Comparison based regression techniques for predicting housing prices

Aashish Deshpande Nachiket Patwardhan

1 Introduction

1.1 Task and Motivation:

During this globalization era, many people are interested in Investments.There are several objects that are used for investments such as gold, property, and stocks etc.

In recent times, Property Investment has increased remarkably. Housing price trends directly indicate the current economic situation. There are various factors which has impact on house prices such as Area, no. of bedrooms, no. of bathrooms etc. Even the Locality has a great impact on the House Prices. It also depends on whether House is in center of city, all the resources are accessible or not, whether it is close to highways or not.

Therefore, we chose to study the house pricing predictions problem, which will help us dig deep into variables and provide a model that will accurately predict the House Prices.

We are going to implement 3 regression techniques and find the one that suits better for this problem and data set.

A comparative study will help us gain necessary insights and help us improve and maintain our model.

1.2 Highlights

Visualize Data:

housing.hist(bins=50, figsize=(15,15))

plt.show()

The insights that we can gather from this dataset in the form histogram:

- households — Almost all of the districts have around 200-600 houses.
- housing median age — At 38 and 18 are two peaks.
- latitude — Major houses are located around latitude of 34 and 36.
- longitude — Major houses are located around longitude of -122 and -116.
- median house value — The median house value is above 510,000.

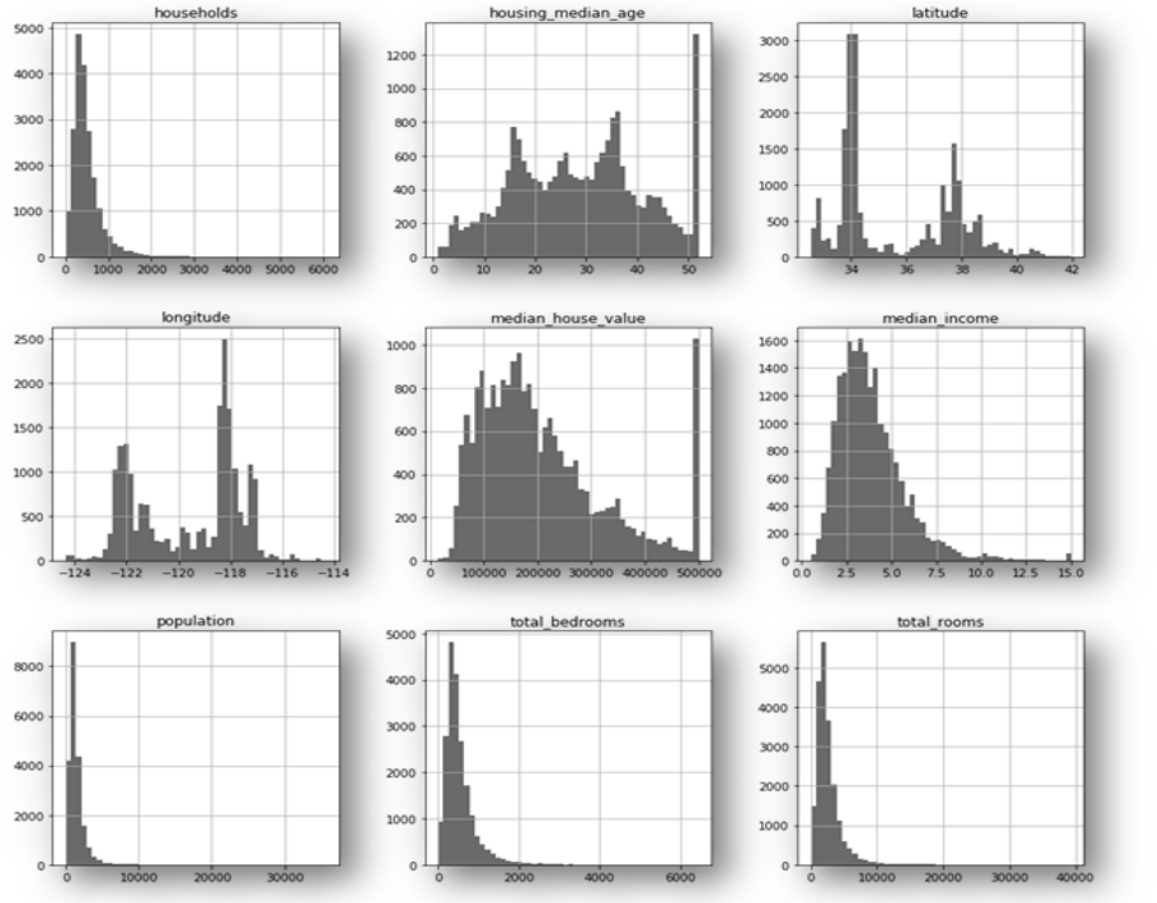

Figure 1: Data Visualization

- median income — There is no income above 15 so some capping has been done. most people have income between 2–5.
- population — Almost all of the districts have population below 3000.
- bedrooms — Most Districts have between 400–700 bedrooms.
- total rooms — Almost all of the districts have around 3200 rooms

Pre-process the data:

- Median income is a very significant feature to find out housing prices.
- We observe that the median income data is continuous.
- We are going to make it discrete.

- Now based on income categories, we'll split our entire data into:
- 80% for training our model
- 20% for testing our model

Data cleaning:

- set missing values to some value like zero, median or mean.
- Here we have set it to median.

1.3 Architecture

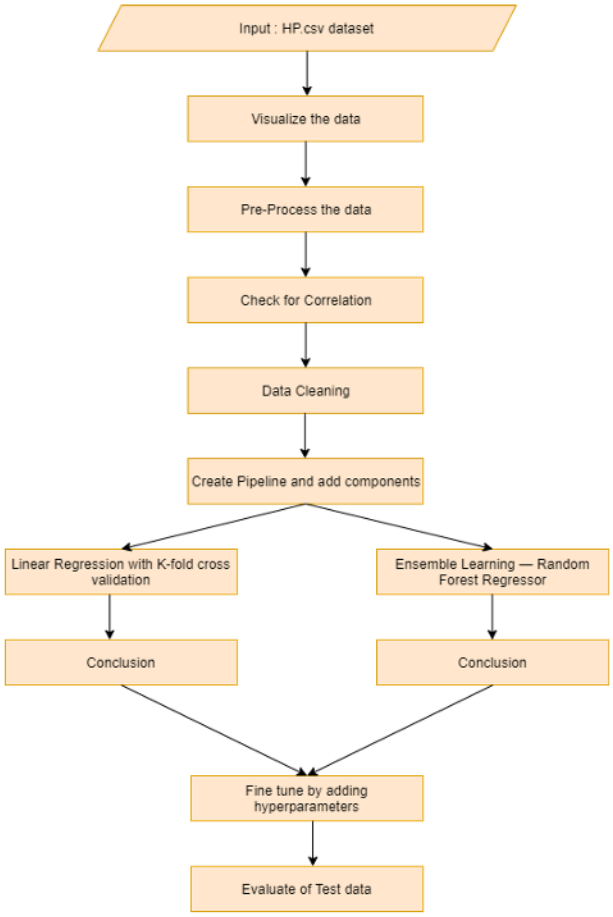

Figure 2: The architecture

2 Methods/Case Study

Methods: We have used regression techniques to predict housing prices via supervised learning models and compared them for accuracy.

2.1 Method 1: Linear Regression with K-fold cross validation

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog).

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

Main Advantage of using Cross Validation is Robustness.

Below diagram shows the cross validation for k = 4.

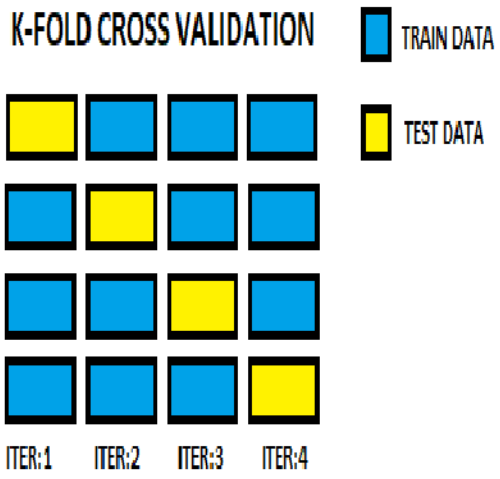

Figure 3: K fold CV

2.2 Method 2: Decision Tree Regressor with K-fold cross validation

Decision trees advantages are that it can be used for both classification and regression problems.

A decision tree is arriving at an estimate by asking a series of questions in True/False form to the data, each question narrowing our possible values until the model gets confident enough to make a single prediction. The order of the question as well as their content are being determined by the model.

Decision tree with k fold cross validation where k=10, divides our training data in 10 folds.Then trains and evaluates our decision tree model 10 times and results in an array of 10 scores.

Then we analyze the scores.

During training, the model is fitted with any historical data that is relevant to the problem domain and the true value we want the model to learn to predict. The model learns any relationships between the data and the target variable.

After the training phase, the decision tree produces a tree , by calculating the best questions as well as their order to ask in order to make the most accurate estimates possible. When we want to make a prediction the same data format should be provided to the model to make a prediction. The prediction will be an estimate based on the trained data.

2.3 Method 3:Ensemble Learning — Random Forest Regressor

Random forest is an ensemble of decision tree algorithms.

Ensemble methods are learning models that achieve performance by combining the opinions of multiple learners.

In doing so, you can often get away with using much simpler learners and still achieve great performance. Moreover, ensembles are inherently parallel, which can make them much more efficient at training and test time, if you have access to multiple processors.

In bagging, a number of decision trees are created where each tree is created from a different bootstrap sample of the training dataset. A bootstrap sample is a sample of the training dataset where a sample may appear more than once in the sample, referred to as sampling with replacement.

Bagging is an effective ensemble algorithm as each decision tree is fit on a slightly different training dataset, and in turn, has a slightly different performance. Unlike normal decision tree models, such as classification and regression trees (CART), trees used in the ensemble are unpruned, making them slightly overfit to the training dataset. This is desirable as it helps to make each tree more different and have less correlated predictions or prediction errors.

Predictions from the trees are averaged across all decision trees resulting in better performance than any single tree in the model.

3 Results and Discussion

After building the data, we decided to train the data on 3 different models. Firstly, we decided to train on Linear regression with k fold cross validation with k =10. To evaluate the goodness of this model, we calculated the root mean squared error between predictions and actuals. It measures the standard deviation of the errors the system makes in its predictions.

Mean: 69052.463

Standard Deviation: 2731.674

For example, an RMSE of 52,000 means that about 67 % of the system's predictions fall within 49,500 of the actual value and about 93 of the predictions fall within 100,000 of the actual value. As our median housing values lie within 120,000 and 265,000, a typical prediction error of $ 69,052 with ±2731 is not very satisfying. This can be classified as an example of a model underfitting the training data.

Comparison table:

	Linear Regression with K-fold cross validation	Decision Tree Regressor with K-fold cross validation	Ensemble Learning — Random Forest Regressor
Mean	69052.463	71487.986	52607.956
Standard Deviation	2731.674	2493.962	1734.313

Next, we trained the data on Decision Tree Regressor with k fold cross validation with k =10. This model performed a little worse than the Linear Regression model as we had an error of 71487 with ±2493.

Finally, we decided to train our model on Ensemble Learning- Random forest Regressor. Ensemble learning turned out to be the best model among 3 models and also very promising.

4 Conclusion

In this project we did a comparative study on supervised learning models based on regression techniques. With the aim of predictive analysis we understood the important tools needed to

develop a powerful ML model we found that; Ensemble learning approach used by random forest builds the "forest", which is an ensemble of decision trees, usually trained with the "bagging" method which uses the idea that a combination of learning models increases the overall result instead of using a standalone singular learning model.

References

- Breiman, "Random Forests", Machine Learning, 45(1), 5-32, 2001
- Hands On Machine Learning with SciKit Learn and Tensor Flow - Aurélien Géron
- http://ciml.info/
- https://scikit-learn.org/
- https://gdcoder.com/decision-tree-regressor-explained-in-depth