number of missing values, the existence of the loan status variable and his balanceness level, and the multicollinearity analysis. Overall, the adequacy of the data with the objective of predicting the risk of default is what we are looking for.

3. Number of missing value and datapoint

Machine learning algorithms generally require a large set of data with low missing value to be efficient. The first dataset contains information on the loan holder of a german's bank with 1000 observations and more than 30% missing value. The second dataset (credit-risk) contains information on loan holders. It has more than 32000 observations with less than 10% missing value for certain variables. The last dataset records information on credit card holders with 30% missing value for one variable and more than 77000 observations

4. Loan status variable

The german's bank dataset doesn't have a column that gives information on the status of the loan. On the other hand, the credit card dataset does have information on the credit status (whether or not the holder has repaid the credit) however the information provided is only on past due credit payment duration. The credit-risk dataset contains the loan status even though it is unbalanced (the positive class 'default' accounts for 23%).

5. Multicollinearity analysis

In all the different dataset, the level of correlation coefficient between the numeric variables is really low. Regarding the categorical variables, although the Khi2 test shows the existence of correlation, the Cramer's V statistic shows that the intensity is really low (<0.1) for most of those variables. Last but not least, to assess the intensity and the direction of the relationship between the numeric and categorical variables, we use Kendall's tau coefficient and the box plot methodology. It shows the existence of correlation between the variables even though the intensity is really low. Despite the fact that the result of the Kendall's tau should be mitigated because all the categorical variables are not "ordinal", combine with the box plot it bring useful information

In conclusion, based on those criteria we decided to work with the credit-risk dataset.

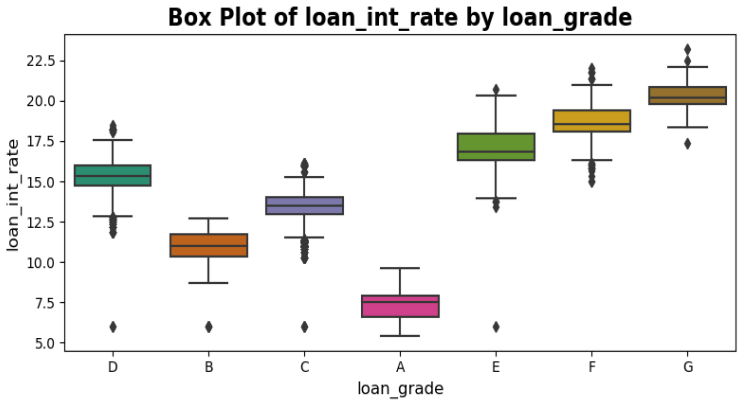
6. Treatment of the data

- Preprocessing

We have 12 variables in our dataframe, including 8 numeric and 4 non-numeric variables for a total of 32000 loans. The dataset present information on age, annual income, home ownership, employment length (in years), loan intent, loan grade, loan amount, interest rate, loan status (0 is non default 1 is default), percent income (the ratio of credit amount to annual salary), historical default, credit history length.

- Missing values

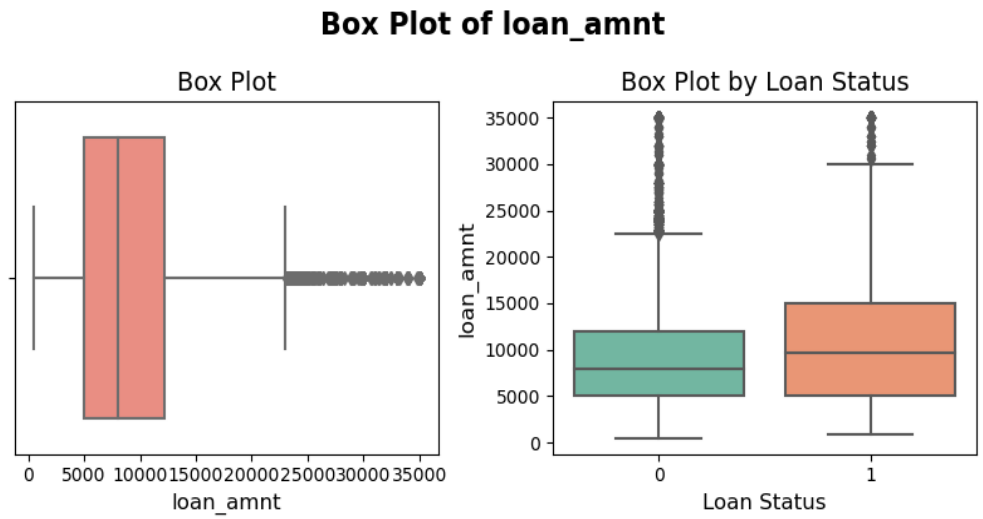
The dataset presents around 10% missing value, mainly on the interest rate and person employment length.



The boxplot shows that the loan interest rates were different across the loan grade label. Thus we replace the missing value of the interest rate by the average of interest rates per loan grade.

- Outliers

The outliers were handled differently for each variable. For the variable age we decided to remove every observation with an age higher than 80 years old (it has a direct incidence on the missing value of the employment length variables). Also, some individuals had an employment length longer than the time they spent on earth and we decided to remove it. For the remaining variables (incomes, loan amount, ratio of loan amount to the annual income) we decide to winsorize them to reduce the effect of the outlier on the model.

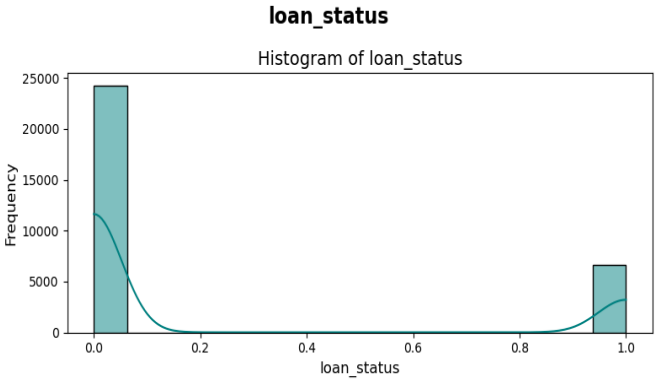


- **Data analysis**

Exploring the dataset provides information on the structure of the dataset and gives insight on the methodology to analyze it moving forward. The table below presents the summary statistics after treating the missing values and outliers.

Numerical variables

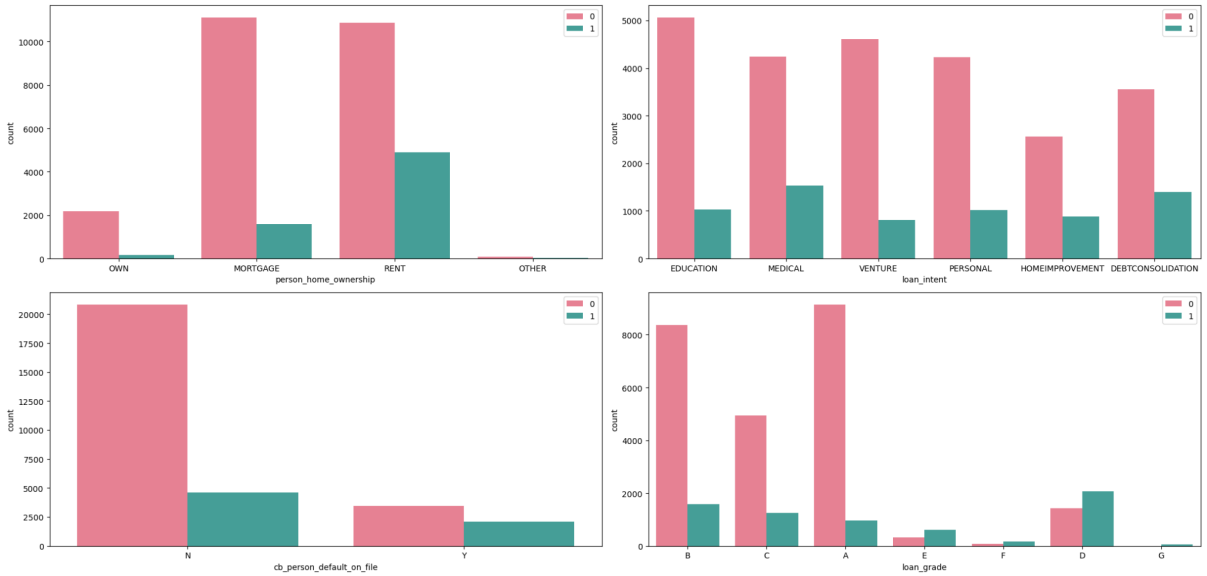
	person_age	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_cred_hist_length
count	30931	30931	30931	30931	27966	30931	30931	30931
mean	27,79	66087,17	4,66	9643,40	11,04	0,22	0,17	5,84
std	6,22	47388,08	3,96	6339,20	3,23	0,41	0,11	4,07
min	20	4000	0	500	5,42	0	0	2
25%	23	39000	2	5000	7,9	0	0,09	3
50%	26	55680	4	8000	10,99	0	0,15	4
75%	30	80000	7	12375	13,48	0	0,23	8
max	80	948000	41	35000	23,22	1	0,83	30



The plot presents the repartition of the observations according to the loan status. It shows that the dataset is highly unbalanced in favor of the 'no-default(0)' case. This will have a direct impact on the model ability to predict accurately

The plot below shows the classification of the observation according to the loan status for different characteristics. We can see that the unbalancedness is spread across the whole variables, with the predominance of the 'non-default' case for all the labels of the categoricals variable except for the loan grade E.

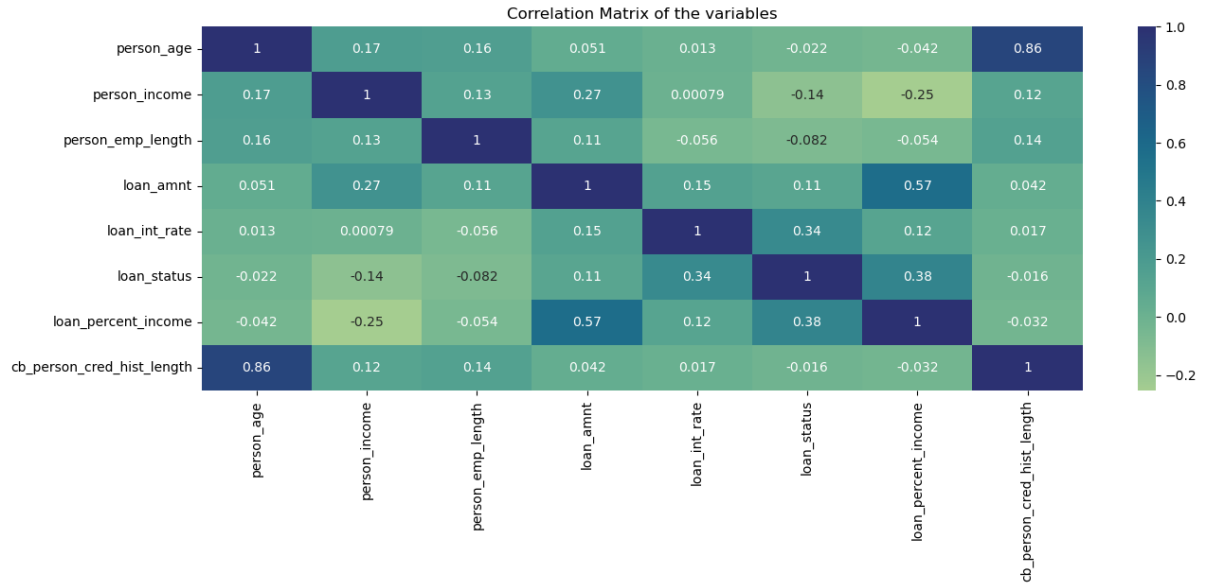
Categorical variables per loan status



- **Bivariate analysis and features selection**

The bivariate data analysis is done with the objective of efficiently choosing the different variables to put together in the machine learning model, so as to reduce the overfitting risk. To investigate the relationship between the numeric variables, we use the correlation matrix to analyze the correlation level and direction, and the Variance Inflation Factor to assess the existence of multicollinearity in a regression. It quantifies the extent to which the variance of the estimated regression coefficient is inflated due to the correlation between predictor variables.

Correlation matrix



As shown in the plot, the overall correlation is low except for the individual age and the credit history length which are positively correlated. The older an individual becomes, the

higher his credit history is. Also the loan amount and the loan percent income (the ratio of credit amount to annual salary) are positively correlated (which is expected).

	Variable	VIF
0	const	65,7986
1	person_age	4,43114
2	person_income	2,145506
3	person_emp_length	1,071535
4	loan_amnt	2,917501
5	loan_int_rate	1,029122
6	loan_percent_income	2,803383
7	cb_person_cred_hist_length	4,381836

From the table below we can confirm the existence of multicollinearity among the predictors. However, it was to be expected as the correlation matrix showed a high correlation for the individual age and the credit history length, and also the loan amount and the loan percent income.
For the qualitative variable we use the chi-square technique to investigate the correlation between the differences. It is a statistical test used to determine if there is a significant association between two categorical variables. It is based on comparing the observed frequencies of categories in a contingency table with the expected frequencies that would occur if the variables were independent. The null hypothesis of the test is that "there is no association between the variables". The formula is the following:

$$\chi^2 = \sum [(\text{Observed} - \text{Expected})^2 / \text{Expected}]$$

We then compute the Cramer's V to have a look at the intensity of the relationship shown by the chi-square test. Cramer's V is a measure of association between two categorical variables. It is derived from the Chi-Square test statistic and helps assess the strength of the association. Cramer's V ranges from 0 to 1, where 0 indicates no association, and 1 indicates a perfect association. It takes into account the number of categories and the size of the sample when interpreting the strength of association.

Although the Chi-square test showed the existence of a significant association between all of the variables, the Cramer's V proved that the intensity is really low (<0.1).

The last methodology that we use is Recursive Feature Elimination (RFE). It is a feature selection technique based on recursion, which aims to select the most relevant features for a machine learning model. The main idea of RFE is to repeatedly build a model and eliminate the least important features, until a desired number of features is reached. It presents the following advantages:

- RFE takes into account relationships between features, unlike filtering methods,
- RFE is compatible with various machine learning models and is easy to use,
- RFE evaluates the model's performance at each stage, enabling the optimum number of features to be chosen.
- RFE can be slow, especially for datasets with a large number of features, as it has to build a model for each subset of features.
- RFE performance is highly dependent on the machine learning model used to assess feature importance.
- RFE may not be effective if the machine learning model used does not accurately reflect feature importance (such as neural network).

By combining all these methods, we went from 12 variables to 6 variables which are: annual income, home ownership, employment length (in years), loan intent, loan grade, loan percent income.

- **Model selection**

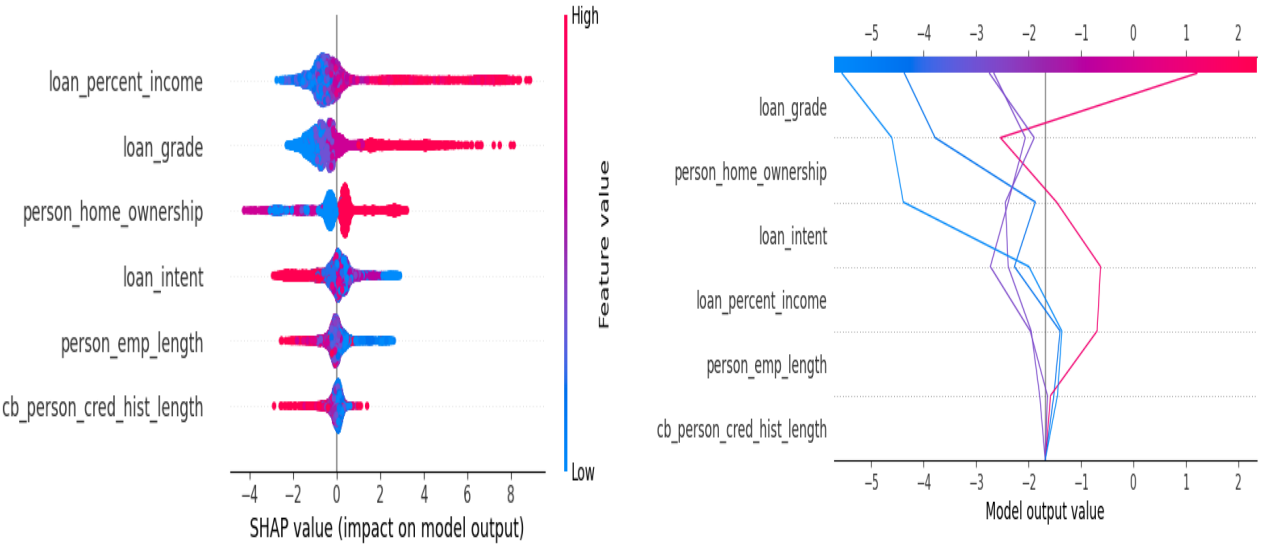
We tried five different models and XGBoost appeared to be the one with the overall best performance compared to the other. The performance metrics for the different models are presented on the table below.

	f1 score	Accuracy	Rmse
Logistic Regression	0,539	0,843	0,3952
Decision Tree	0,672	0,857	0,377
Random Forest	0,744	0,906	0,744
XGradientBoosting	**0,758**	**0,914**	**0,29**
Neural Network	0,737	0,906	0,306

7. Explainable AI

The XAI algorithm that we use is SHapley Additive exPlanations (SHAP). The algorithm iterates through all possible combinations of features, calculating the difference in predictions when including or excluding each feature. These differences represent the contribution of each feature to the prediction. The resulting SHAP values represent the attributions or contributions of each feature to the final prediction.

After applying the SHAP algorithm on the XGBoost model we obtain the result present in the graphic below.

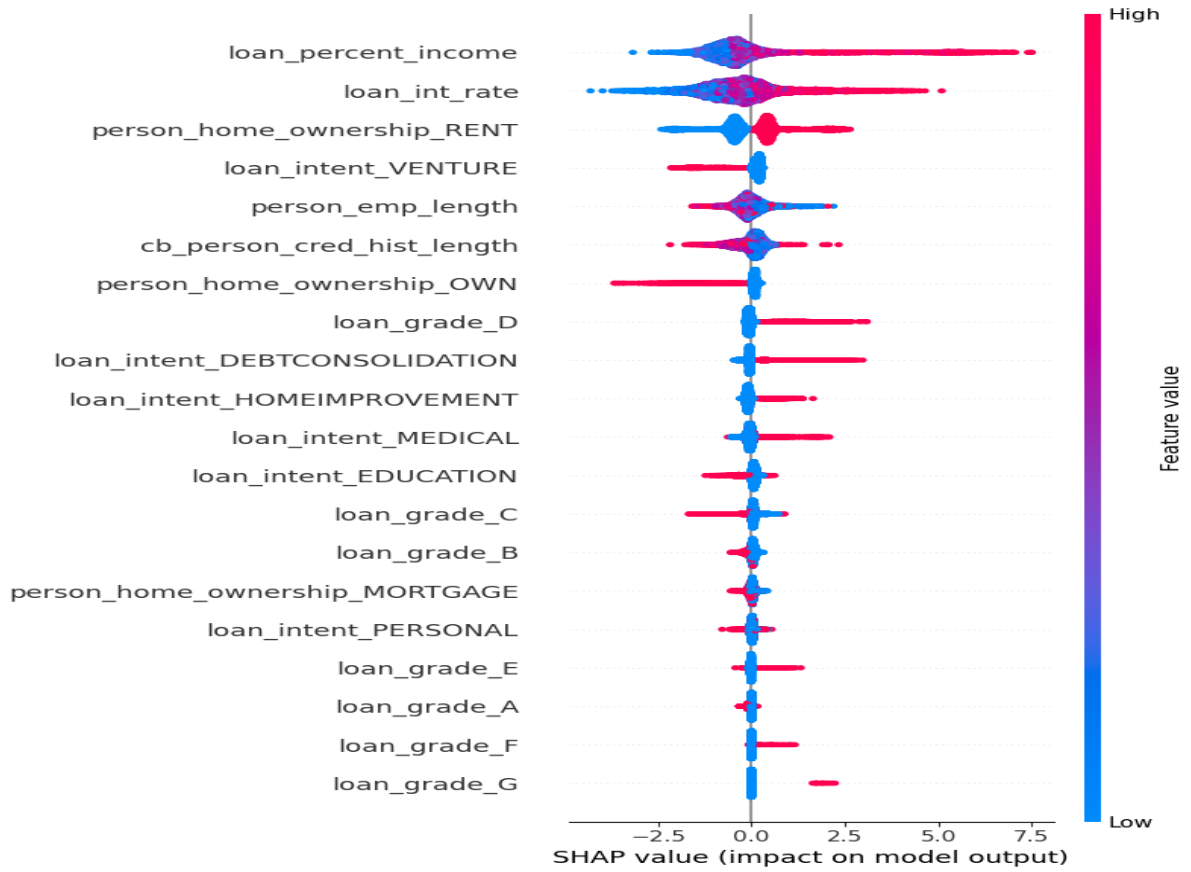


The features are ordered from the most important to the least important.

Loan percent income: is the most important feature. It shows that the higher the ratio of credit amount to annual salary is the more likely it is to default.

Loan grade: The loan grade is ranked from A to G with A being the best rate and G being the worst. By using the LabelEncoder method we were able to keep the rank by taking into account the distance component. Thus, the XAI results which shows that the higher the loan grade is, the more likely it is to default makes perfect sense.

To have a better view at the remaining categorical features we used a different type of encoder method (OneHotEncoder) that doesn't impose a hierarchy on the label. The graphic is shown below.



Thus regarding the home ownership status, it shows being an owner impacts the prediction toward the no-default class, while renting is more even on both classes of loan status. The loan intent variable pushes the prediction toward the default class with the debt consolidation and home improvement being the main cause.

- Robustness check

To further improve the XAI analysis we realized a robustness check based on the stability and comprehensibility metrics for these SHAP values. Here are our results.

Robustness Metrics	**Results**
Stability	0.08
Comprehensibility	1.00

Stability refers to the consistency or reliability of the results across different scenarios, variations, or data samples. Comprehensibility accounts for the interpretability or understandability of the results or models used in the analysis. While we get a high score of comprehensibility, we can see that our score for stability is very low. This can be due to several causes such as data model sensitivity or model complexity for example. The problem is that to improve these results, we need to change parameters, treatment of the data.. on our

champion model : XGB. So we have to get back to our model in order to improve this score. That’s where the use of SHAP isn’t an end in itself, machine learning models still need to be precisely known in order to be fine tuned.

To conclude, Explainable AI is a great tool, it allows us to provide insights and precision on models. Using SHAP can give some insights on how the model categorized specific observations, how variables weigh in the model. It’s a good point for explaining to people not aware of this kind of subject. It doesn't take work away from data scientists. Models still need to be precisely understood in order to be fine tuned. Explainable AI won’t be helpful for this purpose.