Final Report

Healthcare Cost Prediction

Table of Contents

1. Introduction

1.1 Dataset Overview

The dataset under consideration is an extensive collection of records, each representing an individual's profile and associated medical insurance charges. Sourced from Kaggle, this dataset encompasses diverse demographic and lifestyle attributes, offering a holistic view of factors that may influence the cost of medical insurance. The key features include age, gender, BMI (Body Mass Index), number of children, smoking status, region, and, most importantly, the target variable: medical insurance charges incurred by each individual that we want to predict. The dataset aims to capture the diverse characteristics of individuals that may influence the cost of their medical insurance. Understanding these factors is crucial for both insurance providers and policyholders, as it can aid in making informed decisions regarding insurance coverage and pricing strategies.

1.2 Dataset and Source

This dataset contains 1338 rows, each of which represents a comprehensive collection of health-related data designed to provide insights into factors influencing medical costs for individuals. The dataset contains 1338 observations on 7 variables and the variables of interest are given in the table below.

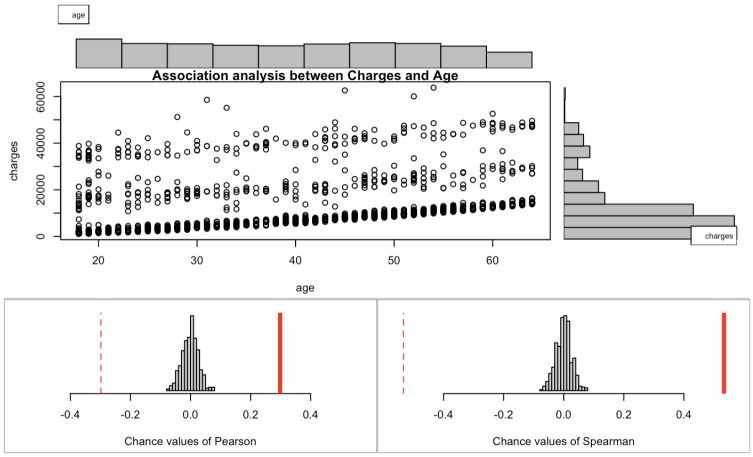
Variable	Description
Age	The age of the insured individuals.
Sex	Gender of the insured person (Male/Female)
BMI(Body Mass Index)	A measure of body fat calculated from an individual's weight and height.
Children	Number of children or dependents covered by the insurance.
Smoker	Smoking status of the insured (Yes/No)
Charges	Medical insurance charges incurred by each individual.
Region	Geographic region of the insured person.

2. Real Problem

The primary problem is to understand how lifestyle factors, including smoking habits and BMI, contribute to variations in individual medical costs. Can we predict medical costs considering the most influential variables?
This project aims to explore the relationships between various features and insurance costs, providing valuable insights for both individuals seeking insurance and insurance providers. Addressing these questions not only benefits the insurance industry by optimizing pricing and risk assessment but also empowers individuals to make informed decisions about their health and insurance coverage. Through data-driven insights, this project aims to contribute to fair and transparent practices within the healthcare and insurance sectors

3. Association Analysis

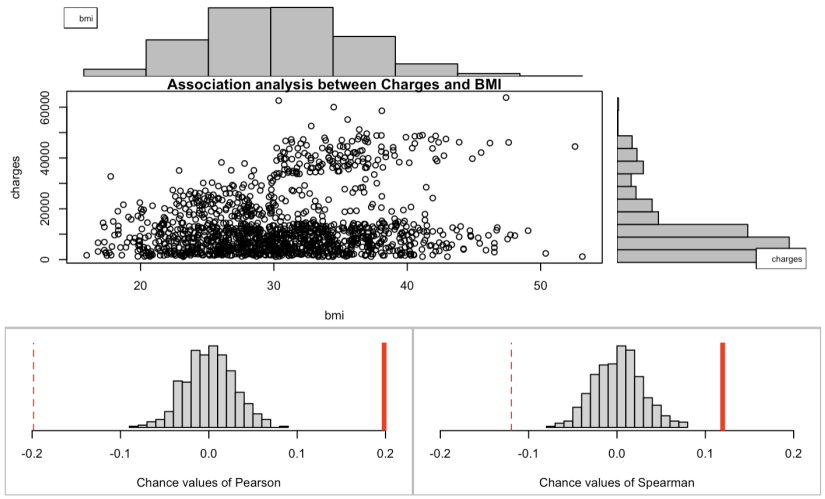
Association Analysis between Charges and Age




```
Association between age (numerical) and  charges (numerical)
 using 1337 complete cases
Permutation procedure:
                                Value Estimated p-value
Pearson's r                 0.2983082                 0
Spearman's rank correlation 0.5335233                 0
With 500 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.007
 the p-value of Spearman's rank correlation is between 0 and 0.007
Note:  If 0.05 is in this range, increase the permutations= argument.



Advice: If stream of points is well described by an ellipse, use Pearson's r.
Otherwise, as long as stream is monotonic, use Spearman's rank correlation
or try logs, e.g. associate( log10(y)~log10(x) )
```

For the association between Charges and Age, we get the p-value as 0, and the correlation is statistically significant and conclusive. We consider Pearson's r correlation as there is a linear correlation between charges and age. r = 0.2983082

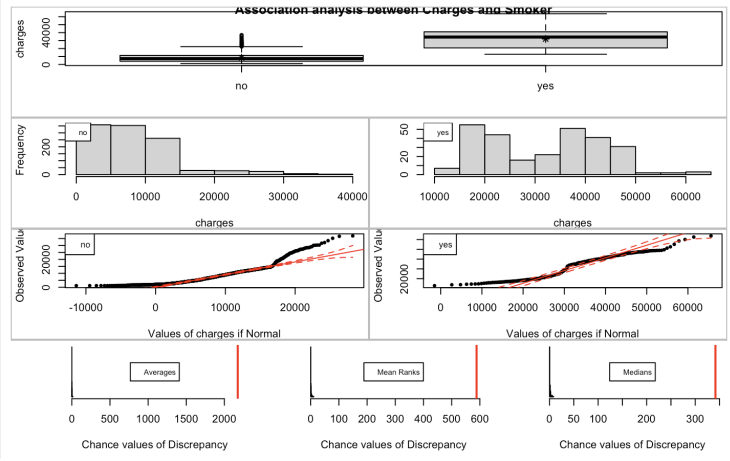
Association analysis between Charges and BMI(Body Mass Index)




```
Association between bmi (numerical) and  charges (numerical)
 using 1337 complete cases
Permutation procedure:
                              Value Estimated p-value
Pearson's r                0.1984008                 0
Spearman's rank correlation 0.1195850                 0
With 500 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.007
 the p-value of Spearman's rank correlation is between 0 and 0.007
Note:  If 0.05 is in this range, increase the permutations= argument.



Advice: If stream of points is well described by an ellipse, use Pearson's r.
Otherwise, as long as stream is monotonic, use Spearman's rank correlation
or try logs, e.g. associate( log10(y)~log10(x) )
```

For the association between Charges and BMI(Body Mass Index), we get the p-value as 0, and the correlation is statistically significant and conclusive. We consider Pearson's r correlation as there is a linear correlation between Charges and BMI. r = 0.1984008

Association analysis between Charges and Smoker




```
Association between smoker (categorical) and  charges (numerical)
 using 1337 complete cases

Sample Sizesx
  no  yes
1063  274

Permutation procedure:
With 500 permutations, we are 95% confident that
 the p-value of ANOVA (means) is between 0 and 0.007
 the p-value of Kruskal-Wallis (ranks) is between 0 and 0.007
 the p-value of median test is between 0 and 0.007
Note:  If 0.05 is in a range, change permutations= to a larger number



Advice: If it makes sense to compare means (i.e., no extreme outliers and the
distributions aren't too skewed), use the ANOVA.  If there there are
some obvious extreme outliers but the distributions are roughly symmetric, use
Rank test.  Otherwise, use the Median test or rerun the test using, e.g., log10(y)
instead of y
```

For the association between Charges and Smoker, there is a noticeable difference among bars, so there is an association between charges and smokers. Since the distributions have extreme outliers, we will compare the medians. The estimated p-value for the medians is 0, which is less than 0.05. We can conclude there is a significant relation between charges and if the person is a smoker.

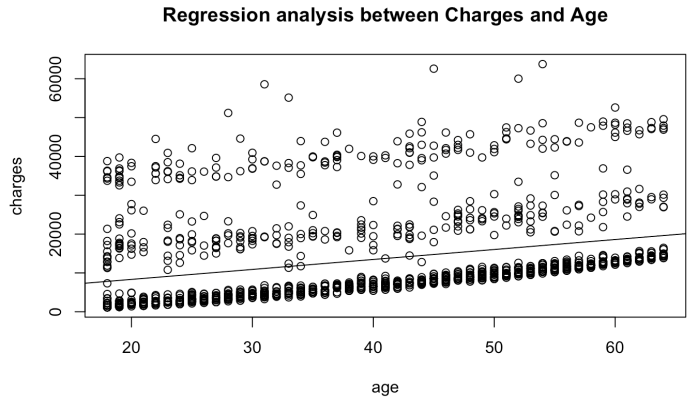
Analysis Summary

Variable	R^2	RSE(Residual Standard Error)	P-Value
AGE	8%	11560	<2e-16
BMI	4%	11870	2.47e-13
SEX	0.33%	12090	0.0338
SMOKER	6.1%	7471	<2.2e-16
REGION	0.6%	12080	0.03276

Based on the analysis summary, we observe the following key insights:
The age of the insured individuals explains 8% of the variability in medical insurance charges. The relationship is statistically significant, as indicated by the extremely low p-value. Body Mass Index contributes to 4% of the variability in insurance charges, with a strong and statistically significant association (p-value < 0.05). Gender shows a modest impact, explaining only 0.33% of the variability in insurance charges. The association is statistically significant, but the effect is relatively small. Smoking status stands out as a substantial factor, explaining 6.1% of the variability in insurance charges. The relationship is highly significant (p-value < 0.05), emphasizing the notable impact of smoking on medical insurance costs. The geographic region contributes minimally (0.6%) to the variability in insurance charges, with a statistically significant association (p-value < 0.05). The analysis underscores that smoking status is a particularly influential factor, explaining a substantial portion of the variability in medical insurance charges. This is evident in the high R^2 value (6.1%) and the exceptionally low p-value (<2.2e-16). Understanding and considering the impact of smoking status is crucial for accurate predictions and fair pricing in the insurance context.

4. Regression Analysis

Regression analysis between Charges and Age

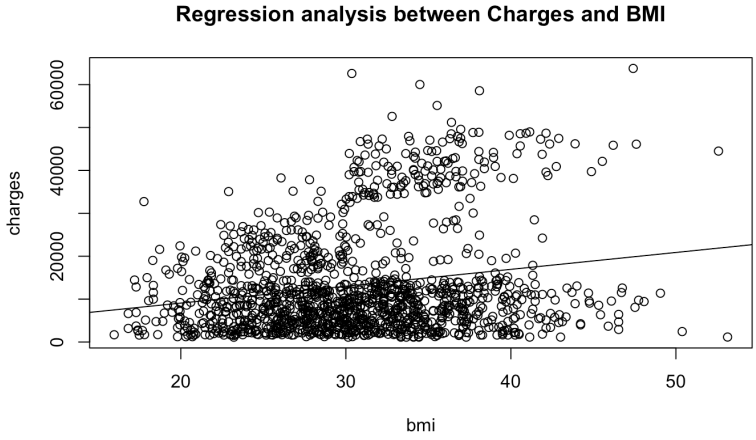



```
Residuals:
   Min     1Q Median     3Q    Max
 -8064  -6684  -5943   5466  47828

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)  3190.02     938.40   3.399             0.000695 ***
age           257.23      22.53  11.419 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11560 on 1335 degrees of freedom
Multiple R-squared:  0.08899,   Adjusted R-squared:  0.08831
F-statistic: 130.4 on 1 and 1335 DF,  p-value: < 0.00000000000000022

                2.5 %    97.5 %
(Intercept) 1349.1302 5030.9133
age          213.0402  301.4193
```

For the regression analysis, we get r^2 as 8%; it is a relatively weak correlation. The intercept of the line is 3190.02 and the slope of the line is 257.23. Since the confidence interval of the slope doesn't include 0, the relationship between charges and age is statistically significant. The p-value for the coefficient of age is very low (< 2e-16), indicating that there is strong evidence to reject the null hypothesis that the coefficient for age is zero.

The model suggests that there is a significant relationship between the age and charges variable. The coefficient for age is positive, indicating that as age increases, charges are expected to increase as well.

Regression Analysis between Charges and BMI(Body Mass Index)

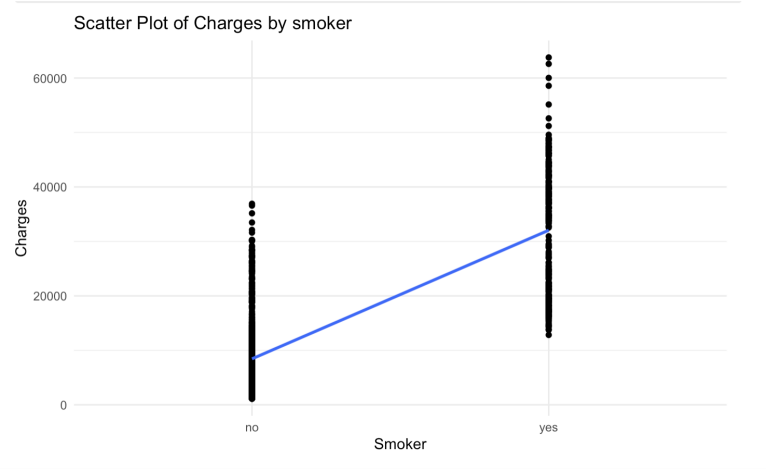



```
Residuals:
   Min     1Q Median     3Q    Max
-20964  -8112  -3762   4717  49433

Coefficients:
            Estimate Std. Error t value          Pr(>|t|)
(Intercept)  1202.14    1664.86   0.722             0.47
bmi           393.86      53.25   7.396 0.000000000000247 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11870 on 1335 degrees of freedom
Multiple R-squared:  0.03936,   Adjusted R-squared:  0.03864
F-statistic:  54.7 on 1 and 1335 DF,  p-value: 0.0000000000002468

                 2.5 %    97.5 %
(Intercept) -2063.8800 4468.1608
bmi           289.3899  498.3219
```

For the regression analysis, we get r^2 is 4% which is a relatively weak correlation. The intercept of the line is 1202.14 and the slope of the line is 393.86. Since the confidence interval of the slope doesn't include 0, the relationship between charges and BMI is statistically significant. The p-value for the coefficient of 'BMI' is very low (2.47e-13), indicating that there is strong evidence to reject the null hypothesis that the coefficient for BMI is zero.

The model suggests that there is a significant relationship between the BMI and charges. The coefficient for BMI is positive, indicating that as BMI increases, charges are expected to increase as well. However, the overall explanatory power of the model is relatively low.

Regression Analysis between Charges and Smoker




```
Residuals:
   Min     1Q Median     3Q    Max
-19221  -5048   -923   3702  31720

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)   8440.7      229.1   36.84 <0.0000000000000002 ***
smokeryes    23609.6      506.2   46.65 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7471 on 1335 degrees of freedom
Multiple R-squared:  0.6197,    Adjusted R-squared:  0.6195
F-statistic:  2176 on 1 and 1335 DF,  p-value: < 0.00000000000000022

                2.5 %    97.5 %
(Intercept)  7991.153  8890.167
smokeryes   22616.623 24602.521
```

For the regression analysis, we get r^2 as 6.1% which is a strong correlation suggesting that the model explains a substantial proportion of the variance in the dependent variable. The intercept of the line is 8440.7 and the slope of the line is 23609.6 Since the confidence interval of the slope doesn't include 0, the relationship between charges and age is statistically significant. The p-value for the coefficient of 'smokeryes' is very low (< 2e-16), indicating that there is strong evidence to reject the null hypothesis that the coefficient for 'smokeryes' is zero.

The model suggests that there is a significant relationship between the smoker variable and the charges variable. The coefficient for ‘smokeryes’ is positive, indicating that being a smoker is associated with higher charges.

Other Techniques:

5. Backward Elimination Method

The study used a backward elimination strategy to refine a linear regression model for predicting healthcare charges. Initially, a comprehensive model (fit1) was fitted to explore potential relationships, followed by a more focused model (fit2) featuring specific variables like age, BMI, number of children, and smoking status. A new data frame named "Ryan" is created with specific values for each predictor variable. The predict() function is then used to estimate healthcare charges for Ryan based on the second linear regression model (fit2) and we get the Healthcare charges for Ryan as $7159.48.

```
Call:
lm(formula = charges ~ ., data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11305.1  -2850.3   -979.9   1395.0  29992.8

Coefficients:
                 Estimate Std. Error t value             Pr(>|t|)
(Intercept)     -12066.04    1000.20 -12.064 < 0.0000000000000002 ***
age                256.76      11.91  21.555 < 0.0000000000000002 ***
sexfemale          129.48     333.20   0.389             0.697630
bmi                339.25      28.61  11.857 < 0.0000000000000002 ***
children           474.82     137.90   3.443             0.000593 ***
smokeryes        23847.33     413.35  57.693 < 0.0000000000000002 ***
regionnorthwest   -349.23     476.82  -0.732             0.464053
regionsoutheast  -1035.27     478.87  -2.162             0.030804 *
regionsouthwest   -960.08     478.11  -2.008             0.044836 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6064 on 1328 degrees of freedom
Multiple R-squared:  0.7507,    Adjusted R-squared:  0.7492
F-statistic:   500 on 8 and 1328 DF,  p-value: < 0.00000000000000022
```

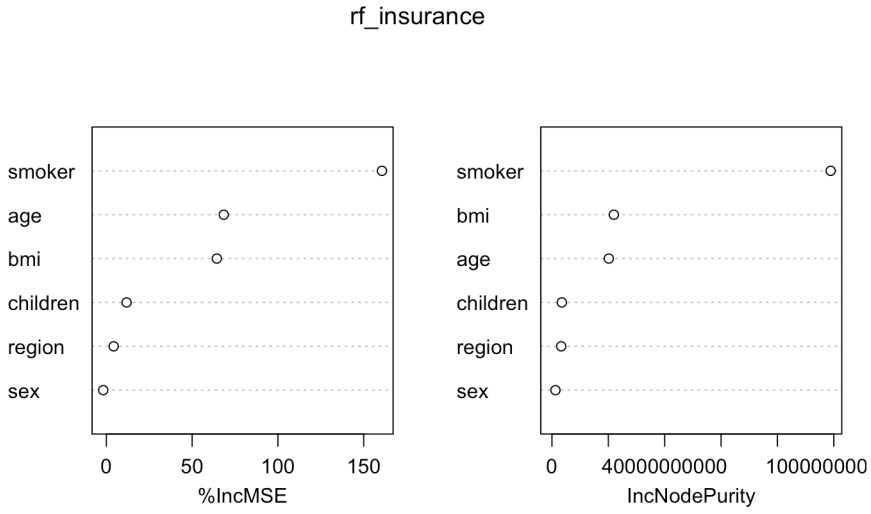
Backward Elimination Method Implemented

```
Call:
lm(formula = charges ~ age + bmi + children + smoker, data = insurance)

Residuals:
   Min     1Q Median     3Q    Max
-11898  -2921   -986   1395  29510

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) -12098.82     942.63  -12.84 < 0.0000000000000002 ***
age             257.77      11.91   21.64 < 0.0000000000000002 ***
bmi             321.87      27.39   11.75 < 0.0000000000000002 ***
children        472.98     137.88    3.43             0.000621 ***
smokeryes     23810.40     411.41   57.88 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6070 on 1332 degrees of freedom
Multiple R-squared:  0.7495,    Adjusted R-squared:  0.7488
F-statistic: 996.5 on 4 and 1332 DF,  p-value: < 0.00000000000000022

[1] "Health care charges for Ryan: 7159.48"
```

Variables with higher p-values were systematically removed, resulting in a more streamlined model. This deliberate selection of predictors aligns with the principles of backward elimination, focusing on the retention of relevant variables for optimizing the model's predictive capabilities. The resulting model, fit2, demonstrates the importance of key predictors in explaining healthcare charge variability.

6. Random Forest

A Random Forest regression model is constructed to predict healthcare charges using the "insurance" dataset. The dataset is divided into training and testing sets, with 80% of the data used for training the model. The Random Forest model is built with the randomForest function, considering all available predictor variables. The model's performance is then evaluated on the test set, calculating the Root Mean Squared Error (RMSE) as a measure of prediction accuracy. Variable importance is assessed using the importance function and visualized with a varImpPlot, and we find that the smoker is the most important variable. Additionally, a hypothetical individual named Ryan is introduced with specific characteristics, and the model is used to predict Ryan's health care charges, and we get it as $18649.19. The code concludes by printing the predicted charge for Ryan and reiterating the health care charges prediction using the Random Forest model. The careful consideration of factors such as age, gender, BMI, smoking status, and region contributes to the model's ability to make accurate predictions. The Random Forest approach provides a robust framework for capturing complex relationships within the data, making it valuable for predicting healthcare charges in a diverse population.



7. Comparison of RMSE

METHOD	RMSE(ROOT-MEAN-SQUARE DEVIATION)
Validation test approach	6058.642
Random Forest	5317.114

Random forest is the best method to predict the insurance charges as it has lower RMSE compared to the Validation test approach.

8. Conclusion

In conclusion, our analysis of the "insurance" dataset revealed key predictors, such as age, BMI, and smoking status, significantly influencing health care charges. The application of both linear regression and Random Forest models demonstrated the latter's superior predictive performance, showcased by a lower Root Mean Squared Error. This robust model, capturing intricate relationships within the data, provides valuable insights for healthcare planning and cost estimation. The personalized prediction for an individual named Ryan further underscores the model's adaptability. In summary, our findings offer practical tools for stakeholders in healthcare and insurance, facilitating data-driven decision-making and enhancing understanding of factors influencing healthcare charges.

9. References

Dataset:
Medical Cost Personal Datasets. (2018, February 21). Kaggle.
https://www.kaggle.com/datasets/mirichoi0218/insurance/data