Introduction

This project focuses on a dataset that describes the details about over 1000 users of a U.S telecommunications company.

The goal of this project and research is to conduct an exploratory data analysis on the telecom dataset and then prepare a summary about the factors that are likely to cause customers to cancel their mobile and internet service. This project also includes the development of machine learning algorithms in order to predict the likelihood of a customer cancelling their service in the future.

This project used R language and RStudio to analyze and perform exploratory data analysis and machine learning on the dataset and draw meaningful inferences from the dataset.

Questions to be answered

Using the above-mentioned variables and data, this project focuses on the evaluation of the data and determining the factors that primarily impact a customer's cancellation of their plan.

To attain this goal, the below mentioned questions have been formulated on the dataset, which will be answered by the support of appropriate visualizations and summary tables in the subsequent sections of this report.

1. Does the person having more average call and international call minutes make them more likely to retain their cell service?
2. Is it more likely for people to cancel their service early on, and the people who do not cancel their service to stay on the service for longer periods?
3. Does the type of internet service a person is enrolled in have an impact on them retaining their cell service?
4. Does a person having online security and device protection enabled have a positive effect on them not canceling their cell service?
5. Does the type of contract a person is registered with have an impact on them staying with the company for longer?
6. Is a person with lesser monthly charges more likely to not cancel their service?

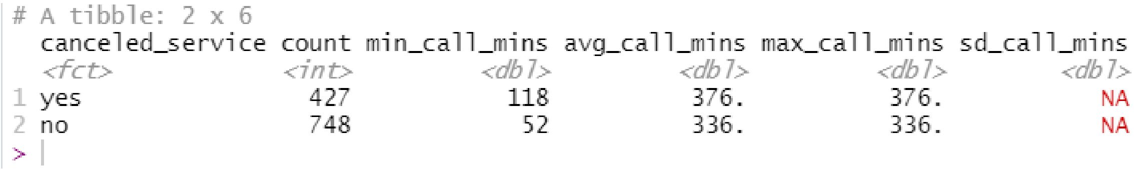
Question 1

Question: Does the person having more average call and international call minutes make them more likely to retain their cell service?

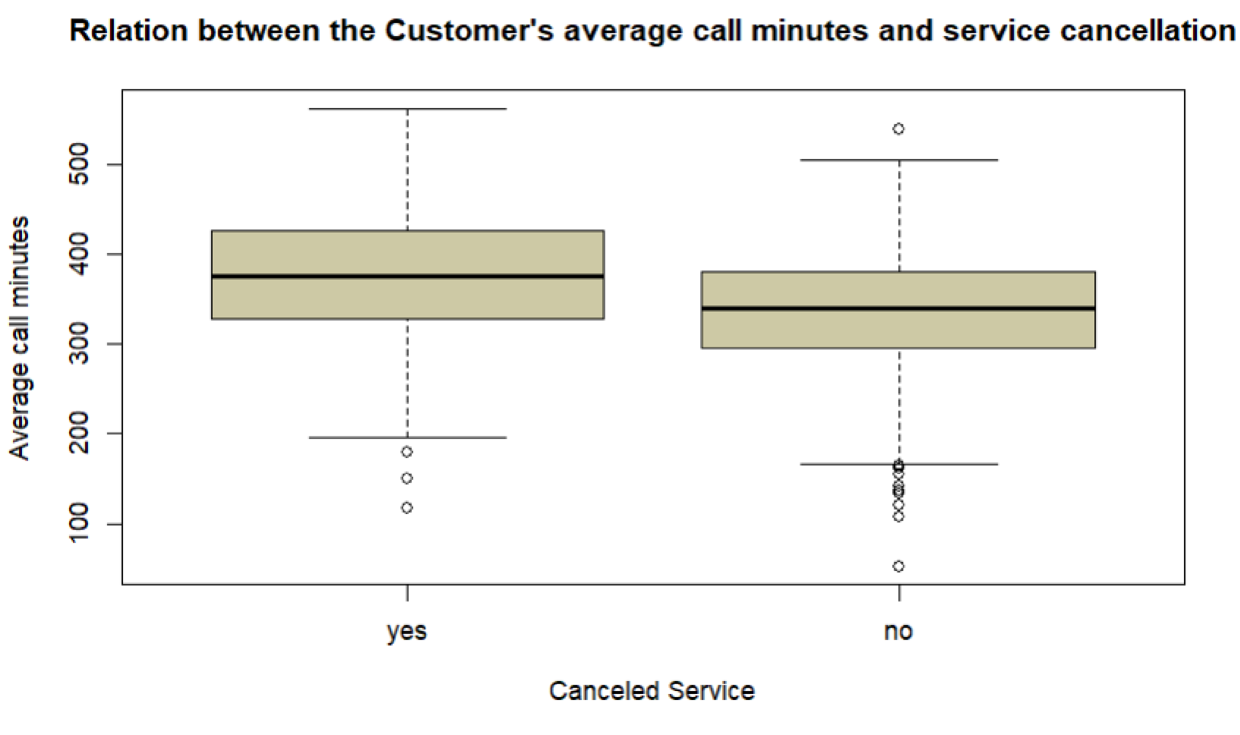
Answer:

It can be observed from the below boxplots that the customers with more call minutes are more likely to cancel their service. This could be attributed to the tendency of the customers to upgrade to a better service to match their needs.

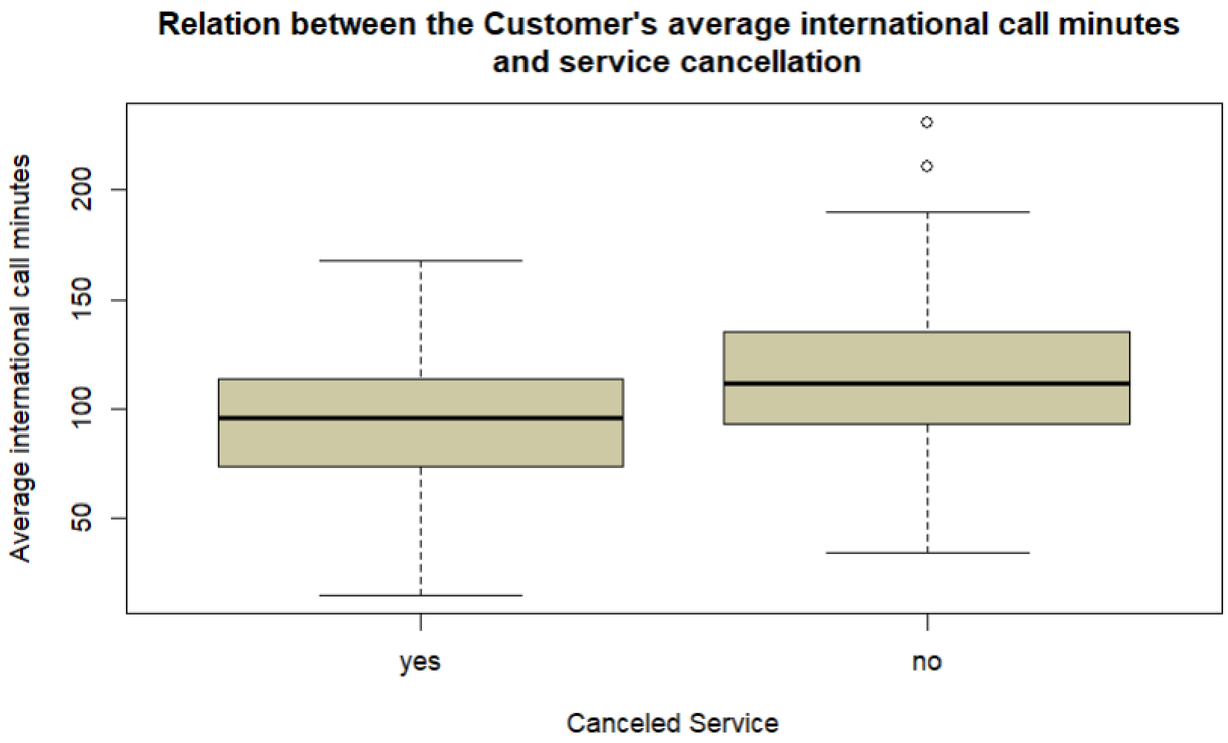
The same data is being conveyed by the summary table seen below: the average monthly call minutes of the customers who have canceled their service is higher than that of those who have not.



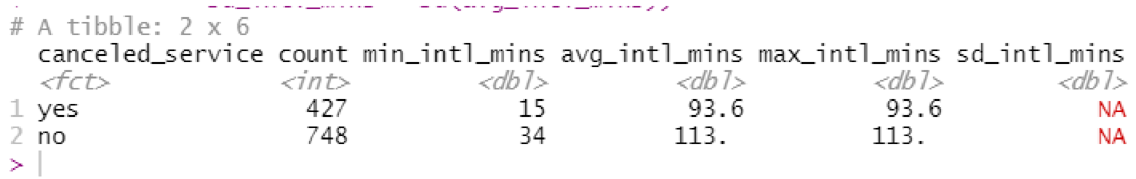
```
# A tibble: 2 x 6
  canceled_service count min_call_mins avg_call_mins max_call_mins sd_call_mins
  <fct>            <int>         <dbl>         <dbl>         <dbl>        <dbl>
1 yes                427           118          376.          376.           NA
2 no                 748            52          336.          336.           NA
> |
```



Contrary to the average call minutes, the customers with higher monthly international call minutes are more likely to retain their current plan, as can be seen in the below boxplot.



This can also be seen in the below summary table: the average monthly call minutes of the customers who have canceled their service is lower than that of those who have not. This could be attributed to the fact that there might not be very many plans that suit the customer's international call requirements.



```
# A tibble: 2 x 6
  canceled_service count min_intl_mins avg_intl_mins max_intl_mins sd_intl_mins
  <fct>            <int>         <dbl>         <dbl>         <dbl>        <dbl>
1 yes                427            15          93.6          93.6           NA
2 no                 748            34         113.          113.            NA
> |
```

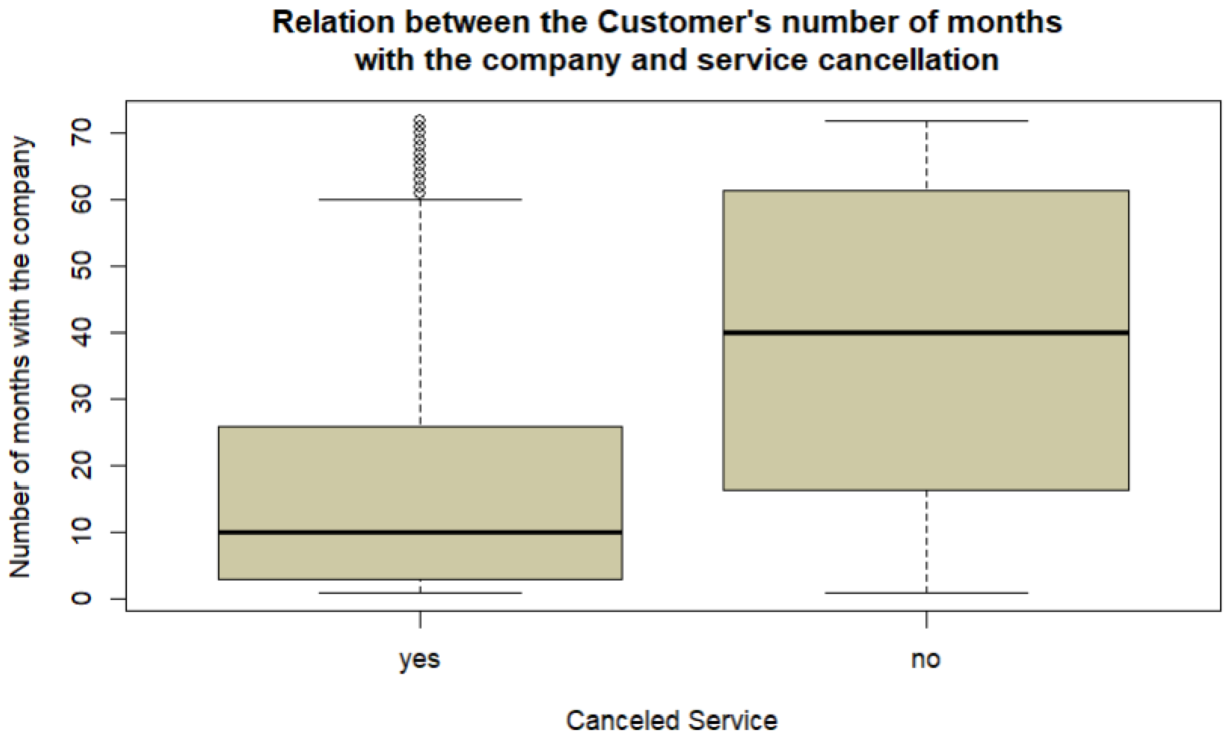
Question 2

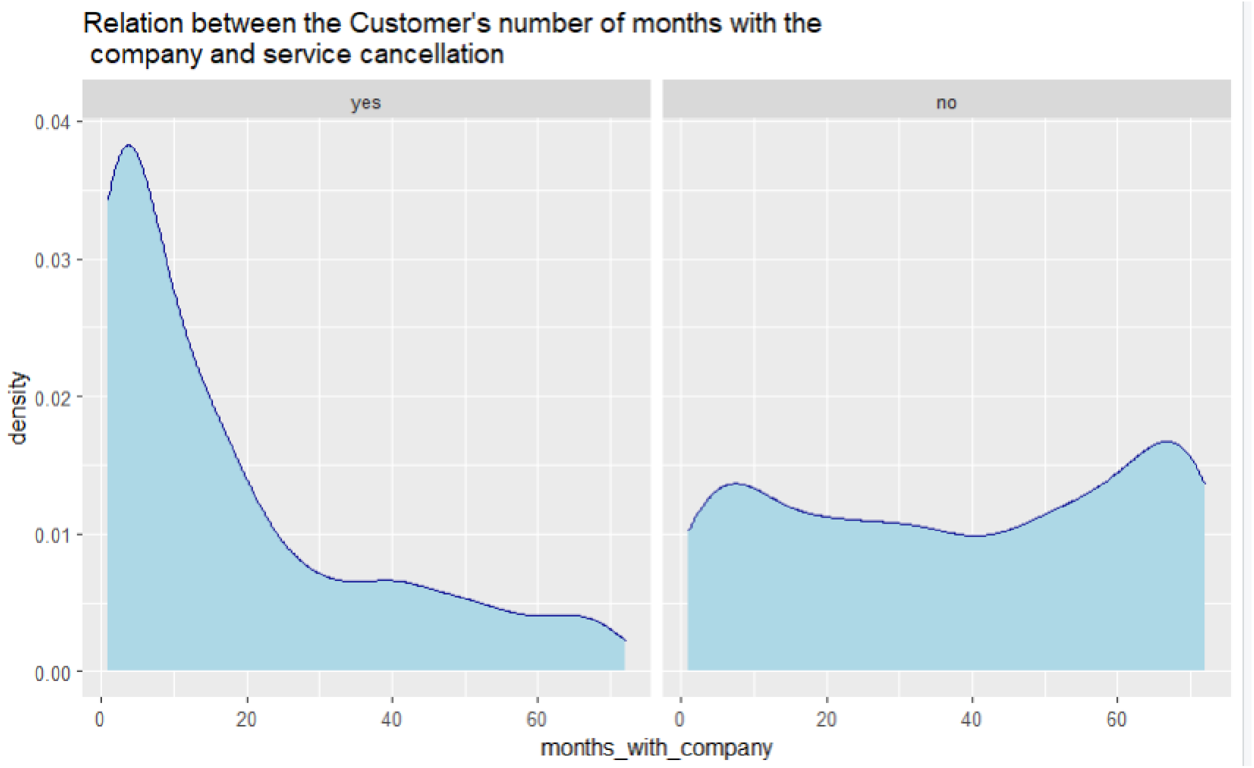
Question: Is it more likely for people to cancel their service early on, and the people who do not cancel their service to stay on the service for longer periods?

Answer:

There is a significant gap between the number of customers who cancel their service with respect to the duration for which they have stayed with the company.

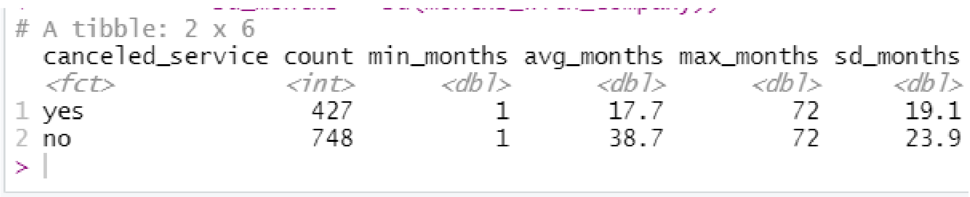
It can be seen from the below boxplots and density plots that majority of the customers who have canceled their service have done so in very short times of joining the company and the customers who have note canceled their service have stayed with the company for very long durations, significantly higher than the customers who have cancelled their service.





The same information can be seen in the below summary table:

The average number of months with the company of the customers who have canceled their service is significantly lower than those who did not.



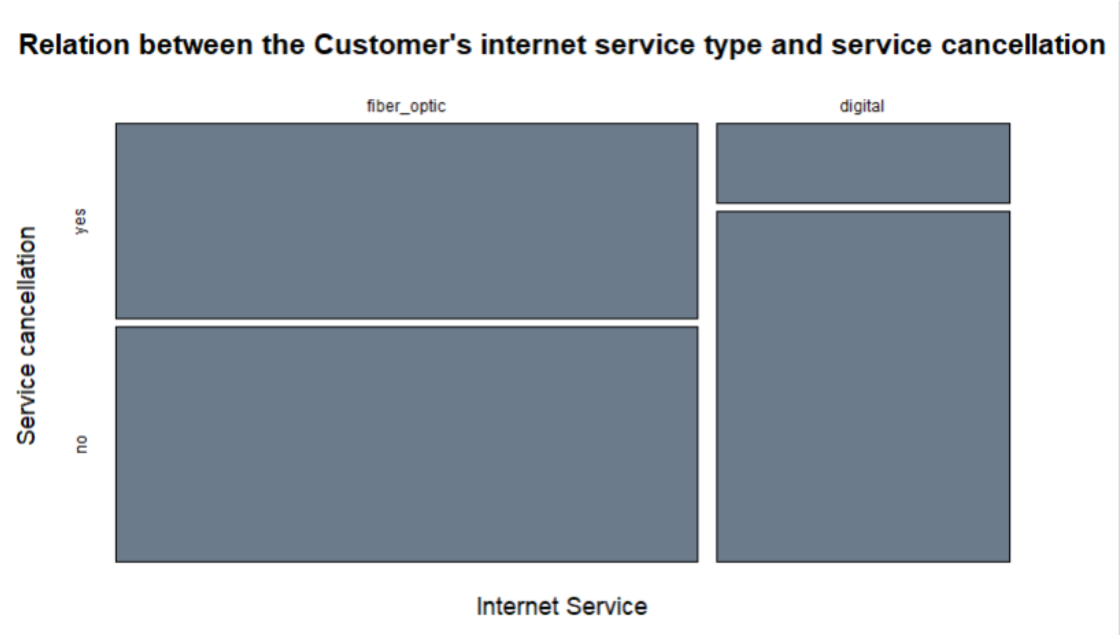
```
# A tibble: 2 x 6
  canceled_service count min_months avg_months max_months sd_months
  <fct>            <int>      <dbl>      <dbl>      <dbl>     <dbl>
1 yes                427          1       17.7         72      19.1
2 no                 748          1       38.7         72      23.9
> |
```

Question 3

Question: Does the type of internet service a person is enrolled in have an impact on them retaining their cell service?

Answer:

As can be seen from the below mosaic plot, the proportion of customers who have retained their service is higher among the customers who are enrolled with a digital internet service. This shows that the customers who have digital internet connection are less likely to cancel their service as compared to the customers who are enrolled with a fiber optic internet connection.

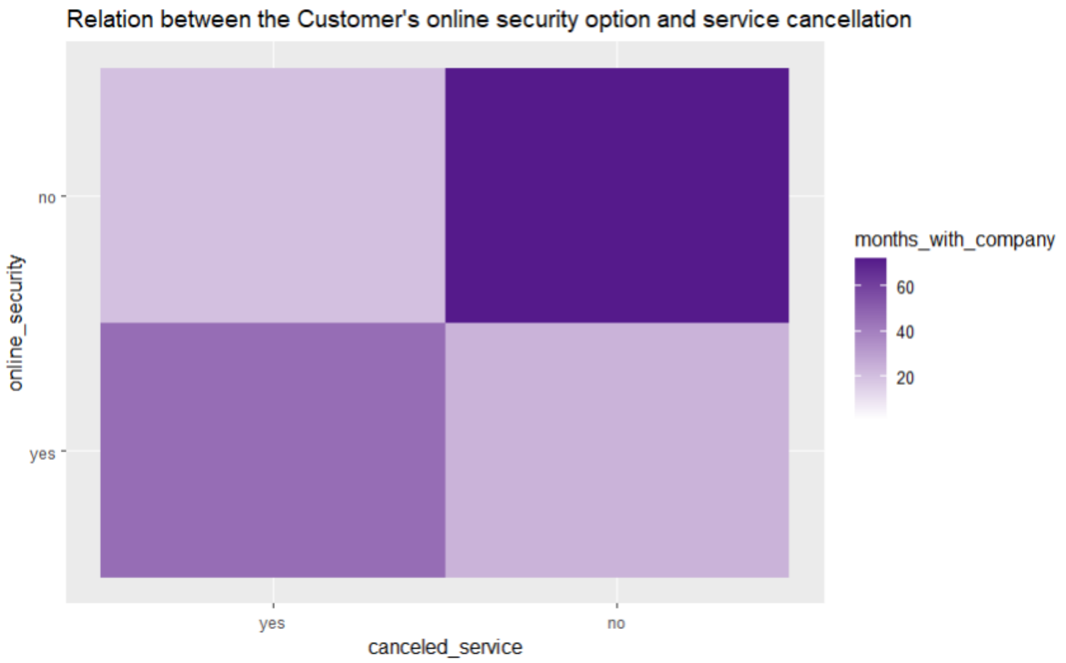

Relation between the Customer's internet service type and service cancellation

Question 4

Question: Does a person having online security and device protection enabled have a positive effect on them not canceling their cell service?

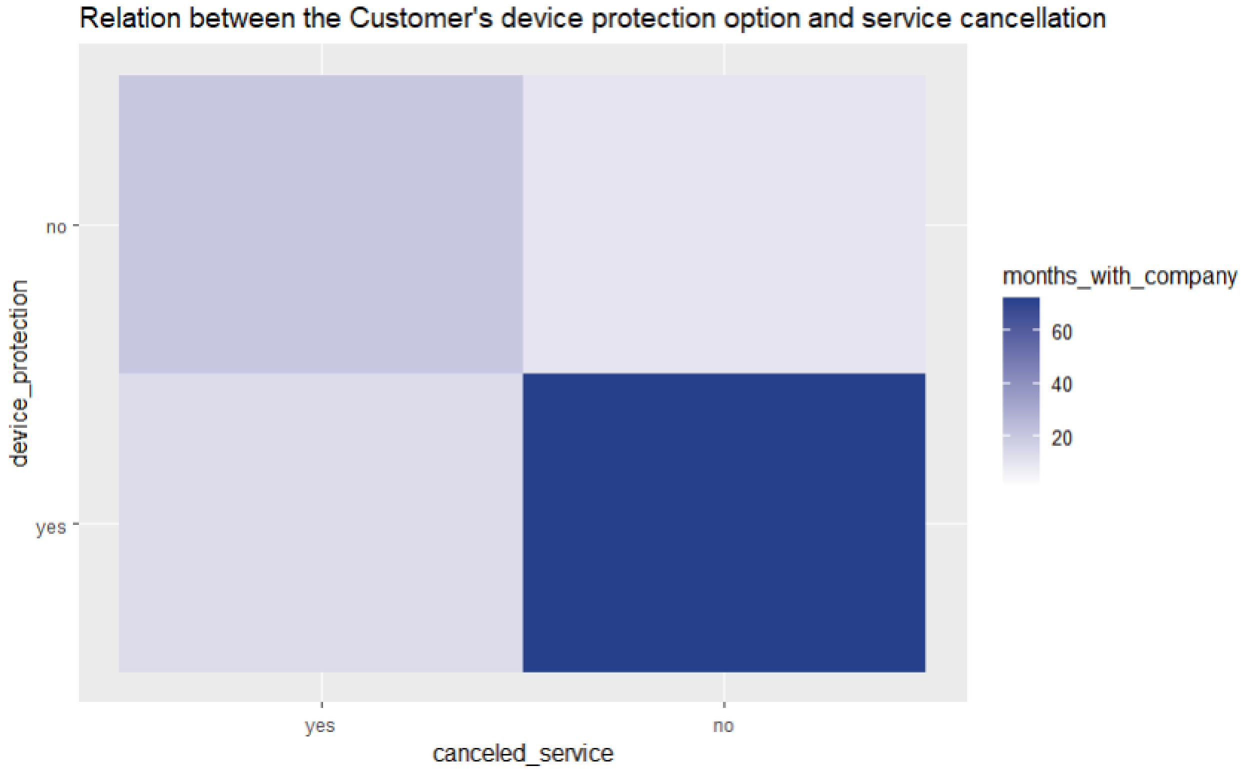
Answer:

The below heatmap shows that the customers with online security option enabled are more likely to cancel their plan and stay with the company for lesser durations as compared to the customers who do not have this option enabled.



Relation between the Customer's online security option and service cancellation

The below heatmap demonstrates that the customers with device protection option enabled are far less likely to cancel their plans and stay with the company for longer durations as compared to the customers who do not have this option enabled.



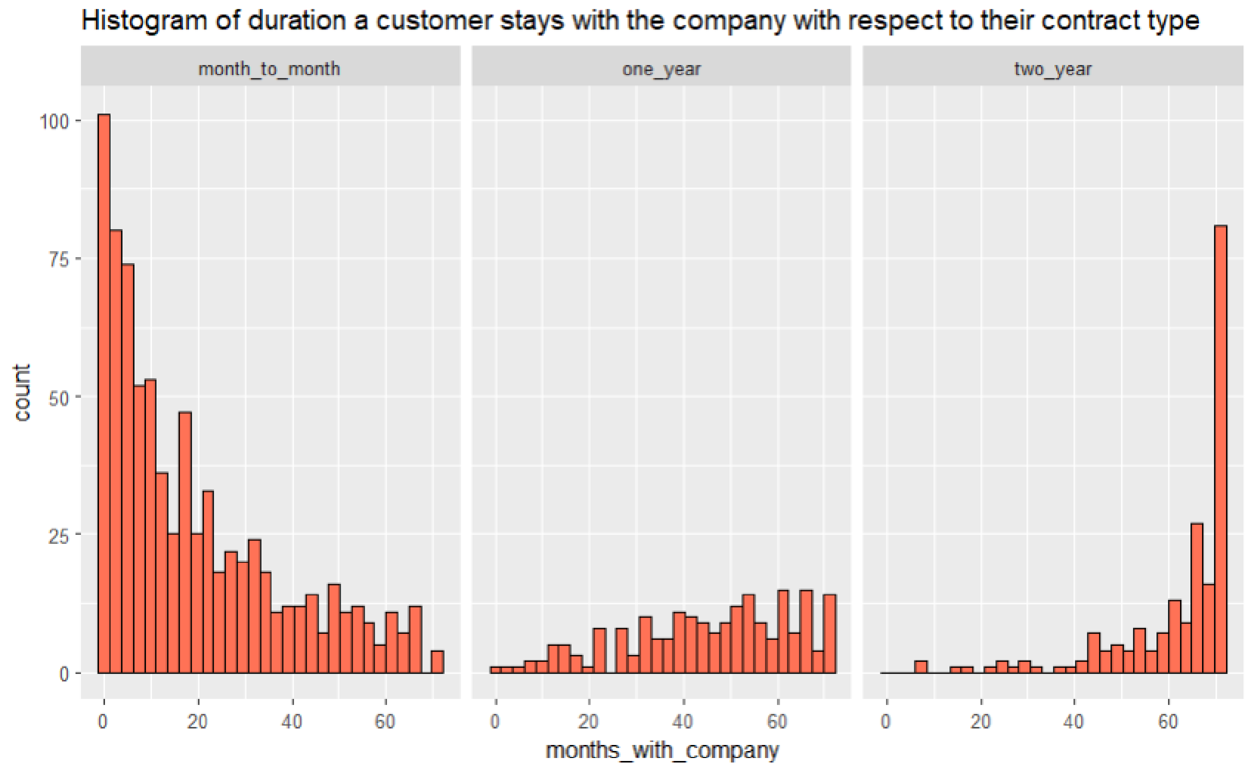
Relation between the Customer's device protection option and service cancellation

Question 5

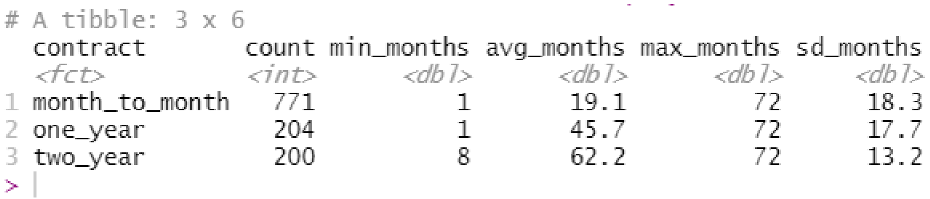
Question: Does the type of contract a person is registered with have an impact on them staying with the company for longer?

Answer:

It can be clearly illustrated from the below histograms that most of the customers with a month-to-month plan cancel their plans with the company after a very short period of time. A sharp contrast with this phenomenon can be observed with the customers who are enrolled in one year and two year plans, who tend to stay with the company for longer periods of times.



The same data can be seen from the below summary table, where the average number of months with the company of the customers enrolled in month to month plans is much lesser than that of the customers who are enrolled in one year and two year plans.



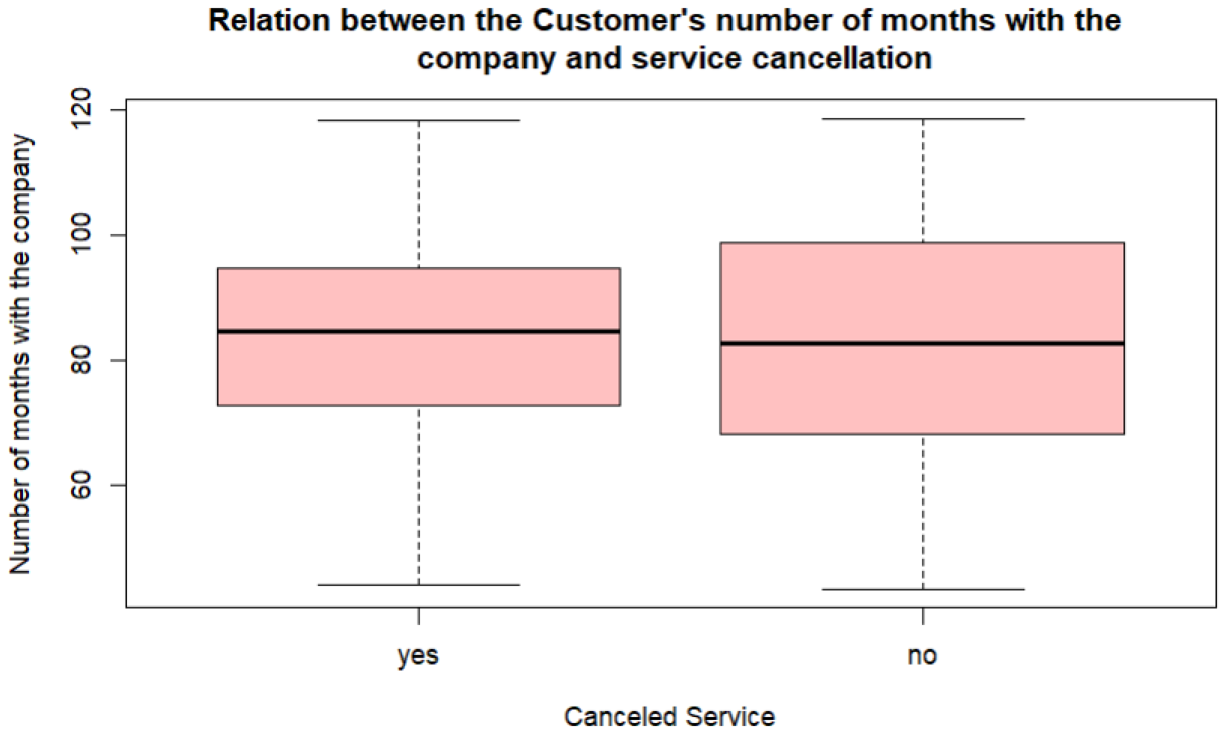
```
# A tibble: 3 x 6
  contract       count min_months avg_months max_months sd_months
  <fct>          <int>      <dbl>      <dbl>      <dbl>     <dbl>
1 month_to_month   771          1       19.1         72      18.3
2 one_year         204          1       45.7         72      17.7
3 two_year         200          8       62.2         72      13.2
> 
```

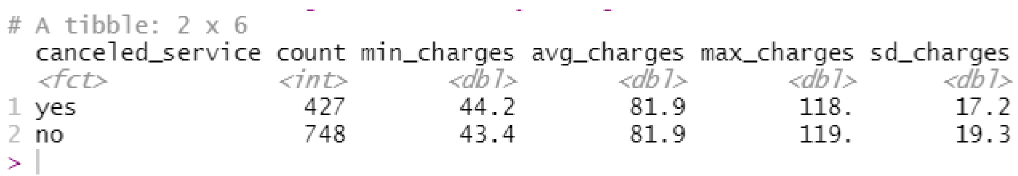
Question 6

Question: Is a person with lesser monthly charges more likely to not cancel their service?

Answer:

There is no significant difference between the customers who cancel their service and those who do not, with respect to their monthly charges. But it can be observed that the minimum monthly charges of the customers who do not cancel their service is slightly lesser than that of those who did, from the below boxplots and summary table.





```
# A tibble: 2 x 6
  canceled_service count min_charges avg_charges max_charges sd_charges
  <fct>            <int>       <dbl>       <dbl>       <dbl>      <dbl>
1 yes                427        44.2        81.9        118.       17.2
2 no                 748        43.4        81.9        119.       19.3
> 
```

Machine Learning

As the first step of performing machine learning on the telecom dataset, the data has been split into training and testing datasets in a 80-20 proportion respectively.

After the above step, as a part of feature engineering, a recipe has been developed for the dataset with the below steps:

- All the numeric predictors have been centered
- All the numeric predictors have been scaled

- All the numeric predictors have been normalized
- All the variables with high large absolute correlations with other variables have been removed
- All the nominal variables have been converted into numeric binary model items

After the above steps, the data is ready to be modeled using classification algorithms.

The below classification algorithms have been applied to the dataset to predict the response variable, canceled_service:

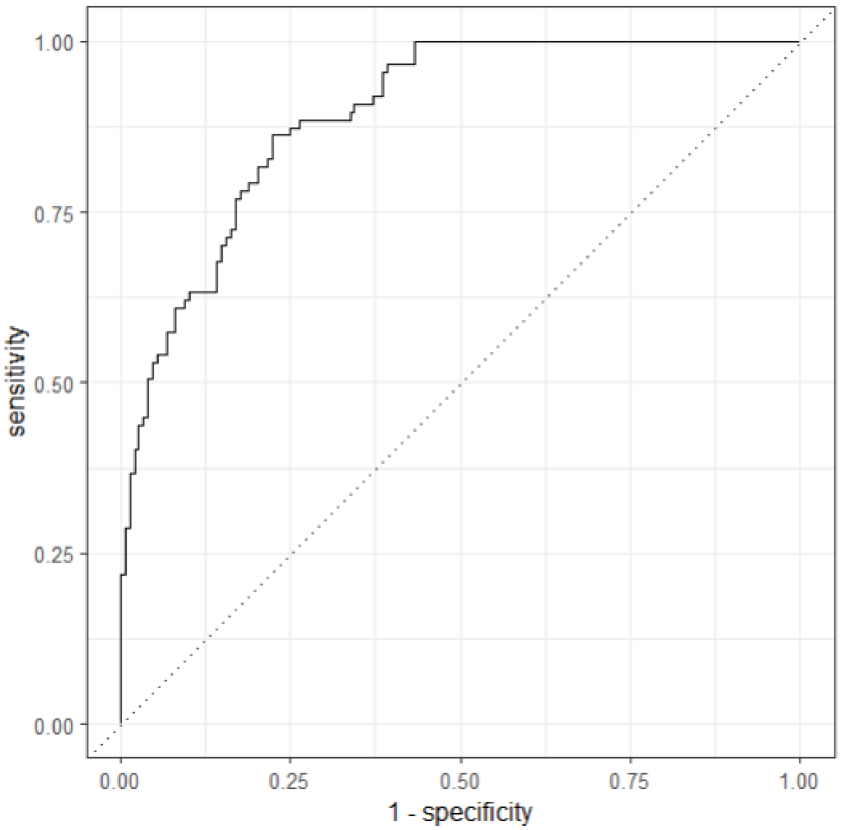
- Logistic Regression
- Random Forests
- Decision Trees
- K-Nearest Neighbors

For all the models, the model will be first applied on the training dataset and then validated on the testing dataset.

Logistic Regression

ROC Curve:

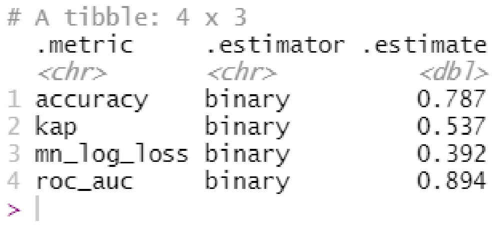
The ROC curve obtained for the Logistic Regression model can be viewed below:



Metrics of the model/Area under the curve:

The performance of the Logistic Regression model can be evaluated based on the metrics by testing the model on the testing dataset.

As can be seen from the below metrics table, the accuracy of the model is 78.7% and the area under the ROC curve for this model is 89.4.

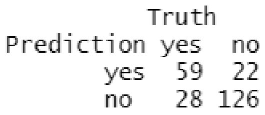


```
# A tibble: 4 x 3
  .metric     .estimator .estimate
  <chr>       <chr>          <dbl>
1 accuracy    binary         0.787
2 kap         binary         0.537
3 mn_log_loss binary         0.392
4 roc_auc     binary         0.894
> |
```

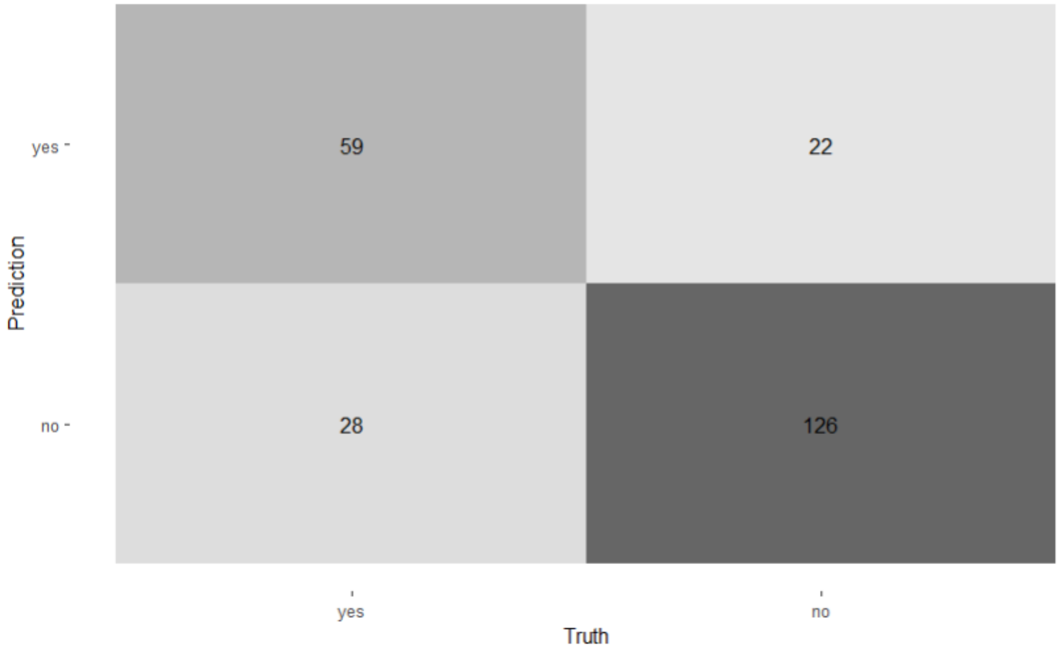
Confusion Matrix of the model:

The confusion matrix of the model can be seen below:

It can be seen that the model has accurately classified the canceled_service variable in the testing dataset for most of the observations.



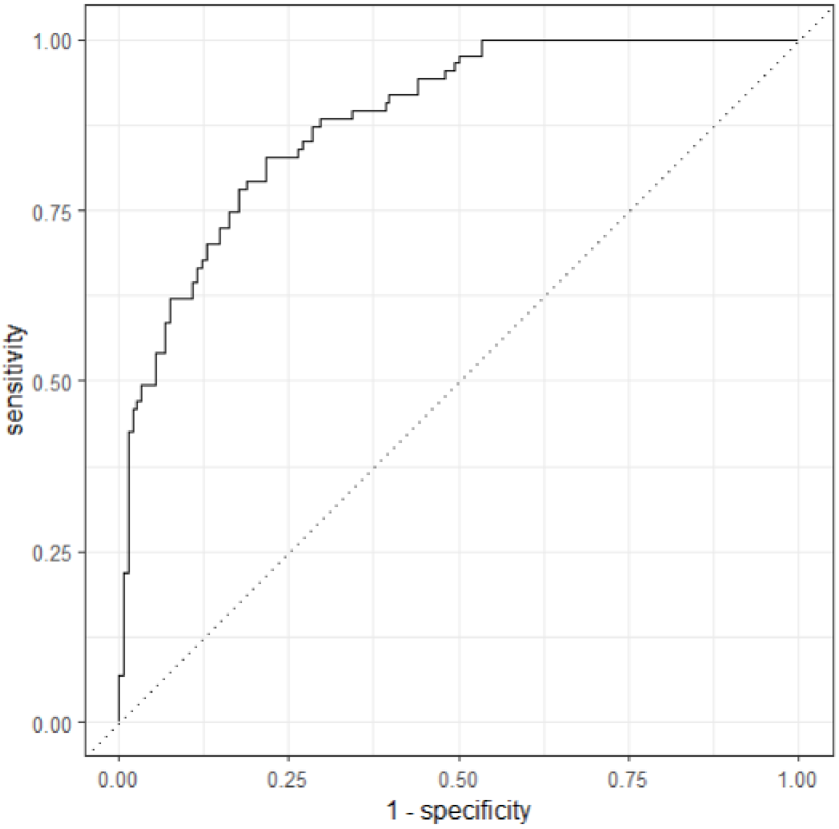
```
          Truth
Prediction yes  no
       yes  59  22
       no   28 126
```



Random Forests

ROC Curve:

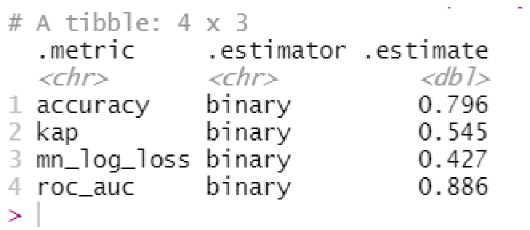
The ROC curve obtained for the Random Forests model can be viewed below:



Metrics of the model/Area under the curve:

The performance of the Random Forests model can be evaluated based on the metrics by testing the model on the testing dataset.

As can be seen from the below metrics table, the accuracy of the model is 79.6% and the area under the ROC curve for this model is 88.6.

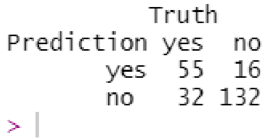


```
# A tibble: 4 x 3
  .metric     .estimator .estimate
  <chr>       <chr>          <dbl>
1 accuracy    binary         0.796
2 kap         binary         0.545
3 mn_log_loss binary         0.427
4 roc_auc     binary         0.886
> |
```

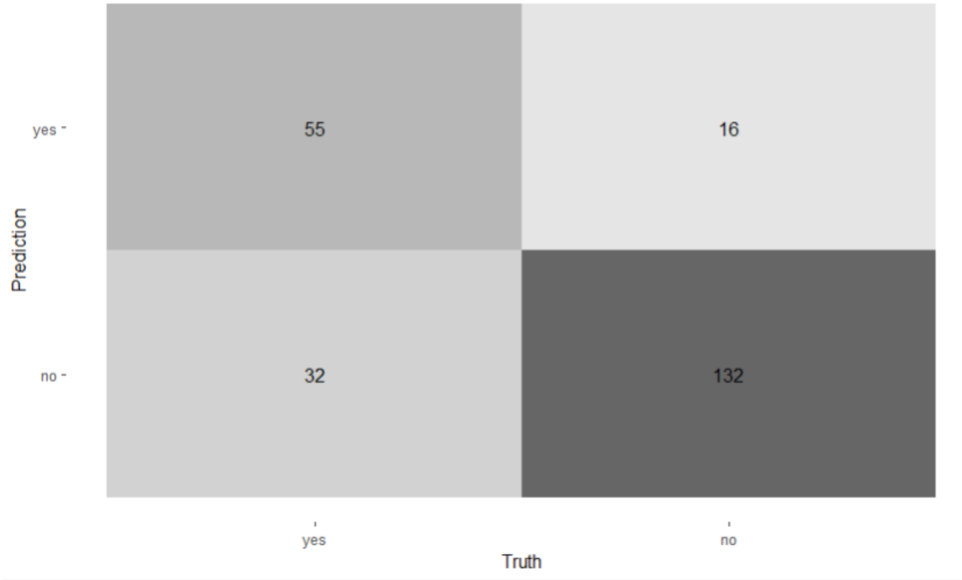
Confusion Matrix of the model:

The confusion matrix of the model can be seen below:

It can be seen that the model has accurately classified the canceled_service variable in the testing dataset for most of the observations.



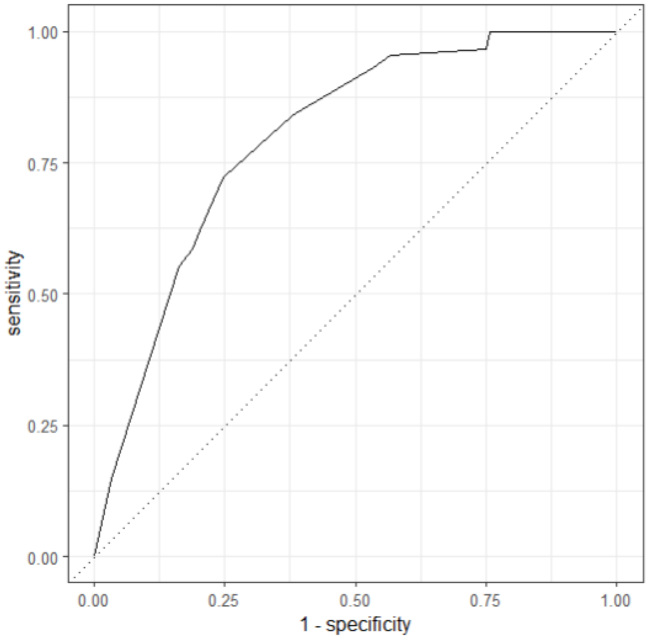
```
          Truth
Prediction yes  no
       yes  55  16
       no   32 132
> 
```



Decision Trees

ROC Curve:

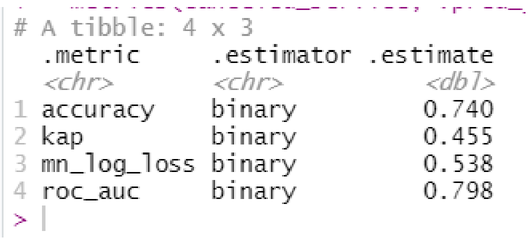
The ROC curve obtained for the Decision Trees model can be viewed below:



Metrics of the model/Area under the curve:

The performance of the Decision Trees model can be evaluated based on the metrics by testing the model on the testing dataset.

As can be seen from the below metrics table, the accuracy of the model is 74% and the area under the ROC curve for this model is 79.8.

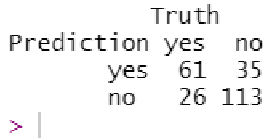


```
# A tibble: 4 x 3
  .metric     .estimator .estimate
  <chr>       <chr>          <dbl>
1 accuracy    binary         0.740
2 kap         binary         0.455
3 mn_log_loss binary         0.538
4 roc_auc     binary         0.798
> 
```

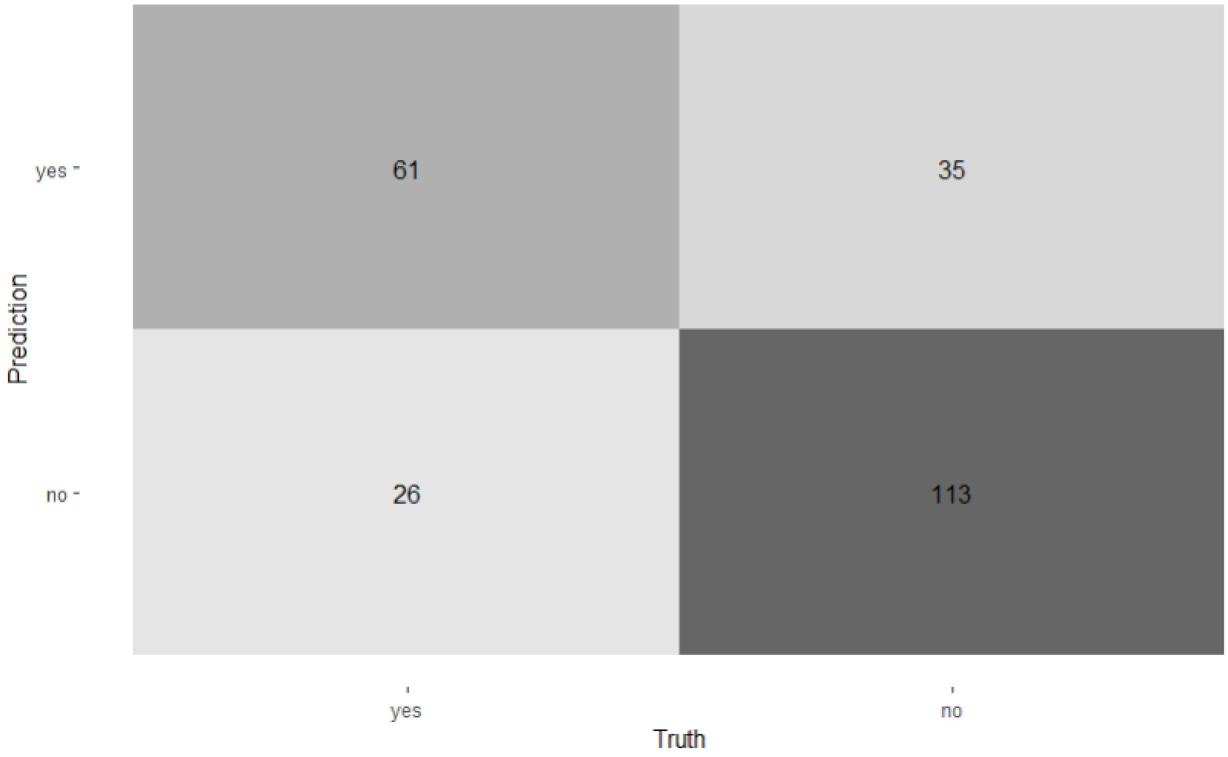
Confusion Matrix of the model:

The confusion matrix of the model can be seen below:

It can be seen that the model has accurately classified the canceled_service variable in the testing dataset for most of the observations.



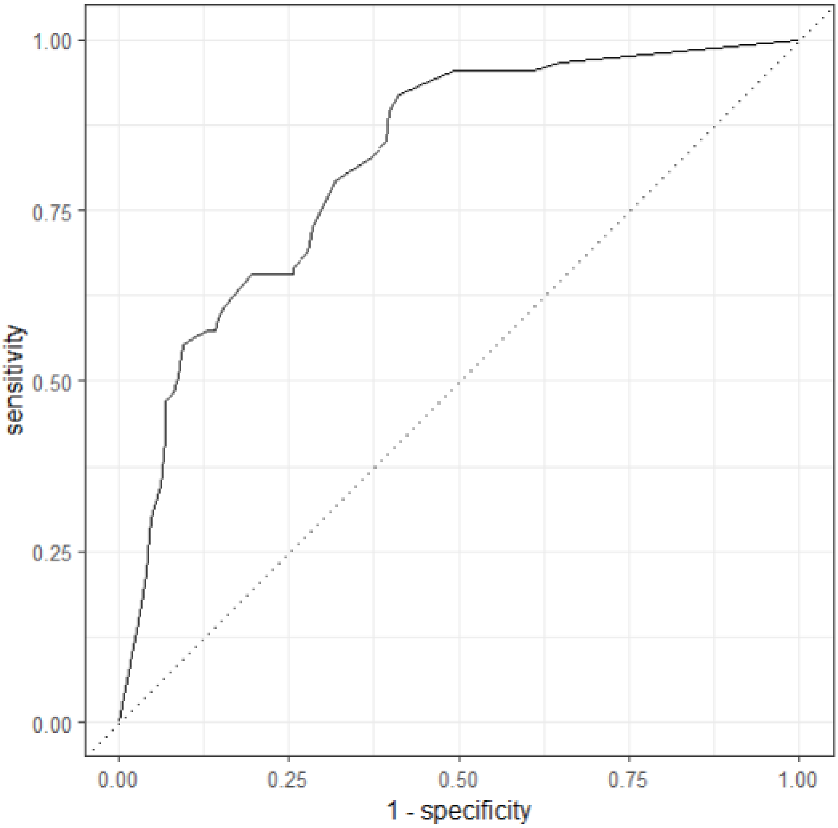
```
          Truth
Prediction yes  no
       yes  61  35
       no   26 113
> 
```



K-Nearest Neighbors

ROC Curve:

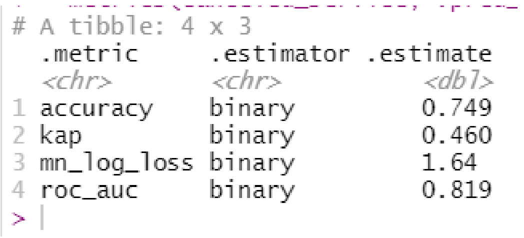
The ROC curve obtained for the K-Nearest Neighbors model can be viewed below:



Metrics of the model/Area under the curve:

The performance of the K-Nearest Neighbors model can be evaluated based on the metrics by testing the model on the testing dataset.

As can be seen from the below metrics table, the accuracy of the model is 74.9% and the area under the ROC curve for this model is 81.9.

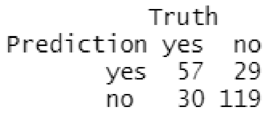


```
# A tibble: 4 x 3
  .metric     .estimator .estimate
  <chr>       <chr>          <dbl>
1 accuracy    binary         0.749
2 kap         binary         0.460
3 mn_log_loss binary         1.64 
4 roc_auc     binary         0.819
> |
```

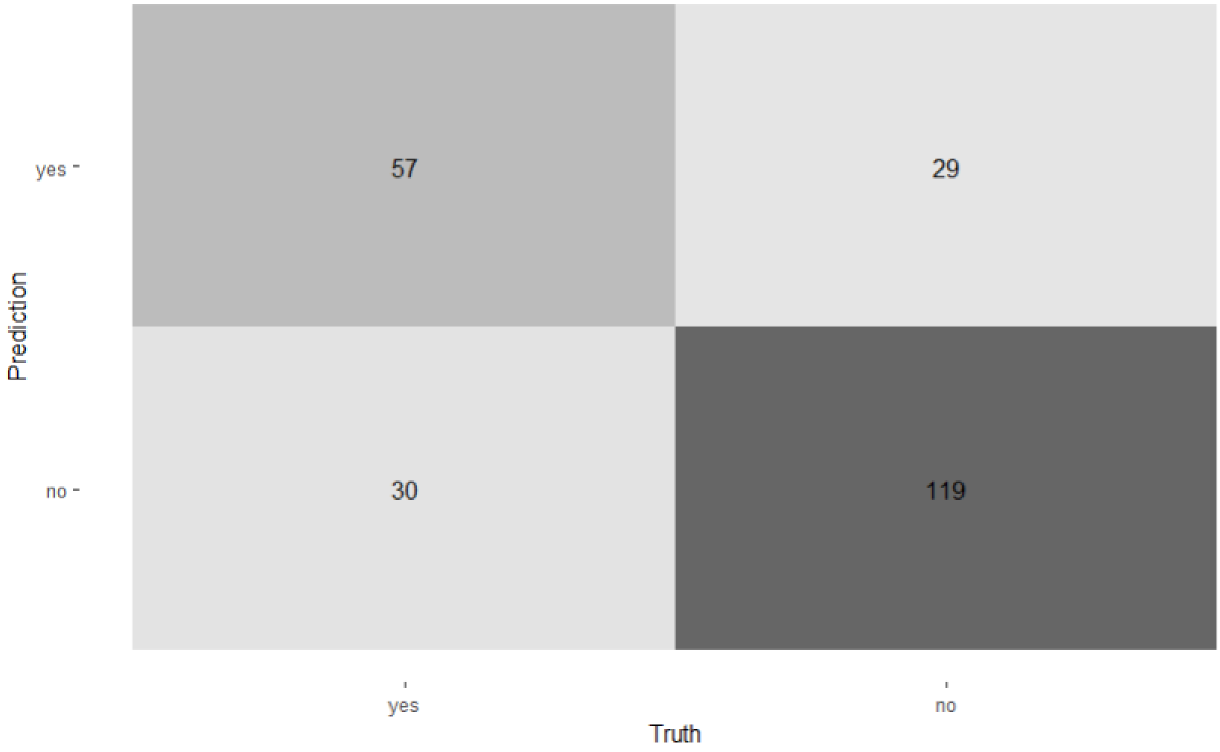
Confusion Matrix of the model:

The confusion matrix of the model can be seen below:

It can be seen that the model has accurately classified the canceled_service variable in the testing dataset for most of the observations.



```
          Truth
Prediction yes  no
       yes  57  29
       no   30 119
```



Highlights and Key Findings

Below are the key findings that have been observed in this exploratory data analysis:

1. The customers with more average monthly call minutes are more likely to cancel their service.

2. The customers with more average monthly international call minutes are more likely to not cancel their service and stay with the company.

3. Most of the customers who cancel their service do so in very short times of joining the company and the customers who do not cancel their service stay with the company for very long durations, significantly higher than the customers who have cancelled their service.

4. The customers who have digital internet connection are less likely to cancel their service as compared to the customers who are enrolled with a fiber optic internet connection.

5. The customers with online security option enabled are more likely to cancel their plan and stay with the company for lesser durations as compared to the customers who do not have this option enabled.

6. The customers with device protection option enabled are far less likely to cancel their plans and stay with the company for longer durations as compared to the customers who do not have this option enabled.

7. Most of the customers with a month-to-month plan cancel their plans with the company after a very short period of time.

8. The customers who are enrolled in one year and two-year plans tend to stay with the company for longer periods of times.

9. The monthly charges of a customer does not much impact the likelihood of them cancelling their service.

Comparing the models

The above models can be compared by their accuracies and areas under their ROC curves as below:

Model	**Accuracy**	**Area under ROC curve**
Logistic Regression	78.7%	89.4
Random Forests	79.6%	88.6
Decision Tree	74%	79.8
K-Nearest Neighbors	74.9%	81.9

The Area Under the Curve (AUC) is the measure of the ability of a model to differentiate between classes in a categorical variable. It is used as a summary of the ROC curve. The performance of a classifier model is directly proportional to its ROC AUC. The higher the AUC of a model, the more reliable it is at differentiating between the different classes of a categorical variable.

For this dataset, Random Forests model has the highest accuracy of all the models and Logistic Regression has the highest area under the curve of all the four models.

Both Logistic Regression and Random Forests are reliable models in order to predict the canceled_service variable in this dataset, based on their high accuracies and high areas under the ROC curves, as compared to other models.

Recommendations

Below are the key recommendations and suggestions that can be given to the Telecommunications Company in order for them to analyze and make necessary changes to reduce customer attrition rates:

1. The monthly prices that each customer has to pay has to be decided on a dynamic basis based on their average monthly call minutes. The customers who in general have more monthly call minutes could be charged slightly less or more incentives could be provided to them in order to encourage them to stay in their plans, making them less likely to cancel their plans and looking for other options.

2. More importance must be given to the customers who have newly enrolled with the company to ensure that their needs are being adequately met, in order to make them long-term customers who are less likely to cancel their plans, according to the results obtained above. The monthly plan prices could be made lesser for a fixed number of months after a customer has joined the company, in order to make them stay with the company longer, which would eventually be profitable to the company in the long run.
3. Encourage more customers to take up digital internet connection as opposed to fiber optic internet connection. This could be done by making the digital internet price affordable. Also look into the factors as to why the customers with fiber optic internet connection are cancelling their service and take remedial measures to rectify these factors.
4. Make the online security option more affordable, in order to encourage the customers who have this option enabled to stay on their current plan. The same pricing strategy that is being used for the device protection option could be used for the online security options as well, as the customers with device security option enabled are far less likely to cancel their services.
5. Encourage more customers to enroll in the one-year and two-year plans as opposed to in the monthly plans, in order to make them stay with the company longer. This can be done by setting the prices of the plans in such a way that a customer will pay less for a one-year plan as opposed to twelve month-to-month plans.