**Housing Price Prediction.**

**Introduction.**

This is a report of a regression model built from a housing features data set, aiming to identify the importance of each feature and predict the housing price. The provided train data set contains 3000 entries and 12 columns and has the price as the predictor variable. The data is clean, as there is no missing data, hence data cleaning is not required. The test data contains 999 entries and 12 columns and requires no cleaning as there are also no missing values.

The purpose is to run a regression model to determine each feature's importance in the model, predict the house price in the test data set, showing the estimated value.

Python tool was used in the Google Colab environment for the analyses of the data.

**Data Analyses.**

For the analyses, the first step is to import the needed packages and the data sets into the google colab environment and read the data to the data frame to enable manipulation. The packages imported are ; pandas, numpy, matplotlib.pyplot, and seaborn.

The CSV file containing the data sets, both the train and the test, was uploaded and read on google colab and after that, the head was confirmed to ensure that the data had been uploaded and can be assessed. The data info was viewed to know the data type and check for missing values. It shows that the data quantity is 3000 rows, with 12 columns, all integers.

**Features Correlation:**

The features were then correlated with the given price, and it was found that among all the features, the room has the highest correlation with the price with a value of 65%, followed by furnished which is 45%, and wood floor, which has 43 %.

See the table below.

A screenshot of a computer

Description automatically generated with low confidence

The other packages needed for the task are then uploaded. They include train\_test\_split, cross\_val\_score, GridSearchCV, StandardScaler, PolynomialFeature, Lasso, LogisticRegression, LinearRegression, Pipeline, mean\_squared\_error, mean\_absolute\_error, r2\_score, RandomForestRegressor, SVR and SelectFromModel.

X and y data frames for training and testing the models were then created. Price was dropped from X and used as y, for both training and testing data, since it’s the predictor variable.

The X and y data frames were then split into training and testing sets to the ratio of 70: 30 using the train test split Function, and a random stat parameter of 20 was set to ensure that the same split is performed each time the same code is run.

**Lasso Algorithm**: The lasso regression model module is then used to create a lasso model and fit it into the training data to enable it to make predictions.

Predictions. The model’s predictions are as follows.

CV: 0.9999999988142042

R2\_score (train): 0.9999999988357038

R2\_score (test): 0.9999999988823783

RMSE: 0.0758122186667998

All the scores are good. Cross-validation and R2 for trains and testing are very high; close to 1 showing that the model fits and performs very well. The RMSE shows that the error rate is low, at 0.075, this indicates that the model is good.

**Importance of each feature.**

**Answer number 1.**

**C**oefficient Feature importance is used to determine the effect/importance of each feature on the target variable(housing price/ value). The importance of each feature in predicting the target variable is determined by how large the coefficient value is.

The Lasso Regression model was used to predict the importance of each feature and the result is shown as follows.

A picture containing text, screenshot, font, line

Description automatically generated

Furnished has the highest importance score, followed by wood floor, then by solar power, rooms, club access, backyard, kitchen, security, green paint, bathroom, and French door has the least importance.

**Linear Regression Algorithm**: Linear regression was also used for comparison’s sake.

The linear regression module was fit to the training data, and the prediction is as follows.

CV: 1.0

R2\_score (train): 1.0

R2\_score (test): 1.0

RMSE: 2.899388340230146e-12

The R2 score for both the test and train data set gave a perfect score of 1,0. This means that the model is perfectly fit. However, it could sometimes mean that the model is overfitted, hence cross-validation was used to check for that, and it gave a perfect score of 1.0, indicating that the model is not overfitted.

And the feature importance was predicted as plotted below.

A picture containing text, screenshot, font, number

Description automatically generated

Also shows that furnished has the highest importance, then followed by woodfloor, Solar power, rooms, club access, backyard, kitchen, security, green paint, bathroom, and the French door has the least feature.

Both algorithms’ models give good scores and hence are good for prediction. The logistic regression model is preferred to the Lasso Regression model because it gives a perfect figure of 1.0 on Cross-validation and R2, which indicates that it performed optimally, as against the Lasso model which gives a score of 0.99.

**The predicted housing prices and values.**

**Answer number 2.**

The housing prices were also predicted and below are the estimated values in the data frame of the test dataset (first twenty rows). The full score is on a CSV file attached to the assignment.

A screenshot of a computer

Description automatically generated with low confidence

The RMSE of the estimated values was compared with the prices in the test data to see how well the model performs on unseen data. An RMSE of 13 is gotten which shows that the estimated values predicted by the model are about 13 pounds different from the original price. The r2\_score is 0.99 which is also close to 1 and shows that the model is a very good fit for the data, and is able to make accurate predictions.

Mae: 13.000000000000002

Mse: 169.00000000000006

Rmse: 13.000000000000002

R2: 0.9999656095212318

References.

app.croneri.co.uk. (n.d.). *Green Transport Policy | Croner-i*. [online] Available at: <https://app.croneri.co.uk/topics/corporate-social-responsibility-transport-policy-vehicle-emissions/green-transport-policy>.

Drucker, P. (2016). *Management challenges for the 21st century.* Routledge.

Andriy Burkov (2019). *THE HUNDRED-PAGE MACHINE LEARNING BOOK*. Andriy Burkov.

‌

Appendix: The code is attached to the file.

‌