Capstone Project Summary

Kornel Kovacs – MS in Business Analytics 2020

Table of content

I. Executive Summary

This project aims to build a system for powerplant outages in Europe. They have an influence on the supply side, therefore on the price of energy. This is the reason why it is worth monitoring outages. The purpose was to design and implement the entire pipeline from data collection to visualizing the relevant aspects of it. The system was built using Python, Oracle SQL and R. With this product, you may perform automated data collection for all European countries which are piped to a database. In order to have a visual aggregation of so much information, a dashboard has been designed.

II. Data retrieval

The key aim of the data retrieval process was to find a reliable source of information. One platform has been discovered where all outages are publicly available. You may register for a free API key and start using it. However, the *reliable* part is somewhat debatable. In order to leverage on the site, one needs to implement quite complicated solutions.

Regarding the hardships, we obviously have an input to describe some spatial feature (that is, which country.) Moreover, the API requires timewise inputs as well, specifically a start and an end date. This is fine so far. The very first issue was that the API has a limit of 200 entries to return. As you need to provide the start plus end date a priori, you have no idea of the

distribution of outages, therefore you cannot adjust your query. To overcome this issue, I have implemented a recursive algorithm that would always split the date range in case of failure.

Another discovery I have made while conducting my project was that it is sometimes very cumbersome to implement an IT-related work in a corporate environment. Restrictions are made on an extreme scale. The firewall will not let your code pass, you may not be allowed to install any software unless absolutely necessary. The definition of necessary may not be the same for a data scientist and a system administrator. These are the kind of situations when a data scientist needs to also be a great consultant.

III. Data cleaning & upload

The importance of data cleaning and exploratory analysis should not be underestimated. As we were using and official API, I thought it would return nice and clean entries. It was not the case. As many countries contribute to this database on their own, the database is full of duplicates and unreal values (e.g. negative outage values). Not only we had to implement subset-based duplicate filters, but also needed to find insights to uncover in what ways mistakes were made from different countries. As my client had more domain knowledge than I did, they helped me a lot identifying major factors and eventually come up with quite complicated filtering rules.

Some countries in Europe has quite unique characters in their ABC. Germany has ß which sounds similar to *s*. The Czech Republic has many letters with a sort of apostrophe on the top of them (e.g. Č). In order for them to fit the database, proper encoding needed to be done. Furthermore, the right choice in terms of the type of the column was crucial.

As for the database design part, my client used Oracle DB, therefore I designed the schema to fit the general rules of a relational database. Obviously, it contained multiple tables as this is the standard. It turned out to be too difficult to query for end-users, therefore I was asked to redesign the structure in a way that it only has one table. This is how the final structure came to be.

IV. Data visualisation

As for data visualisation, there were two clear objectives and end goals in mind. One sort of visual is for a broad audience, that potential customers may subscribe for. The purpose of this is to give them a hint of the product to make them interested and to justify that this product

indeed has the features they are looking for. The other part of visualisation is a dashboard. This has advanced metrics and customisable features, plots to give insights of what may be going on in the energy market. This is for the phase of sales when a customer is engaged with the product and already expressed their high level of interest.

As for the former, again, the clear purpose is customer acquisition. I designed a system that would automatically generate weekly aggregates of powerplant outages for specific countries building on the database. These plots get sent out to potential customers who signed up for them. This process serves the purpose of making them interesting and engaged with the product.

The latter part is an R Shiny dashboard. This is under further development. However, an interactive and nice-looking dashboard is definitely more appealing than a simple plot is. The dashboard is connected to the Oracle DB, therefore receives the updates real-time helping the decision-makers do their work more efficiently.

V. Pipeline

As a closing chapter, I intend to specify some technical aspects of my work and pipelines. This is clearly an engineering heavy project without any fancy conclusion for decision makers. This is a system to help decision makers better apply their domain knowledge and hence make more profitable decisions.

The scraper system runs on a simple windows machine. We are using windows scheduler to automatically run the scripts in every necessary period. The code base performs downloading, cleaning and upload to the Oracle DB. Those entries that are filtered out are not lost. They are piped to a CSV file with an option of later including them to the database. This part of coding is implemented in Python.

The aggregated plots are also generated in Python querying the database. The legal issues about subscriptions are not yet resolved, therefore, it is not yet connected to a user database. However, I have already implemented functions to send e-mails with attachments. Everything is set and ready to go live as soon as legal worries disappear.

As for the dashboard, this is in an early stage. It is not yet deployed, no Dockerfile is written. It is just the option of showing it from and Rstudio setup. However, I am convinced that this is the right way to go. We are looking for partners who could potentially make the design more insightful for energy professionals.