Assignment IV– Individual Take Home Assignment

Rahul S. Kalubowila

Department of Statistics, University of Colombo

IS 3005

2016S16026 | S13581

Contents

Acknowledgment

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. A special gratitude I give to the course lecturer, Dr. Nadeeka Basnayake, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my project especially in writing this report. Furthermore, I would also like to acknowledge with much appreciation the crucial role of the staff who gave the permission to use all required equipment and the necessary materials to complete the task. I have to appreciate the guidance given by supervisors especially in our analysis that has improved our analysis skills and thank their comment and advices.

Executive Summary

Following is the executive summary from data of 396 student on the comparison of Multiple Linear Regression (MLR) model & Artificial Neural Network (ANN) to predict the final grade of a mathematics examination.

Out of the 2 models MLR emerged superior in both Mean Squared Error (MSE) and Runtime. Additionally, when it comes to convenience and interpretability MLR is better.

Hence, Empirically MLR model is superior to ANN regression model

Introduction

Outline

The purpose of this study is to compare empirically ANN regression model & multiple linear regression model. For this study we are analyzing a dataset that consists of student achievement in secondary education of a foreign school. We an ANN regression model and a multiple linear regression model to predict the final grade and compare performance.

Data

Data Source – Data was presented in a text file in .csv format, Delimited by a comma. And contained 396 entries of data in 5 continuous variables.

Column Name	Description	Data Type
Age	Student's Age	Continuous
Absences	Number of school absences	Continuous
G1	1st period grade	Continuous
G2	2nd period grade	Continuous
G3	3rd period grade	Continuous

Methodology

Analysis was carried out using Python programming language.

Objectives

Compare MAE

Compare Runtime

Model Analysis

Multiple Linear Regression Model

The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. The following is the summary for the MLR model that was obtained for the school grades data set.

OLS Regression Results

Dep. Variable:	G3	R-squared:	0.835
Model:	OLS	Adj. R-squared:	0.833
Method:	Least Squares	F-statistic:	342.9
Date:	Thu, 27 Feb 2020	Prob (F-statistic):	1.06e-104
Time:	20:18:41	Log-Likelihood:	-565.82
No. Observations:	276	AIC:	1142.
Df Residuals:	271	BIC:	1160.
Df Model:	4		
Covariance Type:	nonrobust		

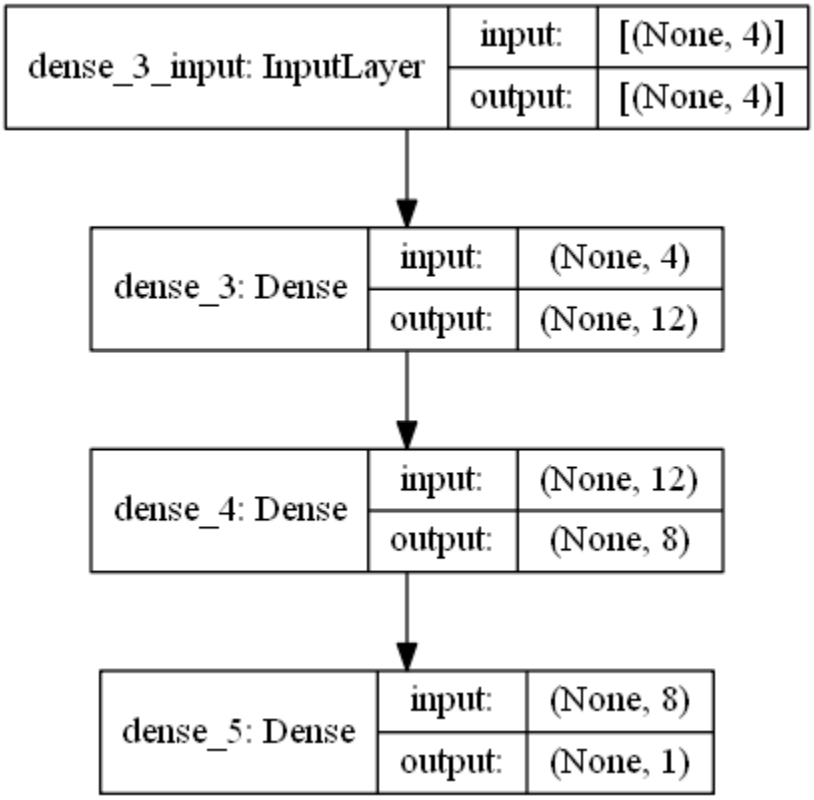
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.3030	1.625	0.186	0.852	-2.896	3.502
age	-0.1506	0.093	-1.619	0.107	-0.334	0.033
absences	0.0488	0.015	3.217	0.001	0.019	0.079
G1	0.1799	0.067	2.671	0.008	0.047	0.312
G2	0.9669	0.061	15.953	0.000	0.848	1.086

Omnibus:	156.800	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	895.667
Skew:	-2.353	Prob(JB):	3.22e-195
Kurtosis:	10.466	Cond. No.	338.

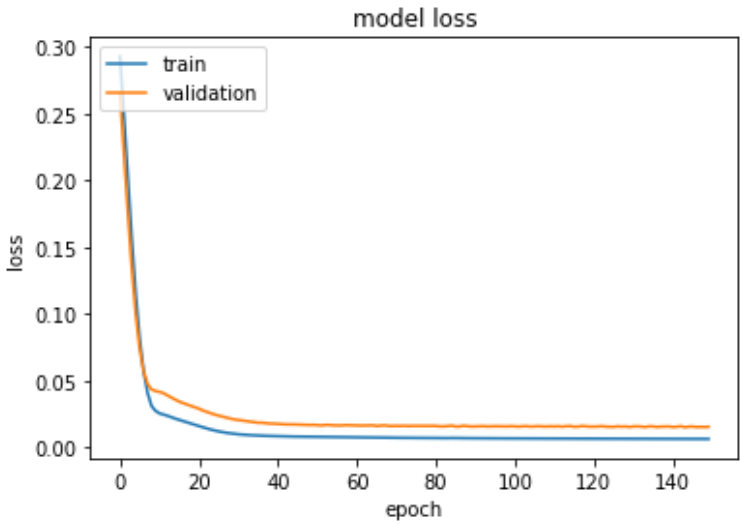
We can see that the model gives an adjusted R^2 value of .835. Which indicates that the model explains 83.5% of the variability of the response data around its mean. We can see that age has a negative relation with the final grade. Surprisingly, number of school absences are positively correlated with the final grade.

3-layer Deep Artificial Neural Network Regression Model

The Artificial Neural Network (ANN) is one of the well-known prognostic methods used to find a solution when other statistical methods are not applicable. The advantages of this tool, such as the ability to learn from examples, fault tolerance, operation in a real-time environment, and forecasting non-linear data, all make it a widely used statistical tool. Following is the summary for the ANN model that was obtained for the school grades data set.



We created an ANN model with 3 hidden layers. For 1st & 2nd layers, Rectified Linear Unit (ReLU) activation function was used. And for the output layer a linear function was used since this a regression problem. Given below is the loss function with increasing epochs for the training and validation data set.



Comparison

	MLR	ANN Regression
MSE	1.4293	3.8331
Runtime	0.0169 seconds	2.1950 seconds

Mean Squared Error (MSE)

The mean squared error tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the "errors") and squaring them. From the 2 models MLR has a lesser MSE value. This means that errors in estimation are lesser in MLR. Therefore, MLR is superior in terms of MSE.

Also, it is important to note that when calculating the MSE for ANN model the scaled MSE had to be unscaled to get a fair comparison.

Runtime

In computer science, runtime, run time or execution time is the time when the CPU is executing the machine code. Here, we want to see how much time is needed to run the algorithm and relate that to resource consumption. We can see that the MLR algorithm is approximately 130 times faster than the ANN model. Factors like number of layers, Units in layers and iterations increases the runtime for an ANN model.

Conclusion

In conclusion we can deduce that MLR is superior to ANN regression model when analyzing the school grades data set in terms of both MSE and Runtime.

Factors like type of data, number of layers, Units in layers and number of iterations effects performance significantly.