Code.Hub REGENERATION®

ReGeneration Academy on Big Data &
Artificial Intelligence (powered by Microsoft)

Group Project

A case study for predicting the price of an
Airbnb listing in Athens using Microsoft Azure

March 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.

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 in a company that is involved in short-term rentals as a Data Scientist. You are tasked with building a POC of a service that will predict the price of a listing, given its attributes. This service will be marketed by your company to various Airbnb hosts, who will pay to see a suggested price of their listing. They can then use this suggestion to tweak their asking price as they see fit.

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.).

Your own project will focus exclusively on creating the model for predicting the house prices. The steps you should follow regarding the data flow are left up to you. Your data 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-D04** 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 a file called listings.csv and 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.

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

Try to answer the following questions. By doing that you will get a feel of the dataset, as well as a hint of the various preprocessing steps you may need to perform for the following deliverables.

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. What is the distribution of score ratings? Are there lots of reviews scoring < 50?

11. How many identified hosts are there? What is their percentage over all hosts?

12. What are the top-20 most common amenities provided by the hosts?

13. Can you identify the top-10 rated listings? Are they by the same host?

14. Can you identify the top-5 rated locations/neighbourhoods?

15. What is the distribution of price for each room type?

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

D01: a notebook containing the answers to the aforementioned questions

D02: a notebook containing any other 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.

D03: a notebook showing the preprocessing steps as you applied them.

5.3 Configure Azure workspace

- 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 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 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.

D04: the production-level code

6 Project deliverables

You will need to push the deliverables **D01**-**D04** to the **master** 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. Details on the content will also be discussed during the contact sessions.