FLIP ROBO

NAME OF THE PROJECT

HOUSING PROJECT PREDICTION USING MACHINE LEARNING MODEL

Submitted by:

VARSHA VASANT SHINDE

ACKNOWLEDGMENT

We first would like to thank god for giving us the opportunity to be here and make this project. Secondly, we thank our families for their unwavering support and great sacrifices in order to help us reach this moment. We would also like to thank the faculty of Mohd Kashif sir and FLIP ROBO for their work and efforts to give us the best education possible in order to reach our full potential, especially our SME Mr.Mohd Kashif for his guidance and assistance in not only this but Throughout our entire journey

INTRODUCTION

The shelter is one of the three essential requirements of life. It protects an individual and makes him feel safe. Purchasing a house is a dream of every Indian, but sadly for many, it is not attainable. The rising prices of residential properties worry a ton of residents. People pay a fortune to buy their Dream House.

Due to a lack of proper framework, prices have surged and thus the development of negative sentiment of the market. This is a concerning issue for many individuals as if not handled, buying a house will become impossible for many citizens of India. We aim to fill the gap by using machine learning to predict future prices of residential properties, which will help potential buyers to make informed purchasing decisions and buy their dream home at the right Price.

Thus, eliminating surge gains and promoting a healthy market. In India, an inadequate amount of work has been done for valuation in real estate. As a result, sellers use this to their advantage and escalate the prices. Thus, there is a biased procedure to purchase residential property in India as there is no standardized list to aid potential buyers in making a viable buying decision. A typical man cannot contemplate the different market patterns and their impact on the property costs in detail.

Hence, a device that understands these patterns and the impact of different parameters on property costs is required. Different machine learning algorithms can be utilized to foresee future estimates. We require building a model that predicts future housing prices considering precision accuracy and different error metrics.

Analytical Problem Framing

Data collection is the process of gathering information on variables in systematic manner. Data collection is the way toward social events and estimating data on focused factors in a built up framework, which at that point empowers one to address pertinent inquiries and assess results. Before any kind of machine learning analysis, data collection is must. However validity of the Dataset is required otherwise there is no point in analyzing data. We gathered data using various sources like kaggle, magic bricks, 99acres also ready reckoner rates which is a government sites which provides adjusted prices of property.

Data cleaning is the process of cleaning our data. It is process of detecting and removing errors to increase the value of data. There are many garbage values present on the dataset. These values can be removed by checking whether any missing values are present in the dataset or not. We also need to check the validity of the dataset. The values need to be present in the given range. If a variable had many missing values we can drop those values. Data cleaning can be carried out using various wrangling tools. It finds the deficient information and replaces the messy information.

Data pre-processing is done in order to transform the dataset into a clean dataset for better machine learning models. Data pre-processing techniques are applied to data in raw format, which is not feasible for analysis. As in our case, the data is collected from different property websites where property agents entered it, so there are missing values, data in various formats, and incorrect data. We performed data integration to combine the data from various sectors of the capital into an integrated dataset. Data transformation methods were applied to transform the data records to a format that is good for machine learning analysis.

Model/s Development and Evaluation

TRAIN THE DATA

Since the data is divided into two modules: Training set and Test set, we will be firstly training the model. Target variable will be present in the training set.

VALIDATION OF MODEL

Validation is the process of checking whether the applied algorithm tests the given dataset or not. Thus the accuracy of the model should be as high as possible. After applying the algorithm we can check how well our model tests data and also we can apply two or more models to check the model or which tests our dataset best. The model is viewed as input-output transformation for these tests. The validation test compares the outputs from the system that is under consideration to the outputs that are obtained from the model provided that the same input parameters are given to the model. The output values obtained from the model are recorded.

LINEAR REGRESSION

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. In linear regression, there is connection between vector and target variable. After using the parameters which are free, we can anticipate the object variable. upper, lower and normal in our dataset. The upper section comprises of possibilities of the estate that are high in price, same as normal and lower section comprises estimations of centre price range and low range price house.

DECISION TREE

Decision Tree is a supervised learning technique that is used for classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the Features of a dataset, branches represent the decision rules and each leaf node represents the outcome . There are various algorithms in Machine Learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand. The logic behind the decision tree can be easily understood because it shows a tree-like structure. It is simple to understand as it follows The same process which a human follow while making any decision in real-life. It can be very useful for solving decision-related problems. It helps to think about all the possible outcomes for a problem. There is less requirement of data cleaning compared to other algorithms. This algorithm is a class of data mining methods like linear regression. Mainly, Decision trees are uncomplicated but best form of analysis. The decision tree contains nodes which form a tree with the root node, which means it is a tree with a node called root that has no incoming edges which have only outgoing edges. All other nodes can have one incoming edge. A node with outgoing edges is called an intermediate node. All other nodes which have no outgoing edges are called leaves or sometimes also called as decision nodes. According to the input attribute values, each intermediate node is divided or splits into two or more sub trees. In the elementary and most of the cases, each test considers a single attribute, such that from the root node of the tree down to a

RANDOM FOREST

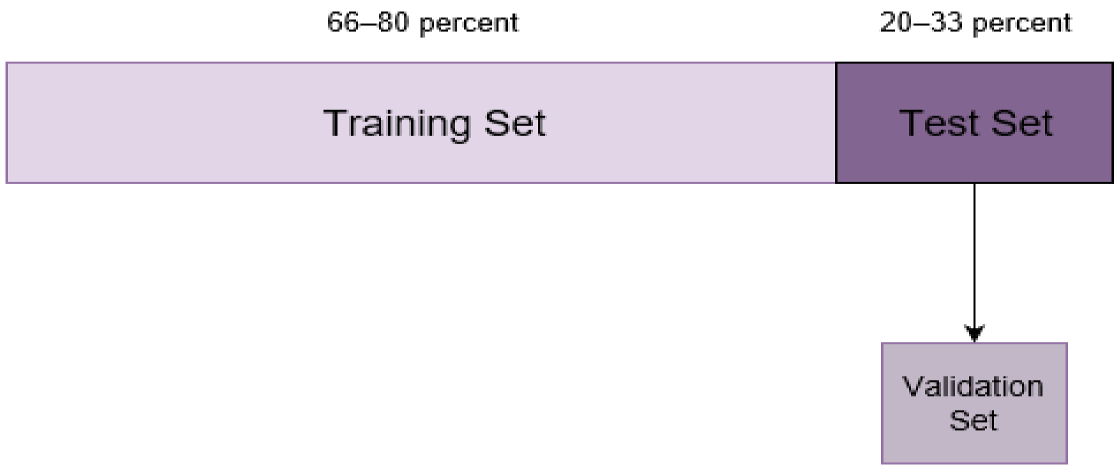
Random forests for regression are formed by growing trees, are determined on a random vector. The output values are algorithmic, and we consider that the training set is independently drawn from the distribution of the random vector X and Y. The random forest predictor is formed by taking the moderate over k of the trees. The number of trees in the forest is created randomly and can go to infinity. It has extremely high accuracy. Scales well. Computationally, the algorithm scales well when new features or samples are added to the dataset. Random forest comparatively is easy to use. The output values are algorithmic and we consider that the training set is independently drawn from the distribution of the random vector X and Y.

Modelling During this stage, a data scientist trains numerous models to define which one of them provides the most accurate predictions.

Model training

After a data scientist has pre-processed the collected data and split it into three subsets, he or she can proceed with model training. This process entails “feeding” the algorithm with training data. An algorithm will process data and output a model that is able to find a target value (attribute) in new data — an answer you want to get with predictive analysis. The purpose of model training is to develop a model.

Two model training styles are most common — supervised and unsupervised learning. The choice of each style depends on whether you must forecast specific attributes or group data objects by similarities.



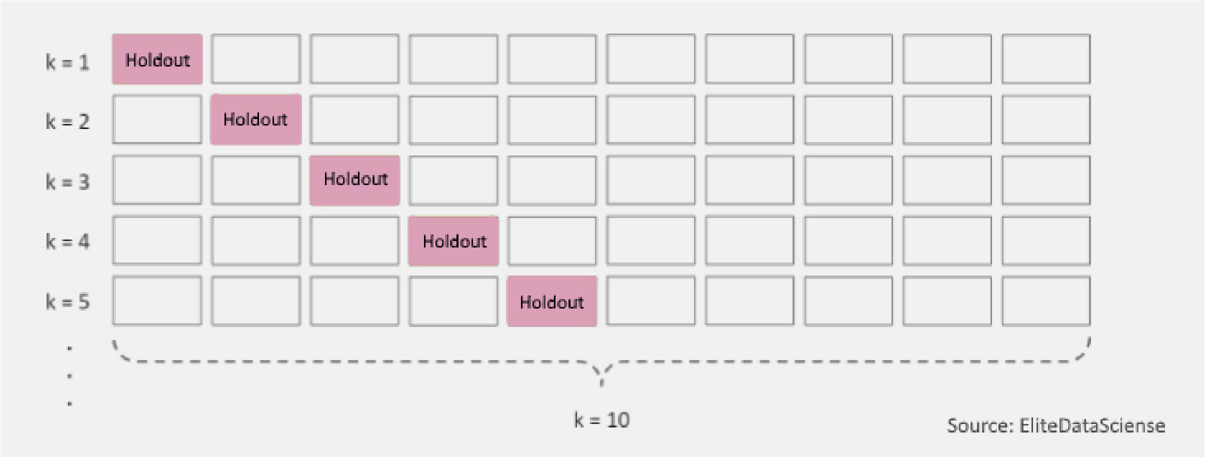
Supervised learning. Supervised learning allows for processing data with target attributes or labelled data. These attributes are mapped in historical data before the training begins. With supervised learning, a data scientist can solve classification and regression problems.

Unsupervised learning. During this training style, an algorithm analyzes unlabeled data. The goal of model training is to find hidden interconnections between data objects and structure objects by similarities or differences. Unsupervised learning aims at solving such problems as clustering, association rule learning, and dimensionality reduction. For instance, it can be applied at the data pre-processing stage to reduce data complexity.

Model evaluation and testing

The goal of this step is to develop the simplest model able to formulate a target value fast and well enough. A data scientist can achieve this goal through model tuning. That’s the optimization of model parameters to achieve an algorithm’s best performance. One of the more efficient methods for model evaluation and tuning is cross-validation.

Cross-validation. Cross-validation is the most commonly used tuning method. It entails splitting a training dataset into ten equal parts (folds). A given model is trained on only nine folds and then tested on the tenth one (the one previously left out). Training continues until every fold is left aside and used for testing. As a result of model performance measure, a specialist calculates a cross-validated score for each set of hyper parameters. A data scientist trains models with different sets of hyper parameters to define which model has the highest prediction accuracy. The cross-validated score indicates average model performance across ten hold-out folds.



Source: EliteDataSciense

Then a data science specialist tests models with a set of hyper parameter values that received the best cross-validated score. There are various error metrics for machine learning tasks.

CONCLUSION

Regardless of a machine learning project's scope, its implementation is a time-consuming process consisting of the same basic steps with a defined set of tasks. The distribution of roles in data science teams is optional and may depend on a project scale, budget, time frame, and a specific problem. For instance, specialists working in small teams usually combine responsibilities of several team members.

Even though a project's key goal development and deployment of a predictive model is achieved, a project continues. Data scientists have to monitor if an accuracy of forecasting results corresponds to performance requirements and improve a model if needed. Make sure you track a performance of deployed model unless you put a dynamic one in production. One of the ways to check if a model is still at its full power is to do the A/B test. Performance metrics used for model evaluation can also become a valuable source of feedback. The faster data becomes outdated within your industry, the more often you should test your model's performance.

Satisfaction of customers by expanding the exactness of their decision and diminishing the danger of putting resources into a home. The sales prices will be calculated with better accuracy and precision. The system will satisfy customers by providing accurate output and preventing the risk of investing in the wrong house. That would make it even easier for the people to select the houses that best suits their budgets.