Machine Learning Engineer Nanodegree – capstone proposal

Predicting apartment prices in Bucharest

Adrian Ciubotaru
December 6th 2021

Domain Background

This project is about the world's largest asset class, estimated to be worth $277 trillion at the end of 2020. I am talking about real estate. Naturally, such a vast and lucrative domain has attracted problem solvers from within the machine learning community, who attempted to answer one of the two main questions for the industry: what's the right value for the price and what's the right value for the rent?

Most academic studies, including this 2020 study, focus on finding quality features to be used with algorithms such as Random Forrest, XGBoost and Linear Regression. There are however, efforts to improve predictions using computer vision, both for satellite images and property photos, and NLP for the text within the listings among other things.

Problem Statement

It is difficult to figure out how to appropriately price an apartment in Bucharest. The largest website only offers the range for a neighborhood and consulting an appraiser is both expensive and time consuming.

Datasets and Inputs

Around 2500 listings scraped from www.imobiliare.ro. All apartments are from Bucharest, from a subset of all existing neighborhoods. The listings include information such as rooms, floor, building age, as well as sections of free text, where other amenities are mentioned. The dataset needs to be cleaned and go through feature selection.

For each of the 2500 listings there are 12 clearly defined features such as area, number of rooms, number of bathrooms, floor, etc. Each listing contains a free text area, which includes things such as heating system, AC availability, materials used for floors & walls and other facilities. The dependent variable, the price, is to be obtained after data cleaning, because it is expressed in 3 different currencies and may or may not include V.A.T.

Solution Statement

I will create a regression model that, based on inputted characteristics of the apartment, will return a reasonable price.

Benchmark Model

Currently the website offers a range for all properties within a neighborhood. The range goes from the minimum listing price of any property in the area all the way to the max listing price of any property in the area. A reasonable benchmark model would be *(Min Price + Max Price)/2*.

This would be a naïve model, which always predicts the same price for each neighborhood.

Evaluation Metrics

For this exercise, I will use RMSE (root-mean-square error) as an evaluation metric.

Project Design

1. Clean data & create features. This includes, but is not limited to:
    - Converting values to appropriate data types (ex: strings to float)
    - Removing clearly erroneous values (ex: an apartment listed for 500$)
    - Adding VAT to the price where this is stated to be necessary
    - Adding features derived from other features (ex: bathroom/room)
    - Dealing with missing data appropriately (ex: if the value for parking spots is not present it could either be 0 or unknown)
2. Use Sagemaker AutoML to determine the best algorithm
3. Expose final model endpoint
4. Create a simple web page with a form where any user can input apartment features in order to receive a price prediction