Model Analysis - Linear Regression

Rugerio Karen, ITC A01733228, Tecnológico de Monterrey Campus Puebla

Abstract: This document presents the implementation and explanation of an analysis of a linear regression model using Scikit learn library with a dataset of total payment for all the claims in thousands Swedish Kronor. Data split, cross-validation, r2 metrics, and learning curves were applied to validate the performance of the model.

Introduction

When a car insurance claim is raised, the insurance company decides to either approve or reject the claim. The decision is made based on the survey report of the surveyor and documents submitted by the insured. Some of the most common reasons to reject a claim are:

1. Delay in claim intimation.
2. Fraudulent claim.
3. Drive under the influence of a substance.
4. Drive without a valid driving license
5. Violate policy terms and conditions of the insurance company.
6. Insurance policy has lapsed.
7. Car is modified without informing the insurance company.
8. Driver is not the owner of the car driving.
9. Consequential loss happened.
10. Driver negligence

Additionally, the claims fluctuate depending on the damage and the repair needed. Some of the repair costs go beyond the policy limit. For most insurance companies, it is imperative to know and predict how much money they will pay based on the number of claims. Statistical and machine learning algorithms can be applied to facilitate this task.

Dataset

The dataset selected is the Swedish Committee on Analysis of Risk Premium in Motor Insurance [college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/slr/frames/slr06.html]

This data set is represented in a comma separated values file, where:

x = number of claims.
y = total payment for all the claims in thousands Swedish Kronor.

Data split

While preparing the data, a stage of training was performed. The course of action was the next:

1. Split the set into two different sets: A train set and a test set.
2. Use the r2 metric to calculate a score for the train set and a score for the test set.
3. Perform a pre-validation stage of the score using Cross-Validation.
4. Creating a learning curve to visualize the Training score vs

Cross-validation score performance.

It is imperative to have separate sets for training and testing to prove the performance of the model and to avoid overfitting (explained below) when facing predictions. Afterwards, it is important to do a Cross-Validation to do a pre-validation stage using a set different from the training set to avoid reducing the number of elements of this set.

To create the first fata split, the *train_test_split* function from *sklearn.model_selection (scikit learn API)* was used. This function receives the model, X and Y arrays, the test size (for this scenario, the test size was 20% of the full set due to the fact that the dataset size was small. Leaving an 80% of the dataset used for training), and a random seed (1).

Linear Regression

The model selected to make predictions over the dataset previously mentioned, was a linear regression. Linear Regression is a machine learning algorithm which is based on supervised learning. It models a target prediction value of independent variables. It is mainly used to find out the relationship between variables and forecasting. Linear Regressions performs the task to predict a dependent variable value (y) based on a given independent variable (x).

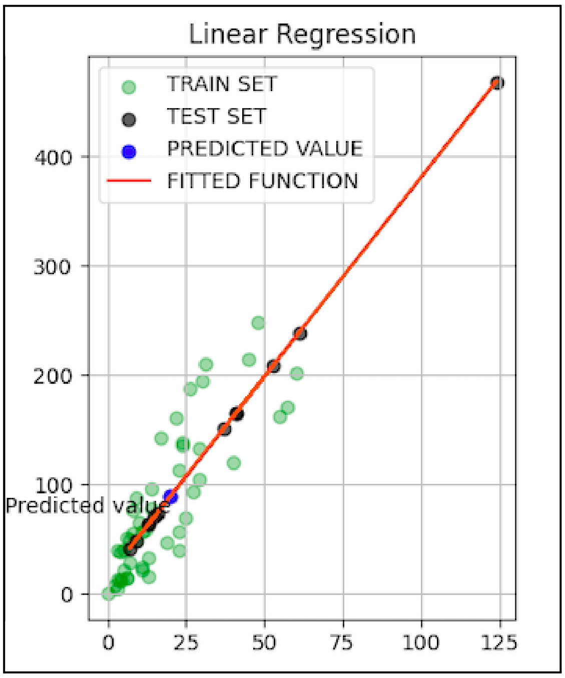
The hypothesis function for linear regression is:

$$y = \theta_1 + \theta_2 . x$$

where:

x = the input training data.
y = labels to data
θ1: intercept
θ2: coefficient of x

(The model gets the best regression fit line by finding the best θ1 and θ2 values)



Running the model

```
$ python3
linearRegressionSci.py -p 20
Predicted values:
 [125.2274803   44.85991831
161.7581903  234.8196103
41.20684731
 190.9827583  121.5744093
37.55377631  37.55377631
128.8805513
  26.59456331  44.85991831
63.12527331 114.2682673
96.00291231
  77.73755731  55.81913131
48.51298931  41.20684731
26.59456331
  33.90070531 106.9621253
59.47220231 121.5744093
223.8603973
```

```
  22.94149231  15.63535031
30.24763431 103.30905431
26.59456331
  99.65598331 110.6151963
48.51298931  30.24763431
66.77834431
  55.81913131 180.0235453
66.77834431  63.12527331
37.55377631
  85.04369931 103.30905431
99.65598331  55.81913131
33.90070531
  52.16606031  99.65598331
216.5542553   30.24763431]

The r2 using train set is:
0.7032179694407252

The r2 using test set is:
0.8758201871131815

Accuracy of 0.70 with a
standard deviation of 0.13
Accuracy of 0.71 after
applying Lasso with a standard
deviation of 0.22
Prediction for 20.0 claims =
88.69677030579116 Swedish
Kronor
```

Standardization

When models are not performing as expected, either when they present an underfitting or overfitting state, a standardization method can be used to improve accuracy. There are different types of standardization methods. In this work, Lasso regression is used. Lasso is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

The accuracy obtained before and after applying Lasso regression were:

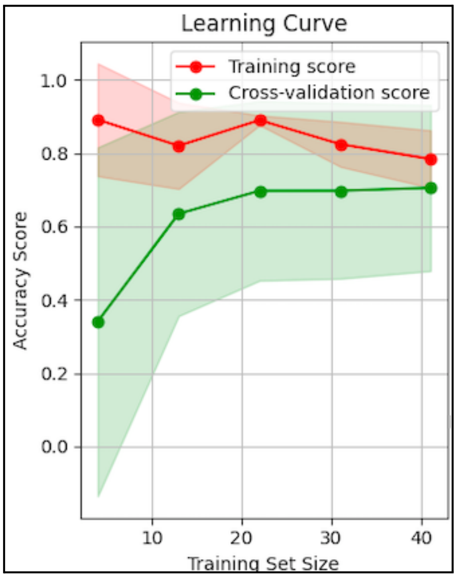
Accuracy of 0.70 with a standard deviation of 0.13 before applying Lasso regression.

Accuracy of 0.71 with a standard deviation of 0.22 after applying Lasso regression.

Even when the original linear regression model was accurate enough, there was a little improvement after applying the Lasso regression on the model.

Results analysis

After running and training the linear regression model, it was evaluated using the techniques mentioned previously. The r2 score using the train set was around 0.7 and the r2 score using the test set was around 8.7. To precisely describe the performance of the model, a cross-validation stage was used along with a learning curve plot to visualize it. A learning curve shows the validation and training score of an estimator for varying numbers of training samples and it used to find out how much the model is benefitted from adding more training data and whether the estimator suffers more from a variance error or a bias error.



As it can be seen from the previous figure, adding more training samples will most likely increase generalization. The training score and cross-validation score are around 0.8, indicating that the model is performing well and adding more data is beneficial.

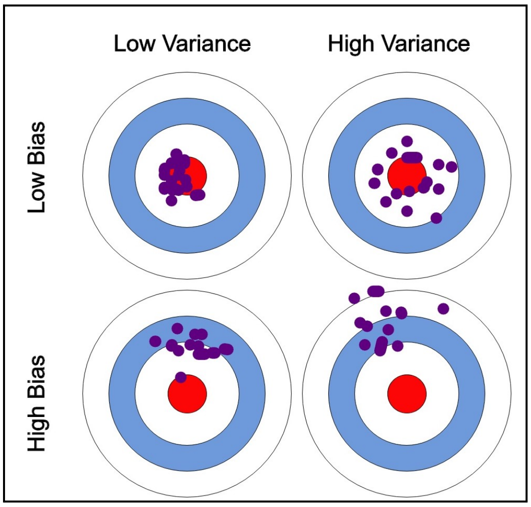
Bias - Variance tradeoff

1. High Bias and Low Variance is a sign of Underfitting. Predictions are mostly consistent but inaccurate on average. It happens when the model is too simple with very few parameters.

2. High Bias and High Variance is a sign of inconsistent predictions and inaccurate on average.

3. Low Bias and Low Variance is the ideal model. Difficult to achieve.

4. Low Bias and High Variance is a sign of Overfitting. Predictions are mostly inconsistent but accurate on average.



The accuracy of the model is high (0.70 before applying the standardization), for which it is possible to determine that the bias level and the variance are low enough. The r2 score is also high (0.70 using the training set and 0.87 using the test set) Therefore, there is neither overfit nor underfit state on the model.

Conclusion

When doing machine learning, it is important to have a good understanding of how to build effective models with high accuracy. There are many models that can be applied to a dataset. Linear regressions model target prediction values of independent variables and are mainly used to find out the relationship between variables and forecasting.
Adding a pre-evaluation stage (cross-validation) and evaluating the model using metrics such as the r2 score and adding is extremely important to understand how well the model performs, the model capacity and if the model is overfitting or underfitting. Standardization methods can be used to improve the accuracy of the model. To additionally improve the results of this model, it would be useful to have more data.

References

Repository: *github.com/KarenRugerioA/Machine-Learning/tree/mod2/UseOfFrameworkML*

Dataset: *Auto Insurance in Sweden. (s. f.). Retrieved September 8th, 2022* *https://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/slr/frames/slr06.html*

Pant, A. (2021, December 7th). Introduction to Linear Regression and Polynomial Regression. Medium. Retrieved September 8th, 2022 *https://towardsdatascience.com/introductio* *n-to-linear-regression-and-polynomial-regression-f8adc96f31cb*

Bias-Variance. (s. f.). Machine Learning Master. Retrieved September 8th, 2022

https://nvsyashwanth.github.io/machinelearningmaster/bias-variance/