Codebase creation for strategic supplier spend allocation and time-series forecasting of non-preferred spend with different machine learning algorithms

Project Summary for CEU Business Analytics MSc

Sabina Umarova

June 2022

Contents

1 Abstract

The following report summarizes the work done, key outcomes and learning experience during the Capstone Project for the CEU MSc in Business Analytics Program. Due to the non-disclosure requirement, some details regarding the client and the project are removed from this document.

2 Background and Client Introduction

Spend analytics is a crucial part of any procurement strategy but not all organizations implement it in their strategies. To gain the benefits from reviewing spend data it is essential to look at its lowest levels that are transactions, payments, and other information. The typical challenges that companies may face in reviewing such data and implementing the spend analytics include incomplete, not updated, or fragmented datasets making it difficult to synchronize information, manual or rules-based data classification leading to static and inaccurate results, visualization with proprietary tools that are not always user friendly and difficult to navigate. As a result, a company misses or misclassifies half of the total spend data and data insights which leads to the loss of savings opportunities.

The client team of my capstone project provides data driven insights to its external clients to discover saving opportunities. To enhance the effectiveness of procurement organization in companies, the team always strives to identify new ways to gain important business insights which can enhance sourcing decision making. Three levels of team's delivery are Data Curation, Use Case Visualizations, and Insights, which mainly includes evaluation of data, provision of strategy and roadmap, data cleaning and normalization, spend categorization, strategic decision-making visualizations, and insight models.

In my capstone project I work with the team to help them with one of the use case visualization tools, mainly to create a strategic supplier spend allocation codebase and perform time-series forecasting of spending from non-preferred suppliers for it.

3 Project Description

The core idea of the project is to use multiclient dataset to identify top preferred suppliers and define how much of client's spending from non-preferred suppliers can be allocated to market preferred suppliers so that bolstering the client's negotiation capabilities regarding discounts and other bonuses from the strategic suppliers based on historical and forecasted values of spending from non-preferred suppliers. A supporting codebase should be created in such a way that it can be reproduced for any client with minimum updates in the code, that is to make it as automatic as possible.

The crucial part of the project includes the time-series forecasting analysis which includes training and fitting prediction models for total non-preferred spending amounts of each spend category with three different methods supported by Python programming language, that are Holt-Winters Exponential Smoothing, Autoregressive Integrated Moving Average and Prophet, an open-source library for univariate time series forecasting developed by Facebook.

Data processing and analysis can be performed on a computer using the following open source and free tools:

SQL - The project contains an SQL script to communicate with the database, efficiently process the data and mainly create a codebase.

Python - The project contains a Python script for data exploration and time-series forecasting analysis.

Cognos – The project contains an interactive dashboard in Cognos as a final deliverable.

There are two main data sources for this project. The first one is the multiclient dataset which includes historical data from multiple clients and will be used to identify preferred suppliers in the market. The second data source is a dataset of one of the company's clients spending. Data in both datasets is mainly coming from invoices, purchase orders and other procurement-related documents. Although the data cleaning, data normalization and spend categorization were performed as part of the data curation process by the company team, some transformation pipeline is still needed for this project.

To communicate with the client database and download a multiclient and a client data I am using DBeaver which is a SQL client software application and a database administration tool. The multiclient dataset used as market data for this project was decided to be restricted for the invoice period from 1 July 2019 to 30 June 2020. The client dataset used for the strategic spend allocation was restricted for the fiscal year of 2021. The time-series forecasting analysis includes client's data for the invoice period from 2019 to 2021.

4 Codebase Creation for Use-case visualization

The codebase mainly includes SQL queries creating master data table and supporting tables from the existing data archive. The master table is going to be used for the use-case visualization, that is the Supplier Consolidation Dashboard. The dashboard is designed to assist Procurement Professionals to identify preferred and non-preferred spend under a certain Category based on the multiclient dataset. This dashboard will enable them to achieve additional savings by applying rebates.

The final master table for the Strategic Supplier Spend Allocation Codebase includes following variables: Supplier Name, Supplier Country, GEO, Category, Client Preferred Spend, Total Non-preferred Spend, Consolidation Portion, Total Spend, Rebate, Saving, Saving Rank and Saving Flag. The codebase includes all the instructions and also includes the possibility of adding suppliers as preferred in the list of Top 10 suppliers when it is not there, such that the code can be easily run with updated data for different clients.

As a result, the interactive dashboard was created in Cognos based on the created master table with the historical values, where filters for the country level, geo level, category and saving opportunity can be made.

5 Time-Series Forecasting

To extend our dashboard with the predicted values I performed forecasting analysis for total non-preferred spend. Under pattern decomposition each spending category resulted in different trends and seasonality, that is why it was decided to conduct forecasting analysis separately for each spending category. The future period for the forecast was defined as the first quarter of 2022. An assumption made for the dashboard is that preferred spend amount equals the preferred spend amount in the first quarter of 2021.

The forecasting analysis includes Stationarity Test, Building and Fitting Models, Evaluation and Comparison of the Models, and Forecasting for future periods for the all 13 spend categories. Time-series data for 5 of the categories resulted in non-stationarity, thus I used algorithms that can handle stationarity. As a result SARIMA method performed the best for 8 categories, Prophet for 3 categories and Holt-Winters for 2 categories. Since we had a tiny data for this analysis, we should keep in mind that the results obtained with SARIMA, Facebook Prophet and Holt-Winters Exponential Smoothing methods can change significantly for all the three models with a larger dataset.

6 Summary, Next Steps and Learning Experience

The SQL queries I wrote for the use-case visualization can already be used for other external clients. The resulting dashboard shows possible savings from switching from non-preferred to preferred suppliers, rebate portion, savings and saving rank filtering for the supplier's country and different consolidation portions of non-preferred spend.

With the efforts of machine learning algorithms, I was able to forecast the desired values of non-preferred spend by category levels. As the next step, with a larger amount of historical data, forecasting analysis can also include a country level. Current analysis can be easily reproduced with the notebooks created in Python and can be used as a basis for forecasting with other algorithms.

Overall, the capstone project provided me with the unique opportunity to implement the knowledge gained during the program. Being fully integrated to the client's team helped me to get a better understanding of companies providing data driven insights to external clients and to learn more about the procurement data and spend analysis.

Most classes of the Business analytics program were R language based and I had only one SQL class and one Python class. However, for myself I find Python a very useful language and also it is used by my client, therefore my current project was conducted mainly with it. Moreover, doing additional online courses and reading related articles helped me with the coding part for the time-series forecasting with Python.

I think the capstone project played a crucial role of CEU MSc in Business Analytics Program, as it provided me with a chance to observe a new business area, deepen programming knowledge, improve communication skills and finally it gave me with a chance to gain an experience in the leading data analytics company and be a part of the of professionals.