“Use Of Machine Learning Techniques to Predict the Likelihood of Default”

Public Project Summary

Nuno Gomes

This is a summary version of the Capstone Project Report *"Use Of Machine Learning Techniques to Predict the Likelihood of Default"*, completed as part of the MSc. Business Analytics at Central European University.

Supervisor: Zoltan C. Toth

Introduction

The purpose of this project is to apply Machine Learning models and techniques allied to traditional statistical methods to a work database in order to predict customer's likelihood to default, based on contract details and previous payment behavior. It is focused on commercial entities whose financed billing was analyzed, resembling traditional industry leases and loans.

Customer's behavioral data, complemented with Default Flag and derivative features is to be used as an input in order to understand, based on dates, financed amounts and contract details, which are the predictors that Companies may be able to use to identify and prevent potential default situations.

Accurately predict the likelihood of default as well as identifying the key drivers that determine this likelihood, equips the Company with better tools to understand their current and potential customers, improving their financial planning and better informing future strategies and actions.

The subject

In a competitive Business environment as today's, if on one side we have companies that on their quest to survive and thrive, need direct access to capital and financing solutions, on the other side, we also have companies willing to take the risk and make that capital, directly or indirectly, available, expecting on their turn to be properly compensated, for a deal involving a risk as lower as possible, under an agreed set of conditions.

On the lending side, in order to assess viability, profitability and associated risk, the lenders want to know if a potential customer is expected to pay on time or if it will be delinquent and ultimately even default, so when a business applies for a Loan, the lender must evaluate if the potential borrower can reliably repay the loan principal and the associated interest. Ultimately it will depend on Lenders appetite for risk and leverage level and on Borrower's creditworthiness.

Data

The data used during the analysis was real production data, property of the current employer, that due to confidentiality reasons should not be shared, other than with the CEU for the present project. The data was deanonymized, till some extent, in order to prevent unauthorized use and disclosure.

Due to the Company's complexity on its processes and structure, the data was natively stored in different databases and subsystems, to which different accesses and authorizations were necessary, namely:

- Contract information databases
- Accounts Receivable and collection databases
- Billing databases

- Credit databases
- Customer databases

The goal is to understand if based on the billing and contract details, allied to the payment behavior, a given entity may be identified as more prone to default. That said, it was Author's initial assumption that small nuances related to the payment behavior could be a good predictor of such an event. Original considerations and nuances that the feature engineering attempted to capture, taking into account the limitations concerning data availability.

Data Exploration and Feature Engineering

Using some of those available fields, upon exploration and assessment, different levels of aggregation and calculation were performed during the feature engineering process, initially via SQL code and later already on RStudio, according to the package, model and purpose. Ultimately the process led to the creation of 233 derivative features.

Modeling

In order to put in practice what was taught throughout the course, the decision was to train 3 different solutions, firstly to test acquired knowledge and secondly to understand if any of them would significantly considerably outsmart the others, or if on the other hand they would give similar results for the same data.

GLM – Generalized Linear Model

It is considered the go-to method for Binary classification problems (cases with two class values), being a multiple regression with an outcome variable (also known as dependent variable). Although able to move with non-liner function, working at some extent with linearly and non-linearly separable problems, it is not flexible enough to naturally capture more complex relationships. It is still used as credit Industry reference and benchmark in several scenarios and even complex ensemble models end up comprising a linear model amongst the algorithms.

Upon getting a consolidated database, several iterations were performed and considering correlation, P-values and packages like caret(varImp) it was create a list of relevant features that demonstrated better prediction abilities, already excluding any possible collinearity.

ANN – Artificial Neural Network

ANN may be briefly described as computing systems that attempt to mimic biological neural networks. They have the ability to digest a considerable amount of inputs and process them to infer hidden as well as complex non-linear relationships. Currently they are broadly applied on image analysis. Although they are able to translate complex non-linear relationships into an insightful output, it is very complicated to deconstruct and explain how they were able to retrieve that output, being known as "black box" due to the lack of visibility on its inner operations.

Conclusion

Due to report's nature and limitations in terms of time and access to data that could be disclosed, conditions and features that would have been able to improve the prediction capabilities were left outside of the scope.

Nonetheless, payment behavior seems to be a good indicator of a customer's ability to fulfill its credit responsibilities and repay the Borrower during the duration of the contract.

The approaches discussed on the report present acceptable results – Accuracy: GLM 0.79 and Neuralnet 0.85 -, but nonetheless it is understandable that to improve default prediction accuracy and the credit risk modeling as a whole, data dimension should be the focus area. Apart from the financial information, payment behavioral data, transactional data, social media data, geographical information and other sources not so conventional as the ones used till now, can potentially add a considerable degree of insight.