Analyzing MBA student's data to Predict the salary

Nikhil Lobo
W3124829
nikhil.lobo@my.uwrf.edu
University of Wisconsin River Falls
May 8th, 2019

Abstract:

Background: Base salary of MBA Students has always captured the attention of graduates and starting salary has become one of the main indicators of how well the course. **Objective**: Analysis of MBA Starting Salaries dataset. **Method**: T-test, K-NN and Regression models used to analyze and predict the starting salary of the graduates. **Result:** The students data set has been analyzed and salary can be predicted using various parameters like work_experience, age and gmat_scores. **Conclusion:** After running the t.test, we can't say that the average salary of males is greater than females because the p value is more than 0.05. When used predictor variables gmat_vpc, gmat_qpc and age in the model shows better relationship with the dependent variable based on R^2 value.

Introduction:

Obtaining a Master of Business Administration (MBA) degree is a significant accomplishment, offering many benefits. The acquired knowledge and training can position the graduate to introduce and implement sound business practices into their career. MBA degree can open up several career advancement opportunities in addition to providing desired salary increases. For many students, perhaps the greatest concern is whether earning this prestigious degree is cost-effective.

According the GMAC's 2011 Global Management Education Graduate Survey, employees who plan to stay with their current employer, expect a 39% increase in salary after earning their MBAs. In this research project, the main goal is to be predicting the starting salary or expected increases in salary for the experience candidates after earning an MBA degree.

Method:

Data

The MBA graduate's data downloaded from the uci.edu. which has 274 observations and 13 attributes describe the graduate students academic and work details.

Analysis

Raw data was prepared for analysis using tools from the tidyverse (Wickham, 2018).

Data set had various not related data. Like some of the observations with salary column had unwanted entries those observations have been removed from the data to keep the data simple.

Before the prediction of graduate's salary t.test has been performed on the average salary of males and females and the result from the T_Test is below.
Observations from T_Test

- Can't say that the average salary of males is greater than females because the p value is more than 0.05.

For the prediction, data was summarized. Then three methods were used and compared to predict the salary of MBA graduates.

- The first method was the Mean as a model.

  $\hat{y} = (y1 + y2+\ldots. yn) / n$

- The second method was the Linear Regression model. Better fitted compare to mean as a model.

  $Y = a + bX$

- Third model is the multiple regression model

  $y_i = B_0 + B_1x_{i1} + B_2x_{i2} + \ldots B_{p}x_{ip}$ for $i = 1,2, \ldots n$.

The best model is selected among the models listed above based the R^2 value which is the total sum of the errors.

To classify graduate student is employed and not used two methods and compared.

- K-NN
  To determine which of the K instances in the training dataset are most similar to a new input, a distance measure is used. For real-valued input variables, the distance measure is Euclidean distance.

  EuclideanDistance (x, xi) = sqrt (sum ((xj – xij) ^2))

- Logistic Regression
  To calculate the probability on categorical variable the formula used is:

$$P= \frac{e\text{^}a+bX}{1+e\text{^}a+bX}$$

Result:

Many observations can be made by looking at figure 1 and 2. It shows the count of number of male and female employed.

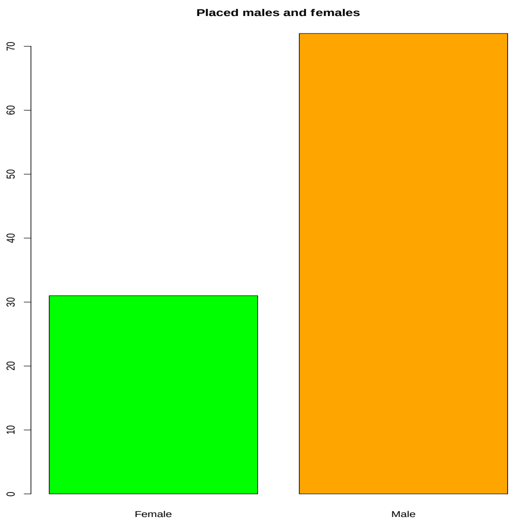

Figure:1

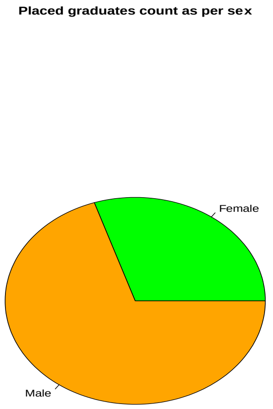

Figure: 2

Each person has his own perspective towards the job. In the dataset there is one column which tells the job satisfaction of the graduates. Its rated as 1 - Low and 7 - High. Figure-3, the histogram shows the graphical representation of job satisfaction of graduates from the entire dataset.

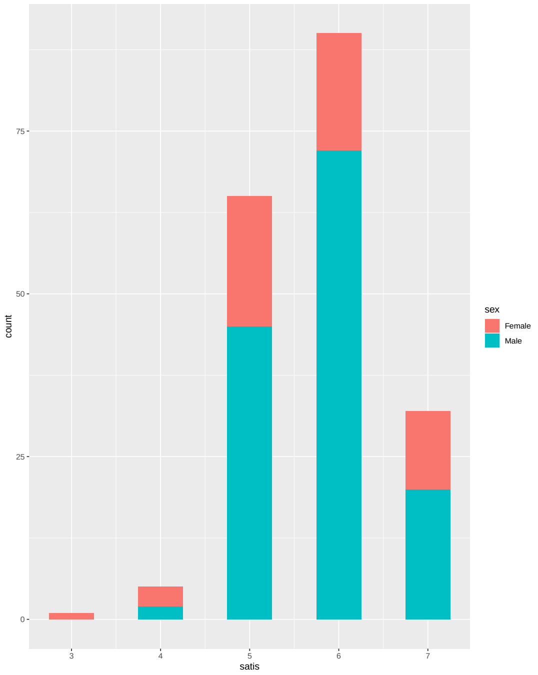

Figure-3

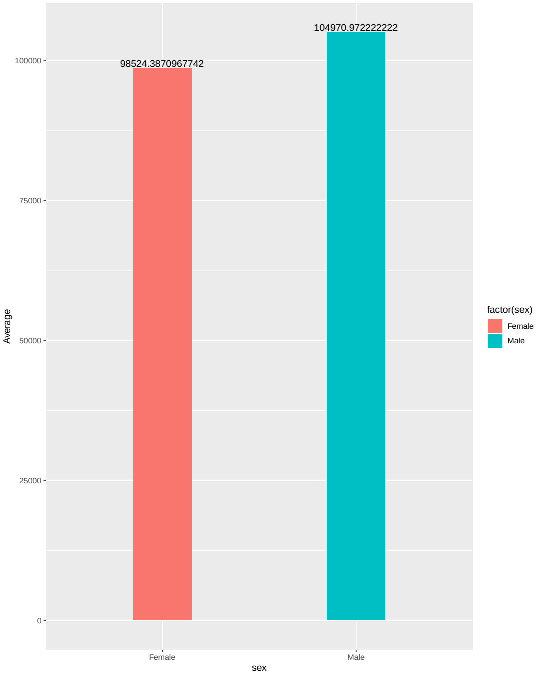

Figure: 4

From the dataset we made two groups employed and not employed using the salary column of the dataset. In the figure-4, the bar plot compares the average salary of males and females from the employed data.
To predict the salary we used mean model, which is not the best fit for the data. Then used linear regression model as in the figure: 5 to show how the line best fitted to the data.

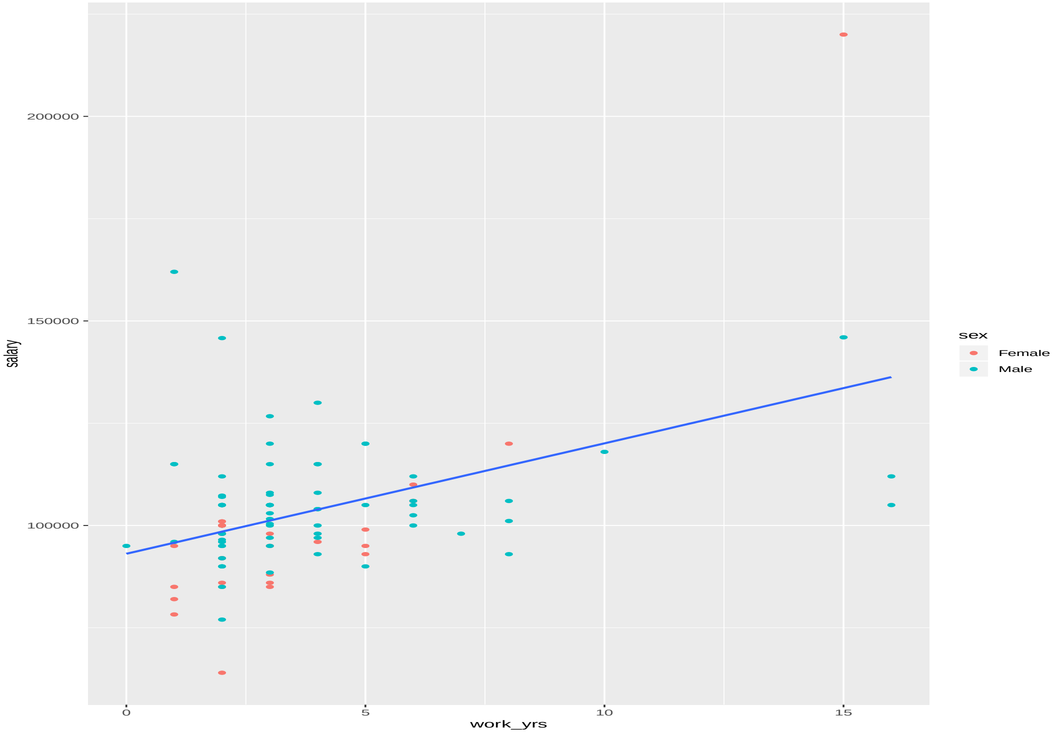

Figure: 5

Then used multiple regression model to compare with linear regression model, which is best fit than linear model. The analysis is made based on the R^2 and P-values.
For multiple regression model, initially inputted all the variables. Then I removed one by one variable in step wise, removing the highest p value variable and keep doing regression until maximum adjustable R-squared is obtained. Which is actually best fit to regression.

Used K-NN and Logistic Regression methods for classifying the graduates as employed and not employed and compared both the k-NN and logistic regression model.
K-NN mis-rate is 0.02590674 which is good. K-NN does not tell us which predictors are important, we don't get a table of coefficients with p-values.
Logistic Regression tells the which predictors are important using the p-values.

Discussion:

To conclude, the salary of the graduate students can be predicted based on their age, work experience and competitive exam scores. There is no much difference in the average salary of male and females. As compared in the figure-4.
After comparing all the models (mean, linear regression and multiple regression) based on the R^2 value, the multiple regression model best suitable for this problem.
For future analysis if the dataset has the specific attributes about the students like their interests and studying hours can helps to predict the salary accurately.

References:

deJong, J. (2006). Why do the Companies Prefer MBAs? Monster.com. Retrieved September 19, 2006 from: http://featuredreports.monster.com/mba/mbapreferred/

Schoenfeld, G. (2006). The Corporate Recruiters Survey 2006 Survey Report. McLean, VA: Graduate Management Admission Council.

Smith, A. (1994). The Wealth of Nations. (E. Cannon, Ed.) New York: Random House, Inc. (Original work published 1776.)

QS TopMBA.com (2008) JOBS & SALARY TRENDS REPORT

Graddy, K., and L. Pistaferri. 2000. "Wage Differences by Gender: From recently graduated MBAs." Oxford Bulletin of Economics and Statistics 62: 837-854.

David Joseph Hoppck (Feb 17, 2019) What is the Average Salary for an MBA Graduate? FROM: https://www.investopedia.com/articles/personal-finance/031215/what-average-salary-mba-graduate.asp