Luther Nicholaus
CIS 607
Project

Title of the Project

Predicting the house price of unit area.

Problem Statement.

The problem statement is hinged on the prediction of house price of unit area in the Sindian District, New Taipei City, Taiwan. The goal is to establish a solid model for predicting the house price of unit area based on independent variables. The independent variables not only describe the location and age of the house, but also describe the market and neighborhood qualities surrounding the given house.

Dataset Link

The dataset used for analysis is titled "Real estate valuation data set". It is sourced from the UCI Machine Learning Repository. It is a sample of 414 houses collected from Sindian District, New Taipei City, Taiwan between years 2012 and 2013. The dataset details houses' transactional dates, age, distance to the nearest MRT station, and number of convenience store(s) in the living circle on foot.

Link: https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set

Three Research Questions/Hypothesis and the Technique to Test Each

Primary Research:

How does each variable affect or weigh towards the prediction of house prices in Taipei City?

Correlation analysis will be deployed to explore the strength of potential linear relationships between independent variables and the dependent variable. The underlying goal is to determine variables that strongly influence house prices in the given market. Eventually, using regression analysis, a regression equation to predict house prices based on select independent variables will be derived and tested.

Research Hypothesis 2:

Is there enough evidence to fail to reject Chen's (2015) claim that the average house price of unit area is $20,667?

T-test statistical analysis will be used to test the claim applicability towards the population. The test will confirm if there is a significance variation between claimed mean and sample mean. A t-

test is to be applied because the population standard deviation is unknown, and the sample size is above 30.

Research Question 3:

Which houses are fairly-priced and overpriced?

Classification analysis will be used to predict two value category/classes of houses (cheap or expensive) based on specific attributes. A decision tree algorithm is likely to be used in the classification model. Using the model, one can establish or predict house that are underpriced or overpriced.

Data Analysis and Results

The software used for analysis is Python. The modules imported are pandas, seaborn, numpy, sklearn, and scipy. These modules will assist with data collection, data cleaning, and data analysis. Part of data cleaning involves dropping all rows with missing values. Besides, to execute the primary analysis and research question 2, a new column that holds location groups is created by clustering the longitude and latitude values of each house. The outcome is two location clusters. After the latter, the columns, longitude and latitude, are dropped from the data frame.

```
#import libraries
import pandas as pd
import seaborn as sns
import seaborn as sb
import numpy as np
import sklearn.cluster as cluster
import sklearn.metrics as metrics
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.cluster import KMeans
from matplotlib import pyplot as plt
from sklearn import tree
from scipy import stats as st
#%matplotlib inline

#import dataset
df = pd.read_csv('C:/Users/18166/OneDrive/Documents/Real estate valuation data set.csv')
df = df.drop(['No'], axis=1)

#drop missing values
df = df.dropna()

kmeans = KMeans(n_clusters=2, init='k-means++')
```

```
kmeans.fit(df[['X5 latitude','X6 longitude']])
df['Location Clusters'] = kmeans.labels_
```

Primary Research. A linear regression model is built and tested to predict house prices based on select independent variables. The dependent variables (y) is the house price attribute. The independent variables (x) are house age (M1), distance to the nearest MRT station (M2), number of convenience stores (M3), and location clusters (M4). The established set of variables are split into two groups: 80% for training the model and 20% for testing the model.

```
#Extract x and y variables
x = df[['X2 house age', 'X3 distance to the nearest MRT station', 'X4 number of convenience stores',
'Location Clusters']]
y = df[['Y house price of unit area']]

#spliting data into training and test dataset
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=1)

linreg = LinearRegression()
a = linreg.fit(x_train, y_train)

r_sq = a.score(x_train, y_train)
print('coefficient of determination:', r_sq)

print(linreg.intercept_)
print(linreg.coef_)
score = linreg.score(x_test,y_test)
print(score)
```

The results of the analysis are a coefficient of determination of 51.6%. The model training score is 84.3%. The prediction model is y = 42.39 - 0.23(M1) - 0.003(M2) + 1.33(M3) -5.93(M4).

Research Hypothesis 2. The t-test, a statistical analysis method, is applied using the price column to determine where to reject or fail to reject Chen's (2015) claim that the average house price in the city is $20,667. Since the claim contains the statement of equity, it will be the null hypothesis. Therefore, the alternative hypothesis will be: the average house price of unit area is not $20,667. The absence of population standard deviation commands the use of the t test. The level of significance for this test is 0.05. The p value of the two-sided test is significantly less than the significance level of 0.05, as shown below.

```
Test_Results = st.ttest_1samp(df['Price'], 20.667)
print(Test_Results)
```

Research Question 3. The determine if a house is fairly-priced, a new column titled “Fair” is extracted using the linear regression from the primary research and house price values. The linear regression is used to predict the value of the homes based on established variables in the research. Using the latter values as home values and the price values as the actual home prices, the determination of whether a house is overvalued or fairly-valued is made. Overvalued homes are those whose actual prices exceed their respective values. These homes have the value of 0 in the new “Fair” column, while fairly-priced homes have the value of 1.

```
y_pred = linreg.predict(x)
y_pred = pd.DataFrame(y_pred, columns = ['Value'])
df = pd.concat((df, y_pred), axis=1)

df = df.drop(['X5 latitude', 'X6 longitude'], axis=1)
df = df.rename(columns ={"Y house price of unit area":"Price", "X1 transaction date":"Transacttion
Date", "X2 house age": "House Age", "X3 distance to the nearest MRT station": "Distance to
nearest MRT", "X4 number of convenience stores":"Number of Stores"})

df['price-value'] = df['Price'] - df['Value']
df['Overpriced/Fairy-priced'] = np.where(df['price-value'] > 0, 'Over', 'Fair')

#Convert the non-numeric columnn to numeric
ValueDummy = pd.get_dummies(df['Overpriced/Fairy-priced'])
df = pd.concat((df, ValueDummy), axis=1)
df = df.drop(['Overpriced/Fairy-priced', 'Over'], axis=1)
df =df.rename(columns = {"Fair":"Overpriced/Fairy-priced"})
# NOTE: 1 - "Fairly-Priced & 0 - Over-Priced
```

The x and y data values are defined using a clean data frame. The x array values are from all columns, except the “Fair” column. Recall the goal is to create a model to predict if home is fairly-priced or overpriced. Based on the latter analytical objective, the y attribute holds values from the “Fair” column. The x and y set of values are split into 80%-to-20% chunks. The 80% chunk is used to train the classification model, while the 20% chunk is used to test and score the model. Lastly, the classifier is tested and scored for prediction using the train sample set of x and y values.

```
#Extract x and y variables
x = df[['House Age', 'Distance to nearest MRT', 'Number of Stores', 'Price', 'Location Clusters']]
y = df[['Overpriced/Fairy-priced']]

#Spliting the dataset into 80% Training and 20% Test
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2,random_state=0)

#Build and Train Decision Tree Classifier
```

dt_clf =tree.DecisionTreeClassifier(max_depth=5)

dt_clf.fit(x_train, y_train)

#Prediction
score = dt_clf.score(x_test, y_test)
print(score)

```
In [2]: runfile('C:/Users/18166/.spyder-py3/untitled3.py', wdir='C:/Users/18166/.spyder-py3')
coefficient of determination: 0.51615662879011
[42.39065044]
[[-2.30396265e-01 -3.89122724e-03  1.33164626e+00 -5.93689572e+00]]
0.715399582111498
0.8433734939759037
Ttest_1sampResult(statistic=25.88995607129923, pvalue=1.645179367270192e-88)
```

Figure: Analysis Results

Conclusions

Primary Research. From the R square score, the input variables explain about 51.6% of the variation in the dependent variable. With a p-value less than the 0.5 level of significance, the model to predict home prices is significant. The accuracy of the model to predict prices using the test x and y set of values is 71.53%. This means the model is a moderate predictor of house prices in the given market.

The prediction model **is y = 42.39 - 0.23(M1) - 0.003(M2) + 1.33(M3) -5.93(M4)**

Research Hypothesis 2. Since all the two-sided p values from five random samples are considerably less than the significance level of 0.05, the test rejects the null hypothesis. At a 0.05% level of significance, there is sufficient evidence to reject the claim that the average house price of unit area is $20,667.

Research Question 3. The accuracy of the model to predict if a house is fairly-priced or not is 84.34%. This score is based on the mean accuracy of the given test labels and data. With an 84.34% score, the decision tree classifier appears to be a solid predictor of if a house is fairly-priced or not.

How the Results Provide Solution for the Selected Problem

The regression model can help individuals and businesses predict house prices in the given market. The model's price predictions are dependent on the location of a given house, house's age, distance to the nearest MRT station, and the number of convenience stores. The classification model with an outstanding score of 84.3% aids decision making regarding whether a home is overpriced or not. Aforementioned decisions range from investment decisions to purchasing decisions.

References

Chen, Y.-L. (2016, July 28). The factors and implications of rising housing prices in Taiwan. *Brookings*. Retrieved September 10, 2021, from https://www.brookings.edu/opinions/the-factors-and-implications-of-rising-housing-prices-in-taiwan/

Yeh, I. C. (2018). Real estate valuation data set Data Set. *UCI Machine Learning Repository*. Retrieved from https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set