California Housing Price Prediction

Objective

The aim of this project is to build a model of housing prices so that we can predict the median house values in California from the provided dataset using linear regression.

Acquire/Import data

To acquire data we can use pandas which will import our xlsx dataset in a pandas dataframe. We can subsequently explore the dataset using df.head() to get the first 5 rows so that we get familiar with the columns and values. df.shape() will give us a touple where the first item is the number of items in the dataframe and the second is the number of features.

Data munging

From looking at the dataframe we notice that there are some missing values we can check that by using df.isnull().sum() this will give us a list of feature with the amount of null values present, in our dataframe the total_bedroom feature has 207 missing values, to clean the data we will need to replace the missing value with the mean of total_bedroom, to do so we can use SimpleImputer which is a transformer for completing missing values.

Now that we have solved the null values on total_bedroom we can move onto ocean_proximity which has 5 different types of values and these values are categorical values which means that we need to convert them into dummy variables. First we can simplify the amount of values from 5 to 4 because in the dataset there are only 5 ISLANDS under ocean_proximity and ISLANDS and NEAR_BAY is similar, we can group ISLANDS with NEAR_BAY which has 2290 items, we can do so by substituting the ISLANDS value with NEAR_BAY, once this is done we can proceed to convert categorical values in numerical values, we need to do so because linear regression accepts only numerical values, we can do so by using pd.get_dummies and then remove ocean_proximity from the dataframe and concat the result of pd.get_dummies.

Exploration

To explore the data we can visualise the dataframe using a histogram for each feature this will show how are the values distributed for each feature in the dataset. From this visualisation we can notice that most of the histograms are right skewed which is expected because there are less big houses.

Model Building

We'll be using Linear Regression as the model of choice because this a supervised problem where we have the Y which in this case is median_house_value (dependent variable). Before starting training our model we need to split the data in train (80%) and test (20%). Test data this will allow us to check that our model is predicting the correct value and we can also check how well the model is doing. Once we've split the data then we need to fit and transform the data with StandardScaler which will transform the data is such a way that it has mean of 0 and standard deviation of 1.

Now we are ready to fit the model with the train data. From fitting we can then calculate the score of the model on the test data which in this case is about 63% which is not bad.

To go even further and check how is our model doing we can use Root Mean Squared Error (RMSE) which will measure the average of error squares, which basically calculates standard deviation of the residuals: how far is the regression line from the datapoints on average. We can use sklearn's mean_quarred_error to do so.

Model visualisation

To visualise the model we can use pandas inbuilt function plot and plot a scatter graph, on a dataframe that will include the Y_test sent and the predictions on the X_test data. The graph that comes out suggests that there is a linear relationship between the predicted values and the test values.