House Price Prediction in Phnom Penh

Project Overview

The goal of this project is to build a model that can predict the price of a house given a set of features. The features can include the location of the house, the size of the house, the number of bedrooms and bathrooms, and the condition of the house. The model will be trained on a dataset of historical house prices, and it will be evaluated on a separate dataset of house prices.

Project Steps

1. Data Collection

The first step is to collect a dataset of house prices. The dataset should include the following information:

```
* The price of the house
* The location of the house
* The size of the house
* The number of bedrooms and  bathrooms
* The condition of the house
```

The dataset can be collected from only sources as Khmer24

2. Data Cleaning

Once the dataset has been collected, it needs to be cleaned. This involves removing any errors or inconsistencies in the data. In our dataset we did something like:

- Identify the columns that contain price data. This can be done by looking at the column names or by inspecting the data.
- Remove the $ and , characters from the price data. This can be done using a regular expression or a simple string manipulation function.

- Check for any invalid or missing values in the price data. This can be done by using a data validation function or by visually inspecting the data.
- Replace any invalid or missing values with a default value. The default value can be zero, the mean of the price data, or some other value that is appropriate for the dataset.
- Bedroom and Bathroom data contain more+ , we change them to average value of bedroom or Bathroom is number 3.

3. Feature Engineering

The next step is to engineer features. This involves creating new features from the existing features. For example, the number of bedrooms and bathrooms can be combined to create a new feature called "total bedrooms and bathrooms."

- Identify the categorical features in the dataset. This can be done by looking at the column names or by inspecting the data.
- Calculate the number of observations for each category. This can be done using a pandas DataFrame method such as `value_counts()`.
- Remove categories that are present less than a certain threshold. The threshold can be chosen based on domain knowledge or experimentation.
- Drop the rare categories from the DataFrame. This can be done using a pandas DataFrame method such as `drop()`.
- Scaling features: Features can have different scales, which can make it difficult for machine learning algorithms to learn from the data. You can scale features by normalising them to have a mean of 0 and a standard deviation of 1.
- Feature selection: You can select the most important features by using a feature selection algorithm. Feature selection algorithms can help you to reduce the number of features in your dataset, which can improve the accuracy of your machine learning model

4. Model Selection

There are many different machine learning models that can be used to predict house prices. Some of the most popular models include linear regression, decision trees, and random forests. The best model for a particular dataset will depend on the specific features of the dataset.

5. Model Training

The model is trained on the dataset of historical house prices. The model learns the relationship between the features and the price of the house.

6. Model Evaluation

The model is evaluated on a separate dataset of house prices. The evaluation metric is the root mean squared error (RMSE). The RMSE measures the difference between the predicted price and the actual price.

7. Model Deployment

The model can be deployed to a production environment. This means that the model can be used to predict the price of a house given a set of features.

Project Deliverables

The project deliverables include the following:

- A report that describes the project
- The code for the model
- The dataset of historical house prices
- The dataset of house prices used for evaluation
- The evaluation results

Project Timeline

The project should be completed within 4 weeks.

Project Team

The project team consists of five people. The team members should have experience in data science and machine learning.

I hope this project guideline is helpful! Let me know if you have any other questions.