Assignment I

CMP5130 (1) Machine Learning and Pattern Recognition 20/21 (3)

Ahmet Ali Beylihan - 2018021

Bahcesehir University

Keywords: Linear Regression, Regularization, Kfold, Cross Validation, Machine Learning

---

1. Linear Regression

1.1. Univariate Linear Regression

Before applying linear regression, we check the types of columns in our dataset. When we check, we see that ocean_proximity attribute is of type object. All other attributes are of float64 type. So first we need to encode the "ocean_proximity" attribute. Using 'LabelEncoder' we encode our ocean_proximity attribute. Thus, all our attributes are of numeric type. Before applying linear regression, we check the missing values. When we check, we can see that the total_bedrooms attribute contains 207 missing NaN values. We drop the total_bedrooms attribute. After dropping, we apply the StandardScaler to our dataset. After this step, we apply linear regression.

When we apply linear regression for all our attributes, we get the lowest mean absolute percentage error for the median_income attribute which is 38.54359328023379%.

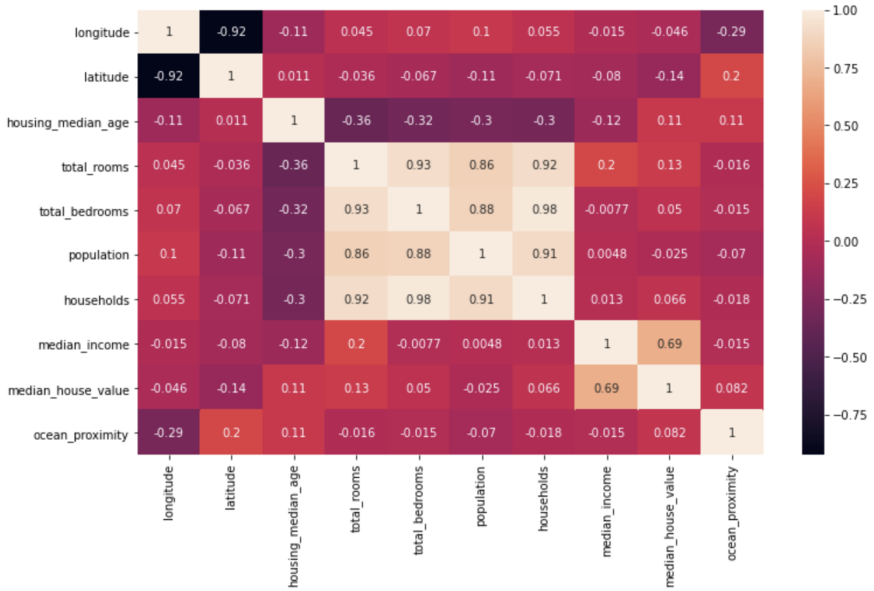

Figure 1: Correlation Graph

When we check the correlation between **median_house_value** and the other attributes, we can clearly see the most correlated attribute is 0,69 positively, that is **median_income**. The graph is at Figure 2. For **median_income**, mean absolute percentage error is 38.54359328023379% which is the minimum percentage error in all features.

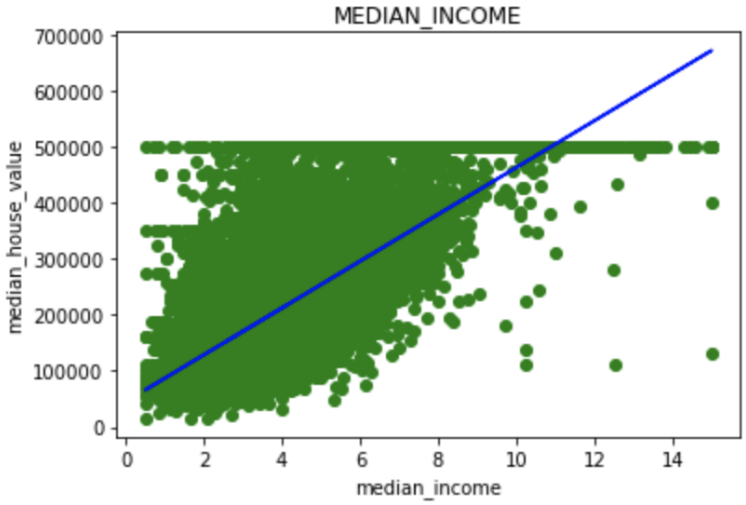

Figure 2: Median Income vs Median House Value

1.2. Multivariate Linear Regression

When we apply multivariate linear regression for our dataset, mean absolute percentage error is 30.70891508061274%. When we print the r-square, it prints 0.6317087499065234. It's seen that the linear model, which we created using all attributes/features, was more successful than the model we obtained when we applied the univariate feature with the median_income attribute.

2. Regularization and Kfold Cross Validation

Cross-validation was used to see the data that the algorithm did not see and make a more objective estimation. After applying cross-validation, ridge and lasso regression were used for various alpha values. According to the regression estimation results, regularization coefficient with min error is 0.00010999999999999999 for Ridge Regression and regularization coefficient with min error is 0.0006799999999999999 for Lasso Regression.