Final Project

Insurance Premium Charges in US | Kaggle

Group No. 1

B. Thanushan	S14025
W. S. S. Fernando	S13990
W. M. C. B. Weerakoon	S14028

ABSTRACT

Insurance companies deal with risk. To be successful they should manage that risk factor very well. Analyzing the data from the customers of the company and making an inspired decision is the best approach to manage risk. The insurance companies make revenue by establishing prices for insurable risks.

The dataset we used to analyze is of an insurance company in the United States. We developed a statistical learning model to predict the charges of customers using several factors of the customers and also, we found out the most important factors that are affecting the charges.

Using exploratory data analysis, we found out that the charges of smoking customers are significantly higher than those of non-smoking customers. The age and the charges of the customers have a strong positive relationship and also, we found out that the median BMI value of people from the southern region is higher than other regions.

We used Multiple Linear Regression using Stepwise Selection Method, Ridge Regression, and Lasso Regression methods for advanced analysis. All the fitted regression models gave a large RMSE value and the R^2 values were also not that significant and therefore, we used advanced regression techniques like Random Forest Regressor and Gradient Boosting Regressor to fit the model.

The Gradient Boost Algorithm gave the best test RMSE and adjusted R^2 value out of all the models which are 4635.075 and 84.26% respectively and it was selected as the most suitable model. Smoker, BMI, and Age were selected as the most important variables in predicting the insurance charges.

A website was created as the data product where the user can enter the data into the website and it predicts the relevant insurance charges according to the entered data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

01. INTRODUCTION

The insurance industry always collects and uses the data of customers to make decisions. In considering risk underwriting, it is important to use past data to make decisions to avoid losses. When creating and underwriting insurance policies it mainly considers the revenue. The profitability of insurance depends on how well it understands the risk it is insured against and how well it handles the costs. The premium must be sufficient to cover the expected claims for an insurance company to be profitable. Hence, it is clear that having knowledge about premium charges and the factors that the customers are insured against are important.

02. DESCRIPTION OF THE PROBLEM

Insurance companies mostly rely on data-driven decision making. Therefore, the problem that has been chosen for the final project is to predict the health insurance premium charges of the customers in a US health insurance company through their current customer data. Our main objective is to build a statistical learning model that helps to predict the premium charge, a customer should pay annually to the health insurance company when the details of the customer are given. And also, find the factors which have a major impact on the premium charges by doing a proper exploratory data analysis.

03. DESCRIPTION OF THE DATASET

The dataset is taken from the Kaggle Data sets. This data set contains 1338 details of insured customers where the insured premium charges are given against some attributes of the customers. There are 7 columns including the premium charges and the details of the attributes are given below.

Variable	**Type of Variable**	**Description**
Age	Integer | quantitative	Age of primary beneficiary
Sex	String | qualitative	Gender of the customer, Male or Female
BMI	Float | quantitative	Body mass index of the customer
Children	Integer | quantitative	No of children covered by the health insurance / No of dependents
Smoker	String | qualitative	Smoker - yes / non-smoker - no
Region	String | qualitative	The beneficiary's residential area in the US; Northeast, Southeast, Southwest, Northwest
Charges	Float | quantitative	Individual annual premium charges billed by health insurance

Table 03-1: Description of the dataset

04. METHODOLOGY

- **Adjusted R-Squared**

Adjusted R^2 shows how well terms fit a curve or line but adjusts for the number of terms in a model. In other words, adjusted R^2 tells us the percentage of variation explained by only the independent variables (Age, Sex, Children, BMI, Region, Smoker) that affect the dependent variable (Charges). Adjusted R^2 was used over R^2 since the addition of variables increases the R^2 value even though the added variables are not that significant.

- **Root Mean Squared Error (RMSE)**

The RMSE is the square root of the variance of the residuals. It indicates how close the observed data points are to the model's predicted values. Whereas R^2 is a relative measure of fit, RMSE is an absolute measure of fit. RMSE is a good measure of how accurately the model predicts the response. RMSE has the useful property of being in the same units as the response variable. Lower values of RMSE indicate a better fit.

- **Random Forrest Regressor**

A random forest is an estimator that fits several classifying decision trees on sub-samples of the dataset and improve the predictive accuracy and control over-fitting.

- **Gradient Boost Regressor**

Gradient boosting is a technique for building predictive models such as decision trees in regression and classification problems.

05. DESCRIPTIVE ANALYSIS

05.01. Analysis of Charges

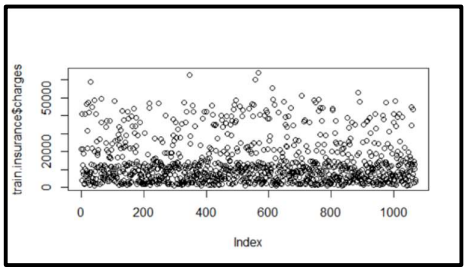


Figure 05-2: Scatterplot of Charges

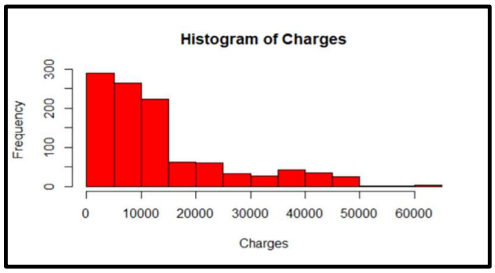


Figure 05-1: Histogram of Charges

The observations divide into 3 clear clusters according to the above figures as $0 to $15000, $15000 to $50000, and $50000 and above. This insurance company might have 3 different health insurance schemes with different premium charges and the clusters can be a result of that.

05.02. Analysis of Age

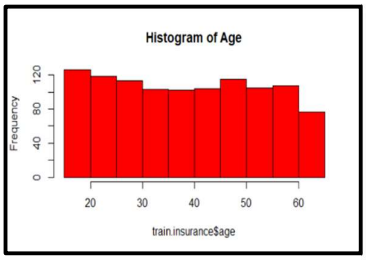


Figure 05-3: Histogram of Age

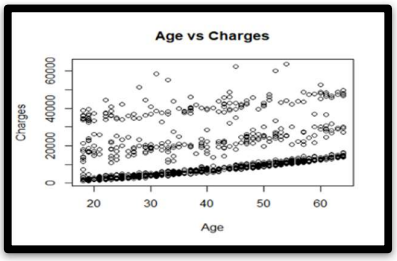


Figure 05-4: Scatterplot of Charges vs Age

According to *Figure 05-3*, A clear positive linear relationship can be observed between age and charges. This can be seen since the premium amount increases average about 8% to 10% for every year of age [1].

05.03. Analysis of Smoker

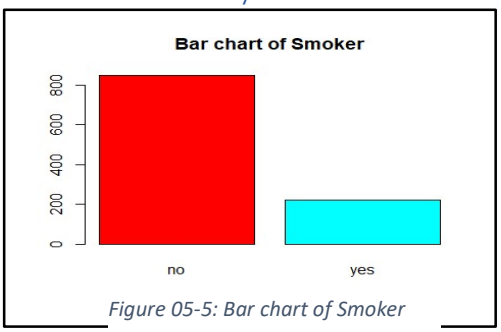


Figure 05-5: Bar chart of Smoker

The prevalence of smoking among U.S adults declined from 20.9 percent to 16.8 percent from 2005 to 2014 [2] and also, 80% of U.S adults believed that smoking cigarettes increased a person's risk of getting cancer [3]. The clear difference between the number of non-smoking customers and the number of smoking customers where the number of non-smoking customers is higher according to *Figure 05-5*, can be a result of the above-mentioned factors.

The prevalence of cigarette smoking was higher among males (18.8%) [2]. That can be clearly seen from *Figure 05-6,* where the percentage of smoking males is higher than the percentage of smoking females. The average number of years lost from average life expectancy due to smoking is 13.2 years [4] and therefore due to increased risk, premiums for smokers are 40% to 100% higher than for non-smokers, depending on your health history [5].

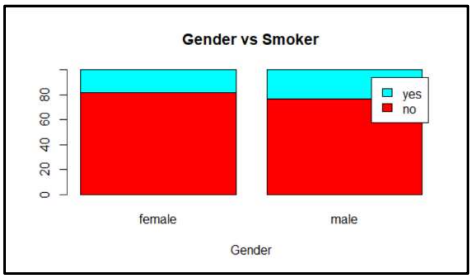


Figure 05-6: Stacked bar chart of Gender and Smoker

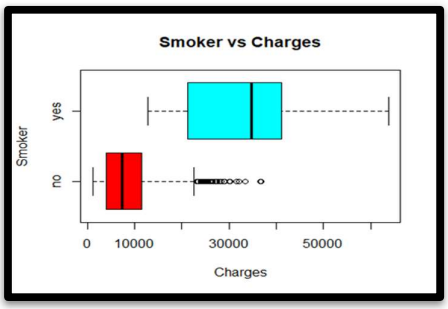


Figure 05-7: Group boxplots of Smoker vs Charges

The higher median insurance charge of people who smoke than the people who do not smoke justifies those facts. Moreover, all the charges paid by smoking customers are higher than 75% of charges paid by non-smokers. This means the insurance company should focus on attracting more people who smoke.

05.04. Analysis of BMI

The standard categories are underweight (BMI less than 18.5), normal (18.5–24.9), overweight (25–29.9), and obese (30 or more). According to these criteria, about one in three Americans are overweight but not obese, and an additional one in five are obese [6]. *Figure 05-9* also shows that 75% of people in our dataset are either overweight or obese.

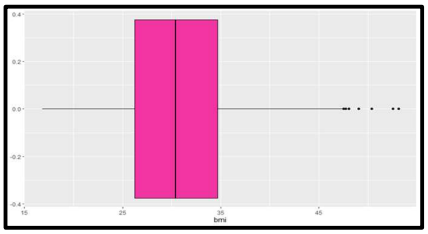


Figure 05-9: Boxplot of BMI

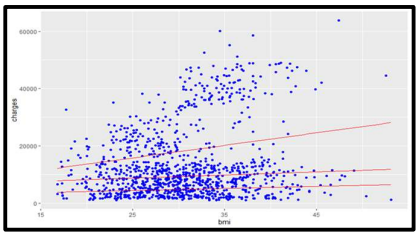


Figure 05-8: Scatterplot of BMI

A higher BMI, beginning in the upper range of the normal weight category, is associated with increased mortality and increased risk for coronary heart disease, osteoarthritis, diabetes mellitus, hypertension, and certain types of cancer [6]. Due to that, the premium charges are higher for people with high BMI. This fact is evidently proved with our dataset.

05.05. Analysis of Region

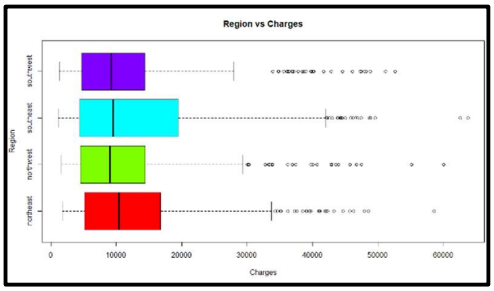


Figure 05-10: Group boxplots of Region vs Charges

The Northeast region consists of 6 out of 7 regions that do not increase insurance premium due to smoking [7]. Hence, the median charges of the northeast region should be lower than other regions. However, according to *Figure 05-10*, the northeast region has the highest median charges among all four regions. This can happen since the dataset consists a much smaller number of smokers than non-smokers.

There are significant health needs within the southern region. Southerners as a group are generally more likely than those in other regions to have a number of chronic illnesses and experience worse health outcomes [8]. A broad array of factors contributes to the high rates of chronic disease and poor health outcomes in the South and one factor is obesity. Our analysis also shows that Southeast and Southwest regions have a higher BMI value which means they have obesity.

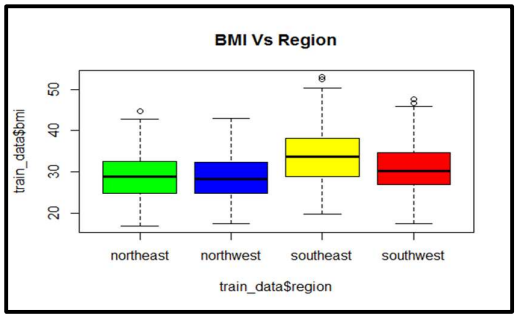


Figure 05-11: Group boxplots of BMI vs Region

05.06. Analysis of Children

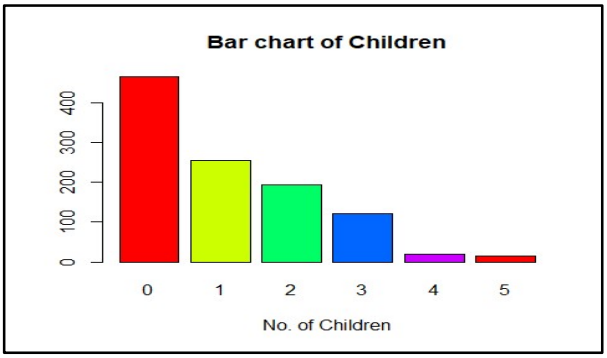


Figure 05-12: Bar chart of Children

Most customers do not have children and also when the number of children a customer has is increasing, the number of customers for each category is decreasing. Child care is too expensive, want more time for the children I have, worried about the economy are the main reasons for more Americans to have a lesser no. of children [9].

05.07. Analysis of Gender

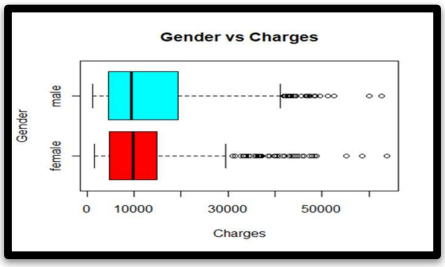


Figure 05-13: Group boxplots of Gender vs Charges

The figure shows that the median charges between male and female customers are approximately equal. But the interquartile range of the charges of male customers is higher than the female customers. Therefore, we can say the spread of the charges of male customers is higher.

05.08. Correlation Matrix

It is clear that from the above plots, the response variable which is "Charges" is highly correlated with the variable "Smoker" and the variables "age" and "Charges" are also correlated. There is some correlation between "Region" and "BMI" too. Apart from that, any substantial correlations cannot be seen among variables and hence we used multiple linear regression model for the advanced analysis of the dataset. Although there are no substantial correlations between the predictor variables, we used ridge and lasso regression methods in advanced analysis too.

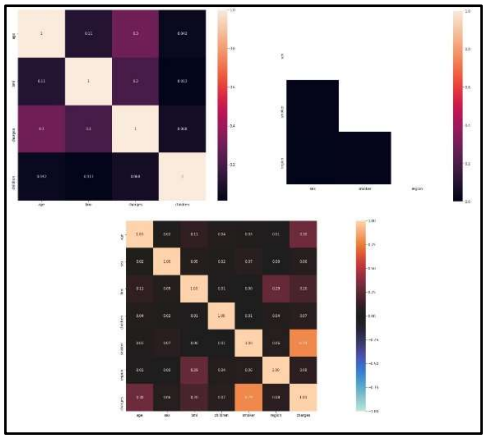


Figure 05-14: Correlation Matrices

06. ADVANCED ANALYSIS

It has been concluded to use regression techniques such as multiple linear regression with stepwise selection method, ridge regression, lasso regression, and elastic net regression from the results of the explorative data analysis. Root mean squared error of test and adjusted R^2 were selected as the evaluation metrics to compare the models.

```
> coef(best_ridge)
11 x 1 sparse Matrix of class "dgCMatrix"
                        s0
(Intercept)     -13670.8906
(Intercept)           .
age                259.3119
sex               -441.9567
bmi                358.8193
children           532.8738
smoker           23851.0867
region_northeast  1488.1838
region_northwest   942.6338
region_southeast  -326.0331
region_southwest  -106.9119
```

Table 06-2: Coefficients of Ridge Regression Model

The test RMSE and adjusted R^2 of the Stepwise Selection Model were 6056.135 and 0.7527322. The test RMSE and adjusted R^2 of Ridge Regression were 6098.291 and 0.7275261. The test RMSE and the adjusted R^2 value of Lasso Regression were 5616.884 and 0.755102 respectively. The test RMSE and the adjusted R^2 value of Elastic Net Regression were 5616.705 and 0.755118 respectively.

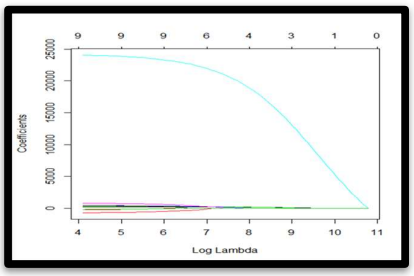


Figure 06-1: Elastic Net Regression

```
> coef(best_mod)
     (Intercept)              age              sex              bmi         children           smoker region_northeast
      -13899.9787         260.2115        -445.5107         357.4661         536.1679       23903.2970        1708.5439
region_northwest
       1165.5179
```

Table 06-1: Coefficients of Stepwise Selection Method

All the regression techniques that have been used gave a large RMSE and mediocre adjusted R^2 values and therefore, advanced models like Random Forest Regressor and Gradient Boosting Regressor were used to fit the model. Since our main objective is not the interpretation of predictors the advanced models were used.

	shrinkage	interaction.depth	n.minobsinnode	bag.fraction	optimal_trees	min_RMSE
1	0.30	5	3	0.65	16	4311.660
2	0.30	5	5	0.65	16	4311.660
3	0.30	5	7	0.65	16	4311.660
4	0.10	7	3	0.80	34	4314.806
5	0.10	7	5	0.80	34	4314.806
6	0.10	7	7	0.80	34	4314.806
7	0.10	5	3	0.65	39	4322.266
8	0.10	5	5	0.65	39	4322.266
9	0.10	5	7	0.65	39	4322.266
10	0.01	5	3	0.65	457	4325.361

Table 06-3: Grid search summary of Gradient Boosting Model

The RMSE values of the Random Forest Model and Gradient Boost Model were 4688.107 and 4635.075 respectively. The adjusted R^2 value of the Random Forest Model and Gradient Boost Model were 0.8361868 and 0.8425937 respectively.

```
Random Forest

1072 samples
   9 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 965, 965, 964, 964, 964, 965, ...
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
  1     8559.526  0.8019009  6350.588
  2     5459.933  0.8436181  3834.227
  3     4689.164  0.8544573  2883.696
  4     4555.189  0.8567914  2592.965
  5     4567.610  0.8553574  2542.447
  6     4584.637  0.8543098  2535.855
  7     4619.417  0.8523107  2555.753
  8     4643.816  0.8510409  2566.653
  9     4678.044  0.8491204  2587.891

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 4.
```

Table 06-4: Grid search summary of Random Forest Model

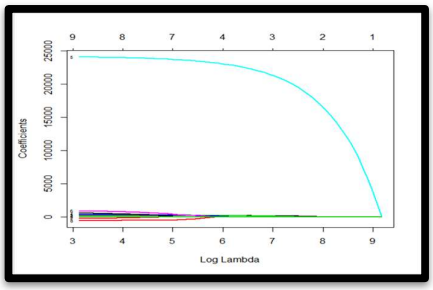


Figure 06-2: Lasso Regression

By comparing all the test RMSE and adjusted R^2 values of models, The Gradient Boost Algorithm gave the best test RMSE and adjusted R^2 value. This variable importance plot was drawn for the best model selected. The smoker, BMI, and age have a stand out importance from other variables. Hence, we can conclude that the annual premium charge of a customer is highly relying on those three factors. The variable region and sex have less importance in the model.

MODEL	RMSE	R-SQUARED
Stepwise Selection	6056.135	0.7527322
Ridge Regression	6098.291	0.7275261
Lasso Regression	5616.884	0.7551026
Elastic Net Regression	5616.705	0.7551182
Random Forest	4688.107	0.8361868
Gradient Boosting	4635.075	0.8425937

Table 06-5: Summary of all models

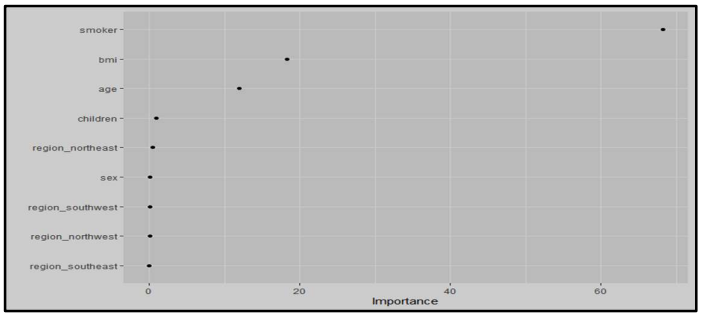


Figure 06-3: Variable importance Plot

07. DISCUSSION AND CONCLUSION

The insurance charges of the company to which the dataset belongs depend on various factors. There is a clear difference in the number of smokers and non-smokers in the dataset. Even though the number of non-smokers is higher than the number of smokers, the insurance charges of smokers are much superior to the charges of non-smokers with even the median charge of smokers is higher than the maximum charge of non-smokers. The charges of the company mainly depend on whether he is a smoker, the BMI of the customer, and the age of the person.

A proper website is created to predict the annual health insurance premium charges of US citizens using the Final statistical learning model. This website can be used by any health insurance company to get information on the estimated annual premium charge to be collected from a health insurance customer. This will improve their decision-making time and processing time.

The model can be improved by getting more observation since only three factors contributes to the annual premium charges. We can also improve the model by considering more factors such as pre-existing health condition, medical history, profession, etc.

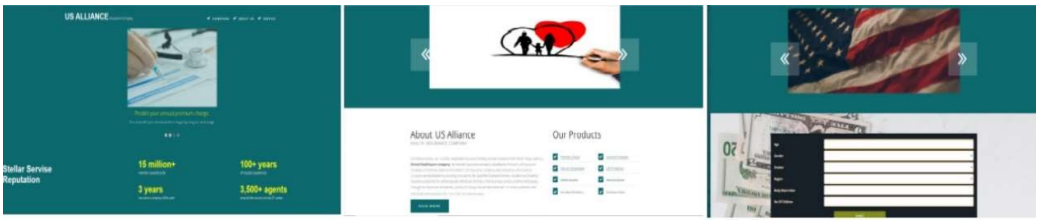


Figure 07-1: The Website

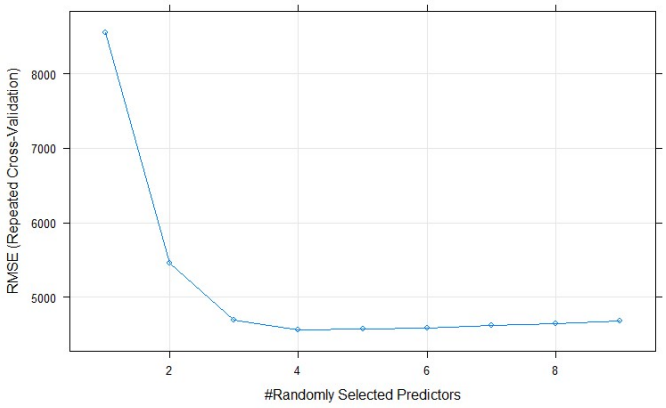
08. REFERENCES

1. Investopedia. 2020. *Life Insurance: How Much Does Age Raise Your Rate?*. [online] Available at: <https://www.investopedia.com/articles/personal-finance/022615/how-age-affects-life-insurance-rates.asp>.
2. Cdc.gov. 2015. *Current Cigarette Smoking Among Adults — United States, 2005–2014*. [online] Available at: <https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6444a2.htm>.
3. Cdc.gov. n.d. *What Are the Risk Factors for Lung Cancer? | CDC*. [online] Available at: <https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm>.
4. Statista. 2020. *Topic: Smoking*. [online] Available at: <https://www.statista.com/topics/1600/smoking/>.
5. Cooperators.ca. n.d. *Does smoking affect life insurance?*. [online] Available at: <https://www.cooperators.ca/en/Resources/protect-what-matters/life-insurance-for-smokers.aspx>.
6. Healthaffairs.org. 2020. *The Effects Of Obesity, Smoking, And Drinking On Medical Problems And Costs | Health Affairs Journal*. [online] Available at: https://www.healthaffairs.org/doi/10.1377/hlthaff.21.2.245
7. HealthMarkets. n.d. *What You Need to Know About Smoking and Health Insurance*. [online] Available at: <https://www.healthmarkets.com/content/smoking-and-health-insurance>.
8. KFF. 2016. *Health and Health Coverage in the South: A Data Update*. [online] Available at: <https://www.kff.org/racial-equity-and-health-policy/issue-brief/health-and-health-coverage-in-the-south-a-data-update/>.
9. Nytimes.com. 2018. *Americans Are Having Fewer Babies. They Told Us Why. (Published 2018)*. [online] Available at: <https://www.nytimes.com/2018/07/05/upshot/americans-are-having-fewer-babies-they-told-us-why.html>.

09. APPENDIX AND R CODE

```
library("caret")
library("glmnet")
library("vip")
library("ISLR")
library("dplyr")
library('fastDummies')
insurance_data = read.csv('F:/statCS/3rd year/Sem2/statistical learning/finalproject/insurance.csv')
View(insurance_data)
dim(insurance_data)
sum(is.na(insurance_data))
#no missing values
insurance_data$sex<-recode(insurance_data$sex,"male"=1, "female"=0)
insurance_data$smoker<-recode(insurance_data$smoker, "yes"=1, "no"=0)
insurance_data <- dummy_cols(insurance_data, select_columns = 'region')
insurance_data<-subset(insurance_data, select = -region )
set.seed(1234)
train_index <- createDataPartition(insurance_data$charges,p=0.8,list = FALSE,times = 1)
train_data <- insurance_data[train_index,]
test_data <- insurance_data[-train_index,]
Train_x = model.matrix(charges~. ,train_data)
Train_y =train_data$charges
Test_x = model.matrix(charges~.,test_data)
Test_y = test_data$charges
########################################## RIDGE REGRESSION#####################################
# Setting the range of lambda values
lambda_seq <- 10^seq(2, -2, by = -.1)
Ridge.fit = glmnet(Train_x,Train_y,alpha = 0,lambda  = lambda_seq)
summary(Ridge.fit)
lambdas <- Ridge.fit$lambda
#Doing k fold cross validation to select the best lambda
# Using cross validation glmnet
ridge_cv <- cv.glmnet(Train_x,Train_y,alpha = 0, lambda = lambdas)
# Best lambda value
best_lambda <- ridge_cv$lambda.min
best_lambda
best_fit <- ridge_cv$glmnet.fit
head(best_fit)
# Rebuilding the model with optimal lambda value
best_ridge <- glmnet(Train_x,Train_y, alpha = 0, lambda = best_lambda)
coef(best_ridge)
#predicting the test using ridge model
predicted_Ridge <- predict(best_ridge,s=best_lambda,Test_x)
predicted_Ridge
################################## STEP wise selection ###################################
# Set seed for reproducibility
set.seed(123)
# Set up repeated k-fold cross-validation
train.control <- trainControl(method = "cv", number = 10)
# Train the model
step.model <- train(charges~., data = train_data, method = "leapSeq", tuneGrid = data.frame(nvmax = 1:9), trControl = train.control)
step.model$results
step.model$bestTune
step.model
summary(step.model$finalModel)
coef(step.model$finalModel,7)
best_mod <- lm(charges~age+sex+bmi+children+smoker+region_northeast+region_northwest,data = train_data)
summary(best_mod)
View(test_data)
text_x <- test_data[,-6]
text_y <-test_data[,6]
predicted <- predict(step.model,text_x)
eval_results(text_y, predicted, test_data)–
```

```
######################################## RANDOM FOREST ##########################################
#Grid search
set.seed(1234)
mtry <- sqrt(ncol(Train_x))
tunegrid <- expand.grid(.mtry=mtry)
control <- trainControl(method="repeatedcv", number=10, repeats=3, search="grid")
tunegrid <- expand.grid(.mtry=c(1:9))
rf_gridsearch <- train(charges~., data=train_data, method="rf", tuneGrid=tunegrid, trControl=control)
print(rf_gridsearch)
plot(rf_gridsearch)
ran_mod_tuned <-randomForest(charges~.,data = train_data,mtry = 4, importance = TRUE, na.action = na.omit)
predicted_RF_tuned <- predict(ran_mod_tuned,Test_x)
SSE <- sum((predicted_RF_tuned - Test_y)^2)
SST <- sum((Test_y - mean(Test_y))^2)
R_square_RF <- 1 - SSE / SST
RMSE = sqrt(SSE/nrow(test_data))
R_square
RMSE
```




```
######################################## GRADIENT BOOSTING ########################################
# grid search
for(i in 1:nrow(hyper_grid)) {

  # reproducibility
  set.seed(123)

  # train model
  gbm.tune <- gbm(
    formula = charges ~ .,
    distribution = "gaussian",
    data = train_data,
    n.trees = 1000,
    interaction.depth = hyper_grid$interaction.depth[i],
    shrinkage = hyper_grid$shrinkage[i],
    bag.fraction = hyper_grid$bag.fraction[i],
    train.fraction = .75,
    n.cores = NULL, # will use all cores by default
    verbose = FALSE
  )
  # add min training error and trees to grid
  hyper_grid$optimal_trees[i] <- which.min(gbm.tune$valid.error)
  hyper_grid$min_RMSE[i] <- sqrt(min(gbm.tune$valid.error))
}
hyper_grid %>%
  dplyr::arrange(min_RMSE) %>%
  head(10)
set.seed(123)
# train GBM model
gbm.fit.final <- gbm( formula = charges ~ .,distribution = "gaussian",data = train_data, n.trees = 16,interaction.depth = 5, shrinkage = 0.3, n.minobsinnode = 3,
bag.fraction = 0.65,train.fraction = 1,n.cores = NULL,verbose = FALSE
)
#PREDICT
pred <- predict(gbm.fit.final, n.trees = gbm.fit.final$n.trees, test_data)
pred
Test_y
# results
caret::RMSE(pred, Test_y)
SSE <- sum((pred - Test_y)^2)
SST <- sum((Test_y - mean(Test_y))^2)
R_square_gbm <- 1 - SSE / SST
R_square_gbm
vip(gbm.fit.final, num_features=10, geom ="point")
```

```r
#helper packages
library(recipes) #for feature engineering

#Modeling packages
library(glmnet) #for implementing regularized regression
library(caret) #for automating the tunining process
library(pdp) #for partial dependency

#Model interpretability packages
library(vip) #for variable importance

data = read.csv('D:/3_year/3_year/2_sem/STAT/StatLearning_1/Final_Project/insurance.csv')

attach(data)
summary(data$charges)

sum(is.na(data))
data$sex<-recode(data$sex,"male"=1, "female"=0)
data$smoker<-recode(data$smoker, "yes"=1, "no"=0)
data <- dummy_cols(data, select_columns = 'region')
data<-subset(data, select = -region )

#training and testing datasets
#Using caret package
set.seed(1234)
train_index <- sample(1:nrow(data), 0.8 * nrow(data))
test_index <- setdiff(1:nrow(data), train_index)
train.insurance <- data[train_index,]
test.insurance <- data[test_index,]
```

```r
#### Lasso Regression ###

x1 = model.matrix(charges~.-1, train.insurance)
dim(x1)
tx1 = model.matrix(charges~.-1,test.insurance)
y1 = (train.insurance$charges)
ty1 =(test.insurance$charges)


fit.lasso1 = glmnet(x1,y1,alpha = 1)
plot(fit.lasso1,xvar = "lambda",label = TRUE, lw=2)
#plot(fit.lasso,xvar = "dev", label= TRUE , lw =2)

#Doing cross validation to select the best lambda
set.seed(5)
cv.lasso1 = cv.glmnet(x1,y1,alpha=1)
plot(cv.lasso1, main =" Lasso penalty\n\n")
lam.best1 =cv.lasso1$lambda.min #select lambda that min training MSE
#coefficient vector under the one std of the best lambda
min(cv.lasso1$cvm) #min MSE
cv.lasso1$lambda.min #lambda for this min MSE


plot(fit.lasso1,xvar="lambda", main="Lasso penalty\n\n")
abline(v=log(cv.lasso1$lambda.min),col ="red", lty= "dashed")
abline(v=log(cv.lasso1$lambda.1se),col ="green", lty= "dotted")

lasso_pred1 = predict(fit.lasso1, s = lam.best1, newx =tx1) #use best lambda to predict test data

caret::RMSE(lasso_pred1, ty1)

d1= data.frame(lasso_pred1,ty1)
RSS1 = sum((lasso_pred1-ty1)^2)
TSS1 = sum((ty1-mean(ty1))^2)
1-(RSS1/TSS1)
```

```r
#### Elastic Net Regression ###
enet1 = glmnet(x1,y1, alpha=0.2)
plot(enet1,xvar ="lambda")

set.seed(123)

cv_glmnet1 = train(x1,y1,method ="glmnet", preProcess = c("zv","center","scale"),
                   trcontrol= trainControl(method="cv", number="10"), tuneLength =10)
cv_glmnet1
cv1 = cv_glmnet1$bestTune$lambda
cv2= cv_glmnet1$bestTune
cv2
fit.en1= glmnet(x1,y1,alpha = 0.1)
pred1 = predict(fit.en1,s= cv1,newx = tx1)

#mean((pred1-ty1)^2)
caret::RMSE(pred1, ty1)

out_e1 = glmnet(x1,y1,alpha=0.1) #fit full data set
enet_coeff1 = predict(out_e1, type="coefficients", s= cv1)[1:10,]
enet_coeff1

RSS_e1 = sum((pred1-ty1)^2)
TSS_e1 = sum((ty1-mean(ty1))^2)
1-(RSS_e1/TSS_e1)

#### Feature interpretation ###

vip(cv_glmnet1, num_features=20, geom ="point")

pl1 = partial(cv_glmnet1,"smokeryes",grid.resolution =20, plot=TRUE)
pl2 = partial(cv_glmnet1,"age",grid.resolution =20, plot=TRUE)
pl3 = partial(cv_glmnet1,"bmi",grid.resolution =20, plot=TRUE)
pl4 = partial(cv_glmnet1,"children",grid.resolution =20, plot=TRUE)
grid.arrange(pl1,pl2,pl3,pl4,pl5,pl6,ncol=2,nrow=3)
```

```r
library("caret")
library("glmnet")
library("vip")
library("ISLR")
library("dplyr")
library('fastDummies')
library("VIF")
library("xgboost")
library("gbm")
insurance_data = read.csv('D:/3_year/3_year/2_sem/STAT/StatLearning_1/Final_Project/insurance.csv')


insurance_data$sex<-recode(insurance_data$sex,"male"=1, "female"=0)
insurance_data$smoker<-recode(insurance_data$smoker, "yes"=1, "no"=0)
insurance_data <- dummy_cols(insurance_data, select_columns = 'region')
insurance_data<-subset(insurance_data, select = -region )

fix(insurance_data)

set.seed(1234)
train_index <- createDataPartition(insurance_data$charges,p=0.8,list = FALSE,times = 1)
train_data <- insurance_data[train_index,]
test_data <- insurance_data[-train_index,]

Train_x = model.matrix(charges~. ,train_data)[,-1]
Train_y =train_data$charges
Test_x = model.matrix(charges~.,test_data)[,-1]
Test_y = test_data$charges
```

```r
hyper_grid <- expand.grid(
  shrinkage = c(.01, .1, .3),
  interaction.depth = c(3, 5, 7),
  n.minobsinnode = c(3,5,7),
  bag.fraction = c(.65, .8, 1),
  optimal_trees = 0,               # a place to dump results
  min_RMSE = 0                     # a place to dump results
)

# grid search
for(i in 1:nrow(hyper_grid)) {

  # reproducibility
  set.seed(123)

  # train model
  gbm.tune <- gbm(
    formula = charges ~ .,
    distribution = "gaussian",
    data = train_data,
    n.trees = 1000,
    interaction.depth = hyper_grid$interaction.depth[i],
    shrinkage = hyper_grid$shrinkage[i],
    bag.fraction = hyper_grid$bag.fraction[i],
    train.fraction = .75,
    n.cores = NULL, # will use all cores by default
    verbose = FALSE
  )

  # add min training error and trees to grid
  hyper_grid$optimal_trees[i] <- which.min(gbm.tune$valid.error)
  hyper_grid$min_RMSE[i] <- sqrt(min(gbm.tune$valid.error))
}
```

```r
hyper_grid %>%
  dplyr::arrange(min_RMSE) %>%
  head(10)


set.seed(123)

# train GBM model
gbm.fit.final <- gbm(
  formula = charges ~ .,
  distribution = "gaussian",
  data = train_data,
  n.trees = 16,
  interaction.depth = 5,
  shrinkage = 0.3,
  n.minobsinnode = 3,
  bag.fraction = 0.65,
  train.fraction = 1,
  n.cores = NULL, # will use all cores by default
  verbose = FALSE
)

pred <- predict(gbm.fit.final, test_data)

# results
caret::RMSE(pred, Test_y)
SSE <- sum((pred - Test_y)^2)
SST <- sum((Test_y - mean(Test_y))^2)
R_square_gbm <- 1 - SSE / SST
R_square_gbm

saveRDS(gbm.fit.final,"model1.rds")
```

Website

```
from flask import Flask,request,render_template,redirect,url_for
import numpy as np
import joblib

model = joblib.load("GBR.joblib")

# start flask
app = Flask(__name__)

# render default webpage
@app.route('/')
def home():
    return render_template('home.html')

@app.route('/about')
def about():
    return render_template('index.html')

@app.route('/service')
def service():
    return render_template('services.html')

@app.route('/predict', methods=['POST'])
def get_data():
        age = request.form['Age']
        sex = request.form['sex']
        bmi = request.form['bmi']
        noc = request.form['noc']
        smoker = request.form['smoker']
        region = request.form['region']
                                age = int(age)
        bmi = float(bmi)
        noc = int(noc)

        if sex == "Male":
                                sex_male = 1
        else :
                                sex_male =0

        if smoker == "Yes":
                                smoker_yes = 1
        else :
                                smoker_yes =0

        if region == "southeast":
                                region_northwest = 0
                                region_southeast = 1
                                region_southwest = 0
        elif region == "southwest":
                                region_northwest = 0
                                region_southeast = 0
                                region_southwest = 1
        elif region == "northwest":
                                region_northwest = 1
                                region_southeast = 0
                                region_southwest = 0
        else :
                                region_northwest = 0
                                region_southeast = 0
                                region_southwest = 0



        arr = np.array([[age,bmi,noc,sex_male,smoker_yes,region_northwest,region_southeast,region_northwest]])
        pred = model.predict(arr)

        return render_template('services.html',data=pred)

if __name__ == '__main__':
    app.run(debug=True,port=5000)


##########################################################################################################
#################################################################

<!DOCTYPE html PUBLIC " //W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">

<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<title>US Alliance</title>
<meta name="keywords" content="" />
<meta name="description" content="" />
<script type="text/javascript" src="http://ajax.googleapis.com/ajax/libs/jquery/1.7.1/jquery.min.js"></script>
<script type="text/javascript" src="../static/jquery.slidertron-1.3.js"></script>
<link href="http://fonts.googleapis.com/css?family=Open+Sans:400,300,600,700,800" rel="stylesheet" />
<link href="../static/default.css" rel="stylesheet" type="text/css" media="all" />
<link href="../static/fonts.css" rel="stylesheet" type="text/css" media="all" />

<!--[if IE 6]><link href="default_ie6.css" rel="stylesheet" type="text/css" /><![endif]-->

</head>
<body>
                                <div id="header-wrapper">
                                                <div id="header" class="container">
                                                                <div id="logo">
                                                                                <h1><a
href="#">US Alliance</a></h1>
                                                                                <span
style="color: white">Insurance Company</span>
                                                                </div>
                                                                <div id="menu">
                                                                                <ul>
                                                                                <li class="icon icon-ok current_page_item"><a href="/" accesskey="1" title="">Homepage</a></li>
                                                                                <li class="icon
icon-ok"><a href="/about" accesskey="3" title="">About Us</a></li>
                                                                                <li class="icon
icon-ok"><a href="/service" accesskey="5" title="">Service</a></li>
                                                                                </ul>
                                                                </div>
                                                </div>
</div>
                                                <!--
**********************************************************************************************************
***** -->
                                <div id="slider">
                                                <a href="#" class="button previous-
button"><span class="icon icon-double-angle-left"></span></a>
                                                <a href="#" class="button next-
button"><span class="icon icon-double-angle-right"></span></a>
                                                <div class="viewer">
                                                                <div
class="reel">
                                <div class="slide">
                                                <img src="../static/images/bg6.jpg" alt="" style="width: 100%;height:
100%" />
                                </div>
                                <div class="slide">
                                                <img src="../static/images/bg1.jpg" alt="" style="width: 100%;height:
100%"/>
                                </div>
                                <div class="slide">
                                                <img src="../static/images/bg5.jpg" alt="" style="width: 100%;height: 100%
"/>
                                </div>
                                                                </div>
                                                </div>
                                                <script type="text/javascript">

                                $('#slider').slidertron({

                                viewerSelector: '.viewer',

                                reelSelector: '.viewer .reel',

                                slidesSelector: '.viewer .reel .slide',

                                advanceDelay: 3000,

                                speed: 'slow',

                                navPreviousSelector: '.previous-button',

                                navNextSelector: '.next-button',

                                slideLinkSelector: '.link'
                                                                });
                                                </script>
                                                <!--
**********************************************************************************************************
***** -->

<div id="wrapper">
                                <div id="page" class="container">
                                                <div id="content">
                                                                <div class="title">
                                                                                <h2>About US
Alliance</h2>
                                                                                <span
class="byline">Health Insurance Company</span> </div>
                                                                <p>
                                                                                US Alliance
Group, Inc. is a fully integrated insurance holding company based in Fort Worth, Texas, and is a <b>United Healthcare company</b>.
Its licensed insurance company subsidiaries <i>Freedom Life Insurance Company of America</i>, <i>National Foundation Life
Insurance Company</i>, and <i>Enterprise Life Insurance Company</i> are dedicated to providing innovative Life, Specified
Disease/Sickness, Accident and Disability Insurance solutions for self-employed individuals, families, small business owners and their
employees. Through its insurance companies, USHEALTH Group has served more than 15 million customers with individually tailored
plans for more than 50 collective years.
                                                                                </p>
                                                                <a href="#" class="button">Read More</a>
</div>
                                                <div id="sidebar">
                                                                <div class="title">
                                                                                <h2>Our Products</h2>
                                                                                </div>
                                                                <div id="stwo-col">
                                                                                <div
class="sbox1">


                                                <ul class="style2">

                                                                <li class="icon icon-ok"><a href="#">Premier Choice</a></li>

                                                                <li class="icon icon-ok"><a href="#">Secure Advanatage</a></li>

                                                                <li class="icon icon-ok"><a href="#">Health Access</a></li>

<li class="icon icon-ok"><a href="#">Accident Protector</a></li>


                                                </ul>
                                                                                </div>
                                                                                <div
class="sbox2">


                                                <ul class="style2">

                                                                <li class="icon icon-ok"><a href="#">Income Protector</a></li>

                                                                <li class="icon icon-ok"><a href="#">Life Protector</a></li>

                                                                <li class="icon icon-ok"><a href="#">Secure Dental</a></li>

                                                                <li class="icon icon-ok"><a href="#">Premier Vision</a></li>


                                                </ul>
                                                                                </div>
                                                                </div>
                                                </div>
                                <div id="banner-wrapper">
                                                <div id="banner" class="container">
                                                                <div class="box-left">
                                                                                <h2>HEALTH
INSURANCE</h2>
                                                                                <span>Let us
help you customize your health coverage today.</span> </div>
                                                                <div class="box-right"> <img
src="../static/images/LOGO.png" alt="" style="width: 500px;height: 250px" /></div>
                                                </div>
                                </div>
                                <div id="page" class="container">
                                                <div id="content">
                                                                <div class="title">
                                                                                <h2>Legal
Notice</h2> </div>
                                                                <p>
                                                                                All products are
underwritten and issued by Freedom Life Insurance Company of America, National Foundation Life Insurance Company and Enterprise
Life Insurance Company, wholly owned subsidiaries of US Alliance, Inc. All products not available in all states. Contact a Licensed
Insurance Agent for additional information.
                                                                                </p></div>
                                                <div id="sidebar">
                                                                <div class="title">
                                <h2>Contact Details</h2>
                                                                                </div>
                                                                <div id="stwo-col">
                                                                                <div
class="sbox3">
                                <ul class="style2">

                                                <li><img src="../static/images/location.png" alt="" style="width:
25px;height: 25px" /><a href="https://goo.gl/maps/mzoTsvRbABeNs7tR7">300 Burnett St, Suite 200 Fort Worth, TX
76102,USA</a></li>

                                                <li><img src="../static/images/phone.png" alt="" style="width:
25px;height: 25px"/> &amp;nbsp; &amp;nbsp; +1-866-722-8744</li>

                                                <li><img src="../static/images/email.png" alt="" style="width: 25px;height:
25px" > <a href="#">info@ushealthgroup.com</a></li>

                                                <li><img src="../static/images/clock.png" alt="" style="width: 25px;height:
25px" /> &amp;nbsp; &amp;nbsp; 7:00 AM - 7:00 PM CT Monday - Friday</li>

                                                <li><img src="../static/images/megaphone.png" alt="" style="width:
25px;height: 25px" />&amp;nbsp; &amp;nbsp; Media Inquiries : <a href="#">media@ushealthgroup.com</a></li>

                                                </ul>
                                                                                </div>
                                                                </div>
                                                </div>
</div>
<div id="copyright" class="container">
                                                <p>&copy; Untitled. All rights reserved. | Photos by <a href="http://fotogrph.com/">Roffliexgraphy</a>
| Design by <a href="http://templated.co" rel="nofollow">UOC FOS</a>.</p>
</div>
</body>
</html>

##########################################################################################################
#################################################################################

<!DOCTYPE html PUBLIC " -//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">

<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<title>US ALLIANCE</title>
<meta name="keywords" content="" />
<meta name="description" content="" />
<script type="text/javascript" src="http://ajax.googleapis.com/ajax/libs/jquery/1.7.1/jquery.min.js"></script>
<script type="text/javascript" src="../static/jquery.slidertron-1.3.js"></script>
<link href="http://fonts.googleapis.com/css?family=Open+Sans:400,300,600,700,800" rel="stylesheet" />
<link href="../static/h1.CSS" rel="stylesheet" type="text/css" media="all" />
<link href="../static/default.css" rel="stylesheet" type="text/css" media="all" />
<link href="../static/fonts.css" rel="stylesheet" type="text/css" media="all" />

<!--[if IE 6]><link href="default_ie6.css" rel="stylesheet" type="text/css" /><![endif]-->

</head>
<body>
<div id="header-wrapper">
                                <div id="header" class="container">
                                                <div id="logo">
                                                                <h1><a href="#">US Alliance</a></h1>
                                                                <span style="color: white">Insurance
Company</span>
                                                </div>
                                                <div id="menu">
                                                                <ul>
                                                                <li class="icon
icon-ok current_page_item"><a href="/" accesskey="1" title="">Homepage</a></li>
                                                                <li class="icon
icon-ok"><a href="/about" accesskey="3" title="">About Us</a></li>
                                                                <li class="icon
icon-ok"><a href="/service" accesskey="5" title="">Service</a></li>
                                                                </ul>
                                                </div>
                                </div>
</div>
                                <!--
**********************************************************************************************************
***** -->
                                <div class="slideshow-container">
                                <div
class="mySlides fade">
                                <div class="numbertext">1 / 4</div>
                                <img src="../static/images/image_6.jpg" style="width:100%">
                                <div class="text">What is US ALLIANCE?</div>
                                <p style="text-align: center; color:white;">US ALLIANCE is US's largest and undisputed leader in life
insurance. Going from strength to strength with innovative products and services,
                                                whilst reinforcing our trust and reputation amongst the people, we have
grown to become a household name in every corner of the country and a benchmark of life insurance in US.
                                </p>
                                <div
class="mySlides fade">
                                <div class="numbertext">2 / 4</div>
                                <img src="../static/images/image_5.jpg" style="width:100%">
                                <div class="text">Better level of innovative health coverage for people just like you</div>
                                <p style="text-align: center; color:white;">US ALLIANCE insurance company offers a full portfolio of
plans that let you tailor health coverage to YOUR specific needs.
                                                With over 50 collective years of insurance experience, we provide a broad
choice of flexible solutions and secure insurance plans.
                                                All you need to do is contact an agent to find the health coverage that best
suits you and your family's individual needs!
                                </p>
                                <div
class="mySlides fade">
                                <div class="numbertext">3 / 4</div>
                                <img src="../static/images/image_4.jpg" style="width:100%">
                                <div class="text">Customize your package.</div>
                                <p style="text-align: center; color:white;"> Our insurance plans offer an array of options for customers
on a limited budget, as well as, those who are looking for additional benefits.
                                                No matter the type of plan you choose, we work with you to build the
health coverage that's best for you and your family. For more details you can refer our About us page.
                                </p>
                                <div
class="mySlides fade">
                                <div class="numbertext">4 / 4</div>
                                <img src="../static/images/image_3.jpg" style="width:100%">
                                <div class="text">Predict your annual premium charge.</div>
                                <p style="text-align: center; color:white;"> You can predict your annual premium charges by using our
service page.
                                </p>
                                                                <br>
<div style="text-align:center">
                                <span class="dot"></span>
                                <span class="dot"></span>
                                <span class="dot"></span>
                                <span class="dot"></span>
                                                                </div>

                                                                <script>
                                var slideIndex = 0;
                                showSlides();

                                function showSlides() {
                                                var i;
                                                var slides = document.getElementsByClassName("mySlides");
                                                var dots = document.getElementsByClassName("dot");
                                                for (i = 0; i < slides.length; i++) {
                                                                slides[i].style.display = "none";
                                                }
                                                slideIndex++;
                                                if (slideIndex > slides.length) {slideIndex = 1}
                                                                for (i = 0; i < dots.length; i++) {
                                                dots[i].className = dots[i].className.replace(" active", "");
                                                                }
                                                slides[slideIndex-1].style.display = "block";
                                                dots[slideIndex-1].className += " active";
                                                setTimeout(showSlides,9000); // Change image every 10 seconds
                                }
                                                                                </script>
                                                                                <div id="page"
class="row">
                                <div class="column left" >
                                                <br>
                                                <h1 style="color:white;"> Stellar Servise </h1>
                                                <h1 style="color:white;"> Reputation </h1>
                                </div>
                                <div class="column right" >
                                                <h1 style="color:yellow;">15 million+ </h1>
                                                <p style="color:white; ">member covered so far</p>
                                                <br>
                                                <h1 style="color:yellow;">3 years</h1>
                                                <p style="color:white; ">insurance company of the year</p>
                                </div>
                                <div class="column left1" >
                                                <h1 style="color:yellow;"> 100+ years </h1>
                                                <p style="color:white; ">of industry experience</p>
                                                <br>
                                                <h1 style="color:yellow;">3,500+ agents </h1>
                                                <p style="color:white; ">around the country across 31 states</p>
                                </div>
                                                                                </div>

                                                <!--
**********************************************************************************************************
***** -->

<div id="wrapper">
                                <div id="banner-wrapper">
                                                <div id="banner" class="container">
                                                                <div class="box-left">
                                                                                <h2>HEALTH
INSURANCE</h2>
                                                                                <span>Let us
help you customize your health coverage today.</span> </div>
                                                                <div class="box-right"> <img
src="../static/images/LOGO.png" alt="" style="width: 500px;height: 250px" /></div>
                                                </div>
                                <div id="page" class="container">
                                                <div id="content">
                                                                <div class="title">
                                                                                <h2>Legal
Notice</h2> </div>
                                                                <p>
                                                                                All products are
underwritten and issued by Freedom Life Insurance Company of America, National Foundation Life Insurance Company and Enterprise
Life Insurance Company, wholly owned subsidiaries of US Alliance, Inc. All products not available in all states. Contact a Licensed
Insurance Agent for additional information.
                                                                                </p>
                                                                                </div>
                                                <div id="sidebar">
                                                                <div class="title">
                                <h2>Contact Details</h2>
                                                                                </div>
                                                                <div id="stwo-col">
                                                                                <div
class="sbox3">
                                <ul class="style2">

                                                <li><img src="../static/images/location.png" alt="" style="width:
25px;height: 25px" /><a href="https://goo.gl/maps/mzoTsvRbABeNs7tR7">300 Burnett St, Suite 200 Fort Worth, TX
76102,USA</a></li>

                                                <li><img src="../static/images/phone.png" alt="" style="width:
25px;height: 25px"/> &amp;nbsp; &amp;nbsp; +1-866-722-8744</li>

                                                <li><img src="../static/images/email.png" alt="" style="width: 25px;height:
25px" > <a href="#">info@ushealthgroup.com</a></li>

                                                <li><img src="../static/images/clock.png" alt="" style="width: 25px;height:
25px" /> &amp;nbsp; &amp;nbsp; 7:00 AM - 7:00 PM CT Monday - Friday</li>

                                                <li><img src="../static/images/megaphone.png" alt="" style="width:
25px;height: 25px" />&amp;nbsp; &amp;nbsp; Media Inquiries : <a href="#">media@ushealthgroup.com</a></li>

                                                </ul>
                                                                                </div>
                                                                </div>
                                                </div>
</div>
<div id="copyright" class="container">
                                                <p>&copy; Untitled. All rights reserved. | Photos by <a href="http://fotogrph.com/">Roffliexgraphy</a>
| Design by <a href="http://templated.co" rel="nofollow">UOC FOS</a>.</p>
</div>
</body>
</html>
```