Predicting house values with Linear Regression and Random Forest

Gerli Poopuu GERL0016@STUD.KEA.DK

Jakub Kriz JAKU0526@STUD.KEA.DK

Vaidaras Pranskaitis VAID004@STUD.KEA.DK

Grzegorz Goraj GRZE0203@STUD.KEA.DK

Abstract

This paper describes a data analysis of a housing dataset, together with a prediction of a housing price using Linear Regression and Random Forest. The features of the dataset are described and analyzed, and their predictive strength of house value is evaluated.

1. Introduction

In this paper, we attempt to predict the median house values in California in the 1990s and examine the importance of given features on the median house price. The dataset (Géron, 2017) used for this analysis is taken from the StatLib repository. The research question is framed as a multiple regression problem, where we aim to achieve the lowest root mean squared error (RMSE) (Powers et al., 2010) and examine the predictive strength of the different features.

2. Methods

2.1 Statistical and machine learning methods

Statistical methods were used to examine the dataset. In order to understand the relationship between the dependant and independent variables correlation analysis was used. Linear Regression and Random Forest were used to predict the median house value. This is done using Python data analysis library pandas and Python machine learning library scikit-learn.

2.2 Visualization

The importance of features is presented through visualization. Together with the statistical analysis, different types of graphs like histograms, scatter-plots, and bar-charts are used to communicate the findings and information about the dataset. This is done using the Python plotting library matplotlib.

3. Dataset

The original dataset is the California housing data from StatLib repository (Pace et al., 1997), which consists of all the block groups from California from 1990. A block group

is a geographically compacted area. An average block group consists of 1,500 individuals. For this paper a slightly manipulated dataset was used, which was taken from the Hands-On Machine Learning with Scikit-Learn and TensorFlow book (Géron, 2017). This set contains an extra categorical variable called *ocean proximity* which was inferred from the geological location. The given dataset has 10 variables (Table 1), each consisting of 20,640 observations (except for *total bedrooms*, which has only 20,433 instances). Since our main goal is to predict median housing values in California, we have *median house value* as the dependant variable.

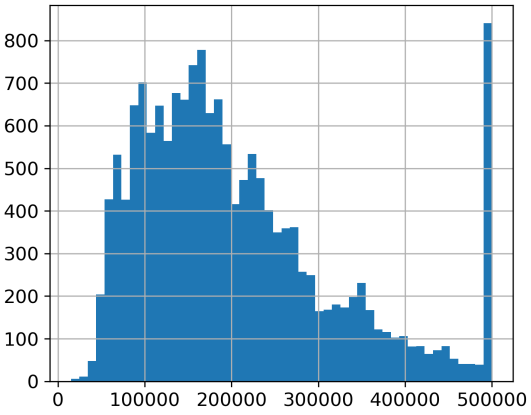

Figure 1: Median house value distribution

Variable	Data type
Median house value	float64
Median income	float64
Housing median age	float64
Total rooms	float64
Total bedrooms	float64
Population	float64
Households	float64
Latitude	float64
Longitude	float64
Ocean proximity	categorical

Table 1: Variables in the dataset.

4. Analysis

4.1 Dependent variable

The dependent variable is *median house value*, expressed in dollars. In Figure 1 it can be seen that the it has been capped at the value of 500,000. This causes a significant problem, since we will not be able to predict house values larger than that. The value of 500,000 is over-distributed in the data, which might cause some problems for our predictions later.

4.2 Correlations with dependent variable

We look at the Pearson correlations (Zou et al., 2003) of other variables with the dependent variable shown in Table 2. We can clearly see that *median income* has the strongest correlation, which can be seen in Figure 2.

Variable	Correlation
Median income	0.69
Total rooms	0.13
Housing median age	0.11
Households	0.07
Total bedrooms	0.05
Population	-0.02
Longitude	-0.05
Latitude	-0.14

Table 2: Correlations with the dependent variable.

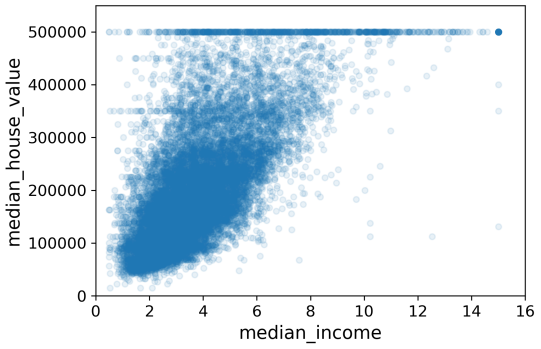

Figure 2: Median house value with median income

It is expected that *longitude* and *latitude* should not have significant correlations, since these variables only make sense together. Their effect on the dependent variable is examined in section 4.3.

4.3 Location

Location of a district is determined by latitude and longitude. These variables do not make sense alone, and have to be considered together. In Figure 3 we can clearly see two dominant areas corresponding to San Francisco and Los Angeles. Note that the data point distribution on the map corresponds to different districts, which gives us no information about population density.

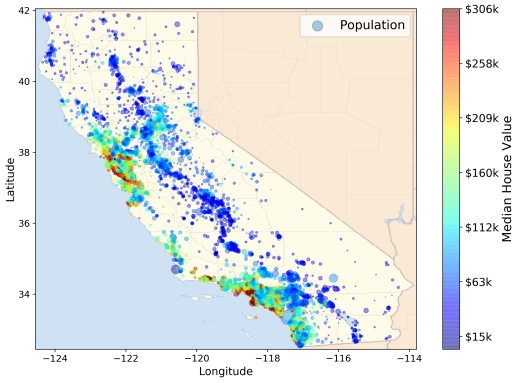

Figure 3: House values plotted on a map of California

4.4 Ocean Proximity

The only variable which helps us with some location information is *ocean proximity*. We have five categories:

- Inland
- Less than 1 hour to ocean
- Near bay
- Near ocean
- Island

Grouping all districts by these categories and calculating the average *median house value* we find that island districts have the highest and inland districts have lowest median house prices as shown in Figure 4.

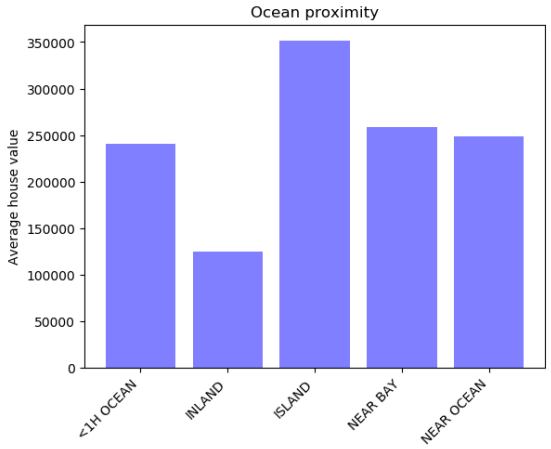

Figure 4: Average median house price per ocean proximity category

However, it is worth noting that island districts are sparse and when compared in terms of total estimated housing value per district are in total worth much less than the other ocean proximity categories. Figure 5 shows that most of the value is concentrated in properties in the *less than 1 hour from ocean* category. This is most likely due to the fact that this area is the most densely populated.

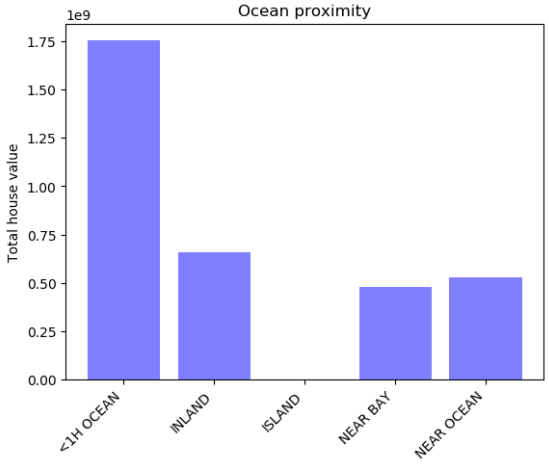

Figure 5: Estimated total house value for each ocean proximity category

5. Housing Price Prediction

5.1 Data preprocessing

5.1.1 Train test split

To know about the generalization ability of our models, we need to split the data into a test set and training set. We perform stratified sampling (Imbens and Lancaster, 1996) on a *median income*, which is the highest correlating feature with the dependent variable. The distribution is split into four strata, and each of them is sampled with same proportions in both train and test set.

5.1.2 Replacing empty values

The value of *total bedrooms* is missing for 207 data points. Based on the correlation analysis done earlier in section 4.2, we expect *total bedrooms* not to be a very important feature. Therefore we expect the 207 data points to still give us meaningful information despite the missing values. Due to that we fill in the missing values with a median of the total bedrooms distribution.

5.1.3 Additional features

To reinforce our machine learning models, we add some additional features that might help predict the housing price. Namely these are *rooms per household*, *bedrooms per room* and *population per household*. In Table 3 we can see weak, but noticeable correlations with the dependent variable for two out of three of these new features.

Variable	Correlation
Rooms per household	0.14
Bedrooms per room	-0.25
Population per household	-0.02

Table 3: Correlations of added features with the dependent variable

5.1.4 Encoding ocean proximity

Being the only categorical variable, *ocean proximity* has to be encoded so that it can be used by the machine learning algorithms, for which we use one-hot encoding. One-hot encoding splits the feature into 5 different features (corresponding to ocean proximity categories) each indicating presence (1) or an absence (0) of the category.

5.1.5 Scaling

Our numerical features have different scales. We apply standardization scaling (Mohamad and Usman, 2013) technique to our data, transforming it so that each numerical feature has a mean of 0, and a variance of 1.

5.2 Linear Regression

We use Linear Regression (Zou et al., 2003) to predict the housing prices. We obtain an $R^2 = 0.65$ and $RMSE = 68628$. If we take a look at the regression coefficients, shown in Figure 6, we see that *island* coefficient has the largest absolute value. Island houses are more expensive (Figure 4), but they are also very sparse in our data so it

is difficult to conclude the causality. *Population* and *households* have moderate coefficients even though they do not correlate with the dependent variable (Table 2). Same goes for *latitude* and *longitude*, for which we concluded in section 4.3 that these variables alone should not have a high impact on the dependent variable alone. This suggests that Linear Regression has a high bias and that a non-linear model is required to discover all the underlying relationships in the data.

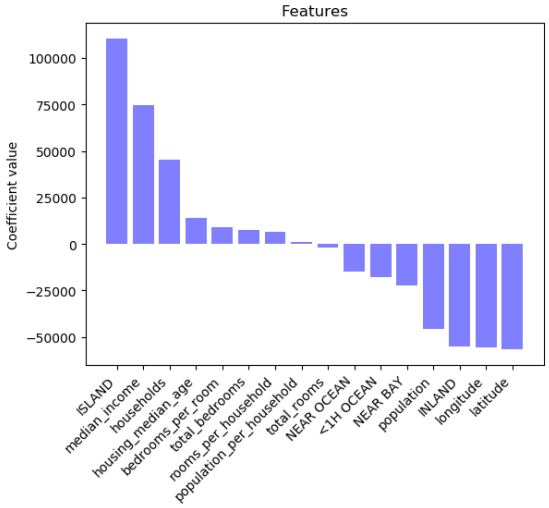

Figure 6: Coefficients of Linear Regression

5.3 Random Forest

We apply Random Forest (Liaw and Wiener, 2002) on our problem as an attempt to capture non-linear relationships. With 100 estimators we obtain $RMSE = 18603$. The estimated feature importance can be seen in Figure 7. The feature importance seems more reasonable than Linear Regression coefficients, with median income being the most important feature. With ocean proximity, inland category seems to be only significant predictor of the house price. This is in accordance with the average prices per ocean proximity Figure 4, where inland is significantly lower than others. Rooms per household and house median age have noticeable importance as well, which corresponds to our previous correlation analysis.

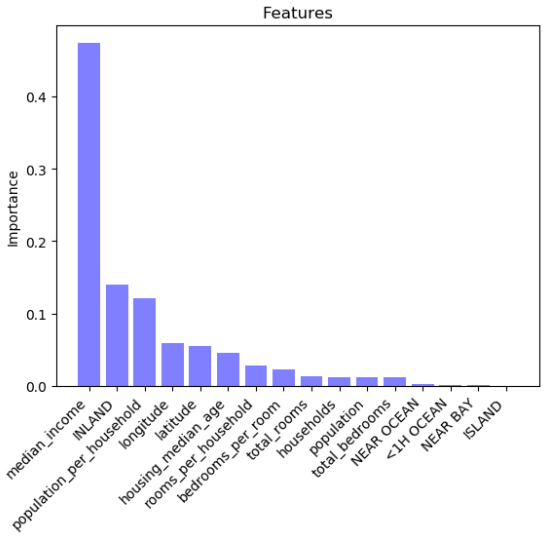

Figure 7: Importance of features in Random Forest regression

6. Findings

The main goal of our research was to predict median house values by looking into 9 independent predictors and examine the strength of these predictors. We developed a Random Forest regression model and achieved $RMSE = 18603$. Through our correlation analysis and feature importances in Random Forest regressor, we discovered that *median income* is the best predictor of the house price. Location of a district is a good predictor as well, with inland districts being less priced than districts near the ocean.

Acknowledgments

We wish to thank our teacher Henrik Strøm for his support and knowledge sharing.

References

Aurélien Géron. Hands-On Machine Learning with Scikit-Learn and TensorFlow:Concepts, Tools, and Techniques to Build Intelligent Systems. In *Hands-On Machine Learning with Scikit-Learn and TensorFlow:Concepts, Tools, and Techniques to Build Intelligent Systems*, chapter 2, page 44. 1st edition, 2017. ISBN 978-1491962299.

Guido W Imbens and Tony Lancaster. Efficient estimation and stratified sampling. *Journal of Econometrics*, 74(2):289–318, 1996.

Andt Liaw and Matthew Wiener. Classification and Regression by randomForest. *R news*, 2002. ISSN 16093631. doi: 10.1177/154405910408300516.

Ismail Bin Mohamad and Dauda Usman. Standardization and its effects on K-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, 6 (17):3299–3303, 2013.

Pace, R. Kelley, and Ronald Barry. California Housing Dataset, 1997. URL `http://www.dcc.fc.up.pt/{~}ltorgo/Regression/cal{_}housing.html`.

Rob Powers, Soumya Ray, Arindam Banerjee, Hendrik Blockeel, Luc De Raedt, Adwait Ratnaparkhi, Johannes Fürnkranz, Claude Sammut, Katharina Morik, Yoav Shoham, Thomas Stützle, Marco Dorigo, Geoffrey I. Webb, Ying Yang, Philip K. Chan, Christophe Giraud-Carrier, Ricardo Vilalta, Jorma Rissanen, Rohan A. Baxter, Ivan Bruha, Stephen Scott, Mauro Birattari, Pavel Brazdil, Luís Torgo, William Uther, Jiawei Han, Xin Jin, Hanhuai Shan, Susan Craw, Prasad Tadepalli, and Carlos Soares. Mean Squared Error. In *Encyclopedia of Machine Learning*. 2010. doi: 10.1007/978-0-387-30164-8_528.

Kelly H Zou, Kemal Tuncali, and Stuart G Silverman. Statistical Concepts Series: Correlation and Simple Linear Regression. *Radiology*, 2003.