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 simple linear regression model with Scikit learn library using a dataset of total payment for all the claims in thousands Swedish Kronor.

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.

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)

To get the best θ1 and θ2 for the best, the cost function is needed(J). Cost function of linear regression is the calculation of the MSE:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

It is also possible to decompose the bias and the variance from this process.

$$\begin{aligned}\text{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\ &= \mathbb{E}(\hat{\theta}^2) + \mathbb{E}(\theta^2) - 2\theta\mathbb{E}(\hat{\theta}) \\ &= \text{Var}(\hat{\theta}) + (\mathbb{E}\hat{\theta})^2 + \theta^2 - 2\theta\mathbb{E}(\hat{\theta}) \\ &= \text{Var}(\hat{\theta}) + (\mathbb{E}\hat{\theta} - \theta)^2 \\ &= \text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta})\end{aligned}$$

Bias and variance of a MSE are important to determine the accuracy and the consistency of the model being evaluated.

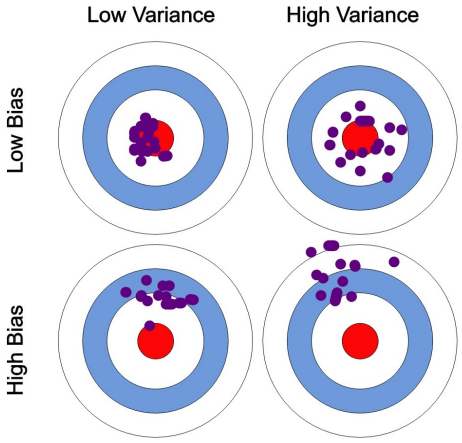
Bias - Variance tradeoff

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.

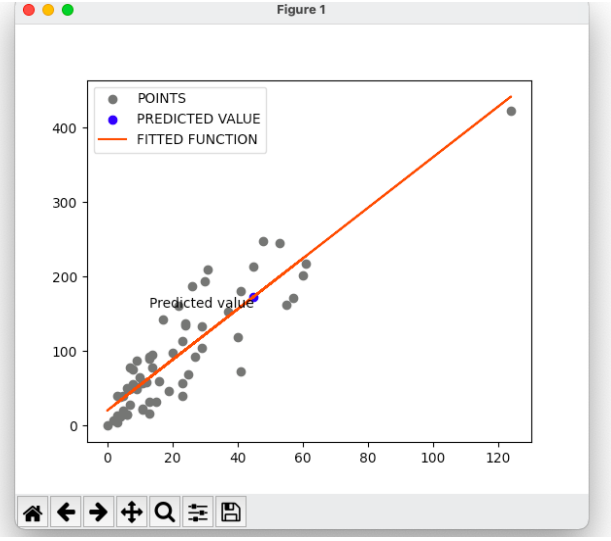
4. Low Bias and High Variance is a sign of Overfitting. Predictions are mostly inconsistent but accurate on average.



Running the model

```
user@system:~% python3
modelAnalysis.py

Y predictions: [ 49.49886882
67.58483121  70.59915828
43.47021469 396.14648132
 145.95733491 182.12925969
61.55617708 206.24387621
145.95733491
 133.90002665  61.55617708
82.65646654]
MSE: 1671.7795956092089
Bias: 1384.4849490557654
Variance: 287.2946465534371
```



Result analysis

In this scenario, the Bias is way larger than the Variance: This is an indicator of an underfitting model. Moreover, both the Bias and the Variance are high. Which could lead to some inconsistent and some inaccurate predictions.

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.
Bias-variance decomposition is extremely important to understand how well the model performs, the model capacity and if the model is overfitting or underfitting. To improve the results of this model, it would be useful to have more data and/or more parameters.

References

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

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/introduction-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/