Group 4
STAT 515 Final Project

MEDICAL INSURANCE CHARGES PREDICTION

Introduction

Health insurance is a policy that budgets for and pays for one's medical expenses. It protects you financially in the event of a major accident or sickness. A broken leg, for example, will cost up to $7,500, having insurance will cover it all. Everyone needs to have a secured health insurance policy. Health insurance will help you avoid big and unanticipated costs. The Insurance firms usually collect a higher premium than the amount charged to the insured individual to make a profit. As a result, insurance firms devote a significant amount of time, effort, and resources to developing models that reliably forecast healthcare costs.

Dataset

The dataset selected contains details regarding healthcare finance. It allows one to forecast a person's medical insurance costs. This dataset contains data on the insurer, their dependents, and their medical costs over a year. The prediction is based on the dataset's other parameters, such as BMI, smoker, children, age. To predict the costs, we use the linear regression model, Decision Tree, and Random Forest Model, to see which model is a better fit.

The data contains the following data item:

1.Age

2.Sex

3.BMI: Body Mass Index of the insurer

4.Children: Determines whether the insurer has any children who are to be covered

5.Smoker: Determines whether he/she is a smoker or non-smoker

6.Region: Residence of the insurer; Northeast, Northwest, Southeast, Southwest

7.Charges: Yearly medical charge

Preliminary Analysis

It is very important to understand each variable’s significance by plotting individually before creating a model. The below plots show the primary visualizations.

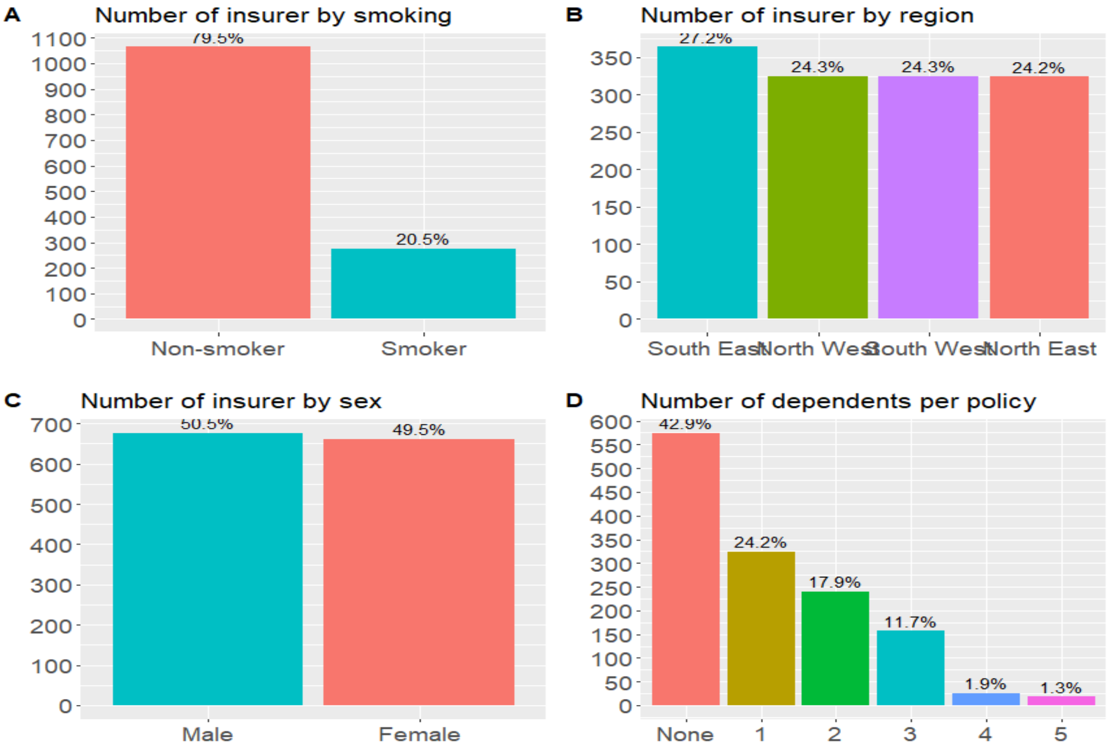

Figure 1 - Visualizations for Nominal Data Items

From the above plot, it is observed that smoking status has considerable nonsmokers (up to 80%) that outnumber smokers (by 20%). The region of residence of policyholders is similarly spread across the countries, with the South East being the most populated (27 %), and the remaining regions each having about 24 % of policyholders. In terms of gender, there are marginally more men (51%) than women (49%) in the study. Whereas most policyholders (43%) do not have dependents protected by their insurance. The majority of those who have dependents protected by their insurance has only one (24 %). The number of dependents that can be insured is limited to five (1 %).

It is derived from the above plot that the youngest policyholder is 18 years old, while the oldest is 64 years old. Apart from the youngest and oldest policyholders, all ages in the range are reasonably evenly represented. The most populous age group is 18-23 years old, while the least populous age group is 60-64 years old. There are no anomalies in the data. BMI is usually distributed, with the least common values being the smallest and highest, and the median and mean being nearly equal. On a broader scale, there are a few outliers. The lowest reported BMI score is 16 and the highest is 53.1. Charges are overwhelmingly skewed to the right, with several outliers on the larger hand.

This ensures that most of the charges are modest, with a few exceptions. The smallest charge is $1,122 and the highest charge is $63,770.

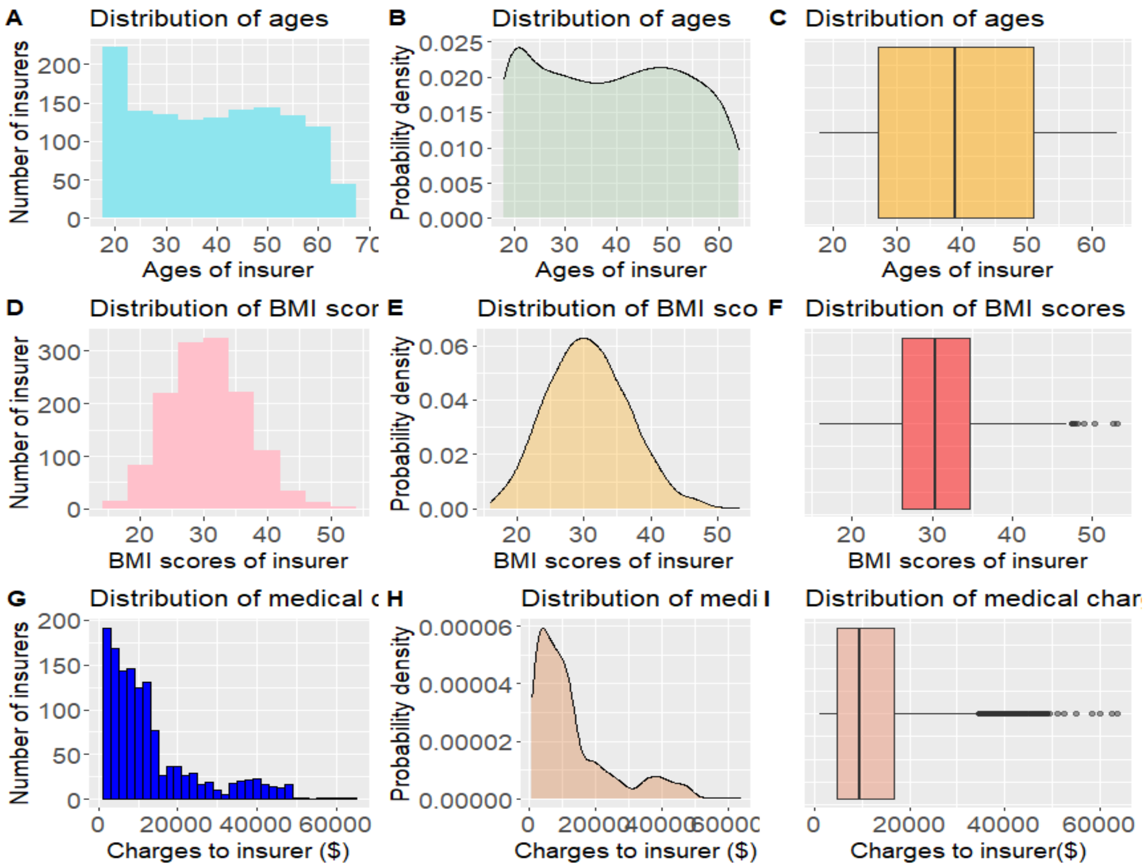

Figure 2 - Visualizations for continuous Data items

Linear Regression

In this dataset, medical insurance charge is the dependent variable and all the other variables age, BMI, smoker, children, gender, region are independent variables. The prediction of charges is dependent on these variables so to formulate the relationship between these independent variables and charges, first, we try to fit a basic linear model to see the fit and performance. This model is with all the independent variables.

From the summary of this linear model, for the Pr(>|t|) column, we see that the regression coefficients for age, BMI, smoker, children are significant, while gender and region are least significant. The multiple R-squared is 0.7509 which tells that the model accounts for 75.09% of the variance in the insurance charges. We then plotted below diagnostic graphs.

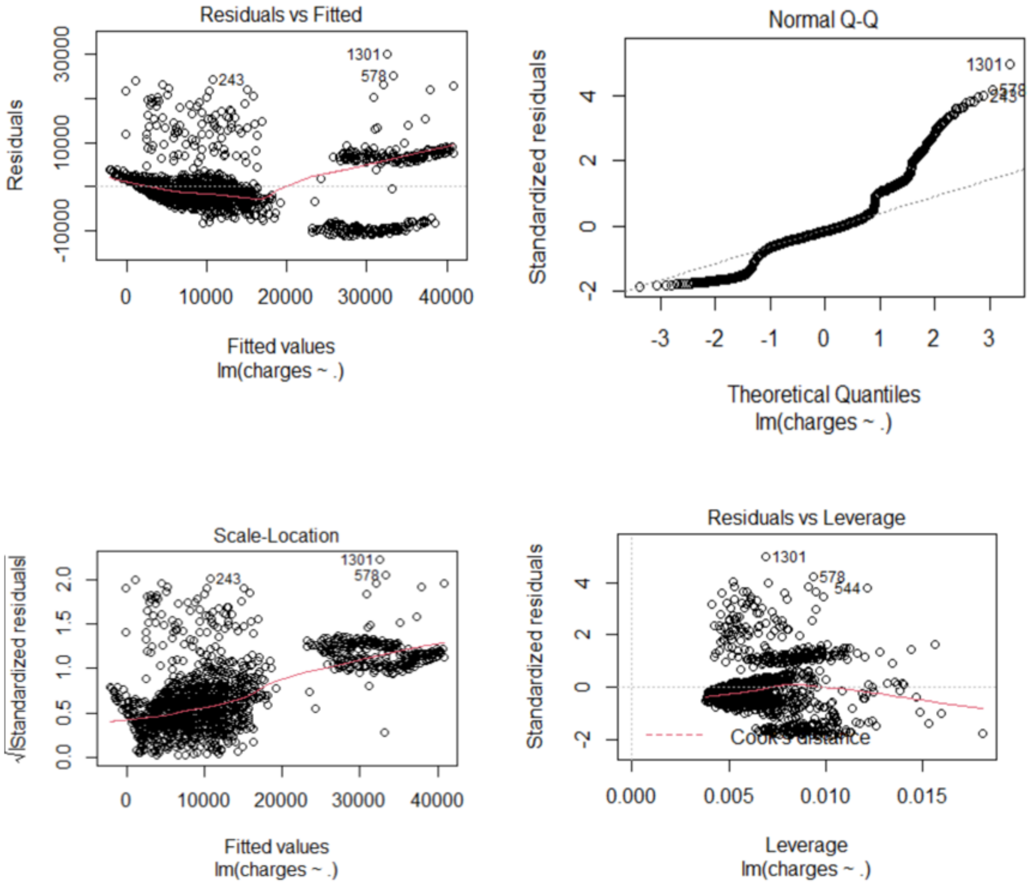

Figure 3 – Diagnostic Plots

In the first plot, for a good linear relationship, the red horizontal line should be approximately flat, and the residuals should be equally spread along the line. But this is not the case here, there are three groups in this plot, so for each, the variance is different across the residuals. It is the case of a curved relationship, which tells us that we may have to add a quadratic term in our regression. From the Q-Q Normalized plot, we must see if the residuals are normally distributed for a good linear relationship the residuals should be along the straight line without much deviation, but in our graph, we see the deviation of residuals at the head and tail of the line. So, residuals do not satisfy normality. The scale location plot should show the constant variance of the residuals. But this is not the case here the residuals are not equally spread along with the predictors' range. The last plot, Residuals vs leverage, is used to how much each data point influences the regression.

<u>Tackling Non – Linear Relationship</u>

To handle the above nonlinear relationship, we introduced a quadratic term for age, as it is one of the continuous variables with significance. In general increase in age increases the medical insurance charges as older people tend to have more risk compared to young people. After adding this quadratic term, we see the curved line is changed to a straight line as shown in the below plot.

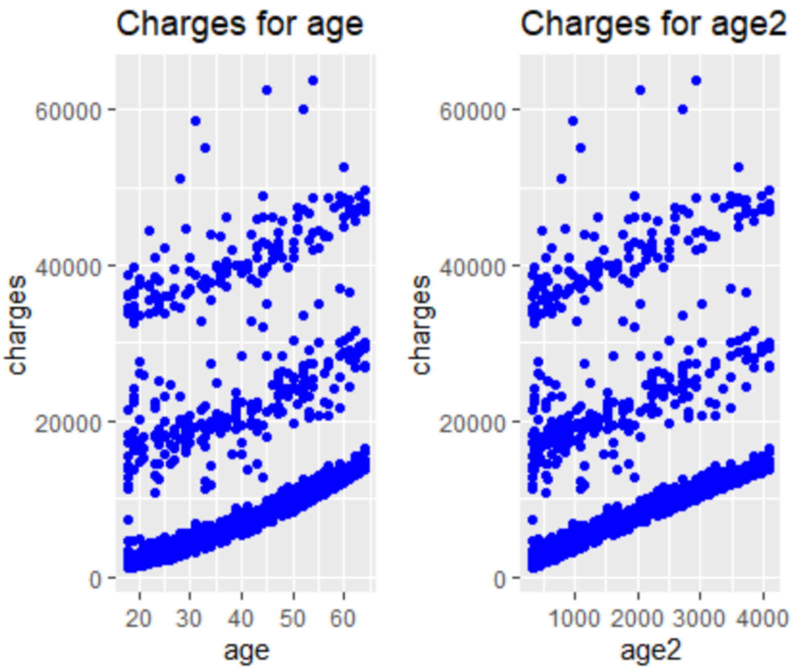

Figure - 4 Scatter plots for age, age2 with charges

So, this variable will be added in our next model, along with this we added another variable BMI 30. BMI (Body Mass Index) determines the obesity of a person. BMI over 30 indicates obesity so insurance charges are expected to increase for BMI over 30 as it can be considered as the threshold value for health risk. In this model, we removed insignificant variables gender, and region.

For the second regression model, we split the data into train and test, 70% for the train for fitting the model and 30% of data for testing and predicting, then we use age, BMI, smoker, children, age 2, BMI 30 to predict the insurance charges. From the summary of this model, we saw that this model has a multiple R-squared of 0.757 which tells that the model accounts for 75.7% of the variance in the insurance charges. As there is more than one independent variable, the regression coefficients indicate the increase in the dependent variable for a unit change in a predictor variable, holding all other independent variables constant. For example, for BMI along with BMI 30, the estimate is around 2317 so it suggests that an increase of 1 in BMI is associated with a 2317 increase in the charges, keeping age, children, smoking constant 75.7 % of the Multiple R – Squared indicates a good fit, so we use this model to predict test data.

After passing the test data to the model and predicting, we have the below plot for actual charges in test data vs predicted charges.

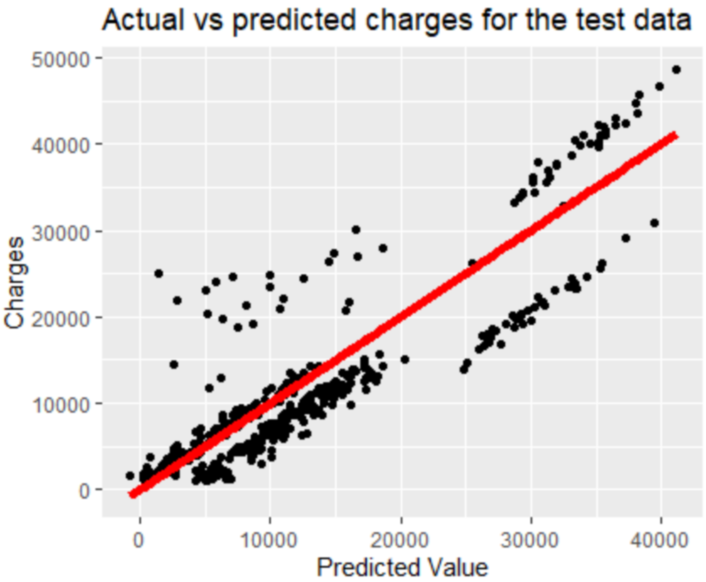

Figure 5 – Actual vs predicted charges in the test data

From this plot, we see a linear line. Mostly, with the increase in actual charges, the Predicted charges are increased. Although this model is a good fit, we can see the test MSE (39300) to be high, and thus, we are implementing other models to reduce the MSE(Mean Squared Error).

Decision Tree

Basic regression trees divide a data set into smaller subgroups and then fit a simple constant to each of the subgroup's observations. Based on the different predictors, successive binary partitions (also known as recursive partitioning) are used to partition the data. The average response values for all observations in that subgroup are used to calculate the constant to forecast.

The below figure represents the regression tree for our data set. The first split is made by age parameter at a value of 41.5 years. Branching towards the left are the people having age < 41.5 and towards the right are people having age > 41.5 years. On the left side, they are further split again with BMI parameter at value 30.6. The predicted insurance cost will be 7934 dollars (present at the terminal node) which is the least when a person < 41.5 years of age and has a BMI < 30.6. This reason being people < of 41.5 years and BMI < 30.6 are less risky of health failures, hence respective insurance charges are less.

For BMI > 30.6 and age < 41.5, the insurance charges would be 12110 dollars as BMI > 30 is obese and have a greater risk of health failure. When we branch towards the right from the split at age<41, there is another partition with age < 59, where the left branches with people < 59 years of age and right with people > 59 years of age. People with age < 59 years are further partitioned with BMI < 34. Those people with age < 59 years (and > 41 years) and BMI < 34 have an insurance premium of 14250 dollars.

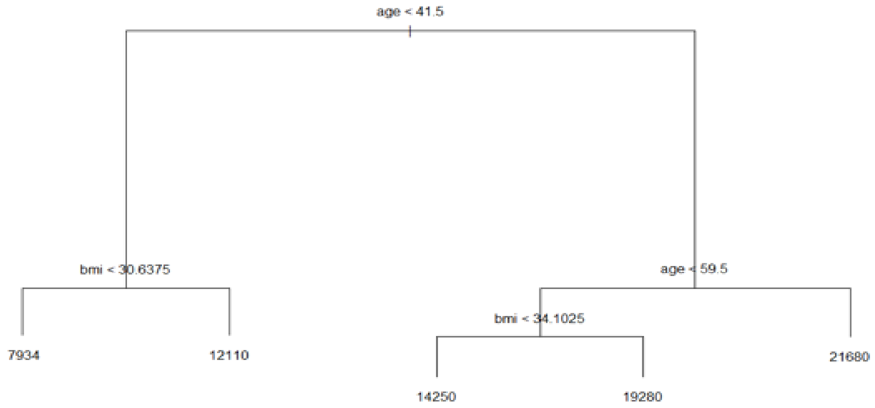

Figure 6 - Regression tree Model

People with BMI > 34 and are in the age between 41 and 59 years have a premium of 19280 dollars. Finally, people with age > 59 years have the highest premium 21680 because they are the people with the highest risk of health failure. Now we will see whether pruning the tree will improve the performance or not. This is done using cv.tree() function. The figure below depicts the deviance with the size of the tree.

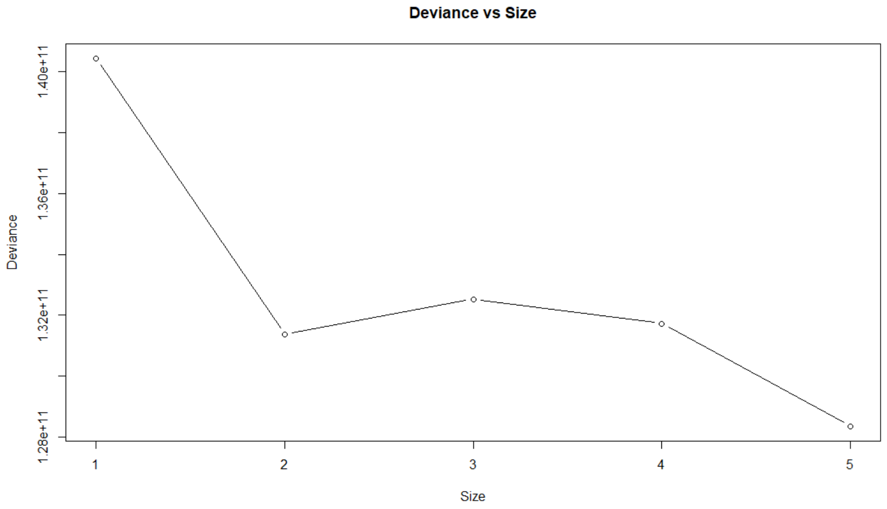

Figure 7 - Cross-Validation

When we look at figure 7 above, at size = 5, the deviance value is the lowest so we will prune the tree to size 5. When we prune with size =5 and plot, the plot is the same as the plot in figure 6. Hence initially plotted the optimum model. We will now predict our model on test and compare it with our predicted values.

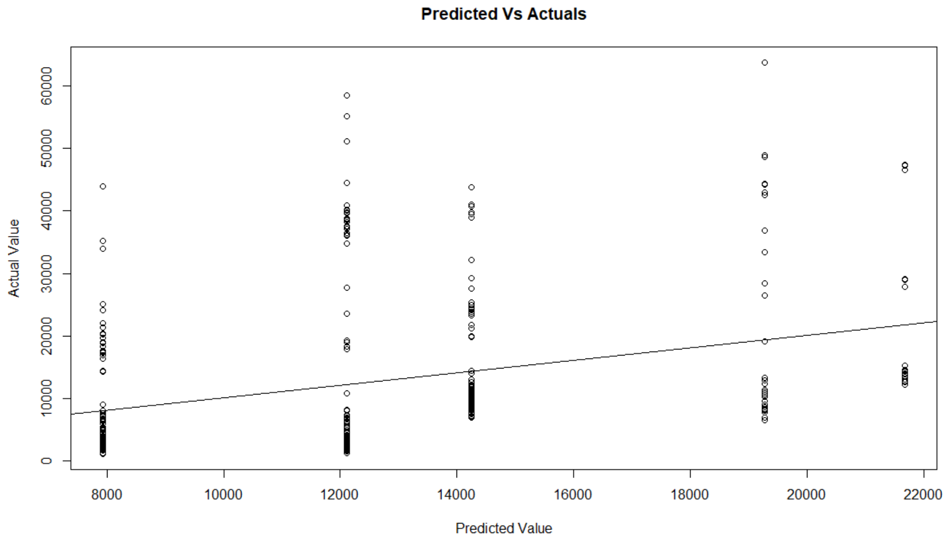

Figure 8 - Predicted vs Actual values

From the above plot, we can infer that both predict(yhat) and actual(test) values are almost similar for the insurance cost instances 8000. However, it is not the case for the rest of the predictors, and we can see most predicted charges below the actual charge's values. The value of MSE is 152976 which is higher than the linear regression model. Therefore, we are implementing another model to have better accuracy and lesser mean squared value.

Random Forest

We choose random forest because it is more robust, and it can capture those non-linear features in our data. By default, randomForest() uses p/3 variables when building a random forest of regression (James et al, 2017). Here we use mtry=3. The test set MSE is 30381 which improved from both linear regression and decision tree. We ordered the importance of the variables (Figure 9). "smoker" is the most important variable across all the trees considered in the random forest, followed by "age", "BMI" and "children".

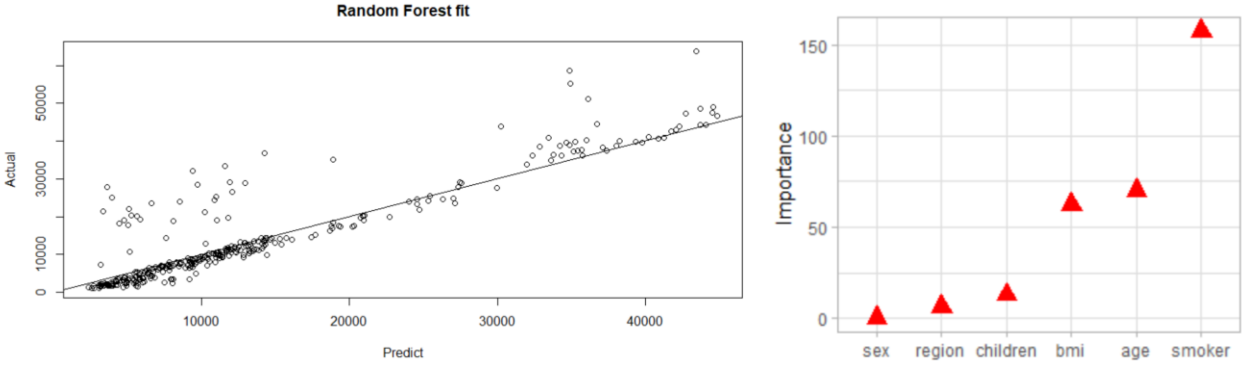

Figure 9 - Random Forest Plots

According to James et al(2017), "Partial dependencies plot illustrates the marginal effect of the selected variable on the response after integrating out the other variables". In our case, the insurance cost increases when age increases. When the BMI crosses above 30 the insurance charges shoot up close to 14500 dollars. The number of children is positively correlated to the insurance cost, i.e., as the number of dependents increases the insurance charges increase which is true since more people are to be covered in the given insurance.

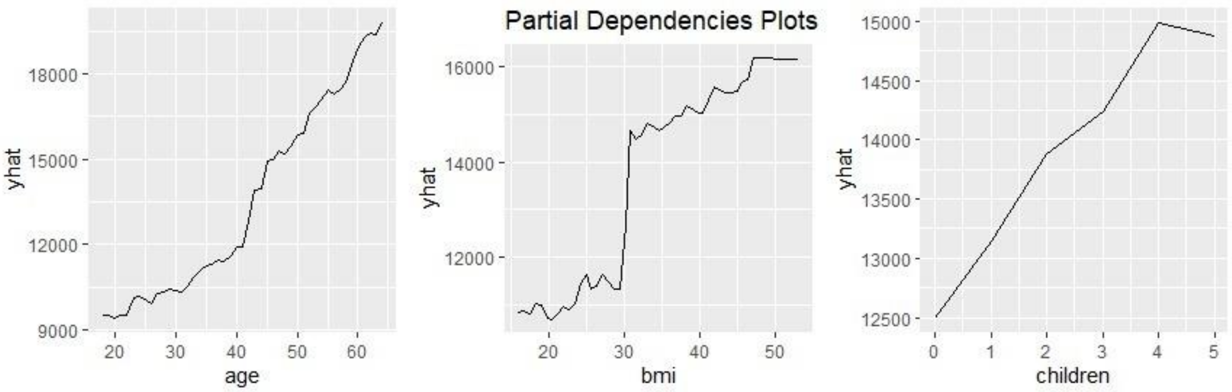

Figure 10 - Partial Dependencies plots

Conclusion

Comparing the models obtained from Linear Regression, Decision tree, and Random Forest, the better fit is seen for Random Forest and then linear model. The independent variables age, BMI, smoker, and children have high significance. An increase in these variables increases insurance charges.

References

Stedy. (n.d.). *stedy/Machine-Learning-with-R-datasets*. GitHub. https://github.com/stedy/Machine-Learning-with-R-datasets.

Centers for Disease Control and Prevention. (2021, March 3). *Defining Adult Overweight & Obesity*. Centers for Disease Control and Prevention. https://www.cdc.gov/obesity/adult/defining.html.

pdp: An R Package for Constructing Partial Dependence Plots. • pdp. (n.d.). https://bgreenwell.github.io/pdp/articles/pdp.html.

APPENDIX

```
# Import the libraries
library(tidyverse)
library(PerformanceAnalytics)
library(caret)
library(caTools)
library(gridExtra)
library(grid)
library(ggplot2)
library(randomForest)
library(vip)
library(pdp)
library(funModeling)
library(magrittr)
library(skimr)
library(caret)
library(cowplot)
options(scipen = 999)
options(repr.plot.width=12, repr.plot.height=8)


# Transform variables to factor
df$sex<- factor(df$sex)
df$smoker<- factor(df$smoker)
df$region<- factor(df$region)
df$children <- factor(df$children)

# check for missing values
df %>%
  is.na() %>%
  sum()
# check data types
df %>%
  str()
skim(df)

figsize <- options(repr.plot.width=12, repr.plot.height=12) # set plot size for this plot

# Smoker count plot
smoker <- df %>%
  ggplot(aes(x=smoker, fill=smoker)) +
  geom_bar(show.legend = FALSE) +
  # add percentages on top of bars
  geom_text(
    stat='count',
    aes(label=paste0(round(after_stat(prop*100), digits=1), "%"),group=1),
    vjust=-0.4,
    size=4
  ) +
  # add labels
  labs(
    x = "",
    y = "",
    title = "Number of insurer by smoking"
  ) +
  # rename x-ticks
  scale_x_discrete(
    labels = c("no" = "Non-smoker", "yes" = "Smoker")
  ) +
  # adjust y-ticks
  scale_y_continuous(
    breaks=seq(0,2000,100)
  ) +
  # resize text
  theme(
    plot.title = element_text(size=16),
    axis.text.x = element_text(size=14),
    axis.text.y = element_text(size=14)
  )
smoker

# Region count plot
region <- df %>%
  ggplot(aes(x=forcats::fct_infreq(region), fill=region)) +
  geom_bar(show.legend = FALSE) +
  # add percentages on top of bars
  geom_text(
    stat='count',
    aes(label = paste0(round(after_stat(prop*100), digits=1), "%"), group=1),
    vjust=-0.4,
    size=4
  ) +
  # add labels
  labs(
    x = "",
    y = "",
    title = "Number of insurer by region"
  ) +
  # rename x-ticks
  scale_x_discrete(
    labels = c("northeast" = "North East", "northwest" = "North West",
               "southeast" = "South East", "southwest" = "South West")
  ) +
  # adjust ticks
  scale_y_continuous(
    breaks=seq(0,350,50)
  ) +
  # resize text
  theme(
    plot.title = element_text(size=16),
    axis.text.x = element_text(size=14),
    axis.text.y = element_text(size=14)
  )

region

# Sex count plot
sex <- df %>%
  ggplot(aes(x=forcats::fct_infreq(sex), fill=sex)) +
  geom_bar(show.legend = FALSE) +
  # add percentages on top of bars
  geom_text(
    stat='count',
    aes(
      label=paste0(round(after_stat(prop*100), digits=1), "%"), group=1),
    vjust=-0.4,
    size=4
  ) +
  # add labels
  labs(
    x = "",
    y = "",
    title = "Number of insurer by sex",
    fill = "Sex"
  ) +
  # rename x-ticks
  scale_x_discrete(
    labels = c("male" = "Male", "female" = "Female")
  ) +
  # adjust y-ticks
  scale_y_continuous(
    breaks=seq(0,700,100)
  ) +
  # resize text
  theme(
    plot.title = element_text(size=16),
    axis.text.x = element_text(size=14),
    axis.text.y = element_text(size=14)
  )

sex

children <- df %>%
  ggplot(aes(x=forcats::fct_infreq(children), fill=children)) +
  geom_bar(show.legend = FALSE) +
  # add percentages
  geom_text(
    stat='count',
    aes(label=paste0(round(after_stat(prop*100), digits=1), "%"), group=1),
    vjust=-0.4,
    size=4
  ) +
  # add labels
  labs(
    x = "",
    y = "",
    title = "Number of dependents per policy"
  ) +
  # rename x-ticks
  scale_x_discrete(
    labels = c("0" = "None")
  ) +
  # adjust y-ticks
  scale_y_continuous(
    breaks=seq(0,600,50)
  ) +
  # resize text
  theme(
    plot.title = element_text(size=16),
    axis.text.x = element_text(size=14),
    axis.text.y = element_text(size=14)
  )

children
```

```
# Plot grid
cowplot::plot_grid(
  smoker, region, sex, children,
  labels="AUTO",
  ncol = 2,
  nrow = 2
)

options(figsize)
figsize <- options(repr.plot.width=20, repr.plot.height=16)

# Age distribution
age_hist <- df %>%
  ggplot(aes(x=age))+
  geom_histogram(
    binwidth = 5,
    show.legend = FALSE,
    fill="cadetblue2"
  )+
  labs(
    x = "Ages of insurer",
    y = "Number of insurers",
    title = "Distribution of ages"
  )+
  # resize text
  theme(
    plot.title = element_text(size=16),
    axis.text = element_text(size=14),
    axis.title = element_text(size=14)
  )
age_hist

age_dens <- df %>%
  ggplot(aes(x=age)) +
  geom_density(
    alpha=.3,
    fill="darkseagreen"
  )+
  labs(
    x = "Ages of insurer",
    y = "Probability density",
    title = "Distribution of ages"
  )+
  # resize text
  theme(
    plot.title = element_text(size=16),
    axis.text = element_text(size=14),
    axis.title = element_text(size=14)
  )
age_dens

age_box <- df %>%
  ggplot(aes(y=age)) +
  geom_boxplot(
    alpha=.5,
    fill="orange"
  )+
  coord_flip() +
  theme(
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank()
  )+
  labs(
    y = "Ages of insurer",
    x = "",
    title = "Distribution of ages"
  )+
  # resize text
  theme(
    plot.title = element_text(size=16),
    axis.text = element_text(size=14),
    axis.title = element_text(size=14)
  )
age_box


bmi_hist <- df %>%
  ggplot(aes(x=bmi))+
  geom_histogram(
    binwidth = 4,
    show.legend = FALSE,
    fill = "pink"
  )+
  labs(
    x = "BMI scores of insurer",
    y = "Number of insurer",
    title = "Distribution of BMI scores"
  )+
  # resize text
  theme(
    plot.title = element_text(size=16),
    axis.text = element_text(size=14),
    axis.title = element_text(size=14)
  )
bmi_hist

bmi_dens <- df %>%
  ggplot(aes(x=bmi)) +
  geom_density(
    alpha=.3,
    fill="orange"
  )+
  labs(
    x = "BMI scores of insurer",
    y = "Probability density",
    title = "Distribution of BMI scores"
  )+
  # resize text
  theme(
    plot.title = element_text(size=16),
    axis.text = element_text(size=14),
    axis.title = element_text(size=14)
  )
bmi_dens

bmi_box <- df %>%
  ggplot(aes(y=bmi)) +
  geom_boxplot(
    alpha=.5,
    fill="red"
  )+
  coord_flip() +
  theme(
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank()
  )+
  labs(
    y = "BMI scores of insurer",
    x = "",
    title = "Distribution of BMI scores"
  )+
  # resize text
  theme(
    plot.title = element_text(size=16),
    axis.text = element_text(size=14),
    axis.title = element_text(size=14)
  )
bmi_box



charges_hist <- df %>%
  ggplot(aes(x=charges)) +
  geom_histogram(
    binwidth = 2000,
    show.legend = FALSE,
    fill = "blue",
    col = "black"
  )+
  labs(
    x = "Charges to insurer ($)",
    y = "Number of insurers",
    title = "Distribution of medical charges"

  )+
  # resize text
  theme(
    plot.title = element_text(size=16),
    axis.text = element_text(size=14),
    axis.title = element_text(size=14)
  )
charges_hist

charges_dens <- df %>%
  ggplot(
    aes(x=charges)
  ) +
  geom_density(
    alpha=.3,
    fill="chocolate"
  ) +
  labs(
    x = "Charges to insurer ($)",
    y = "Probability density",
    title = "Distribution of medical charges"
  ) +
  # resize text
  theme(
    plot.title = element_text(size=16),
    axis.text = element_text(size=14),
    axis.title = element_text(size=14)
  )
charges_dens


charges_box <- df %>%
  ggplot(aes(y=charges))+
  geom_boxplot(
    alpha=.5,
    fill="darksalmon"
  )+
  coord_flip()+
  # remove ticks from y-axis
  theme(
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank()
  )+
  labs(
    y = "Charges to insurer($)",
    x = "",
    title = "Distribution of medical charges"
  )+
  # resize text
  theme(
    plot.title = element_text(size=16),
    axis.text = element_text(size=14),
    axis.title = element_text(size=14)
  )
charges_box

cowplot::plot_grid(
  age_hist, age_dens, age_box,
  bmi_hist, bmi_dens, bmi_box,
  charges_hist, charges_dens, charges_box,
  labels="AUTO",
  ncol = 3,
  nrow = 3
)
```

```
#Linear Regression
# Reading the CSV file
df=read.csv("insurance.csv")
head(df)

#Basic Model
lm.all <- lm(charges~., data=df)
lm.all
summary(lm.all)
plot(lm.all)


# New variables to handle non- linear relationship
df$age2<- df$age^2
df$bmi30 <- ifelse(df$bmi >= 30 , 1, 0)
# Plots for age and age2
plot.age <- ggplot(df, aes(x = age, y = charges)) +
  geom_point(colour="blue")+ggtitle("Charges for age")
plot.age2 <- ggplot(df, aes(x = age2, y = charges)) +
  geom_point(colour="blue") +ggtitle("Charges for age2")

grid.arrange(plot.age, plot.age2, ncol=2)


# Split dataset
set.seed(123)
split<- sample.split(df$charges, SplitRatio = 0.7)
df_train<- subset(df, split == T)
df_test<- subset(df, split == F)

# Define training control
train.control <- trainControl(method = "repeatedcv",
                              number = 10, repeats = 3)



model2 <- train(charges ~ age +age2  +  children + bmi+bmi30+ smoker,
                df_train, method="lm", trControl=train.control)
# Summary model 2
summary(model2)

#RMSE
print(model2)
```

```
# Applying the train model to the test data

df_test$pred <- predict(model2, df_test)

#PLot coorelation
df_test %>%
  ggplot(aes(pred, charges)) +
  geom_point() +
  geom_line(aes(y=pred), color = "red", size = 2) +ggtitle("Actual vs predicted charges for the test
data")+
  labs(x="Predicted Value", y="Charges")

# Random Forest and Decision tree code



### Splitting Train and Test######
set.seed(1)
train = sort(sample(1:nrow(df), nrow(df)*3/4))
df.test=df[-train,"charges"]

### Fitting Random Forest Model ########
rf.insuranceCost <- randomForest(charges ~ .,data=df, subset = train, mtry = 2, importance =
TRUE)
yhat.rf = predict(rf.insuranceCost,newdata=df[-train,])

### Predicted vs actual Values and calculating mse #######
plot(yhat.rf,df.test, main = "Random Forest fit", xlab = "Predict", ylab = "Actual")
abline(0,1)
dev.off()
mean((yhat.rf-df.test)^2)

## Plotting parameter importance
importance(rf.insuranceCost)


vip(rf.insuranceCost, num_features = 6, geom = "point", horizontal = FALSE,aesthetics = list(color
= "red", shape = 17, size = 4)) +
  theme_light()
#### Partial Dependence Plots ##########
p1 <- partial(rf.insuranceCost, pred.var = "age", plot = TRUE, plot.engine = "ggplot2")
p2 <- partial(rf.insuranceCost, pred.var = "bmi", plot = TRUE, plot.engine = "ggplot2") +
ggtitle("Partial Dependence Plots")
p3 <- partial(rf.insuranceCost, pred.var = "children", plot = TRUE, plot.engine = "ggplot2")
```

```
grid.arrange(p1, p2, p3, ncol = 3)


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


library(tree)

##### Tree Model applied#######
tree.df=tree(charges~.,df,subset=train)

####### Plotting Tree Model#######
plot(tree.df)
text(tree.df,pretty=0)


# Now we use the cv.tree() function to see whether pruning the tree will
# improve performance.
cv.df=cv.tree(tree.df)
plot(cv.df$size,cv.df$dev,type='b', main = "Deviance vs Size", xlab = "Size",
    ylab = "Deviance")

#### Pruning with size = 5 as deviance was least at this size.####
prune.df=prune.tree(tree.df,best=5)
plot(prune.df)
text(prune.df,pretty=0)

#### Making prediction on Test set########
yhat=predict(tree.df,newdata=df[-train,])
#df.test=df[-train,"charges"]
plot(yhat,df.test, main = "Predicted Vs Actuals", xlab = "Predicted Value",
    ylab = "Actual Value")
abline(0,1)
mean((yhat-df.test)^2)
```