Project Final Report

Group 11

Caleb Neale (can4ku), Rachel Lee (rl6ug), Chenlin Liu (cl2trg), Chethan Shivaram (cbs9md)

1 Executive Summary

The questions of interest our group is trying to answer:

1. How can we best predict the future amount of data a customer will use?
   Response variable: `DataUsage`

2. How can we best predict whether a customer will churn (cancel service) or not?
   Response variable: `Churn`
   Categorization:
   - 0: the customer did not cancel the service
   - 1: the customer canceled the service

Answer provided by analysis in Milestone 5:

1. To predict a customer's future data usage, it's essential to look into whether they have a data plan, their average monthly bill, and their average monthly charge. The most important of these features is the existence of a data plan. With a data plan, a customer will use more data. Average monthly bill and average roaming minutes are both positively associated with their data usage. The higher the average monthly bill and average roaming minutes, the higher the data usage.
2. To predict whether a customer will churn, the most helpful predictors to know are day minutes, number of calls to customer service, the customer's monthly charge, and whether they've renewed their contract in the past year.

The following are variables used in our dataset:

- `Churn` (categorical): 1 if customer canceled service, 0 if not
- `AccountWeeks` (numerical): number of weeks customer has had an active account
- `ContractRenewal` (categorical): 1 if customer recently renewed contract, 0 if not
- `DataPlan` (categorical): 1 if the customer has a data plan, 0 if not
- `DataUsage` (numerical): gigabytes of monthly data usage
- `CustServCalls` (numerical): number of calls into customer service
- `DayMins` (numerical): average daytime minutes per month
- `DayCalls` (numerical): average number of daytime calls
- `MonthlyCharge` (numerical): average monthly bill
- `OverageFee` (numerical): largest overage fee in the last 12 months
- `RoamMins` (numerical): average number of roaming minutes.

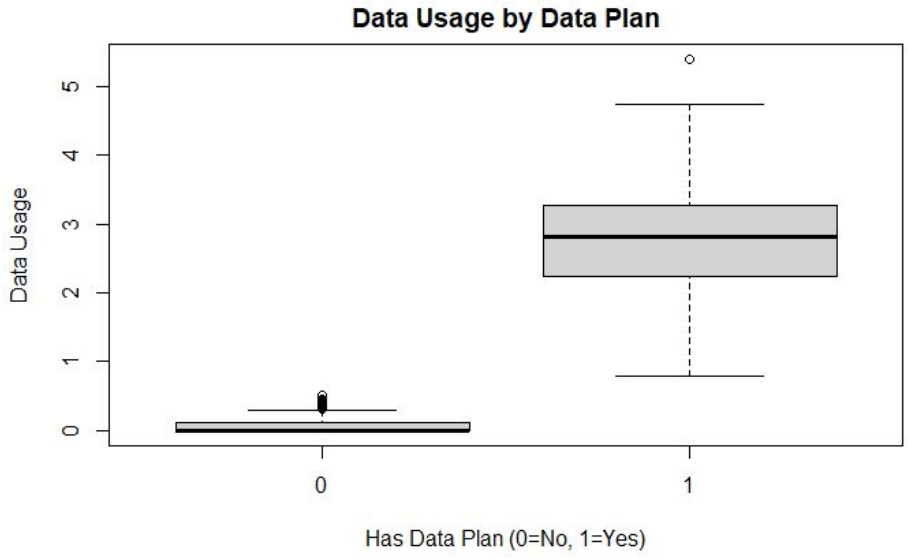
2 Data Cleaning & Processing

After reading in the data, we converted the categorical variables in the data set from dummy coding to factors. No further data cleaning is required because our data set did not contain any missing values, contained numerical and integer values, and were in reasonable ranges..

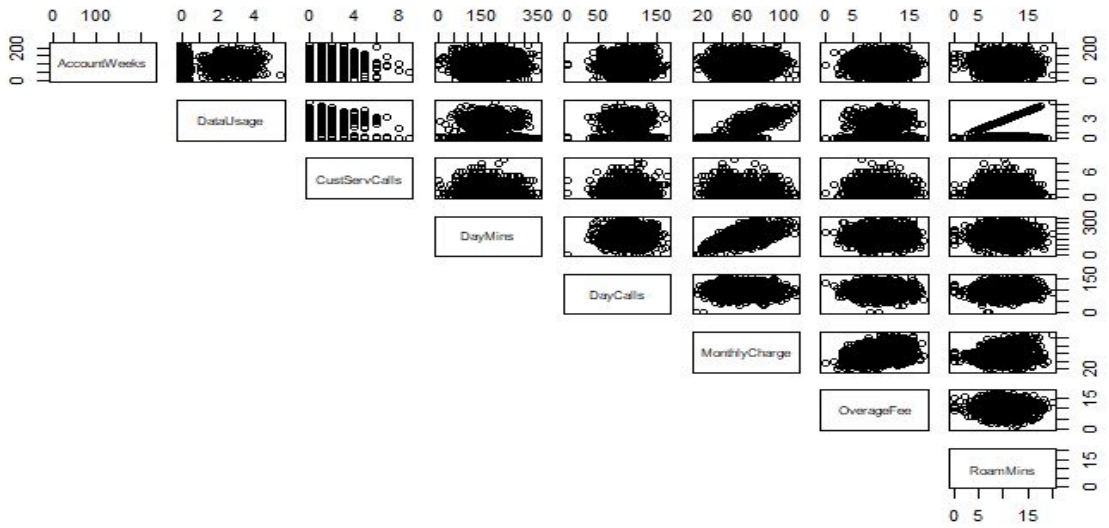
Prior to completing the regression model and the regression tree, we removed observations of zero data usage. This allows our data to meet regression assumptions, and also make sense in the context of our question, predicting future data usage. Subsetting the data is not needed for the classification section.

3 Regression Question

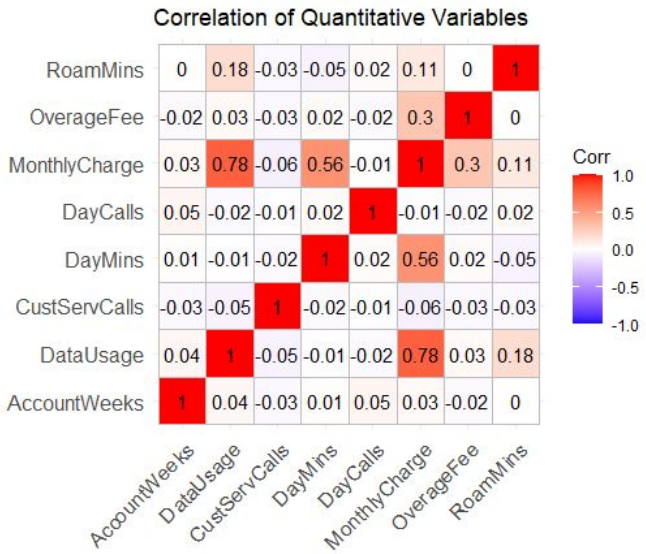
3.1 Exploratory Data Analysis



This chart illustrates the (expected) idea that if a customer has a data plan they use more data. It appears that the minimum data usage by customers with a data plan is above the maximum for those without. It is interesting that there are customers without a data plan that use data at all; this suggests that a data plan is not necessary for using data, but is related to a higher amount of data usage.



This pairplot doesn't immediately show any trends which may be useful in answering our regression question. There is an interesting relationship between data usage and roaming minutes, with a portion of the data appearing almost perfectly linear and a portion showing no relationship. When looking at the data usage row, there does seem to be a pattern of points clustered towards the bottom of the graph, suggesting there is a subset of customers which do not use any data regardless of any attributes.

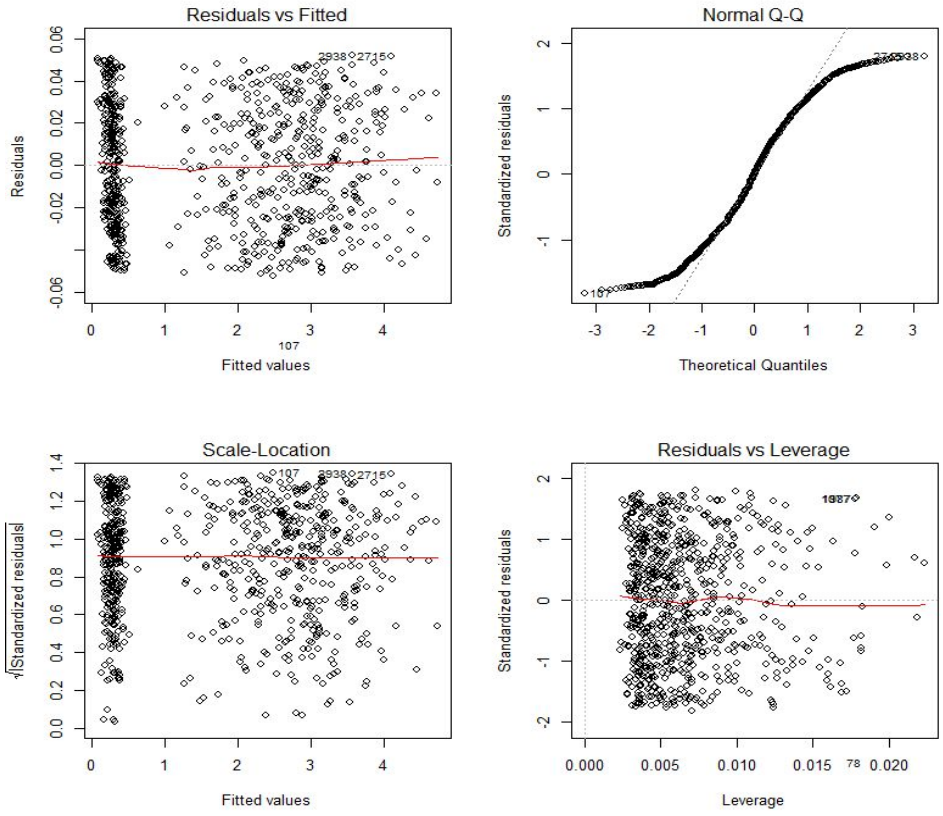


This correlation matrix shows that MonthlyCharge is by far the most strongly linearly correlated variable with data usage. This leads us to believe that it will be a significant predictor of data usage when making the model. There are fairly low correlations between other predictors and data usage.

3.2 Regression Model

To make our regression model meet the regression assumptions, we ran the regression on customers who have used data (DataUsage > 0). The best regression model we got includes DataPlan, DayMins, MonthlyCharge, and OverageFee as the predictors.

Diagnostic plots of the regression model:



The red line in the first plot for this regression shows that the average of the residuals is about zero, and the residuals are not correlated because they are randomly distributed according to the first plot. The s-shaped distribution of points in the QQ plot indicates that the residual distribution is light-tailed, so the residuals are not normally distributed. Because the vertical spread of the third plot is constant, it indicates that the residuals have constant variance. The last plot shows that there are no influential outliers in the data. Although the problem of non-normality of the residuals is not solved, other assumptions about residuals (mean of zero, no autocorrelation, constant variance, and no influential outliers) are not violated. The normality condition is not as important as other assumptions as long as we do not perform confidence intervals based on this regression model.

Summary Output for the Linear Regression Model:

```
Call:
lm(formula = DataUsage ~ DataPlan + DayMins + MonthlyCharge +
    OverageFee, data = datatrain)

Residuals:
      Min        1Q    Median        3Q       Max
-0.052334 -0.024769  0.000125  0.024800  0.051948

Coefficients:
               Estimate Std. Error  t value Pr(>|t|)
(Intercept)   -0.0545957  0.0056746   -9.621   <2e-16 ***
DataPlanTrue   0.0041044  0.0051032    0.804    0.421
DayMins       -0.0169514  0.0000372 -455.643   <2e-16 ***
MonthlyCharge  0.0998801  0.0001863  536.259   <2e-16 ***
OverageFee    -0.1697558  0.0005247 -323.519   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02894 on 755 degrees of freedom
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.9995
F-statistic: 4.154e+05 on 4 and 755 DF,  p-value: < 2.2e-16
```

DataUsage = -.0546 + .0041*DataPlan - .0170*DayMins + .0999*MonthlyCharge - .1698*OverageFee

The regression model shows that a customer is predicted to use more data if they have a data plan or a higher average monthly bill. If a customer has more average daytime minutes per month or higher overage fee, the customer will be predicted to use less data. Since our main point of interest is predicting future data use of our customer base, an R^2 value of .9995 indicates an extremely competent model.

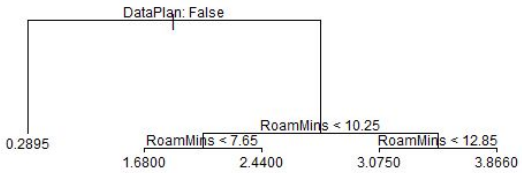
3.3 Regression Tree

R Output from 5-Fold Cross Validation:

```
$size
[1] 5 4 3 2 1

$dev
[1]   28.97537   85.90101   89.79704  241.83803 1397.62713
```

Based on the 5-fold Cross Validation, a tree with 5 terminal nodes should be used since it has the smallest deviance, so the pruned tree is the same as the tree from recursive binary splitting. Thus, we chose to present the tree built with recursive binary splitting.

Graphical Output for the Regression Tree Using Recursive Binary Splitting:

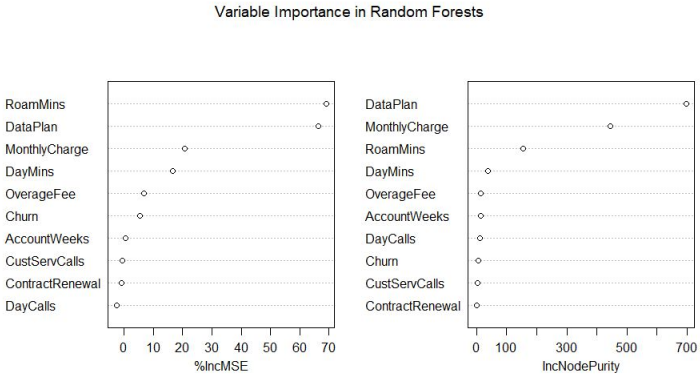


The resulting tree has five terminal nodes, and the predictors used in the tree were `DataPlan` (the existence of a data plan) and `RoamMins` (the average number of roaming minutes).

In answering the question about predicting the future amount of data a customer will use, the existence of a data plan (`DataPlan`) is the most important factor. Among those who do have a data plan, average roaming minutes (`RoamMins`) is the next (and, only other) important factor. Among customers who have a data plan, roaming minutes is positively associated with data usage, the more roaming minutes, the more data usage.

R Output from the importance() and the varImpPlot() functions with Random Forests:

	%IncMSE	IncNodePurity
Churn	5.51	7.02
AccountWeeks	0.68	13.03
ContractRenewal	-0.79	0.75
DataPlan	66.43	696.54
CustServCalls	-0.51	4.00
DayMins	16.58	37.90
DayCalls	-2.51	11.01
MonthlyCharge	20.65	443.34
OverageFee	6.76	15.01
RoamMins	69.00	155.44



Based on the increase in node purity, `DataPlan`, `MonthlyCharge`, and `RoamMins` are the top three important factors in predicting the data usage of a customer.

To predict a customer's future data usage, it's essential to look into whether they have a data plan, their average monthly bill, and their average number of roaming minutes. The most important of these features is the existence of a data plan. With a data plan, a customer will use more data. Average monthly bill and average roaming minutes are both positively associated

with their data usage. The higher the average monthly bill and average roaming minutes, the higher the data usage.

3.4 Summary of Findings

The test MSE of the linear regression model is 0.0007721683.

The test MSE of the trees is 0.04251291.

The test MSE of random forests is 0.0227786.

A comparison of the test MSE with linear regression model, regression tree (built with recursive binary splitting or pruning), and random forests, and comment on these values.

We can see that the test MSE of the linear regression model is almost 100 times less than that of the tree-based models. Because of this the linear regression model is by far the best predictive model of the three. Between the two tree-based models, the RF test MSE was around 2% while that of the RBS model was around 4%. This was to be expected as RF decreases the variance of the errors of the predictions.

Commentary on Similarities and Differences of Findings From 3.1 to 3.3

`DayMins, MonthlyCharge, DataPlan,` and `OverageFee`	Lin. Reg.
`DataPlan` and `RoamMins`	RBS
`DataPlan, MonthlyCharge,` and `RoamMins`	RF

The existence of a data plan, DataPlan, was an important factor in each of our methods. This makes sense in the context of our question, how can we best predict the future amount of data a customer will use? With a data plan, a customer will use more data. In addition, MonthlyCharge was an important factor in both our linear regression model and our tree produced using random forests. Typically, more data usage results in a higher monthly charge. RoamMins, the amount of roaming minutes, was a significant factor in both of our regression trees. With more roaming minutes, typically, there is a higher data usage, and also a higher monthly charge. Although our methods did not produce the same factors for each, there is overlap and correlation within the factors.

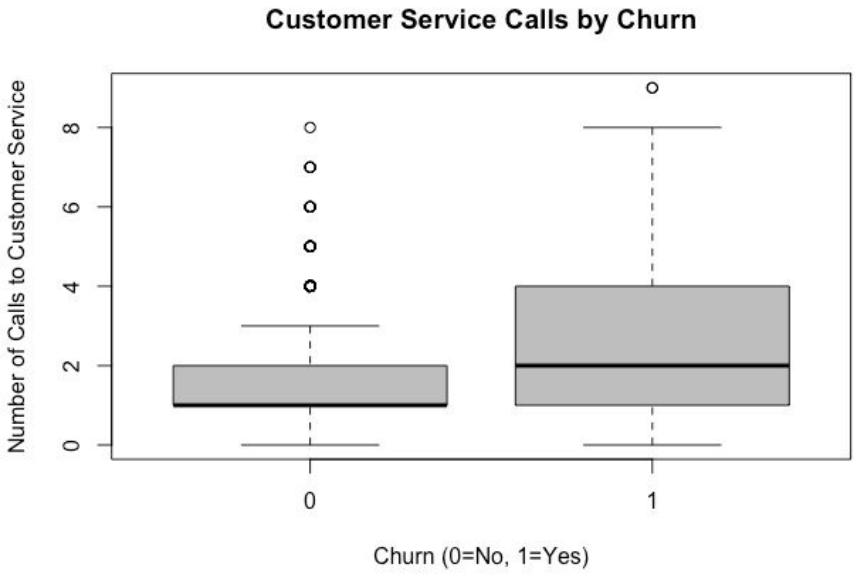
Commentary in Which Model was Better in Answering Our Question of Interest

All of our regression models had low test MSE's at .07%, 4.3%, and 2.2% for our linear regression model, regression tree using recursive binary splitting, and our tree using random forests, respectively. Our linear regression model, containing the predictors DayMins,

MonthlyCharge, DataPlan, and OverageFee, had the best overall accuracy on the test data at roughly .07%.

4 Classification Question

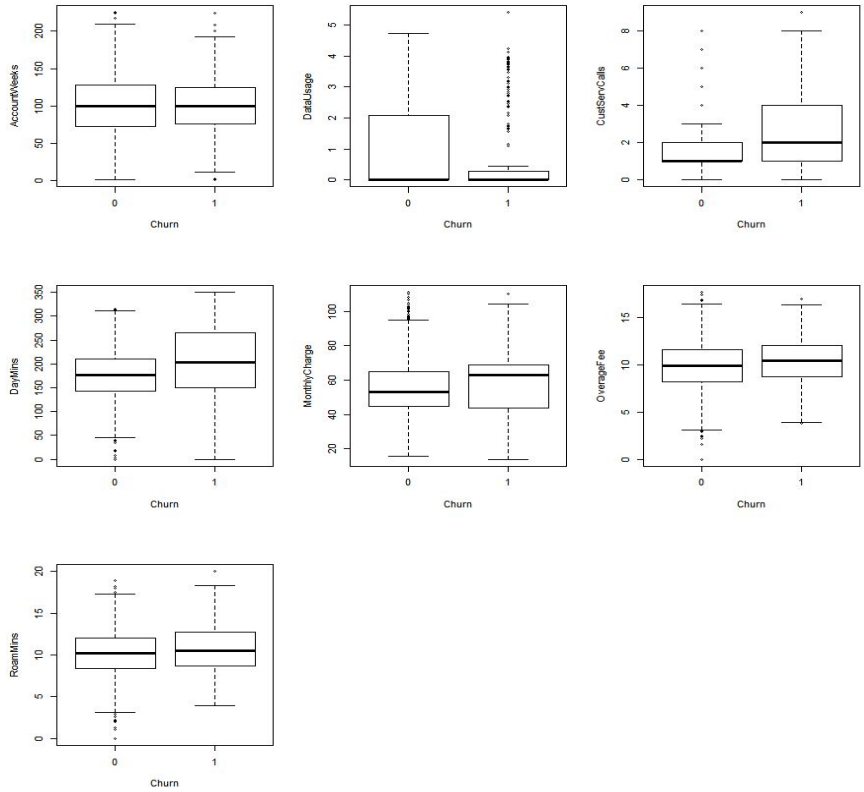
4.1 Exploratory Data Analysis



From this graphic, it is obvious that customers who canceled their service and those who did not have different distributions of the number of calls to customer service. Most of the customers who did not churn called the customer service less than three times while half of those who churned called the customer service more than twice. Therefore, it is likely that the number of calls to customer service from a customer helps us predict their probability of churn. We will perform more explorations about the dataset to determine the models and predictors for the response variables in the future.

We did not complete any data cleaning in this section of our analysis as our data did not contain any missing values, contained numerical and integer values, and were in reasonable ranges. Prior to analysis, we set seed using set.seed(69420).

Boxplots: Quantitative Predictors vs. Churn



Since our response variable is binary categorical and most of our predictors are numerical, we felt that multiple boxplots would be the best way to view the data.

Applying the above logic to the other plots, we think it is likely that CustServCalls, MonthlyCharge, and DayMins will drive our model's prediction of Churn.

4.2 LDA and Logistic Regression

We chose to not present the LDA model because we tried different combinations of the predictors, but none of them satisfy the multivariate normality assumption for LDA.

The most satisfactory logistic regression model we got uses all variables in the data set as the predictors for predicting whether a customer is going to cancel their service. We attempted to improve the model by removing insignificant predictors from the initial model. It showed improvement from the initial model, but we figured the reduced model, with only 2 predictors, would not be as accurate. Ultimately, we proceeded to complete analysis on the model containing all predictors.

Summary Output for Logistic Regression Model:

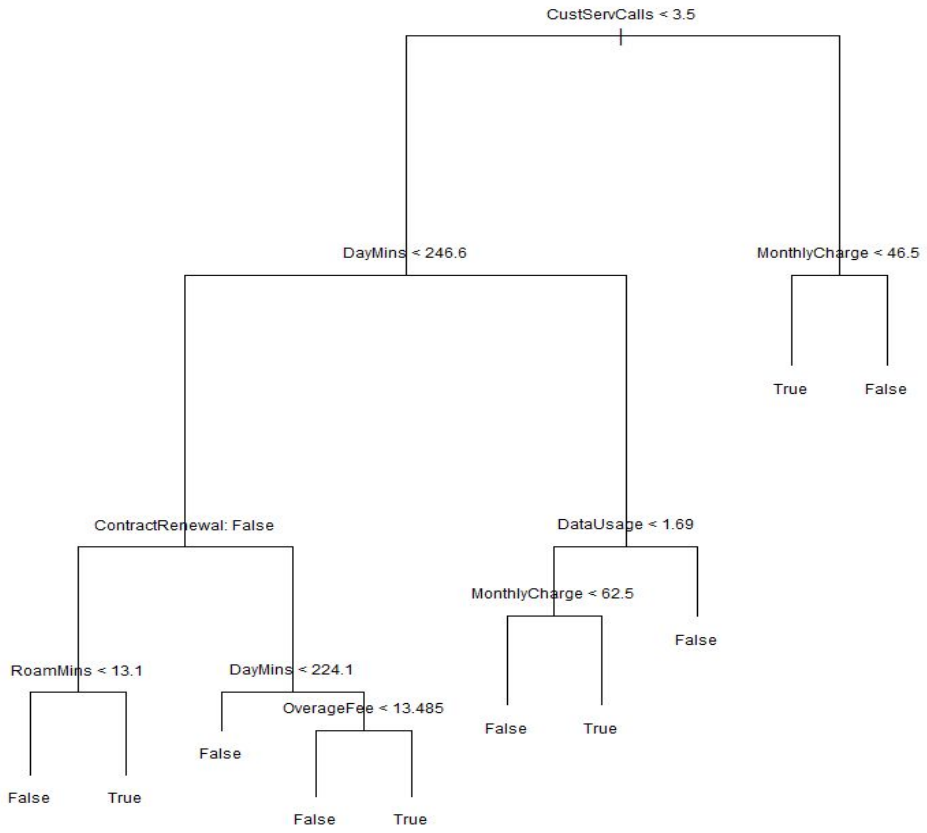
```
Call:
glm(formula = Churn ~ AccountWeeks + DataUsage + CustServCalls +
    DayMins + DayCalls + MonthlyCharge + OverageFee + RoamMins +
    DataPlan + ContractRenewal, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9025  -0.5211  -0.3509  -0.2003   3.0340

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.695e+00  7.551e-01  -7.542 4.61e-14 ***
AccountWeeks     6.940e-04  1.960e-03   0.354  0.72329
DataUsage        1.027e+00  2.689e+00   0.382  0.70253
CustServCalls    5.894e-01  5.538e-02  10.643  < 2e-16 ***
DayMins          3.077e-02  4.549e-02   0.676  0.49880
DayCalls        -1.427e-05  3.753e-03  -0.004  0.99697
MonthlyCharge   -1.028e-01  2.674e-01  -0.384  0.70064
OverageFee       2.873e-01  4.565e-01   0.629  0.52908
RoamMins         8.444e-02  3.019e-02   2.797  0.00516 **
DataPlan        -1.078e+00  7.520e-01  -1.434  0.15162
ContractRenewal -1.711e+00  2.079e-01  -8.231  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1419.2  on 1665  degrees of freedom
Residual deviance: 1124.2  on 1655  degrees of freedom
AIC: 1146.2

Number of Fisher Scoring iterations: 6
```

In this summary output for Logistic Regression with all predictors, we see the significant predictors CustServCalls, RoamMins, and ContractRenewal. As shown by the boxplots in the EDA, it was expected that the number of calls into customer service would have an impact on whether a customer churned or not. It’s reasonable to understand that if a customer has to make many calls to customer service, they are likely experiencing issues with service. We also see ContractRenewal as a significant predictor. It is expected that whether a customer recently renewed a contract would have an impact on whether they choose to cancel the service. If a customer recently renewed their contract, they are likely to be satisfied with their service.

4.3 Classification Trees

The tree produced from recursive binary splitting is the same as the tree produced from pruning. Thus, we chose to present the tree built with recursive binary splitting.

Graphical Output for the Classification Tree Using Recursive Binary Splitting:

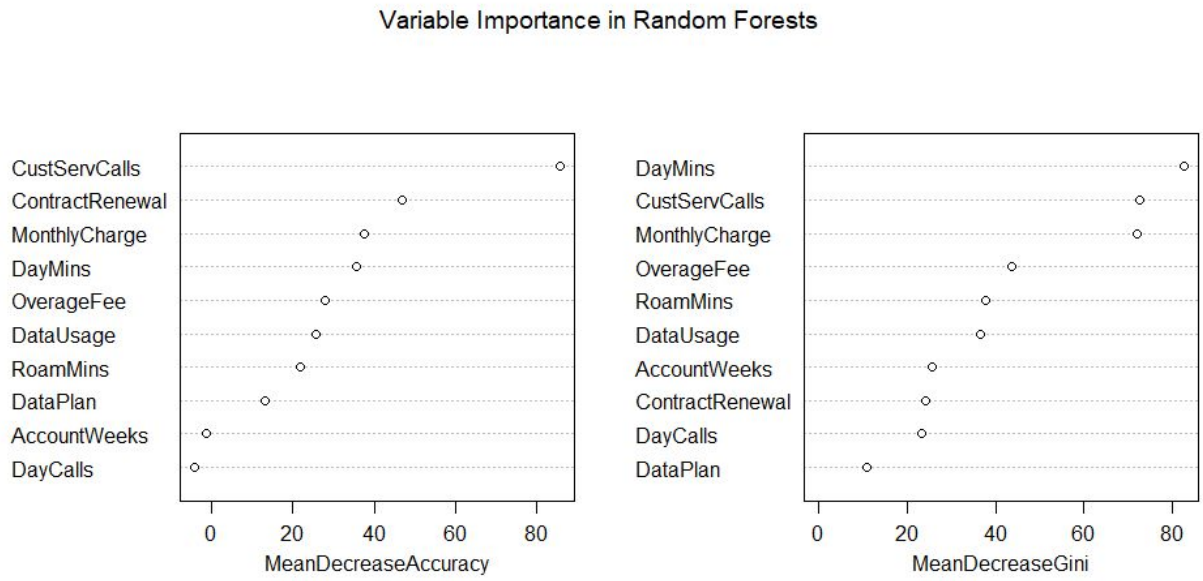


The resulting tree has ten terminal nodes, and the predictors used in the tree were `CustServCalls, DayMins, MonthlyCharge, ContractRenewal, DataUsage, RoamMins,` and `OverageFee.`

How Tree Classifies Customer Churn and Relation of Predictors to Response.

This tree predicts whether a customer will cancel service based on the relevant predictors listed above. We can see that when a customer makes greater than or equal to 3.5 calls to customer service per month, the only variable needed to predict churn is how much that customer pays per month (Over $46.5 means stayed with service under $46.5 means cancelled service), whereas if the customer's number of customer service calls is less than 3.5, the prediction of whether or not they churn is based on a multitude of factors, including ContractRenewal, DataUsage, MonthlyCharge, RoamMins, DayMins, and OverageFee. CustServCalls is the most important predictor of customer churn, followed by DayMins and MonthlyCharge as is shown by the order of splits in the tree.

R Output from the varImpPlot() Function with Random Forests:



`DayMins`, `CustServCalls`, `MonthlyCharge`, and `ContractRenewal` are the most important factors in predicting customer churn depending on whether we evaluate based on MeanDecreaseAccuracy or MeanDecreaseGini.

Between random forests, RBS, and pruning, we found that a tree fitted with random forests gave us the lowest overall error rate. However, we found that in all cases we could lower the FNR at minimal cost to the FPR by lowering the threshold. Since we are primarily interested in predicting whether a customer will cancel service, we are not as concerned about the FPR. Some challenges faced were determining the practicality of our predictions, interpreting false positives and false negatives in context, and understanding how the output of the model may be contextualized in context of how the company may prioritize its needs.

4.4 Summary of Findings

Confusion Matrix of Logistic Regression (Threshold = 0.5)

	predicted	
actual	FALSE	TRUE
False	1399	38
True	173	57

Overall Error Rate = .1266

False Positive Rate = .0264

False Negative Rate = .7522

Discuss if the threshold needs to be adjusted

Yes. With a threshold of 0.5 the false negative rate is higher than the true positive rate. A false negative, in context, means predicting that a customer won't churn when they actually will. Since we are presumably interested in maximizing the company, making false negative predictions will cause the loss of an opportunity to prevent customer churn, and thus should be minimized. As a false positive will simply result in effort to keep a customer, there is less of an opportunity for revenue loss and thus false positives are less detrimental to the company. As such, the threshold should be lowered to reduce false negatives.

Confusion Matrix of Classification Tree (Threshold = 0.5)

	predicted	
actual	FALSE	TRUE
False	1414	23
True	111	119

Overall Error Rate = .0789

False Positive Rate = .0160

False Negative Rate = .4826

Discuss if the threshold needs to be adjusted

Yes. With a threshold of 0.5 the false negative rate is roughly equal to the true positive rate. For the same reasoning provided in the previous analysis, the threshold should be lowered to reduce the false negative rate.

Confusion Matrix of Random Forest (Threshold = 0.5)

	predicted	
actual	FALSE	TRUE
False	1414	23
True	91	139

Overall Error Rate = .0684

False Positive Rate = .0160

False Negative Rate = .3957

Discuss if the threshold needs to be adjusted

Though the false negative rate with random forests is lower than in the logistic regression and the classification tree with recursive binary splitting, it is still higher than we would like. We think a lower threshold would allow us to better answer our question of interest.

Confusion Matrix of Random Forest (Threshold = 0.2)

	predicted	
actual	FALSE	TRUE
False	1306	131
True	54	176

Overall Error Rate = .1110

False Positive Rate = .0912

False Negative Rate = .2348

After lowering the threshold from 0.5 to 0.2, we get a lower false negative rate but higher overall error and false positive rate. According to our goal described above, we believe that we achieve a better prediction by lowering the threshold.

Commentary on Similarities and Differences of Findings From 4.1 to 4.3

`CustServCalls`, `RoamMins`, and `ContractRenewal`	Log. Reg.
`CustServCalls`, `Daymins`, and `Monthly Charge`	RBS
`DayMins`, `CustServCalls`, `MonthlyCharge`, and `ContractRenewal`	RF

The number of calls to customer service a customer made was a significant predictor in every classification method we performed. This was consistent with our hypothesis from the EDA that CustServCalls would be a driver of our predictions. In addition, MonthlyCharge (which measures how much a customer is charged per month) and DayMins (which measures how many minutes a customer spends on phone calls each day) were both significant predictors in the RBS and RF models, while ContractRenewal (which measures whether or not a customer recently renewed his or her contract) was significant in the RF and logistic regression models. We can see that although the decision algorithm for each classification method is different, they produce similar results.

Commentary in Which Model was Better in Answering Our Question of Interest

Although logistic regression resulted in similar predictors to the tree-based methods, the error rates of the predictions were much different. The overall error rate for LR was around 13% while the error rate for the tree-based methods was around 7-8%, almost half that of the LR. Additionally the FNR, which we are the most interested in minimizing in order to answer our question of interest, was over 75% in LR, compared to 40% in RF and 48% in the classification tree with RBS. Because we are trying to predict which customers WILL Churn, and the FNR measures the proportion of customers that were predicted to not Churn when in reality they did, we decided that the RF model, which has the lowest FNR, best answered our question of interest.

5 Further Work

If given more time on this project, our group could have explored more questions of interest. In addition, unsupervised learning could be implemented to further explore our data set. Lastly, it was somewhat difficult to contextualize our results given that none of us have a business background or know a lot about the company. More research could have been done on this front to make our conclusions more thorough.