Modelling report

Loan Prediction by

Zhaklin Yanakieva

Date: 06/04/2021

Table of contents

Versioning table

Time	Changes	Comments	Version
Week 6 - 9	Creating the document - 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.

I will compare the models by calculating the MAE, MSE, RMSE and the accuracy.

1. Logistic regression:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).[1]

[1] https://en.wikipedia.org/wiki/Logistic_regression

2. Decision tree:

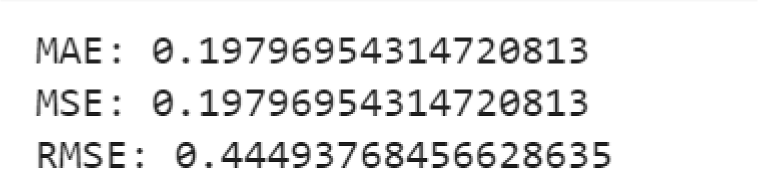
It is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. [2]

3. Random forest:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees.

Results

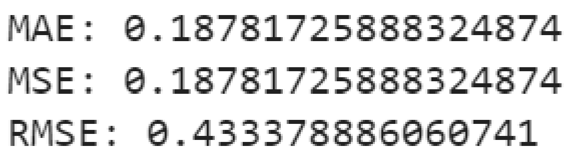
1. Logistic Regression



```
MAE: 0.19796954314720813
MSE: 0.19796954314720813
RMSE: 0.44493768456628635
```

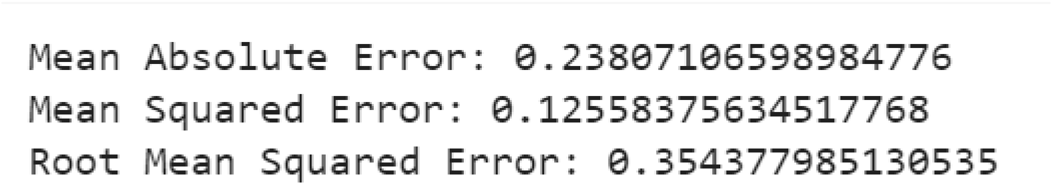
2. Decision tree classifier

[2] https://en.wikipedia.org/wiki/Decision_tree



```
MAE: 0.18781725888324874
MSE: 0.18781725888324874
RMSE: 0.433378886060741
```

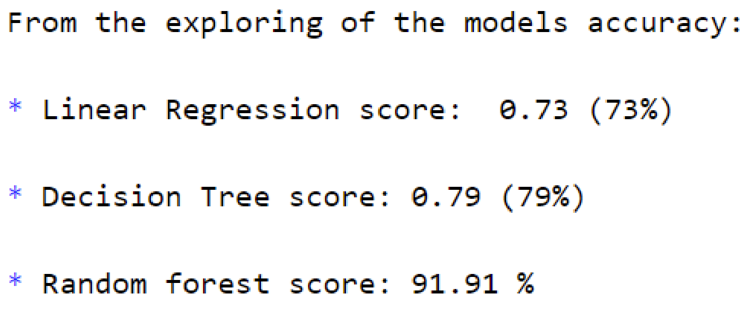
3. Random forest classifier



```
Mean Absolute Error: 0.23807106598984776
Mean Squared Error: 0.12558375634517768
Root Mean Squared Error: 0.354377985130535
```

Evaluation of the models

In this paper, I used three machine learning algorithms: Logistic regression, Decision Trees and Random Forest to find the model with the best performance that will work for the loan 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:  0.73 (73%)

* Decision Tree score: 0.79 (79%)

* Random forest score: 91.91 %
```

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.