Capstone Project Proposal

Domain Background

The real estate industry is one of the fastest-growing sectors in developing countries, especially in Vietnam. The price of real estate in Hanoi, the capital of Vietnam, is very expensive, which leads big challenge to possessing real estate in Hanoi. More than 50% of Hanoi's population is people who come from other cities to work and study. Most of them must rent a house or apartment to make accommodations. The rental price is affected by multiple factors, including brokers. Naturally, it becomes increasingly crucial to make use of AI and Bigdata technology to enable renters to self-estimate the price of rental houses.

Problem Statement

The rental price is affected by multiple factors, including brokers. Brokers can push the rental price higher than the actual price to get more commission. Currently, there are not any tools that can help renters estimate the price of rental houses or find the price of similar rental houses. In the capstone project, I build an AI model that can estimate the price of rental houses in Hanoi. In this scope of the capstone project, I limit rental houses to rental apartments.

Datasets and inputs

The dataset is crawled from nha.chotot.com, which is one of the biggest websites about real estate in Vietnam. I only collect news that relevant rental apartments in Hanoi. The dataset has some useful features i.e. area, no bedroom, no bathroom, location, … and detail description which contain a lot of information.

Solution statement

The plan of this project is to build a predictive model based on AI/ML to predict the price of rental apartments in Hanoi. The model is actually a regression model. The raw dataset is rental-apartment news crawled from the nha.chotot.com website. The target value of the dataset is the price of rental apartments on this website. The features to build the model are based on the features of apartments. i.e. area, no bedroom, no living room, and location,…

Benchmark model

It's always a good practice to set a benchmark while building further machine learning models. However, there aren't any AI/ML models built on my dataset. Honestly, I do not know any business metrics to validate the efficiency of AI models. So, I use the RMSE metrics for this project, the lower value the better model. I would try different models to gain as small as possible RMSE value on the test set.

In this project, tree-based models will be selected to build the predictive model due to explainability. Then, hyperparameter tuning will help to find out the optimized parameters used in the final model. In the end, It will be tested with the test set to validate that the model can result in acceptable accuracy.

Evaluation metrics

Since the project is building a regression model, I choose RMSE metrics as the model evaluation metric

Project Design

The project is designed with the following steps:

1. Crawling data. Crawling news that relevant rental apartments in Hanoi. The Scrapy and Selenium tool is used to crawl data. The crawled data is saved in CSV file.
2. Exploratory Data Analysis. EDA is an important step in the machine learning lifecycle. The step explores the dataset which includes how many features, which type of each feature, checking the missing value, visualizing the data distribution, etc. In that way, we can have a better understanding the dataset looks like and how to select the important features for building the model.
3. Data preprocessing. There are some processes to tackle missing values, normalize categorial features, format features to specific types, etc.
4. Feature engineering. The next step will find features that are useful for building the model.
5. Building model + Tunning hyperparameter. The next step is to build a predictive model. Here we choose a decision tree model as a baseline which will help explain the feature importance better so that I can get some insight into what factors affect the price of rental apartments. I also choose a random forest model in order to towards a low RSME value in the training of the model. Then, I tweak hyperparameters to find out the best hyperparameters.
6. Testing model. Evaluate the built model on the test set.
7. Conclusion and further improvement. Review the build process and proposed advanced features that can make the model more stable in production.