REPORT

Most Important Features in the Credit Scoring Process

JAMES

Published by
CrowdProcess Inc.
1177 Avenue of the Americas
10036 NY, USA

Published in July 2017

Design and execution
CrowdProcess, Inc.

Abstract

The aim of this study is to identify the most relevant variables for credit risk assessment in retail lending. In order to do this, we've used James, the Credit Risk AI, to generate a benchmark using data from 15 major European financial institutions that offer retail credit products to individuals.

After cleaning and preparing the data, we ran a Gradient Boosting algorithm for each dataset. Feature importances were then collected and compiled to give a general view of which variables are most likely to help financial institutions assess the creditworthiness of their customers.

The results give a strong insight on what one of the most advanced machine learning algorithms considers to be characteristics with a high predictive power in a credit risk context.

Introduction

In the majority of predictive modelling tasks, features do not contribute equally in the prediction of the target behavior. In fact, some of them have no predictive power whatsoever. Distinguishing between relevant and irrelevant features to predict the performance of a given target is indeed a major step in a data scientist's modelling journey. This modelling step might have significant implications in the model interpretation and in the decision one would make based on a given model.

Feature importance calculation

Individual decision trees intrinsically decide feature selection by picking appropriate split points. The information gathered from those decisions can be used to measure the importance of each feature. Called "Gini importance" or "mean decrease impurity", the feature importance is defined as the total decrease in node impurity, weighted by the proportion of samples reaching that node.

This calculation is made every time a variable is used as a split points of a tree. This can be extended to decision trees ensembles by simply averaging the feature importance of each tree.

Used methodology

To make this study, James used 15 datasets provided by major European financial institutions. After the usual process of data cleaning and preparation, we ran a Gradient Boosting algorithm for each dataset independently. The 10 most important features of each model were extracted and each individual feature was assigned to a broader category in order to be comparable with features from other datasets. Finally, these categories were ranked according to the number of times they appeared in the 10 most important features of each model.

As a concrete example, all variables regarding individuals' incomes were grouped into an Income category. This Income category includes variables such as working income, housing income, other income or, more broadly, financial assets. These categories, described in Table 1, can be understood as types of information.

Table 1 - Feature Categories

Variable	Description
Client details	Type of client, years as client, subscribed products, etc.
Credit Product	When a dataset contains different products
Job details	Years on the job, job seniority, job title, etc.
Income	Working income, housing income, other incomes, financial assets, etc.
Interest rate	Interest rate of the loan
Age	Age of the applicant
Economic activity	Working field
Score	Scores provided by underwriters
Credit history	Credit payment history
Loan term	Term of the loan
Housing details	House type, ownership, housing costs, etc.
Geography	Postal code, city, region, etc.
Consumption behavior	Living costs, expenses, withdrawals, etc.
Nationality	Nationality of the applicant

Limitations and Results

Limitations

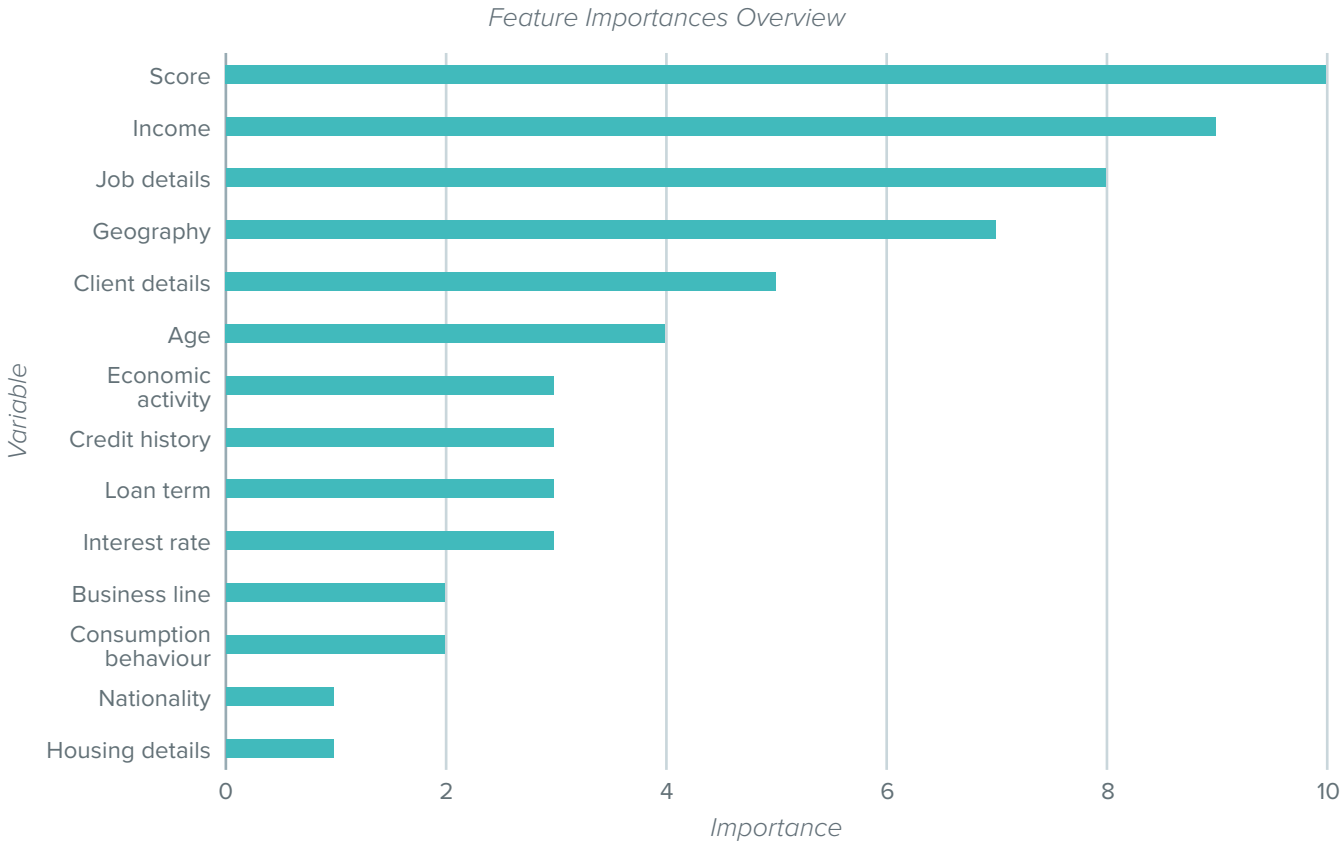
Before presenting and analyzing the results, we need to state the limitations of this study.

Datasets and feature importance specificities were not taken into account. As the presented ranking was only made according to the number of times a feature appeared in the 10 most important variables of each model, the absolute weight of importance (weight given by the Sklearn feature importance calculation) was not considered. If it had been used instead of the count, the ranking might have been slightly different.

Also, the interactions between variables in each datasets were not studied: due to correlation and other statistical interactions, a variable might show different importance levels, depending on the set of variables it's combined with.

Presenting and analysing the results

The results, presented in Graphic 1, show that Score, Income and Job details are presently the most relevant features in predicting a target's behavior.



It is interesting to note that, apart from "Scores provided by underwriters", the top 6 variables are of socioeconomic nature. Namely, "Income", "job details", "geography", "age" and "economic activity" are the category of

variables showing up more frequently in the 10 most important features of James models'.

The client details variables, showing up in fifth position, also seem to be having a significant impact on predicting the target's behavior.

From 8th position downwards, we can observe that most categories are related to the loan itself (loan term, interest rate, credit product), or to the client behavior when contracts a credit (credit history, consumption behaviour)

Nationality and Housing details are showing up one time each in 13th and 14th positions respectively.

Conclusion

The goal of this document is to let lenders know which basic information they should be collecting from their customers in order to create statistical models for default prediction.

This information can be useful not only for financial institutions that are going to launch in the near future but also for established lenders that might have room to improve their data collection and statistical modelling.

As an outcome of our research, we recommend the collection of all relevant data of socioeconomic nature and those describing the terms of the loan itself. Since the quality of the data is one of the most important factors in statistical modelling, we encourage lenders not only to collect this data from clients but also to make sure it is well kept and structured.