Spending pattern embeddings for credit risk assessment

Andro Merćep, mag. ing.
(andro.mercep@fer.hr)
mentor: Prof. Zvonko Kostanjčar, PhD
University of Zagreb Faculty of Electrical Engineering and Computing

1. Introduction

Probability of default (PD) estimation – one of the key factors in the credit risk assessment process

- A debtor is considered to be in *default* if they are more than 90 days past due on their credit obligation
- Usually estimated using socio-demographic features, account balance, client's credit history and overdraft utilization, tenure features, etc.
- Evaluating new loan applications using new data sources such as customer spending data

2. Problem Description

Customer spending data available from 2013 to 2017

- Total amount (in HRK) that a client spent in some store (company) in the 12 months preceding the loan application date
- Data from 2017 are used as an out-of-time test set

Client ID	Application date	Store	Amount
10000	2015-09-30	Spar	1623
10000	2015-09-30	Lidl	781
10000	2015-09-30	dm-drogerie markt	520

Creating a pivot table with 220,000 clients and 1598 stores (i.e. features)

- Clients typically shop in a small number of different stores – 98% of data equal to 0
- Sparsity causes performance issues; logistic regression achieved an out-of-time Gini coefficient of 0.198
- PCA did not provide a meaningful improvement with a score of 0.205

3. Methodology

Proposed approach based on *spending context*

- Sparse high-dimensional vectors occur in NLP when representing words as one-hot encoded vectors
- Such representation lacks any syntactic or semantic word relationships
- Word2vec creates the word's representation by using its context, i.e. the words surrounding it
- Similarly, we can define the spending context on a customer level – each customer will represent a "sentence", and "words" will be the different stores
- Stores that occur in similar contexts will have similar embeddings

4. Results

We trained two models based on different word2vec architectures: continuous skip-gram and continuous bag-of-words

- Note that both approaches only consider the location (store) where the transaction took place, and completely disregard the amount spent

Model	Gini coefficient
Logistic regression	0.198
PCA + logistic regression	0.205
Continuous skip-gram	0.325
Continuous bag-of-words	**0.340**

Out-of-time test set results (2017 data)

5. Conclusion

- We have developed an alternative approach to the credit risk assessment problem using customer spending data
- The proposed models provide a substantial increase in performance when compared to industry standard benchmark
- Both approaches maintain the required high level of model explainability

6. Project Acknowledgement

This work was supported in part by the Croatian Science Foundation under Project 5241, and in part by the European Regional Development Fund under Grant KK.01.1.1.01.0009 (DATACROSS)