Implementation of a full-stack environment

Cyril Tabet, Walid Anabtawi, Marc Violides,
Sophia Gurria, Youssef Abdel Nasser, &
Miquel Amengual

TABLE OF CONTENTS

1 PROJECT OVERVIEW

2 DATA UNDERSTANDING

3 APPROACH

4 NEXT STEPS...

PROJECT OVERVIEW

Goal, scope, steps, timeline & milestones for the project

1

GOAL & PROJECT SCOPE

GOAL

GOAL

Determine the best restaurants in Madrid within 4 different categories and criteria.

SCOPE

SCOPE

Complete flow of data ingestion, processing, analytics and governance.

METHODOLOGY

DATA INGESTION

200 restaurants in Madrid (50 Spanish, 50 Burgers, 50 Italian, 50 Mexican)

DATA PROCESSING

Merge the 4 JSON datasets, and data cleaning.

ML MODELS - SPARK MLib

Train model that predicts the rating of the restaurant (check 3 types of regression and register results in MLFlow).

DASHBOARD

Create a dashboard mapping locations in Madrid to show TOP 10 restaurants by: rating, review count, zip code, and graph the distribution of > 4.5* rated and > 10 reviews restaurant by categories

DATA UNDERSTANDING

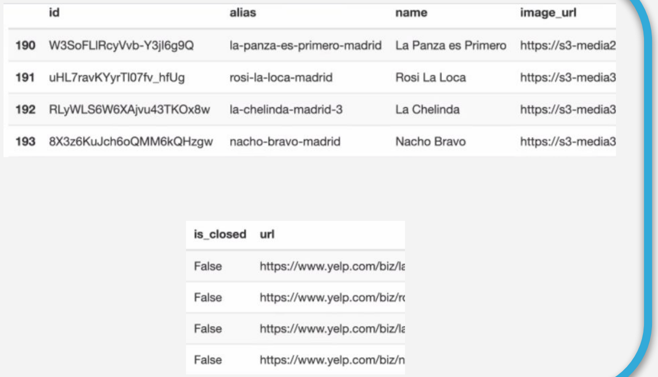
Sample and overview of the data, first insights into the data preprocessing stage

2

DATA OVERVIEW

- **id**: unique for every restaurant (remove)
- **alias**: detailed name of restaurant (remove: same information as name)
- **name**: name of the restaurant (keep)
- **image_url**: url to image of restaurant (remove)
- **is_closed**: all restaurants are available (remove: FALSE for all)
- **url**: url to restaurant (keep for reference)
- **review_count**: change data type to numeric (keep: transform)
- **categories**: create two separate columns, one with an alias and the other with the title (keep: transform)
- **distance**: distance (keep)
- **rating**: average rating of the restaurant (keep)
- **coordinates**: create two separate columns, one for latitude and one for longitude (keep: transform)
- **transactions**: empty list (remove)
- **price**: how expensive the restaurant is (keep: transform)
- **location**: keep the displayed address only (keep)
- **phone**: restaurant's phone number (remove)
- **display_phone**: restaurant's phone number (remove)

SAMPLE OF THE DATA



	id	alias	name	image_url
190	W3SoFLIRcyVvb-Y3jl6g9Q	la-panza-es-primero-madrid	La Panza es Primero	https://s3-media2
191	uHL7ravKYyrTI07fv_hfUg	rosi-la-loca-madrid	Rosi La Loca	https://s3-media3
192	RLyWLS6W6XAjvu43TKOx8w	la-chelinda-madrid-3	La Chelinda	https://s3-media3
193	8X3z6KuJch6oQMM6kQHzgw	nacho-bravo-madrid	Nacho Bravo	https://s3-media3

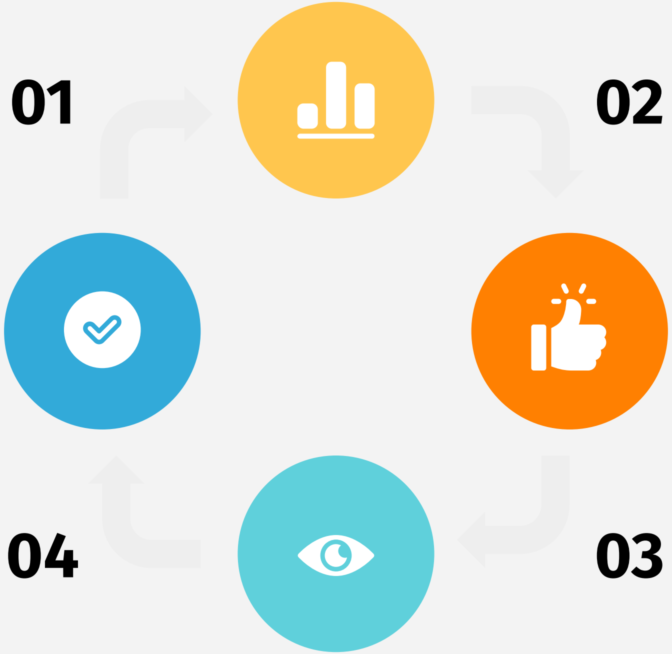
is_closed	url
False	https://www.yelp.com/biz/la
False	https://www.yelp.com/biz/ro
False	https://www.yelp.com/biz/la
False	https://www.yelp.com/biz/n

APPROACH

Data ingestion, processing, analytics,

3

Data Ingestion Strategy



01

Data Sources

Collect data from Yelp using API network protocol

02

Ingest

Data is in JSON format, get a pandas dataframe for each type

03

Combine

Combine the 4 dataframes to get a single one

04

Save as table (bronze)

This will be very beneficial in the long run to get raw data

Data Processing Strategy

Data cleaning

We start by removing variables that are not useful for our analysis (Silver)

Define IVs and DV

With the DV being ratings find features with most correlation

Split data and regression

Train-test split and apply 3 types of regressors in MLflow (Gold)

Get results and store

Get results of regressors compare them and store the best performing one

NEXT STEPS...

PIPELINES

Create the pipelines for the ingestion and preprocessing.

↓

MACHINE LEARNING

Perform Machine Learning on MLFlow and compare the models.

↓

DASHBOARDS

Data visualization for querying with different variables such as rating, review_counts and postal code.