CSC-523

Machine Learning

Assignment - 1

Group 3

Multivariate Linear Regression Assignment

Group Members:

Dip Patel (201501070)
Heet Gorakhiya (201501034)
Kuldeep Jitiya (201501038)
Grishma Shah (201501095)

Problem Statement:

To predict or estimate salary of a person based on set of attributes of the available dataset and find the contributions for each particular parameters and provide a best fit and best features to represent the regression model.

Dataset Used:

https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/

List of available features in the dataset:

Sr. No.	Feature	Sr. No.	Feature
1	Age	16	Marital Status
2	Attrition	17	Monthly Income
3	Business Travel	18	Monthly Rate
4	Daily Rate	19	Num Companies Worked
5	Department	20	OverTime
6	Distance From Home	21	Percent Salary Hike
7	Education	22	Performance Rating
8	Education Field	23	Relationship Satisfaction
9	Environment Satisfaction	24	Stock Option Level
10	Gender	25	Total Working Years
11	Hourly Rate	26	Training Times Last Year
12	Job Involvement	27	Work Life Balance
13	Job Level	28	YearsAtCompany
14	Job Role	29	Years In Current Role

15	Job Satisfaction	30	Years Since Last Promotion
31	Standard Hours	32	Employee Count
33	EmployeeNumber	34	Over 18
35	Years With Current Manager		

Step 1: Selection of Important Features And Data Cleaning:

Out of the list of features, some features like

- "Standard Hours",
- "Employee Count",
- "Employee Number", and
- "Over 18"

have been removed, because either they have no effect whatsoever on the prediction, or they have redundant values (same value in all the cells like All employee are over 18.so,"'Over 18" column only contained value of 'Y').

Step 2: Univariate Case Analysis:

Non-Categorical Data:

The following are the R-Square scores of the discrete as well as ordinal, i.e., non-categorical parameters, (for linear regression model) when mapped with Monthly Income parameter:

Features	**R-Square Value**
JobLevel	0.90306992555980259
TotalWorkingYears	0.59736397010559661

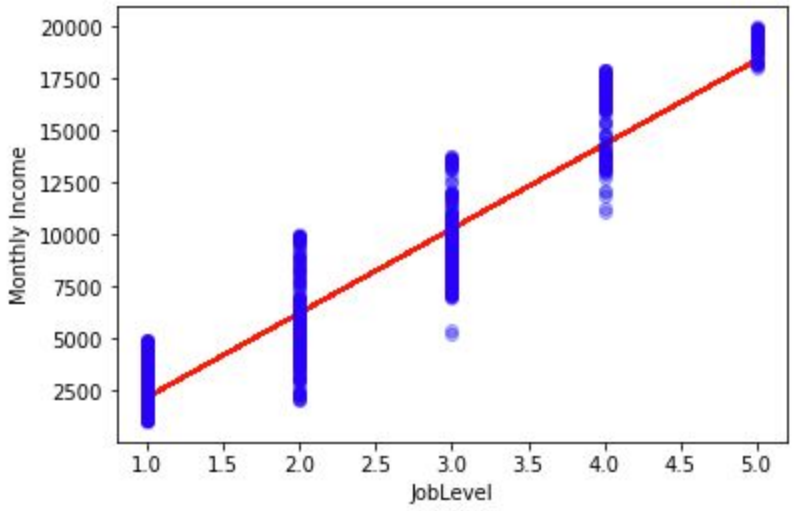
YearsAtCompany	0.26448888197942461
Age	0.24785916980965295
YearsInCurrentRole	0.13236329476944894
YearsSinceLastPromotion	0.11900957083419683
YearsWithCurrManager	0.11839027790456835
Attrition	0.02554869209700783
NumCompaniesWorked	0.022354799812446258
Education	0.0090175301842538857
MonthlyRate	0.0012119885646063588
WorkLifeBalance	0.00094145149382995896
PercentSalaryHike	0.0007435758064530118
RelationshipSatisfaction	0.0006694346975643084
TrainingTimesLastYear	0.00047246573016501703
PerformanceRating	0.00029309913326758252
DistanceFromHome	0.00028949132999012139
HourlyRate	0.00024946005087644885
JobInvolvement	0.0002332184306049756
Daily Rate	5.9398756456663371e-05
Job Satisfaction	5.1218961149013253e-05
OverTime	3.7079396714267972e-05
Stock OptionLevel	2.9242967257259167e-05

For the Univariate Analysis, the above table denotes the value of each parameter and its R square value. We can see that the R square value of job level is highest and so it is most contributing factor. We consider the top 6 influential features based on their R-Squared values,

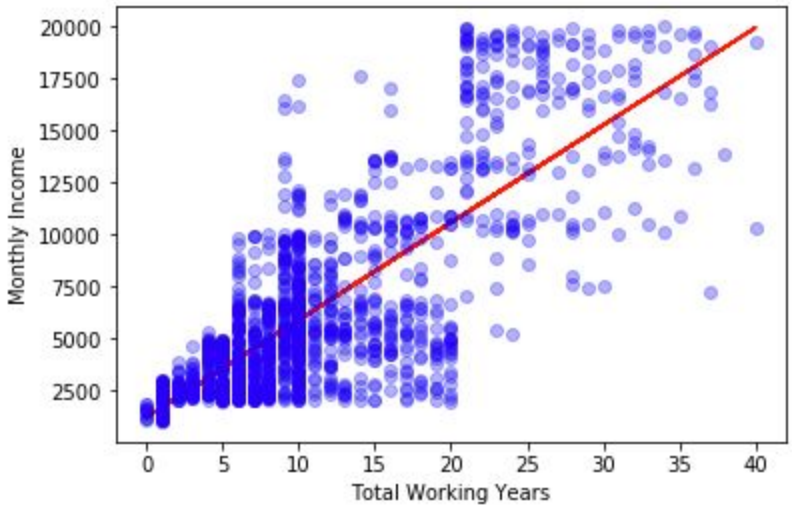
as they contribute most to the prediction of income, and the rest we discard as unimportant.The top 6 contributing factor are Job Level, Total Working Years,Years At Company, Age, Years In Current Role ,Years Since Last Promotion as there R square value is higher than the remaining factors. We can conclude from above observation that the factors like Job Involvement, Hourly Rate , Distance From Home , Relationship Satisfaction etc have the lowest value so they are the least contributing factor.

Plotting the Univariate graphs with **Linear polynomial** fitting for each of the selected features:

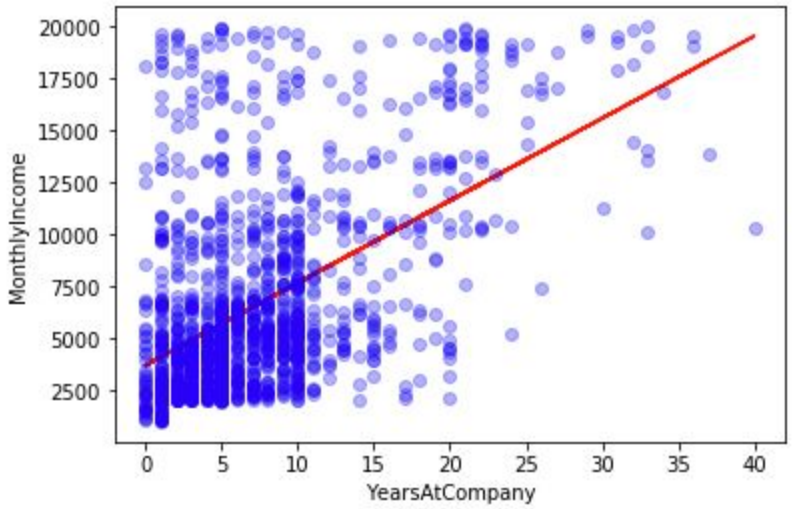
1. Job Level:



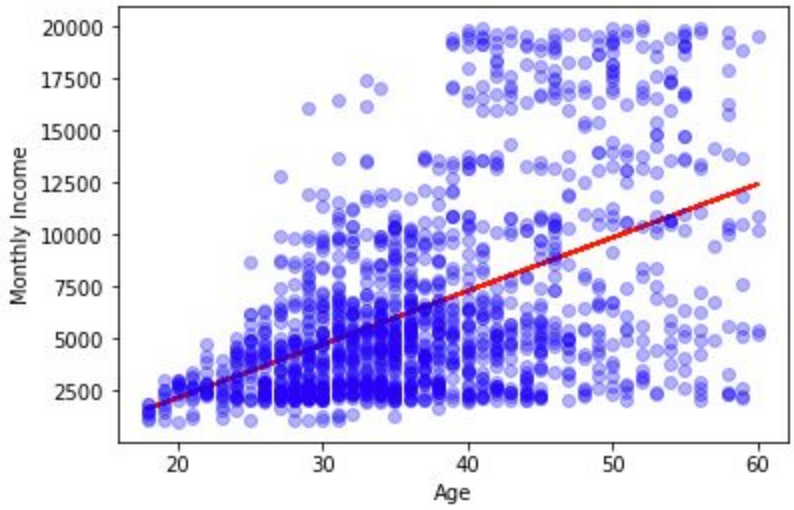
2. Total Working Years:



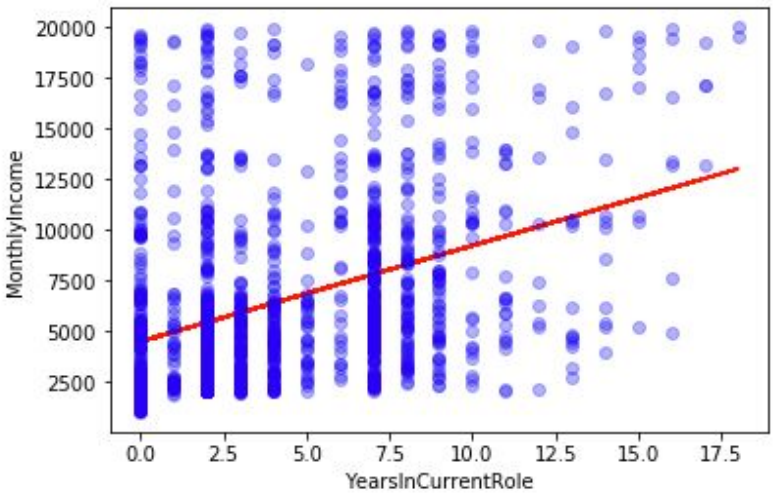
3. Years At Company:



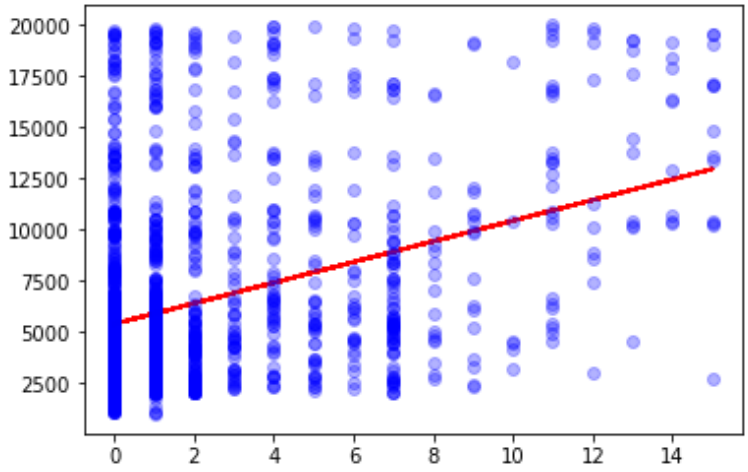
4. Age:



5. Years In Current Role:



6. Years Since Last Promotion:



Features Like “Number Of companies Worked”, “Education”, and “Attrition” have comparatively higher R-Squared Values, but it is evident from the above graphs that the lower the R-Squared value, the higher variance of outliers in the data. Hence, we consider the top 6 non-categorical features for prediction of monthly Income.

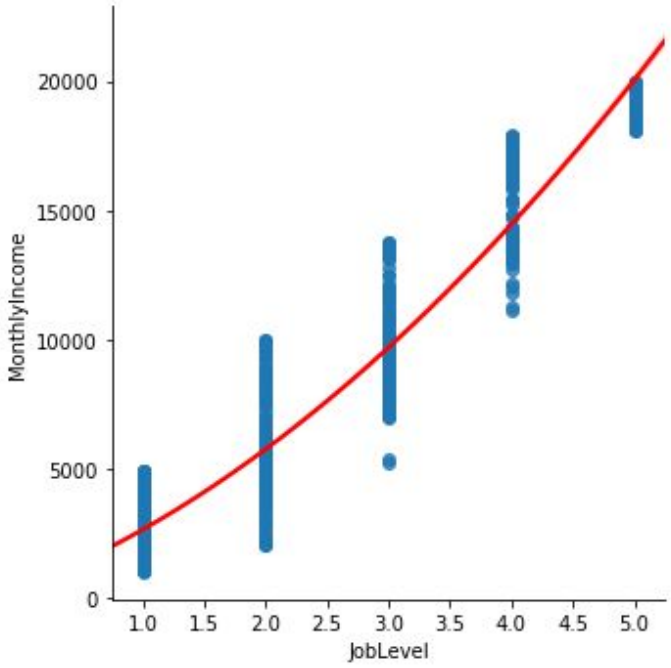
The R-Square values for a **2nd order polynomial** (ax^2 + bx + c) fitting are:

Feature	**R-Square Value**
JobLevel	0.91859927133836505
TotalWorkingYears	0.59830256115117564
YearsAtCompany	0.26826096056801529
Age	0.24809379104125007
YearsInCurrentRole	0.1372154850664524
YearsSinceLastPromotion	0.12308191260353218

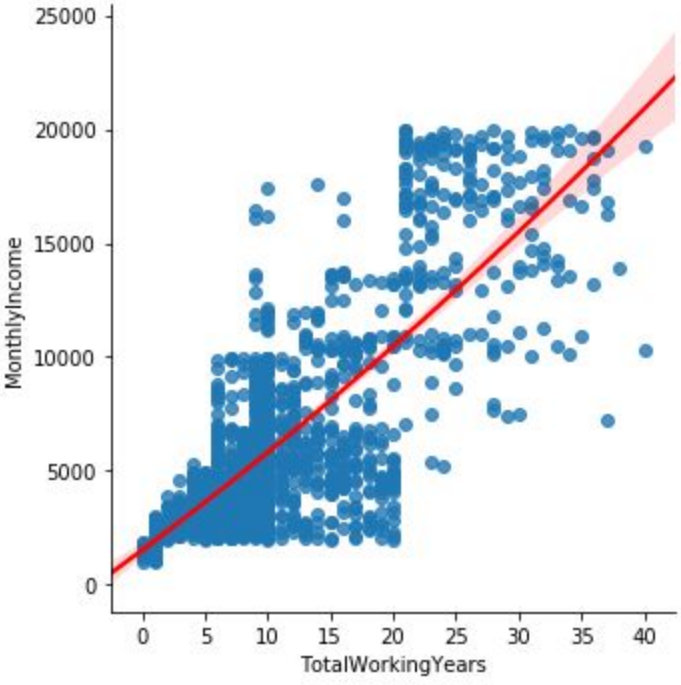
As we can see, the R-Square values are slightly more for each Feature. Hence, we can say that the polynomial of order 2 fits this dataset better than linear polynomial.

Plotting the Univariate by fitting a **2nd order polynomial** graph for each of the selected features:

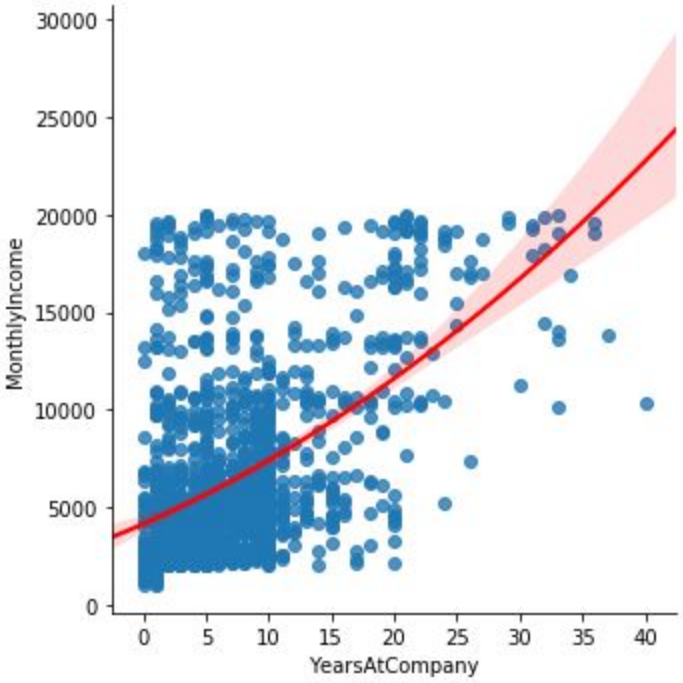
1. Job Level:



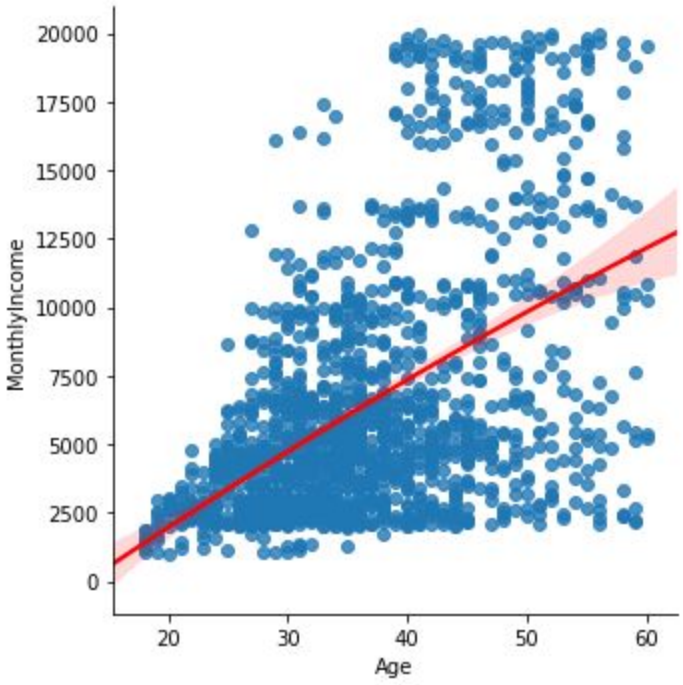
2. Total Working Years:



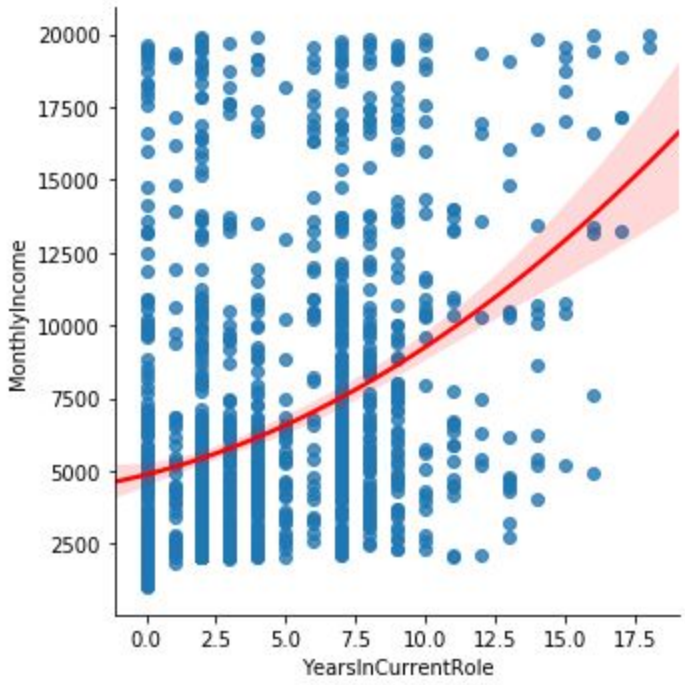
3. Years At Company:



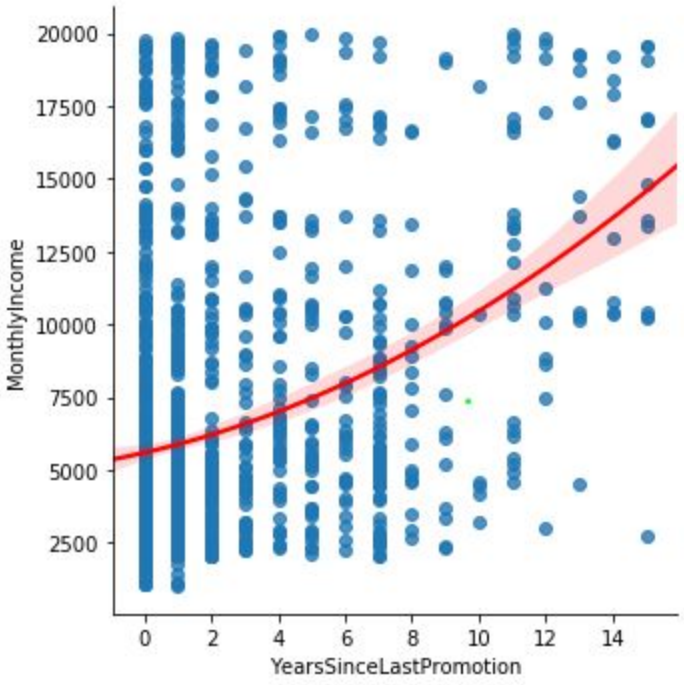
4. Age:



5. Years With Current Role:



6. Years Since Last Promotion:



The R-Square values for a **3rd order polynomial** (ax^3 + bx^2 + cx + d) fitting are:

Feature	R-Square Value
JobLevel	0.92436455282140728
TotalWorkingYears	0.60804037579673742
YearsAtCompany	0.27848170386960991
Age	0.25360658814343207
YearsInCurrentRole	0.13721787878047542
YearsSinceLastPromotion	0.12819665915221479

Here again, we can see that the R-Square values are slightly more for each Feature as compared to 2nd order polynomial. Hence, we can say that the polynomial of order 3 fits this dataset better than 2nd order polynomial.

At higher and higher order, there will be very miniscule improvement in the R-Square values. **For polynomial of order 4, the improvement is of the order of 10^(-3)**, and **for polynomial of order 4, the improvement is of the order of 10^(-5)**. Hence, only computational complexity increases without much improvement in R-Square values. So, it is **insignificant.**

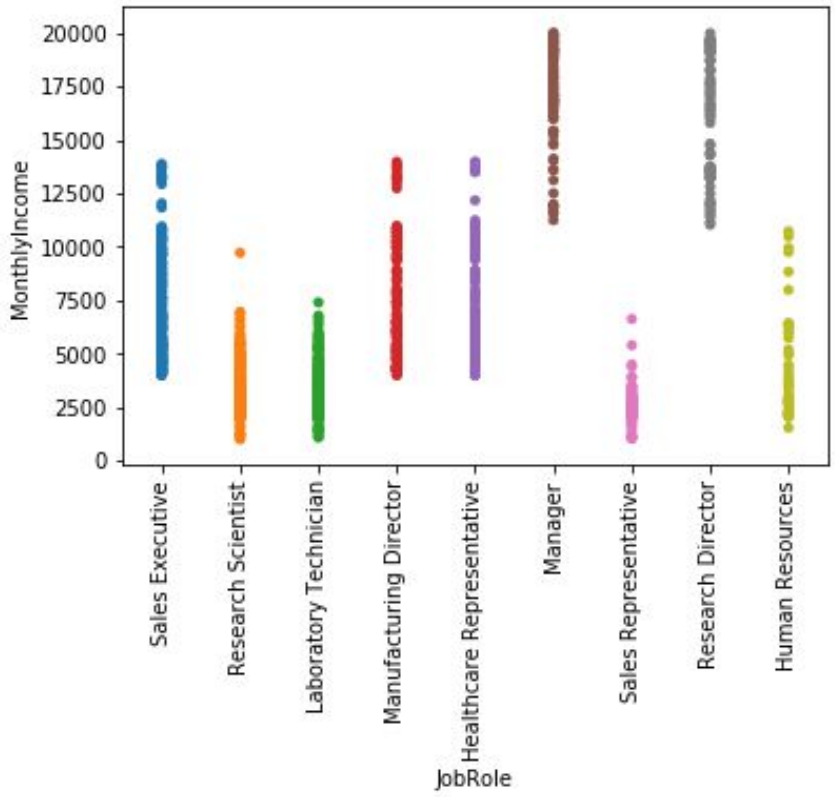
Categorical Data:

The following are the R-Squared Values of Categorical Parameters, when mapped with Monthly Income:

Feature	R-Square Value
JobRole	0.81598594166
MaritalStatus	0.00794013326889
EducationField	0.00660851782305
Department	0.00431533555628
BusinessTravel	0.00147531298601
Gender	0.0010149634995575418

From these R-Square values, we can see that only “JobRole” will have a significant impact on the prediction of Monthly Income in comparison to other parameters So, we can safely neglect all other categorical features since their contributions will be negligible.

Plotting the graph of Job Role vs Monthly Income:



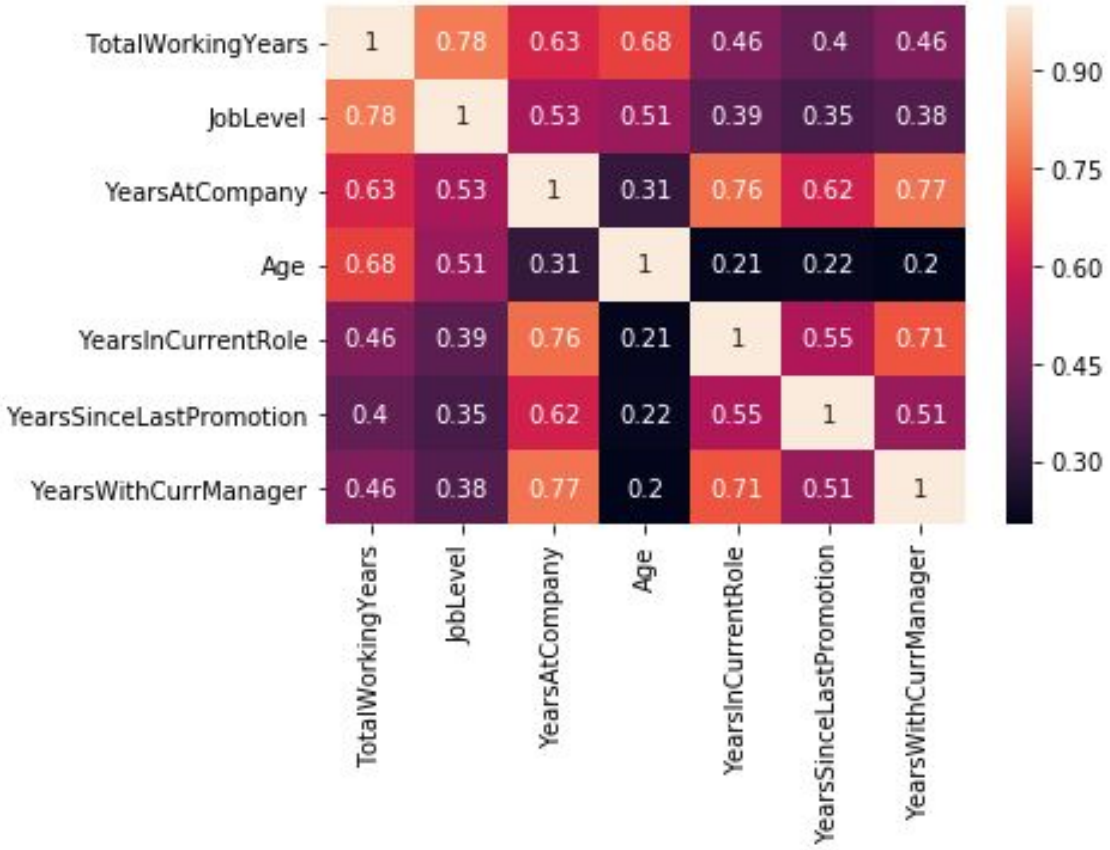
Step 3: Multivariate Case Analysis:

For the multivariate case analysis, we have to combine the most contributing features such that we can predict the value of Monthly income based on multiple features.

We can see from the analysis of univariate features that the most contributing features are:

- Job Level
- Total Working Years (Experience)
- Years at Company
- Age
- Years in Current Role
- Years Since Last Promotion

Given Below is the Covariance Matrix of the non-categorical features:



From this Matrix, we can see that individually, the features like “YearsAtCompany”and “Age” have high R-Square values but their correlation with the most dominant feature “JobLevel” is also high. Hence, these features will become redundant and not contribute much in the Multivariate Analysis.

And, in Categorical data, only one feature is significant:

- Job Role

Considering these all features, we calculate the R-Square value of the Multivariate regression model, which is computed to be: 0.943883623362

When we consider only two features, namely : “JobLevel” and “JobRole”, the R-Square value turns out to be: 0.941536578221

Analysing these two values, we can see that there is maximum Contribution of the two aforementioned features, while the impact of all other significant features combined is: 0.943883623362 - 0.941536578221 = 0.00234704514

Given Below is a table which shows the gradual but comparatively insignificant increase in the value of R-Square:

Features Used	**Value of R-Square in the Multivariate model**
1. JobLevel	0.90306992555980259
1. Joblevel 2. Job Role	0.941536578221
1. Joblevel 2. Job Role 3. TotalWorkingYears	0.943439686801
1. Joblevel 2. Job Role 3. TotalWorkingYears 4. YearsAtCompany	0.943445258559
1. Joblevel 2. Job Role 3. TotalWorkingYears 4. YearsAtCompany 5. Age	0.943471871985
1. Joblevel 2. Job Role 3. TotalWorkingYears 4. YearsAtCompany 5. Age 6. YearsInCurrentRole	0.943478141333
1. Joblevel 2. Job Role 3. TotalWorkingYears 4. YearsAtCompany 5. Age 6. YearsInCurrentRole 7. YearsSinceLastPromotion	0.943633226488

From the table we can conclude that the effect of adding more features to the multivariate regression will only be of the order of (-4). Hence, we conclude that a multivariate model of these 7 features best represent the regression.

Conclusion:

From the given analysis, we can say that for the best (optimal) prediction of Monthly Income, the following features should be incorporated in the regression analysis model:

1. Joblevel
2. Job Role
3. TotalWorkingYears
4. YearsAtCompany
5. Age
6. YearsInCurrentRole
7. YearsSinceLastPromotion

Optimal Features: Only **two features**, namely: "**JobLevel**" and "**JobRole**" are potentially sufficient as the cumulative of other 5 features provides only 0.00234704514 of R-square value improvement.
And, a **3rd degree polynomial** provides the best fit out of the three.

MontlyIncome = Y

JobLevel = x1
(below all variable take only 0 and 1)
JobRole_Healthcare Representative = x2
JobRole_Human Resources = x3
JobRole_Laboratory Technician = x4
JobRole_Manager = x5
JobRole_Manufacturing Director = x6
JobRole_Research Director = x7
JobRole_Research Scientist = x8
JobRole_Sales Executive = x9
JobRole_Sales Representative = x10

Y = 3048.75114639***x1** + -644.48706233***x2** + -911.57462828***x3** + -1174.22587739***x4** + 3427.26291582***x5** + -801.87364856***x6** + 3281.93645466***x7** + -1057.62106851***x8** + -806.71460994***x9** + -1312.70247548***x10** + 632.827738422

The last value: 632.827738422 denotes the intercept.