Data Analysis for Education Consultants

Qaif Fangari

May 26th, 2020

1. Introduction

1.1 Background
Education consultancy is one of the major industry which has been in operation for a quite a some time now, generally their success is measured based on the parameter that how well the student manage to earn after following their advice on the choice of college and the course. Generally, a master's degree in business administration is an ultimate stage of an individual's education. Master of Business Administration (MBA) is the course which is taught all over the globe. The objective of the course is to help students get acquainted with tricks of doing business and the course offer specialization in various fields such as Marketing, Finance and HR. Therefore, it is important for education consultants to have knowledge about what kind of choices prior to and while doing MBA will yield maximum return. Wherein returns are considered to be the salary offered after the completion of MBA.

1.2 Problem
Data that might contribute to determining salary of students after completion of education include his past academic details and records. This project aims at determining ideal career path for MBA enthusiasts.

1.3 Interest
Obviously, Educational consultants would be interested in determining ideal career paths for their students. This will help consultants to provide better service to their clients and would also help students be assured about their career choices.

2. Data acquisition and cleaning

2.1 Data sources
The data set is obtained from Kaggle.com (https://www.kaggle.com/benroshan/factors-affecting-campus-placement) . It contains details about salary offered to students after completing Post Graduation in Business Administration and their educational background.

Our dataset contains following labels:

1. Gender
2. Secondary education percentage
3. Secondary education Board
4. Higher Secondary Education percentage
5. Higher Secondary Education board
6. Under-Graduate percentage
7. Under-Graduate field
8. Work experience
9. Aptitude test percentage
10. MBA specialization
11. MBA percentage
12. Status
13. Salary offered

2.2 Data cleaning

Data downloaded or scraped from the source contained lots missing value. As some people were mot placed, these missing values belonged to the Salary column, I decided to drop those rows with the missing data.

2.3 Feature selection

After data cleaning it was observed that there is some data which is not relevant to our analysis. Those are the board of education, as it has been observed that companies consider quantitative information. Hence we ignore the data on board of Secondary and Higher Secondary Education.

3. Exploratory Data Analysis

3.1 Correlation between non categorical factors

	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
sl_no	1.000000	-0.093480	-0.218428	-0.102250	0.041467	-0.072432	0.063764
ssc_p	-0.093480	1.000000	0.293416	0.380657	0.317892	0.430560	0.035330
hsc_p	-0.218428	0.293416	1.000000	0.221307	0.284672	0.329983	0.076819
degree_p	-0.102250	0.380657	0.221307	1.000000	0.217683	0.494093	-0.019272
etest_p	0.041467	0.317892	0.284672	0.217683	1.000000	0.284143	0.178307
mba_p	-0.072432	0.430560	0.329983	0.494093	0.284143	1.000000	0.175013
salary	0.063764	0.035330	0.076819	-0.019272	0.178307	0.175013	1.000000

Figure 1: Correlation between different parameters

The above table gives information on level of correlation between various parameters mentioned in dataset. Our purpose is to find the impact of non-dependent variables which are percentage in secondary, higher secondary, degree and MBA education on dependent variable which is salary. We calculate the correlation using simple and multiple linear regression.
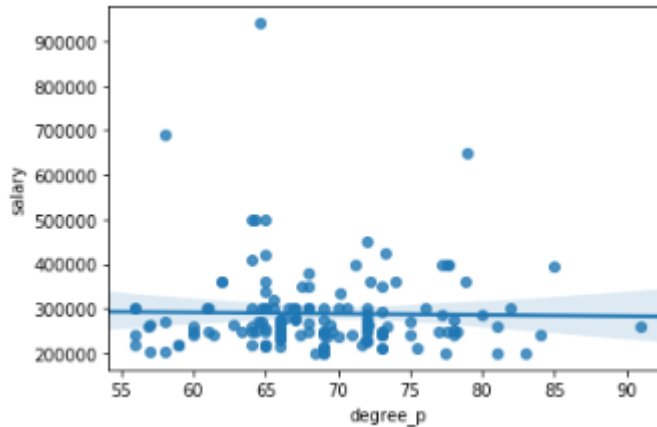
3.1.1 Correlation between MBA percentage and salary
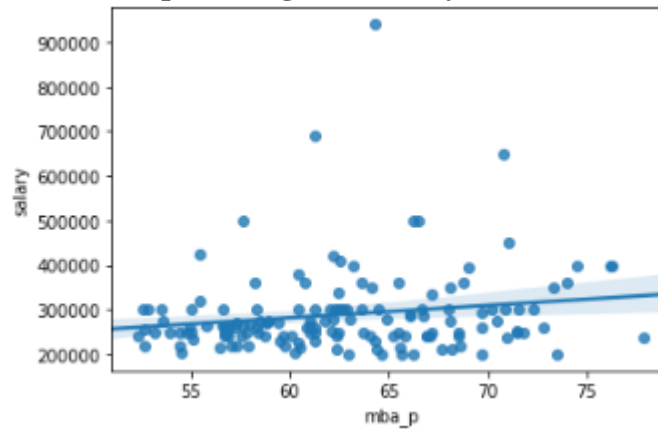
The Pearson Correlation Coefficient is 0.1750129406952748 with a P-value of P=0.03337689255770949. As P-value is less than 0.05; there is a moderate evidence that correlation is significant. The linear regression equation is as follows Yhat= 114715.29018899679 + [2779.51111313] mba_p
The R-square is: 0.030629529410807547
The mean square error of price and predicted value is: 8409560162.271541
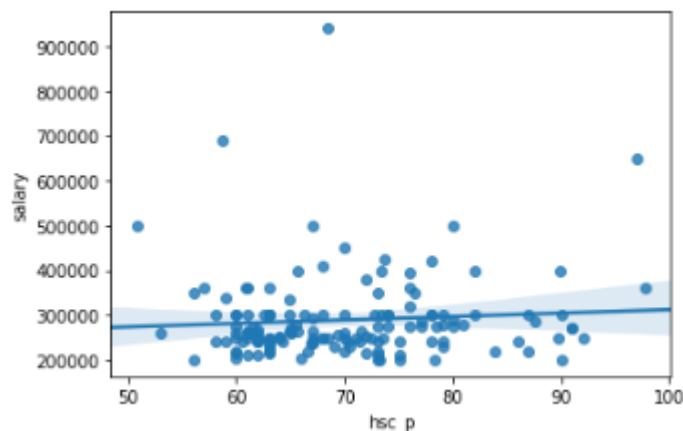
3.1.2 Correlation between Degree percentage and salary

The Pearson Correlation Coefficient is -0.0192722345765629 with a P-value of P=0.8161570666355349. As P-value is greater than 0.1; there is a moderate evidence that correlation is not significant. The linear regression equation is as follows Yhat= 307650.38245922566 + [-276.32859597] degree_p

The R-square is: 0.00037141190255738048

The mean square error of price and predicted value is: 8672057739.205791

### 3.1.3 Correlation between Higher secondary education percentage and salary

The Pearson Correlation Coefficient is 0.07681903294535551 with a P-value of P=0.3534116730065116. As P-value is greater than 0.1; there is a moderate evidence that correlation is not significant. The linear regression equation is as follows Yhat= 234843.62750858403 + [769.54711447] hsc_p

The R-square is: 0.005901163822659261

The mean square error of price and predicted value is: 8624085655.297733

### 3.1.4 Correlation between Secondary education percentage and salary
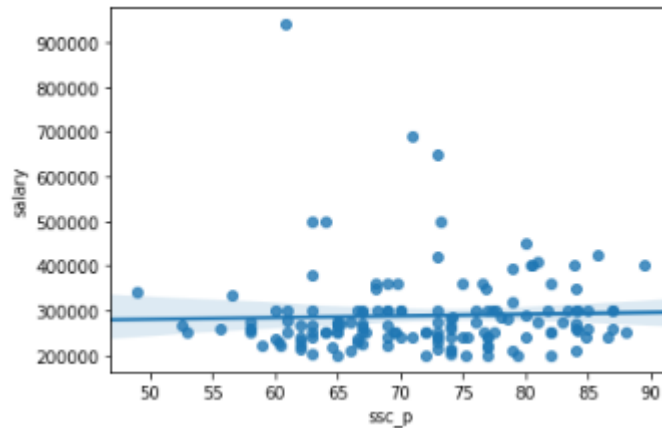
The Pearson Correlation Coefficient is 0.03533034131020021 with a P-value of P=0.6698885767074778. As P-value is greater than 0.1; there is a moderate evidence that correlation is not significant. The linear regression equation is as follows Yhat= 261483.4135696075 + [378.85427599] ssc_p

The R-square is: 0.0012482330170948686

The mean square error of price and predicted value is: 8664451132.406292

### 3.1.5 Correlation between Aptitude percentage and salary
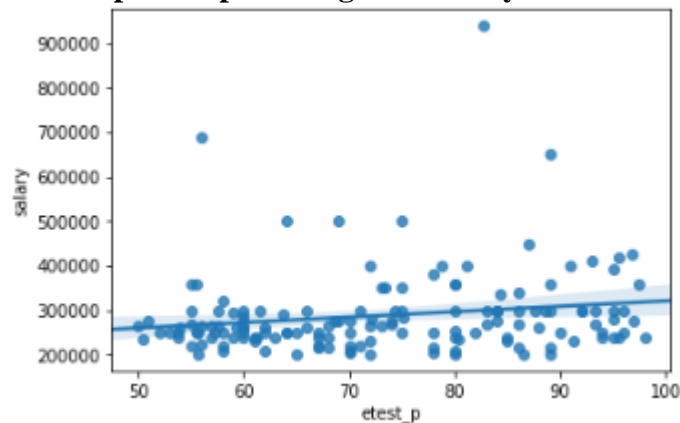
The Pearson Correlation Coefficient is 0.17830730897634872 with a P-value of P=0.030142277073780595. As P-value is less than 0.05; there is a moderate evidence that correlation is significant. The linear regression equation is as follows Yhat= 199761.83197581192 + [1213.76231223] etest_p

The R-square is: 0.03179349643438667

The mean square error of price and predicted value is: 8399462422.543881

3.1.6 Correlation between Multiple factors and salary

Independent variable used in Multiple linear regression are MBA, Undergrad, Higher secondary, Secondary and aptitude test percentage. The predicted value curve is fitted on actual value curve to know the efficiency of model. If we remove outliers from our dataset then efficiency of model will improve drastically.

Actual vs Fitted Values for salary
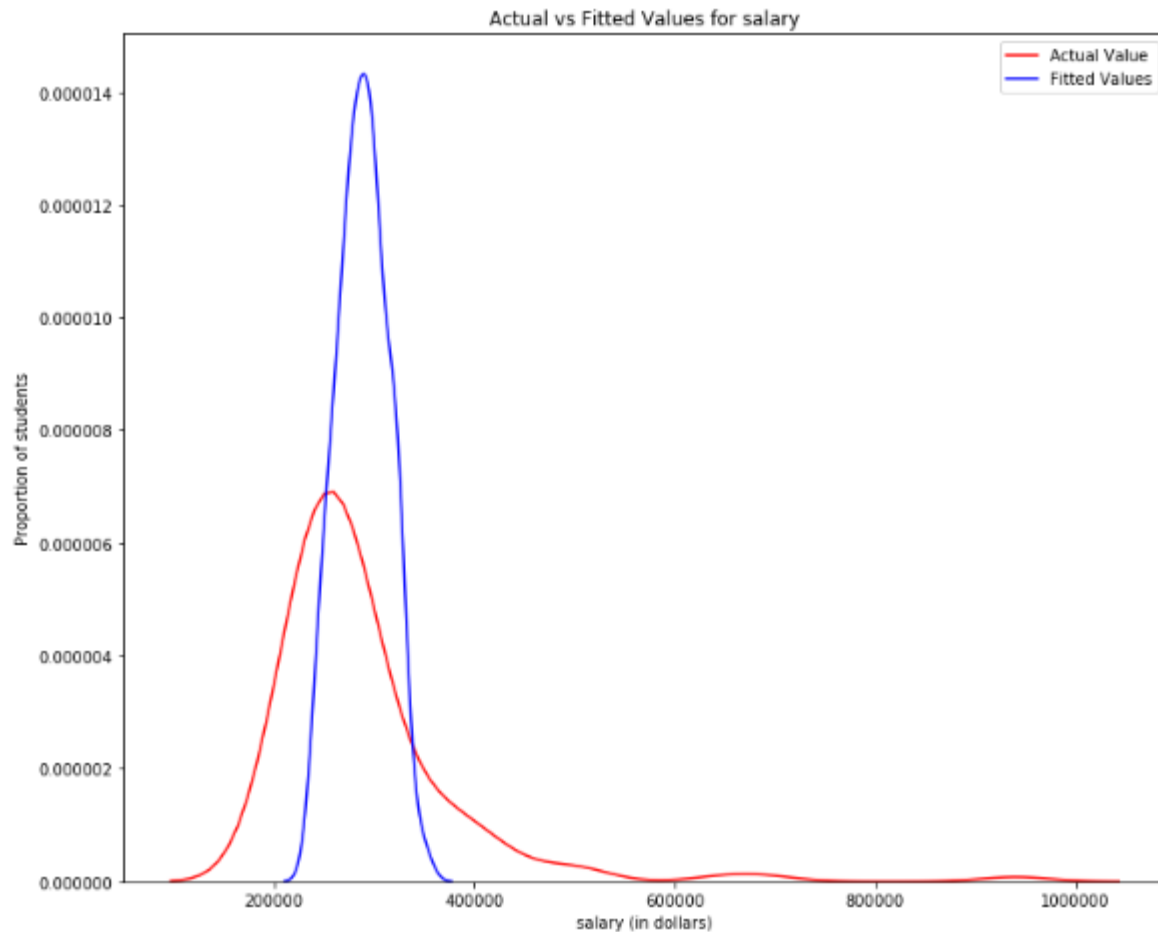
The R-square is: 0.06935803897750015
The mean square error of price and predicted value using multifit is: 8073579501.546181

3.2 Impact of categorical factors on salary

Categorical factor like choice of specialization and academic background plays an important role in estimating salaries offered after the graduation. We have proved this by using box plot for the various decision that students make in their academics and what are the returns on that decision.

3.2.1 Spread of salary based on MBA specialization

It have been observed that students with specialization in Marketing and Finance have been offered higher packages after the graduation as compared to student with specialization in Marketing and HR. There are lots of high amount of outlier for students with specialization Marketing and Finance.
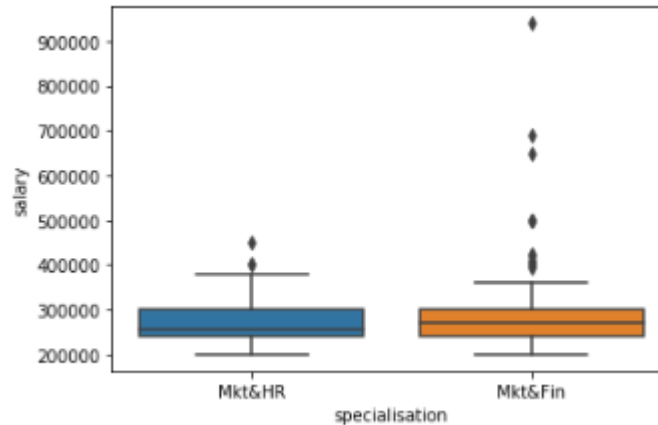
3.2.2 Spread of salary based on Undergraduate degree field

It have been observed that students with Science and tech background are generally offered highest packages and students with commerce and management background comes in second. While the interquartile range for Science & technology students is more the highest package is offered to student from Commerce and management background, this can be considered to be outlier.
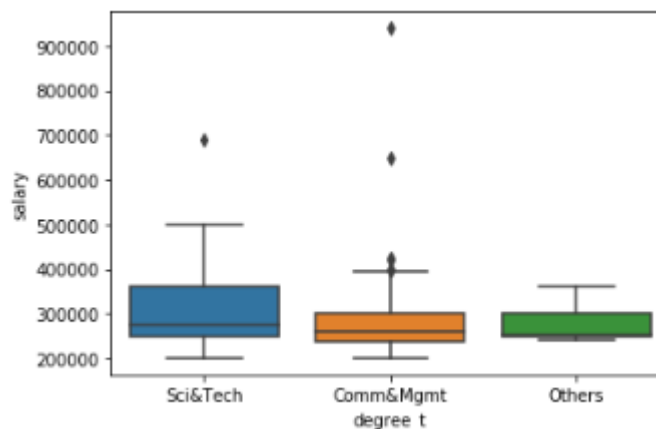

3.2.3 Spread of salary based on Work experience

It have been observed that students with some amount of work experience are preferred by recruiters over freshers. While the inter quartile range for students with or without work experience is more or less the same. The mean salary offered to students with work experience is more as compared to students with no work experience.
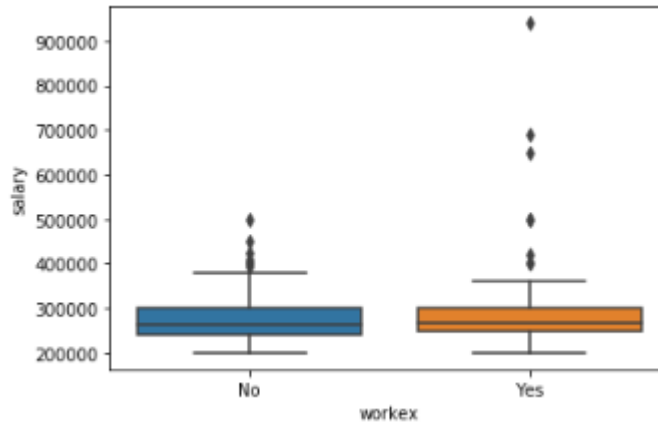
3.2.4 Spread of salary based on Higher Secondary Education

It have been observed that students with science background are offered higher packages with respect to students from Arts and commerce. The interquartile range for students with Arts background is very less and their average salary is also pretty less as compared to students from Science and Commerce background.
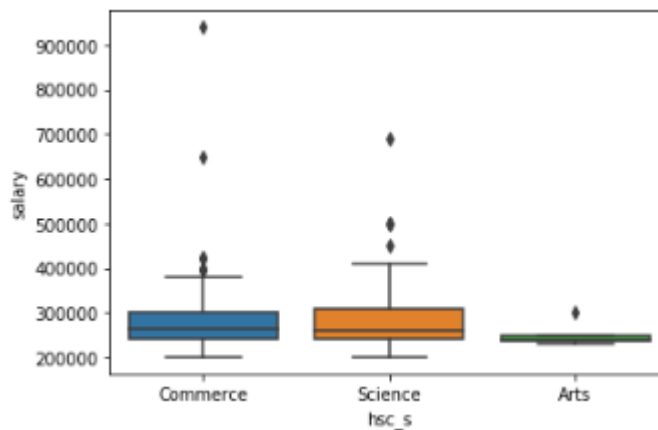
3.2.5 Correlation matrix

Following matrix helps us summarize the impact of different choices on salary offered to students after post-graduation. From the below two matrix, it can be safely concluded that student with Science and Technology background, having some amount of work experience and Completed MBA with Marketing and Finance specialization are generally offered the highest salaries.
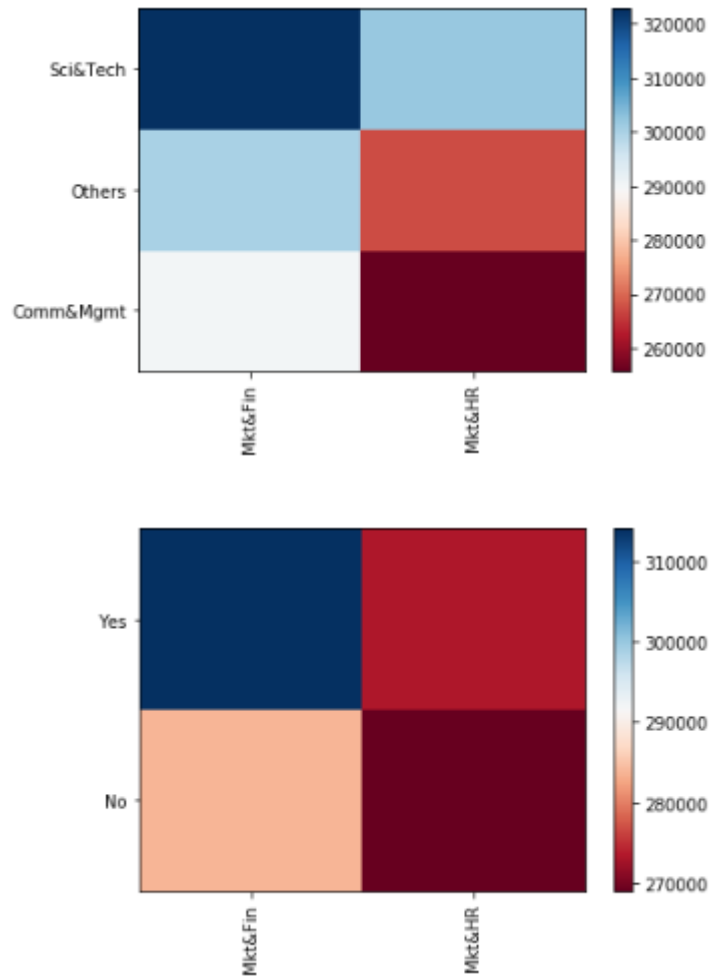
Figure 6. Scatter plot of player improvement and that of last season

4. Conclusions

In this study, I analyzed the relationship between student qualifications and the salary offered to them after ultimate stage of their educations. I built both regression model and correlation model to justify the findings for my project. The model can be really helpful to students in making career defining decision in many ways.