ING - Lion's Den preliminary task

Kamil Kulesza

March 9, 2023

Introduction

Hello, my name is Kamil and I'm a master's student in applied mathematics at Politechnika Białostocka following a career path in data science. I'm coming from an engineering background with a degree in Automatic Control and Robotics which build my interest in data-driven approaches to problem-solving.

In the first section of this report, I will present the modeling approach for the probability of default using a combination of Random Forest for feature selection with logistic regression for interpretability. The second section will consider the application of ESG in credit risk modeling.

1 Task 1

In the given data set for the credit scoring task, we can observe almost 150k observations, of which almost one-third contain missing observations. Because of this, I have decided to consider two modeling approaches:

- **Full case analysis** - a baseline in which we perform the modeling on observations with missing values dropped
- **Imputed case analysis** - trying to gain some information and justifications for possible value imputation.

Also, I will stay away from dropping "outliers" as it should be consulted with an expert or further justification provided for why some values are outstanding. Also, we have almost 150k variables so these shouldn't make our model perform poorly.

Because the values come from very wide ranges due to referring to large monetary values, I have decided to divide all values by a million. Furthermore, the logarithm transformation is applied, but because at the same time, we have both large negative and positive values I preprocess them using the following formula:

$$x = x + |\min x| + 1$$

where x is the i-th column and $\min x$ is the smallest observation in that column.

This transformation allows us to move all the observations on the positive side of the axis and the addition of 1 escapes the logarithm of 0 for the smallest observation when we transform the data. After this, z-score standardization is applied.

Also, an additional variable *obs_count_n* is created out of *obs_date* variable and was one hot encoded to give information about the companies that occur in the consecutive years in the data set. More on this and how it was created can be found in the attached *analysis.ipynb*. As for modeling to preserve the interpretability of the parameters the logistic regression was chosen as it provides information about the classification probability which is given by an equation:

$$\mathrm{logit} p(x) = \beta_0 + \sum_i \beta_i x_i$$

where β_i are the parameters associated with the variables. The probability $p(x)$ can be further calculated using the inverse logistic function as the parameters are originally on the log odds scale.

For the feature selection algorithm, I have chosen the Random Forest feature importance to select 10 best-performing variables. This also gives us a way to look at the complexity of data, we can measure the performance of linear vs complex algorithms given these two models. In the random forest feature importance can be calculated by measuring the purity of constructed nodes, the purity of a variable represents how well it separates one class from another.

The performance of full and imputed case models were collected in Table 1. Additionally, in figures 1 and 2 the AUC and ROC curves are presented. The model obtained in the imputed case analysis provides a better prediction trade-off between falsely classified default and non-default observations. The full case model has in general problem with keeping the prediction clean from the miss classified cases with is represented by the drop in the PR curve. The variable selection process in both cases results in a similar variable set, but there was a lot of information in the imputed data that helps the model better classify the observations.

The final set of variables obtained from the RF feature importance involved variables: [Var_02, Var_18, obs_count_0, Var_11, Var_19, Var_01, Var_17, Var_16, Var_21, and Var_03], interestingly most of them were associated with different kinds of assets, but we also have a piece of information about liabilities. The appearance of variable obs_count_0 informs us that the companies that occur for the first time are more volatile agents.

Something that I started to worry about when tried to interpret the parameters is how the variable such as Assets Current Total, Assets Total, and IFRS_Assets covariate together, and after further investigation, it can be found that the variables Var_02 = Var_18, Var_11 = Var_19, Var_01= Var_16 provide exactly the same information. So the final set of variables can be reduced to [Var_02, obs_count_0, Var_11, Var_01, Var_21, Var_03] removing the colinearity problem and providing a clearer interpretation of the parameters while maintaining the same performance as the model from the imputed case analysis.

More about modeling can be found in the *modeling.ipynb* notebook. I also tried to use PCA to check if there was some interesting pattern after dimensionality reduction to approach detection using clustering methods, but the structure doesn't give much information to separate the classes, this can be found in the *pca.ipynb*.

Table 1: Model evaluation metrics

Model	Precision	Recall	AUC
RF - full case	0.28	0.76	0.82
RF - imputed case	0.27	0.79	0.81
LR - full case	0.20	0.65	0.74
LR - imputed case	0.20	0.75	0.78

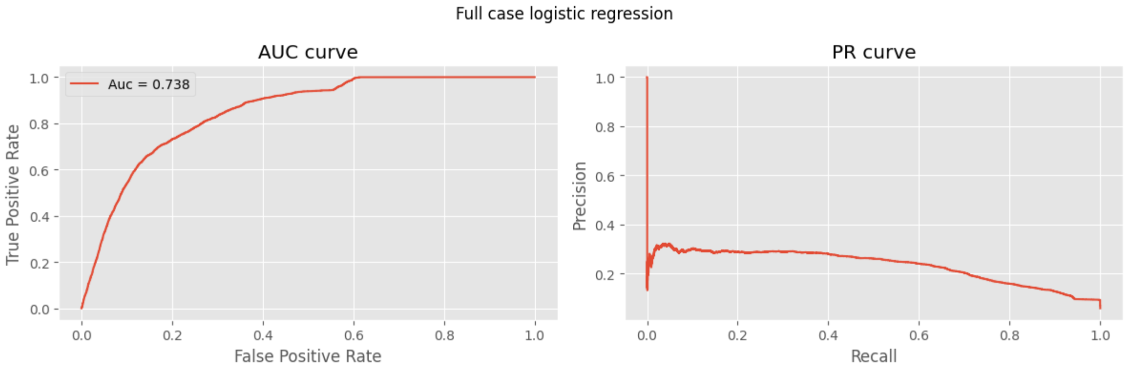

Figure 1: AUC and RC curves for the full case model

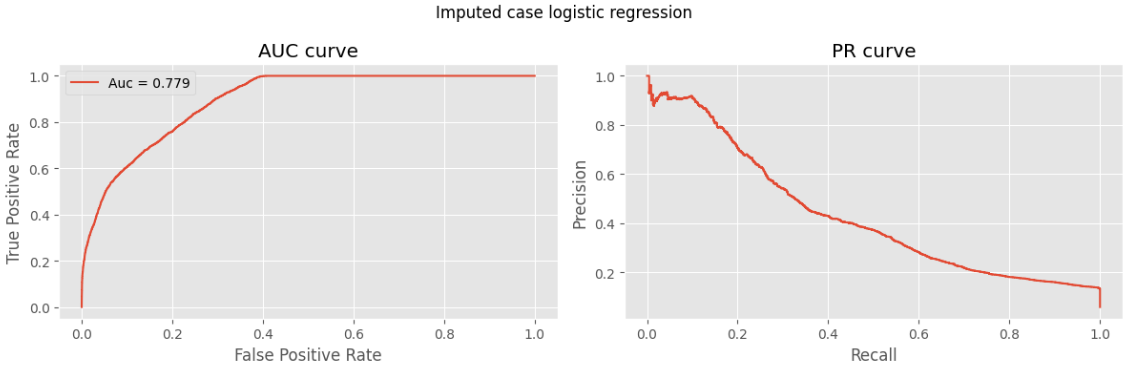

Figure 2: AUC and RC curves for the imputed case model

2 Task 2

From the perspective of a company, having a good public image is a very important part of the operating strategy. As we educate the population about the environmental risk, social acceptance of all groups of people, or fair working conditions any action against these statuses can result in social backlash. This can directly affect the company sales, public strikes against the company, or even lawsuits that could result in a company failing to pay its debts putting the bank at risk.

I do think that there are some interesting possibilities to incorporate ESG factors into credit risk modeling. When the Russian invasion of Ukraine began, many companies escaped the Russian market. To this day there is a lot of social criticism targeting companies like Auchan which did not leave the Russian market. Incorporating such factors could be partially achieved using social media posts or search popularity with the combination of semantic understanding.

Some other interesting factors could involve the analysis of historical trends where the companies change their products to incorporate more eco-friendly aspects and how it impacts their competitors which didn't follow the change.

Thinking about the possible effects of ESG risk in our data could be partially observed via profits and liabilities. As customers decide to strike by not buying the company's products, the sales drop, and they may fail to pay their debts. However, this is really confounding as there are many unobserved cases that affect profits and would require a careful study to assess their impact.

As proposed before, I think it would be interesting to add some long-term effects to see how people look at the company on social media. Using the information about the age of users could provide information on how different groups of people see the company, and how an action that the company makes impacts its social image.