Department of Economics and Business
CENTRAL EUROPEAN UNIVERSITY

Master of Science in Business Analytics

CAPSTONE PROJECT

Public Project Summary

Parsing client documents for risk assessment process

Professor: György Bőgel

Supervisor: Eszter Windhager-Pokol

Student: Abduvosid Malikov

Summary

This summary is dedicated to describing the capstone project completed with the Fintech startup. The main operations of the start-up include digital factoring. In factoring, the creditor conducts some service with the debtor. To receive the payment for the conducted service earlier, the creditor approaches the Fintech startup, uploads its sales ledger invoice PDF document that shows the amount and date of payment, and easily gets the money.

The aim of the project is to automate the process of analyzing the uploaded invoice. These documents contain transactional data. Also, this is in line with the recent introduction of the PSD2 directive which soon obliges companies to provide the last 3-month transactional data. Therefore, it is aimed to automate the following workflow:

- extract the transactional data from the invoice documents
- transform data
- analyze the relationship between the transactional data and goodness of the client (whether the client is good or not)

The end goal is to validate and assess these materials according to the risk methodology. More clearly, by using certain tools, documents should be read, processed, and analyzed.

Benefits to the client

Currently, one of the executives of the startup is involved in reviewing these documents. Solving the problems brings value to the business. For instance, it enables to save time and process much more documents at a given time. It also helps seniors to focus on other strategic issues rather than daily operations. Last but not least, clients may approach this startup again to see whether the debtor is a well-performing company or not. This can potentially increase customer retention.

The company stores its data in the cloud bucket. The web application that accesses this cloud bucket and performs other operations already built. The solution project that simplifies the document reading and data analysis process will be one component of this big ecosystem.

The reason to prioritize this task over others is that analyzing the document is the business's core activity. By making this activity more efficient, a startup gains a competitive advantage in the market.

Key outcomes

Stages

After several meetings and discussions with Chief Risk Officer, it was decided to divide the project into 3 phases:

- 1st phase is to build a sales ledger reader that extracts meaningful data from PDF documents
- 2nd phase is to build a sales ledger transformer that transforms extracted data

- 3rd phase is to make a sales ledger analysis based on the transformed data. The output of the analysis is the report that indicates the positive performance of the client based on its historical transactions or indicates warning sign.

External libraries

Reading PDF documents is almost no different than easily reading any other popular files, such as CSV or Excel files. However, for this case, we need a tool that can read tabular data from PDF efficiently. There are already some libraries developed to solve this task. The list includes such libraries as Tabula, pdfplumber, pdftables, and PDFminer. However, some of these libraries are not maintained by their developers anymore or are not efficiently solve the current task at hand. There is one more library for reading tables from PDF files named Camelot. It is a Python library that can help to extract tables from PDFs. When the developer of the library Vinayak Mehta compared this library to other libraries, it won in several categories like reading header row correctly, identifying column separators, and assigning the columns based on the text edge. Camelot is reliable as well. It has been starred by more than 1100 users on GitHub and is built on top of the PDFMiner tool (4600 stars on GitHub). Also, the python package camelot-py was scanned for known vulnerabilities and missing license, and no issues were found in the health analysis review by Snyk Advisor. Thus, the package was deemed as safe to use. Consequently, I decided to use this library.

Besides this, I used standard libraries for data analysis such as Pandas (working with data, data manipulation) and NumPy (data types, working with numbers).

The data extraction outputs text data. We need some tool for parsing this text data to find, replace, or delete a certain string. In this case, Regular Expressions should do the trick. Regular Expressions, also known as regex, are a sequence of characters used to check whether a pattern exists in a given text (string) or not. They help in working with textual data, which is often the first step in data science projects that includes text mining. With regex, we can find and match the necessary data (journal number, debtor name) from the text and assign them to the corresponding group. Thus, I imported the re module as well. Also, I used https://regex101.com/ to check whether the regular expression pattern I built matches the necessary text.

One more module I used is collections and it's namedtuple datatype. It is needed for storing the output in the necessary format.

Evaluation of output

The evaluation of output was one of the difficult parts of the project. When a PDF document contains only one page, with quick eyeballing we can verify manually that the data in PDF matches with the extracted dataset. However, when a PDF document contains dozens of pages this technique does not work. More clearly, the output of the first phase of the project is the dataset extracted from the PDF document. But how to ensure that the data in this dataset totally matches with the data in PDF? The fact that each invoice PDF document contains the total amount comes in handy in this context. A solution suggested by Python for CPAs recommends checking whether the total sums from the PDF document matches with the one in the dataset (Python for CPAs, 2019). This validation technique was used: the total sums numeric columns in the extracted dataset matched with the total sum given in the PDF document.

Lessons learned

At the beginning of the project, it was clear that the Project Sponsor wants a system for Credit Risk Assessment. By reading one PDF document, I could extract only one dataset. However, building, training, and testing a model for Credit Risk Assessment required many hundreds of more data.

Data Science projects are difficult both for business owners and data scientists. Business owners have a problem at hand, but they do not know how to technically implement it. Data scientists have that practical experience however they need clear business goals. When these two parties collaborate, they can solve the problem at hand.

This project has served as a clear example of that phenomenon. I started everything by having a first meeting with the company representatives. The business problem and most importantly technical solution for it were unclear for me at this stage. However, I was curious about this task and decided to dig deeper in any way. With weekly status meetings, feedback, tons of Googling, reading the documentation remedied the situation. In the end, I could arrive with the solution to that task.

On this last note, I want to highlight how having a positive attitude important and promoting a Growth Culture are important in the Data Science process. According to Tony Schwartz, focusing on building a culture of growth is more effective than building higher performance cultures. The author provides these components of such culture: “an environment that feels safe, a focus on constant learning, time-limited experiments, and continuous feedback”. In my personal experience, I am very thankful to my Project Sponsor for managing me under such a culture. Because I was submitting other academic assignments, sometimes I could not perform as expected. In those situations, he would understand my status and encourage me to move forward with the project. As a result, I could continue the project with big motivation and finished it.

To conclude, it is worth applauses when the company makes data-driven decisions. It can appear in the form of automating its processes or building an infrastructure that predicts some outcome. Be it the former or the latter, it is much more productive when two parties organize it around the culture of growth, as this project proved.