Introduction

This application is used by our industry partner SumUp to enhance information about possible leads collected at their sign-up website. This enhanced data is then used to predict the possible value that lead could bring to SumUp using a machine learning model. The application is split into two major parts, namely the base data collector (BDC) and the estimated value predictor (EVP).

Base Data Collector

General Description

This component is used to enhance data on possible client leads. The initial data of leads only includes the first and last name of the lead, the phone number of the lead, the lead's email address, and the lead's company name. This data is not sufficient to predict the lead's value, hence this component is supposed to query multiple data sources, including various APIs to enhance the provided lead data.

Desing

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 particular 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
- retrieve multiple data from the Google Places API
- analyze the sentiment of Google reviews using GPT
- inspect the surrounding areas of the business using the Regional Atlas API

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, preprocessed 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.

Estimated Value Predictor

This component is used to predict the possible value a given lead might give to SumUp. The EVP includes an ML model that can perform regression to predict the lead value. The model is trained on confidential data from SumUp and cannot be shared with the general public. It also transforms the output of the BDC into feature vectors that can be used for prediction by the ML model. The used ML model can easily be swapped out to accommodate for a quick testing and exploration phase on which model might perform best. Feature vector creation is not yet possible in this application as the focus for the mid-project release was a stable data collection phase.

Data Field Definitions

This document outlines the data fields obtained for each lead. The data can be sourced from the online *Lead Form* or be retrieved from the internet using APIs. It is currently unfinished, and will be updated once we finalise what data points will be used for the AI model.

The data types selected are on the assumption that we're using the PostgreSQL database.

Data Field Table

Field Name	Data Type	Description	Validation Rules	Data Source	Sample Data (if available)	Name Convention
First Name	text	First name of business owner		Lead Form		first_name
Last Name	text	Last name of business owner		Lead Form		last_name
Email Address	text	Owner's email address (doesn't specify business or personal)		Lead Form		email_address
Telephone Number	varchar	Owner's telephone number (doesn't specify business or personal)	Length dependent on country code	Lead Form		phone_number
					Categories: Keine 0 – 35.000 35.000 - 60.000 60.000 - 100.000	

Field Name	Data Type	Description	Validation Rules	Data Source	Sample Data (if available)	Name Convention
Annual Income from Card Payments	enum	Enumerated income-ranges that indicate how much of the company's income is comprised of card payments		Lead Form	100.000 - 200.000 200.000 - 400.000 400.000 - 600.000 600.000 - 1 Mio. 1 Mio. – 2 Mio. 2 Mio. – 5 Mio. Mehr als 5 Mio.	annual_income
Products of Interest	enum	Enumerated categories indicating SumUp products the owner is interested in		Lead Form	Categories: Keine Alle Kartenterminals Kassensystem Geshäftskonto Andere	products_of_interest
Email Domain	text	Domain of the email address provided by the lead form		Pre-processing		domain

Links to Data Sources:

Lead form: https://www.sumup.com/de-de/kontaktieren-vertriebsteam/
Google Places API: https://developers.google.com/maps/documentation/places/web-service/overview
OpenAI API: https://platform.openai.com/docs/overview
Meta API: https://developers.facebook.com/docs/graph-api/overview