Housing Price Predictions with Melbourne Housing Dataset

Project Overview

Do you want to solve one or more problems using Machine Learning methods in one project? A project is waiting for you where you will evaluate the results you have obtained using traditional Machine Learning methods. You need to solve the problems in the project using regression and tree-based methods. You should complete this project, where you can learn the methods of examining the data, preparing the data for the model, and exploratory data analysis. You should have a project like this in your portfolio!

Dataset

Melbourne is the capital and largest city of the Australian state of Victoria, and the second-most populous city in both Australia and Oceania. The dataset contains several attributes of the houses in Melbourne along with their prices.

The variables in the data set:

- Suburb
- Address
- Rooms: Number of rooms
- Price: Price in Australian dollars, target variable

- Method: S - property sold; SP - property sold prior; PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; W - withdrawn prior to auction; SA - sold after auction; SS - sold after auction price not disclosed. N/A - price or highest bid not available.
- Type: br - bedroom(s); h - house,cottage,villa, semi,terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential.
- SellerG: Real Estate Agent
- Date: Date sold
- Distance: Distance from CBD in Kilometres
- Regionname: General Region (West, North West, North, North east ...etc)
- Propertycount: Number of properties that exist in the suburb.
- Bedroom2 : Scraped # of Bedrooms (from different source)
- Bathroom: Number of Bathrooms
- Car: Number of carspots
- Landsize: Land Size in Metres
- BuildingArea: Building Size in Metres
- YearBuilt: Year the house was built
- CouncilArea: Governing council for the area
- Lattitude
- Longtitude

Steps of the Project

1. Creating a Google Colaboratory File

- Make sure your project has .ipynb extension.
- Make sure that there are comment lines explaining the details in your project.
- When submitting the project, submit the cells of this .ipynb file so that the cells are run and the results are visible.

2. Importing Required Libraries

- Import the required libraries for the project to the Colab environment.
- Import NumPy, Pandas, Seaborn, and Matplotlib libraries for data analysis
- Import sklearn.model_selection, sklearn.metrics, sklearn.ensemble, sklearn.linear_model, sklearn_tree, sklearn.neighbour libraries and modules for modellling and evaluating performance of the model

3. Project Definition

For this project, we need to load the Melbourne Housing dataset into our project. The quality and amount of data we collect will determine how good our predictive model can be. For this reason, we need to examine the dataset very carefully. We will estimate the price of a house using the Melbourne Housing dataset, which is a real-life example. Before evaluating any cost, we will start by analyzing the data using preprocessing techniques. We will then build our models and measure their performance to complete the project.

4. Gathering and Observing Data

- Load the dataset to the project with the help of read_csv() and observe the first 5 columns
- Find the shape, number of columns and size of the dataset
- Show the information of the dataset, which contains the number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non-null values).

5. Exploratory Data Analysis

- Examine the descriptive statistics of dataset
- The values of some variables are given as objects. At the same time, we observe that there are also categorical values. This might give us trouble when examining the dataset. Therefore, in such cases, we need to define the variables categorically.
- Check for duplicate data. If there are duplicate data, clear them from the dataset.
- Clear outlier data in the dataset. When you examine the dataset, you will observe that the outlier data is generally in the "Landsize" and "Buildingarea" variables.
    - We expect you to use the z-score method when detecting outliers in the dataset!
- Find and remove the missing values on the dataset
    - You can observe from the dataset that the missing values are in the bathroom and car variables. We expect the missing values in the dataset to be filled using the mode method. You can use the code below for this.

HINT:

```
for column in categorical_columns:
 data[column] =
data[column].fillna(data[column].mode().iloc[0])
```

- **Data Visualization:**
    - Build a Histogram to visualize price distribution
    - Draw a pair plot to see the relationship between all numerical variables and the price variable.
    - Draw a correlation matrix by using a heatmap on seaborn
    - Implement Label Encoder and One Hot encoder for categorical variables

6. Model Selection

- Since we are going to make a price estimation, we need to determine our x and y variables correctly.
- Splitting our data into train-test in order to increase the performance of model training
- Train your models using preprocessed data with the models mentioned below

HINT: Please use this dictionary

```
models = {
    'Lasso': {
        'model': Lasso()
    },
    'LinearRegression': {
        'model': LinearRegression()
    },
    'Ridge': {
        'model': Ridge()
    },
    'ElasticNet': {
        'model': ElasticNet()
    },
    'KNeighborsRegressor': {
        'model': KNeighborsRegressor()
    },
    'RandomForestRegressor': {
        'model': RandomForestRegressor()
    },
    'GradientBoostingRegressor': {
        'model': GradientBoostingRegressor()
    },
    'AdaBoostRegressor': {
        'model': AdaBoostRegressor(n_estimators = 5, learning_rate = 1.2, loss = 'exponential', random_state = 2)
    },
```

7. Model Evaluation

- Comparing models in each other
- Choose the best performing model by using evaluation metrics(MAE, MSE, RMSE, R2)

Example:

```
MAE:  245102.450828939
MSE:  111946245183.27266
RMSE:  334583.6893562994
R2:  0.49244357398844063
```

8. Project Delivery

- For the project, you need to prepare a code file with the extension of .ipynb and run all the cells.
- You need to add these files that you have prepared to a GitHub repo and add the link of this repo to the form that is given down below.
- The project will be done as a team. The teams created should be a maximum of 5 people.
- You can send information about your project team via this form.
- Form Link: ##

- **Deadline: ##**