Gergely Kinizsi – Capstone project summary

Introduction

For my capstone project, I've taken part in an internal project of General Electric: project Janus. This seemed to be the optimal choice for me, as I'm an employee of GE Global Operations Services in Budapest and my designated tasks were in scope of what I've learned in the Business Analytics MSc course.

Prerequisite - Data lake

GE started to build its data lake infrastructure in 2017. This was a prerequisite for any company wide data related project, as the company used 100+ ERP instances from various providers: SAP, Oracle, Peoplesoft just to mention the bigger ones. Decision was made to cut this number and use only strategic ERPs (certain instances of SAP and Oracle). Also, ESCOA was introduced and strategic ERPs were allowed to use only this type of Chart of Accounts.

Project Janus

Project Janus' goals are: transformation of cash in bank recording, reconciliation and reporting activities. It focuses on standardizing the work, digitize to reduce effort and centralize responsibility. The scope is over 7000 bank accounts within the company. Strategic tools in focus are ATM, OCM and SAP CM. The expected result (after 2 years) is to halve the FTE working in the field and al bank accounts to have standard, automated, centralized process. It's clearly visible from the above, this is a huge effort and I was involved in a very small section of it.

The aim of the project is to standardize processes throughout the finance departments of the company. Before the initiative was started, GE experienced the following problems:

- Different reconciliation methods
- No common source of truth
- Difficulty in collecting „open items"

Automated Transaction Matching tool

One deliverable of the project is ATM tool, which will provide solution to the above-mentioned problems, plus new features as well:

- Standard reconciliation method
- Data refresh, reference tracking automatic
- Automation of booking/rec via rules
- More efficient process (ATM efficiency can be measured via auto booking and auto match ratio)

ATM is a homegrown tool of GE, developed for some years now. Its aim is to work as a cash management module, where the ERP lack this ability.

Categorization

My task was to categorize the movements in ATM, based on the data available, to 28 predefined types. These movements are bank transactions, coming through all affected bank accounts. The types cover all transactions, incoming and outgoing as well. None should remain without type. This task was suitable for me, as I have worked previously in the Cash accounting team, which handles the bank related bookkeeping. Also, my current role is in Treasury, so I have perspective from bank technical side as well.

Machine learning

I tried to acquire the source code of the algorithm, but I was rejected as it's sensitive and classified information. However, I received the technical details from the data scientist. The code was written in Python, she used RandomForest classification as method. There was need for further data preparation: bag of key words was created based on all the text fields (mainly description, brtn, brcr, etc), which excludes words – stop words, short length words, low frequent words, etc.

Using the results

The results of the machine learning are not directly applied to new data. The ATM tool works with "rules". These are simple functions, checking given criteria. Ie.: If a transaction comes through bank account X, with the description containing Y and the movement is on the Debit side, then it's considered to be Z type of movement.

Therefore, the result of the machine learning was used for rule creation. Based on validation sample (140k), totally there are 6000+ rules, in which top 1000 rules with high accuracy (95%) cover ~80% total transactions and ~90% total correct predictions.

There are plans in place to abolish rules and apply the machine learning continuously on data. However current setup and infrastructure, this is not currently feasible.

As I learned during the MSc course, 80-90% of the time is data preparation. In this case it was 100% for me.

page intentionally left blank

page intentionally left blank

page intentionally left blank

page intentionally left blank

page intentionally left blank

page intentionally left blank

page intentionally left blank