Performance of Linear Regression in Predicting Graduates' Salaries

Raj Choudhary*
**School of Computer Science and Engineering, VIT Chennai, Tamil Nadu, India 600127*
Email: raj.choudhary2016@vitstudent.ac.in *

Abstract: **Linear Regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.**

The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Index terms **– Linear Regression, Regression, Employability, Forecasting**

1. Introduction

In this paper, salary slab classification is attempted using the AMEO dataset which contains the employability outcomes of engineering graduates. The features include standardized test scores, academic performance metrics and other background details of the candidate. The target variable is the first salary of the candidate. Simple Linear Regression, regularized Linear Regression as well as Polynomial Regression is used to predict the salaries of the graduates.

2. Methodology

Simple Linear Regression is used to predict the salaries of fresh graduates' using the 'Quant' feature of the dataset as from the studying the important features from Decision Tree Classifier, it was found that it was the most deciding feature. After this, regularized linear regression and polynomial regression were implemented to see if the performance of the model could be improved.

3. Dataset – AMEO 2015

For every engineer, AMEO [1] dataset provides anonymised bio data information along with their respective skill scores and employment outcome information. Specifically, the following information is available for every engineer:

1) Scores on Aspiring Minds' AMCAT - a standardized test of job skills. The test includes cognitive, domain and personality assessments.
2) Personal information like gender and date of birth.
3) Pre-university information like 10th and 12th grade marks, board of education and 12th grade graduation year.
4) University information like GPA, college major, college reputation proxy, graduation year and college location.
5) The following employment outcome information is available for every engineer: First job, annual salary, First job title, First job location, Date of joining and leaving of first job.

AMEO 2015 has gained traction since its public release. Aspiring Minds annually publishes the National Employability Report, a data-driven commentary on graduates and their employability. A recent NER was based on an extension of this dataset.

4. Analysis and result

From the R2 scores measured by the different regression model on the dataset, clearly the model is not able to perform efficiently on the dataset based on just one feature. Whether it is simple regression, regularized linear regression using Lasso or polynomial regression, all score 0.05 which is very poor. This could be possible as we are only using 1 feature out of the 70 available features. So we might be able to recognize the pattern as we might be missing on important features.

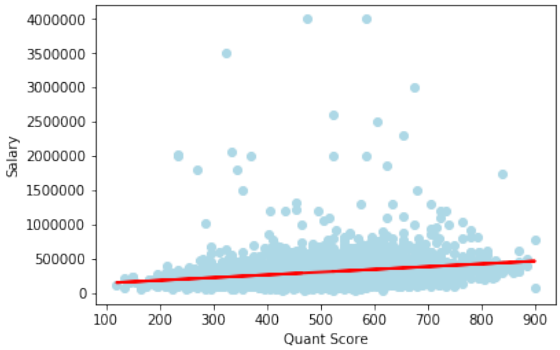

Figure 1. Simple Linear Regression

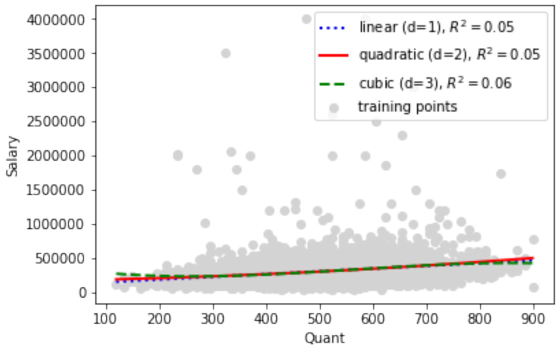

Figure 2. Polynomial Regression with multiple values of d

5. Conclusion

Here, we confirm that univariate regression whether it be simple, regularized or with polynomial features will not be able to predict the salaries correctly based on just 1 feature as it is not able to recognize the pattern in the dataset. So, we need to employ multivariate regression so that our model can recognize the pattern in the dataset and predict the salaries correctly.

6. Reference

[1]. V. Aggarwal, S. Srikant, and H. Nisar, “Ameo 2015: A dataset com- prising amcat test scores, biodata details and employment outcomes of job seekers,” in *AMEO 2015: A Dataset Comprising AMCAT Test Scores, Biodata Details and Employment Outcomes of Job Seekers*. ACM, 2016.