Modelling report

Housing Price Prediction by

Zhaklin Yanakieva

Date: 06/04/2021

Table of contents

Versioning table

Time	Changes	Comments	Version
Week 6 - 9	Creating the documents - initial changes	The document was created and the structure was done	1.0.0
Week 10 - 12	Adding the models and explaining; adding results	The modelling will have its representation	2.0.0

Abstract

Machine learning research typically focuses on optimization and testing on a few criteria, but training the models requires more. In this report, I describe the implementation of the modelling phase of the project. My decision of what models and how they will be trained to find the most suitable for the deployment phase is further explained and the evaluation is made after that.

Introduction

I decided to use several models and eventually I can decide which one performed the best in order to use in the next phase - Deployment. I will explore the machine learning algorithms: Logistic Regression, Decision tree, Random Forest. All three will show different results for accuracy. I decided to use these four models so as to check more features for comparing and different aspects.

Modelling

Machine learning is about predicting and recognizing patterns and generating suitable results after understanding them. ML algorithms study patterns in data and learn from them. An ML model will learn and improve on each attempt. To gauge the effectiveness of a model, it’s vital to split the data into training and test sets first. So before training our models, we split the Lending Club data into a Training set which was 80% of the whole dataset and Test set which was the remaining 20%. Then it was important to implement a selection of performance metrics to the predictions made by our model. In this case, we tried to identify whether an individual is going to default on a loan or not. Model accuracy might not be the sole metric to identify how our model performed - the F1 score and confusion matrix should be important metrics to analyse as well. What is important is that the right performance measures are chosen for the right situations. I used 4 algorithms for the modelling purpose: [1]

1. Linear Regression:

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).[2]

2. Lasso Regression:

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).[3]

3. Ridge Regression:

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where independent variables are highly correlated. It has uses in fields including econometrics, chemistry, and engineering.[4]

4. Random forest regressor:

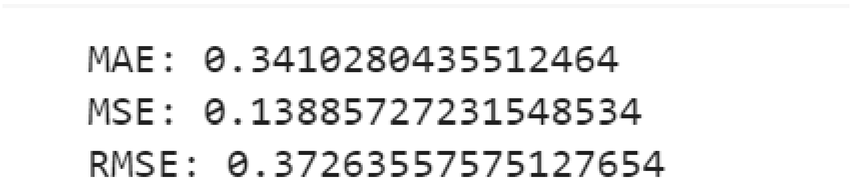
Random Forest Regression is a supervised learning algorithm that uses ensemble learning methods for regression. A Random Forest operates by constructing several

[1] https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012042/pdf
[2] https://en.wikipedia.org/wiki/Linear_regression
[3] https://www.statisticshowto.com/lasso-regression/
[4] https://en.wikipedia.org/wiki/Ridge_regression

decision trees during training time and outputting the mean of the classes as the prediction of all the trees.[5]

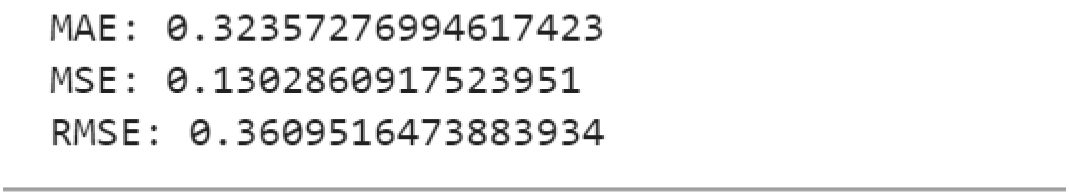
Results:

1. Linear Regression:



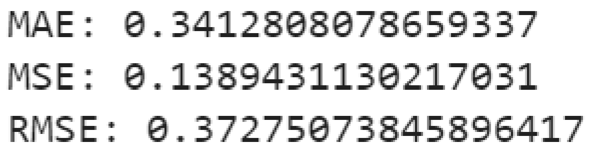
```
MAE: 0.3410280435512464
MSE: 0.13885727231548534
RMSE: 0.37263557575127654
```

2. Lasso Regression:



```
MAE: 0.32357276994617423
MSE: 0.1302860917523951
RMSE: 0.3609516473883934
```

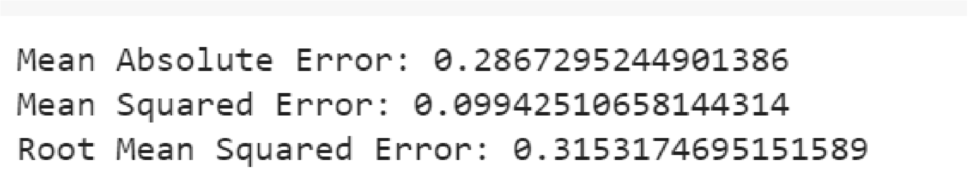
3. Ridge Regression:



```
MAE: 0.3412808078659337
MSE: 0.1389431130217031
RMSE: 0.37275073845896417
```

[5] https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

4. Random forest regressor:

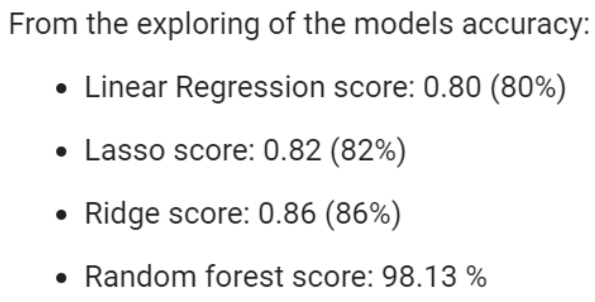


```
Mean Absolute Error: 0.2867295244901386
Mean Squared Error: 0.09942510658144314
Root Mean Squared Error: 0.3153174695151589
```

Evaluation of the models

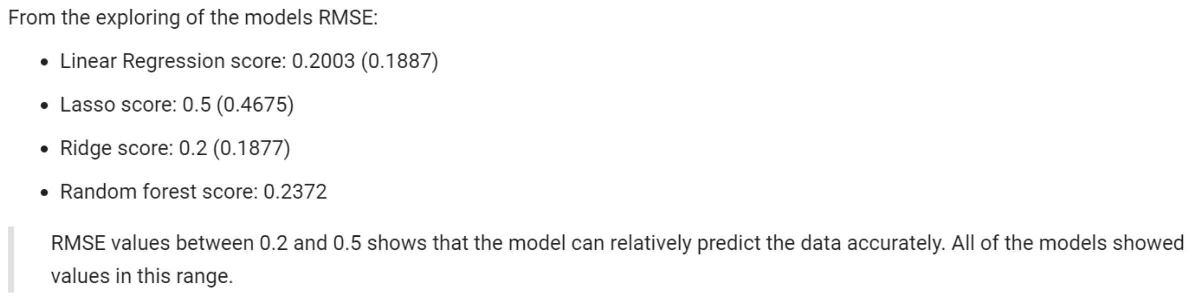
In this paper, I used three machine learning algorithms: Linear, Lasso and Ridge regression, and Random Forest to find the model with the best performance that will work for the housing price prediction.

The results of both the models are shown in the modelling part above and to get a better understanding of the scores of the two models, I will explain more about them here. The results show that the Random forest is the best fit for this project and should be used for the deployment.



From the exploring of the models accuracy:

- Linear Regression score: 0.80 (80%)
- Lasso score: 0.82 (82%)
- Ridge score: 0.86 (86%)
- Random forest score: 98.13 %



From the exploring of the models RMSE:

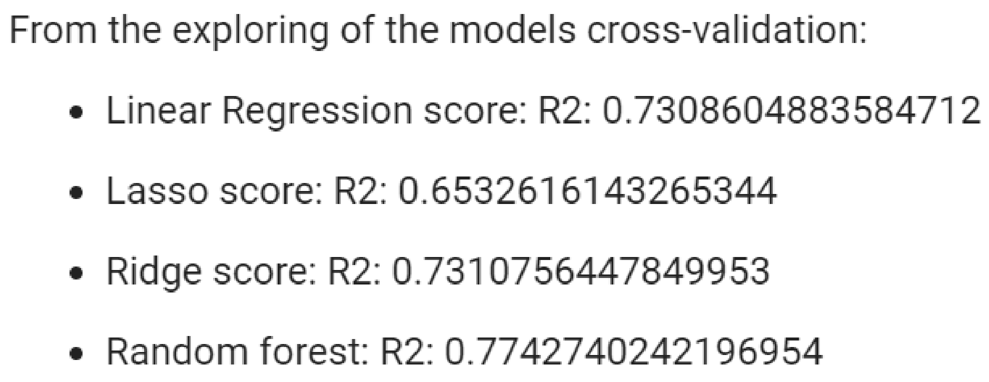
- Linear Regression score: 0.2003 (0.1887)
- Lasso score: 0.5 (0.4675)
- Ridge score: 0.2 (0.1877)
- Random forest score: 0.2372

> RMSE values between 0.2 and 0.5 shows that the model can relatively predict the data accurately. All of the models showed values in this range.

Random forest turns out to be the more accurate model for predicting the house price.

All of the models showed RMSE values between 0.2 and 0.5 so that they show relatively accurate predictions of the data.

In the end, I tried three different models and evaluated them using Mean Absolute Error. I chose MAE because it is relatively easy to interpret and outliers aren’t particularly bad for this type of model. The one I will be using for the deployment is the Random forest



From the exploring of the models cross-validation:

- Linear Regression score: R2: 0.7308604883584712
- Lasso score: R2: 0.6532616143265344
- Ridge score: R2: 0.7310756447849953
- Random forest: R2: 0.7742740242196954

Conclusion

All of the models showed RMSE values between 0.2 and 0.5 so that they show relatively accurate predictions of the data.

I evaluated the models performances with F1 score metric and the one that is overfitting the least is the Random forest. In the end, I tried three different models and evaluated them using Mean Absolute Error. I chose MAE because it is relatively easy to interpret and outliers aren’t particularly bad for this type of model. The one I will be using for the deployment is the Random forest.