AI for finance: Project 4

Giovanni Scialla

May 2023

Scenario

Predicting loan defaults based on credit score data

Feature Variables

Age Age of the person

Home ownership categorical (MORTGAGE, RENT, OWN, OTHER)

Employment length length of the employment period of the person (in years)

Loan intent categorical (DEBT CONSOLIDATION, EDUCATION, HOME IMPROVEMENT, MEDICAL, PERSONAL, VENTURE)

Loan grade categorical (A, B, C, D, E, F, G)

Loan amount

Interest rate

Percent income percent income of the loan

Historical default binary variable (YES, NO)

Credit history length

Loan Status Loan status to predict

Problem setting

Binary Classification predicting loan status: (0 = non default , 1 = Default)

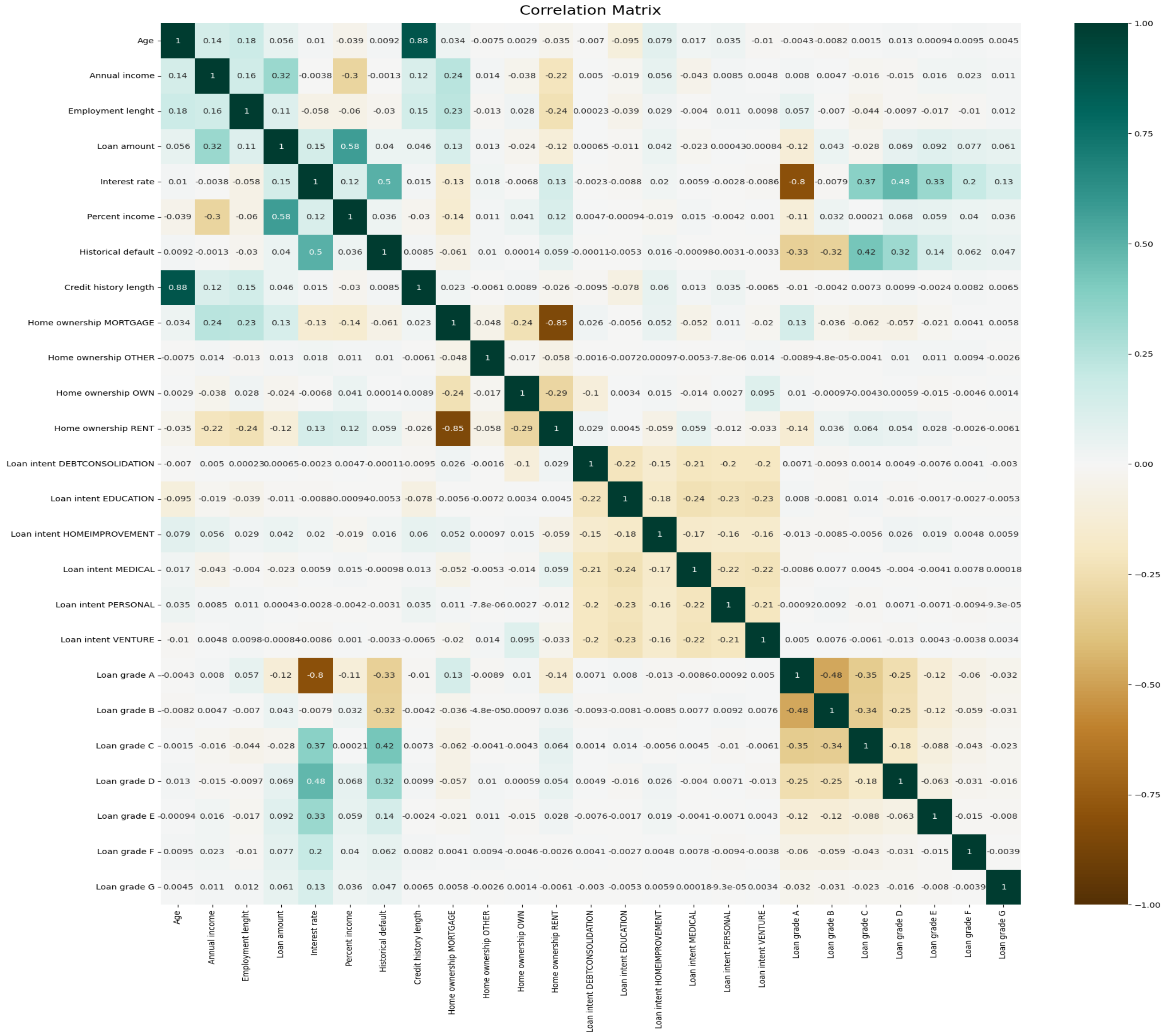
Dataset split 70% of data for training and 30% for testing

Exploratory Data Analysis

By plotting some statistics about our dataset we can see how there are some missing data and errors in our dataset, for example the maximum age found in the data is 144 which it can't be possible. Same thing for the employment length. So as a first data cleaning step we are going to filter out those wrong and missing data

	Age	Annual income	Employment lenght	Loan amount	Interest rate	Loan status	Percent income	Credit history length
count	32581.000000	3.258100e+04	31686.000000	32581.000000	29465.000000	32581.000000	32581.000000	32581.000000
mean	27.734600	6.607485e+04	4.789686	9589.371106	11.011695	0.218164	0.170203	5.804211
std	6.348078	6.198312e+04	4.142630	6322.086646	3.240459	0.413006	0.106782	4.055001
min	20.000000	4.000000e+03	0.000000	500.000000	5.420000	0.000000	0.000000	2.000000
25%	23.000000	3.850000e+04	2.000000	5000.000000	7.900000	0.000000	0.090000	3.000000
50%	26.000000	5.500000e+04	4.000000	8000.000000	10.990000	0.000000	0.150000	4.000000
75%	30.000000	7.920000e+04	7.000000	12200.000000	13.470000	0.000000	0.230000	8.000000
max	144.000000	6.000000e+06	123.000000	35000.000000	23.220000	1.000000	0.830000	30.000000

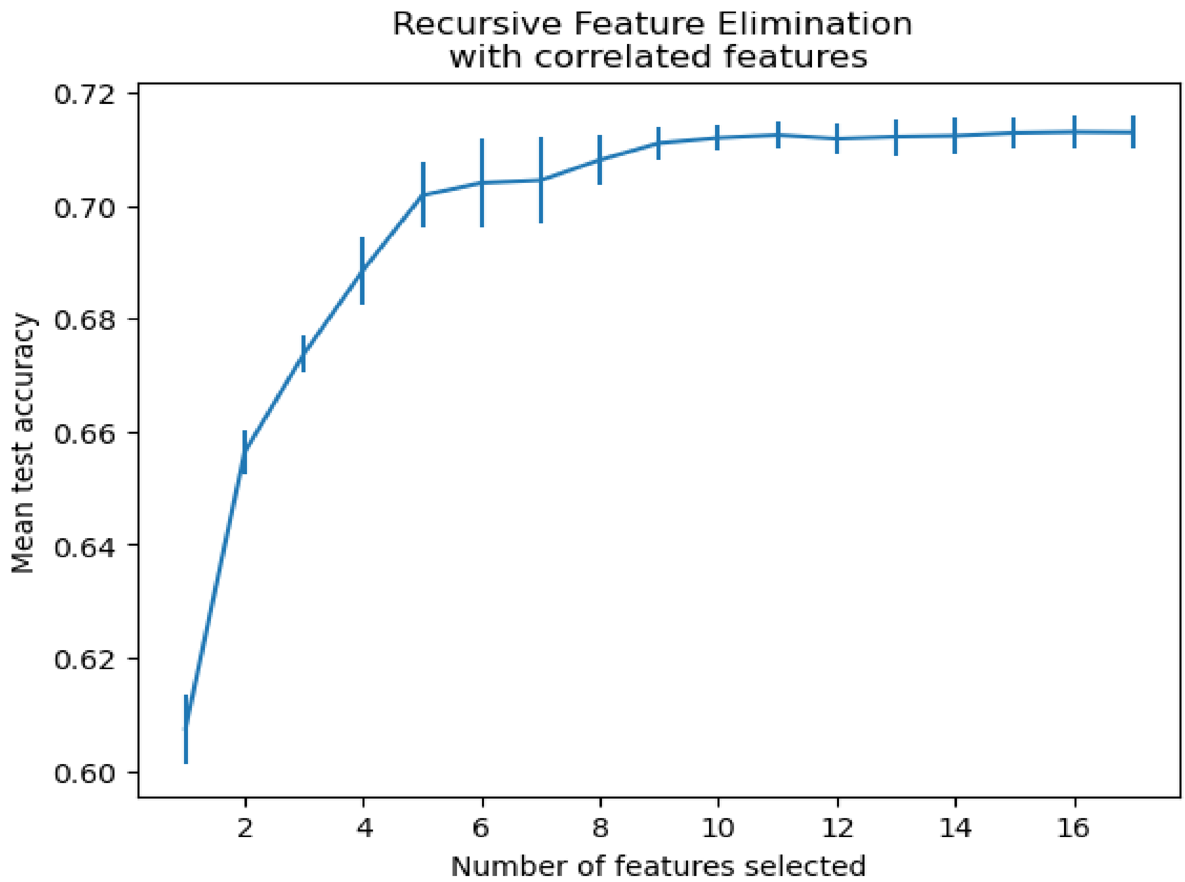
Next step will be to encode all the categorical variables with the one-hot encoding and also normalize the numerical ones since there is an huge order gap between different columns. After this preprocessing steps we are ready to plot the correlation matrix to see if there are some multicollinear variables.



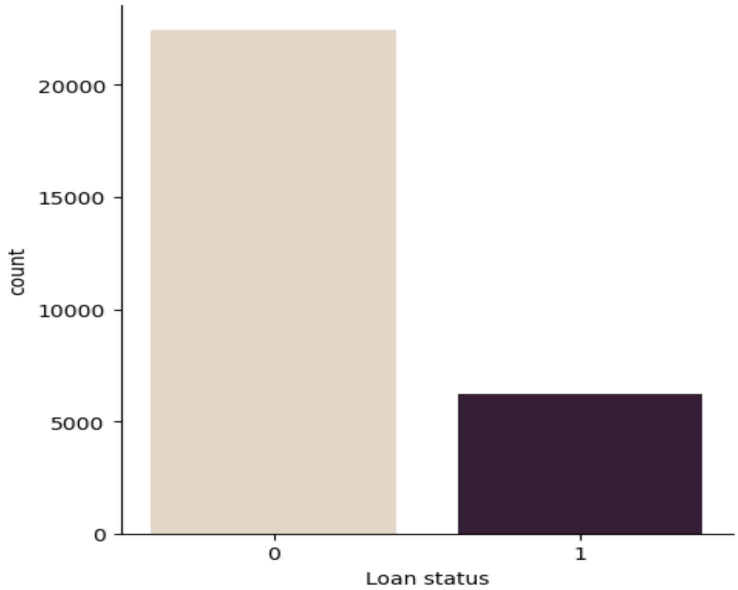
Since the Loan grade is reflected in the interest rate we can drop it from our dataset, along with the age because is correlated to the credit history length.

Feature selection

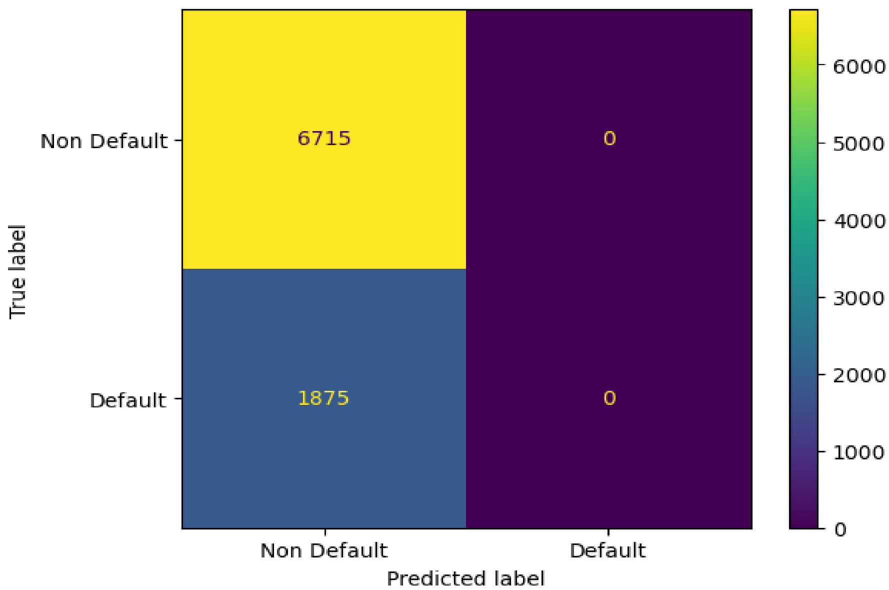
By Recursive feature elimination we see that our dataset is good enough to be fitted to our models. after 12 features our accuracy stays pretty much stable without dropping.



Dummy Classifier as the baseline



Next we plot the number of observations belonging to each class for the loan default. Since our predicting variable is highly unbalanced we can't rely on the plain accuracy, but we have to evaluate our models on a weighted accuracy based on class frequency. If we run a dummy classifier that always predicts the most frequent class we can see how the accuracy is still pretty high due to the unbalance. The balanced accuracy can deal with it.



Accuracy	Balanced Accuracy	F1-score
0.781	0.500	0.000

Predicting Models

Models used

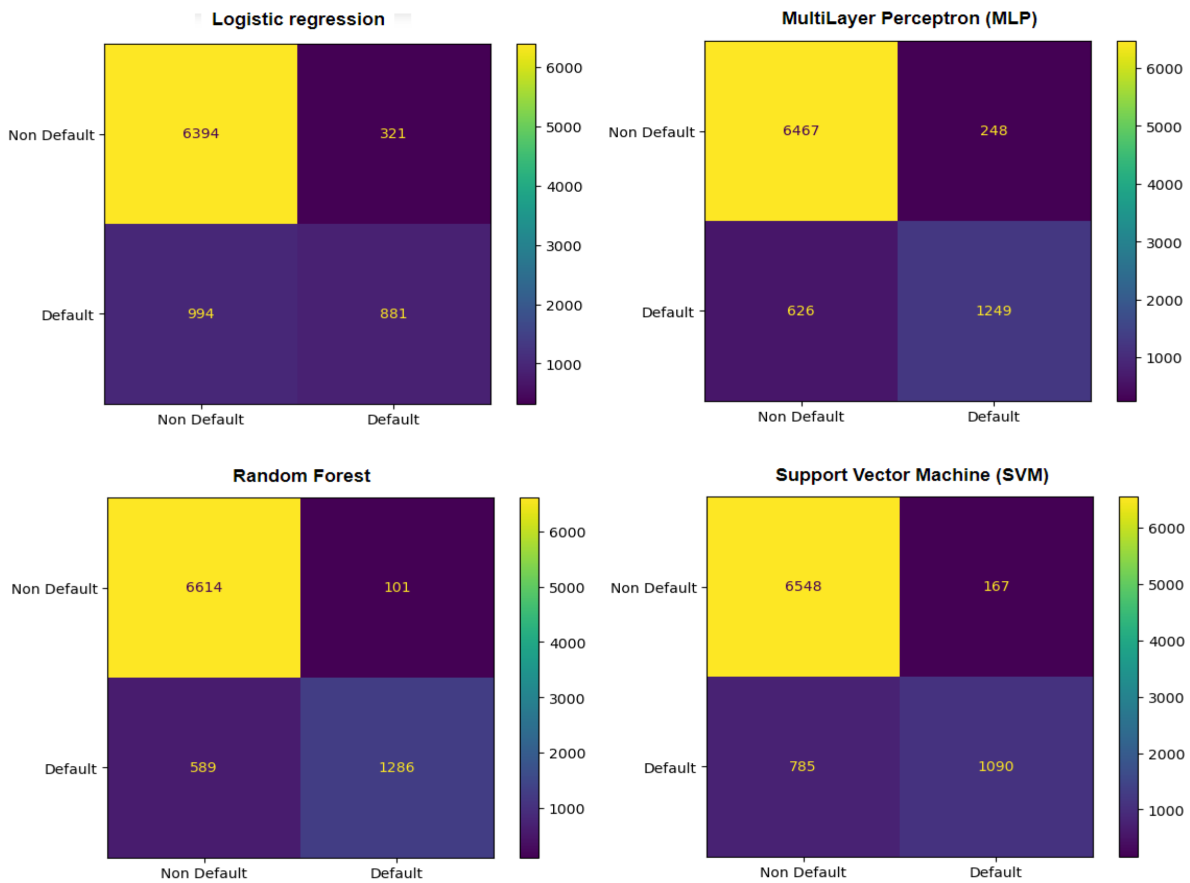
Logistic Regression solver = linear, loss function = l2, regularizer = 10.0

Random Forest loss = entropy, number of estimators = 200

Multilayer Perceptron (MLP) hidden layer size = (30,30,20)

Support Vector Machine (SVM) kernel = poly, degree = 3, regularizer = 10.0

Evaluation with balanced accuracy and confusion matrix



	Accuracy	Balanced Accuracy	F1-score
Logistic	0.846	0.711	0.572
MLP	0.898	0.814	0.740
Random Forest	0.919	0.835	0.788
SVM	0.889	0.778	0.696

Feature Importance

Plotting the feature ranking for the best model , which is the Random forest classifier.

We see how the most determinant factors in a loan default are the person's percent income and the interest rate of the loan