DETERMINANTS

OF HEALTH INSURANCE PREMIUMS

WIDJAJA Edbert - 3036240503

QUESTION TO ANSWER

What is the order of significance of attributes that determine health insurance premium costs ?

QUANTIFIED USING MACHINE LEARNING

WHY IS IT IMPORTANT?

Inform prospective policyholders on their estimated premium, improve financial planning

Provide details on an insurance company's decision-making and risk management process, which builds transparency

EDA and Data Visualization to analyse initial relationships using R

LET'S FIRST ANALYZE OUR DATASET !

Loading Dataset and Libraries

age (double)	sex (character)	bmi (double)	children (double)	smoker (character)	region (character)	charges (double)
19	female	27.900	0	yes	southwest	16884.924
18	male	33.770	1	no	southeast	1725.552
28	male	33.000	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.471
32	male	28.880	0	no	northwest	3866.855
31	female	25.740	0	no	southeast	3756.622
46	female	33.440	1	no	southeast	8240.590
37	female	27.740	3	no	northwest	7281.506

```
library(readr)
library(ggplot2)
library(dplyr)
library(ggthemes)
library(gridExtra)
library(reshape2)
library(randomForest)
library(caret)
```

Check for NAs

```
          NoOfNull
age              0
sex              0
bmi              0
children         0
smoker           0
region           0
charges          0
```

No Null Values to remove..

Convert charges & bmi column to 2 d.p. & 1 d.p. respectively

What can we observe?

For Factor Variables

Premium costs are more or less similar in each group except for smoking / non-smoking

What can we observe?

- Graphs still support previous arguments
- Age (+), Charges (+)

For non-Factor Variables

ENCODING DATA

For an ML-algorithm to process the dataset, we must change each non-numerical column to numerics

```r
#encode factors to numeric for machine learning evaluation
insurance <- insurance |>
  mutate(sex = ifelse(sex == "male",0,1),
         smoker = ifelse(smoker =="yes",1,0),
         region = case_when(
           region == "northwest" ~ 0,
           region == "northeast" ~ 1,
           region == "southeast" ~ 2,
           TRUE ~ 3
         )) |>
  mutate_at(c("sex","children","smoker","region"),as.numeric)
```

	age	sex	bmi	children	smoker	region	charges
1	19	1	27.9	1	1	3	16884.92
2	18	0	33.8	2	0	2	1725.55
3	28	0	33.0	4	0	2	4449.46
4	33	0	22.7	1	0	0	21984.47
5	32	0	28.9	1	0	0	3866.86
6	31	1	25.7	1	0	2	3756.62
7	46	1	33.4	2	0	2	8240.59
8	37	1	27.7	4	0	0	7281.51
9	37	0	29.8	3	0	1	6406.41
10	60	1	25.8	1	0	0	28923.14
11	25	0	26.2	1	0	1	2721.32

corr values

Major Takeaways

- Smoker ~ Charges : 0.79
- Age ~ Charges : 0.3
- bmi ~ Charges : 0.2
- children ~ Charges : 0.07
- Sex ~ Charges :-0.06
- Region ~ Charges : 0.01

Seems like we‘ve answered our question...

HAVE WE ANSWERED THE QUESTION?

No

- Correlation does not imply causation
- Might not be a good indicator

But, Yes

- Identifying relationships & associations
- Provides insight for further analysis

Random Forest Modeling and Fitting

How it works?

- Multiple decision trees searching for best feature among a random subset of features

RANDOM FOREST REGRESSION

Model Fitting and Prediction on Existing Dataset

Adding a new column : prediction

	age	sex	bmi	children	smoker	region	charges	prediction
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	19	1	27.9	1	1	3	16885.	17629.
2	18	0	33.8	2	0	2	1726.	4235.
3	28	0	33	4	0	2	4449.	5023.
4	33	0	22.7	1	0	0	21984.	5024.
5	32	0	28.9	1	0	0	3867.	4660.
6	31	1	25.7	1	0	2	3757.	4029.

```r
set.seed(42)
rf_model <- randomForest(formula = insurance$charges ~ . , data = insurance, ntree=1250,mtry = 3,
                         keep.forest= FALSE, importance = TRUE )
print(rf_model)
predicted_insurance <- insurance |>
  mutate(prediction = rf_model$predicted)
```

importance = TRUE ?

The model is trained with all of the existing data and the prediction is done on the same dataset

ACCURACY OF MODEL

Determining accuracy and error term of model

```
Call:
lm(formula = predicted_insurance$prediction ~ predicted_insurance$charges)

Residuals:
     Min       1Q   Median       3Q      Max 
-23531.0   -659.6    436.3   2046.9  12432.9 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  2.160e+03  1.730e+02   12.49   <2e-16 ***
predicted_insurance$charges 8.480e-01  9.631e-03   88.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4265 on 1336 degrees of freedom
Multiple R-squared:  0.853,	Adjusted R-squared:  0.8529 
F-statistic:  7752 on 1 and 1336 DF,  p-value: < 2.2e-16
```

```
               Type of random forest: regression
                     Number of trees: 1250
No. of variables tried at each split: 3

          Mean of squared residuals: 21566862
                    % Var explained: 85.28
```

85.3% ACCURACY

We see that most predicted points = actual point

3

Conclusion and Possible Improvements

CONCLUSION & IMPROVEMENT

Conclusion

```
importance <- as.data.frame(varImp(rf_model))
importance <- importance |>
  mutate(Overall = round(Overall,2)) |>
  arrange(desc(Overall)) |>
  rename(Importance_Score = Overall)
```

	Importance_Score
smoker	364.07
bmi	179.65
age	176.18
children	28.98
region	15.04
sex	-6.12

As suspected, the smoker column is most significant

Improvements

- **Experiment with other models for maximum accuracy**
- **Using GridSearch to find optimal hyper parameters**

REFERENCES

- https://www.kaggle.com/datasets/simranjain17/insurance/data
- https://www.kaggle.com/code/adepvenugopal/ml-regression-predict-insurance-charges
- https://www.geeksforgeeks.org/medical-insurance-price-prediction-using-machine-learning-python/