Leveraging Machine Learning Framework to Predict Cost of Treatment

Master of Professional Studies in Informatics

Healthcare Analytics

Submitted by: - Swapnesh Tiwari

Date: 12/01/2022

TABLE OF CONTENTS

TABLE OF FIGURES

Introduction

In this report we will be analyzing the medical cost dataset, which consists of Age, Gender, Tobacco use in yes or no, BMI which is a calculated variable of height and weight, Area, and Cost.

In this report, we will describe the data, form a linear regression of different groups such as males and females, predict the outcome, and provide metrics of accuracy or visualization. We will be using cost as a dependent variable on other variables, by doing so we will be predicting cost as a function of age, gender, BMI, and tobacco. Some of the variables in the medical cost dataset are located below:

1. "Age": the beneficiary's age.
2. "sex": female or male
3. "BMI" stands for "Body Mass Index," which is an objective measure of body weight in meters/height in kg based on the ratio of height to weight, with a range of 18.5 to 24.9 being optimal.
4. "Children": the total number of dependents and children with health insurance
5. "Tobacco": whether a tobacco consumer
6. "area" refers to the beneficiary's home state's northeast, southeast, southwest, or northwest.
7. "cost": Individual medical expenses that health insurance companies bill

Analysis

Exploratory Analysis

Table 1 shows a descriptive analysis of the whole dataset

	Age	**Gender***	**BMI**	**Children**	**Tobacco***	**Area***	**Cost**
vars	1	2	3	4	5	6	7
n	1338	1338	1338	1338	1338	1338	1338
mean	39.21	1.51	30.66	1.09	1.2	2.52	13270.42
sd	14.05	0.5	6.1	1.21	0.4	1.1	12110.01
min	18	1	15.96	0	1	1	1121.87
max	64	2	53.13	5	2	4	63770.43
range	46	1	37.17	5	1	3	62648.55
se	0.38	0.01	0.17	0.03	0.01	0.03	331.07

Table 2 shows a descriptive analysis of the male population

	Age	Gender*	BMI	Children	Tobacco*	Area*	Cost
vars	1	2	3	4	5	6	7
n	676	676	676	676	676	676	676
mean	38.92	1	30.94	1.12	1.24	2.52	13956.75
sd	14.05	0	6.14	1.22	0.42	1.1	12971.03
min	18	1	15.96	0	1	1	1121.87
max	64	1	53.13	5	2	4	62592.87
range	46	0	37.17	5	1	3	61471
se	0.54	0	0.24	0.05	0.02	0.04	498.89

By comparing table 1 and table 2 we can see the sample size differs, the actual whole dataset sample size is 1338 observations, whereas for males the dataset size is 676 observations, summary table gives us an important point to note the BMI, highest among the population for males screened for this dataset was 53.13, whereas the highest cost for male is 62592.87.

Table 3 shows a descriptive analysis of the female population

	Age	Gender*	BMI	Children	Tobacco*	Area*	Cost
vars	1	2	3	4	5	6	7
n	662	662	662	662	662	662	662
mean	39.5	1	30.38	1.07	1.17	2.51	12569.58
sd	14.05	0	6.05	1.19	0.38	1.11	11128.7
min	18	1	16.82	0	1	1	1607.51
max	64	1	48.07	5	2	4	63770.43
range	46	0	31.25	5	1	3	62162.92
se	0.55	0	0.23	0.05	0.01	0.04	432.53

From this table, the sample size of the female population in this dataset is 662 observations, whereas the highest BMI for the female patient is 48.07, and the highest cost is around 63770.43.

Comparing male and female statistics: By comparing two different populations it is clear that males have higher BMI levels than females, but at the same time it is interesting to note that females have a higher cost of medical expenses than the male population.

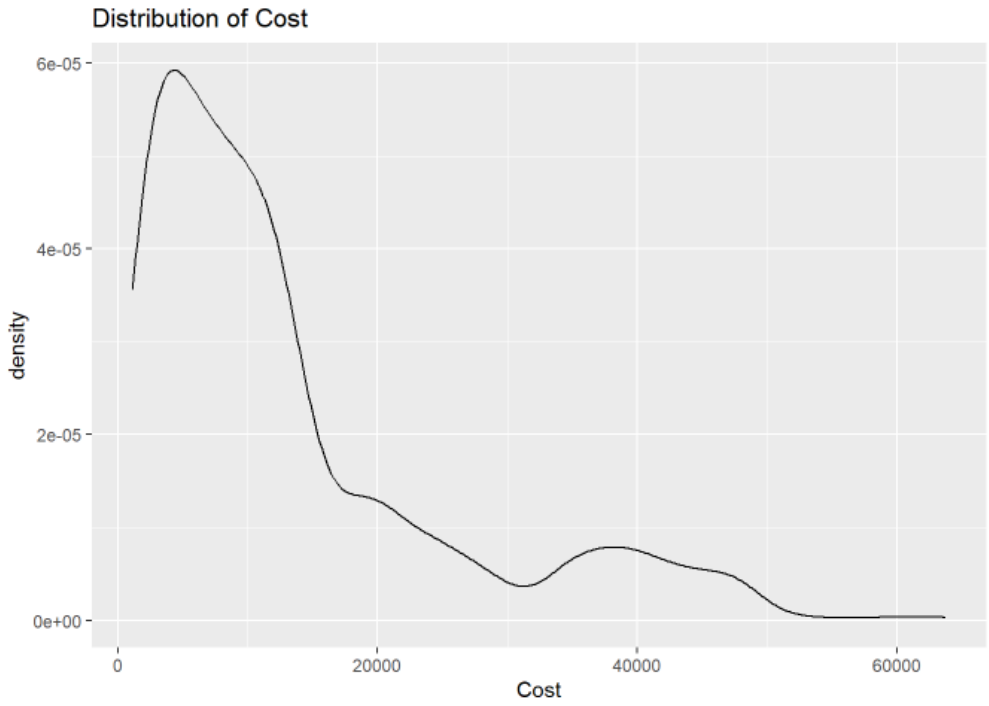

Figure 1 shows the distribution of response variable costs.

In figure 1 the distribution is basically right skewed with a deep depression or tail to the right side. There's a bump at around $30,000 - $40,000, maybe it can be another hidden distribution. To check we need to see other visualizations with other categorical variables.

There are multiple graphs, but I will be including only the one relevant to this report

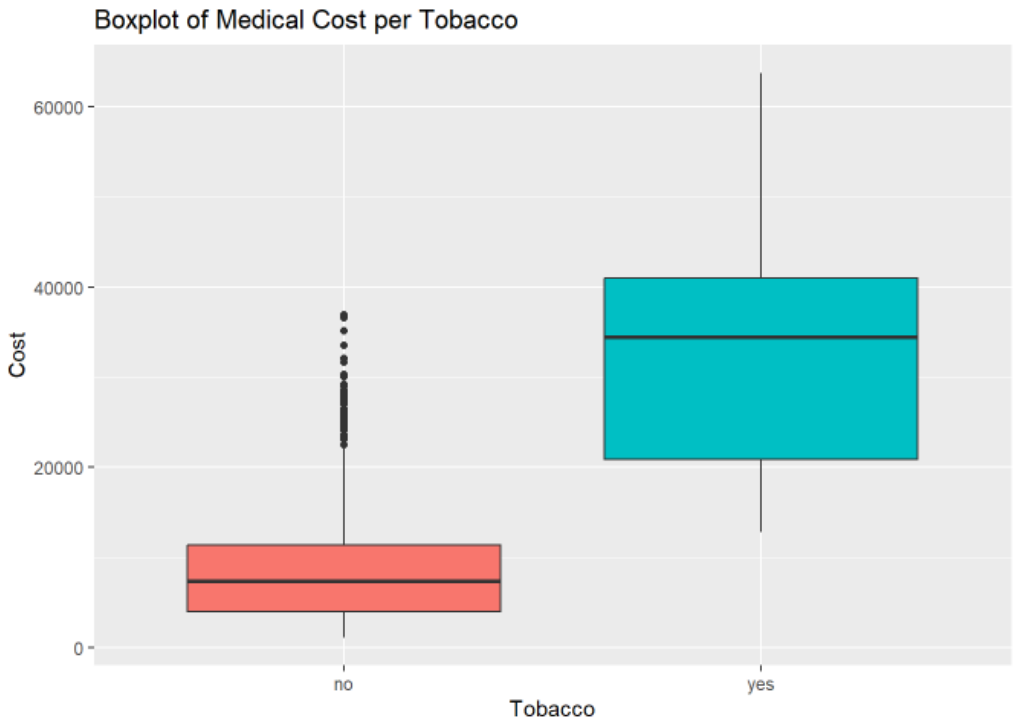

Figure 2 shows the medical cost per tobacco consumption.

Since tobacco consumption has a high influence on the medical cost of patients, we can see that In the following graph tobacco consumption seems to make a significant difference in medical costs.

Since the above plots suggest that tobacco consumption can have a big influence on hospital costs so we will plot the distribution of Costs for tobacco consumption

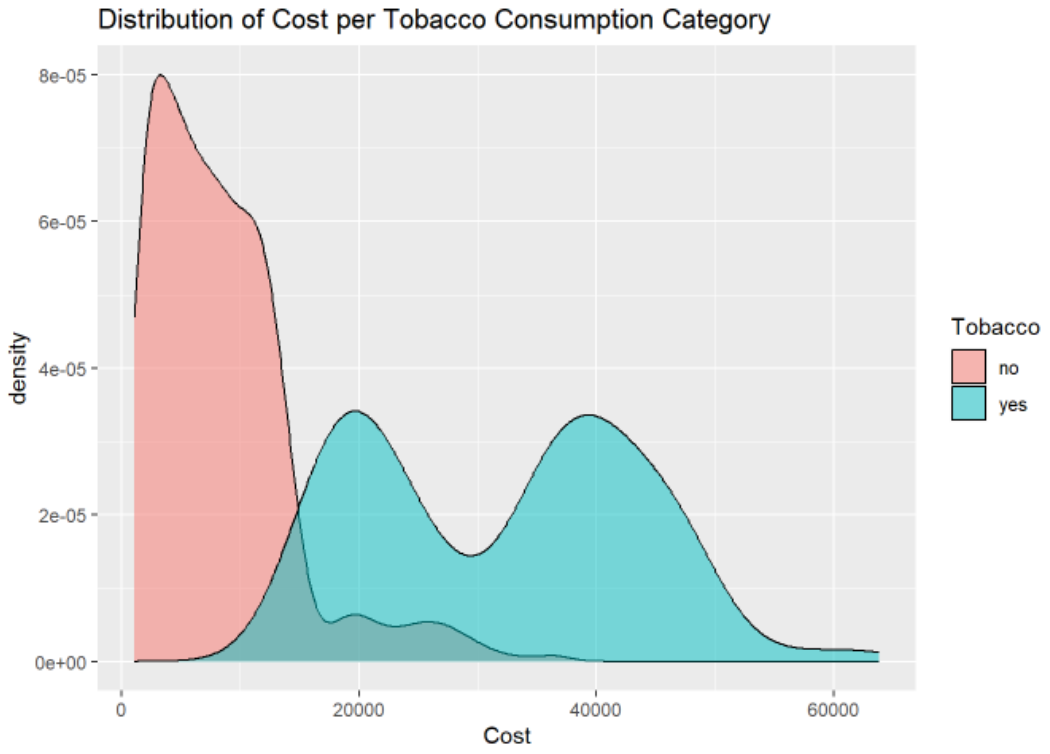

Figure 3 shows the distribution of tobacco consumption.

Therefore, proven that medical cost is very high among patients who consume tobacco products.

Before we can create model predictions it is important to check the correlation between variables.

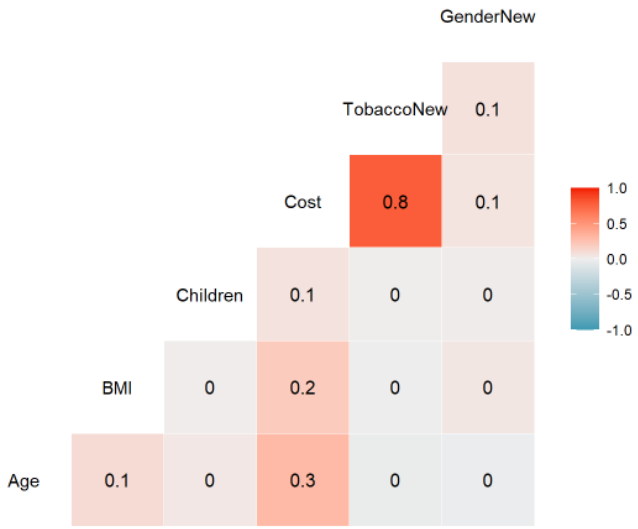

Figure 4 shows a correlation between different variables.

Therefore, now we know tobacco consumption and medical cost is highly related.

Predictive Analysis

a. Model 1 with age on cost.

Call:

lm(formula = Cost ~ Age, data = DF1)

Table 4 shows the residuals of model 1:

Min	1Q	Median	3Q	Max
-8059	-6671	-5939	5440	47829

Table 5 shows the coefficients of model 1:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3165.9	937.1	3.378	0.000751 ***
Age	257.7	22.5	11.453	< 2e-16 ***

F-statistic: 131.2 on 1 and 1336 DF, p-value: < 2.2e-16

b. Model 2 with age and gender on cost

Call:

lm(formula = Cost ~ Age + Gender, data = DF1)

Table 6 shows the residuals of model 2:

Min	1Q	Median	3Q	Max
-8821	-6947	-5511	5443	48203

Table 7 shows the Coefficients of model 2:

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	2343.62	994.35	2.357	0.0186 *
Age	258.87	22.47	11.523	<2e-16 ***
GenderM	1538.83	631.08	2.438	0.0149 *

F-statistic: 68.8 on 2 and 1335 DF, p-value: < 2.2e-16

c. Model 3 with age, tobacco, gender on cost.

Call:

lm(formula = Cost ~ Age + Tobacco + Gender, data = DF1)

Table 8 shows the residuals of model 3:

Min	1Q	Median	3Q	Max
-16122.4	-2048.5	-1318.9	-228.2	28725.3

Table 9 shows the Coefficients of model 3:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2433.56	558.26	-4.359	1.41e-05 ***
Age	274.93	12.46	22.061	< 2e-16 ***
Tobaccoyes	23847.63	434.89	54.836	< 2e-16 ***
GenderM	81.82	350.98	0.233	0.816

F-statistic: 1151 on 3 and 1334 DF, p-value: < 2.2e-16

Call:

lm(formula = Cost ~ Age + Gender + Tobacco + BMI, data = DF1)

Table 10 shows the residuals of model 4:

Min	1Q	Median	3Q	Max
-12364.7	-2972.2	-983.2	1475.8	29018.3

Table 11 shows the residuals of model 4:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-11633.49	947.27	-12.281	<2e-16 ***
Age	259.45	11.94	21.727	<2e-16 ***
GenderM	-109.04	334.66	-0.326	0.745
Tobaccoyes	23833.87	414.19	57.544	<2e-16 ***
BMI	323.05	27.53	11.735	<2e-16 ***

F-statistic: 986.5 on 4 and 1333 DF, p-value: < 2.2e-16

Table 12 shows all models in one table.

	Model1	Model2	Model3	Model4
call	expression	expression	expression	expression
terms	Cost ~ Age	Cost ~ Age + Gender	Cost ~ Age + Tobacco + Gender	Cost ~ Age + Gender + Tobacco + BMI
residuals	numeric 1338	numeric 1338	numeric 1338	numeric 1338
coefficients	numeric 8	numeric 12	numeric 16	numeric 20
aliased	logical 2	logical 3	logical 4	logical 5
sigma	11560.31	11538.97	6399.019	6094.362
df	integer 3	integer 3	integer 3	integer 3
r.squared	0.0894059	0.09344348	0.7214122	0.7474973
adj.r.squared	0.08872432	0.09208534	0.7207857	0.7467396
fstatistic	numeric 3	numeric 3	numeric 3	numeric 3
cov.unscaled	numeric 4	numeric 9	numeric 16	numeric 25

3. Predict the outcome

Now we will be calculating RMSE, and we will be using one of the easiest and common way to calculate with the use of converting all the values to log but earlier we did not take the log of our feature selections. But, since did not take a log in the earlier Regression model therefore, we will be using this model.

Table 13 shows the Statistics for RMSE calculations.

MAE	RMSE	RMSLE
2918.187	16126.75	0.4376657

Above we can see three main calculations, mean absolute error, root mean square error and root mean square logarithmic error, we can see root mean square should be always equal to or higher than the mean absolute error. These are average differences from the true values.

Polynomial Regression

I have improved my model by using feature engineering and splitting the dataset into two sets, train, and test since developing new features visualizes new interactions within existing features, this is called as a polynomial regression model, a major feature matrix which consist of all the combinations of polynomial regression with 2 degrees of freedom, Ostertagová, E. (2012).

We will be integrating the below formula into the dataset for polynomial model prediction with 2 degrees of freedom from Bass, H. (1985).

"Formula = as.formula(

paste(' ~ .^2 + ',

paste('poly(', colnames(TrainX), ', 2, raw=TRUE)[, 2]', collapse = ' + ')))"

From the above correlation figure 4, we know Gender and Area do not correlate with medical cost, but still, we will be including those in the polynomial regression model.

After applying the formula now our equation is :

"~.^2 + poly(Cost, 2, raw = TRUE)[, 2] + poly(GenderNew, 2, raw = TRUE)[, 2] + poly(Age, 2, raw = TRUE)[, 2] + poly(BMI, 2, raw = TRUE)[, 2] + poly(TobaccoNew, 2, raw = TRUE)[, 2]"

```
##  [1] "(Intercept)"
##  [2] "Cost"
##  [3] "GenderNew"
##  [4] "Age"
##  [5] "BMI"
##  [6] "TobaccoNew"
##  [7] "poly(Cost, 2, raw = TRUE)[, 2]"
##  [8] "poly(GenderNew, 2, raw = TRUE)[, 2]"
##  [9] "poly(Age, 2, raw = TRUE)[, 2]"
## [10] "poly(BMI, 2, raw = TRUE)[, 2]"
## [11] "poly(TobaccoNew, 2, raw = TRUE)[, 2]"
## [12] "Cost:GenderNew"
## [13] "Cost:Age"
## [14] "Cost:BMI"
## [15] "Cost:TobaccoNew"
## [16] "GenderNew:Age"
## [17] "GenderNew:BMI"
## [18] "GenderNew:TobaccoNew"
## [19] "Age:BMI"
## [20] "Age:TobaccoNew"
## [21] "BMI:TobaccoNew"
```

Figure 5 now shows we have 16 total columns.

a) The constant term in the polynomial is represented by the column $(Intercept)$, which is composed of the constant 1.
b) The original features include age, BMI, gender, and smokers.
c) The original features are the square of $age2, bmi2, gender, smoker, and 2.
d) There are six interactions between pairings of four features: age times BMI, age times gender, age times smoker, BMI times smoker, and gender times smoker.
e) The objective feature is $Cost$.

I have made a new model using stepwise selection using the backward technique which has been shown in an R markdown file supplied along with this report.

Call:

lm(formula = Cost ~ Age + Gender + BMI + Tobacco, data = DF1)

Table 14 shows residuals for the original model:

Min	1Q	Median	3Q	Max
-12364.7	-2972.2	-983.2	1475.8	29018.3

Table 15 shows the coefficients for the original model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-11633.49	947.27	-12.281	<2e-16 ***
Age	259.45	11.94	21.727	<2e-16 ***
GenderM	-109.04	334.66	-0.326	0.745
BMI	323.05	27.53	11.735	<2e-16 ***
Tobaccoyes	23833.87	414.19	57.544	<2e-16 ***

Multiple R-squared: 0.7475, Adjusted R-squared: 0.7467

F-statistic: 986.5 on 4 and 1333 DF, p-value: < 2.2e-16

We have a total of 4 features in this model, given that all other characteristics are constant, a unit change in tobacco usage will result in a larger change in cost than a unit change in any other feature because patients who use more tobacco have the largest coefficient of all features. In this instance, assuming all other characteristics remain constant, a patient who does not use tobacco would incur

lower medical costs than a patient who does, Additionally, this model's modified R-squared value of 0.7475 indicates that it adequately accounts for 74% of the variation in medical costs.

Let's compare the original model with a new predicted polynomial regression model

Call:

lm(formula = Cost ~ BMI + GenderNew + TobaccoNew + `poly(Age, 2, raw = TRUE)[, 2]` + `poly(BMI, 2, raw = TRUE)[, 2]` + `BMI:TobaccoNew` + `GenderNew:TobaccoNew`, data = TrainPoly)

Table 16 shows residuals for the polynomial model:

Min	1Q	Median	3Q	Max
-9319.6	-2021.8	-1336.1	-242.2	29773.7

Table 17 shows the coefficients for the polynomial model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1611.379	3270.280	-0.493	0.6223
BMI	326.428	207.598	1.572	0.1162
GenderNew	-724.785	362.013	-2.002	0.0456 *
TobaccoNew	-21192.162	2134.570	-9.928	<2e-16 ***
`poly(Age2 raw = TRUE)[ 2]`	3.150	0.145	21.726	<2e-16 ***
`poly(BMI2 raw = TRUE)[ 2]`	-5.245	3.250	-1.614	0.1070
`BMI:TobaccoNew`	1471.893	68.512	21.484	<2e-16 ***
`GenderNew:TobaccoNew`	206.307	827.225	0.249	0.8031

Multiple R-squared: 0.827, Adjusted R-squared: 0.8257

F-statistic: 633.7 on 7 and 928 DF, p-value: < 2.2e-16

Except for gender: tobacco, all of our eight features are important for "Cost." According to the equations, a patient who does not smoke and has a BMI of 0 will be charged by health insurance $211,992 (which we know this scenario is impossible). Additionally, given that all other features

are fixed, a change in tobacco costs more to treat than a change in any other feature because it has the highest coefficient of all the features. As a result of this research, we are aware that adding additional characteristics to our model via polynomial combinations may cause the assumptions underlying it to alter, which may lead to incorrect interpretation.

This model's adjusted R-squared is 0.827, which indicates that the model's features account for 82% of the variation in cost, compared to the preceding Linear Regression model, the Polynomial Regression model captures higher cost variance.

Because we anticipate receiving residuals close to zero, the residuals from the linear regression model should be normally distributed. We can visualize this by plotting the histogram of the residuals.

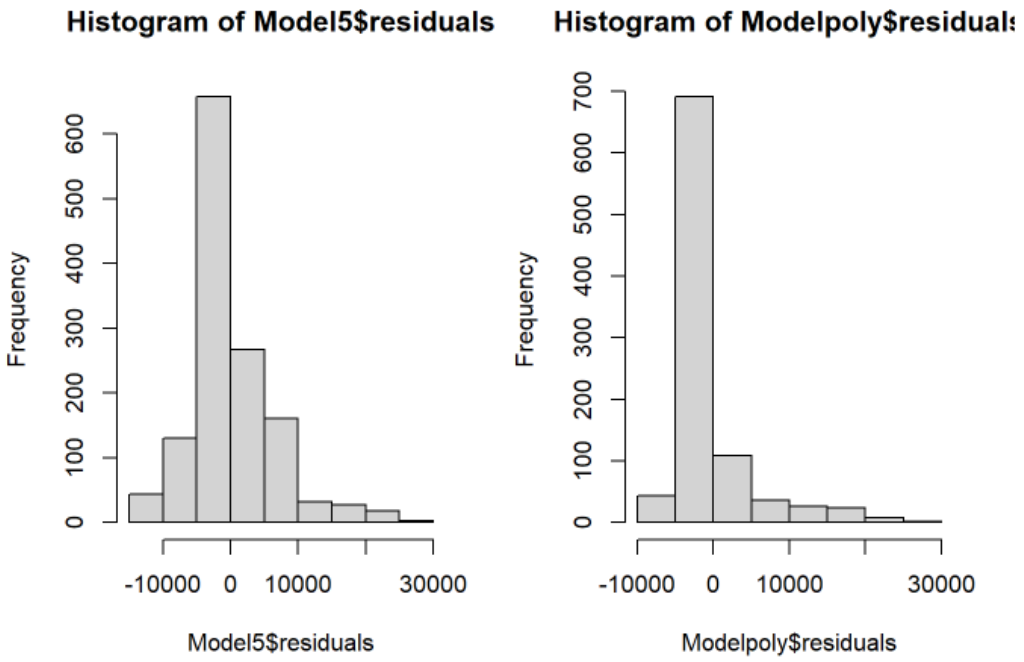

Figure 6 shows the visual histogram for both models to compare.

From the above histogram, we now confirm that most of the residuals are located close to zero.

Conclusion and Recommendations

This report consists of different sets of analyses, such as exploratory and predictive analyses. This has been done using various models in this report, one is a linear regression model and the other is a polynomial regression model, I divided the dataset into two parts using feature selection which made it possible to create more accurate models. Then using polynomial techniques predicted values that influence cost at a high level such as tobacco consumption. From this report now we know, patients with high consumption of tobacco will have higher medical costs in the future and at the same time high premiums.

Recommendations are to increase tobacco consumption awareness so that people consume less tobacco and have an insight into increasing medical expenses, secondly polynomial regression model should be the preferred model for this test since it gives a good prediction on medical datasets.

References

Ostertagová, E. (2012). Modeling using polynomial regression. *Procedia Engineering*, *48*, 500-506.

Bass, H., & Meisters, G. (1985). Polynomial flows in the plane. *Advances in Mathematics*, *55*(2), 173-208.