NAME:SATHEESH M

REG NO:422221104036

COLLEGE CODE :4222

COLLEGE NAME:SURYA GROUP OF INSTITUIONS

Random forest

Random Forest is a supervised learning algorithm that employs ensemble learning method for classification and regression. It runs *n* number of regression trees, and combines them into a single model to make more accurate prediction than one single tree. RF constructs many decision trees at training, and predictions from all trees are combined to make the final prediction. Employing random sampling with replacement (bagging in machine learning terminology), RF helps data scientists reduce the variance associated with those algorithms that have high variance, typically decision trees. Given a training set of feature *X* and output *Y*, bagging repeatedly selects a random sample of the training set for *B* times $(b=1,2,...,\beta)$ $=1,2,...,$ and fits the trees to these samples. For every tree, we obtain a sequence of instances which are randomly sampled replacement from the training set. Each sequence of instances corresponds to a random vector $\emptyset_k\emptyset$ forming a specific tree. Since all the sequences will not be exactly the same, the decision trees constructed from them will also be slightly variant. De Aquino Afonso et al. suggest that the prediction of the *K*-th tree for an input *X* can be represented :

$$h_k(X)=h(X,\emptyset_k),\forall k\in\{1,2,...,K\}h\text{ }=h\text{ },\emptyset\text{ },\forall\in1,2,...,\text{ }(9)$$

where *K* is the number of trees. As a tree splits, each of which randomly selects features to avoid correlations among features. Al paydin points out that a node *S* can be split into two subsets, *S1* and *S2* by selecting a threshold *c* that minimises the difference in the sum of squared errors..

$$SSE=(\sum_{i\in S_i}(v_i-1|S_1|\sum_{i\in S_1}v_i)2+\sum_{i:i\in S_2}(v_i-1|S_2|\sum_{i\in S_2}v_i)2)\text{ }=\sum\in\text{ }-1\text{ }1\sum\in\text{ }1\text{ }2+\sum:\in\text{ }2\text{ }-1\text{ }2\sum\in\text{ }2\text{ }2(10)$$

By following the same decision rules, we can predict any subtree as the mean or median output of instances. Finally, we can obtain the final prediction as an average of each tree's output.

:

$$h(X)=1K\sum_{Ki=1}h_k(X)h\text{ }=1\text{ }\Sigma\text{ }=1\text{ }h\text{ }$$

EVALUATION

In machine learning, it is a general practice to rely on a correlation matrix to decide what features to be incorporated into our models. Table 3 presents the correlations between LRP_i and each feature, and shows that housing floor area has the largest correlation with prices, followed by property age, travelling time to Central District and floor level. In comparison, correlations between each orientation and prices are very close to zero and so they are excluded from our estimation. Then, we present the data visualisation to portray the relationship between property prices and each selected feature in figure 3. All the charts exhibit the expected relationship between the features and the housing prices.

Figure 3. Data visualisation