Python B Course
ASSESSMENT 2 – PROJECT

Moaz Mohamed [*1]

Abstract

In this project the Beijing housing dataset have been downloaded and preprocessed for data modelling. the data have been cleaned from url columns, deleting non numerical characters and an advanced clustering algorithm have been used to extract new feature. Three separate data modelling algorithms have been used. after extensive hyper-parameters tuning the model Multi-layer Perceptron regressor preformed the best but with computation trade off.

1. Introduction

The objective of this project is to predict the price of a house in Beijing. The dataset provides ample of information and features that numerous predictive models can be utilized. But before using any predictive model Careful analysis and cleaning should be conducted to ensure optimum performance. In addition to selective preprocessing steps such as implementing standardization of the data to make sure underlying assumptions of certain models are met.

In this project i have used:

- Stochastic gradient descent regressor.
- Passive Aggressive Regressor.
- Multi-layer Perceptron regressor.

The rational for my choice are as follows.

- A predictive model suitable for large datasets.
- A model that can be implemented online.
- Has the ability to adjust and be tunable according to incoming data.

By my selection for the above models. I am hoping to achieve a fast model that can be implemented online. The ability to predict a prices with reasonable margin of error.

2. Materials and Methods

2.1. Data Cleaning

- DeleteD none numerical columns and no important features like (url, ID).
- Some features had dtype as object. astype function was used to convert the features from object to numerical date type.
- Regular expressions methodology was used to extract numerical numbers from the floor feature. And eliminating unidentified characters.
- (to numeric) function was used to convert some features that astype function failed to convert.

2.2. Feature selection and Creation

- Longitude and latitude features alone don't provide significant information for a predictive model. But it is common knowledge that the location of house can provide some insight into its value. Hence by the utilization of HDBSCAN clustering algorithm more than 175 clusters have been identified and added into the dataset as a new computed feature..
- With the addition of clusters feature. A correlation matrix have been computed. Features that had low correlation value have been deleted from the dataset due to the minute significance that they provide to the predictive models.
- Median value for each feature was used instead of deleting the rows with empty values.

2.3. Data Segregation and Baseline Estimators

2.3.1. Data Segregation

Training and testing sets have been created with a random seed of 42. The split ratio is 80:20 accordingly. Gridsearch function was also utilized for hyper-parameter search and its ability to pick the best scoring predictor taking into consideration the cross validation to counter over-fitting and memorization. Which the above mentioned steps provide a good estimate for the predictor's ability to generalize for unseen data

2.3.2. BASELINE ESTIMATORS

A baseline estimator have been developed by calculating the mean of continuous target and using that constant value as a predictor.

In addition to the mean valued predictor. A dummy regressor was also developed by using employing sklearn library. The regression strategy was selected as median. A function "results" have been implemented for the convenience of printing the results of the regressor against the two baseline estimators. Another function was also developed to plot the true price against the model predicted price for evaluation.

2.4. Model Selection and Pipeline Preparation

2.4.1. STOCHASTIC GRADIENT DESCENT REGRESSOR(SGDR)

An SGDR was created using sklearn library. an adaptive learning rate was selected with a 42 as the random seed. early stopping method activated in the parameters of the predictor with a validation set size of 0.2.A pipeline was made for SGDR with standard scaler. Dence SGDR is sensitive to scaling. A standardization method is a must to satisfy the assumptions of the SGDR.

2.4.2. PASSIVE AGGRESSIVE REGRESSOR(PAR)

An PAR was created using sklearn library. early stopping method was selected with a validation set size or 0.2. A dedicated Pipeline was created with the utilization of standard scaler and Principle component analysis (PCA) as a dimensionality reduction method. To reduce the size of the dataset and select a combination of features with the highest variance.

2.4.3. MULTI-LAYER PERCEPTRON REGRESSOR(MLPR)

An MLPR was created using sklearn library. an adaptive learning rate was used and a random seed of 42. A pipeline was implemented with the usage of standard scaler. Hence MLP algorithms do assume that data are standardized.

2.4.4. MODEL TRAINING AND HYPER-PARAMETER TUNING

All the models were trained and tuned with Grid Search function with an appropriate and conservative range of values for parameter tuning. All models were trained and evaluated by grid search function with the default cross validation value of 5.

3. Results and Discussion

The three models or predictor did perform better than the naive and dummy regressor. What is important to note is the performance of MLPR model is significantly better than the other two models As this can be seen in Table 1.

A deeper analysis of the models performance on unseen test data provides more information.

- SGDR performance in Figure 1. Results in mismatch combination of bias and variance. the model seems to be exhibits high bias at some data points and overshoots at other data points indicative of high variance but either behaviours aren't consistent all over the data points. Which could be because the penalty choice of the grid Search function of l2. Elastic-net (combination of l1 and l2) seems to be more appropriate to adjust the predictor.
- PAR performance in Figure 2. Unlike the mismatch combination of SGDR predictor. PAR seems to be high in bias and low and variance in relative to the other models. This is either could be from bad hyper-parameters choice or due to the nature of how the model works. Hence its developed to be an online learning model. Adjusted and developed for contentious training and fast inference time. As the last prediction of this models would be the most accurate for an online setting. Due to the aggressive nature of the predictor at every iteration it change the weight vector so the last iteration is classified correctly and is outside the margin. And the passive nature is the stream of data is classified correctly there is no change to the weight vector. This consist shift between passive and aggressive seems not so suitable for a closed problem or non online problem. But its most suitable for online inference like regression or classification of twitter feed. This is when PA predictor preforms best.
- MLPR performance in Figure 3. Seems to provide the best combination or trade off between bias and variance between all other models. Of course neural networks are notorious for its robustness and effectiveness in finding patterns assuming good hyper-parameters tuning is preformed. But there are a known downside to MLP models. which is inference time. According the deployment requirements and the specification of the targeted computing unit. MLP models in general are at a disadvantage due to the trade off between the ability to increase its capabilities by adding more layers and enhance its abilities to find pattern and generalize better on unseen data but at the cost of computation power needed and the tight inference time. Hence a careful consideration of the deployment requirement is advisable.

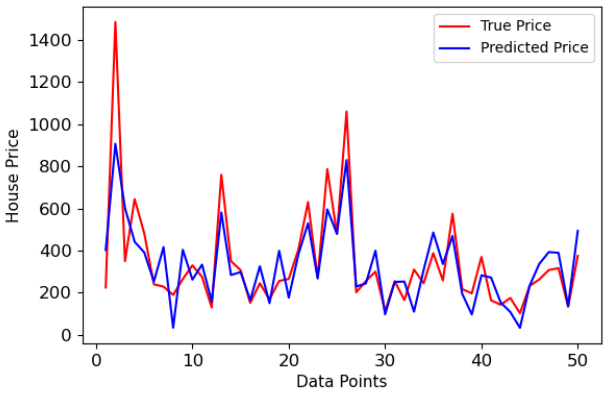

Figure 1. SGDR algorithm performance on unseen test data.

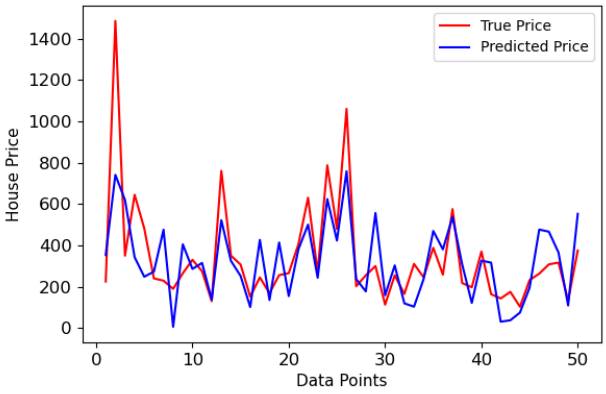

Figure 2. PAR algorithm performance on unseen test data.

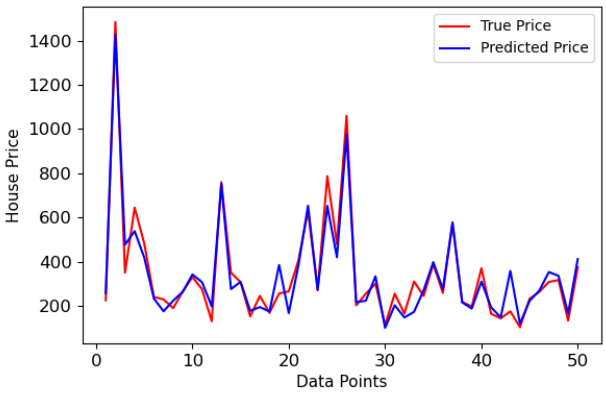

Figure 3. MLPR algorithm performance on unseen test data.

4. Conclusions

The selected dataset have been cleaned and processed. feature creation by adding new feature using HDBSCAN clustering algorithm. in addition to feature selection according to the correlation matrix. Multiple Pipeline have been implemented with appropriate steps and function to satisfy the predictor or model assumptions about the dataset. A dimensionality reduction technique also have been implemented. Three separate models with three custom hyper-parameters tuning with cross validation in the loop. All three models have been properly evaluated using a separate test set. MLPR performed the best but with computation requirement limitation. SGDR came second but with a trade of in bias and variance mismatch. PAR came last but its geared towards online regression. All advantages and disadvantages for all three algorithms have been presented.

Table 1. RMSE of models Against dummy and naive regressors

MODEL	TEST	TRAIN
SGDR	132.378	129.425
PAR	159.215	156.362
MLPR	80.625	79.096
NAIVE BASELINE	231.712	N/A
DUMMY BASELINE	231.712	N/A