Assignment 1

August 2021

Problem Statement

The focus of this Capstone project is to use the Ames, Iowa Dataset regarding homes in Ames, Iowa from 2006 to 2010. The data focuses on Sales price and quantitative and qualitative data about the houses and their corresponding sales prices. The data mainly cover descriptive elements of the homes themselves like garage size, bathroom numbers and size, and year sold. The focus of this capstone project is to predict future sales price of homes based upon this data that has been provided as well as other economic and census data to give an accurate prediction of sales price in todays market. To do this, I will bring in economic and census data, combine these new datasets with the original Ames, Iowa dataset and test and build a few different machine learning models to find the best predictor of current sales prices of Ames, Iowa homes.

Assessment

In order to best predict current and future home prices in Ames, Iowa it will be imperative to build a highly successful model that predicts the sales prices of the homes in the original dataset. To do this I will have to, as in all machine learning projects, spilt the data into a test and train datasets. Before this, I will conduct exploratory data analysis to find the quality of they data, decide what to do with null or mossing values, or data fields with little relevant data. I will also conduct correlation matrices to help with feature selection and focus in on the features that are most relevant to the sales price. There are also a large number of categorical variable with which I will use one hot encoding and dummy variables to turn into quantitative data. I plan on using algorithms mainly based around regression as this seems to be the best choice for the outcome we are seeking. I plan to experiment with ridge, lasso, elasticnet, svr, gradient boosting and light gbm regression models. I will decide which as the best output and choose this model as the predictor model. As mentioned previously, I am introducing new data to the case problem around economic statistics and census population data to 1) see if they have an impact on the prediction of sales price, and 2) use them as aids in predicting future sales price. I am first going to

run experiments and algorithms on the original data set. Once the best algorithm has been established, I will add the other data which includes unemployment percentage, population, wages, household income, and property values. This data will seek to clarify and provide a better picture of the landscape in Ames, Iowa and act as benefits in creating a more accurate model. Once I have set on a model with the new blended data, I will use said predictor and current values to predict current home sales prices in Ames, Iowa.

Assessment

I am currently on track to complete this project on time. I have gathered my data and identified my sources. I have the original Ames, Iowa dataset as well as data for the economic and census features that I have chosen to use. I have also conducted the EDA, performed one hot encoding to turn the categorical variables into numeric ones, and split the data into test-train datasets. I have also performed a correlation matrix to identify which features are best to help predict sales price. I am currently working on testing the algorithms on the data set. Once I have completed these tests I will introduce the new data and perform more testing on the blended dataset. Lastly, I will use these findings to predict current sales price with the best model from my testing.

Conclusion

This capstone project will build on the famous Ames, Iowa dataset and Kaggle challenge to predict sales prices and expand it to further include economic and census data to better predict current and future home prices.