2023

Olin Group

IMPACT PROJECT MCSBT

ALVAREZ, EDGARDO

BARCA, FEDERICO

MARINO, ADRIAN

VIEHHOFER, SEBASTIAN

ZABALLA, JOSE LUIS

Executive Summary

Olin Group, a Spanish startup, born in 2011 with the main purpose of consolidating the Telco market in Spain identified after their first acquisitions a pressing need for data cleaning and standardization during their several acquisition processes. This requirement arose as a critical means to enhance operations and customer relationship management. In response, our team developed a solution integrating algorithms, and public APIs.

Applying the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology helped fashion a solution to standardize and verify data. This approach steered the team through understanding the data and devising an exhaustive cleaning plan. The data was characterized by a lack of standardization, frequent typographical errors, and missing details, which necessitated considerable correction. The team disassembled addresses into their respective components, rectified discrepancies using regular expressions, and employed Python's Pandas libraries for data cleaning and manipulation.

For the data modeling phase of CRISP-DM, Fuzzy matching algorithms were explored. These algorithms function by approximating strings and calculating the operations required to make them identical. The best-known Fuzzy Algorithms are: Levenshtein, Jaro-Winkler, and Hamming Distance. The Levenshtein algorithm, due to its capability to manage variations in address strings and compare longer strings, proved most effective.

However, the algorithms alone did not deliver acceptable results, prompting us to incorporate several public APIs. These included Open Street Maps, TomTom, Google Address Validation, Google Places, and Google Geocoding API for matching and validating our addresses. While OSM and TomTom, being free or less expensive, were our initial choices, Google Maps APIs ultimately outperformed them, delivering better results (over 70% accuracy).

Implementation, the final phase of the CRISP-DM model, involved building an authenticated web app using Python and Streamlit. This application enables Olin employees to upload a CSV file with raw data, which our solution then cleans and validates using Python libraries and Google's Geocoding and Places APIs. Additional features were added to the app, such as a form for validating new customer addresses and a geolocation map for pinning the provided addresses.

Significant improvements notwithstanding, there is a clear roadmap for future enhancements. These include developing an API for integrating our service with other applications, transitioning to a microservice architecture for improved maintainability, synchronizing our app with Olin’s internal authentication, and adding a fourth tab on the web app to provide client based KPIs. These KPIs could identify business opportunities by filtering clients by region, socioeconomic power, home size, etc.

In conclusion, our solution, built upon the CRISP-DM framework, not only resolved Olin’s address validation challenge but also laid a foundation for future enhancements. A tool has been effectively created for Olin to address not only their current issue of registering new clients from acquired businesses but also to tackle similar challenges with future acquisitions.