MSA Phase1 Data Science Pathway

Udit Sharma

email:- uditsharma16@gmail.com

Contents

1 Executive Summary

The following report contains the steps of data collection, cleaning and model fitting that has been done for MSA phase-1 under data science pathway. The aim of this assignment is to build the model that can be used to predict capital value of the property. The dataset provided in this assignment contains information of property and koordinates API and deprivation index dataset by Otago university is added to it. The resulting dataset contains 16 features. All the features define the different features of the property like land area, number of bedroom and bathrooms and how many people live there and what is the location of the property. The dataset has only a total of 3 NaN values. After initial analysis it can be seen that number of bedroom, bathrooms and deprivation score are the correlated factors to 'CV'. After fitting the model with default python parameters it can be seen that highest R^2 score is for linear regression was .51.

2 Initial Data Analysis and Cleaning

2.1 Dataset building

The dataset for this assignment is built from combination of 3 datasets. The first dataset has been provided in the assignment which contains the house prices and different parameters related to it. The second dataset is accessed using the 'koordinates' API which contains the census population data for the year 2018 available at https://koordinates.com/services/query/v1/vector.json. The API requires api key, layer id, latitude and longitude as the required parameters. The data has been joined by using the 'SA1' columns. 'SA1' stands for statistical area 1 which is a division of an area. The third dataset is available at the Otago university site https://www.otago.ac.nz/wellington/departments/publichealth/research/hirp/otago020194.html which contains the depriviation index and is also joined with other dataset using 'SA1' column.

2.2 Initial Data Analysis

	Bedrooms	Bathroom	Land area	CV	Latitude	Longitude	SA1	0-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60+ years	Suburbs	CU18	ep2018_S	NZDep2018
count	1051	1051	1051	1051	1051	1051	1051	1051	1051	1051	1051	1051	1051	1051	1051	1051	1051
mean	3.777355	2.075167	856.9895	1387521	-36.8937	174.7993	7006319	47.549	28.96384	27.04282	24.12559	22.6156	29.36061	93.97146	179.9144	986.5033	5.063749
std	1.169412	0.992861	1588.156	1182939	0.1301	0.119538	2591.262	24.6922	21.03744	17.97541	10.94277	10.21058	21.80503	50.50923	71.05928	94.28725	2.913471
min	1	1	40	270000	-37.265	174.3171	7001130	0	0	0	0	0	0	0	3	849	1
25%	3	1	321	780000	-36.9506	174.7208	7004416	33	15	15	18	15	18	53	138	918	2
50%	4	2	571	1080000	-36.8931	174.7986	7006325	45	24	24	24	21	27	92	174	959	5
75%	4	3	825	1600000	-36.8558	174.8809	7008384	57	36	33	30	27	36	138	210	1031	8
max	17	8	22240	18000000	-36.1777	175.4924	7011028	201	270	177	114	90	483	189	789	1380	10

Figure 1: Description of Data

2.3 Data Types

It is important to determine what data types exist in the dataset. This is important as most python machine learning models specially the ones being used in this assignment that are contained in the 'sklearn' python library only can work with integer, float or binary data. So it is important to clean that data by either dropping non-numerical data or converting it using encoding techniques.The following methods are used to clean the data in this case:

1. **Land Area**: Land area contains numerical value but due to presence of m^2. This is solved by extracting the numerical part from it.
2. **Address**: The address column is dropped. The address is a unique column and thus contains different values. It can either be label encoded that would create a unique continuous id values. It is better to drop that column as we already have the suburbs, latitude and longitude values which contain the location information so address is dropped.
3. **Suburbs**: The suburbs column is label encoded. The suburb can be divided in a categorical data by label encoding. As suburbs might affect the price of the house thus it is not dropped.

```
Bedrooms             int64
Bathrooms          float64
Address             object
Land area           object
CV                   int64
Latitude           float64
Longitude          float64
SA1                  int64
0-19 years           int64
20-29 years          int64
30-39 years          int64
40-49 years          int64
50-59 years          int64
60+ years            int64
Suburbs             object
CU18                 int64
NZDep2018_Score    float64
NZDep2018          float64
dtype: object
```

Figure 2: Datatypes of the final data

2.4 Null Values Replacement/Imputation

It is also important to clean the data and replace the null values. The null values need to be handled as the python machine learning can't work with them. This can be solved by either dropping the null values or replacing them. Replacement can be done by mean values/ median value or an imputation method can be used to replace them. In this assignment only 2 NaN values exist in 'Bathrooms' column and 1 exist in 'Suburbs' column. For this assignment K-nearest neighbour imputer is used to replace the null values as bathroom column is categorical and knn-imputer works well with categorical data.

```
Bedrooms            0
Bathrooms           2
Address             0
Land area           0
CV                  0
Latitude            0
Longitude           0
SA1                 0
0-19 years          0
20-29 years         0
30-39 years         0
40-49 years         0
50-59 years         0
60+ years           0
Suburbs             0
CU18                0
NZDep2018_Score     0
NZDep2018           0
dtype: int64
```

Figure 3: Showing Null Values

3 Analysis of correlations and patterns in the data

Analysing correlations and pattern in data is important as it can help improve accuracy of the model. This can help in selecting useful features and removing noisy features.

3.1 Correlation Analysis

Correlation analysis show that the number of bedroom, bathrooms,deprivation index and number of people between 30-39 years are the most related parameter to capital value of the property and 4 most important parameters that might help in building a high accuracy model. The pair plot has is included in the code.

```
Bedrooms              0.224836
Bathrooms             0.375237
Land area             0.112220
CV                    1.000000
Latitude              0.120609
Longitude             0.018317
SA1                  -0.109920
0-19 years           -0.156010
20-29 years          -0.182458
30-39 years          -0.214312
40-49 years          -0.044565
50-59 years           0.131029
60+ years             0.083532
Suburbs               0.062061
CU18                 -0.128474
NZDep2018_Score      -0.344391
NZDep2018            -0.378120
Name: CV, dtype: float64
```

Figure 4: Correlation of Columns with capital value 'CV'

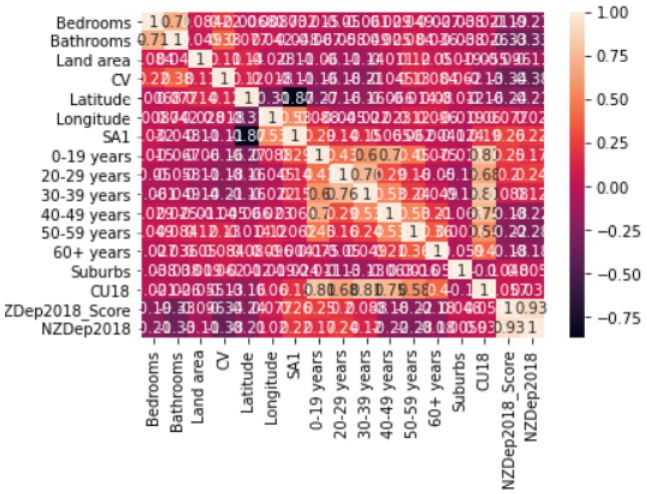

Figure 5: Heat map for correlation between different features

3.2 Feature Selection

For initial model fitting address is dropped for the reason explained in 'section 2.2'. After fitting the parameter with remaining parameters the accuracy of linear regression model was only around '.2'. Thus from correlation analysis the 4 major dominating parameters bedroom, bathroom, deprivation index and people between 30-39 is selected. In case this didn't offer good model accuracy the features can be selected by other feature selection algorithms like f_regression and SelectKBest etc. Land area can also be considered for model fitting as it make sense that it should be an important parameter even though it doesn't have very high correlation so selecting it would be purely intuition based. Similarly as suburbs can effect the price it can also be selected.

4 Building the Model

For this section 4 different regression models have been considered Linear, Mlp, decision tree and KNN regressor. In case there is need for parameter optimization it can be done using pipeline and feature selection to improve accuracy further.

4.1 Model building

Regression models are considered for this assignment. For supervised learning classification and regression models both can be used but it depends on the data. As 'CV' is continuous thus a regression model is considered. First the results by choosing high correlation features is taken and then f_regression is used from feature selection to build the model. Following is the table that contains the results for different regression models.

Regression Model	R^2 **Score**
Linear Regression	0.4021
Multilayer perceptron Regression	0.3704
Decision Tree Regression	-0.9131
K neighbors Regression	0.2525
Linear Regression with feature and logarithmic 'CV'	0.5105
Decision Tree Regression with feature selection and logarithmic 'CV'	-0.4079
K neighbors Regression with feature selection and logarithmic 'CV'	0.2358

4.2 Analysis

It can be seen that linear regression with feature selection f_regression(for 6 features) produces best R^2 score. Multilayer preceptron regression performs close by whereas decision tree regression fails to perform at all. The score can be increased by using better data cleaning and feature selection. It was also noticed during experimentation that while running the train_test split multiple

times the accuracy of the model varied from 20-50 percent. This is specially strange as such high variations might mean that some parts of data are not appropriate and needs to be removed.

5 Conclusions

In this assignment I have done analysis of data by looking at different correlations and patterns in the property value dataset. It was found that deprivation index, number of bedrooms and bathrooms are the most correlated features to capital value of the property. Also linear regression model with f_regression feature selection and logarithmic('CV') as the target provided highest R^2 score of .51.