1. **Business Case:** High customer churn rates can be costly for telecommunications companies because it takes more resources and money to acquire new customers than to retain existing ones. In the telecommunications industry, the average customer churn rate is about 22% (Liibert, 2023). It costs five times as much to attract a new customer than to retain an existing one and that a 5% increase in customer retention can lead to a 25% to 95% increase in profits (Bernazzani, 2022). By identifying the factors that contribute to customer churn, Telco can improve their services, products and customer support to better meet their customers' needs and preferences which can in turn lead to customer satisfaction and loyalty, higher retention rates and higher revenue.

2. **Business Question:** What are the main reasons that customers leave Telco, and what steps can the company take to improve their services and retain these customers?

3. **Analytics Question:** How do factors like customer gender, senior citizen status, dependents and partner status, tenure, phone service, multiple lines, internet service, online security, online backup, device protection, tech support, TV and movie streaming services, contract type, payment and billing options, monthly and total charges, influence the likelihood of customer churn at Telco?

   3.1 Outcome variable of interest: The outcome variable of interest is customer churn, which is a binary variable (Yes/No). The analytics question we examine is classification which aims to explore which factors are most important in determining the likelihood of customer churn.

   3.2 Main Predictors: Our key predictors include "tenure" (numeric), "MonthyCharges" (numeric), "TotalCharges" (numeric), "OnlineSecurity" (binary), "InternetService" (categorical), "Contract" (categorical), "TechSupport" (binary) and "DeviceProtection" (binary).
   Our analysis explores classification parametric (Logistic Regression and Ridge Regression) and non-parametric (Random Forest) modeling methods. Our analytics goal is to achieve both accuracy and interpretability of the models, in order to make informed decisions based on the insights gained from the analysis that will help reduce customer churn.

4. **Dataset Description:** "Telco Customer Churn" dataset from Kaggle.com contains information on 7,043 Telco customers from different US states. The dataset includes 21 variables that are quantitative, binary, and categorical predictors. The dataset provides information about customer demographics, such as gender, senior citizen status, dependents, partner status, and tenure, as well as details about their phone and internet services, online security, online backup, device protection, tech support, TV and movie streaming, contract type, payment and billing options, monthly and total charges. Additionally, the dataset includes a binary variable for customer churn, indicating whether a customer has left the company or not. For a snapshot of our entire dataset, please refer to [*Appendix 1*]. For the description of the variables in our dataset, please refer to [*Appendix 2*].

5. **Descriptive Analytics:**

   5.1-5.2 Visual and Quantitative Analytics: For our analysis of the Telco dataset, our outcome variable is customer churn, which is a binary variable indicating whether a customer has left the company or not. We explored the outcome variable, churn, to see the distribution of churned customers versus the non-churned customers *[Appendix 3]*. This also helped us see the balance of the dataset and check for outliers.
   Our preliminary analysis of the dataset focused on several key predictors that we believe may influence customer churn, including customer demographics, account information, and service usage patterns. The mean monthly charges were $64.80 and the mean tenure was 32.42 months. The box plot compares the distribution of the numerical variables ("tenure" and "MonthlyCharges") between customers who churned and those who did not *[Appendix 4]*. We also performed ANOVA *[Appendix 5]* to test whether there is a significant difference in means between the groups of churned and non churned customers. Both these variables have p-values less than 0.05 which suggests that monthly charges and tenure may be important factors in predicting churn. Furthermore, we examined the relationships between the categorical predictors

("InternetService", and "Contract") and churn using bar graphs *[Appendix 6]*. We found that customers with fiber optic internet service and yearly contracts had a lower likelihood of churning. Finally, we analyzed the relationships between the binary predictors ("TechSupport", "Device Protection", "OnlineSecurity" and "StreamingMovies") and churn using bar graphs, as well *[Appendix 7]*. We found that customers with tech support, device protection, online security and streaming movies had lower likelihood of churning. By conducting this descriptive analysis, we gained insights into the relationships between the key predictors and the response variable. These insights helped us with variable selection later by allowing us to identify the most important predictors for our model specification.

5.3 Data Pre-Processing and Transformations: For our Telco dataset, we started by checking for missing values and found 11, which we removed to ensure the validity of our analysis. We then eliminated the customerID column as it was not necessary for our analysis. To make our further analysis more straightforward, we converted the values for SeniorCitizen to 'Yes' and 'No'. Additionally, we re-coded the values for the "No phone service" and "No internet service" columns to simply "No" for easier analysis. Then, we converted our categorical variables to factors to run our models.

6. Modeling Methods and Model Specifications

6.0 Initial Set of Predictors: Based on our understanding of the telecommunications industry, we came up with an initial set of predictors that we believe may impact customer churn:
MonthlyCharges: Customers paying higher monthly charges are at a higher risk of churning if they perceive a lack of expected services or features.
Contract: The length and type of contract can play a significant role in customer churn as customers who have long-term contracts may be more committed to the Telco compared to those who have short-term contracts.
SeniorCitizen: Analyzing the impact of age on customer churn is crucial as older customers may have distinct preferences, requirements, and support needs, potentially influencing their service preferences and the need for additional assistance.
TechSupport, OnlineSecurity, and DeviceProtection: These are additional features that we think may influence customer satisfaction and retention.
StreamingMovies: The availability of streaming movies can have a significant impact on younger generations.
Tenure: Customers with longer tenure are less likely to switch providers due to the established relationship they have built.

6.1 Initial Logistic Regression Modeling: We fit a logistic regression using these 8 predictors and it produced a significant F test *[Appendix 8]*. The ANOVA table shows the comparison of a null model (Model 1) and the model with our initial set of predictors (Model 2) *[Appendix 9]*. The table shows that Model 2 has a significantly lower deviance than Model 1, with a p-value that is less than 0.001 which suggests that the addition of the predictor variables significantly improves the model's ability to predict customer churn. Significant predictors at the 0.001 significance level include MonthlyCharges, ContractOneyear, ContractTwoYear, SeniorCitizenYes, TechSupportYes, OnlineSecurityYes and tenure. StreamingMovies was significant at 0.1 level. The predictors in the model help reduce the deviance of the null model by (8143.4 - 6011.4)/8143.4 = 0.2618 or 26.18%. This model seems to be a good fit to predict "customer churn".

6.2 Logistic Assumptions Tested: After fitting our initial logistic regression model, we tested for multicollinearity. Our assessment showed that there was no severe multicollinearity in the model as the highest Condition Index (CI) value was 10.76, which is below the threshold of 30 *[Appendix 10]*. We assessed other logistic regression assumptions and found that they all held. Our logistic regression is BLUE (Best Linear Unbiased Estimator) for our classification model.

6.3 Model Specifications Evaluation (and Variable Selection): For this project, we evaluated two model specifications. The first model specification used in this project was the initial set of 8 predictors using a business perspective. The second model specification for all three models was selected using stepwise variable selection with a p-value threshold of 0.05 to include only the most significant predictors. The full model for this variable selection method included all the predictors. The lower bound was the null model. The stepwise variable selection method reduced the number of predictors to 14: SeniorCitizen, Dependents, tenure, MultipleLines, InternetService, OnlineSecurity, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges and TotalCharges *[Appendix 11]*. All are significant at the 0.001 significance level except the SeniorCitizenYes, DependentsYes, OnlineSecurityYes, TechSupportYes, and PaymentethodElectronicCheck. The other two levels of PaymentMethod (Credit card and Mailed check) were in our output result but showed no significance. We decided to exclude Dependents variable from our second set of predictors. So, for model specification 2, we ended up with 13 predictors and fitted into a logistic regression model *[Appendix 11]*.

6.3a Logistic Assumptions Tested (Second Set of Predictors): After fitting our logistic regression model with the second set of 13 predictors obtained from the stepwise variable selection method, which was more than the initial set of predictors, we tested for multicollinearity using the Condition Index *[Appendix 13]*. We observed that the highest Condition Index (CI) value was 46.66, which is above the threshold of 30. This was for the variable "TotalCharges". Despite the high CI value, we decided to retain the variable since the model remained stable and it was still less than 50. We also assessed other logistic regression assumptions and sound that they all held for this model. Therefore, we concluded that the logistic regression model with our second set of predictors is also BLUE.

6.4 Methods Evaluated: First, we fitted a logistic regression model for both the specifications/set of predictors. Next, we applied ridge regression to both sets of predictors because it helps with feature selection and can improve the model performance and avoid overfitting *[Appendix 14, 15]*. This was done for analytical purposes, to avoid completely dropping variables that may still hold some predictive power and making the model relatively stable. Next, we used a random forest model to evaluate both sets of predictors and determine how much it improved our predictive accuracy and to extract a clear understanding of which predictors are most "important" in predicting customer churn at Telco *[Appendix 16, 17]*.

6.5 Cross-Validation Testing: In all six combinations of model and specification, we tested using the 10-fold cross validation for its proven results as a robust evaluation method. We used the *caret* package and used Accuracy as our primary method of evaluation metric. We also computed values of other Accuracy Statistics like Sensitivity and Specificity as well as the Area Under Curve (AUC). The results are in the table below:

		Accuracy		Sensitivity		Specificity		AUC	
		Specification 1	Specification 2	Specification 1	Specification 2	Specification 1	Specification 2	Specification 1	Specification 2
Model 1	Logistic Regression	0.7940	0.8043	0.8988	0.8948	0.5029	0.5522	0.9296	0.9344
Model 2	Ridge Regression	0.7939	0.8016	0.9113	0.9051	0.4698	0.5115	0.9291	0.9325
Model 3	Random Forest	0.7968	0.8022	0.9289	0.9150	0.4318	0.4907	0.7284	0.7875

6.6 Final Method/Specification Selected: We honored our analytics goal of achieving both accuracy and interpretability by selecting the Logistic Regression model with the second set of predictors/Specification 2 *[Appendix 18]*. It is a robust method and is easy to interpret the coefficients and achieves good performance in predicting binary outcomes. This had the highest accuracy rate of 80.43%. We were also happy with the values of Sensitivity, Specificity and AUC for this model and model specification. We then fit this model on the entire data set with glm() with all the 13 predictors from our Specification 2.

7. **Analysis of Results:** The model appears to be a good fit for predicting the customer churn at Telcol. The predictors in the model help reduce the deviance of the null model by (8143.4 - 5831.6)/8143.4 = 0.2839 or 28.39%. This percentage reduction is not very large, but it is a contribution to deviance reduction nevertheless, and there are several significant predictors that help provide useful interpretations. The output displays all predictors as significant, with the exception of "PaymentMethodElectronic Check" and "PaymentMethodMailed Check".

   Quantitative predictors: all interpretations are "on average and holding everything else constant": The variable tenure has the highest effect on the response variable. It represents the number of months the customer has been with the company. The negative coefficient indicates that longer tenure is associated with lower probability of churn. For every additional month of tenure, the log-odds of customer churn decrease by 0.061 and the odd decrease by a factor of 0.941. The variable MonthlyCharges represents the customer's monthly charges. The negative coefficient indicates that higher monthly charges are associated with lower probability of churn. For every unit increase in monthly charges, the log-odds of churn decrease by 0.030 and the odds decrease by a factor of 0.970. This is a surprising result as one would expect the likelihood of churn to increase if monthly charges increased. It could be because the customers may be more committed to using services provided by the company and may be less likely to switch to a competitor or that they have opted for longer term contracts. The variable TotalCharges represents the customer's total charges. The positive coefficient indicates that higher total charges are associated with higher probability of churn. For every unit increase in total charges, the log-odds of churn increase by 0 and the odds increase by a factor of 1 (no change). It also has a 50% probability which suggests that TotalCharges may not be a strong predictor of churn.

   Binary predictors: all interpretations are "on average and holding everything else constant": Among the binary predictors in our data set, the presence of certain services has a significant effect on the likelihood of customer churn. Surprisingly, customers who have streaming movie or TV services have a higher log-odds of churning compared to those without these services, with increases of 0.497 and 0.511, respectively. The variables MultipleLines and PaperlessBilling are indicators of whether the customer has multiple phone lines and paperless billing as an option and the positive coefficient indicates that customers with multiple lines and paperless billing options are more likely to churn, which is very interesting as well. Customers with online security and tech support services have a decrease in log-odds of churning, with decreases of 0.262 and 0.228, respectively. This makes sense, as having these services provides customers with satisfaction and reassurance. Finally, being a senior citizen also has a positive effect on churn, with the log odds of churn increasing by 0.243 for a senior citizen compared to a non-senior citizen. This could be due to different technology needs or preferences of senior citizens, which if not met, may lead to switching to a different provider.

   Categorical predictors: all interpretations are "on average and holding everything else constant": For our categorical variables, we selected reference levels to compare the other levels. Our reference level for the InternetService variable is DSL, which is the most basic and affordable internet option available. We found that customers with fiber optic internet service have the highest positive effect on churn, with a log-odds of churning being 1.491 higher compared to DSL users, which means that customers with fiber optic internet service are more likely to churn than those with DSL service. On the other hand, customers without internet service are less likely to churn compared to those with DSL service, with log-odds of -1.561, which is very surprising. For the Contract variable, we chose month to month contracts to be our reference level because it is the most common type of contract offered by Telco and it made sense to compare the other contract types to this baseline. Our results showed that having longer-term contracts, such as one-year and two-year contracts, reduces the log-odds of churn by 0.663 and 1.360 units respectively, compared to a month-to-month contract. This makes sense since

long-term contracts often come with incentives such as discounted rates or other perks that may seem attractive to the customers. Lastly, for the PaymentMethod variable, we selected bank transfer as our reference level since it is a commonly used and well-known payment method. Our results indicate that customers who use electronic check as a payment method have log-odds of churning that is 0.306 higher compared to those who use bank transfer. This could be because electronic checks may be perceived as less secure or reliable by some customers.

8. Conclusion and Lessons Learned

8.1 Conclusions from the Analysis: Based on our analysis, we have identified the main factors that contribute to customer churn at Telco, which is a significant concern for the company. The most influential predictors of predicting customer churn at Telco are: IntentServiceFiberoptic, ContractTwoYear and tenure. Customers who have fiber optic internet service are more likely to churn compared to those with other types of internet service or no internet service at all. Similarly, customers who have signed up for a longer-term contract, such as a two-year contract, are less likely to churn compared to those who have a shorter term contract. Finally, the longer a customer has been with Telco, the less likely they are going to churn.

These insights can help Telco identify the main reasons why customers leave and develop strategies to retain them. Telco could focus on improving the quality of its fiber optic internet service, pricing incentives for customers to sign up for longer-term contracts, and by offering loyalty programs to reward long-time customers. By addressing these key factors, Telco could improve its customer retention rate and maintain its competitive edge in the market. Overall, our model and analysis provide Telco with important information to better understand its customers and make data-driven decisions to improve its services and retain its customers.

8.2 Project Issues, Challenges, and Lessons Learned: One major challenge that we faced was the imbalance distribution of churn class in the dataset. This imbalance posed a challenge in analyzing and making decisions from the data. However, we overcame this challenge by combining the predictions of multiple models to arrive at a more accurate final prediction. Another challenge we encountered was that some variable types had options other than “Yes” and “No”, which we treated as “No” to maintain consistency in the data, which may introduce some bias in the analysis. Furthermore, we faced a challenge with the limited information provided by binary variables, which offer less information compared to variables with numeric values. In addition, we had to carefully consider the impact of each variable on customer churn while deciding on the model specification based on business insights. This required us to provide a comprehensive rationale for explaining the importance of the variables.

An important lesson learned was that we should not rely solely on business insights when selecting variables for the model. This was evident after comparing the accuracy of model specification 1 based on business insights and model specification 2 using the stepwise method. The statistical method was more accurate in providing us with significant predictions. Another important lesson learned from our project on predicting customer churn at Telco is that the insights gained from our analysis can help the company identify key factors that influence customer retention. Although our model identified the main factors that contribute to customer churn at Telco, there may be other important variables that we did not account for in our analysis. With additional time and resources, we could investigate other potential predictors of churn, such as customer satisfaction ratings, customer demographics, service quality and pricing strategies. We could also explore how these factors vary across different customer segments and markets which would then allow Telco to tailor their retention strategies to different groups of customers. In summary, the main lesson learned from this project is that predictive analytics can provide valuable insights to businesses looking to improve their services and retain customers. The insights gained from this analysis provide a starting point for Telco to better understand customer churn and develop strategies to resin their customer base.

Appendix Contents

Appendix 1: Snapshot of dataset

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity
1	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No
2	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes
3	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes
4	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes
5	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No
6	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No
7	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No
8	6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes
9	7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No
10	6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes

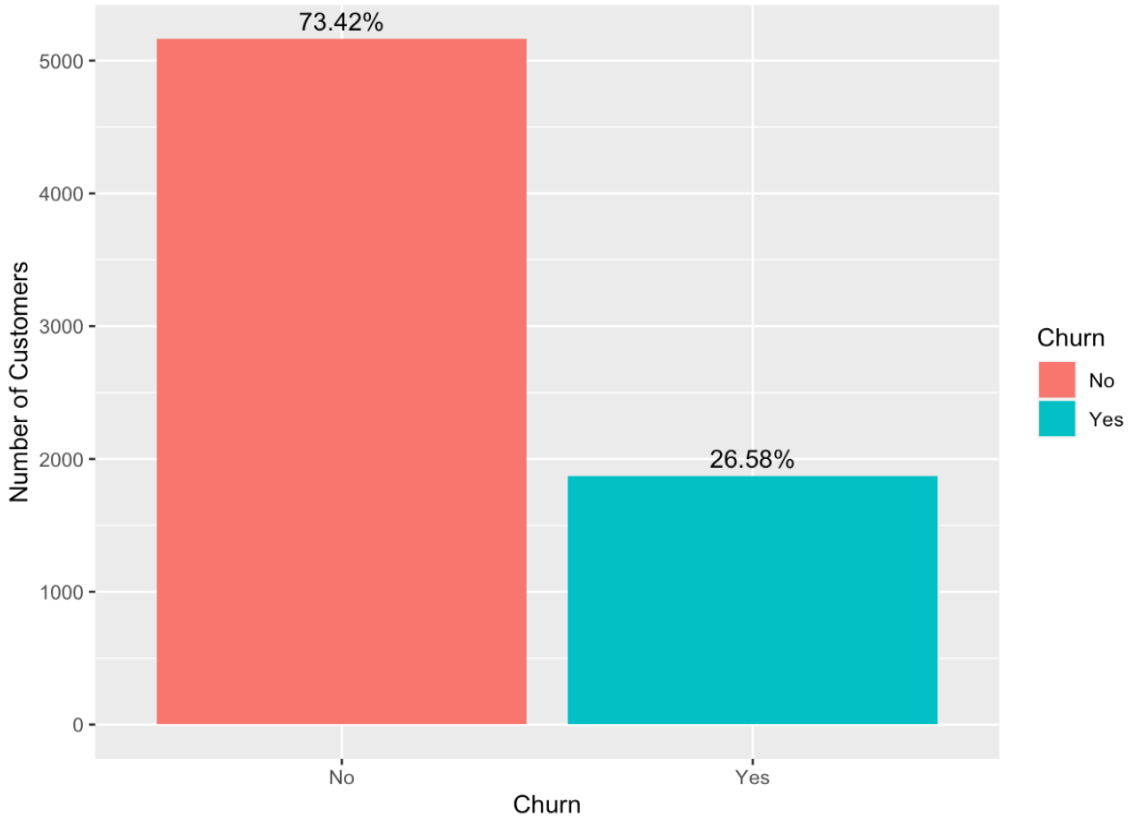
OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod
Yes	No	No	No	No	Month-to-month	Yes	Electronic check
No	Yes	No	No	No	One year	No	Mailed check
Yes	No	No	No	No	Month-to-month	Yes	Mailed check
No	Yes	Yes	No	No	One year	No	Bank transfer (automatic)
No	No	No	No	No	Month-to-month	Yes	Electronic check
No	Yes	No	Yes	Yes	Month-to-month	Yes	Electronic check
Yes	No	No	Yes	No	Month-to-month	Yes	Credit card (automatic)
No	No	No	No	No	Month-to-month	No	Mailed check
No	Yes	Yes	Yes	Yes	Month-to-month	Yes	Electronic check
Yes	No	No	No	No	One year	No	Bank transfer (automatic)

MonthlyCharges	TotalCharges	Churn
29.85	29.85	No
56.95	1889.50	No
53.85	108.15	Yes
42.30	1840.75	No
70.70	151.65	Yes
99.65	820.50	Yes
89.10	1949.40	No
29.75	301.90	No
104.80	3046.05	Yes
56.15	3487.95	No

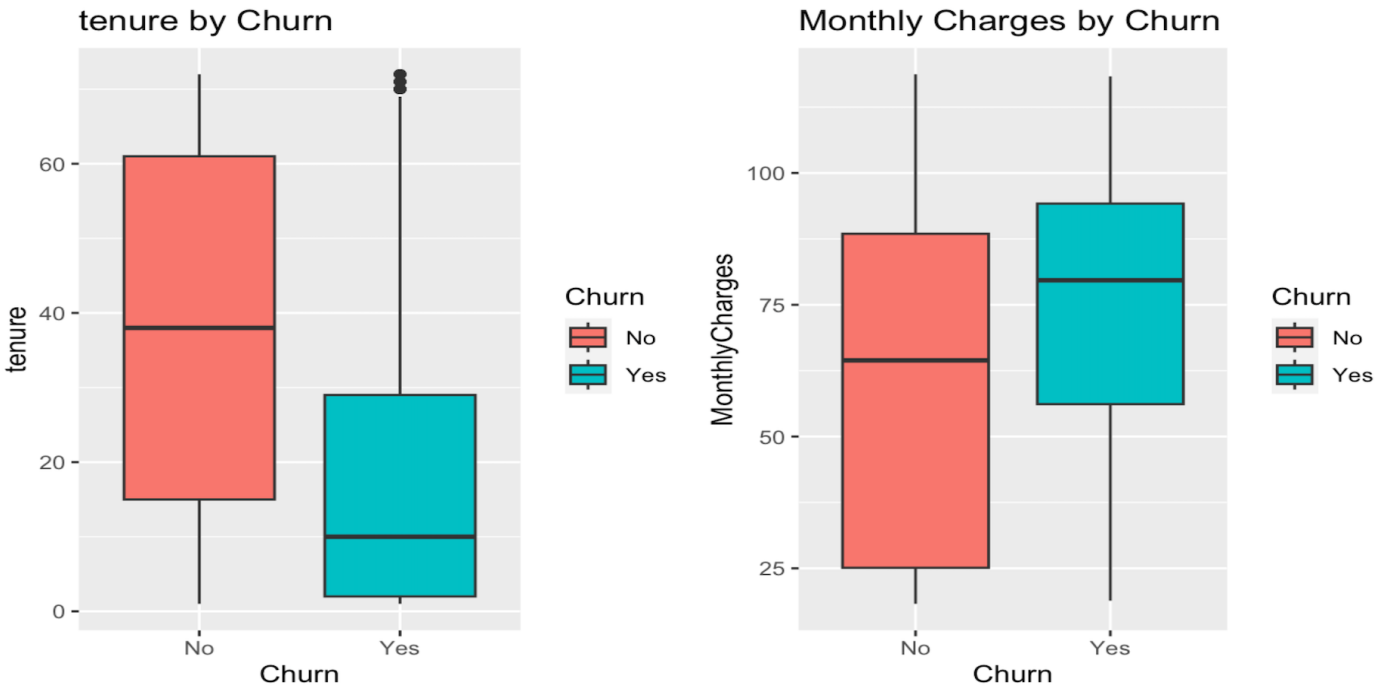
Appendix 2: Description of the dataset

	Variable	Description
2	customerID	Unique identifier for each customer in the dataset.
3	gender	The customer's gender, either male or female.
4	SeniorCitizen	Indicates is the customer is a senior citizen (1) or not (0).
5	Partner	Indicates if the customer has a partner (Yes) or not (No).
6	Dependents	Indicates if the customer has dependents (Yes) ot not (No).
7	tenure	The number of months the customer has stayed with the company.
8	PhoneService	Indicates if the customer has a phone service (Yes) or not (No).
9	MultipleLines	Indicates if the customer has multiple lines (Yes, No, No phone service).
10	InternetService	Indicates if the customer has internet service (DSL, Fiber optic, or No).
11	OnlineSecurity	Indicates if the customer has online security (Yes, No, No internet service).
12	OnlineBackup	Indicates if the customer has online backup (Yes, No, No internet service).
13	DeviceProtection	Indicates if the customer has device protection (Yes, No, No internet service).
14	TechSupport	Indicates if the customer has tech support (Yes, No, No internet service).
15	StreamingTV	Indicates if the customer has TV streaming (Yes, No, No internet service).
16	StreamingMovies	Indicates if the customer has movies streaming(Yes, No, No internet service).
17	Contract	Indicates the contract term of the customer (Month-to-month, One year, or Two year)
18	PaperlessBilling	Indicates if the customer has opted for paperless billing (Yes) or not (No).
19	PaymentMethod	Indicates if the customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), or Credit card (automatic))
20	MonthlyCharges	The amount charged to the customer monthly.
21	TotalCharges	The total amount charged to the customer.
22	Churn	Indicates if the customer has churned (Yes) or not (No).

Appendix 3: Exploring the Churn variable



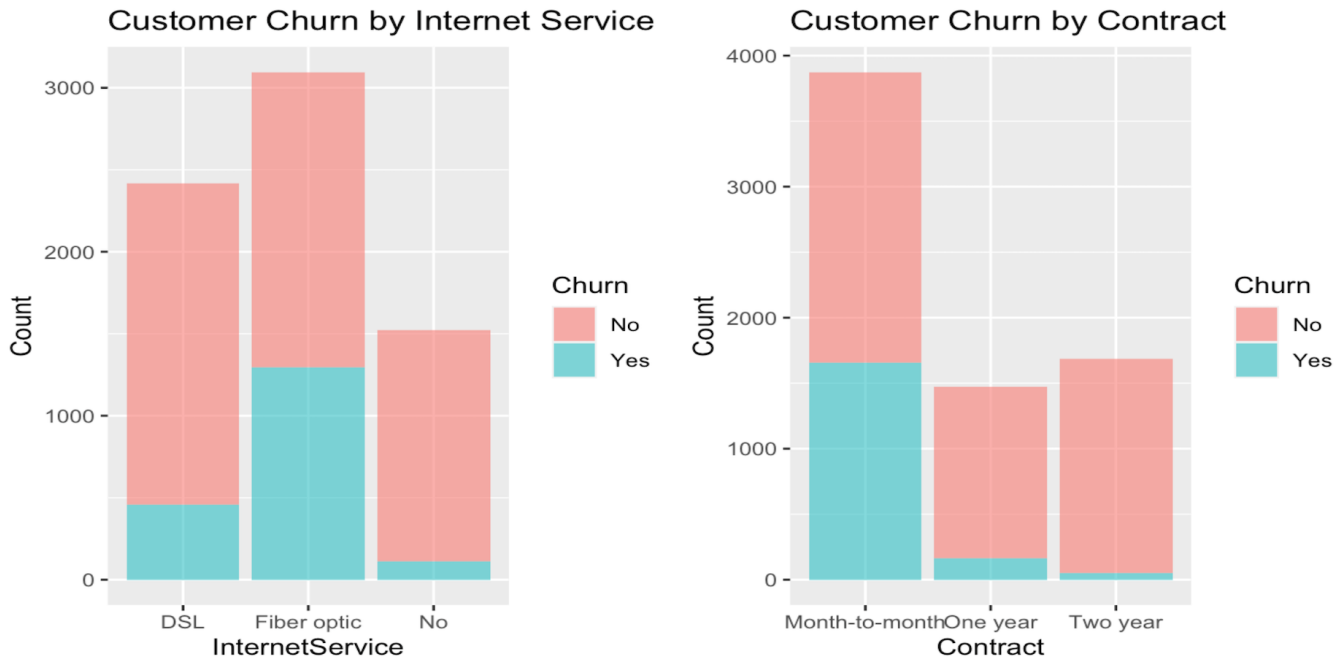
Appendix 4: Box plot of tenure and MonthlyCharges vs Churn



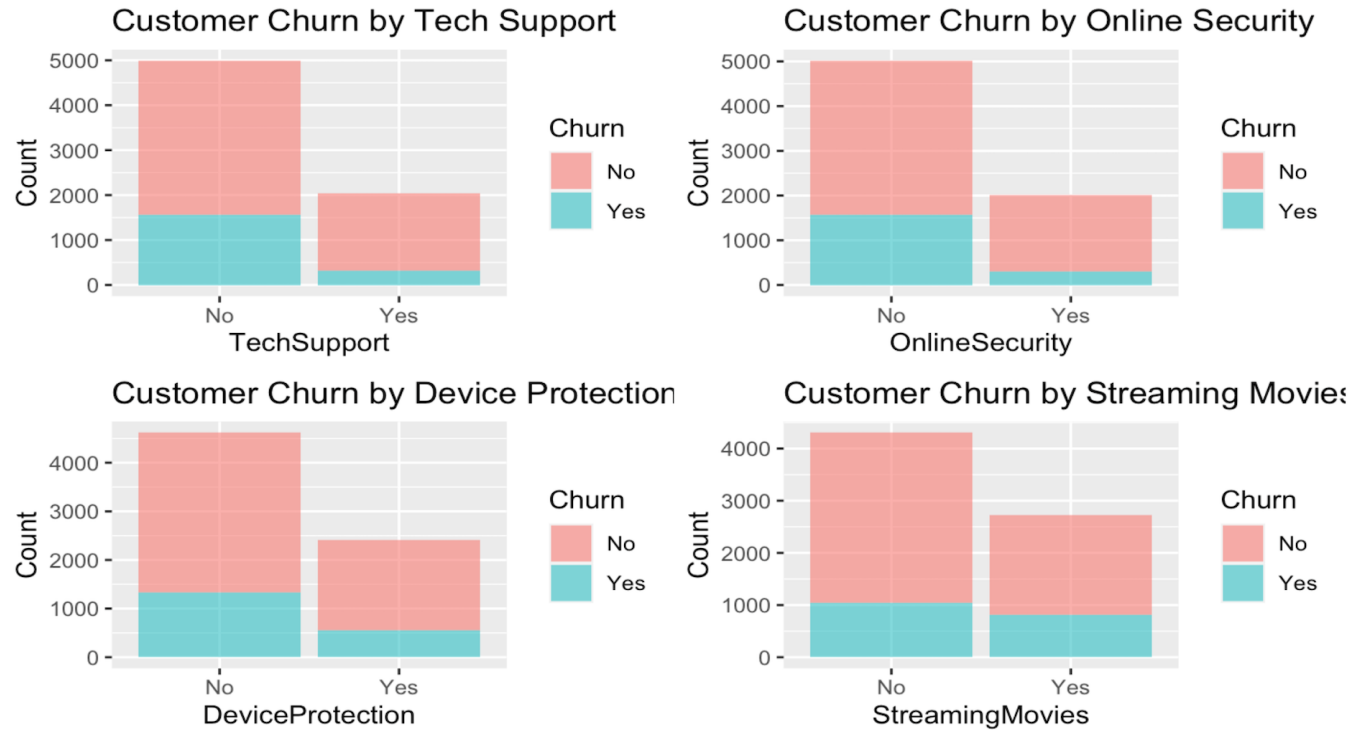
Appendix 5: ANOVA of tenure and MonthlyCharges vs Churn

```
> AOV_tenure
              Df  Sum Sq Mean Sq F value Pr(>F)
Churn          1  530982  530982    1008 <2e-16 ***
Residuals   7030 3704983     527
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> AOV_monthly
              Df  Sum Sq Mean Sq F value Pr(>F)
Churn          1  236713  236713   271.6 <2e-16 ***
Residuals   7030 6127508     872
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 6: Bar graph on InternetService and Contract vs Churn



Appendix 7: Bar graph on TechSupport, DeviceProtection, OnlineSecurity and StreamingMovies vs Churn



Appendix 8: Fitting the Logistic Regression model (initial set of predictors)

```
Call:
glm(formula = Churn ~ MonthlyCharges + Contract + SeniorCitizen +
    TechSupport + OnlineSecurity + DeviceProtection + StreamingMovies +
    tenure, family = binomial(link = logit), data = TelcoCustomerChurn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8907  -0.7216  -0.3048   0.7432   3.1740

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.580885   0.093909 -16.834  < 2e-16 ***
MonthlyCharges       0.029201   0.001636  17.855  < 2e-16 ***
ContractOne year    -0.865593   0.103773  -8.341  < 2e-16 ***
ContractTwo year    -1.624963   0.170954  -9.505  < 2e-16 ***
SeniorCitizenYes     0.371525   0.081728   4.546 5.47e-06 ***
TechSupportYes      -0.503940   0.083846  -6.010 1.85e-09 ***
OnlineSecurityYes   -0.543068   0.082527  -6.580 4.69e-11 ***
DeviceProtectionYes -0.109851   0.079040  -1.390   0.1646
StreamingMoviesYes   0.147597   0.081515   1.811   0.0702 .
tenure              -0.034574   0.002120 -16.309  < 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: 8143.4  on 7031  degrees of freedom
Residual deviance: 6011.4  on 7022  degrees of freedom
AIC: 6031.4

Number of Fisher Scoring iterations: 6
```

	Log-Odds	Odds	Probabilities
(Intercept)	-1.581	0.206	0.171
MonthlyCharges	0.029	1.030	0.507
ContractOne year	-0.866	0.421	0.296
ContractTwo year	-1.625	0.197	0.165
SeniorCitizenYes	0.372	1.450	0.592
TechSupportYes	-0.504	0.604	0.377
OnlineSecurityYes	-0.543	0.581	0.367
DeviceProtectionYes	-0.110	0.896	0.473
StreamingMoviesYes	0.148	1.159	0.537
tenure	-0.035	0.966	0.491

Appendix 9: ANOVA of null model vs Initial Logistic Model

```
Analysis of Deviance Table

Model 1: Churn ~ 1
Model 2: Churn ~ MonthlyCharges + Contract + SeniorCitizen + TechSupport +
    OnlineSecurity + DeviceProtection + StreamingMovies + tenure
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      7031     8143.4
2      7022     6011.4  9   2131.9 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 10: Logit Assumption Test (initial set of Predictors): Multicollinearity

```
> cond.index(model1, data = TelcoCustomerChurn)
 [1]  1.000000  2.267531  2.449930  2.996537  3.161489  3.513081  3.679116  4.143041  6.832414 10.758855
```

Appendix 11: Variable Selection for Second Specification: Stepwise

```
#Model Specification 2: Variable Selection: Stepwise
fit.null <- glm(Churn ~ 1, data = TelcoCustomerChurn, family= binomial(link=logit))
scope = list(lower = fit.null, upper = full.model)
fit.step <- step(full.model, scope = list(lower=fit.null, upper=full.model),
                 direction = "both", test = "F")
summary(fit.step)
```

```
Call:
glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +
    InternetService + OnlineSecurity + TechSupport + StreamingTV +
    StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
    MonthlyCharges + TotalCharges, family = binomial(link = logit),
    data = TelcoCustomerChurn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9338  -0.6797  -0.2869   0.7259   3.4233

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                         8.802e-01  2.792e-01   3.152  0.00162 **
SeniorCitizenYes                    2.184e-01  8.394e-02   2.601  0.00929 **
DependentsYes                      -1.495e-01  8.140e-02  -1.836  0.06636 .
tenure                             -6.048e-02  6.209e-03  -9.742  < 2e-16 ***
MultipleLinesYes                    3.911e-01  8.766e-02   4.462 8.13e-06 ***
InternetServiceFiber optic          1.479e+00  1.924e-01   7.686 1.51e-14 ***
InternetServiceNo                  -1.549e+00  1.781e-01  -8.697  < 2e-16 ***
OnlineSecurityYes                  -2.591e-01  8.963e-02  -2.891  0.00384 **
TechSupportYes                     -2.284e-01  9.129e-02  -2.502  0.01237 *
StreamingTVYes                      4.979e-01  9.769e-02   5.097 3.45e-07 ***
StreamingMoviesYes                  5.072e-01  9.662e-02   5.250 1.52e-07 ***
ContractOne year                   -6.523e-01  1.074e-01  -6.077 1.23e-09 ***
ContractTwo year                   -1.343e+00  1.761e-01  -7.627 2.40e-14 ***
PaperlessBillingYes                 3.411e-01  7.442e-02   4.583 4.58e-06 ***
PaymentMethodCredit card (automatic) -8.644e-02  1.140e-01  -0.759  0.44808
PaymentMethodElectronic check       3.027e-01  9.444e-02   3.206  0.00135 **
PaymentMethodMailed check          -5.923e-02  1.147e-01  -0.516  0.60561
MonthlyCharges                     -2.999e-02  5.793e-03  -5.177 2.25e-07 ***
TotalCharges                        3.310e-04  7.038e-05   4.703 2.56e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8143.4  on 7031  degrees of freedom
Residual deviance: 5828.2  on 7013  degrees of freedom
AIC: 5866.2
```

Appendix 12: Logistic Regression for Specification 2 (stepwise method)

```
logit.model2 <- glm(Churn ~ SeniorCitizen + tenure + MultipleLines + InternetService +
                      StreamingTV + TechSupport + StreamingMovies + Contract + OnlineSecurity +
                      PaperlessBilling + PaymentMethod + MonthlyCharges + TotalCharges,
            data=TelcoCustomerChurn, family = binomial(link=logit))
summary(logit.model2)

# Transforming Coefficients
log.odds <- coef(logit.model2) # To get just the coefficients
odds <- exp(coef(logit.model2))
prob <- odds / (1 + odds) # To convert odds to probabilities
round(cbind("Log-Odds" = log.odds,
            "Odds" = odds,
            "Probabilities" = prob), # All together
     digits = 3)
```

Appendix 13: Logit Assumption Test (Specification 2): Multicollinearity

```
 [1]  1.000000  2.168484  2.479630  2.958797  3.099633  3.519918  3.802771  3.921963  4.140286  4.544790
[11]  5.290028  5.460646  5.516845  6.562624  7.666769 10.521338 17.881465 46.661209
```

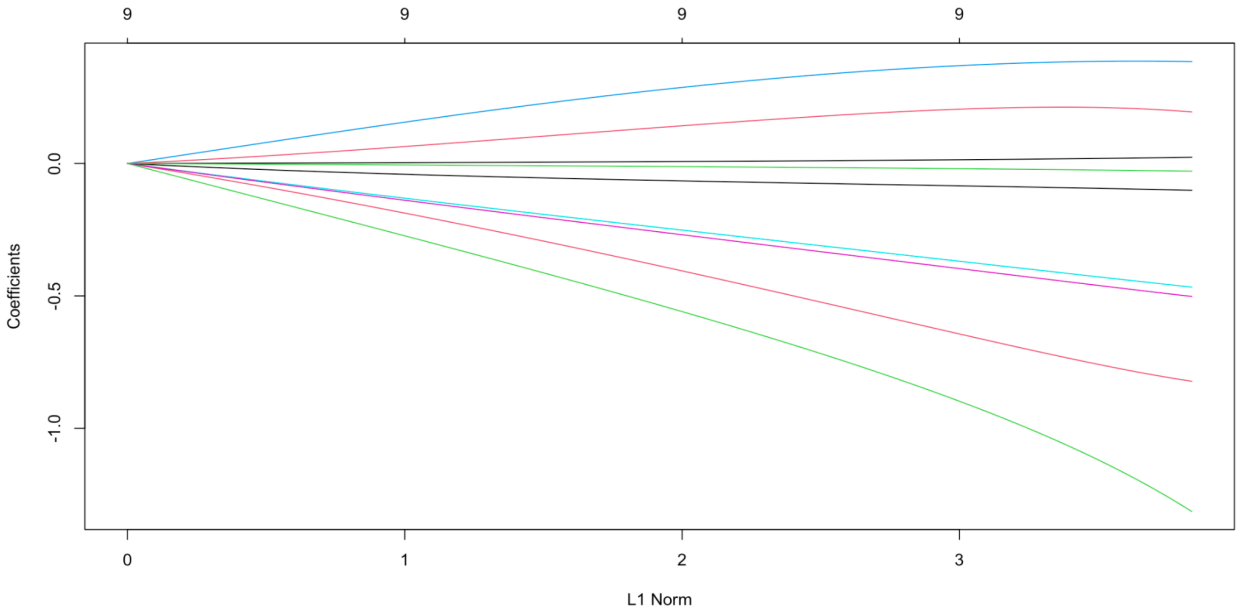
Appendix 14: Model 2: Ridge Regression with initial set of predictors

```
Call:  glmnet(x = x, y = y, family = "binomial", alpha = 0)

    Df  %Dev  Lambda
1    9  0.00 156.400
2    9  0.09 142.500
3    9  0.10 129.800
4    9  0.11 118.300
5    9  0.12 107.800
6    9  0.13  98.220
7    9  0.14  89.500
8    9  0.15  81.550
9    9  0.17  74.300
10   9  0.18  67.700
11   9  0.20  61.690
12   9  0.22  56.210
13   9  0.24  51.210
14   9  0.27  46.660
15   9  0.29  42.520
16   9  0.32  38.740
17   9  0.35  35.300
18   9  0.38  32.160
19   9  0.42  29.310
20   9  0.46  26.700
21   9  0.51  24.330
22   9  0.55  22.170
23   9  0.61  20.200
24   9  0.66  18.410
25   9  0.73  16.770
26   9  0.80  15.280
27   9  0.87  13.920
28   9  0.95  12.690
29   9  1.04  11.560
30   9  1.14  10.530
31   9  1.25   9.597
32   9  1.36   8.744
33   9  1.49   7.967
34   9  1.62   7.260
35   9  1.77   6.615
36   9  1.93   6.027
37   9  2.10   5.492
38   9  2.29   5.004
39   9  2.49   4.559
40   9  2.71   4.154
```

```
40   9  2.71   4.154
94   9  2.57   3.785
19   9  2.43   3.449
46   9  2.30   3.142
75   9  2.17   2.863
06   9  2.06   2.609
39   9  1.94   2.377
74   9  1.84   2.166
12   9  1.74   1.974
51   9  1.64   1.798
93   9  1.56   1.638
36   9  1.47   1.493
82   9  1.40   1.360
31   9  1.33   1.239
81   9  1.26   1.129
33   9  1.20   1.029
87   9  1.14   0.938
43   9  1.08   0.854
00   9  1.03   0.778
59   9  0.97   0.709
18   9  0.93   0.646
79   9  0.88   0.589
41   9  0.84   0.536
02   9  0.80   0.489
65   9  0.76   0.445
27   9  0.73   0.406
88   9  0.70   0.370
50   9  0.67   0.337
10   9  0.64   0.307
69   9  0.62   0.280
27   9  0.59   0.255
84   9  0.57   0.232
39   9  0.55   0.212
92   9  0.53   0.193
43   9  0.51   0.176
93   9  0.50   0.160
40   9  0.48   0.146
85   9  0.47   0.133
28   9  0.46   0.121
68   9  0.45   0.110
06   9  0.44   0.100
```

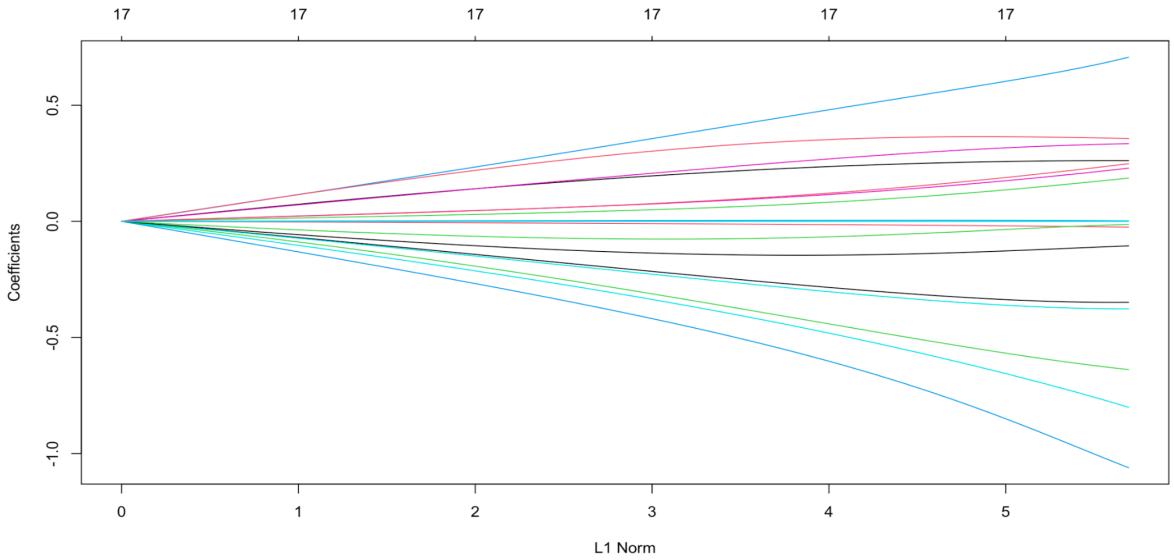
```
80   9 22.06   0.100
81   9 22.42   0.092
82   9 22.76   0.083
83   9 23.08   0.076
84   9 23.37   0.069
85   9 23.64   0.063
86   9 23.90   0.058
87   9 24.13   0.052
88   9 24.34   0.048
89   9 24.54   0.044
90   9 24.72   0.040
91   9 24.88   0.036
92   9 25.03   0.033
93   9 25.17   0.030
94   9 25.29   0.027
95   9 25.39   0.025
96   9 25.49   0.023
97   9 25.58   0.021
98   9 25.66   0.019
99   9 25.72   0.017
100  9 25.78   0.016
```



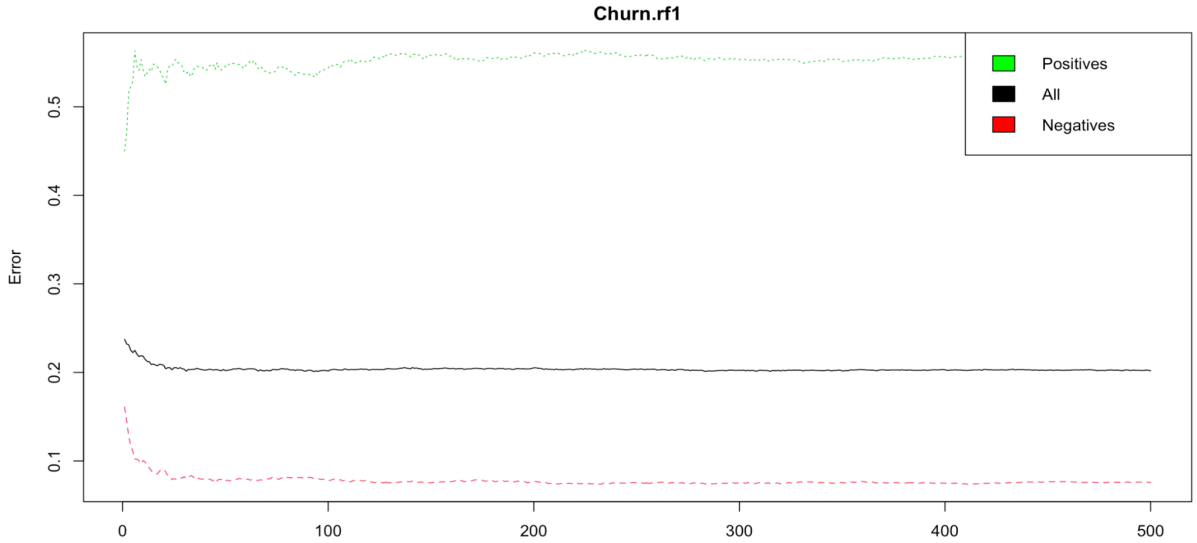
Appendix 15: Model 2: Ridge Regression with Specification 2

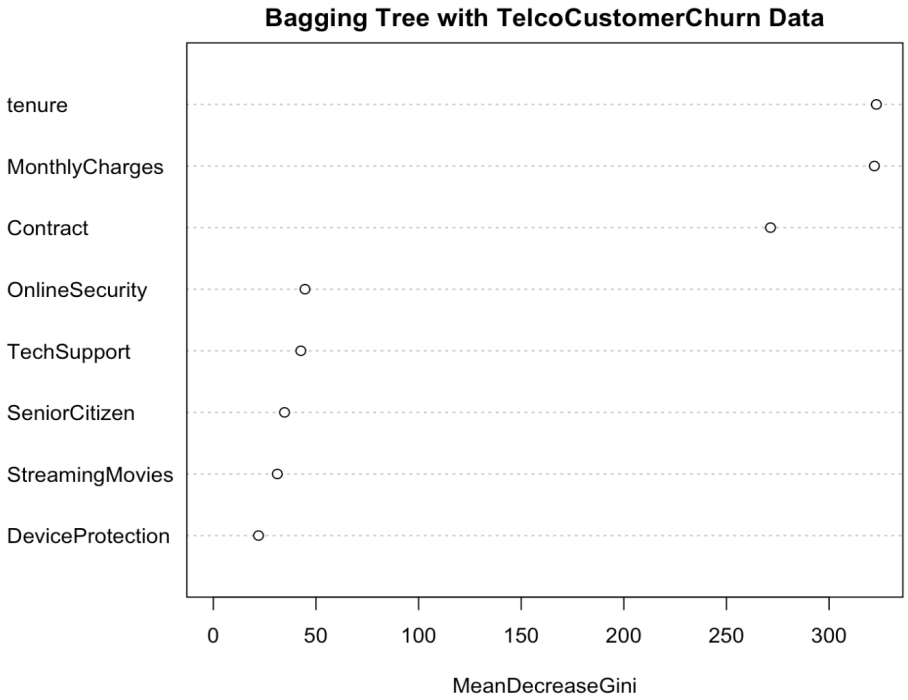
```
Call:  glmnet(x = x2, y = y2, family = "binomial",
alpha = 0)

    Df  %Dev  Lambda
1   17  0.00 156.400
2   17  0.17 142.500
3   17  0.18 129.800
4   17  0.20 118.300
5   17  0.22 107.800
6   17  0.24  98.220
7   17  0.27  89.500
8   17  0.29  81.550
9   17  0.32  74.300
10  17  0.35  67.700
11  17  0.39  61.690
12  17  0.42  56.210
13  17  0.46  51.210
14  17  0.51  46.660
15  17  0.56  42.520
16  17  0.61  38.740
17  17  0.67  35.300
18  17  0.73  32.160
19  17  0.80  29.310
20  17  0.88  26.700
21  17  0.96  24.330
22  17  1.05  22.170
23  17  1.15  20.200
24  17  1.25  18.410
25  17  1.37  16.770
26  17  1.50  15.280
27  17  1.63  13.920
28  17  1.78  12.690
29  17  1.95  11.560
30  17  2.12  10.530
31  17  2.32   9.597
32  17  2.52   8.744
33  17  2.75   7.967
34  17  2.99   7.260
35  17  3.25   6.615
36  17  3.53   6.027
37  17  3.83   5.492
38  17  4.15   5.004
39  17  4.49   4.559
40  17  4.86   4.154
40  17  4.86   4.154
41  17  5.25   3.785
42  17  5.66   3.449
43  17  6.10   3.142
44  17  6.56   2.863
45  17  7.05   2.609
46  17  7.56   2.377
47  17  8.09   2.166
48  17  8.65   1.974
49  17  9.22   1.798
50  17  9.82   1.638
51  17 10.43   1.493
52  17 11.06   1.360
53  17 11.70   1.239
54  17 12.36   1.129
55  17 13.02   1.029
56  17 13.69   0.938
57  17 14.35   0.854
58  17 15.02   0.778
59  17 15.68   0.709
60  17 16.34   0.646
61  17 16.99   0.589
62  17 17.62   0.536
63  17 18.24   0.489
64  17 18.85   0.445
65  17 19.43   0.406
66  17 20.00   0.370
67  17 20.54   0.337
68  17 21.06   0.307
69  17 21.56   0.280
70  17 22.03   0.255
71  17 22.47   0.232
72  17 22.90   0.212
73  17 23.29   0.193
74  17 23.67   0.176
75  17 24.01   0.160
76  17 24.34   0.146
77  17 24.64   0.133
78  17 24.92   0.121
79  17 25.18   0.110
80  17 25.42   0.100
80  17 25.42   0.100
81  17 25.64   0.092
82  17 25.84   0.083
83  17 26.02   0.076
84  17 26.19   0.069
85  17 26.35   0.063
86  17 26.49   0.058
87  17 26.62   0.052
88  17 26.73   0.048
89  17 26.84   0.044
90  17 26.94   0.040
91  17 27.02   0.036
92  17 27.10   0.033
93  17 27.18   0.030
94  17 27.24   0.027
95  17 27.31   0.025
96  17 27.36   0.023
97  17 27.42   0.021
98  17 27.47   0.019
99  17 27.51   0.017
100 17 27.55   0.016
```



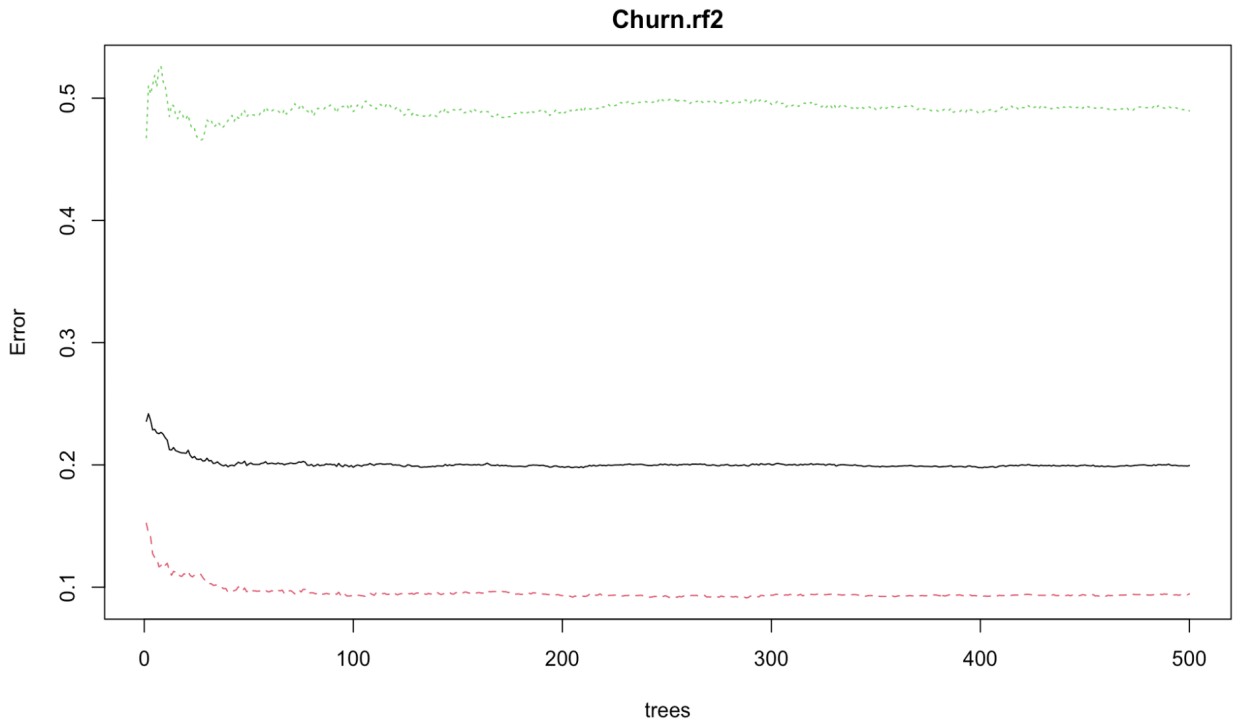
Appendix 16: Model 3: Random Forest Classification Tree (initial set of predictors)

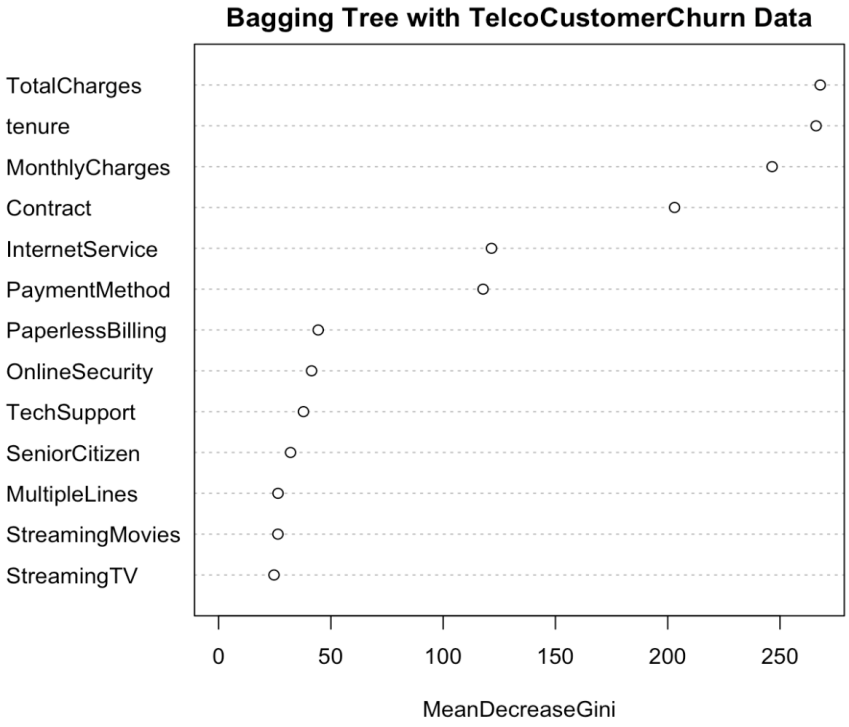




	MeanDecreaseGini
MonthlyCharges	322.06459
Contract	271.45284
SeniorCitizen	34.67836
TechSupport	42.62630
OnlineSecurity	44.64289
DeviceProtection	22.00059
StreamingMovies	31.17470
tenure	323.03703

Appendix 17: Model 3: Random Forest Classification Tree (Specification 2)





	MeanDecreaseGini
SeniorCitizen	32.04698
tenure	266.06418
MultipleLines	26.45818
InternetService	121.52972
StreamingTV	24.68848
TechSupport	37.79506
StreamingMovies	26.41138
Contract	203.04784
OnlineSecurity	41.40989
PaperlessBilling	44.35775
PaymentMethod	117.77813
MonthlyCharges	246.48932
TotalCharges	267.92571

Appendix 18: Fitting the Final Model Choice - Logistic Regression, Specification 2 (full data)

```
Call:
glm(formula = Churn ~ SeniorCitizen + tenure + MultipleLines +
    InternetService + StreamingTV + TechSupport + StreamingMovies +
    Contract + OnlineSecurity + PaperlessBilling + PaymentMethod +
    MonthlyCharges + TotalCharges, family = binomial(link = logit),
    data = TelcoCustomerChurn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9338  -0.6763  -0.2876   0.7264   3.4493

Coefficients:
                                     Estimate  Std. Error z value Pr(>|z|)
(Intercept)                          0.85881129  0.27868968   3.082  0.00206 **
SeniorCitizenYes                     0.24333939  0.08289864   2.935  0.00333 **
tenure                              -0.06118133  0.00620908  -9.854  < 2e-16 ***
MultipleLinesYes                     0.39289727  0.08756651   4.487 7.23e-06 ***
InternetServiceFiber optic           1.49079398  0.19216591   7.758 8.64e-15 ***
InternetServiceNo                   -1.56065054  0.17789015  -8.773  < 2e-16 ***
StreamingTVYes                       0.49712292  0.09763357   5.092 3.55e-07 ***
TechSupportYes                      -0.22812939  0.09126667  -2.500  0.01243 *
StreamingMoviesYes                   0.51141171  0.09652827   5.298 1.17e-07 ***
ContractOne year                    -0.66314539  0.10712672  -6.190 6.01e-10 ***
ContractTwo year                    -1.35996442  0.17578371  -7.737 1.02e-14 ***
OnlineSecurityYes                   -0.26223281  0.08957253  -2.928  0.00342 **
PaperlessBillingYes                  0.34461391  0.07437113   4.634 3.59e-06 ***
PaymentMethodCredit card (automatic) -0.08570932  0.11394346  -0.752  0.45193
PaymentMethodElectronic check        0.30573817  0.09442037   3.238  0.00120 **
PaymentMethodMailed check           -0.05890622  0.11468416  -0.514  0.60750
MonthlyCharges                      -0.03027751  0.00578579  -5.233 1.67e-07 ***
TotalCharges                         0.00033650  0.00007042   4.778 1.77e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8143.4  on 7031  degrees of freedom
Residual deviance: 5831.6  on 7014  degrees of freedom
AIC: 5867.6

Number of Fisher Scoring iterations: 6
```

	Log-Odds	Odds	Probabilities
(Intercept)	0.859	2.360	0.702
SeniorCitizenYes	0.243	1.276	0.561
tenure	-0.061	0.941	0.485
MultipleLinesYes	0.393	1.481	0.597
InternetServiceFiber optic	1.491	4.441	0.816
InternetServiceNo	-1.561	0.210	0.174
StreamingTVYes	0.497	1.644	0.622
TechSupportYes	-0.228	0.796	0.443
StreamingMoviesYes	0.511	1.668	0.625
ContractOne year	-0.663	0.515	0.340
ContractTwo year	-1.360	0.257	0.204
OnlineSecurityYes	-0.262	0.769	0.435
PaperlessBillingYes	0.345	1.411	0.585
PaymentMethodCredit card (automatic)	-0.086	0.918	0.479
PaymentMethodElectronic check	0.306	1.358	0.576
PaymentMethodMailed check	-0.059	0.943	0.485
MonthlyCharges	-0.030	0.970	0.492
TotalCharges	0.000	1.000	0.500

Works Cited

1. Bernazzani, S. (2022, March 11). Here's why customer retention is so important for roi, customer loyalty, and growth. HubSpot Blog. Retrieved April 12, 2023, from *https://blog.hubspot.com/service/customer-retention*

2. Liibert, K. (2023, March 6). Essentials of customer churn and retention. Smartlook Blog. Retrieved April 9, 2023, from *https://www.smartlook.com/blog/customer-churn-retention/#:~:text=The%20average%20churn%20rate%20in,take%20action%20to%20reduce%20it.*