Project Proposal

CS 5233 – Artificial Intelligence

Group #2: Jesus Perez (rxy049), Gideon Koech (cre347), Georgios Filippos Bogdos (oql161)

Topic:

Housing Price Prediction

Problem Statement:

Estimating housing prices is an essential part of real estate. The project aims at building an effective price prediction model for house values in Texas. The model should learn from the data and be able to predict the housing price in a district, given all the other metrics. In order to achieve this, the group will identify important home price attributes which feed the model's predictive power.

Related Study

Housing price prediction research has been done by different researchers. A study done by Prasad and Ravindra compares the performance of several regression models, including linear regression, polynomial regression, and decision tree regression, for predicting house prices in Ames, Iowa. Singh et al., 2017 used several machine learning techniques, which included decision trees, artificial neural networks, and support vector regression, to predict housing prices in Delhi, India. The performance of these techniques was compared and also evaluated the importance of different features of the prediction models. A comprehensive review of the different machine learning techniques used for predicting housing prices, including regression models, artificial neural networks, and support vector machines was conducted by Soni and Goyal, 2018. The challenges faced and future possible research directions was also discussed.

Methods:

After conducting research, the group is choosing to explore the following methods to predict housing Price:

- *Multiple Linear Regression:* Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of linear regression. The variable that we want to predict is known as the dependent variable, while the variables we use to predict the value of the dependent variable are known as independent or explanatory variables. Housing price is affected by multiple factors and features of a specific house. According to the previous research, some analysts have proposed several variables that significantly influence the overall housing price. The most important and common question is whether there is a statistical relationship between an explanatory variable and a response variable. To solve this problem, a typical way is to apply regression analysis to model and quantify this statistical relationship. Many types of regression are adopted in scientific research, depending on the feature and type of given data.

- *Random Forest Regressor:* Random Forests models require minimal data preparation. It is easily hand categorical, numerical and binary features without scaling or normalization required. Random Forests models can help us in performing implicit feature selections as they provide good indicators of the important features. These models are immune to outliers, which is present in our data, and they completely ignore statistical issues because unlike other machine learning models which perform much better after being normalized.

- *Gradient Boost Regression:* This is a form of an ensemble learning that combines various 'weak' models to produce a powerful model with improved prediction capabilities. The model is built by adding new decision trees to the model, with each new tree correcting the mistakes made by the prior trees. Each decision tree that follows is constructed on top of the initial tree to anticipate the errors of the ones that came before it. The ultimate prediction is then created by combining the predictions of all the trees.

Experimental Setup:

The group will be using the standard basic python data science libraries (numpy, pandas, matplotlib, seaborn, etc...) to work with datasets and metrics.

For datasets, we will be working with a California Housing Prices dataset provided from a public domain dataset hosting website. Other datasets could be incorporated along the way, for example Texas Housing Prices. We found the California Housing Prices to be the most useful and consistent, giving us the most options of metrics to load, so we're starting with that dataset. The data contains information from the 1990 California census and pertains to the houses found in a given California district.

The group is using (but not limited to) the following metrics from the dataset:

- Longitude
- Latitude
- Housing Median Age
- Total Rooms
- Total Bedrooms
- Population
- Ocean Proximity

References

Prasad, K. Venkata, and K. Ravindra. "A comparative study of regression models for prediction of house prices in Ames, Iowa." Journal of Construction Engineering and Management 138, no. 4 (2012): 499-508.

Singh, R. K., and S. Singh. "Predicting housing prices with machine learning techniques." International Journal of Computer Applications 173, no. 7 (2017): 1-7.

Soni, N. K., and N. P. Goyal. "Housing price prediction using machine learning techniques: A review." arXiv preprint arXiv:1809.06354 (2018).

Yadaw, S. (2020, June 10). Predicting housing prices using a Scikit-learn's random forest model. Medium. Retrieved March 6, 2023, from https://towardsdatascience.com/predicting-housing-prices-using-a-scikit-learns-random-forest-model-e736b59d56c5