Multiple Regression Analysis Report

on

ABC Company Employees' Salaries

Submitted to Prof. Murat Guner, PhD

Data Analytics & Machine Learning

UMBC

Tamrat Workineh

October 13, 2020

Table of Content

Abstract

This paper describes the application of multiple regression model to predict the annual salaries(target) based on sex, age, work experience and educational backgrounds of an ABC company employees. Based on these numerical values of the features and the target variables, the performance of the model predicted with accuracy of 0.85.

Introduction

Multiple Regression Analysis is a powerful tool for predicting and forecasting variables. Regression allows to study the relationship between dependent and independent variables thereby observe patterns to predict relating variables to each other. There are several projects out there based on this machine learning model on house sale. The researcher was motivated to use this model because it is a reliable method of identifying which variables have impact on a topic of interest. In this project, a practical example of ABC company employees' salaries was predicted based on respective sex, ages, work experiences and educational backgrounds. In the following section. Motivation, Related work, Proposed method, Experiments, Results and discussion, and Conclusion and summary were made based on the study conducted.

Study Objective

The goal of the project is to determine the ability of age, sex, work experience and education to predict the annual income of employees at ABC company by using multiple regression machine learning algorithm.

Motivation:

The motivation of behind using regression is may be due to its linearity, simplicity and it allows one to quantify the effect of each individual factor by also considering the interactions between the factors. Besides, in a regression you one interprets the results as marginal effects.

ABC- Company Dataset

The ABC Company has 500 employees. The dataset of the company originally contains 500 rows and 13 columns. After data wrangling, it is reduced to 500 rows and 6 columns.

Variables	Description
Emp_ID	Employee _ID
Emp_Age	Age of employees
Experience	Work experience of the employees
Sex	Male or Female Employee
Education in Year	Educational Qualification (Diploma/Degree / Masters
Annual_Rate	Annual salary of employee

Table 1 Explanatory variables used in the Study

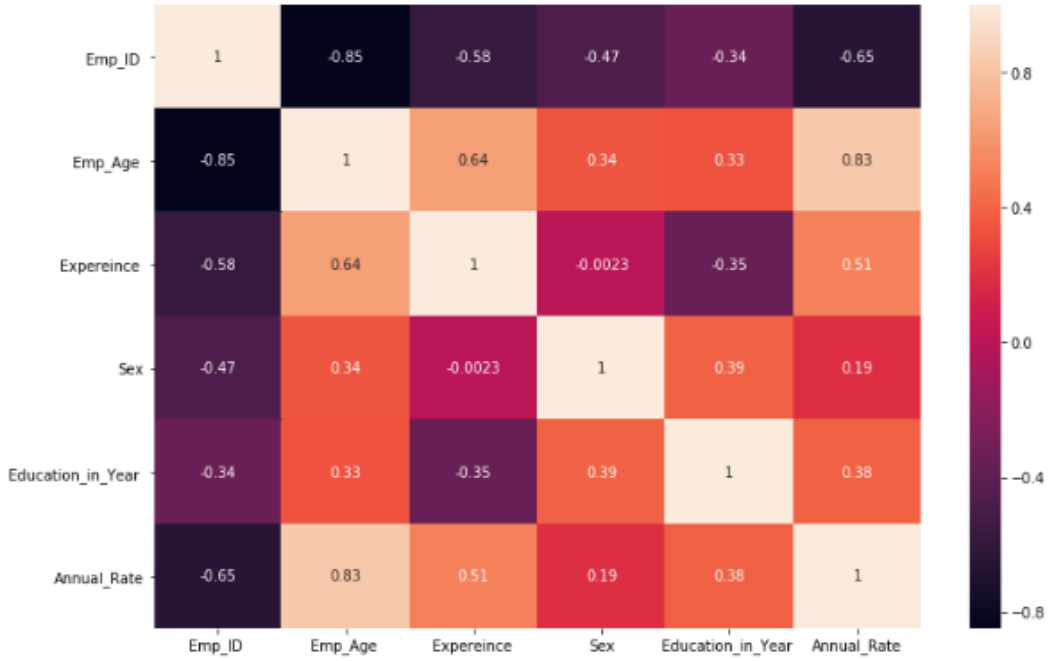

Fig 1 : Heatmap : The heatmap shows the plosive correlation as we go up to lighter color.

Method:

Data Wrangling:

The ABC Company has 500 employees. The dataset of the company originally contains 500 rows and 13 columns. After data wrangling, it is reduced to 500 rows and 6 columns.

Data Wrangling

```
In [7]: # Data need to clean and in good format before applying conducting the anlysis

In [8]: data.drop(['Unnamed: 6', 'Unnamed: 7', 'Unnamed: 8', 'Unnamed: 9',
                   'Unnamed: 10', 'Unnamed: 11', 'Unnamed: 12'],axis=1,inplace=True)  # deleting unecessary columns androws

In [9]: data.head(5)  # data after cleansing
```

Out[9]:

	Emp_ID	Emp_Age	Experaince	Sex	Education_in_Year	Annual_Rate
0	8169	25	3	1	2	47278.4
1	1958	32	3	1	4	47902.4
2	5601	27	3	1	4	47902.4
3	6731	29	3	2	4	47902.4
4	8171	31	5	1	2	40902.4

```
In [21]: plt.figure(figsize=(12,8))
         sb.heatmap(data.corr(),annot=True )
         plt.show()              # From the heatmap, strong correlation the light color show strong
                                 # correlation whilit the dark correlation show us the strong negative coorelation.
```

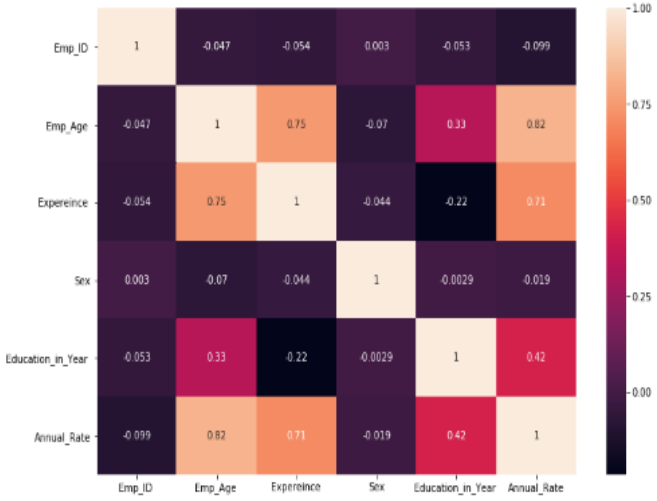

Fig : 2 *The heatmap shows the correlation between the predictor and predictand variables. The lighter color represents more positive correlation.*

Train-Test Split dataset:

This technique was used to take a sample data to evaluating the performance of the linear regression model. To this effect, the feature and target variables were a dataset were split into two subsets 80% to 20 % training and test respectively. Accordingly, the dataset was divided to contains X and Y trains 400 data rows each whereas X and Y test to have 100 rows each.

```
Out[36]: Index(['Emp_ID ', 'Emp_Age ', 'Expereince ', 'Sex', 'Education_in_Year ',
                'Annual_Rate'],
               dtype='object')

In [44]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)

In [46]: X_train.shape

Out[46]: (400, 4)

In [48]: y_train.shape

Out[48]: (400,)
```

Model:

Theory for multiple linear regression

In multiple linear regression, there are p explanatory variables, and the relationship between the dependent variable (Y) and the explanatory variables (X) is represented by the following equation:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p ,$$

Experiments :

In this project, a standard multiple regression analysis was performed to assess the ability of age, sex, work experience and education to predict Annual rate of employees.

Results and Interpretations

A multiple regression model was calculated substituting the formula with the coefficients and Y-intercept to preditct the target employees' annual rate. Based on numerical values of age, sex, work experience and education predictors. The accuracy of the model is found to be 0.85.

Conclusion and Summary:

The goal of the project is to determine the ability of age, sex, work experience and education to predict the annual income of employees at ABC company by using multiple regression machine learning algorithm. The model, based on numerical values of age, sex, work experience and education predictors. The accuracy of the model is found to be 0.85.

Limitations and later work.

Due to ongoing class and limited data model knowledge the study was conducted based on regression model. It has a constraint on generalizability of findings. However, the study will be validated as after additional models are taught.

References:

1. https://www.valuebasedmanagement.net/methods_regression_analysis.html
2. https://www.statisticssolutions.com/sample-size-formula/