HEALTH CARE PREMIUM COSTS PREDICTION

Vinay

2022-06-23

## Abstract

Insurance involves a policy that ensures a decrease or elimination of the loss costs incurred due to several risks. Several factors always influence insurance costs. With machine learning algorithms, implementing insurance policies is more efficient and reliable. Therefore, this paper's main focus is to develop a linear regression model that will help predict the insurance costs and the factors that need to be considered when taking an insurance coverage contract. The project will include using a dataset downloaded from the Kaggle.com website. The project answers the research question of whether there is any relationship between health insurance costs and age. The project results illustrate a strong positive correlation between insurance costs and age.

## Introduction

With the increase in uncertainties and threats, households, companies, people, and properties are exposed to different risks. The risk levels always vary, including health, death, loss of assets, or property. However, the well-being and life of people are the greatest parts and essential. However, these risks cannot be avoided. This is where the finance industry has developed several products that can protect organizations and individuals from such risks through capital reimbursement. Therefore, insurance involves using a policy that removes or decreases the costs of losses incurred when risks occur. The insurance companies need to be sufficiently precise in the case of quantifying and measuring the cost that the policy has covered and the expenses that need to be paid for this insurance. Several variables are used to estimate this, and in case of any omission, it affects the overall policy. Therefore, there needs to be an accurate process in the insurance policy. This has been a challenge for many insurance companies. Therefore, this project objective is to have a linear regression model that will help with a solution for this challenge. The project is implemented in the R programming language.

## Literature Review

In the study performed by Pesantez-Narvaez et al. (2019), the authors made a comparison of the performance of XGBoost and logistic regression techniques in the prediction of the presence of accident claims. The result of their study illustrated that logistic regression was more suitable than the XGBoost technique as it had strong predictability. In another study by Stucki (2019), the author identified Random forest as the best model, with an accuracy of 74%, in predicting the insurance policy determining factors as they tried to understand the customer churn. From the reviewed works, most of the developed models did not consider both the claim severity and cost predicted. The models only make classification of issues related to claims. Therefore, this project will focus on a logistic regression algorithm to help predict health premium costs.

## Theory

H1: Age is a major factor affecting the amount of Health Insurance.

## Data

The data has been retrieved from <https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction/download>. It includes:

#Before performing any analysis, the project involve importing all the necessary libraries

#install.packages("Hmisc")

#install.packages("WVPlots")

library(ggplot2)

library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##

## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':

##

## format.pval, units

library(cowplot)

library(WVPlots)

## Loading required package: wrapr

library(e1071)

##

## Attaching package: 'e1071'

## The following object is masked from 'package:Hmisc':

##

## impute

library(caret)

##

## Attaching package: 'caret'

## The following object is masked from 'package:survival':

##

## cluster

## age sex bmi children smoker region expenses

## 1 19 female 27.9 0 yes southwest 16884.92

## 2 18 male 33.8 1 no southeast 1725.55

## 3 28 male 33.0 3 no southeast 4449.46

## 4 33 male 22.7 0 no northwest 21984.47

## 5 32 male 28.9 0 no northwest 3866.86

## 6 31 female 25.7 0 no southeast 3756.62

Get idea about the data by using the describe()function. We see missing values, mean, lowest, and highest values.

describe(data)

## data

##

## 7 Variables 1338 Observations

## --------------------------------------------------------------------------------

## age

## n missing distinct Info Mean Gmd .05 .10

## 1338 0 47 0.999 39.21 16.21 18 19

## .25 .50 .75 .90 .95

## 27 39 51 59 62

##

## lowest : 18 19 20 21 22, highest: 60 61 62 63 64

## --------------------------------------------------------------------------------

## sex

## n missing distinct

## 1338 0 2

##

## Value female male

## Frequency 662 676

## Proportion 0.495 0.505

## --------------------------------------------------------------------------------

## bmi

## n missing distinct Info Mean Gmd .05 .10

## 1338 0 275 1 30.67 6.894 21.27 23.00

## .25 .50 .75 .90 .95

## 26.30 30.40 34.70 38.63 41.10

##

## lowest : 16.0 16.8 17.2 17.3 17.4, highest: 48.1 49.1 50.4 52.6 53.1

## --------------------------------------------------------------------------------

## children

## n missing distinct Info Mean Gmd

## 1338 0 6 0.899 1.095 1.275

##

## lowest : 0 1 2 3 4, highest: 1 2 3 4 5

##

## Value 0 1 2 3 4 5

## Frequency 574 324 240 157 25 18

## Proportion 0.429 0.242 0.179 0.117 0.019 0.013

## --------------------------------------------------------------------------------

## smoker

## n missing distinct

## 1338 0 2

##

## Value no yes

## Frequency 1064 274

## Proportion 0.795 0.205

## --------------------------------------------------------------------------------

## region

## n missing distinct

## 1338 0 4

##

## Value northeast northwest southeast southwest

## Frequency 324 325 364 325

## Proportion 0.242 0.243 0.272 0.243

## --------------------------------------------------------------------------------

## expenses

## n missing distinct Info Mean Gmd .05 .10

## 1338 0 1337 1 13270 12301 1758 2347

## .25 .50 .75 .90 .95

## 4740 9382 16640 34832 41182

##

## lowest : 1121.87 1131.51 1135.94 1136.40 1137.01

## highest: 55135.40 58571.07 60021.40 62592.87 63770.43

## --------------------------------------------------------------------------------

The described data above looks tidy since there are no missing values.

sapply(data,function(x) sum(is.na(x)))

## age sex bmi children smoker region expenses

## 0 0 0 0 0 0 0

## Methodology

Having the data cleaned, next will do analysis on the dataset and identify patterns in the dataset. First analysis is to find the correlation between expenses, age, and bmi.

x <- ggplot(data, aes(age, expenses)) +

geom\_jitter(color = "blue", alpha = 0.5) +

theme\_light()

y <- ggplot(data, aes(bmi, expenses)) +

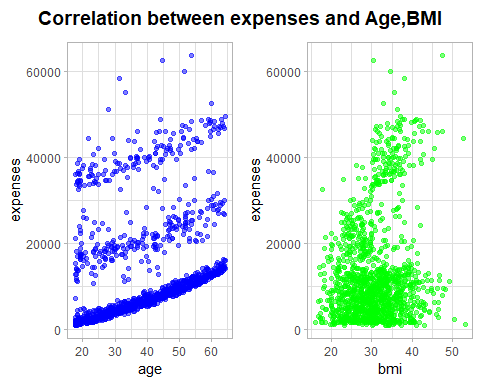
geom\_jitter(color = "green", alpha = 0.5) +

theme\_light()

plot\_1 <- plot\_grid(x, y)

title <- ggdraw() + draw\_label("Correlation between expenses and Age,BMI", fontface='bold')

plot\_grid(title, plot\_1, ncol=1, rel\_heights=c(0.1, 1))



Second analysis involve correlation between expenses and Sex,Children covered

x <- ggplot(data, aes(sex,expenses)) +

geom\_jitter(aes(color = sex), alpha = 0.7) +

theme\_light()

y <- ggplot(data, aes(children,expenses)) +

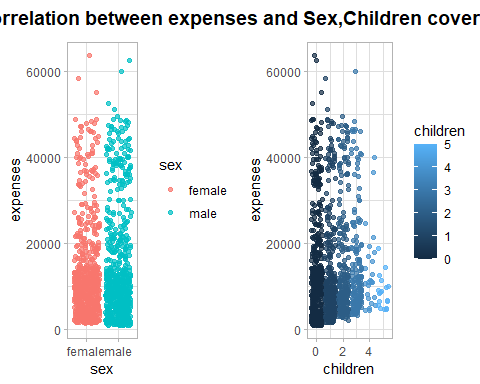
geom\_jitter(aes(color = children), alpha = 0.7) +

theme\_light()

plot\_2 <- plot\_grid(x, y)

title <- ggdraw() + draw\_label("Correlation between expenses and Sex,Children covered", fontface='bold')

plot\_grid(title, plot\_2, ncol=1, rel\_heights=c(0.1, 1))



Next is the correlation analysis between expenses and smokers, region.

x <- ggplot(data, aes(smoker,expenses)) +

geom\_jitter(aes(color = smoker), alpha = 0.7) +

theme\_light()

y <- ggplot(data, aes(region,expenses)) +

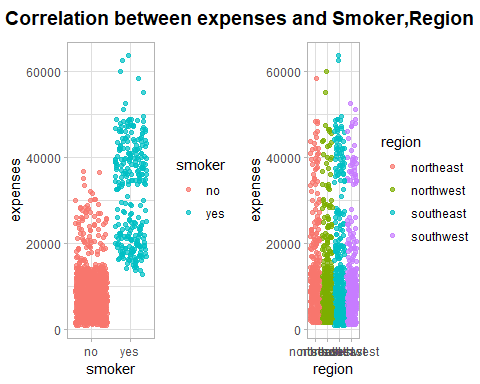
geom\_jitter(aes(color = region), alpha = 0.7) +

theme\_light()

plot\_3 <- plot\_grid(x, y)

title <- ggdraw() + draw\_label("Correlation between expenses and Smoker,Region", fontface='bold')

plot\_grid(title, plot\_3, ncol=1, rel\_heights=c(0.1, 1))



Build the linear regression model. First, split the dataset for modelling.

train\_spl <- round(0.8 \* nrow(data))

ind\_train <- sample(1:nrow(data), train\_spl)

training <- data[ind\_train, ]

testing <- data[-ind\_train, ]

formula\_1 <- as.formula("expenses ~ age + bmi + children + smoker + region")

model\_lr <- lm(formula\_1, data =training)

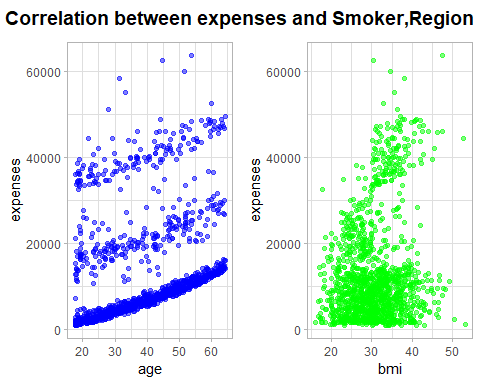
pred <- predict(model\_lr, newdata =testing)

residuals\_1 <- testing$expenses - pred

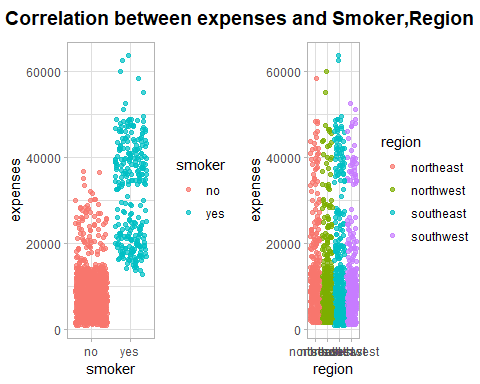
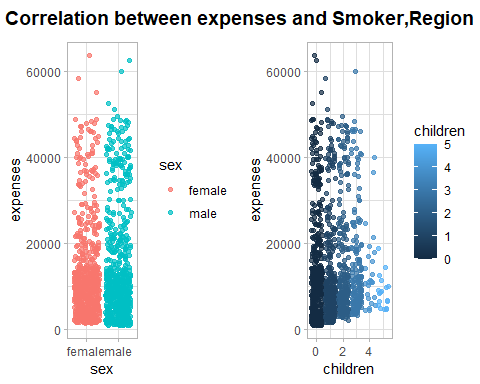
rmse <- sqrt(mean(residuals\_1^2))

## Results

The analysis in the project identified various observations. There is a strong positive correlation between age, BMI, and expenses. This is because from the scatter plot, the increase in values for age and BMI also led to the increase in the values for insurance expenses.



It is also clear that there is no significant relationship between sex and charges. In addition, the insurance expenses of smokers are higher than non-smokers.

The summary of the logistic regression model depicts that smoking greatly impacts insurance charges.

##

## Call:

## lm(formula = formula\_1, data = training)

##

## Residuals:

## Min 1Q Median 3Q Max

## -11069.3 -2859.0 -992.2 1539.7 30179.0

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) -12163.02 1094.10 -11.117 < 2e-16 \*\*\*

## age 251.10 13.18 19.053 < 2e-16 \*\*\*

## bmi 357.14 32.29 11.059 < 2e-16 \*\*\*

## children 577.84 151.78 3.807 0.000149 \*\*\*

## smokeryes 23745.20 460.68 51.544 < 2e-16 \*\*\*

## regionnorthwest -770.49 528.24 -1.459 0.144971

## regionsoutheast -1324.90 527.27 -2.513 0.012127 \*

## regionsouthwest -1411.64 535.22 -2.638 0.008474 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 6017 on 1062 degrees of freedom

## Multiple R-squared: 0.7516, Adjusted R-squared: 0.75

## F-statistic: 459.1 on 7 and 1062 DF, p-value: < 2.2e-16

The model has an RMSE of 6025.65.

print(paste0("RMSE for the model: ", round(rmse, 2)))

## [1] "RMSE for the model: 6250.47"

## Implications

For further analysis, there needs to have an inclusion of more features as this will ensure that there are comparatively more accurate results.

## Conclusion

The project analysis and prediction have determined the main features crucial in health insurance. Smoking has been identified as a major feature that greatly affects insurance expenses. The study has also developed a prediction model useful to insurance companies and their potential customers.

## References

Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. Risks, 7(2), 70.

Stucki, O. (2019). Predicting the customer churn with machine learning methods: case: private insurance customer data