Capstone Project Proposal

For this part, I would like to build a model to predict AirBnb prices in New York City based on different parameters.

Domain background

This project is derived from the home rent valuation domain. It is a subpart of real estate valuation, applied to short stay pricing.

One paper we may look at is: Neloy, Asif & Haque, H. & Islam, Md. (2019). Ensemble Learning Based Rental Apartment Price Prediction Model by Categorical Features Factoring. 350-356. 10.1145/3318299.3318377. It does not apply perfectly to our project as we look at prices per night on AirBnB but techniques could be transferred to our problem.

Problem statement

Indeed, it is crucial for home owners to correctly price the rent to maximize their profit without losing too many clients (not having the home empty too often). Our goal in this project is to determine the best value a home owner can ask regarding its situation, type, location and services offered at the place.

It is definitely a supervised problem for Machine Learning Algorithms as we want to predict a price and not group places together (it is a regression problem here).

Datasets and inputs

The dataset used is from AirBnB open data in New York City. It has 16 features and a little less than 50,000 lines of data. After considering the features, we could drop some like 'last_review', 'name', 'id' and ask the user to provide the other features to then predict the price. Among these features, we could ask for location, type of home, minimum nights and availability.

The features are the following:

- `id` int64
- `name` object
- `host_id` int64
- `host_name` object
- `neighbourhood_group` object
- `neighbourhood` object
- `latitude` float64
- `longitude` float64
- `room_type` object
- `price` int64
- `minimum_nights` int64
- `number_of_reviews` int64
- `last_review` object
- `reviews_per_month` float64
- `calculated_host_listings_count` int64
- `availability_365` int64

Solution statement

In this project, we will benchmark different models, using simple regression up to neural networks to try to predict the price based on the inputs (XGBoost, Recurrent Neural Network, …).

Benchmark model

Considering the data set presented above, we will simply split it into train and test to be able to measure the accuracy of our model. If we want to be even more precise we will do some cross validation.

The goal is to see how far we can improve from the naivest Linear Regression by using more advanced algorithms.

Evaluation metrics

The metric we will use throughout the project is the absolute difference between the predicted price and the true price of the night at one's place. We may look at squared error to evaluate the difference between the absolute error.

Project design

The project will revolve around a Jupyter Notebook to run the training and testing. At the end of the notebook, there will be a cell to run the test and predict the price of the night.

In the meantime, we will have some graphs and histograms to understand the determinants of the price and understand how to correctly market a stay.

Workflow

- Import and preprocess data (delete unnecessary features, remove lines with zeros or NAs) (mainly some pandas commands like df.drop, df.dropna, …)
- Split training ant testing set (SkLearn train_test_split)
- Build a model (using SkLearn for Linear Regression, SVC and TensorFlow for XGBoost and RNN mainly)
- Train the model (train method of the model)
- Test the model (using SkLearn metrics accuracy and choosing between absolute error or squared error)
- Assess accuracy and provide the necessary changes to improve the model