Selection of better Regression Model using R

Name: Abhishek K Singh

Project Description

- Your Company is going to hire a new employee who is having a total experience of 20 years with 2 years of working experience as “Region Manager”.
- Jump from “Region Manager(level-6)” to “Partner(level-7)” level require 4 years
- He is claiming his salary to be “160000”
- As per Your company rule he should be paid as Level=6.5 position.
- But there is not any specific salary criteria for Level=6.5 position.
- Question: Find out he is speaking truth or bluffing about his salary? & Can be hired on that claimed package or not?

* Data Available for Positional Salary

Position	Level	Salary
Business Analyst	1	45000
Junior Consultant	2	50000
Senior Consultant	3	60000
Manager	4	80000
Country Manager	5	110000
Region Manager	6	150000
Partner	7	200000
Senior Partner	8	300000
C-level	9	500000
CEO	10	1000000

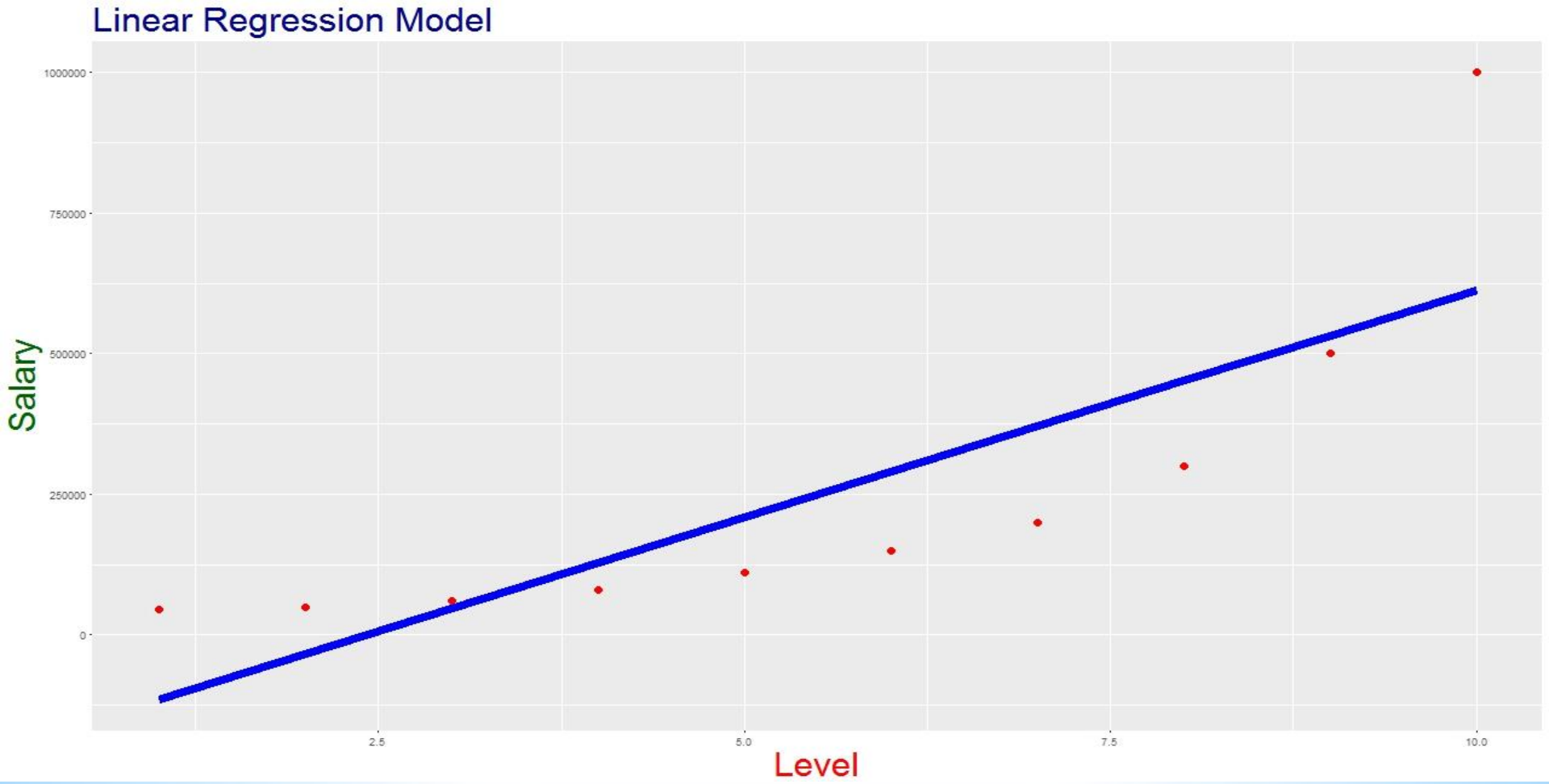
Linear Regression Output

```
Call:
lm(formula = Salary ~ ., data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-170818 -129720  -40379   65856  386545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -195333     124790  -1.565  0.15615
Level          80879      20112   4.021  0.00383 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 182700 on 8 degrees of freedom
Multiple R-squared:  0.669,     Adjusted R-squared:  0.6277
F-statistic: 16.17 on 1 and 8 DF,  p-value: 0.003833

>
```

Pr=0.00383 > 0.001, approximately equal

“Level” is significant to “Salary”, Lets see what graph plot says in next slide.

* Plot of Linear regression model



We can clearly conclude from above graph that although “Level” is significant to “Salary”, But they are not properly correlated &

Linear regression model is not a perfect model for this dataset

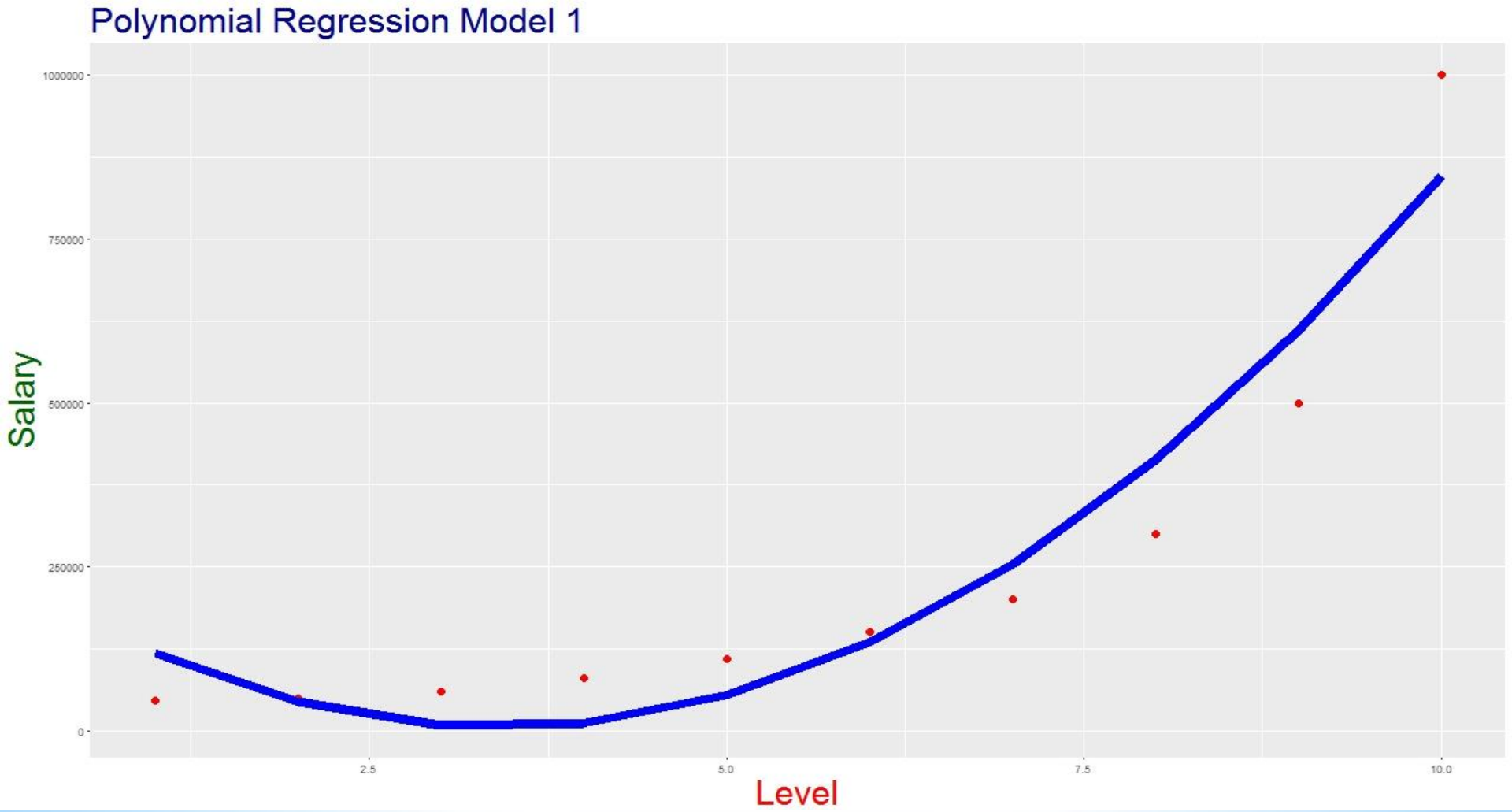
* Polynomial regression model (Level^2)

```
Call:
lm(formula = Salary ~ ., data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-112833  -68852   10682   55186  153364

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   232167     115571   2.009  0.08451 .
Level        -132871      48268  -2.753  0.02839 *
level2         19432       4276   4.544  0.00265 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 98260 on 7 degrees of freedom
Multiple R-squared:  0.9162,    Adjusted R-squared:  0.8923
F-statistic: 38.27 on 2 and 7 DF,  p-value: 0.0001703
```

Pr(Level)=0.0283 > pr(level^2)= 0.0026, that means

“Level^2” is more significant to “Salary”, than Level

* Plot of Polynomial regression model (Level^2)



We can clearly conclude from above graph that although “Level^2” is more significant to “Salary” than “Level”.

But, this model is sill not a perfect model for this dataset

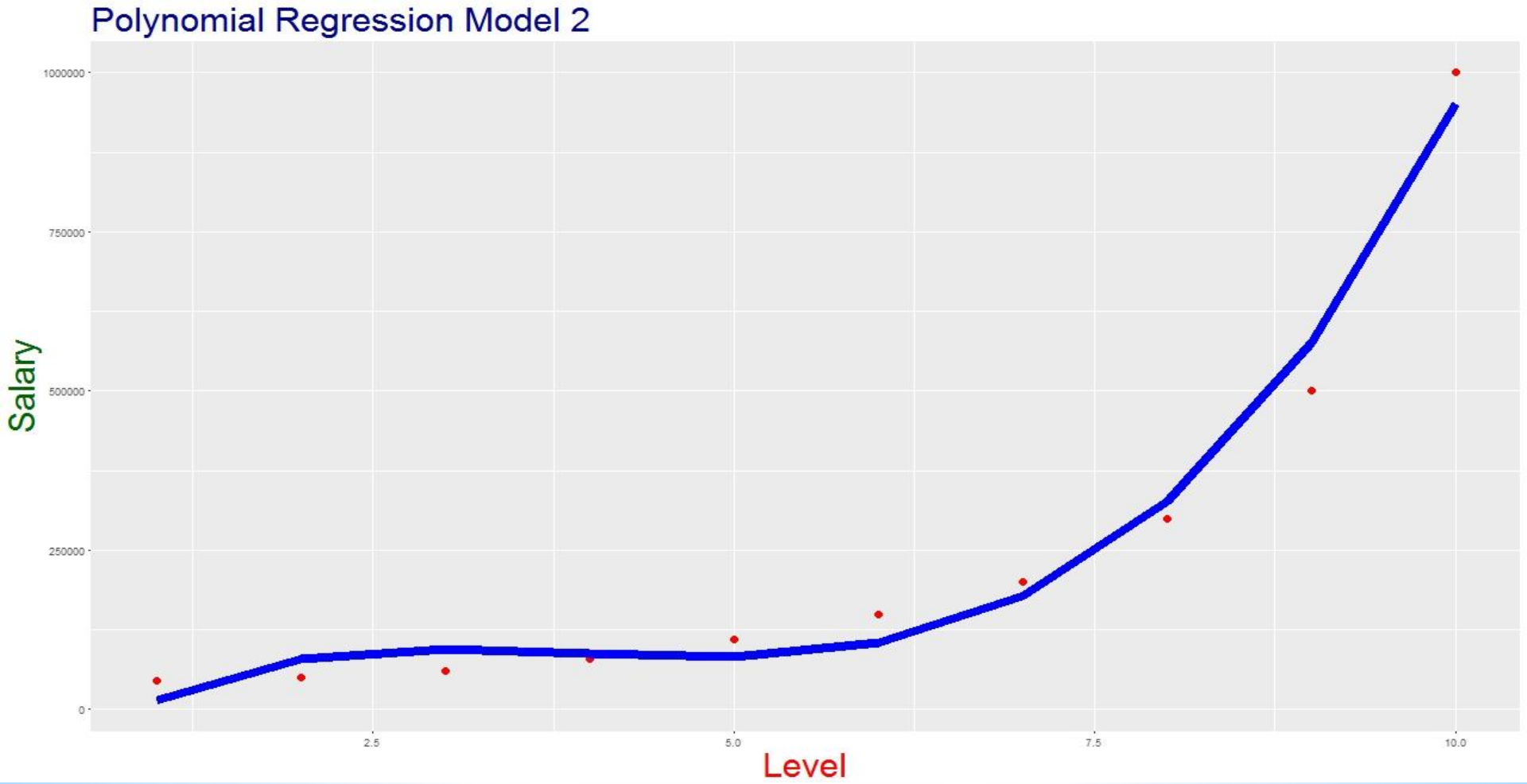
* Polynomial regression model (Level^3)

```
Call:
lm(formula = Salary ~ ., data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-75695 -28148   7091  29256  49538

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -121333.3    97544.8  -1.244  0.25994
Level        180664.3    73114.5   2.471  0.04839 *
level2       -48549.0    15081.0  -3.219  0.01816 *
level3         4120.0      904.3   4.556  0.00387 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50260 on 6 degrees of freedom
Multiple R-squared:  0.9812,    Adjusted R-squared:  0.9718
F-statistic: 104.4 on 3 and 6 DF,  p-value: 1.441e-05
```

Pr(Level^2)=0.0182 > pr(level^3)= 0.0038, that means

“Level^3” is more significant to “Salary”, than Level^2

* Plot of Polynomial regression model (Level^3)



We can clearly conclude from above graph that “Level^3” is more closer to “Salary” than “Level^2”.

This model is better than “Level^2” model for this dataset.

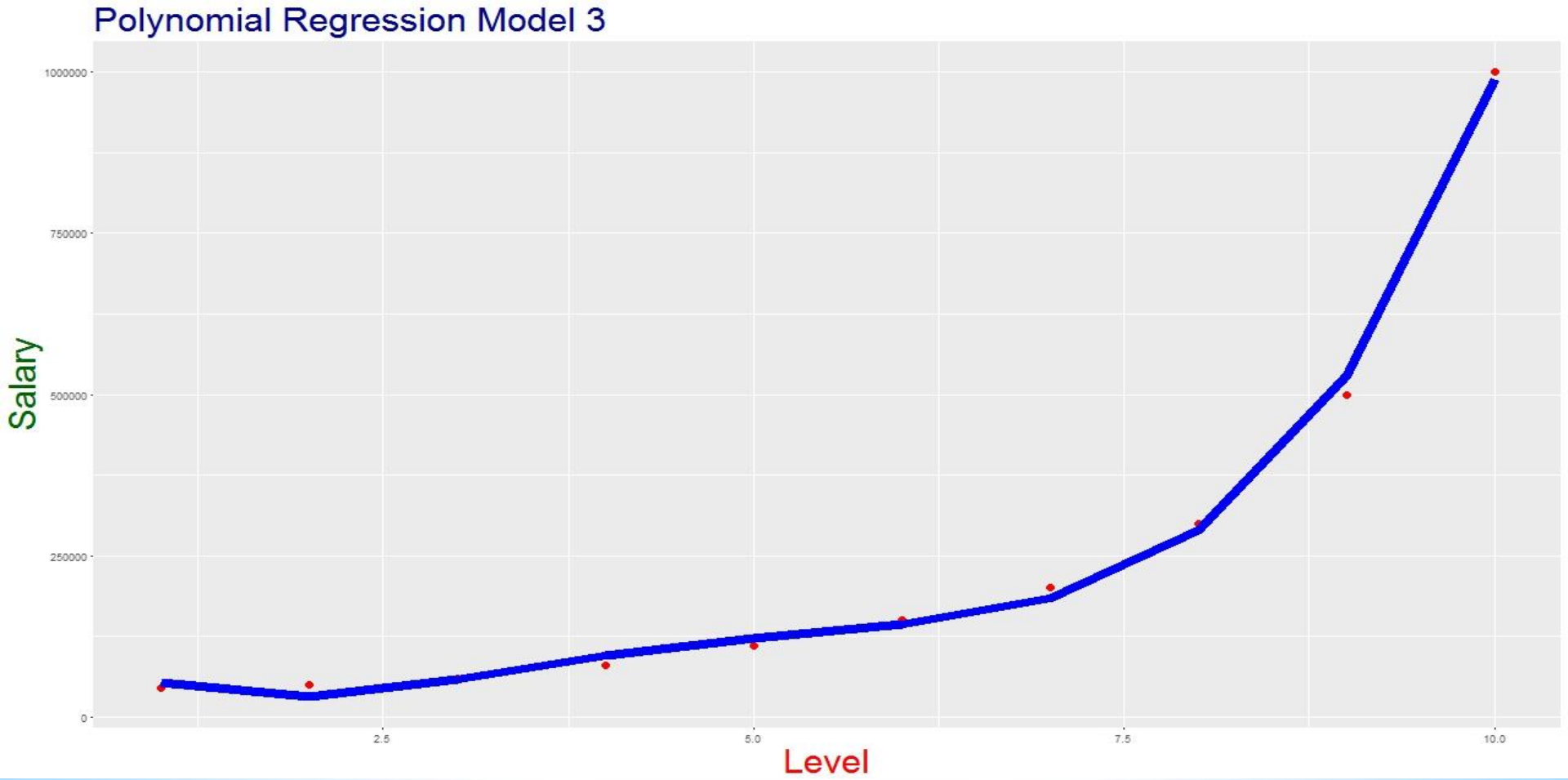
* Polynomial regression model (Level^3)

```
Call:
lm(formula = Salary ~ ., data = dataset)

Residuals:
     1      2      3      4      5      6      7      8      9     10
 -8357  18240   1358 -14633 -11725   6725  15997  10006 -28695  11084

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  184166.7    67768.0   2.718  0.04189 *
Level       -211002.3    76382.2  -2.762  0.03972 *
level2        94765.4    26454.2   3.582  0.01584 *
level3       -15463.3     3535.0  -4.374  0.00719 **
level4          890.2      159.8   5.570  0.00257 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20510 on 5 degrees of freedom
Multiple R-squared:  0.9974,    Adjusted R-squared:  0.9953
F-statistic: 478.1 on 4 and 5 DF,  p-value: 1.213e-06
```

Pr(Level^3)=0.0071 > pr(level^4)= 0.0025, that means

“Level^4” is more significant to “Salary”, than Level^3

* Plot of Polynomial regression model (Level^4)



We can clearly conclude from above graph that “Level^4” is closest to “Salary” than other models.

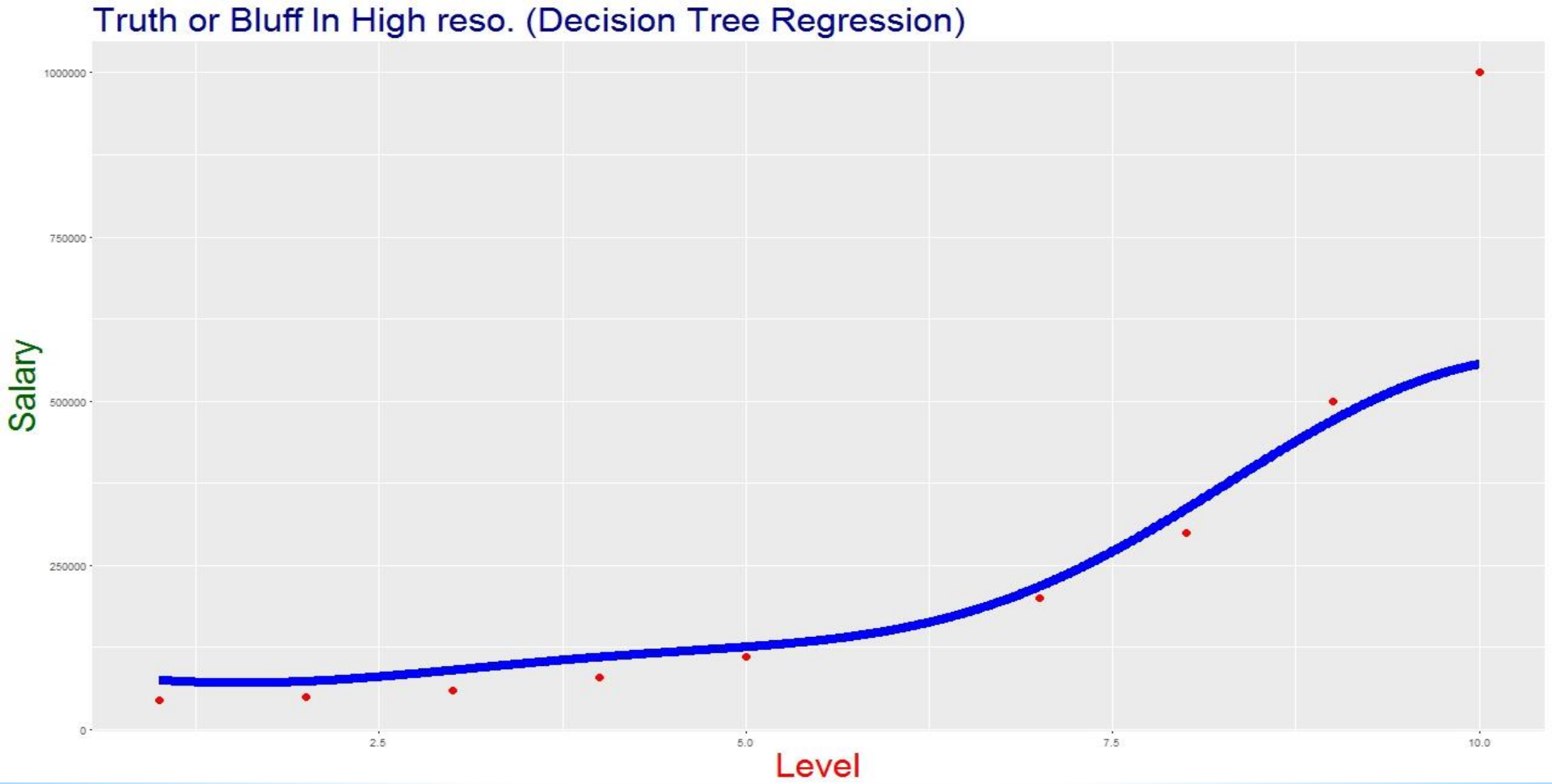
This model is till now best model for this dataset.

* Plot of Support Vector regression(SVR) model

We can clearly conclude from above graph that “SVR” model is closer to “Salary” for the lower value of “Level” Variable.

But, this model is not best model for this dataset when “Level” Variable is High.

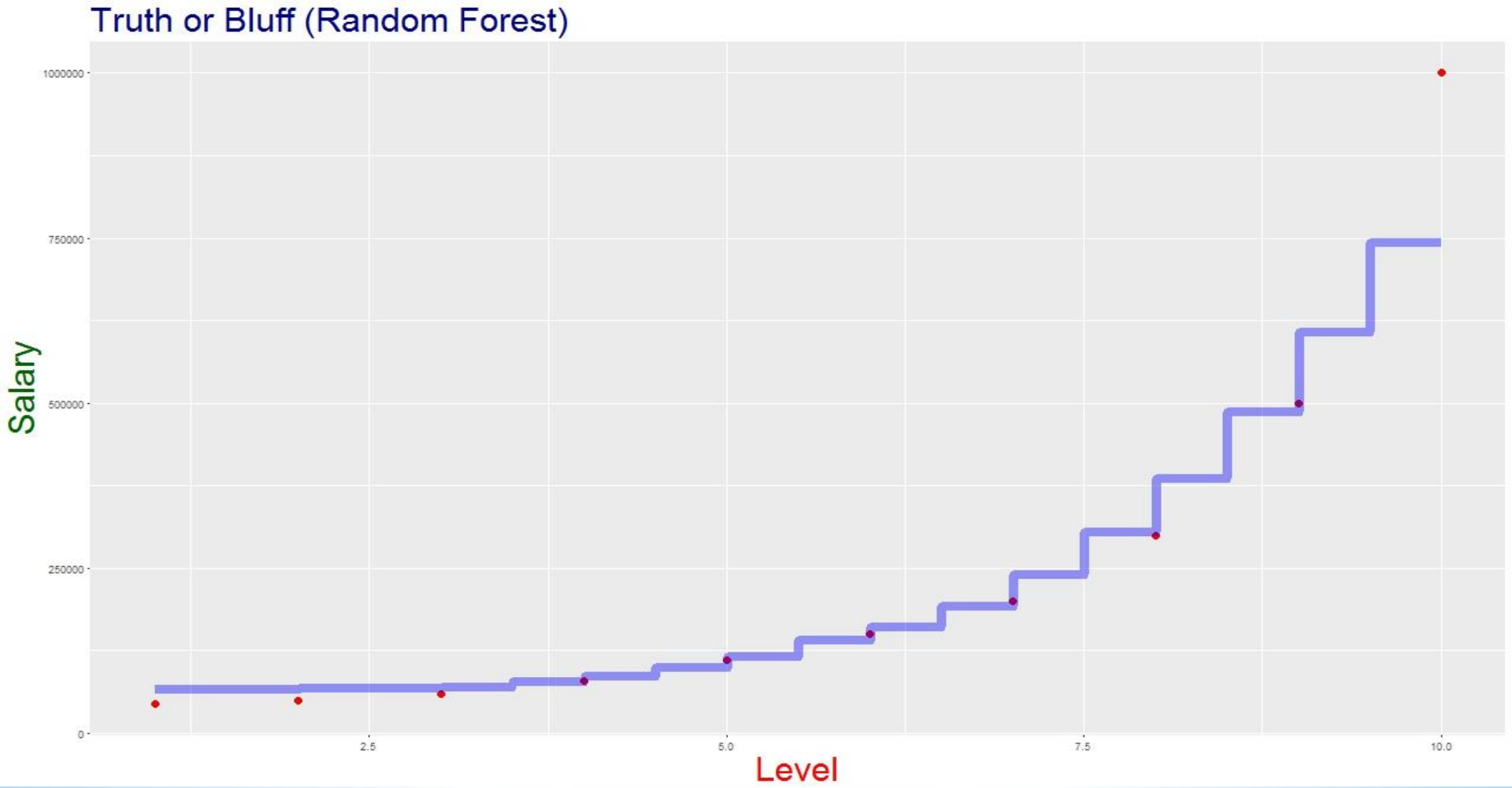
* Plot of Decision Tree regression model



We can clearly conclude from above graph that “Decision Tree” model is closer to “Salary” for the lower value of “Level” Variable.

But, this model is not a good model for this dataset when “Level” Variable is High.

* Plot of Random Forest regression model



We can clearly conclude from above graph that “Random Forest” model is closest to “Salary” for the lower value of “Level” Variable.

But, this model is not a good model for this dataset when “Level” Variable is High.

Prediction Of Salary as per different Regression model for "Level= 6.5"

S.N	Modelling_Type	Level	Salary_Prediction
1	Linear Regression	6.5	330378.7879
2	Polynomial Reg. model1	6.5	189498.1061
3	Polynomial Reg. model2	6.5	133259.4697
4	Polynomial Reg. model3	6.5	158862.4527
5	SVR Model	6.5	177858.6875
6	Decision Tree Regression	6.5	177858.6875
7	Random forest	6.5	162115.5

*Conclusion & Result

- As per different Model graph plot we can conclude that Polynomial Regression Level^4 model prediction was closest to actual “Salary” prediction & is Best suitable model for the given dataset.
- As per previous prediction slide “Salary” for Level=6.5, on the basis of Polynomial Regression Level^4 = 158862
- The new employee claimed his previous Salary = 160000
- Actual Claimed Salary (160000) ≈ Predicated Salary(158862)

Result: Claim about Salary of new employee is “True” & He can be hired on that package.