Boston Housing Price Prediction

Created By: Fahiya Iftekhar
Source of Dataset: Kaggle Hosing Price Prediction Dataset

Project objectives

- Predict housing prices in Boston using data provided
- Go through and get acclimatized with the end-to-end of process of machine learning, from data exploration, feature preprocessing and engineering, to model training and validation

Approach

In this project, house prices will be predicted given explanatory variables that cover many aspects of residential houses. The goal of this project is to create a regression model that is able to accurately estimate the price of a house given the features. I conducted some research before diving into the analysis and implementation to get a sense of how to approach a machine learning project. The following three sites wee helpful to structure the project:

- Hacker Earth - Practical Machine Learning Project in Python
- Kaggle - Comprehensive data exploration with Python
- Machine Learning Mastery - Your First Machine Learning Project in Python

Steps

1. **Data Overview:**

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
Id	MSSubCla	MSZoning	LotFronta	LotArea	Street	Alley	LotShape	LandCont	Utilities	LotConfig	LandSlope	Neighborh	Condition	Condition	BldgType	H
1	60	RL	65	8450	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2
2	20	RL	80	9600	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	Norm	1Fam	1
3	60	RL	68	11250	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2
4	70	RL	60	9550	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam	2
5	60	RL	84	14260	Pave	NA	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	Norm	1Fam	2
6	50	RL	85	14115	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	1Fam	1
7	20	RL	75	10084	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1
8	60	RL	NA	10382	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NWAmes	PosN	Norm	1Fam	2
9	50	RM	51	6120	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Artery	Norm	1Fam	1
10	190	RL	50	7420	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Artery	Artery	2fmCon	1
11	20	RL	70	11200	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1
12	60	RL	85	11924	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	2
13	20	RL	NA	12968	Pave	NA	IR2	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1
14	20	RL	91	10652	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1

   - It is very important to start by thinking through the problem before jumping straight to training models and making predictions. If we don't scrutinize whether the data we are provided with make sense or not and if those variables should be included in the models, we run into the danger of blindly creating models that don't make sense.
   - Before even looking at the data provided, we need to think about the nature of problem, consider whether it's something that can be tackled with data and

modelling. If it can be tackled by data, next step would be to think about what kind of data would be useful and relevant, and what factors could be used for prediction.

- If the subject domain is not something we are familiar with, conducting some quick research to get a sense of what factors could be relevant is useful.
- Going through these steps will help develop an intuition of the problem at hand, and provide the framework & context for analyzing and modelling the data.

2. Data exploration and analysis

- After developing the intuition and framework, the next step is to look at what variables are provided in the data and think about how we can incorporate the data into our framework in order to solve the original problem. For this step, instead of looking at the actual data points first, we just go through the variables that are provided, for example, zoning classification, no. of rooms, bathroom types, Lot area type, kitchen, garage types, etc. and think about logically what we would expect their effects to be. This acts as a preliminary feature selection phase to help us rearrange/ transform or filter out variables that are irrelevant right at the start.
- Finally, after going through the variables provided, we dive into the actual data points and conduct univariate & multivariate analysis to understand their distributions and relationships
- We want to make sure to constantly ask questions - go through each variable, evaluate them using the framework we have created in previous steps, consider if the data we have are fit for use and if they are displaying the relationships that we would expect.

3. Features selection, preprocessing, and engineering

- This is the final step before jumping into modelling. Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

4. Model training, tuning & validation, and Testing:

Using a suitable algorithm, we train the model on the given data set. Once the model is trained, we evaluate the model's performance using a suitable error metric. Here, we also look for variable importance, i.e., which variables have proved to be significant in determining the target variable. And, accordingly we can shortlist the best variables and train the model again Finally, we test the model on the unseen data (test data) set.