Predictive Factors of Insurance Charges

2024-05-09

Introduction

This dataset is the US Health Insurance Dataset. It contains information about health insurance premiums charged to individuals in the United States. It contains 1338 observations and 7 columns.

age: Age of primary beneficiary

sex: The gender of the insurance contractor, either male or female

bmi: (Body mass index), An index calculated from the ratio of height to weight, providing an understanding of body weight relative to height. Ideally falls between 18.5 to 24.9 kg/m^2.

children: How many dependents or children are covered by health insurance

smoker: Shows whether the primary beneficiary is a smoker or non-smoker

region: The residential area of the beneficiary within the United States, categorized into regions such as northeast, southeast, southwest, and northwest

charges: Individual medical costs billed by health insurance, representing the amount charged for healthcare services provided to the beneficiary.

Insurance Premium Prediction:

What are the key factors influencing insurance charges, and how effectively can they be utilized for predicting insurance charges?

By using variables such as age,BMI,smoker status,region and no of children, predict the insurance charges.It acknowledges the need to understand the factors that impact insurance charges and assesses the effectiveness of utilizing these factors for predictive purposes.These are the prediction problems for insurance charges.

Before splitting the data set I have performed data pre-processing steps:

The pre-processing steps include finding missing values,structure of the data how it presents, and converting character datatype to factor data type.The pre processing steps are finding missing values,structure of the dataset,converted data types from character datatype to factor data type.The description of each pre processing explained in the below steps.

Statistical learning strategies and methods:

In this section performed exploratory data analysis using the training set, feature creation and feature selection.Step by step process is described below.

Predictive analysis and results:

Used linear regression and random forest models, cross-validation as resampling methods,and evaluated the performance on the test data.The description of each steps explained in the below .

Reading the dataset

Reading the data set from the read.csv() .The csv file name is "insurance.csv".After reading the data set.

```
# Setting the working directory where insurance.csv located
setwd("C:/Users/jalad/Downloads")

# Read the CSV file
data <- read.csv("insurance.csv")
head(data)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

Structure

Finding the structure of the data by using the str(). This gives the information about the datatypes, no of observations and columns.There are three different types of data integer,numerical and character.In the below shows that result of structure.

```
# structure of the data
str(data)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr  "female" "male" "male" "male" ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

Missing values

Checking the missing values in the data set. written code to count missing values in column by using colSums().But, their is no missing values in the data set.There are no missing values in the data set.

```
# Counting the missing values in each column
missing_count <- colSums(is.na(data))

# Printing the result
print(missing_count)
```

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

Converted data types:

Converted character data type to factor data type.The region,sex,smoker columns has character datatypes.So I converted them to factor.

```
# factor is used to convert character to factor data type
data$sex <- factor(data$sex)
data$smoker <- factor(data$smoker)
data$region <- factor(data$region)
```

Splitting the dataset

Setting the seed for reproducibility: set.seed(123) is used , that whenever I split the data randomly we get the same split each time when we run the code. This is important because it ensures that our results stay consistent even if we run the code multiple times.

Splitting the data into training and testing sets: createDataPartition function from the caret package. Parameter p tells us what proportion of data to put into training set, I have given 0.8 which means 80% of the data in training set and 20% of the data in testing set. The training data is used to train the model.The testing set used for evaluating the model's performance, can be created by selecting the rows that are not included in the training set. After splitting with the train data set I am performing the exploratory data analysis for train data and feature selection.Then using models to predict charges and evaluating the models performance on the test data.

```
# Loading required library
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.2
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

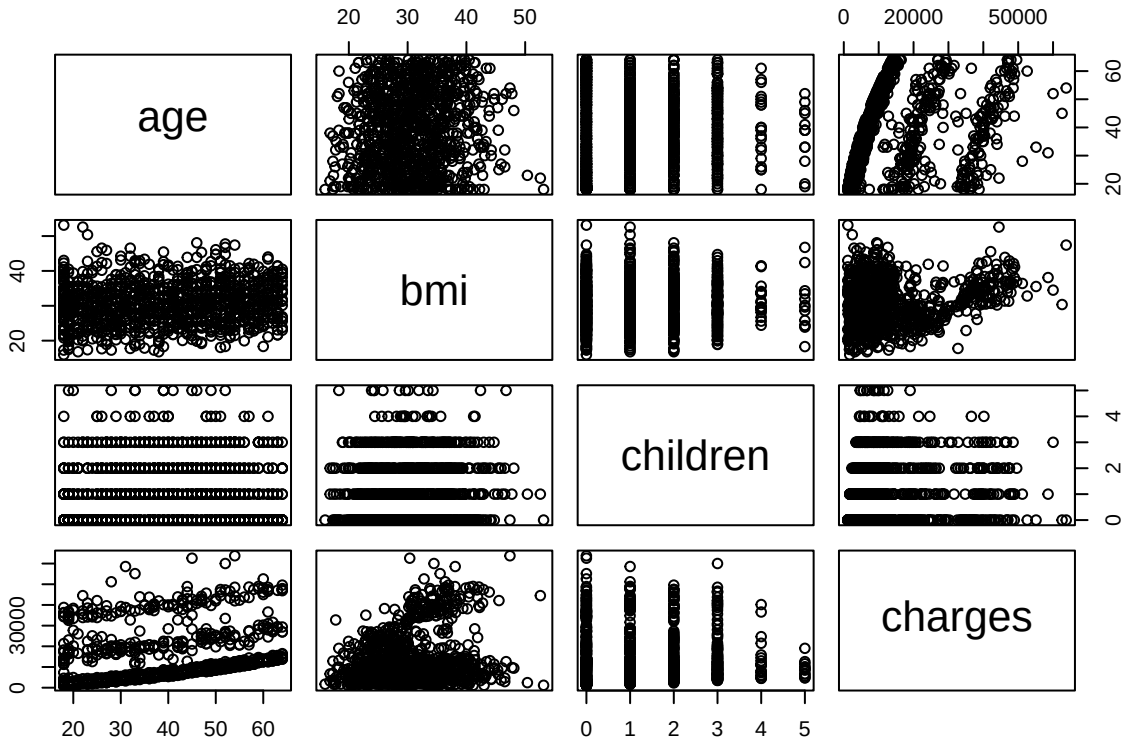
```
# Set seed for reproducibility
set.seed(123)

# Splitting the data into training and testing sets
trainIndex <- createDataPartition(data$charges, p = 0.8, list = FALSE)
trainData <- data[trainIndex, ]
testData <- data[-trainIndex, ]
```

Statistical learning strategies and methods:

Exploratory data analysis using the training set:

Created a pairs plot.Plotted scatter plot for pairs of variables using the pairs() function. (age,bmi,children,charges) these variables are quantitative.

```r
#generates a scatterplot matrix using the pairs()
pairs(trainData[, c("age", "bmi", "children", "charges")],
      panel = function(x, y) {
        points(x, y)
      })
```



age vs bmi: As age increases, there is a tendency for bmi values to increase as well, though the pattern is quite scattered.

age vs children: The number of children tends to rise with increasing age up to a certain point, then decreases at older ages.

age vs charges: As age goes up charges generally tend to be higher though there is substantial spread or variation in the data.

bmi vs children: The relationship between bmi and children count appears relatively weak or non-existent, with children counts scattered across the range of bmi values.

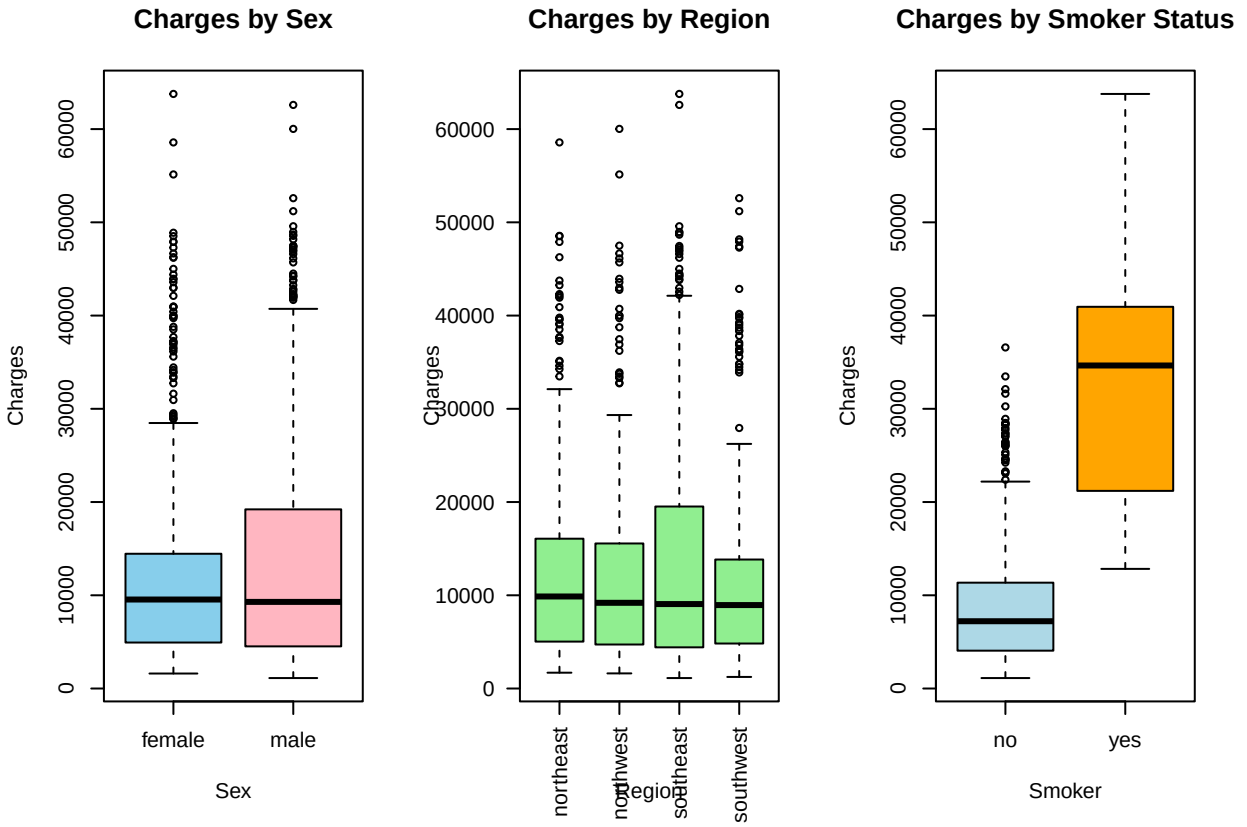
bmi vs charges: Higher bmi values tend to have higher associated charges, but there is significant variability or spread in the charges across different bmi levels.

children vs charges: The plot indicates a subtle positive association between the number of children and charges, suggesting that individuals with more children tend to have slightly higher charges. However, this relationship is not strongly related From this visually exploring the relationships between multiple variables in a dataset.It helps to understand what might affect how much people pay for insurance , Age and BMI appear to be associated with higher charges while the relationship between the number of children and charges is less clear.

Boxplot for qualitative variables

Plotted box plot for qualitative predictor variables and quantitative variable charges(response).To find the relation between the response and predictors.

```
#  1x3 layout
par(mfrow = c(1, 3))

# Boxplot for charges by sex
boxplot(charges ~ sex, data = trainData, xlab = "Sex", ylab = "Charges",
        main = "Charges by Sex", col = c("skyblue", "lightpink"))

# Boxplot for charges by region
boxplot(charges ~ region, data = trainData, xlab = "Region", ylab = "Charges",
        main = "Charges by Region", las = 2, col = "lightgreen")

# Boxplot for charges by smoker
boxplot(charges ~ smoker, data = trainData, xlab = "Smoker", ylab = "Charges",
        main = "Charges by Smoker Status", col = c("lightblue", "orange"))
```



chargers vs sex: it seems that the median charge for females is lower than that for males. This is shown by the horizontal lines inside the boxes.Moreover, the charges for females appear to be more tightly packed and less spread out compared to males.

charges vs region: Looking at the boxes, it seems that the median charges are highest for the southwest region and lowest for the northwest region.

charges vs smoker: The horizontal line within the box represents the median charge for non-smokers. This tells that that smokers tend to have higher medical charges overall.

Summary for trainData:

Describing the summary of the trainData by using the summary(). Finding the trainData the minimum,maximum,quartile and for categorical how many counts are their for each varaible.

```
summary(trainData)
```

```
##       age            sex           bmi           children     smoker
##  Min.   :18.00   female:530   Min.   :15.96   Min.   :0.000   no :850
##  1st Qu.:27.00   male  :542   1st Qu.:26.40   1st Qu.:0.000   yes:222
##  Median :39.00                Median :30.50   Median :1.000
##  Mean   :39.19                Mean   :30.76   Mean   :1.081
##  3rd Qu.:51.00                3rd Qu.:34.80   3rd Qu.:2.000
##  Max.   :64.00                Max.   :53.13   Max.   :5.000
##        region       charges
##  northeast:254   Min.   : 1122
##  northwest:257   1st Qu.: 4745
##  southeast:309   Median : 9382
##  southwest:252   Mean   :13318
##                  3rd Qu.:16604
##                  Max.   :63770
```

age: The ages in the dataset range from 18 to 64 years old. This indicates that the dataset covers a broad range of ages. The median age 39 years falls close to the mean age (39.19 years) shows that the age distribution is approximately symmetric. This means that there are roughly an equal number of individuals younger and older than the median age.

sex:This column has two categories , male (542) and female (532).This shows in data having more male than female.

bmi: The bmi values from 15.96 to 53.13 a BMI below 18.5 is considered underweight, while a BMI of 30 or higher falls into the obese category.With a median BMI of 30.5 shows that the majority of individuals in the dataset are either overweight or obese (obese,overweight,underweight) are considered as by NIH(National Institution of health).

children:The maximum no of children is 5, median no of children is 1.But, the first quartile is 0,this shows that many individuals in the dataset have no children.

smoker:The smoker variable has two categories, 850 members are non-smokers and 222 are smokers. This shows that majority in the train data are non-smokers.

region: There are 4 categories in the region variable.The number of individual from each region varies with southeast having the height representation of 309 and lowest is northeast with 254 members.

charges: Charges differ widely with a minimum of $1122 and maximum of $63770 .The mean is $9382 which shows that there are few expensive charges.

Correlation matrix

it selects only the numerical columns from the trainData dataset using the sapply() function combined with is.numeric(). This creates a new dataset called numerical_trainData containing only the columns with numerical data. Then, it calculates the correlation matrix (cor_matrix) using the cor() function. This function computes the correlation coefficients between all pairs of numerical variables in numerical_trainData

```r
library(ggplot2)

# Selecting numerical columns from trainData
numerical_trainData <- trainData[, sapply(trainData, is.numeric)]

# Computing the correlation matrix
cor_matrix <- cor(numerical_trainData, use = "pairwise.complete.obs")
print(cor_matrix)
```

```
##                 age        bmi   children    charges
## age      1.00000000 0.11862886 0.05195901 0.28859456
## bmi      0.11862886 1.00000000 0.02296727 0.19811147
## children 0.05195901 0.02296727 1.00000000 0.05593239
## charges  0.28859456 0.19811147 0.05593239 1.00000000
```

Age ad bmi: It has weak positive correlation of 0.1186 as the age increases the bmi tends to increase slightly as well , but the relationship is not very strong.

age and children: The connection between age and number of children is not very strong. slightly tends to be found in old aged persons having more children since the association has a low power of strength is 0.0519.

age and charges:this has moderate positive correlation of 0.2886 shows that as age increases the medical charges tend to increase moderately.

BMI and children have a weak positive correlation of 0.0220 implying that there is a slight tendency for people with higher BMI to have more children but the relationship is not strong.

BMI and charges have a moderate positive correlation of 0.1981 indicating that higher BMI is associated with moderately higher medical charges.

Children and charges have a weak positive correlation of 0.0559, suggesting that having more children is associated with slightly higher medical charges, but the relationship is not very strong.

Overall, EDA helps to understand the potential factors influencing medical charges, such as age, BMI, number of children, sex, region, and smoking status.

Create dummy varibles for trainData

In the above created scatter plot for quantitative variables. In the dataset there are qualitative variables . When dealing with categorical variables in statistical modeling, converted the qualitative variables into dummy variables (also known as dummy or indicator variables). Dummy variables are binary variables representing the presence or absence of a particular category.

```r
# Converting qualitative predictors to dummy variables

trainData$sex_female <- ifelse(trainData$sex == "female", 1, 0)
trainData$smoker_yes <- ifelse(trainData$smoker == "yes", 1, 0)

# Creating dummy variables for 'region'
regions <- unique(trainData$region)
reference_region <- regions[1]  # Choosing the first region as the reference level
for (i in 2:length(regions)) {
  col_name <- paste("region_", regions[i], sep = "")
  trainData[col_name] <- ifelse(trainData$region == regions[i], 1, 0)
```

```
}

trainData <- subset(trainData, select = -c(sex, smoker, region))
```

The dummy variable names are sex_female,smoker_yes,region_southeast,region_northwest,region_northeast.

structure of the trainData

Structure of the train data shows the dummy variables by using the str() function. In the train data has 1072 obsevations and 9 variables.The dummy variables are converted into numerical datatype.

```
str(trainData)
```

```
## 'data.frame':    1072 obs. of  9 variables:
##  $ age             : int  19 18 28 33 31 46 37 60 25 62 ...
##  $ bmi             : num  27.9 33.8 33 22.7 25.7 ...
##  $ children        : int  0 1 3 0 0 1 3 0 0 0 ...
##  $ charges         : num  16885 1726 4449 21984 3757 ...
##  $ sex_female      : num  1 0 0 0 1 1 1 1 0 1 ...
##  $ smoker_yes      : num  1 0 0 0 0 0 0 0 0 1 ...
##  $ region_southeast: num  0 1 1 0 1 1 0 0 0 1 ...
##  $ region_northwest: num  0 0 0 1 0 0 1 1 0 0 ...
##  $ region_northeast: num  0 0 0 0 0 0 0 0 1 0 ...
```

#dummy varibles for testdata

```
# Converting qualitative predictors to dummy variables for testData

testData$sex_female <- ifelse(testData$sex == "female", 1, 0)
testData$smoker_yes <- ifelse(testData$smoker == "yes", 1, 0)

# Creating dummy variables for 'region'
regions <- unique(testData$region)
reference_region <- regions[1]
for (i in 2:length(regions)) {
  col_name <- paste("region_", regions[i], sep = "")
  testData[col_name] <- ifelse(testData$region == regions[i], 1, 0)
}
# Removing the original character variables
testData <- subset(testData, select = -c(sex, smoker, region))
```

Forward stepwise selection

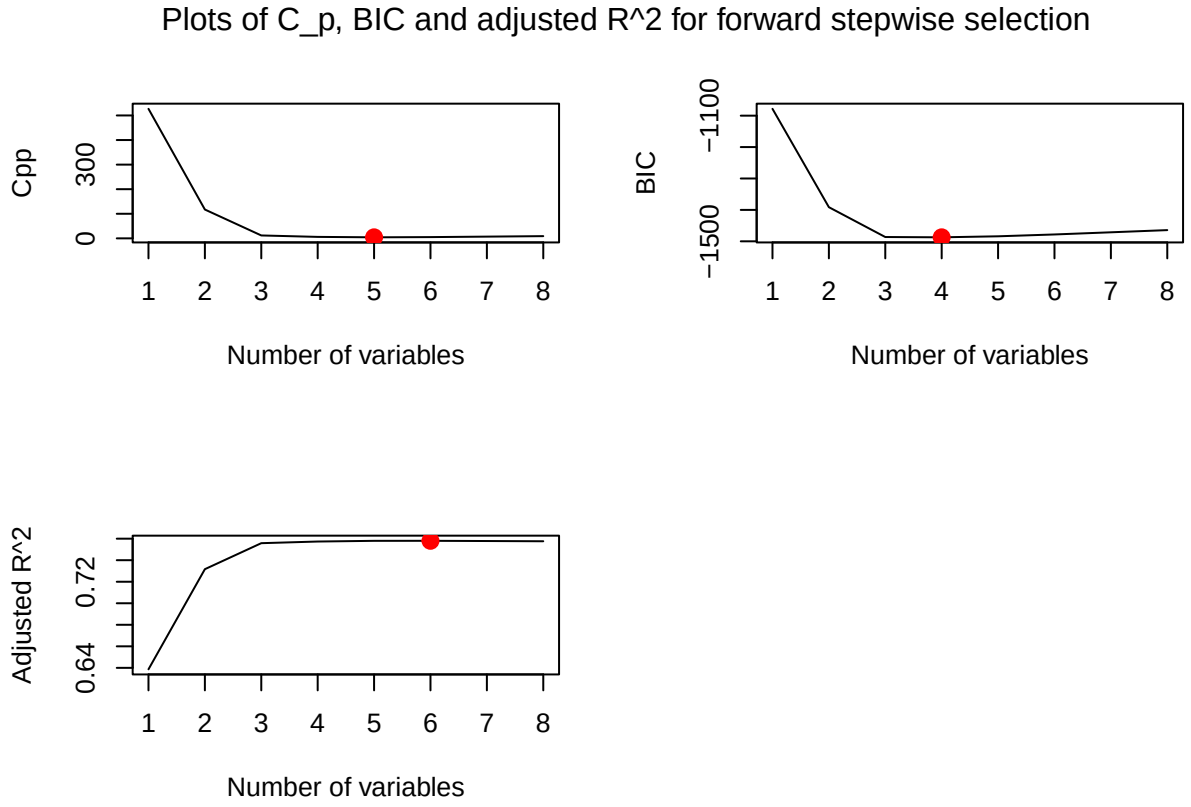
Selection:

In the feature selection used the Forward Step wise selection to select the best subset of predictors for predicting charges (response variable) . In the formula variable charges is the response variable (age ,bmi ,children ,sex_female ,smoker_yes ,region_southeast , region_northwest , region_northeast) these are the predictors. The regsubsets() function performs forward stepwise selection using the specified model formula. It selects the best subset of predictors based on various criteria Cp, BIC, adjusted R^2 while gradually adding predictors to the model.

```r
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.3.3
```

```r
#  formula
model_formula <- charges ~ age + bmi + children + sex_female + smoker_yes +
  region_southeast + region_northwest + region_northeast

# Performing forward step wise selection
regfit_fwd <- regsubsets(model_formula, data = trainData, nvmax = 8, method = "forward")

# Summarizing the results
reg_fwd <- summary(regfit_fwd)

# Plot Cp, BIC, and adjusted R^2 criteria
par(mfrow = c(2, 2))

plot(reg_fwd$cp, xlab = "Number of variables", ylab = "Cpp", type = "l")
points(which.min(reg_fwd$cp),
       reg_fwd$cp[which.min(reg_fwd$cp)], col = "red", cex = 2, pch = 20)

plot(reg_fwd$bic, xlab = "Number of variables", ylab = "BIC", type = "l")
points(which.min(reg_fwd$bic),
       reg_fwd$bic[which.min(reg_fwd$bic)], col = "red", cex = 2, pch = 20)

plot(reg_fwd$adjr2, xlab = "Number of variables", ylab = "Adjusted R^2", type = "l")
points(which.max(reg_fwd$adjr2),
       reg_fwd$adjr2[which.max(reg_fwd$adjr2)], col = "red", cex = 2, pch = 20)

mtext("Plots of C_p, BIC and adjusted R^2 for forward stepwise selection",
      side = 3, line = -2, outer = TRUE)
```



In the plots C_p model selects the 5 variables ,BIC selects the 4 variables and Adjusted R-square selected 6 variables.

Coefficients

```
best_m_cp_fwd <- coef(regfit_fwd, which.min(reg_fwd$cp))
best_m_bic_fwd <- coef(regfit_fwd, which.min(reg_fwd$bic))
best_m_adjr2_fwd <- coef(regfit_fwd, which.max(reg_fwd$adjr2))

print("Coefficients of the best model based on Cp (Forward Stepwise Selection):")
```

```
## [1] "Coefficients of the best model based on Cp (Forward Stepwise Selection):"
```

```
print(best_m_cp_fwd)
```

```
##      (Intercept)              age              bmi         children
##      -11950.5971         247.5264         320.9850         442.4207
##       smoker_yes region_northeast
##       24225.1995         843.0722
```

```
print("Coefficients of the best model based on BIC (Forward Stepwise Selection):")
```

```
## [1] "Coefficients of the best model based on BIC (Forward Stepwise Selection):"
```

```r
print(best_m_bic_fwd)
```

```
## (Intercept)         age         bmi    children  smoker_yes
## -11528.3283    247.9113    313.7158    431.6397  24213.7922
```

```r
print("Coefficients of the best model based on adjusted R^2 (Forward Stepwise Selection):")
```

```
## [1] "Coefficients of the best model based on adjusted R^2 (Forward Stepwise Selection):"
```

```r
print(best_m_adjr2_fwd)
```

```
##      (Intercept)              age              bmi         children
##      -12296.8755         247.1721         327.6921         436.6310
##       smoker_yes region_northwest region_northeast
##       24235.4744         494.4401        1009.6674
```

These are the coefficients of the models. I have chosen the Cp model from the forward stepwise selection process.The Cp model identifies the following key factors influencing charges:Age,bmi,children,smoker_yes,region_northeast.The positive coefficients for age, BMI, children, and smoker_yes indicate that higher values of these variables are associated with higher insurance charges. The positive coefficient for the region_northeast variable suggests that individuals from the Northeast region tend to have higher insurance charges compared to other regions.

Age: For every year a person gets older, insurance charges typically increase by $247.53, on average. BMI: With every one-unit increase in BMI, charges are expected to rise by $320.99, on average. Children: Each additional child is associated with an average increase of $442.42 in insurance charges. Smoker (smoker_yes): Being a smoker is linked to a significant increase in charges, with smokers paying an average of $24,225.20 more than non-smokers. Region (region_northeast): Individuals from the Northeast region typically have insurance charges around $843.07 higher than those from other regions, on average.

In conclusion, the Cp model selected through forward stepwise selection is applicable to prediction problem of insurance charges. It identifies key factors influencing insurance charges while balancing model complexity and prediction accuracy.

Based on the forward stepwise and Exploratory data analysis discussing the applicability of statistical learning methods to the prediction problem of insurance charges: Age: The positive coefficient for age suggests that as individuals get older, their medical charges tend to increase. Age could be significant factor in predicting insurance charges. Linear regression assumes a linear relationship, and this factor meets that assumption.

BMI: The positive coefficient for BMI indicates that higher BMI values are associated with higher medical charges suggests that BMI could be a relevant predictor. Linear regression can handle this linear relationship assumption.

Number of Children: The positive coefficient for the number of children implies that individuals with more children tend to have higher medical charges.However, the relationship is not very strong. Linear regression assumes linearity, but the weak correlation suggests that the relationship might not be entirely linear.

Smoking Status: The substantial positive coefficient for smoker_yes indicates that smokers tend to have significantly higher medical charges compared to non-smokers. This factor seems highly influential and meets the linearity assumption of linear regression.

Region: The positive coefficient for the northeast region suggests that individuals from this region tend to have higher medical charges compared to other regions. However, the effect is relatively smaller compared to smoking status. Linear regression can handle this linear relationship assumption.

Overall, it would seem that these insurance charges can be predicted by linear regression with these factors since they satisfy the assumptions of independence and linearity.However, it should also be noted that this is a simple model which may not capture complex non-linear relationships.In such cases more advanced methods like random forest could be considered.

Predictive analysis:

Model formulation:Defined the model formula including the response variable (insurance charges) and predictor variables (age, BMI, number of children, smoker_yes, and region_northeast).

Forward stepwise selection: The regsubsets functions in R was applied for forward stepwise selection. All potential combinations of predictors are considered in this method with each subset according to Cp criterion.

Model fitting:A linear model and random forest were fitted using the subset of predictors chosen from forward stepwise selection . This step involves finding coefficients that would make these predictors most appropriate best fit the relationship between the predictors and the response variable.

Model evaluation: Analyzed the effectiveness of the adjusted models using re-sampling techniques like 10-fold cross-validation.This required dividing the data into 10 folds, building the model on the basis of 9 folds and examining it on the remaning fold.

Repeated this procedure 10 times and each time rotate the validation fold, then take an average of the outcomes to evaluuate the performance of the model.

Performance Metrics:In order to gauge accuracy and fitness on test data, assess goodness using the RMSE, MAE and r-square. Comparision: comparision between the linear regression and random forest.

Visualization: visualized the actual versus predicted charges to visually inspect the model's performance and assess the alignment between predicted and actual values.

Resampling method: Using the train data and test data. Apply cross validation on the training set to evaluate the model's performance. This involves dividing the training set into k-folds , training the model on k-1 folds, and validating it on the remaining fold. Repeat this process k times, rotating the validation fold each time. Finally, average the performance metrics across all folds.

Linear regression :

Builds a linear regression model to predict charges based on features like age, BMI, children, smoker_yes, and region_northeast using the train() function from the caret package. It then evaluates the model's performance using 10-fold cross-validation and calculates metrics like RMSE, MAE, and R-squared on a test dataset. Finally, it plots the actual vs. predicted charges.

```
# loading required packages
library(caret)

# Formula with selected features
formula <- charges ~ age + bmi + children + smoker_yes+region_northeast
model <- train(formula, data = trainData, method = "lm")

# Performing 10-fold cross-validation on the training set
cv_results <- train(formula, data = trainData, method = "lm",
                    trControl = trainControl(method = "cv", number = 10))

# Printing the cross-validation results
print(cv_results)
```

```
## Linear Regression
##
## 1072 samples
##    5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 964, 965, 966, 964, 965, 964, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   6021.373  0.7569469  4140.696
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
    # Evaluating the model's performance on the test set
    predictions <- predict(model, newdata =testData )

    # Calculating performance measures on test set
    rmse <- sqrt(mean((predictions - testData$charges)^2))
    mae <- mean(abs(predictions - testData$charges))
    rss <- sum((predictions - testData$charges)^2)
    tss <- sum((testData$charges - mean(testData$charges))^2)
    r_squared <- 1 - (rss/tss)

    cat("Test RMSE:", rmse, "\n")
```

```
## Test RMSE: 6302.607
```

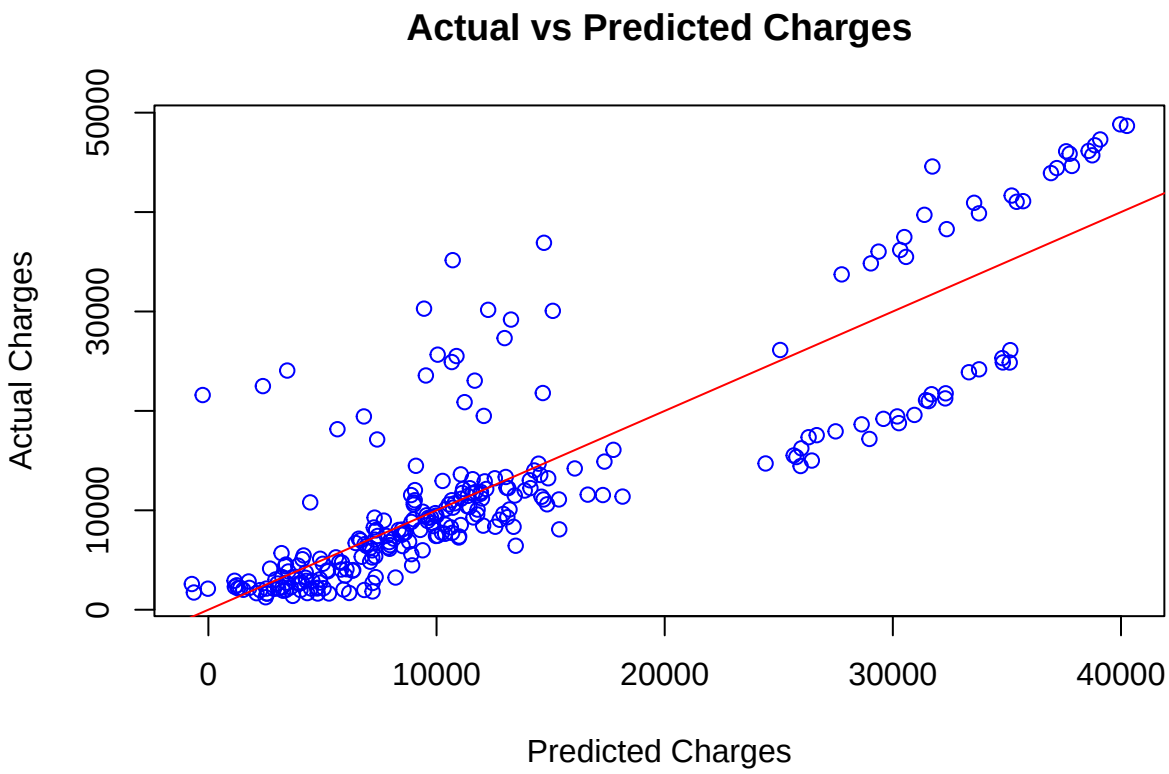
```
    cat("Test MAE:", mae, "\n")
```

```
## Test MAE: 4163.541
```

```
    cat("Test R-squared:", r_squared, "\n")
```

```
## Test R-squared: 0.71023
```

```
  # Plotting actual vs predicted values
  plot(predictions, testData$charges, main = "Actual vs Predicted Charges", xlab = "Predicted Charges",
  abline(0, 1, col = "red")
```



Displaying the predicted and actual values:

Can comparing the result by seeing the actual and predicted values.

```
# Create a dataframe containing actual and predicted values
comparison_df <- data.frame(Actual = testData$charges, Predicted = predictions)

# Print the dataframe
head(comparison_df)
```

```
##       Actual Predicted
## 5   3866.855  5240.296
## 9   6406.411  8510.777
## 14 11090.718 14692.506
## 19 10602.385 14846.579
## 25  6203.902  7088.327
## 28 12268.632 13068.481
```

###Random forest Builds a random forest model to predict charges based on features like age, BMI, children, smoker_yes, and region_northeast using the train() function from the caret package. It then evaluates the model's performance using 10-fold cross-validation and calculates metrics like RMSE, MAE, and R-squared on a test dataset. Finally, it plots the actual vs. predicted charges.

```r
# Loading required packages
library(caret)
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.3.2
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
# Formula with selected features
formula <- charges ~ age + bmi + children + smoker_yes + region_northeast
model_rf <- train(formula, data = trainData, method = "rf")

# Performing 10-fold cross-validation on the training set
cv_results_rf <- train(formula, data = trainData, method = "rf",
                       trControl = trainControl(method = "cv", number = 10))

# Printing the cross-validation results
print(cv_results_rf)
```

```
## Random Forest
##
## 1072 samples
##    5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 967, 965, 964, 964, 965, 966, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared   MAE
##   2     4532.799  0.8612393  2602.505
##   3     4570.566  0.8575948  2492.272
##   5     4705.438  0.8498313  2567.692
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 2.
```

```r
# Evaluating the model's performance on the test set
predictions_rf <- predict(model_rf, newdata = testData)

# Calculating performance measures on the test set
rmse <- sqrt(mean((predictions_rf - testData$charges)^2))
```

```r
mae <- mean(abs(predictions_rf - testData$charges))
rss <- sum((predictions_rf - testData$charges)^2)
tss <- sum((testData$charges - mean(testData$charges))^2)
r_squared <- 1 - (rss/tss)

cat("Test RMSE:", rmse, "\n")
```

```
## Test RMSE: 4911.243
```

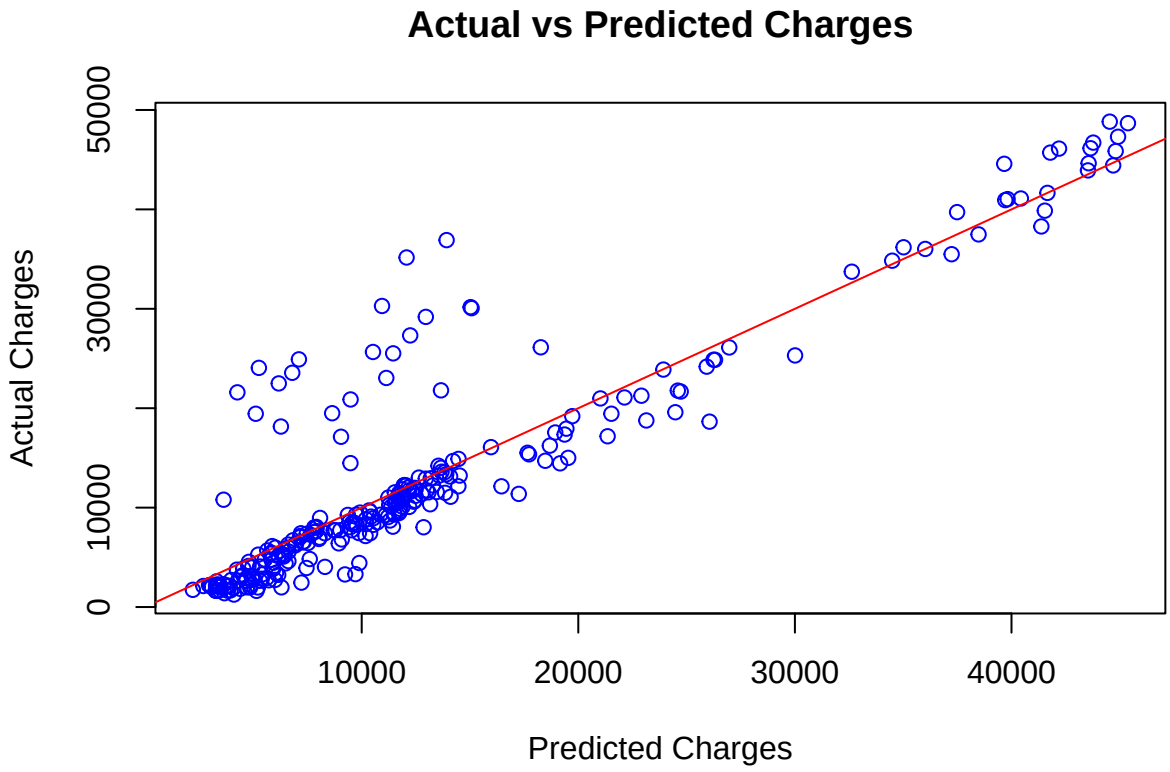
```r
cat("Test MAE:", mae, "\n")
```

```
## Test MAE: 2770.328
```

```r
cat("Test R-squared:", r_squared, "\n")
```

```
## Test R-squared: 0.8240473
```

```r
# Plotting actual vs predicted values
plot(predictions_rf, testData$charges, main = "Actual vs Predicted Charges", xlab = "Predicted Charges"
abline(0, 1, col = "red")
```



Displaying the predicted and actual values: Can comparing the result by seeing the actual and predicted values.

```r
# Create a dataframe containing actual and predicted values
comparison_df1 <- data.frame(Actual = testData$charges, Predicted = predictions_rf)

# Print the dataframe
head(comparison_df1)
```

```
##       Actual Predicted
## 5   3866.855  4535.271
## 9   6406.411  8939.920
## 14 11090.718 11666.379
## 19 10602.385 11721.259
## 25  6203.902  6934.282
## 28 12268.632 11971.277
```

Results:

Base on the cross validation metrics , the linear regression model shows a decent R-squared value(cross validation) which means that approximately 75% of the variability in the dependent variable(charges) has been explained by independent variables.Although RMSE and MAE indicate that there are still errors with predictions as shown differences between predicted charges and actual charges, test performance metrics are 71% , indicating that the model generalizes reasonably well to new unseen data but their is variability.Liner regression plot, that there is considerable deviation from this line, indicating discrepancies between the predicted and actual charges.In random forest also there are some deviation from the actual and predicted values . The random forest on the test data RMSE of 4911.243, an MAE of 2770.328, and an R-squared of 0.8240473.

In comparison to both the cross-validated outcomes and metrics used for testing, the random forest model is better than the linear regression model. Through achieving a lower RMSE and MAE, which shows improved forecast accuracy, it also has a higher R-squared value that shows it is fitting much closer to the data set than the latter. Two is the number of predictors per split that results in the best performance based on all possible divisions; thus mtry should be set at 2 when using Random Forest algorithm. On the whole, Random Forest model proves to be superior.

In conclusion, considering the results, the random forest model seems to perform better when it comes to forecasting insurance premiums than linear regression because it has lower error metrics and larger R-squared values.

Conclusion:

The analysis directly addresses the research question regarding the key factors influencing insurance charges and aims to develop predictive models based on factors.The main key factors are age , bmi , children , smoker_yes , region_northeast . From these factors by using the linear and random forest predicted the insurance charges.

Model performance metrics like RMSE,R-squared,MAE are evaluated through cross-validation so as to offer insights on how well the model could generalize a previously unseen data.Implied in the fact that random forest model outperforms liner regression indicates potential for generalizability across different datasets.For one, insurance industry practitioners and policy makers indicating the potential applicability and generalizability of the developed predictive models in real-world settings.

The analysis is constrained by the inclusion of a specific set of predictors,potentially overlooking other influential factors affecting insurance charges. This limitation could restrict the scope of the analysis and

impede a comprehensive understanding of the determinants of insurance charges. Exploring additional predictors beyond the current set could enhance the predictive accuracy and robustness of the models. Finding outliers in the data and handling the outliers may give the model stability and predictive accuracy. feature transformations could address the limitations of linear regression's assumptions. Techniques such as polynomial regression or spline regression could capture nonlinear relationships more effectively.

The goal of this project will helpful to make of insurance customer engagement insures may find it useful to first understand what factors affect insurance rates.They may then able to explain how the premiums are calculated and suggest personalized recommendations for cost saving.Insurers can also refine their pricing strategies using this information by setting premiums that reflects a accurately the risk of each policy holder.