Hack With CW

Team : 405 Found

Id: 017
Category: Smart City

Title: ML based House Price Prediction System

Description:

Normally house price is decided based on negotiation between seller and buyer. But there is a better way to decide price of a house. There are many factors which affect the price. For example, locality, average price in the locality, availability of transport, availability of health-care nearby, distance from schools, offices etc.. In US, nowadays there is a trend to use algorithms to decide price of a house. There is need of a machine learning algorithm which takes input different decisive parameters and outputs the estimated price.

Plan Of Action

- Data preprocessing.
- Applying ensemble methods (RandomForest, GradientBoostingRegressor, XGBoost) after data cleaning.
- Applying deep learning over processed data.
- Comparing the accuracy of models and deciding the one with best estimates.

Data preprocessing

- Boston House pricing dataset from kaggle was considered for our problem statement.
- Libraries for data visualization like matplotlib, seaborn, sklearn, statsmodel were imported and then methods like info, describe, columns, pairplot, unique were used for exploration and understanding our dataset.
- Dataset was then split into 80% and 20% train set and test set respectively.

Ensemble methods

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.

Three variables giving an R^2 of 0.718 (which is much greater than 0.5 hence statistically significant) : ‘RM’ (avg number of rooms per dwelling), ’LSTAT’ (%lower status of population), ‘PTRATIO’(pupil-teacher ratio per town) were considered as independent variables for predicting our dependent variable ‘price’.

Coefficient of Determination received from various ensemble models:

Gradient Boosting Regressor: 0.8110

Random Forest: 0.8193

XGBoostRegressor: 0.8118

Ensemble methods

We implemented bagging, boosting, and stacking using the implemented algorithms.

After implementation we found that, the **coefficient of determination** was found to decrease.

Thus, we decided to go with averaging.

Prediction = (prediction by GBR + prediction by RF + prediction by XGBR) / 3

- GBR := Gradient Boosting Regressor
- RF := Random Forest
- XGBR := XGBoost Regressor

Deep Learning

MLPRegressor: 0.7852

DL: MAE: 376601.7205038265

MSE: 174715333238.49197

RMSE: 417989.6329318372

- We implemented an mlp regressor from skit-learn library and a deep learning algorithm
- Our deep learning architecture is a sequential mlp

Conclusion :

- We implemented various regression models like linear regression, gradient boosting regressor, random forest, XGBoost, MLP Regressor, and sequential model
- On comparing all the above mentioned models for Boston dataset we found RandomForest giving best predictions with 81.9% coefficient of determination
- By feature extraction we have already taken the best three attributes, thus we would be using them as an input to our model
- We used averaging over gradient boosting regressor, random forest, and XGBoosting regressor, for prediction.

Univ/college: NMIMS/MPSTME

Presented By : Andriya Tiwary
Swapnil Singh