**Final Project Report**

# Definition

## 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](http://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 detailed description which contain a lot of information.

The raw dataset contains **16916** rows, **20** columns

Table

Description automatically generated

## 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

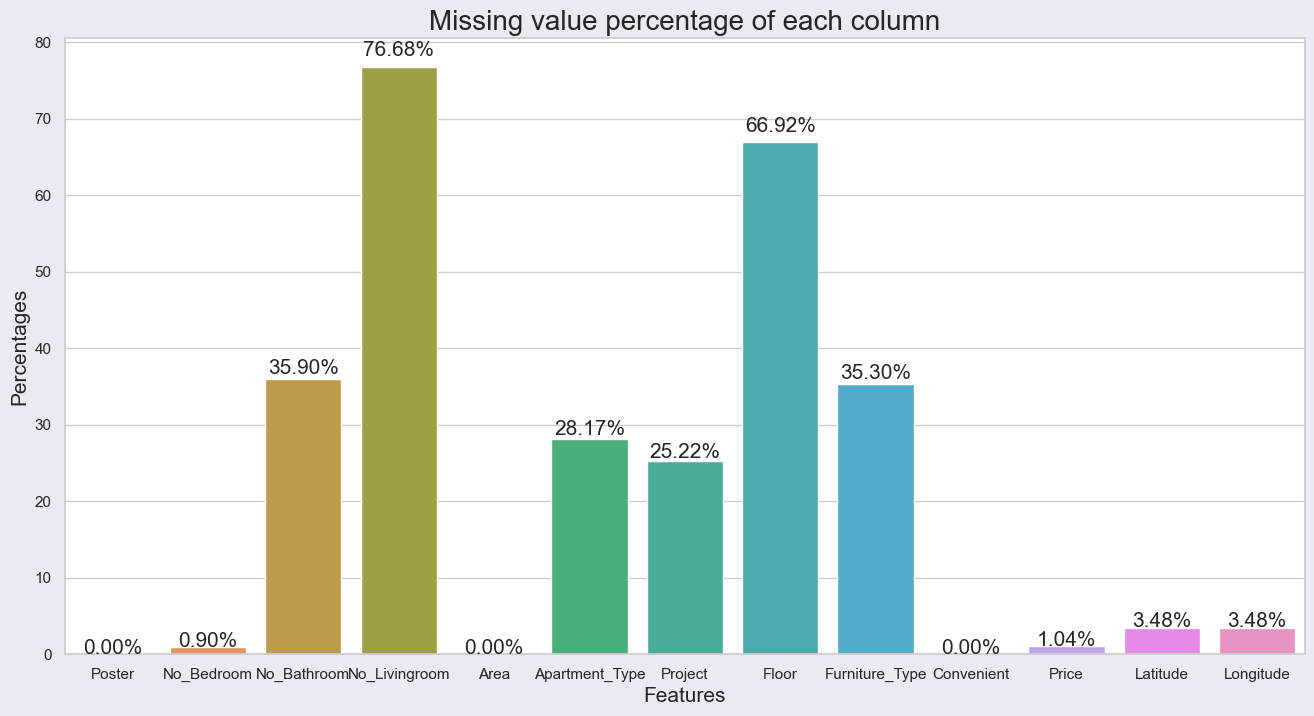
## Project Design

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.

# Exploratory Data Analysis & Data Preprocess

Firstly, I visualize the proportion of missing values in columns.



The above bar chart depicts the proportion of missing values in each column. While there are some columns having a large proportion of missing values such as No\_Livingroom, Floor, No\_Bathroom; Poster, Area, and Convenient feature witness a zero proportion of missing values.

Subsequently, in each feature, I will visualize to get an insight of the feature, then take account into some processes to clean data, and handle missing values, instead of the splitting project into separate steps such as EDA, Data Preprocess,...

### Furniture feature

The furniture feature has 5972 rows missed the value.

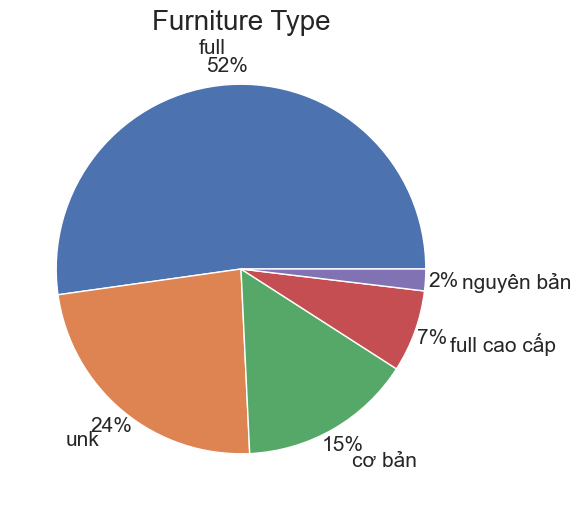
Graphical user interface, application, Excel

Description automatically generatedThe above horizontal bar chart is the frequency value of furniture type. There are 12 types of furniture, which are illustrated in the above figure. However, it has some duplicate furniture types in the dataset i.e. full cao\_cấp == cao\_cấp full, full cơ\_bản == cơ\_bản full, etc. So, I defined a consistent list of furniture types and then normalize them into it.

Text

Description automatically generated

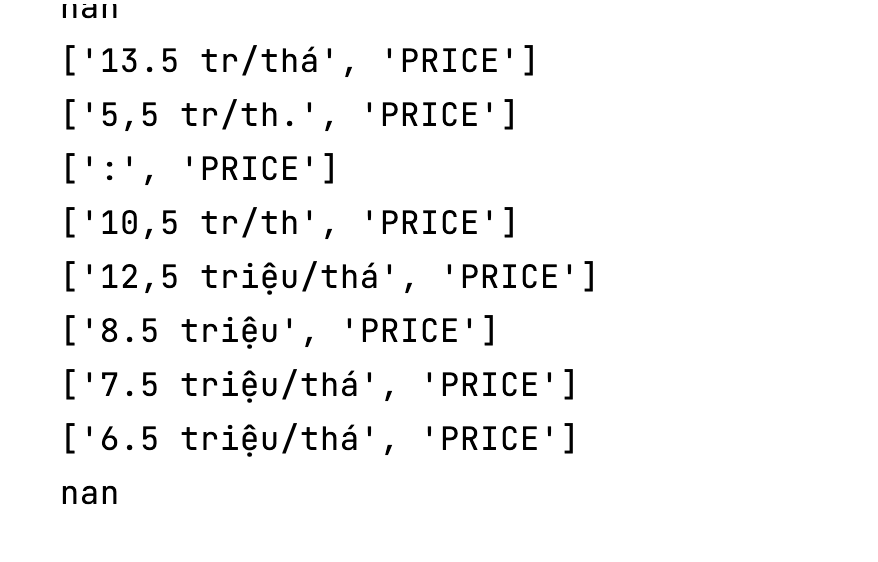
The raw description also contains a lot of information. So, I can extract some information about furniture type from them. Then I fill missing values by the “unk” value.



*The proportion of furniture type after extraction and normalization*

### Area and Price Feature

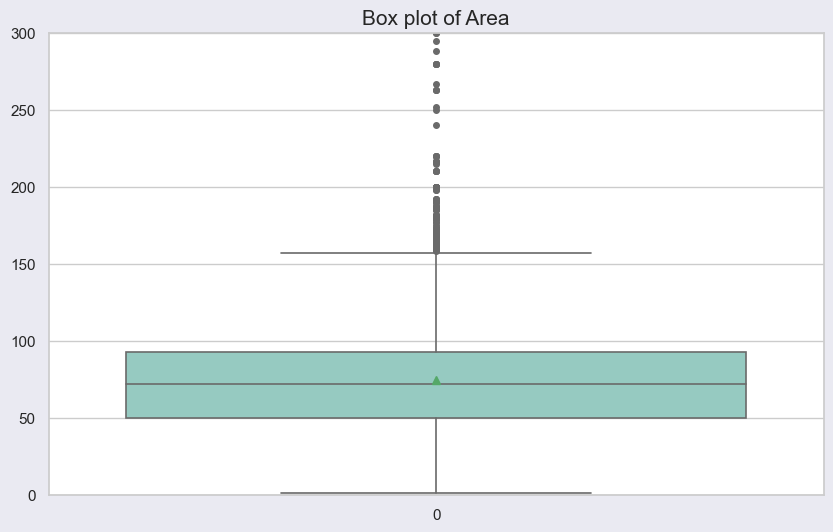
There are a lot of text values in the area and price features.



So, I write a function to extract numbers from text values.



After using the above function to extract numbers, Price and Area features have **192** and **16** missing values, respectively.



The above boxplot is a boxplot of the area feature. It illustrates that there are some outlier values which is larger than 160 m2.

Table

Description automatically generated

All rows have an area that is larger than 160 m2. Apparently, some apartments have an abnormal area i.e. 445 m2

Chart, table, bar chart

Description automatically generated

The boxplot seen above is from the pricing feature. It demonstrates that the pricing will be an outlier if it is more than 20 million Vietnamdong. However, for luxury apartments, the cost of a rental apartment greater than 20 million Vietnamdong is reasonable. I also print out those rows and find out that the price feature's outliner threshold is 50 million Vietnamdong.

### Apartment Type

Graphical user interface, chart

Description automatically generated

The above horizontal bar chart is the frequency value of apartment type. There are many types of apartments. However, some of them are duplicates. So, I defined a consistent list of apartment types and then normalize them into it.

Graphical user interface, text, application

Description automatically generated

*Normalization function*

Chart, pie chart

Description automatically generated

*The proportion of apartment type after normalization*

### No Room Feature

In the dataset, while the No Livingroom feature has **12971** missing values; the No Bathroom and Bedroom features have a smaller proportion of missing values, **6073** and **152**, respectively.

Chart, table, bar chart

Description automatically generated

The boxplot seen above is from the no bedroom feature. It demonstrates that the no bedroom will be an outlier if it is larger than 4.

### Visualize the relationship between different variables

Chart, scatter chart

Description automatically generated

The above scatter figure demonstrates a positive correlation between no bedroom and area.

Chart, scatter chart

Description automatically generated

Apparently, the above scatter illustrates a positive correlation between area and rental price.

Chart

Description automatically generated with medium confidence

*Correlation of different variables*

### Data cleaning

Chart, bar chart

Description automatically generated

*The missing-value proportion of each feature after some above steps*

In this part, I will remove rows that have outliers or nan values in columns such as Area, No Bedroom, Price, Latitude, and Longitude.

Graphical user interface, text, application

Description automatically generated

Finally, the dataset has **15282** rows

# Feature Engineering

Feature engineering is an important step before building an ML model. Based on previous visualization as well as my knowledge, I will drop columns that have a large proportion of missing values or unuseful features.



## Label encoding

Encode categorical values into numerical ones.

Graphical user interface, application

Description automatically generated

## Split train, val, test set

The proportion of train, val, and test sets are 0.7, 0.1, and 0.2, respectively.

Graphical user interface

Description automatically generated

# Model Building

The decision tree and random forest will be selected to build the model due to explainability. Firstly, I experiment with the base decision tree and random forest model without any hyperparameter tuning.

Graphical user interface, application

Description automatically generated

Text

Description automatically generated with low confidence

## Hyperparameters tuning

I define ranges or list values of hyperparameters, then I use random search to find the best set of hyperparameters.

The Random Forest algorithm will be used to build the model.

A picture containing graphical user interface

Description automatically generated

A picture containing graphical user interface

Description automatically generated

# Results

RMSE of models in the validation set

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Hyperparameter tuning** | **RMSE** |
| Decision Tree | No | 3.0623 |
| Random Forest | No | 2.3014 |
| Random Forest | Yes | **2.2805** |

The above table is the experimental results of models on the validation set. It demonstrates that hyperparameter tuning will improve the result slightly and that the random forest produces better results than the decision tree.

The best model gives a result of 2.6557 RMSE on the test set.

Text

Description automatically generated with medium confidence

I think it is an acceptable result.

## Feature Importance

Chart

Description automatically generated

The figure shows that area feature is the most important feature which affects the price of rental apartments in Hanoi.

# Conclusion & Improvements

In the project, I defined the rental apartment problem in Hanoi and built a predictive model that can predict the rental price of an apartment based on several features: Poster, No\_Bedroom, Area, Apartment\_Type, Furniture\_Type, Latitude, Longitude. The best model gives a result of 2.6557 RMSE on the test set. I also find out the area feature is the most important feature that affects the price of rental apartments.

During the AWS Machine Learning Course, I have learnt a lot of knowledge about the machine learning lifecycle as well as several AI-related services of AWS. It will be useful knowledge for my career path.

## Improvements

Apply MLOps concepts to deliver models faster in production and make models more stable as well as immediately re-train models if the model’s accuracy decrease.