Introduction

This application serves as a pivotal tool employed by our esteemed industry partner, SumUp, for the enrichment of information pertaining to potential leads garnered through their sign-up website. The refined data obtained undergoes utilization in the prediction of potential value that a lead could contribute to SumUp, facilitated by a sophisticated machine learning model. The application is branched into two integral components: the Base Data Collector (BDC) and the Estimated Value Predictor (EVP).

Base Data Collector

General description

The Base Data Collector (BDC) plays a crucial role in enriching the dataset related to potential client leads. The initial dataset solely comprises fundamental lead information, encompassing the lead's first and last name, phone number, email address, and company name. Recognizing the insufficiency of this baseline data for value prediction, the BDC is designed to query diverse data sources, incorporating various Application Programming Interfaces (APIs), to enrich the provided lead data.

Design

The different data sources are organised as steps in the program. Each step extends from a common parent class and implements methods to validate that it can run, perform the data collection from the source and perform clean up and statistics reports for itself. These steps are then collected in a pipeline object sequentially performing the steps to enhance the given data with all chosen data sources. The data sources include:

- inspecting the possible custom domain of the email address.
- retrieving multiple data from the Google Places API.
- analysing the sentiment of Google reviews using GPT.
- inspecting the surrounding areas of the business using the Regional Atlas API.
- searching for company-related data using the OffeneRegisterAPI.
- performing sentiment analysis on reviews using GPT-4 model.

Data storage

All data for this project is stored in CSV files in the client's AWS S3 storage. The files here are split into three buckets. The input data and enhanced data are stored in the events bucket, pre-processed data ready for use of ML models is stored in the features bucket and the used model and inference is stored in the model bucket.

Data preprocessing

Following data enrichment, a pivotal phase in the machine learning pipeline is data preprocessing, an essential process encompassing scaling operations, numerical outlier elimination, and categorical one-hot encoding. This preprocessing stage serves transforms the output originating from the BDC into feature vectors, thereby rendering them amenable for predictive analysis by the machine learning model.

Estimated Value Predictor

The primary objective of the EVP was initially oriented towards forecasting the estimated life-value of leads. However, this objective evolved during the project's progression, primarily influenced by labelling considerations. The machine learning model, integral to the EVP, undergoes training on proprietary historical data sourced from SumUp. The training process aims to discern discriminative features that effectively stratify each class within the Merchant Size taxonomy. It is imperative to note that the confidentiality of the underlying data prohibits its public disclosure.

Merchant Size Prediction

In the context of Merchant Size Prediction, our aim is to leverage pre-trained ML models on new lead data. By applying these models, we intend to predict the potential Merchant Size, thereby assisting SumUp in prioritizing leads and making informed decisions on which leads to contact first. This predictive approach enhances the efficiency of lead management and optimizes resource allocation for maximum impact.

Data field definitions

This section highlights the data field definitions obtained for each lead. The acquisition of such data may derive from the online Lead Form or may be extracted from online sources utilizing APIs.

Field Name	Type	Description	Data source	Dependencies	Example
Last Name	string	Last name of the lead.	Lead data	-	Mustermann
First Name	string	First name of the lead.	Lead data	-	Mustername
Company / Account	string	Company name of the lead	Lead data	-	Mustercompany
Phone	string	Phone number of the lead.	Lead data	-	+49123456789
Email	string	Email of the lead.	Lead data	-	musteremail@example.com
domain	string	The domain of the email is the part that follows the "@" symbol, indicating the organization or service hosting the email address.	processing	Email	@musterexample.de
email_valid	boolean	Checks if the email is valid.	email_validator package	Email	True/False
first_name_in_account	boolean	Checks if first name is written in "Account" input	processing	First Name	True/False
last_name_in_account	boolean	Checks if last name is written in "Account" input	processing	Last Name	True/False
number_formatted	string	Phone number (formatted)	phonenumbers package	Phone	+49123456789
number_country	string	Country derived from phone number	phonenumbers package	Phone	Germany
number_area	string	Area derived from phone number	phonenumbers package	Phone	Erlangen
number_valid	boolean	Indicator weather a phone number is valid	phonenumbers package	Phone	True/False
number_possible	boolean	Indicator weather a phone number is possible	phonenumbers package	Phone	True/False
google_places_place_id	string	Place ID used by Google	Google Places API	Company / Account	-
google_places_business_status	string	Business Status	Google Places API	Company / Account	Operational
google_places_formatted_address	string	Formatted address	Google Places API	Company / Account	Musterstr.1
google_places_name	string	Business Name	Google Places API	Company / Account	Mustername
google_places_user_ratings_total	integer	Total number of ratings	Google Places API	Company / Account	100
google_places_rating	float	Average star rating	Google Places API	Company / Account	4.5
google_places_price_level	float	Price level (1-3)	Google Places API	Company / Account	-

google_places_candidate_count_mail	integer	Number of results from E-Mail based search	Google Places API	Company / Account	1
google_places_candidate_count_phone	integer	Number of results from Phone based search	Google Places API	Company / Account	1
google_places_place_id_matches_phone_search	boolean	Indicator weather phone based and E-Mail based search gave the same result	Google Places API	Company / Account	True/False
google_places_confidence	float	Indicator of confidence in the Google result	processing		0.9
google_places_detailed_website	string	Link to business website	Google Places API	Company / Account	www.musterwebsite.de
google_places_detailed_type	list	Type of business	Google Places API	Company / Account	["florist", "store"]
reviews_sentiment_score	float	Sentiment score between -1 and 1 for the reviews	GPT	Google reviews	0.9
regional_atlas_pop_density	float	Population density	Regional Atlas	google_places_formatted_address	2649.6
regional_atlas_pop_development	float	Population development	Regional Atlas	google_places_formatted_address	-96.5
regional_atlas_age_0	float	Age group	Regional Atlas	google_places_formatted_address	16.3
regional_atlas_age_1	float	Age group	Regional Atlas	google_places_formatted_address	8.2
regional_atlas_age_2	float	Age group	Regional Atlas	google_places_formatted_address	31.1
regional_atlas_age_3	float	Age group	Regional Atlas	google_places_formatted_address	26.8
regional_atlas_age_4	float	Age group	Regional Atlas	google_places_formatted_address	17.7
regional_atlas_pop_avg_age	float	Average population age	Regional Atlas	google_places_formatted_address	42.1
regional_atlas_per_service_sector	float	-	Regional Atlas	google_places_formatted_address	88.4
regional_atlas_per_trade	float	-	Regional Atlas	google_places_formatted_address	28.9
regional_atlas_employment_rate	float	Employment rate	Regional Atlas	google_places_formatted_address	59.9

regional_atlas_unemployment_rate	float	Unemployment rate	Regional Atlas	google_places_formatted_address	6.4
regional_atlas_per_long_term_unemployment	float	Long term unemployment	Regional Atlas	google_places_formatted_address	49.9
regional_atlas_investments_p_employee	float	Investments per employee	Regional Atlas	google_places_formatted_address	6.8
regional_atlas_gross_salary_p_employee	float	Gross salary per employee	Regional Atlas	google_places_formatted_address	63.9
regional_atlas_disp_income_p_inhabitant	float	Income per inhabitant	Regional Atlas	google_places_formatted_address	23703
regional_atlas_tot_income_p_taxpayer	float	Income per taxpayer	Regional Atlas	google_places_formatted_address	45.2
regional_atlas_gdp_p_employee	float	GDP per employee	Regional Atlas	google_places_formatted_address	84983
regional_atlas_gdp_development	float	GDP development	Regional Atlas	google_places_formatted_address	5.2
regional_atlas_gdp_p_inhabitant	float	GDP per inhabitant	Regional Atlas	google_places_formatted_address	61845
regional_atlas_gdp_p_workhours	float	GDP per workhours	Regional Atlas	google_places_formatted_address	60.7
regional_atlas_pop_avg_age_zensus	float	Average population age (from zensus)	Regional Atlas	google_places_formatted_address	41.3
regional_atlas_regional_score	float	Regional score	Regional Atlas	google_places_formatted_address	3761.93
review_avg_grammatical_score	float	Average grammatical score of reviews	processing	google_places_place_id	0.56
review_polarization_type	string	Polarization type of review ratings	processing	google_places_place_id	High-Rating Dominance
review_polarization_score	float	Polarization score of review ratings	processing	google_places_place_id	1
review_highest_rating_ratio	float	Ratio of the highest review ratings	processing	google_places_place_id	1
review_lowest_rating_ratio	float	Ratio of the lowest review ratings	processing	google_places_place_id	0
review_rating_trend	float	Value indicating the trend of ratings	processing	google_places_place_id	0