Machine Learning Project

Name: Naga Tejaswi Veluri

G Number:G01335094

```
# Suppress dplyr summarise grouping warning messages
options(dplyr.summarise.inform = FALSE)

## Add R libraries here
library(tidyverse)
library(tidymodels)

library(skimr)
library(Hmisc)
library(ggsci)
library(caret)
library(rpart.plot)
library(rpart)
library(ggpubr)
library(gridExtra)
library(corrplot)

# Load the dataset
telecom_df <-  readRDS(url('https://gmubusinessanalytics.netlify.app/data/tel
ecom_df.rds'))

telecom_df

## # A tibble: 1,175 × 19
##    canceled_service senior_citizen spouse_partner dependents cellular_serv
ice
##    <fct>            <fct>          <fct>          <fct>      <fct>
##  1 no               no             no             no         single_line
##  2 no               no             yes            yes        single_line
##  3 no               no             yes            no         multiple_line
s
##  4 yes              yes            yes            no         multiple_line
s
##  5 no               no             no             yes        multiple_line
s
##  6 no               yes            no             no         single_line
##  7 no               yes            yes            no         multiple_line
s
##  8 no               no             no             no         multiple_line
s
##  9 no               no             yes            yes        single_line
## 10 no               yes            yes            yes        multiple_line
```

```
s
## # … with 1,165 more rows, and 14 more variables: avg_call_mins <dbl>,
## #   avg_intl_mins <dbl>, internet_service <fct>, online_security <fct>,
## #   online_backup <fct>, device_protection <fct>, tech_support <fct>,
## #   streaming_tv <fct>, streaming_movies <fct>, contract <fct>,
## #   paperless_bill <fct>, payment_method <fct>, months_with_company <dbl>,
## #   monthly_charges <dbl>
```

Data Analysis

In this section, you must think of at least 5 relevant questions that explore the relationship between canceled_service and the other variables in the telecom_df data set. The goal of your analysis should be discovering which variables drive the differences between customers who do and do not cancel their service.

You must answer each question and provide supporting data summaries with either a summary data frame (using dplyr/tidyr) or a plot (using ggplot) or both.

In total, you must have a minimum of 3 plots (created with ggplot) and 3 summary data frames (created with dplyr) for the exploratory data analysis section. Among the plots you produce, you must have at least 3 different types (ex. box plot, bar chart, histogram, scatter plot, etc…)

See the Data Analysis Project for an example of a question answered with a summary table and plot.

Note: To add an R code chunk to any section of your project, you can use the keyboard shortcut Ctrl + Alt + i or the insert button at the top of your R project template notebook file.

#divinding the data set based on cancelled service.

```
customers_not_in_Service <- telecom_df %>%
   group_by(canceled_service)%>%filter(canceled_service=="yes")



customers_not_in_Service

## # A tibble: 427 × 19
## # Groups:   canceled_service [1]
##    canceled_service senior_citizen spouse_partner dependents cellular_serv
ice
##    <fct>            <fct>          <fct>          <fct>      <fct>
##  1 yes              yes            yes            no         multiple_line
s
##  2 yes              no             yes            yes        single_line
##  3 yes              no             yes            no         multiple_line
s
##  4 yes              no             no             no         single_line
##  5 yes              no             no             yes        multiple_line
s
##  6 yes              no             no             no         multiple_line
s
##  7 yes              no             no             no         single_line
##  8 yes              no             yes            no         multiple_line
s
##  9 yes              no             yes            yes        multiple_line
s
## 10 yes              no             no             no         multiple_line
s
## # … with 417 more rows, and 14 more variables: avg_call_mins <dbl>,
## #   avg_intl_mins <dbl>, internet_service <fct>, online_security <fct>,
## #   online_backup <fct>, device_protection <fct>, tech_support <fct>,
## #   streaming_tv <fct>, streaming_movies <fct>, contract <fct>,
## #   paperless_bill <fct>, payment_method <fct>, months_with_company <dbl>,
## #   monthly_charges <dbl>
```

```
customers_in_Service <- telecom_df %>%
   group_by(canceled_service)%>%filter(canceled_service=="no")


customers_in_Service
```

```
## # A tibble: 748 × 19
## # Groups:   canceled_service [1]
##    canceled_service senior_citizen spouse_partner dependents cellular_serv
ice
##    <fct>            <fct>          <fct>          <fct>      <fct>
##  1 no               no             no             no         single_line
##  2 no               no             yes            yes        single_line
##  3 no               no             yes            no         multiple_line
s
##  4 no               no             no             yes        multiple_line
s
##  5 no               yes            no             no         single_line
##  6 no               yes            yes            no         multiple_line
s
##  7 no               no             no             no         multiple_line
s
##  8 no               no             yes            yes        single_line
##  9 no               yes            yes            yes        multiple_line
s
## 10 no               no             yes            no         multiple_line
s
## # … with 738 more rows, and 14 more variables: avg_call_mins <dbl>,
## #   avg_intl_mins <dbl>, internet_service <fct>, online_security <fct>,
## #   online_backup <fct>, device_protection <fct>, tech_support <fct>,
## #   streaming_tv <fct>, streaming_movies <fct>, contract <fct>,
## #   paperless_bill <fct>, payment_method <fct>, months_with_company <dbl>,
## #   monthly_charges <dbl>
```

Question 1

Question: Is Cellular services effecting the customer churning

Answer: We observe high count for customers availing multiple lines are more tend to leave the service.

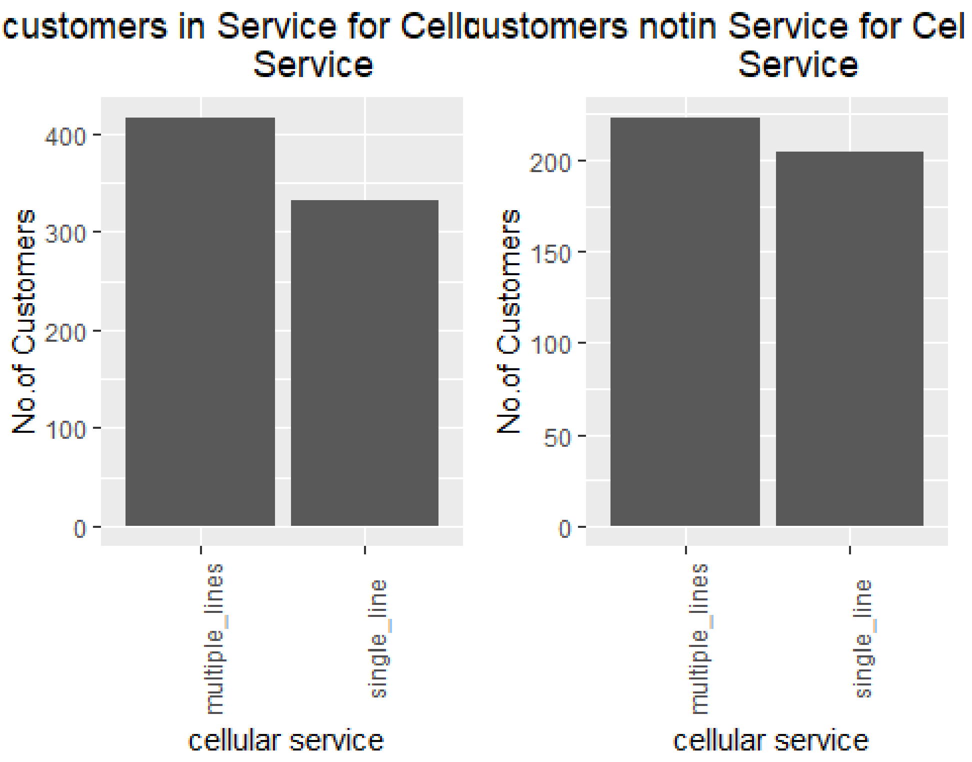
```
Cellular_inService=customers_in_Service %>% group_by(cellular_service) %>%
    summarise(count = n())
Cellular_inService
```

```
## # A tibble: 2 × 2
##   cellular_service count
##   <fct>            <int>
## 1 multiple_lines     416
## 2 single_line        332
```

```
c1=ggplot(Cellular_inService, aes(x=cellular_service, y=count)) +
  geom_bar(position="dodge" ,stat="identity")+
  labs(title = "customers in Service for Cellular in
       Service ", x=" cellular service", y="No.of Customers")+
  theme(axis.text.x = element_text(angle = 90))+
  theme(plot.title = element_text(hjust = 0.5))

Cellular_no_inService=customers_not_in_Service %>% group_by(cellular_service)
%>%
    summarise(count = n())
Cellular_no_inService
```

```
## # A tibble: 2 × 2
##   cellular_service count
##   <fct>            <int>
## 1 multiple_lines     223
## 2 single_line        204
```

```
c2=ggplot(Cellular_no_inService, aes(x=cellular_service, y=count)) +
  geom_bar(position="dodge" ,stat="identity")+
  labs(title = "customers notin Service for Cellular in
       Service ", x=" cellular service", y="No.of Customers")+
  theme(axis.text.x = element_text(angle = 90))+
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(c1,c2,ncol=2,nrow=1)
```



Question 2

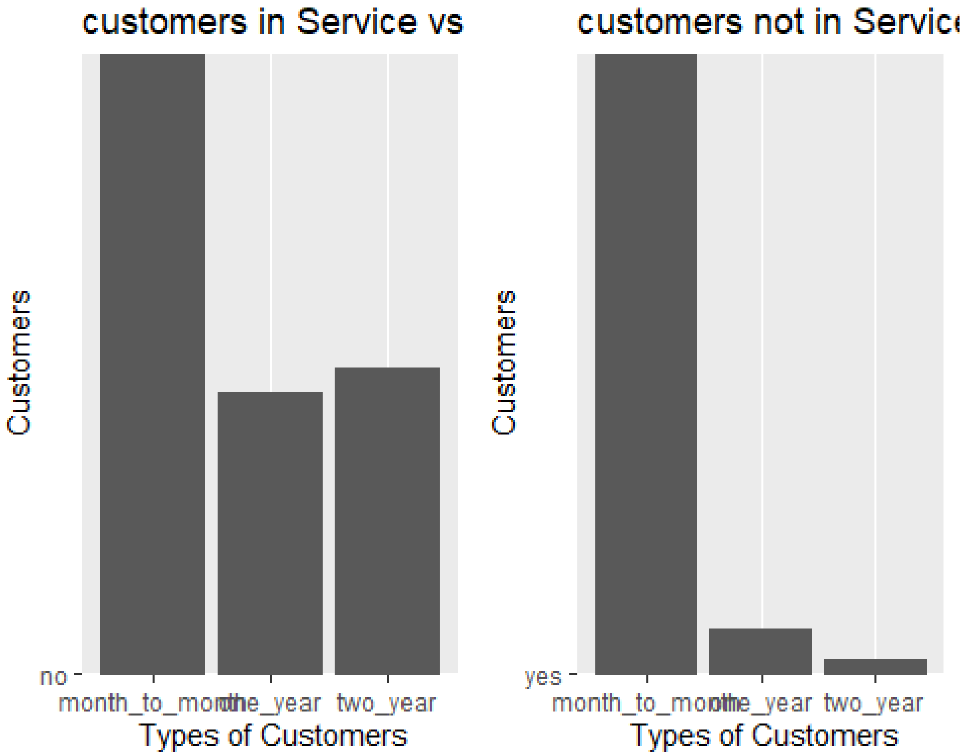
Question: Are long term customers more prone to retain in the service?

Answer: As per the above analysis, we see much profits from one-year and two-year plan customers. Even there are many customers using month-to month plan, there are equalvent customers churning the month-to month plan.

```
contract=customers_in_Service %>% group_by(contract) %>%
    summarise(count = n())
contract

## # A tibble: 3 × 2
##   contract       count
##   <fct>          <int>
## 1 month_to_month   383
## 2 one_year         175
## 3 two_year         190

c1=ggplot(data=customers_in_Service, aes(x=contract,y=canceled_service)) +
geom_bar(stat="identity") +labs(title = "customers in Service vs Contract", x
=" Types of Customers", y=" Customers")

contract=customers_not_in_Service %>% group_by(contract) %>%
    summarise(count = n())
contract
```

```
## # A tibble: 3 × 2
##   contract       count
##   <fct>          <int>
## 1 month_to_month   388
## 2 one_year          29
## 3 two_year          10
```

```
c2=ggplot(data=customers_not_in_Service, aes(x=contract,y=canceled_service))
+
geom_bar(stat="identity") +labs(title = "customers not in Service Vs Contract
", x=" Types of Customers", y=" Customers")
grid.arrange(c1,c2,ncol=2,nrow=1)
```



Question 3

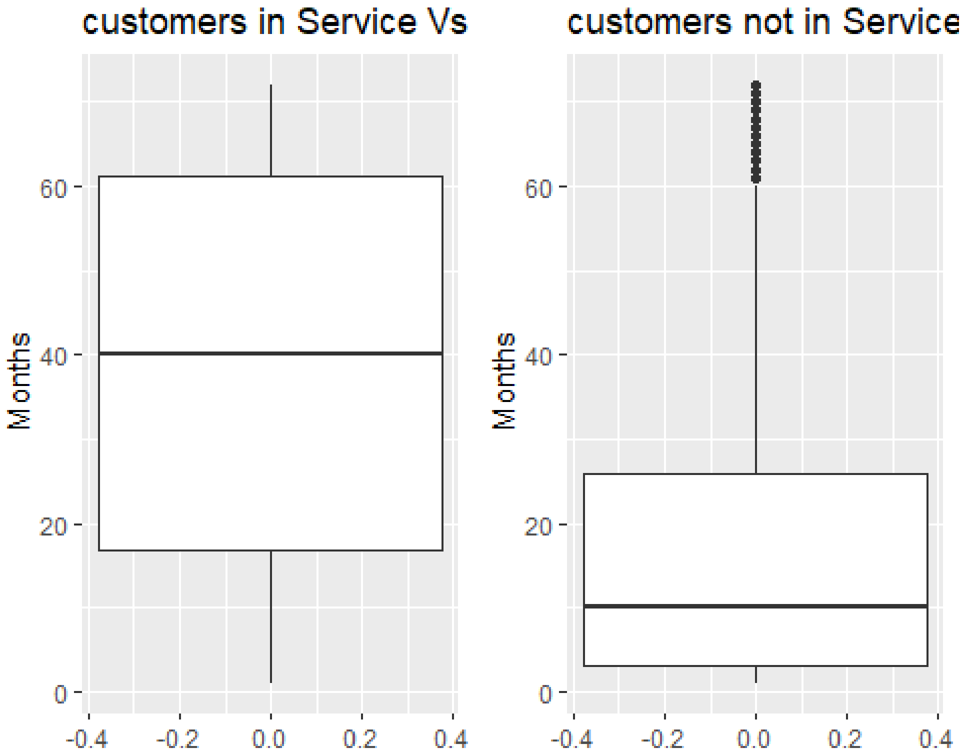
Question: Which Category of people are more prone to leave the service based on their retention history.

Answer: As the months pass on, the customers tend to stayin the services.

```
month=summary(customers_in_Service$months_with_company)
month
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00   16.75   40.00   38.66   61.25   72.00
```

```
month2=summary(customers_not_in_Service$months_with_company)
month2

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.00   10.00   17.69   26.00   72.00

b1 = ggplot(customers_in_Service, aes( y=months_with_company)) +
    geom_boxplot()+labs(title = "customers in Service Vs Months in Company",
y=" Months")

b2= ggplot(customers_not_in_Service, aes( y=months_with_company)) +
    geom_boxplot()+labs(title = "customers not in Service Vs Months in Compan
y", y=" Months")

grid.arrange(b1,b2,ncol=2,nrow=1)
```



Question 4

Question: Is age playing a vital role in retaining the customers?

Answer: No,major customers who are churning the services are not senior citizens.

```
contract=telecom_df %>% group_by(canceled_service,senior_citizen) %>%
    summarise(count = n())
contract
```

```
## # A tibble: 4 × 3
## # Groups:   canceled_service [2]
##   canceled_service senior_citizen count
##   <fct>            <fct>          <int>
## 1 yes              yes              113
## 2 yes              no               314
## 3 no               yes              118
## 4 no               no               630
```

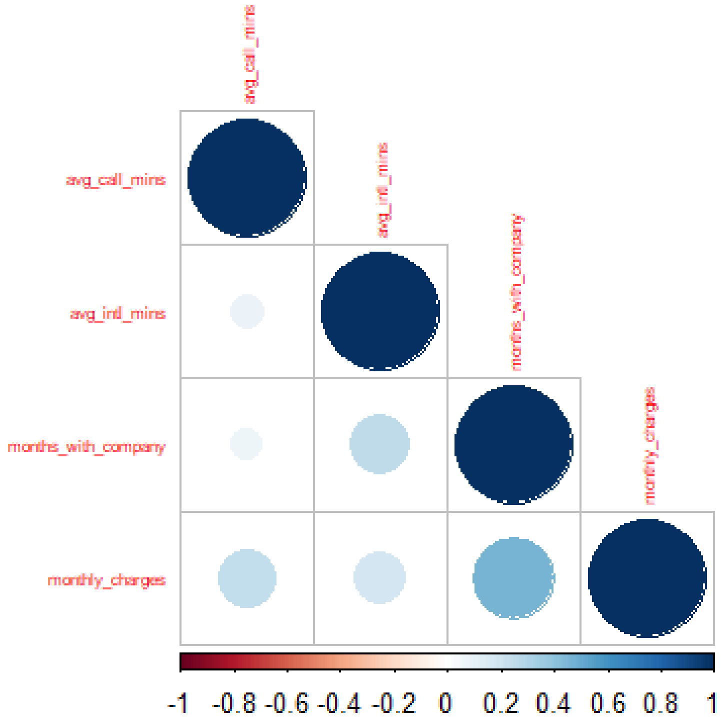
Question 5

Question: What factors are having high influence on customer retaining?

Answer:

We see months in company is nwgatively co-related with the customers leaving the company.

```
#check correlation of all numeric variables
df_num = select_if(telecom_df,is.numeric)
df_num = data.frame(lapply(df_num, function(x) as.numeric(as.character(x))))
res=cor(df_num)
res

##                     avg_call_mins avg_intl_mins months_with_company
## avg_call_mins          1.00000000    0.08486267          0.07687995
## avg_intl_mins          0.08486267    1.00000000          0.25809140
## months_with_company    0.07687995    0.25809140          1.00000000
## monthly_charges        0.24215582    0.18879005          0.46837059
##                     monthly_charges
## avg_call_mins             0.2421558
## avg_intl_mins             0.1887901
## months_with_company       0.4683706
## monthly_charges           1.0000000

corrplot(res, type="lower",tl.cex=0.5 )
```



Question 6

Question: Is lines with dependents and spouse are more stable in network?

Answer: Customers with no dependents and spouse in the line tend to continue for long period.

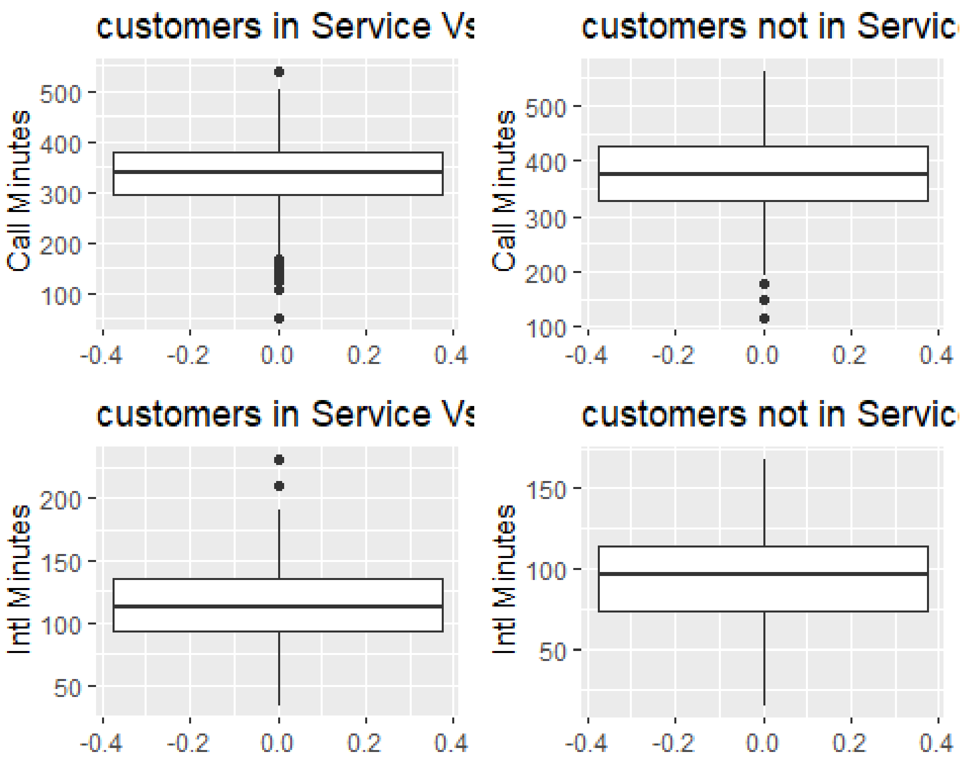
```
dep=telecom_df %>% group_by(canceled_service,dependents,spouse_partner) %>%
    summarise(count = n())
dep

## # A tibble: 8 × 4
## # Groups:   canceled_service, dependents [4]
##   canceled_service dependents spouse_partner count
##   <fct>            <fct>      <fct>          <int>
## 1 yes              yes        yes               65
## 2 yes              yes        no                21
## 3 yes              no         yes               86
## 4 yes              no         no               255
## 5 no               yes        yes              177
## 6 no               yes        no                34
## 7 no               no         yes              214
## 8 no               no         no               323
```

Question 7

Question: Is there any pattern for early detection of churning based on the usage of call minutes and internet call minutes

Answer: Any signs of early detection were not noticed.

```
mint1=summary(customers_in_Service$avg_call_mins)
mint1

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    52.0   294.8   340.0   336.2   380.2   539.0

mint2=summary(customers_not_in_Service$avg_call_mins)
mint2

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   118.0   328.5   375.0   376.2   426.0   562.0

int1=summary(customers_in_Service$avg_intl_mins)
int1

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    34.0    93.0   112.0   113.5   135.0   231.0

int2=summary(customers_not_in_Service$avg_intl_mins)
int2

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   15.00   74.00   96.00   93.57  114.00  168.00

b1 = ggplot(customers_in_Service, aes( y=avg_call_mins)) +
    geom_boxplot()+labs(title = "customers in Service Vs Avg Call Minutes", y
=" Call Minutes")

b2= ggplot(customers_not_in_Service, aes( y=avg_call_mins)) +
    geom_boxplot()+labs(title = "customers not in Service Vs Avg Call Minutes
", y=" Call Minutes")
b3= ggplot(customers_in_Service, aes( y=avg_intl_mins)) +
    geom_boxplot()+labs(title = "customers in Service Vs Avg Intl Minutes", y
=" Intl Minutes")
b4= ggplot(customers_not_in_Service, aes( y=avg_intl_mins)) +
    geom_boxplot()+labs(title = "customers not in Service Vs Avg Intl Minutes
", y=" Intl Minutes")
grid.arrange(b1,b2,b3,b4,ncol=2,nrow=2)
```



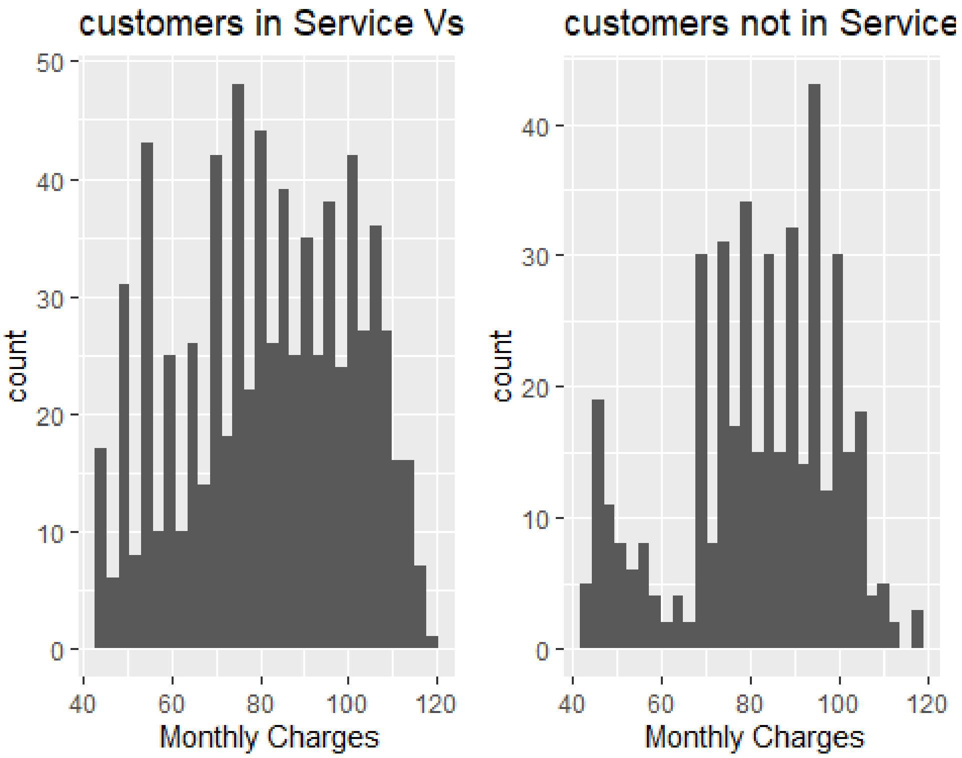
Question 8

Question: Is monthly charges leading the customers to leave the service?

Answer: No,as we don’t see much difference in the price payed by the customers who are in the service and are leaving.

```
cost1=summary(customers_in_Service$monthly_charges)
cost1

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   43.45   68.41   82.72   81.91   98.86  118.60

cost2=summary(customers_not_in_Service$monthly_charges)
cost2

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   44.15   72.72   84.50   81.88   94.72  118.35

c1 = ggplot(customers_in_Service, aes( y=monthly_charges)) +
    geom_histogram()+labs(title = "customers in Service Vs Monthly Charges",
y=" Monthly Charges")+coord_flip()

c2= ggplot(customers_not_in_Service, aes( y=monthly_charges)) +
    geom_histogram()+labs(title = "customers not in Service Vs Monthly Charge
s", y=" Monthly Charges")+coord_flip()
```

```
grid.arrange(c1,c2,ncol=2,nrow=1)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Question 9

Question: Are Streaming options helping the firm to retain customers?

Answer: As per the analysis, a huge portion of customers are not availing the services of streaming.

```
stream=telecom_df %>% group_by(canceled_service,streaming_tv,streaming_movies
) %>%
    summarise(count = n())
stream

## # A tibble: 8 × 4
## # Groups:   canceled_service, streaming_tv [4]
##   canceled_service streaming_tv streaming_movies count
##   <fct>            <fct>        <fct>            <int>
## 1 yes              yes          yes                160
## 2 yes              yes          no                 106
## 3 yes              no           yes                 83
## 4 yes              no           no                  78
## 5 no               yes          yes                140
## 6 no               yes          no                 202
## 7 no               no           yes                170
## 8 no               no           no                 236
```

Question 10

Question: Are security plans adding customer time line in the company?

Answer: As per the analysis, there is no significant impact on customer retantion with the help of these security plans.

```
sec=telecom_df %>% group_by(canceled_service,online_security,online_backup,de
vice_protection,tech_support) %>%
    summarise(count = n())
sec
```

```
## # A tibble: 32 × 6
## # Groups:   canceled_service, online_security, online_backup, device_prote
ction
## #   [16]
##    canceled_service online_security online_backup device_protection tech_s
upport
##    <fct>            <fct>           <fct>         <fct>             <fct>
##  1 yes              yes             yes           yes               yes
##  2 yes              yes             yes           yes               no
##  3 yes              yes             yes           no                yes
##  4 yes              yes             yes           no                no
##  5 yes              yes             no            yes               yes
##  6 yes              yes             no            yes               no
##  7 yes              yes             no            no                yes
##  8 yes              yes             no            no                no
##  9 yes              no              yes           yes               yes
## 10 yes              no              yes           yes               no
## # … with 22 more rows, and 1 more variable: count <int>
```

Question 11

Question: Customers with which internet service type are using the service for long time?

Answer: We observe a grater proportion of customers are opting fiber-optic service service for their internet.

```
int_ser=telecom_df %>% group_by(canceled_service,internet_service) %>%
    summarise(count = n())
int_ser
```

```
## # A tibble: 4 × 3
## # Groups:   canceled_service [2]
##   canceled_service internet_service count
##   <fct>            <fct>            <int>
## 1 yes              fiber_optic        354
## 2 yes              digital             73
## 3 no               fiber_optic        427
## 4 no               digital            321
```

Machine Learning

In this section of the project, you will fit **three classification algorithms** to predict the response variable,`canceled_service`. You should use all of the other variables in the `telecom_df` data as predictor variables for each model.

You must follow the machine learning steps below.

The data splitting and feature engineering steps should only be done once so that your models are using the same data and feature engineering steps for training.

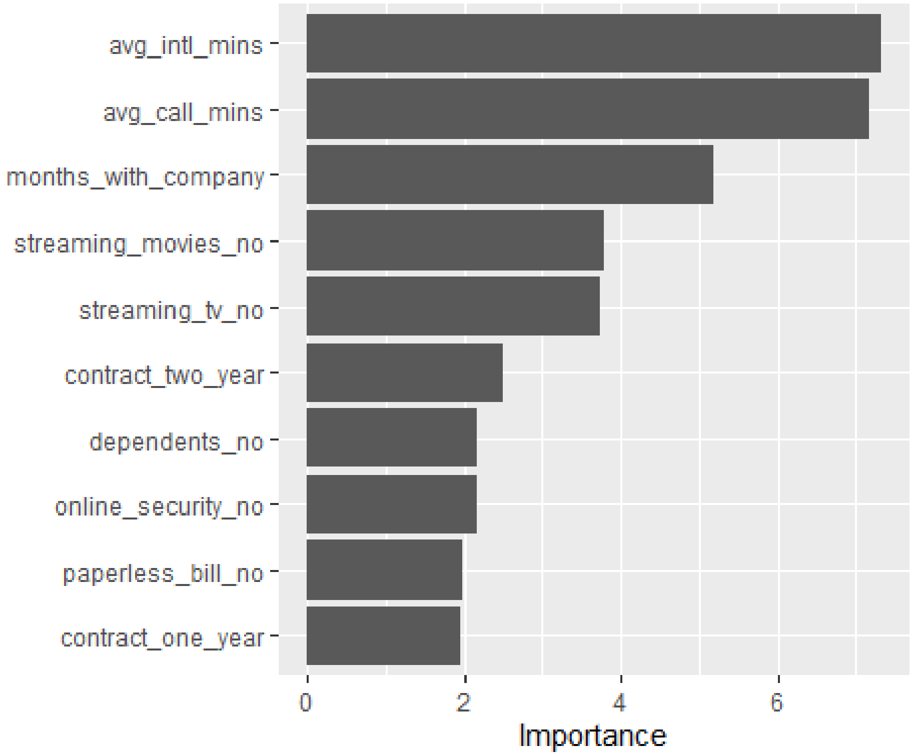
- Split the `telecom_df` data into a training and test set (remember to set your seed)
- Specify a feature engineering pipeline with the `recipes` package
  - You can include steps such as skewness transformation, dummy variable encoding or any other steps you find appropriate
- Specify a `parsnip` model object
  - You may choose from the following classification algorithms:
    - Logistic Regression
    - LDA
    - QDA
    - KNN
    - Decision Tree
    - Random Forest
- Package your recipe and model into a workflow
- Fit your workflow to the training data
  - If your model has hyperparameters:
    - Split the training data into 5 folds for 5-fold cross validation using `vfold_cv` (remember to set your seed)
    - Perform hyperparamter tuning with a random grid search using the `grid_random()` function
    - Refer to the following tutorial for an example - Random Grid Search
    - Hyperparameter tuning can take a significant amount of computing time. Be careful not to set the `size` argument of `grid_random()` too large. I recommend `size` = 10 or smaller.
    - Select the best model with `select_best()` and finalize your workflow
- Evaluate model performance on the test set by plotting an ROC curve using `autoplot()` and calculating the area under the ROC curve on your test data

```r
library(tidymodels)
library(vip)

## Warning: package 'vip' was built under R version 4.2.1

##
## Attaching package: 'vip'

## The following object is masked from 'package:utils':
##
##     vi

library(rsample)
library(recipes)
library(ranger)

## Warning: package 'ranger' was built under R version 4.2.1

library(FIT)

## Warning: package 'FIT' was built under R version 4.2.1

##
## Attaching package: 'FIT'

## The following object is masked from 'package:caret':
##
##     train

## The following objects are masked from 'package:stats':
##
##     optim, predict

library(ranger)

# Create data split object
telecom_Service_split <- initial_split(telecom_df, prop = 0.75,
                     strata = canceled_service)

# Create the training data
telecom_Service_training <- telecom_Service_split %>%
  training()

# Create the test data
telecom_Service_test <- telecom_Service_split %>%
  testing()
set.seed(300)

# Check the number of rows
nrow(telecom_Service_training)

## [1] 881
```

```
nrow(telecom_Service_test)

## [1] 294

telecom_Service_folds <- vfold_cv(telecom_Service_training, v = 5)


my_metrics <- metric_set(accuracy, roc_auc)
#Feature Engineering
telecom_Service_recipe <- recipe(canceled_service ~ ., data = telecom_Service
_training) %>%
                   step_YeoJohnson(all_numeric(), -all_outcomes()) %>%
                   step_normalize(all_numeric(), -all_outcomes()) %>%
                   step_dummy(all_nominal(), -all_outcomes())

telecom_Service_recipe %>%
  prep(training = telecom_Service_training) %>%
  bake(new_data = NULL)

## # A tibble: 881 × 22
##    avg_call_mins avg_intl_mins months_with_com… monthly_charges canceled_s
ervice
##            <dbl>         <dbl>            <dbl>           <dbl> <fct>
##  1        -0.633       -0.0446            0.337          -0.881 no
##  2        -1.13         0.600            -0.775          -1.69  no
##  3         0.234        2.26              1.22            0.440 no
##  4         0.220       -0.540             0.839           1.53  no
##  5        -0.389        1.09             -0.306           1.40  no
##  6        -0.620        1.36              1.20            1.45  no
##  7        -0.834       -1.39              1.35            1.29  no
##  8        -1.35        -1.51              0.589           1.26  no
##  9        -0.834       -0.852             1.33           -0.522 no
## 10        -0.579        0.324             0.269          -0.881 no
## # … with 871 more rows, and 17 more variables: senior_citizen_no <dbl>,
## #   spouse_partner_no <dbl>, dependents_no <dbl>,
## #   cellular_service_single_line <dbl>, internet_service_digital <dbl>,
## #   online_security_no <dbl>, online_backup_no <dbl>,
## #   device_protection_no <dbl>, tech_support_no <dbl>, streaming_tv_no <db
l>,
## #   streaming_movies_no <dbl>, contract_one_year <dbl>,
## #   contract_two_year <dbl>, paperless_bill_no <dbl>, …
```

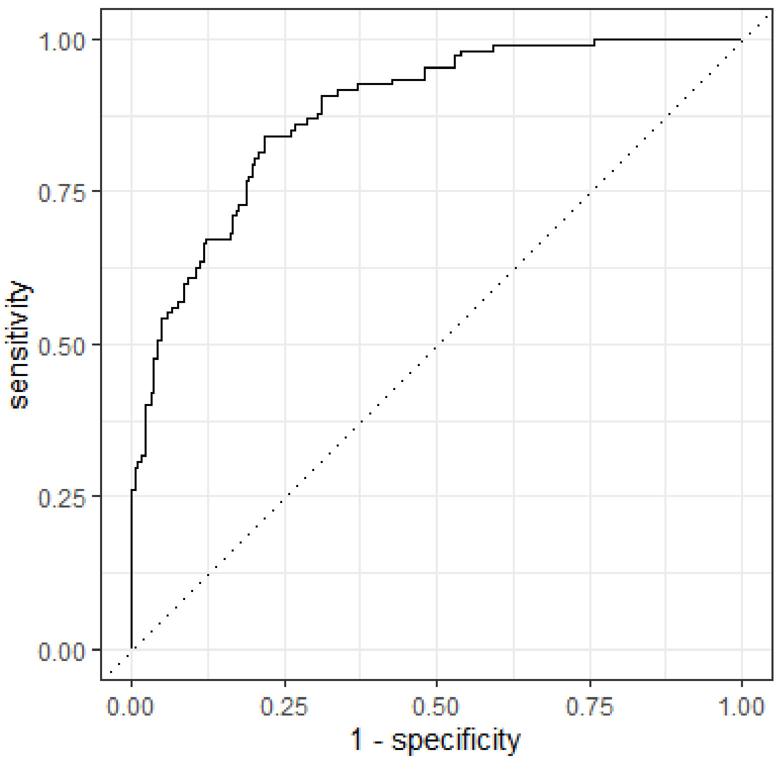
Model 1-logistic regression model

```
# Specify a logistic regression model
logistic_model <- logistic_reg() %>%
  set_engine('glm') %>%
  set_mode('classification')

telecom_Service_workflow <- workflow() %>%
            add_model(logistic_model) %>%
            add_recipe(telecom_Service_recipe)




telecom_Service_fit_model <- telecom_Service_workflow %>%
                  last_fit(split = telecom_Service_split,
                           metrics = my_metrics)
telecom_Service_trained_model <- telecom_Service_fit_model %>%
    extract_fit_parsnip()

vip(telecom_Service_trained_model)
```



```
telecom_Service_metrics<-telecom_Service_fit_model %>%
  collect_metrics()

telecom_Service_metrics

## # A tibble: 2 × 4
##   .metric  .estimator .estimate .config
##   <chr>    <chr>          <dbl> <chr>
## 1 accuracy binary         0.782 Preprocessor1_Model1
## 2 roc_auc  binary         0.881 Preprocessor1_Model1

telecom_Service_fit_results <- telecom_Service_fit_model %>%
                    collect_predictions()
```

```
telecom_Service_fit_results

## # A tibble: 294 × 7
##    id               .pred_class  .row .pred_yes .pred_no canceled_service .
config
##    <chr>            <fct>       <int>     <dbl>    <dbl> <fct>            <
chr>
##  1 train/test spl… no              3    0.0525    0.947 no               P
repro…
##  2 train/test spl… no              6    0.410     0.590 no               P
repro…
##  3 train/test spl… yes             9    0.832     0.168 no               P
repro…
##  4 train/test spl… no             19    0.426     0.574 no               P
repro…
##  5 train/test spl… no             26    0.0467    0.953 no               P
repro…
##  6 train/test spl… no             32    0.0607    0.939 no               P
repro…
##  7 train/test spl… yes            38    0.594     0.406 no               P
repro…
##  8 train/test spl… no             39    0.366     0.634 yes              P
repro…
##  9 train/test spl… no             41    0.0431    0.957 no               P
repro…
## 10 train/test spl… no             52    0.337     0.663 no               P
repro…
## # … with 284 more rows

telecom_Service_fit_results %>%
  roc_curve(truth = canceled_service, estimate = .pred_yes) %>%
  autoplot()
```

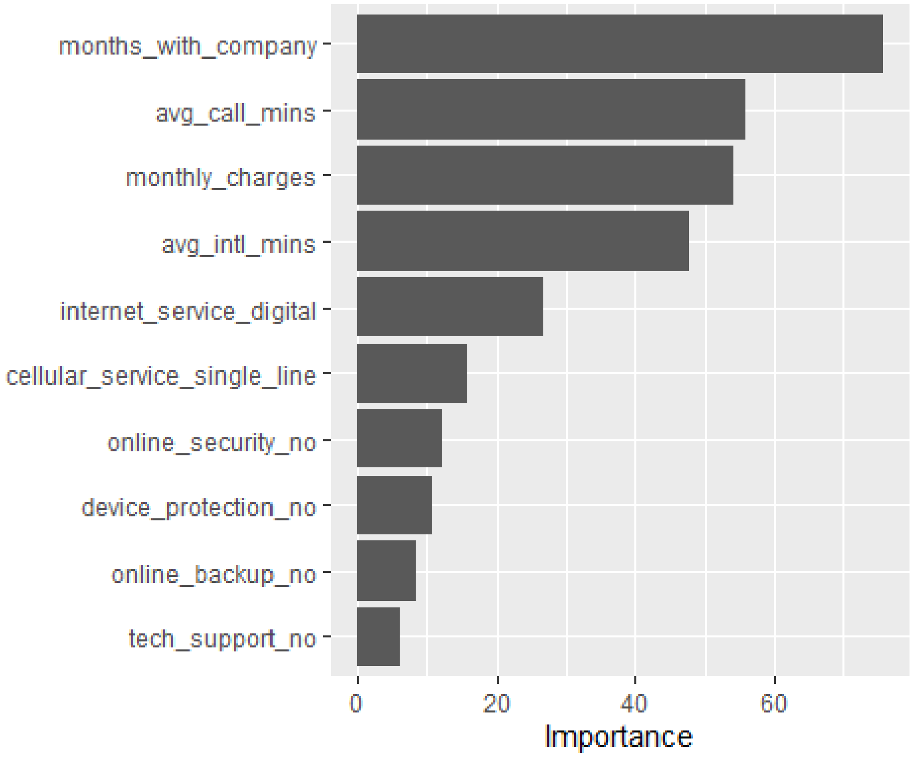


```
conf_mat(telecom_Service_fit_results, truth = canceled_service, estimate = .p
red_class)

##            Truth
## Prediction yes  no
##        yes  72  29
##        no   35 158
```

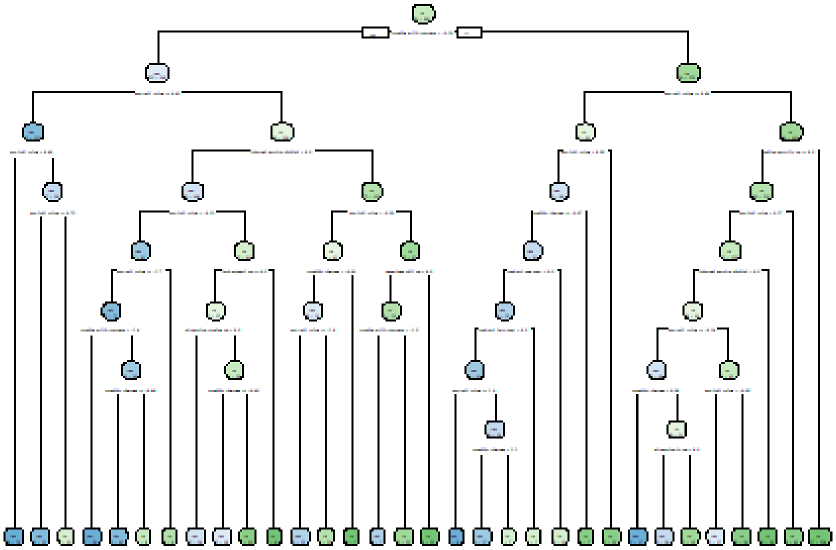
Model 2

```
#Decision tree
telecom_Service_tree_model <- decision_tree(cost_complexity = tune(),
                                            tree_depth = tune(),
                                            min_n = tune()) %>%
              set_engine('rpart') %>%
              set_mode('classification')

telecom_Service_tree_workflow <- workflow() %>%
                 add_model(telecom_Service_tree_model) %>%
                 add_recipe(telecom_Service_recipe)

telecom_Service_tree_grid <- grid_regular(cost_complexity(),
                            tree_depth(),
                            min_n(),
                          levels = 5)
```

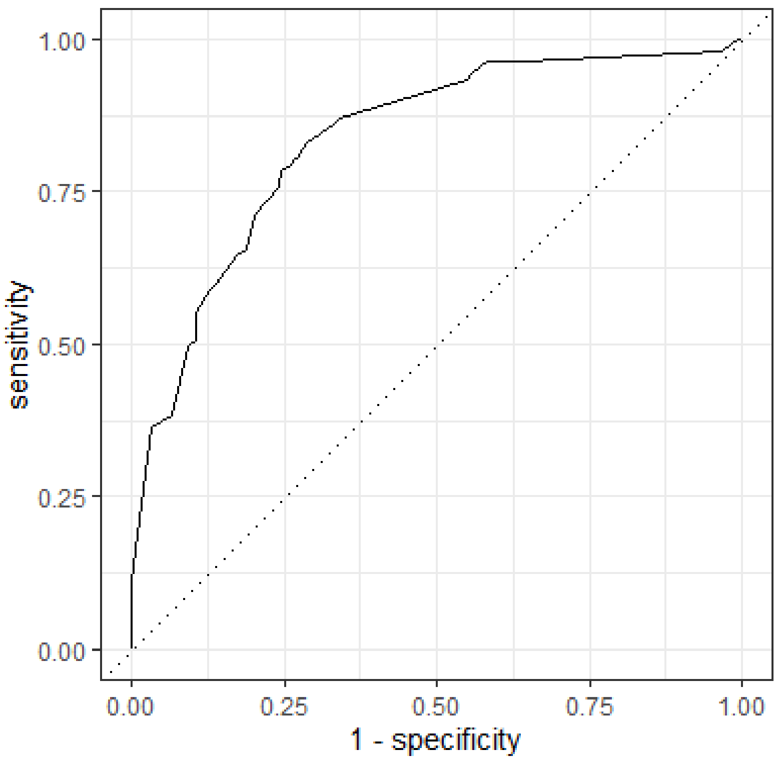
```
set.seed(300)
telecom_Service_tree_tuning <- telecom_Service_tree_workflow %>%
               tune_grid(resamples = telecom_Service_folds,
                         grid = telecom_Service_tree_grid)

telecom_Service_tree_tuning %>% collect_metrics(summarise=FALSE) %>% view()

telecom_Service_best_tree <- telecom_Service_tree_tuning %>%
             select_best(metric = 'roc_auc')

telecom_Service_final_tree_workflow <- telecom_Service_tree_workflow %>%
                       finalize_workflow(telecom_Service_best_tree)

telecom_Service_tree_wf_fit <- telecom_Service_final_tree_workflow %>%
               fit(data = telecom_Service_training)

telecom_Service_tree_fit <- telecom_Service_tree_wf_fit %>%
            extract_fit_parsnip()
vip(telecom_Service_tree_fit)
```



```
rpart.plot(telecom_Service_tree_fit$fit, roundint = FALSE,extra=2)
```



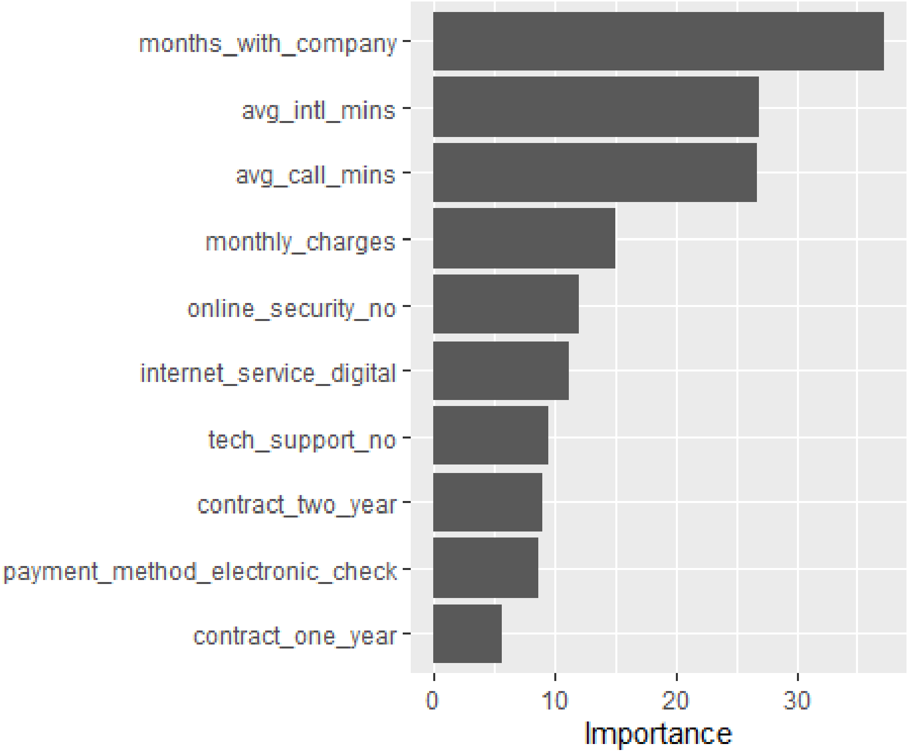
```
telecom_Service_tree_last_fit <- telecom_Service_final_tree_workflow %>%
                 last_fit(split=telecom_Service_split,metrics=my_metrics)

telecom_Service_tree_last_fit %>% collect_metrics()

## # A tibble: 2 × 4
##   .metric  .estimator .estimate .config
##   <chr>    <chr>          <dbl> <chr>
## 1 accuracy binary         0.765 Preprocessor1_Model1
## 2 roc_auc  binary         0.835 Preprocessor1_Model1

telecom_Service_tree_predictions <- telecom_Service_tree_last_fit %>% collect
_predictions()

telecom_Service_tree_predictions %>%
  roc_curve(truth = canceled_service, estimate = .pred_yes) %>%
  autoplot()
```



```
conf_mat(telecom_Service_tree_predictions, truth = canceled_service, estimate
= .pred_class)

##            Truth
## Prediction yes  no
##        yes  76  38
##        no   31 149
```

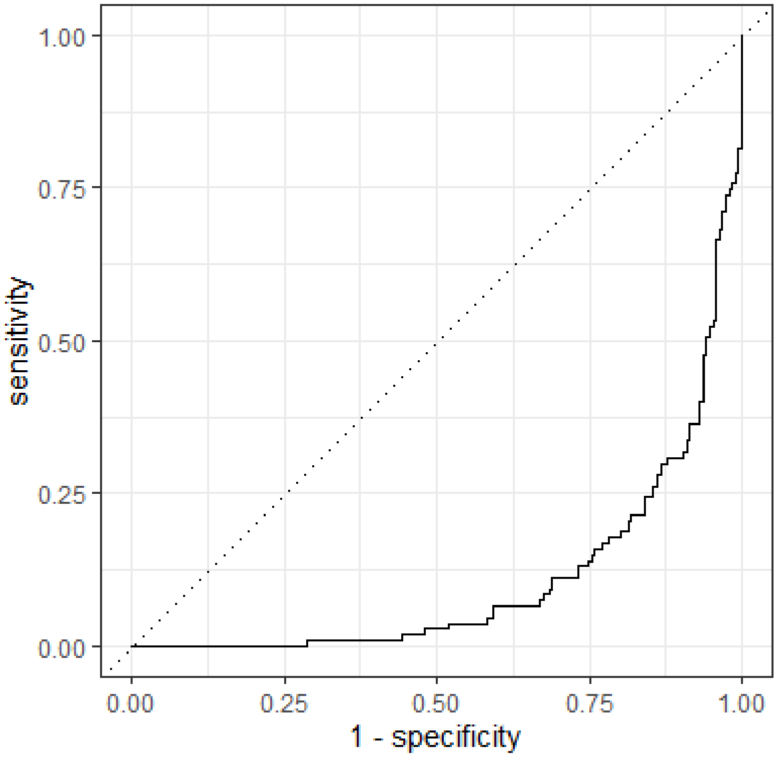
Model 3-Random Forest

```
telecom_Service_rf_model <- rand_forest(mtry = tune(),
                        trees = tune(),
                        min_n = tune()) %>%
            set_engine('ranger', importance = "impurity") %>%
            set_mode('classification')

telecom_Service_rf_workflow <- workflow() %>%
               add_model(telecom_Service_rf_model) %>%
               add_recipe(telecom_Service_recipe)

set.seed(300)

telecom_Service_rf_grid <- grid_random(mtry() %>% range_set(c(2, round(sqrt(n
col(telecom_Service_training))))),
                       trees(),
                       min_n(),
                       size = 10)

set.seed(300)

telecom_Service_rf_tuning <- telecom_Service_rf_workflow %>%
             tune_grid(resamples = telecom_Service_folds,
                       grid = telecom_Service_rf_grid)

telecom_Service_best_rf <- telecom_Service_rf_tuning %>%
           select_best(metric = 'roc_auc')

telecom_Service_final_rf_workflow <- telecom_Service_rf_workflow %>%
                     finalize_workflow(telecom_Service_best_rf)

telecom_Service_rf_wf_fit <- telecom_Service_final_rf_workflow %>%
             fit(data = telecom_Service_training)

telecom_Service_rf_fit <- telecom_Service_rf_wf_fit %>%
          extract_fit_parsnip()
vip(telecom_Service_rf_fit)
```



```
telecom_Service_rf_last_fit <- telecom_Service_final_rf_workflow %>%
               last_fit(split = telecom_Service_split,metrics=my_metrics)

telecom_Service_metrics<-telecom_Service_rf_last_fit %>% collect_metrics()
```

```
telecom_Service_metrics

## # A tibble: 2 × 4
##   .metric  .estimator .estimate .config
##   <chr>    <chr>          <dbl> <chr>
## 1 accuracy binary         0.827 Preprocessor1_Model1
## 2 roc_auc  binary         0.892 Preprocessor1_Model1

telecom_Service_rf_last_fit %>% collect_predictions() %>%
                roc_curve(truth  = canceled_service , estimate = .pred_no) %>
%
                autoplot()
```



```
telecom_Service_rf_last_fit %>% collect_predictions() %>% conf_mat(truth = ca
nceled_service, estimate = .pred_class)

##             Truth
## Prediction yes  no
##        yes  74  18
##        no   33 169
```

Summary of Results

Write a summary of your overall findings and recommendations to the executives at the company. Think of this section as your closing remarks of a presentation, where you summarize your key findings, model performance, and make recommendations to improve customer retention and service at this company.

Your executive summary must be written in a professional tone, with minimal grammatical errors, and should include the following sections:

1. An introduction where you explain the business problem and goals of your data analysis
    - What problem(s) is this company trying to solve? Why are they important to their future success?
    - What was the goal of your analysis? What questions were you trying to answer and why do they matter?
2. Highlights and key findings from your Exploratory Data Analysis section
    - What were the interesting findings from your analysis and **why are they important for the business**?
    - This section is meant to **establish the need for your recommendations** in the following section
3. Your "best" classification model and an analysis of its performance
    - In this section you should talk about the expected error of your model on future data
        - To estimate future performance, you can use your model performance results on the **test data**
    - You should discuss at least one performance metric, such as an F1, sensitivity, specificity, or ROC AUC for your model. However, you must explain the results in an **intuitive, non-technical manner**. Your audience in this case are executives at a telecommunications company with limited knowledge of machine learning.
4. Your recommendations to the company on how to reduce customer attrition rates
    - Each recommendation must be supported by your data analysis results
    - You must clearly explain why you are making each recommendation and which results from your data analysis support this recommendation
    - You must also describe the potential business impact of your recommendation:
        - Why is this a good recommendation?
        - What benefits will the business achieve?

Summary

In the current business world, with cutthroat competitors, The main aim of the telecom company is to retain their customers to maintain the business. As it would be comparatively profitable to retain the customers satisfied rather than investing a great portion in marketing to attract new customers.

During this project, we are analyzing data of telecom company which provides its services throughout United States and predict the patterns of churning customers. In general, telecom companies provide services such as calling and internet services.

This company provides four wide range of features to select from

- Calls
- Internet
- online Security
- Streaming options

The monthly charges vary according to the plan opted.

The ultimate motive of this organization is to generate retain their customers. To do so, multiple factors must be analyzed to see why the customers are churning out.

To understand the patterns and explore the factors, and attract customers, with the help of R- programming, I have completed the exploratory data analysis and developed machine learning algorithms that will predict the customers likelihood of cancelling their services with the company.

Here canceled service = yes means the customer is no more in service, and no means he is continuing the service.

Highlights of the analysis:

The organization provides two types of plan options, one of which is single line: in this type of we see that both enrolment and churning are high when compared to the other plan that is multiple lines in which the enrollment is greater and churning is comparatively low.

On the other hand, we observe see a different pattern in terms of dependents and spouse in plan. Independent and young customers have a prolonged stay in the service.

If we talk in service point of view, one of the key findings is months in the service, as the average period 26 months crosses then the customer is prone to stay in service for long term ,as the period tend increase, we see less attrition.

Customers tend to opt, month-to-month plan when compared to one-year or two-year plans. And we neither have any sort of indications in terms of usage of internet and call minutes nor in terms of monthly price.

Even the offers such as streaming tv and streaming movies is helping the firm to attract customers and neither the online protection nor support offers.

Predictions based on Models:

We used three models to predict the possible outcomes:

- Logistic Regression:
  The outcome variable (canceled_services) has categorical values such as yes/no. With the help of this model, we are measuring the probability of customers cancelling the service with an accuracy of 78.2%.

- Decision Tree:
  With an accuracy of 76.5%, the model uses most significant variables to prune the tree at each level.
- Random Forest:
  Using this model, we are building multiple decision trees and choosing the best tree of accuracy 82.7 %

  During the whole process, factors such as months with company, avg_calls_min, monthly_charges , avg_intl_mints and online_security are top five influencing factors.

  The Random Forest model, performed better in predicting the probability of customers churning in terms of accuracy.
  With the help of confusion matrix, we can see that there are only few true negatives and false negative. Which in turn is efficient in terms of minimizing the error cost of the model when implemented.

Recommendations:

1. The company does not make use of customer service details for their service implementation, making note of complaints or service calls can help understand the pit holes in the process.
2. It is always suggested to know the customers opinion and needs and the problems a customer is facing. To get these details a small survey or customer review calls can be implemented.
3. Also, attracting the customers to enrol in plans can help them to retain in the service. Which can be done by introducing value added promotions while enrolling into the plan.
4. Also giving promotional offers for new customers might attract them to get into the service. We see that as the number of months increase, the customers tend to stay in the service. To enable customer to enrol in the service any introductory offers can attract them to take the service.