Code.Hub

REGENERATION®

Group Project

A case study for creating a service for Airbnb hosts in Athens using Microsoft Azure

November 2021

Overview

This document describes the scope of the case study project that will be undertaken by the different teams. The project involves building a model that predicts the prices of an Airbnb listing in Athens, using Microsoft Azure. Your team is asked to explore the given data, process them as you see fit and build a ML model using the Azure Machine Learning toolkit. You must then deploy the model and use the cognitive services to identify the top-rated listings, by their reviews.

Table of Contents

1 Introduction

This integrative group project aims at encouraging students to apply the knowledge and experience learned in the class towards a real-life business intelligence system.

You are employed as a Data Scientist at Airbnb, a company that is involved in short-term rentals. Airbnb wants to create a service for hosts with top-rated undervalued listings that will suggest they increase their prices. Your team is tasked with building a POC for this service. You must (a) create, train and deploy a model that will predict the price of a listing, given its attributes, (b) identify undervalued listings with this model and (c) identify the top-rated listings from the reviews.

In terms of data content, you are provided with Airbnb data for the region of Athens, where the POC will take place. The data includes information about the **listings** (neighbourhood, amenities, bedrooms and bathrooms, etc.) and **ratings** for those listings.

Your project will focus on the above three tasks. The steps you should follow regarding the data flow, the modelling process and the resource management are up to you. Your code needs to be well documented and organized so that it can be used in production.

2 Project Scope and Deliverables

The main objective of this project is to make a model that predicts the **price** of a listing, given its attributes.

Several subtasks can be spawned from this objective. The main categories are:

a. **Explore** the given data. See what they describe and gather valuable insights about their properties.
b. **Preprocess** the data so that they can be used for predicting the listing price.
c. **Model** the data through the sklearn estimators or Azure services.

Your **project deliverables** which will support the objectives are identified as deliverables **D01-D03** in the following sections. You will collect all deliverables and submit them as your **project portfolio** work.

3 Overview of Project Work

For running this project, you are advised to frequently meet as a team, and discuss and agree on your implementation plan and actions. This means that you must end up with a clear understanding of

a. the roles and responsibilities of the team members
b. the project requirements
c. the data requirements
d. the way you will run your project
e. the tools you will use for the technical work
f. the tools you will need for the running of your team
g. the deliverables of your work

You will use some of the above decision content in the deliverables outlined next.

4 Data Description

The dataset is provided in two files called *listings.csv* and *ratings.csv*.

The first contains nominal information about the listings, like its neighbourhood, its description, amenities, bedrooms, bathrooms and more. Some are useful, some not so much. These are in a very raw form and need to be processed in order to be used by the model.

The second file contains various ratings for the listings above in free text. These need to be analyzed by Azure's cognitive services. **Pay attention here**: this service is not cheap and the resource management among the different members of the team is considered to be part of the exercise. You could also think creatively, do all ratings need to be processed by Azure? Also it's very important to not do duplicate work!

Note: The dataset is provided by Airbnb on a Creative Commons CC0 1.0 Universal (CC0 1.0) "Public Domain Dedication" license, so it is free to use in this project.

5 Detailed Objectives

5.1 Exploratory Data Analysis

Investigate your data. Try to get a feel of the dataset and decide on the preprocessing steps you may need to perform in the next step. The following questions can guide you through this exploration.

1. How many samples and features does each file have?
2. What are the types of your features?
3. Are there any missing values? If yes, how many and how many rows are affected?
4. How many listings per neighbourhood are there?
5. How many listings per room type are there?
6. How many listings per room number are there?
7. What is the distribution of listings per host? What are the most listings that a single host has?
8. When was the first host registered?
9. What year had the most hosts registered?
10. How many identified hosts are there? What is their percentage over all hosts?
11. What are the top-20 most common amenities provided by the hosts?
12. What is the distribution of price for each room type?
13. How many ratings do I have?
14. Do all listings have ratings?

Additionally we encourage you to perform **your own exploration** on the dataset and identify anything you find interesting.

D01: a notebook containing any EDA you performed.

5.2 Preprocessing

In this step you must bring the dataset in a format understandable by most machine learning algorithms. Some steps you might want to consider:

- Handling missing values in the dataset.
- Encoding categorical features.
- Scaling the features.
- Cleaning erroneous values.
- Handling outliers.
- Feature selection/extraction.

Note: Not all of these steps are mandatory. You should do what you think better suits your needs.

D02: a notebook or well structured script showing the preprocessing steps as you applied them.

5.3 Configure Azure workspace

Create and configure the Azure resources you will need in your project. For example:

- Create a Resource Group for this project
- Create a workspace
- Create a compute target
- Create a datastore
- Upload the data
- Register the dataset
- Log the changes you make to the dataset through versioning.
  Note: you don't need to make a new version after every small change, make sure you have a version for the original and one for the final dataset.

5.4 Modelling

This task is where you must build a model that accurately predicts the price of a given new listing. The metrics you should use for evaluating the results are **Mean Absolute Error**, **Mean Absolute Percentage Error** and any other way you see fit!

For this task you must:

- Use AutoML to perform a quick regression on the processed dataset. You should use the scores of the best AutoML run as a benchmark.
- Examine different models, while performing hyperparameter tuning for each. For comparison purposes you should run these as Experiments in Azure.

5.5 Deploy the model as a service

This task is where you deploy your trained model as an API. The main idea behind this task is that Airbnb wants to help **new** hosts with their onboarding on the platform. For that purpose they want to provide them with an educated guess on the correct price of their listing.

For this task you must:

- Create an API (using FastAPI, or any other web framework of your choice) that will hold your trained model. There should be an endpoint for making predictions based on the features you used to train the model.
- Write a Dockerfile that will create an image with all your code ready for deployment on the Azure Cloud.
- Create all the necessary resources (ACR, ACI) that will allow the deployment of the service.

5.6 Identify top-rated, undervalued listings

Airbnb wants to create a service for its hosts that identifies if their listing is undervalued (compared to other listings), while being highly rated and suggests that they increase their price.

We want you to identify such listings. To do this you must:

- Utilize the model from 5.4 to find undervalued listings.
- Make appropriate use of Azure's cognitive services to find what listings are the top rated.

5.7 Deliver code

For the final handover, we ask you to deliver a well-written, documented and organized code. For this we ask:

- Remove all nonessential code (no print, describe, etc).
- Refactor your code into functions.
- Write documentation and type hints for each function.
- Group related functions into files (e.g. one script for training, one for scoring).
- If a function is used in more than one script it should be imported (not redefined).
- Write a readme file explaining how the code is organized, its dependencies and how it should be run.
- Optionally create unit tests for each script.

D03: the production-level code.

6 Project deliverables

You will need to push the deliverables **D01**-**D03** to the **main** branch of your team's private git repo. This branch should be as clean as possible, containing only the deliverables and no former experimentation!

You will also need to prepare a **presentation** on your project work. This is a presentation that you will give as a team at the end of the course.
Hint: keep your presentation at a high-level and entertaining. Say what you did, but not how you technically did it. The presentation isn't a place to showcase your code.

More details on the content and a more personalized opinion on your presentation can also be provided during the Project Development sessions.