6414 Group Project - Telecom Customer Churn Modeling

Jared Babcock, Rishi Bubna, Marta Bras

2019-12-05

Contents

1 Abstract

The telecommunications industry is facing a challenging future, with the emergence of new technologies and the increasing levels of competition resulting in unprecedented levels of customer attrition and price competition between the companies in the sector.

In a highly price sensitive market, customer service and product features play an important role. For telecommunication companies to better tailor their products to customer expectations, understanding the reasons for customer attrition is a crucial step. In that sense, customer churn analysis is one of the vital measures for subscription-based business models such as telecom services and internet providers.

In this report, we developed a model to explore the reasons why a specific telecommunications company's customers churn. In our analysis, we were particularly interested in understanding differences between groups, and to leverage that information to suggest future customer segmentation strategies. We used modelling to also predict which customers are likely to churn in the future.

Additionally, we use modelling to predict each customer's lifetime value (CLTV), as a measure of a particular customer's net worth to the company, during his relationship with the company.

We argue that if a company is able to predict if a customer is likely to churn, while also being able to identify if the same customer is worth reatining (based on a predicted value for CLTV), then the company can choose to increase engagement with the customer in order to retain him.

Preserving "at risk" valuable customers, while leveraging on information about differences in groups to develop better segmentation strategies, can potentially help a telecommunication company differentiate from the competition and hence increase revenues in the long run.

2 Introduction

2.1 Reasons for our analysis

Understanding customers' preferences is essential for any business, playing an even more important role when competition and price elasticity of demand is high. This is the case for the telecommunication industry, in which customers frequently change among telecom operators, resulting in high churn rates and competitive pressures for the companies.

Customer segmentation and targeting are marketing strategies that allow companies to differentiate by tailoring their products to different groups' preferences. In our analysis, we were interested in using modeling to understand why different groups churn and how different factors influence churn rate. We believe that the information provided by our models can be used to support strategic marketing decisions of the company in the medium/long run.

Another important aspect in strategic decision making is provided by the concept of CLTV. Customer lifetime value is a prediction of the net profit of a particular customer during the future relationship with the company. In that sense, it is a good indicator of which customers are worth investing marketing efforts to retain and which are not. In our analysis, we use modelling to predict the CLTV of each customer.

We combine our predictive model for churn with our predictive model for CLTV, to provide a tool for the company to proactively identify customers to target their marketing efforts, in an attempt to not lose them in the future.

We believe that using analytical modeling will help telecommunications' companies ehnance their business model and marketing strategies and further differentiate from competitors.

2.2 Project goals

Considering what was already mentioned, our goals with this project are:

1. Understanding which factors influence churn rate for different groups.
2. Building a predictive model for churn rate that best identifies which customers are likely to churn.
3. Building a predictive model for CLTV that best identifies how much a customer is worth for the company and the factors that contribute most to CLTV.
4. Perform customer segmentation to identify high value customers that are likely to churn.

2.3 A Priori Expectations

We hypothesized different groups will have different churn rates and that that information might be useful for strategic decision making. A logistic regression model may be appropriate for predicting churn probability. We also hypothesize that a linear regression model may be useful for predicting CLTV based on demographic and product specific explanatory variables.

3 Methods

3.1 Description of the data

The IBM Business Analytics Community provides a fictional dataset of over 7,000 customers for a telecom company that contains information about which customers have left, stayed, or signed up for their service. The dataset also contains major demographic information for customers, along with Satisfaction Score, Churn Score, and Customer Lifetime Value (CLTV) index.

The database has data from 7,043 telecom customers, all located in California (USA). The average tenure of the customers is 32 months with an average churn score (determined by the company) of 59% and an average CLTV (determined by the company) of 4,400$.

Table 1: Overview of data

number observations(#)	7043
average tenure (months)	32
min tenure (months)	0
max tenure (months)	72
average churn score(%)	59
min churn score(%)	5
max churn score(%)	100
average CLTV($)	4400
min CLTV($)	2003
max CLTV($)	6500

From the customers, 5,174 have not churned (73.4%). We categorized the different reasons for churn that were provided in the feature "Churn Reason" in the database in 5 subcategories: competitors, customer service, produce features, price and others (see table 2). Interestingly, even though competitors and price play a big role, bad customer service was reported as being the second main reason for churn. Additionally, if we consider bad customer service and bad product features together, these two reasons had a more important role in customer churn than price and competition together.

Table 2: Top 5 churn reasons

Reason	# customers
Competitors offer	621
Customer service	455
Product features	381
Price	199
Other	59

3.2 Methods used

The implementation and transformations performed for each of the methods used are summarized in the tables below.

Table 3: Description of methods used for clustering and classification

Method	Response	Implementation	Transformations
Clustering	Churn reason	1. Calculate distance using Gower distance, 2. Choose a clustering algorithm - partitioning around medoids (PAM) algorithm, 3. Select the number of clusters, 4. Interpret the clusters	1. Transform categorical variables to dummy variables
Logistic Regression	Churn(Y/N)	1. Split data 75/25 into train and test sets, 2. Run Full model, 3. Aggregate data for categorical variables, 4. Check for model assumptions 5. Transform the variables and remove outliers, 6. Variable selection models, 7. Predict on test dataset, 8. Compute the confusion matrix for different threshold levels	1. Numerical variable to bins(categorical variable)
KNN	Churn(Y/N)	1. Transform categorical variables to dummy variables, 2. Use train.kknn function available in package CAR with multiple kernels, a maximum of 30k and as probability, 3. Cross validation to find optimal k and kernel, 4. Predict on test dataset, 5. Compute of the confusion matrix for different threshold levels	1. Transformation of categorical variables to dummy variables

For more information on the clustering model, please refer to attachment 6.4 - Clustering procedure.

Table 4: Description of methods used for regression

Method	Response	Implementation	Transformations
Linear Regression	CLTV value	1. Prepare Data (remove outliers and multicollinearity), 2. Split data 75/25 into train and test sets, 3. Perform variable selection (methods used: forward stepwise, lasso, elastic net), 4. Check for model assumptions 5. Perform transformations to improve goodness of fit, 6. Interpret coefficients, 7. Create gradient boosted model, 8. Performance measures in the testing dataset	1. Combine categorical variables exhibiting multicollinearity, 2. Logarithmic transformation of response, 3. Box-Cox transformation
Decision Tree	Churn(Y/N)	1. Split data 75/25 into train and test sets, 2. Build a decision tree to predict churn on training data, 3. Cross validation to find optimal tree length, 4. Prune tree to the optimal tree length for prediction, 5. Predict on test dataset, 6. Compute the confusion matrix for different threshold levels	1. Transformation of categorical variables to dummy variables
Random Forest	Churn(Y/N)	1. Create a random forest classifier to predict churn on the transformed data, 2. Use the random forest to make predictions and to obtain performace measures such as accuracy and out-of-bag error estimate.	1. Transformation of categorical variables to dummy variables

For more information on the gradiant booster model, please refer to attachment 6.5 - Gradient booster.

4 Results

4.1 Churn Rate - Insigths

4.1.1 Insigths from clustering

As we have seen before, there are different reasons why customers churn - either because customers provided a better offer, the product features were not aligned to customers' interests, the customer service was bad, among others.

From a marketing standpoint, it is interesting to understand if it is possible to group customers based on the reasons why they have churned. To do so, we have implemented a clustering algorithm. The results are presented in the graph and table below.

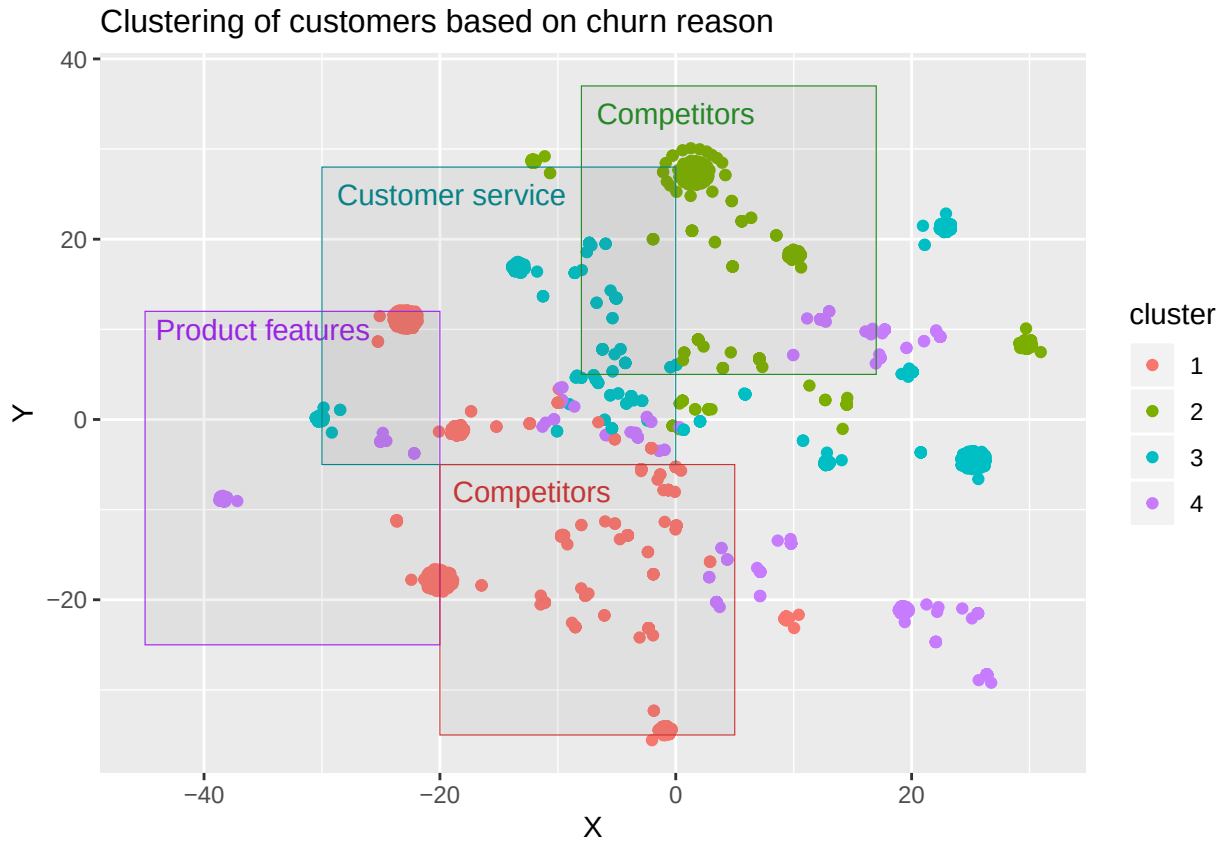


Table 5: Clustering segments

	Gender (F/M)	Age group	Partner	Dependents	Main Service	Main Reasons
cluster1	100% M	92%young	86%No	95%No	62% Fiber	49% competitors, 24% customer service
cluster2	100% F	80%young	100%No	96%No	69% Fiber	54% competitors, 23% product features
cluster3	88% F	86%young	67%Yes	91%No	72% Fiber	63% customer service, 16% competitors
cluster4	76% M	77%senior	73%Yes	95%No	81% Fiber	42% product features, 18% price

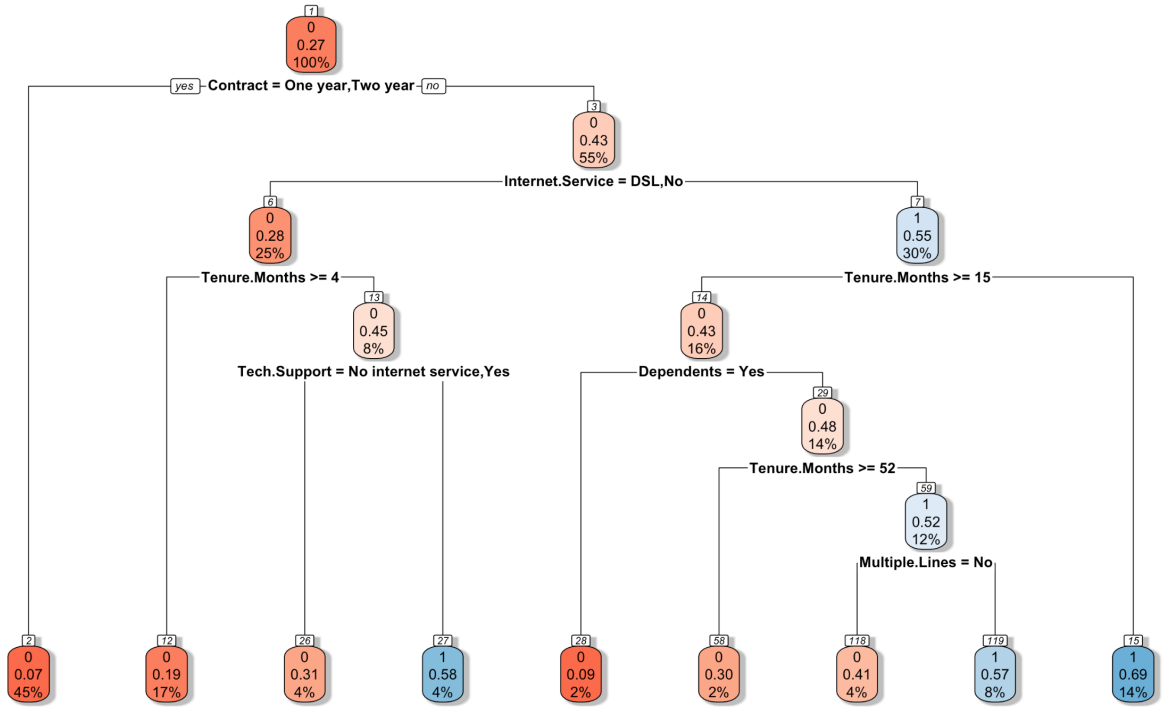
Understanding that different customer groups have different needs is important in defining the company's marketing strategy. One insigth from the clustering analysis is that, even though customer service is important for all groups, female customers might be interested in a more personalized interaction than the other groups.

Another important insigth from this clustering analysis is that the product features might be too complex

for senior citizens that do not have dependents to help them successfully use the products. To address this limitation, the company can provide better assistance not only in the moment of sale, but throughout the product lifetime. Additionally, the company can also chose to develop a more basic option that is easier to use for this group of customers.

4.1.2 Insights from decision tree model

To understand how the different probabilities of churn change for different customer and product features, we developed a decision tree model. The results are presented below. The first number in the node correspondes to the classification of the node (0 if not churn and 1 if churn). The second number in the node correspondes to the % of the customers on the other classification. The third value in the node measures the total % of customers that are included in that node.



The most important insigths are:

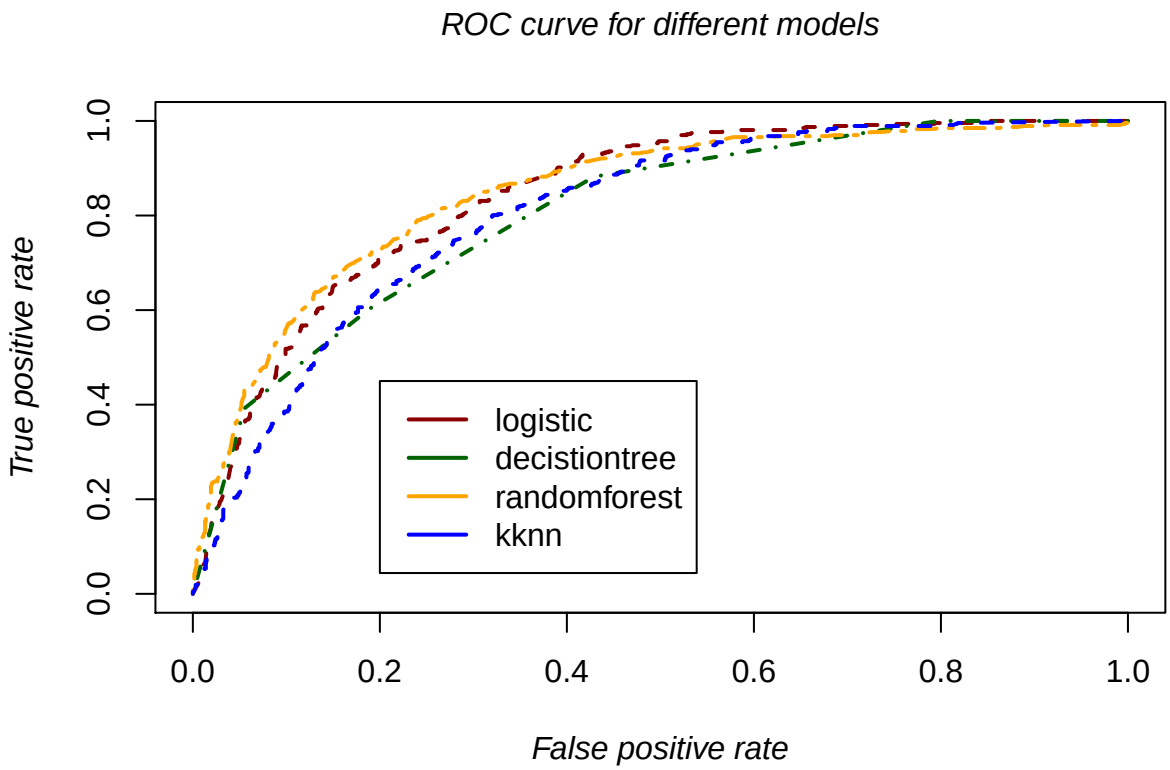
- The most important variable in determining churn rate is duration of contract. If the contract is 1 or 2 years the probability they will not churn is 93%. The probabilities of not churning are much lower if the contract is month-to-month. 45% of the total customers in the testing dataset fall in this category.

- If the customer has a month-to-month contract, has fiber optic, is in default for more than 15 months and has dependents, the probability it will churn is only 9%. Only 2% of customers are in this node.
- The higher churn occurs for month-to-month contracts, fiber optic, tenure higher than 15 months but lower than 52 months, no depedents and multiple lines. In that case, churn rate is 63%.
- Overall, the probabilities of churn are high for month-to-month contracts. The company can create incentives for customers to subscribe to longer contracts.

4.2 Churn Rate - Predictive model

We developed 4 classification models to predict churn. To measure the performance and choose the best model for prediction, we computed the ROC curve. The ROC curve provides one of the most important evaluation metrics for checking any classification model's performance. ROC is a probability curve and Area Under the Curve (AUC) represents degree or measure of separability. It tells how much model is capable of distinguishing between classes.

The results are presented below.



The higher the AUC, the better the model is at predicting churn as churn (true positive - TP) and not churn as not churn (true negative - TN). Looking at the plot above, we can see that random forest has a higher area under the curve for false positive (FP) values lower than 0.4 - in 40% of the cases we predicted as churn when the customer did not churn - and true positive rates lower than 0.8 (in 80% of the cases the customer churn when we predicted so). As the false positive rates increases, logistic model becomes better than random forest. Eventually KKNN and decision tree also become better than random forest for higher false positive rate.

The choice of threshold is cruccial in this case. The question we are trying to answer is:

- Do we want to identify more customers as likely to churn when in reality they will not churn, in order to not miss customers that will eventually churn

or

- Would we rather not identify some of the customers that are likely to churn in order to not incurr in uneccessary costs?

In our analysis we are combining a classification model with a model that computes CLTV. This allows us to filter which customers to target before actually incurring the costs associated with the extra efforts to retain the customers.

For that reason, we started by using a small threshold of 20%. This threshold implies that we are willing to take a lot of false positives, in order to have a high sensitivity rate (TP/(FN+TP)). Because we are getting a lot of false positives, our specificity (TN/(TN+FP)) will be low and our accuracy (TN+TP)/(TN+TP+FN+FP) will also be low.

Table 6: Results for a threshold of 0.2

	Accuracy	Sensitivity	Specificity
logistic	0.712	0.865	0.656
decision tree	0.654	0.882	0.571
random forest	0.557	0.966	0.410
KKNN	0.652	0.863	0.576

From the table, we can conclude random forest performs better for a low threshold meaning that it will be the best model predicting true positives. On the ohter hand it is the worst model at predicting false positives which makes it the worst model in terms of accuracy.

It is not just the cost of acquiring a customer that is relevant for the company. There are other costs associated with the process, such as the computational cost of predicting CLTV for each customer that is likely to churn. For that, we decided to measure the sensibility of the models to other threshold values - 0.5 and 0.6. The accuracy of the models for these levels are summarized in the table below.

Table 7: Accuracy for different thresholds

	0.2T	0.5T	0.6T
logistic	0.712	0.796	0.789
decision tree	0.654	0.798	0.798
random forest	0.557	0.807	0.801
KKNN	0.652	0.767	0.768

We can see that the accuracy of random forest is higher for higher thresholds. This happens because for higher thresholds, random forest is predicting less false positives than the other models.

The choice of the best model is highly dependent on the implementation cost and on on the threshold level. In this case, we chose random forest as the best model for the following reasons:

- We are assuming we will filter the data before targeting the customers so we would rather have higher false positives in order not lose potential good customers;
- Goodness-of-fit for logistic model is not ideal (see attachment)
- Random forest is easier to implement than decision tree and normally does not overfit as much.

4.3 CLTV - Linear Regression Model

We tested multiple linear regression models, including a model with the full set of predictors, a model selected using forward stepwise regression, a model selected using lasso regression, and a model selected using elastic net regression.

Table 8: Metrics for different linear regression models

	adj.rsq	Cp	AIC	BIC
full	0.1850975	23	86068.34	86225.43
step	0.1845654	8	86056.74	86115.65
lasso	0.1847821	13	86060.37	86152.00
elnet	0.1850975	23	86068.34	86225.43

The forward stepwise regression model has the best metrics for Mallow's CP, AIC, and BIC, and has only slightly smaller adjusted R^2 than the larger and more complex full model and elastic net model, so we will choose the forward stepwise regression model as our preferred model. Though this model is preferred, it was a somewhat poor fit; namely, the linearity, constant variance, and independence assumptions did not appear to hold. We performed logarithmic transformations on both predictors and the response variable and the Box-Cox transformation to improve goodness of fit. We eventually settled on a model with a log transformation of the response, with a Box-Cox transformation afterwards. This somewhat improved goodness of fit, and the visual analytics for goodness of fit for this model can be found in the appendix.

Table 9: Significant Predictors of CLTV

Predictor	Coefficient Value	Significance Level
Tenure Months	17.40518808	0.001
Total Charges	0.04729367	0.01
Device Protection	-94.06929302	0.01
Internet Service (Fiber Optic)	-63.92423944	0.1

We can see from the above table that Tenure Months, Total Charges, Device Protection, and Internet Service significantly explain CLTV at differing signficance levels. The Tenure Months coefficient can be interpreted to mean that for each additional month a customer is tenured, their lifetime value increases by ~17. The Total Charges coefficient can be interpreted to mean that for each additional dollar charge, a customer's lifetime value increases by ~0.05. The baseline for Device Protection is No protection, so that coefficient can be interpreted to mean that if a customer accepts device protection, their lifetime value decreases by ~94. The baseline for Internet Service is DSL, so that coefficient can be interpreted to mean that if a customer switches from DSL to fiber optic internet service, their lifetime value decreases by ~64. All interpretations with respect to a particular coefficient assume that all other predictors are held constant.

Table 10: Comparison of Metrics for Forward Stepwise Regression Model and Boosted Regression

	MSPE	MAE	MAPE	PM
stepwise	0.278	0.426	0.027	0.830
boostedTree	1019764.834	850.535	0.224	0.774

We can see from various prediction metrics that the linear regression model obtained from forward stepwise regression performs signficantly better than an out-of-the-box gradient boosted regression model (gradient boosted regression further explained in the appendix).

4.4 Customer segmentation

The goal of the telecommunication company to retain high-value customers can be achieved through customer segmentation. Based on the approach mentioned earlier, the telecommunication company should focus to increase engagement with customers that have the highest chances of churning, and also have the highest customer lifetime value.

The first step in this direction will be to identify customers that are likely to churn using the churn prediction classification model (discussed in section 4.2). This will provide the probability that a particular customer may churn for every active customer. Customers can be sorted based on their probability to churn to filter customers that are most likely to churn based on a chosen threshold.

The choice for this threshold is subjective. A lower threshold may give a higher number of false positives, which could be beneficial for the company to increase engagement with a larger number of customers that are even less likely to churn. This strategy would be helpful for companies that are expecting high competition and is willing to spend more on customer retention. Whereas, a higher threshold is expected to give a higher number of false negatives, where the company would miss to identify certain customers who are likely to churn. This strategy would be better if a company is trying to curb down retention expenses.

For demonstration, consider a customer base of 10 users with the following churn probabilities. Customers that are likely to churn can be identified based on a particular threshold (0.6 in this case) and are highlighted in red in Table 11 below.

Table 11: Sample Customer Churn Prediction

cust_id	prob_churn
1	**0.65**
2	0.10
3	**0.75**
4	**0.90**
5	0.55
6	0.25
7	0.20
8	0.56
9	**0.86**
10	**0.60**

Once a customer segmentation is obtained based if a customer is likely to churn or not, the second step is to segment them by their customer lifetime value (CLTV). This can be achieved by predicting their customer lifetime value (CLTV) using the linear regression model (discussed in section 4.3). The list of customers

that are likely to churn can then be sorted by their predicted customer lifetime value (CLTV). Based on the retention program and it's associated costs, the telecommunication company can choose to increase engagement with a customer that is likely to churn based on their predicted customer lifetime value. Ideally, high-value customers should be focused on first before other low-value customers.

Continuing the demonstration for a sample customer base in Table 11, the customer lifetime value (CLTV) is predicted for every customer that is likely to churn. The customers are then sorted based on their customer lifetime value (Table 12). In a hypothetical scenario that a company only has a budget to retain 3 customers, based on the table below, customers 3, 1 and 9 will be chosen for higher customer engagement towards retention.

Table 12: Customers likely to churn sorted by customer lifetime value (CLTV)

cust_id	prob_churn	CLTV
3	0.75	5789
1	0.65	4832
9	0.86	3022
4	0.90	2915
10	0.60	2454

5 Discussion

5.1 Subject Matter Implications

The implications of our project are very important for companies looking to retain important customers, not just telecommunications companies. The framework for our Linear Regression model shows that regression can be used to predict customer lifetime value using whatever data a company may have about a particular customer. The Logistic Regression section shows that companies could also produce classification models to predict the probability of a customer leaving. Both of these pieces of information are valuable on their own, but combining them could add even more value to the company. For example, a company could use classification to determine which customers are most likely to churn, then sort those customers by lifetime value (found from the other model) to determine the high-priority customers to reach out to. This information helps the company focus its retention efforts and stay profitable because the most valuable customers are retained.

5.2 Limitations and next steps

5.2.1 Limitations:

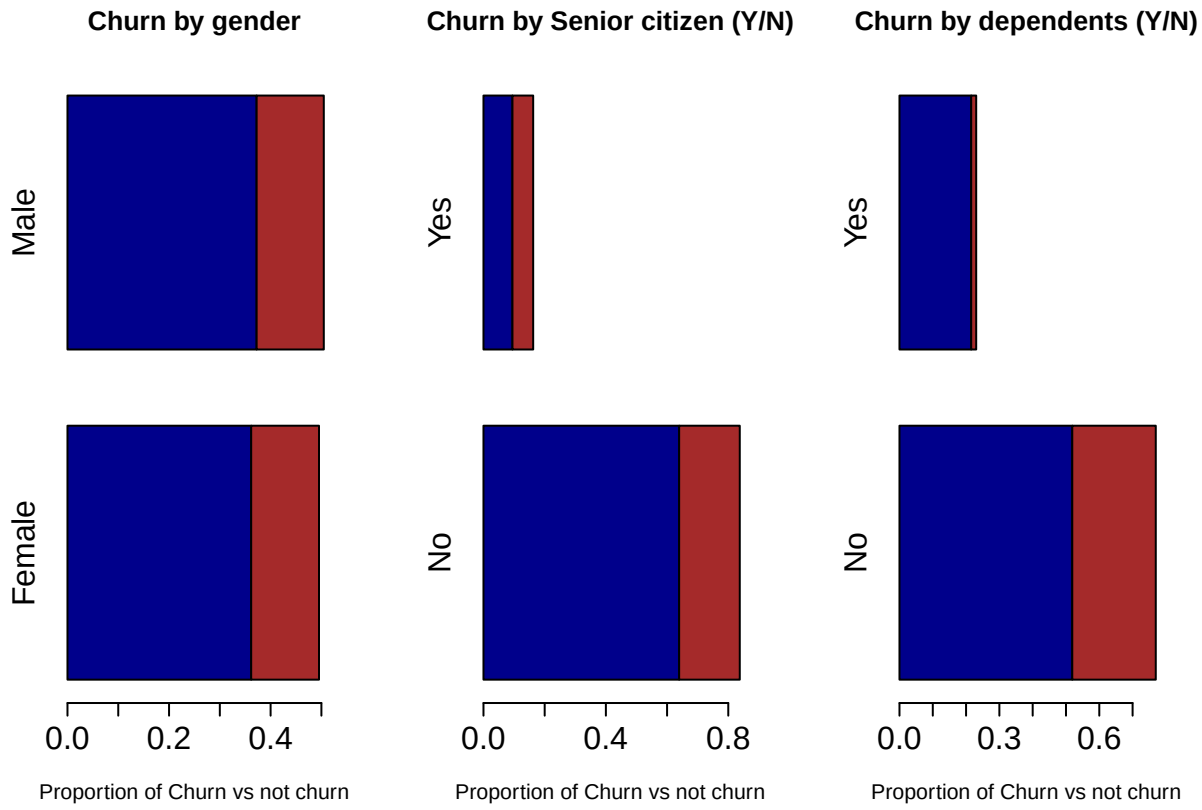
- The linearity, constant variance and normality assumptions were somewhat violated in the linear regression model, even after logarithmic and box-cox transformations. (see appendix 6.3 for visualizing the goodness of fit)
- The hyperparameters for the gradient boosted regression model was not tuned; default values were used.
- The linear regression model had a relatively low R^2 value (~.18), but this may be expected when working with real-life data.
- Not all the assumptions were met when doing the goodness of fit for logistic regression. Namely, there were discrepancies from a normal distribution (see appendix 6.2 for visualizing the goodness of fit).
- The time dedicated to prune the parameters in the random forest and kknn algorithms was limited.
- The choice of the best model to predict churn depends on a cost function defined by the company. The sensibility to different threshold values is subjective and involves decision making.

5.3 Next steps:

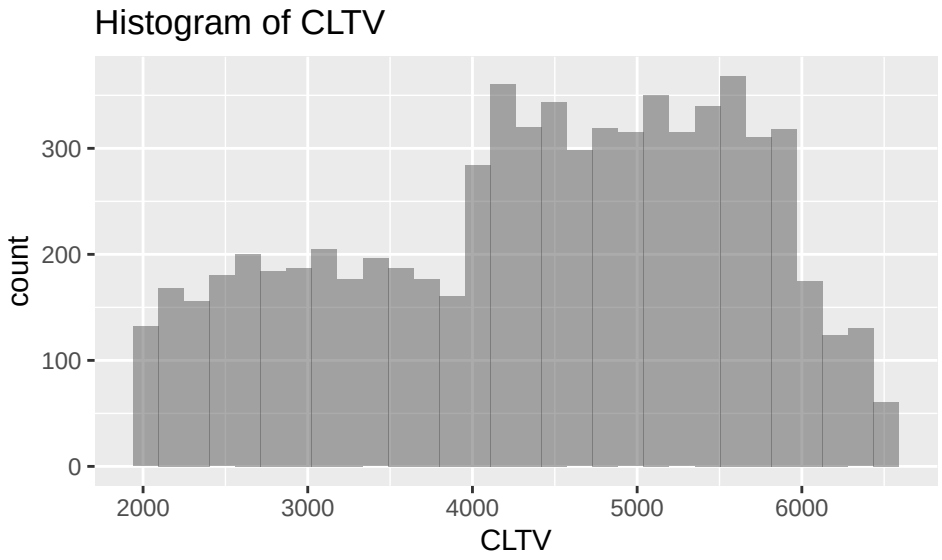
- Try additional transformations to improve goodness of fit for linear regression model.
- Spend more time finding models that fit well to the shape of our dataset.
- Look for other predictors to include to increase R^2 of the linear regression model.
- Improve goodness of fit for logistic model.
- Test different tunning techniques for the models.

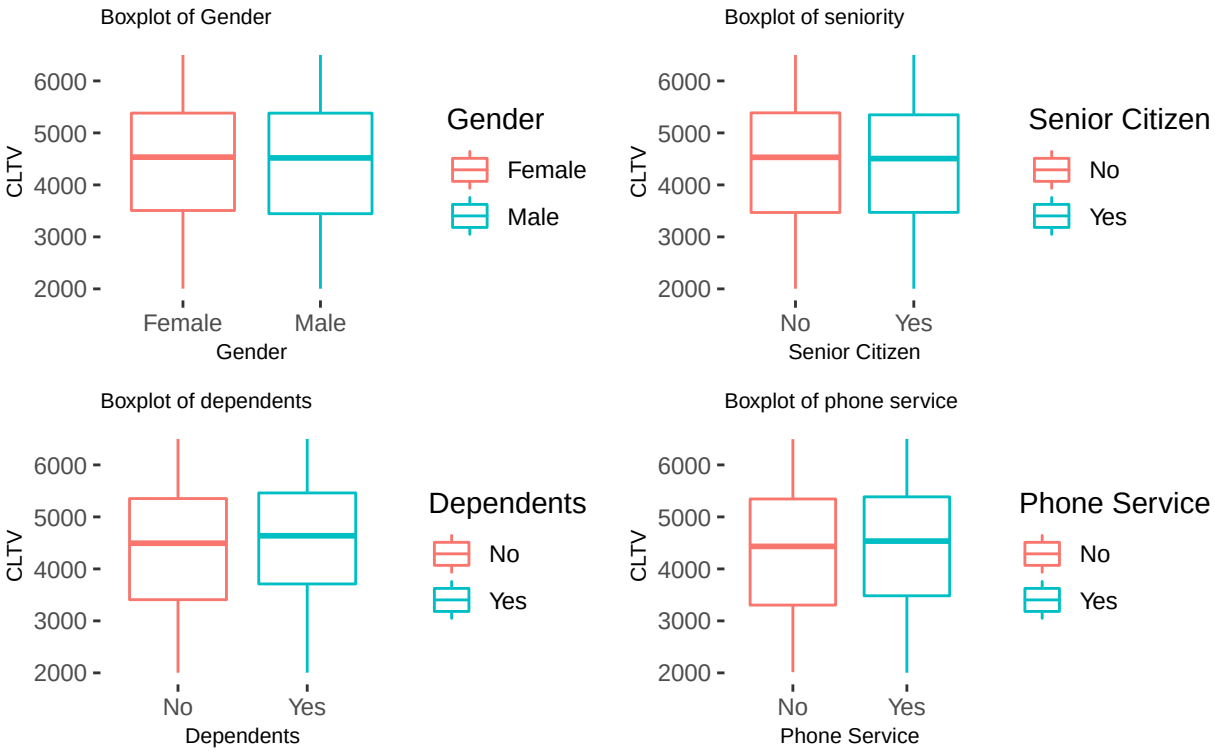
6 Appendix

6.1 Exploratory Data Anaysis

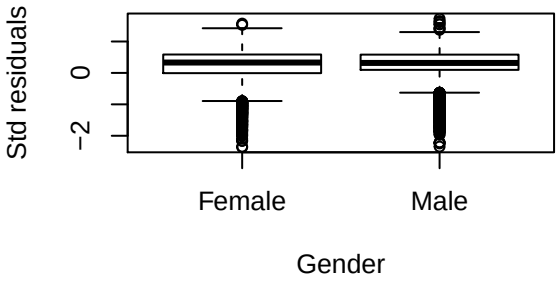


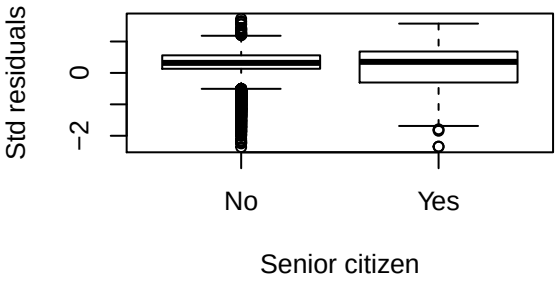

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

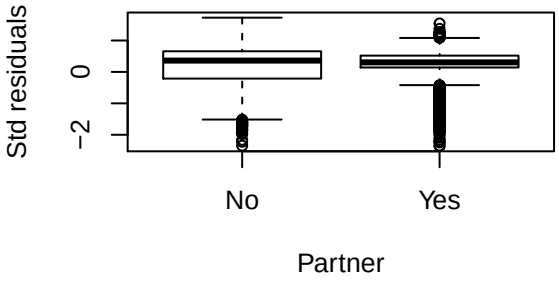


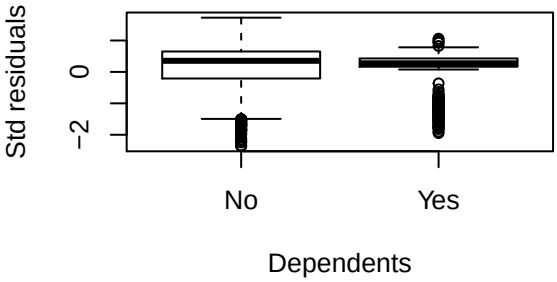

Boxplot of Gender
CLTV
6000
5000
4000
3000
2000
Female
Male
Gender
Gender
Female
Male
Boxplot of seniority
CLTV
6000
5000
4000
3000
2000
No
Yes
Senior Citizen
Senior Citizen
No
Yes
Boxplot of dependents
CLTV
6000
5000
4000
3000
2000
No
Yes
Dependents
Dependents
No
Yes
Boxplot of phone service
CLTV
6000
5000
4000
3000
2000
No
Yes
Phone Service
Phone Service
No
Yes

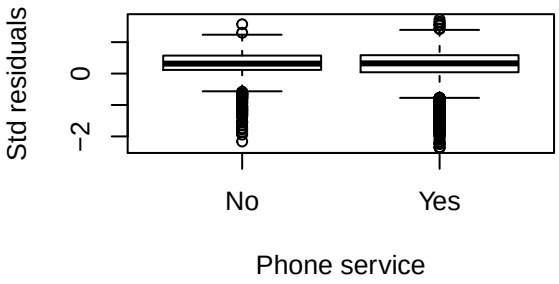
6.2 Goodness of Fit for Logistic regression

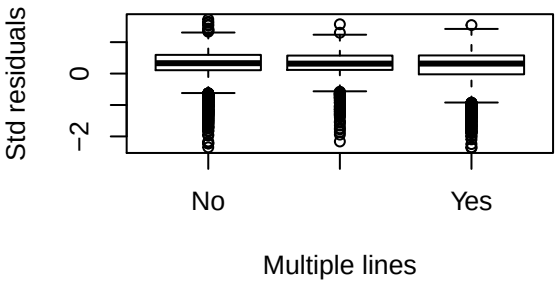


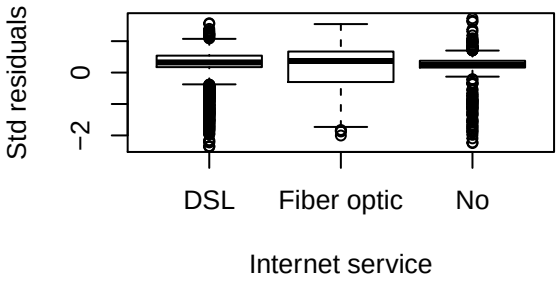


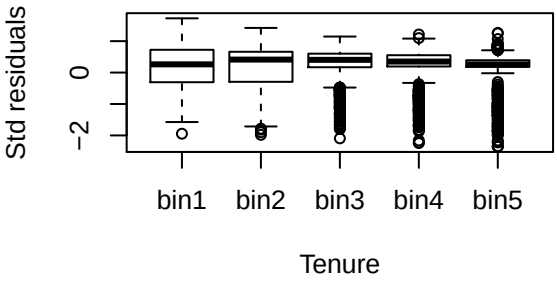


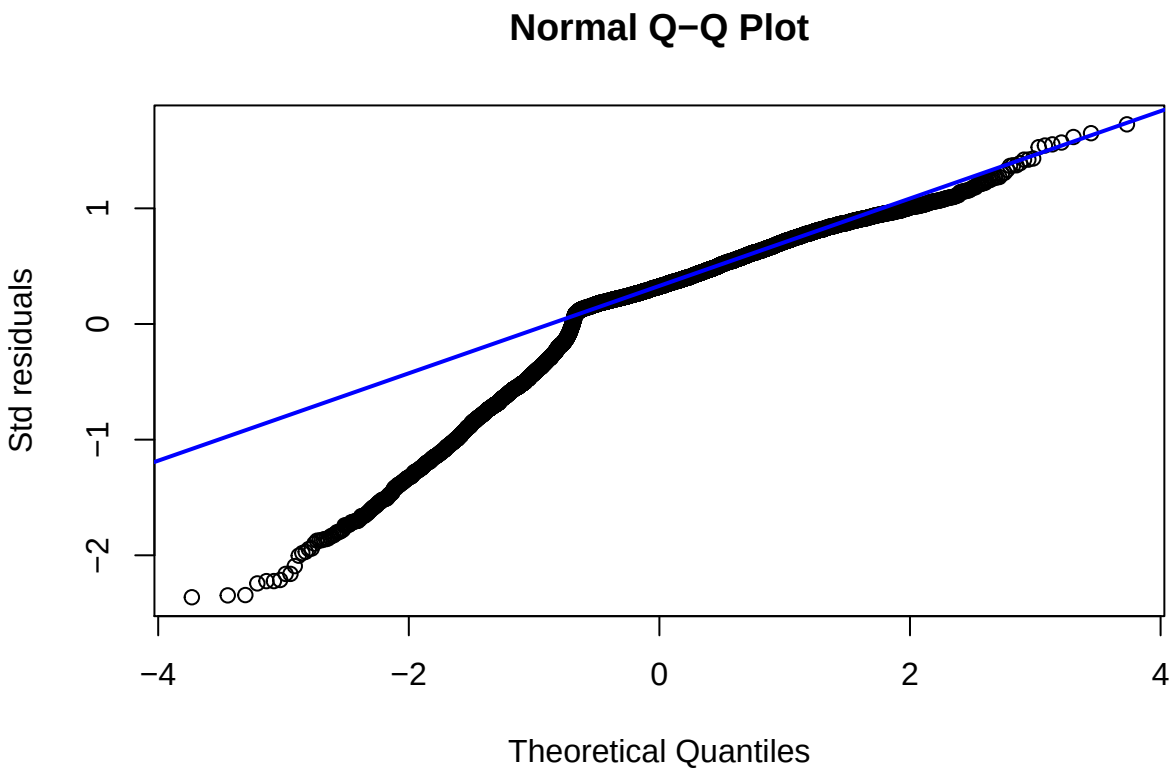






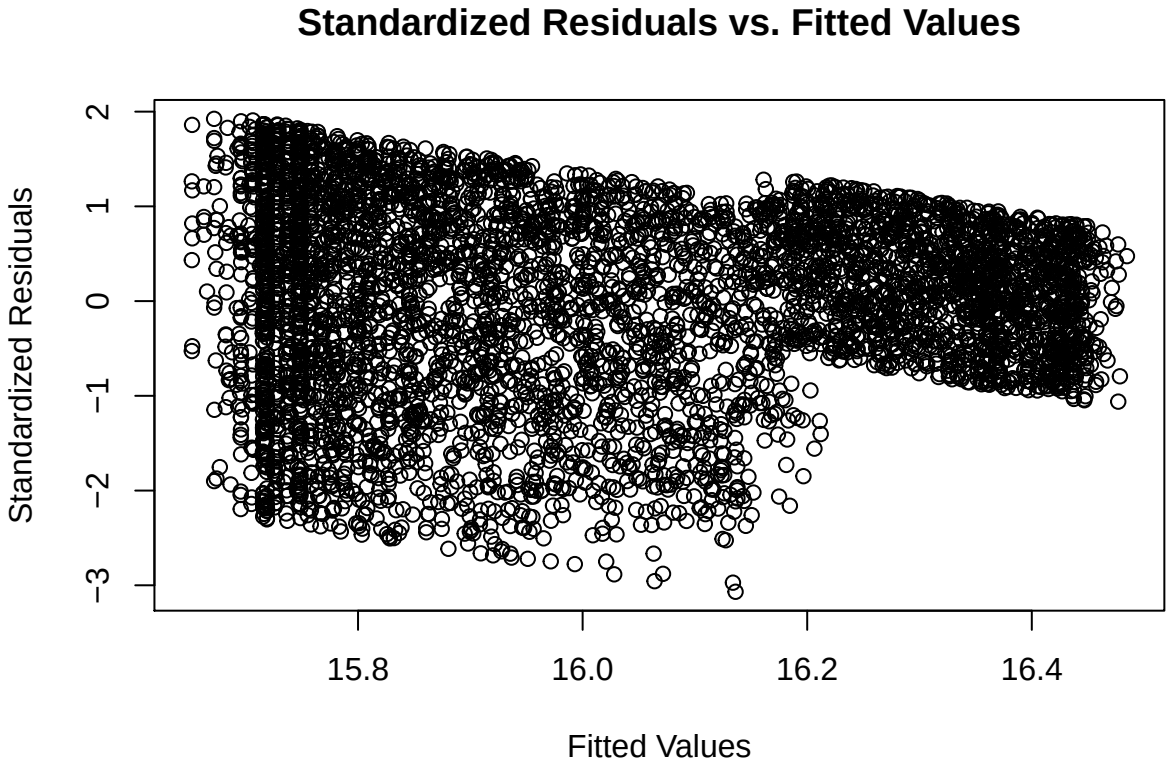




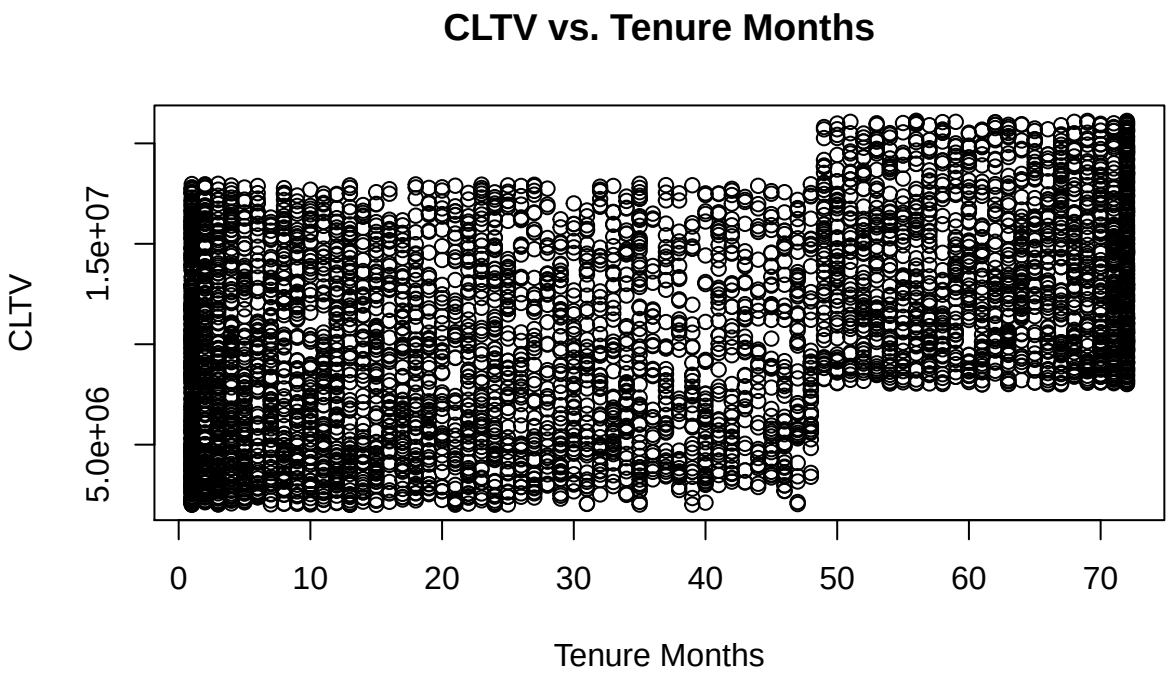


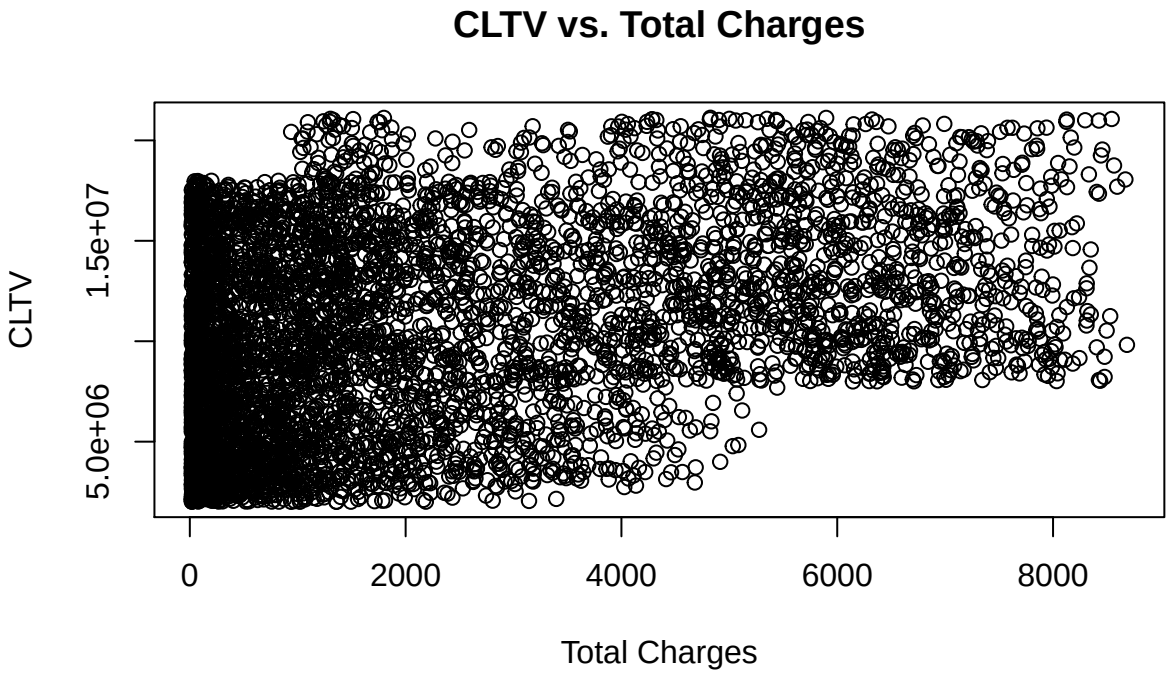
The QQplot shows that the normality assumption does not hold as there is a heavy tale on the rigth side of distribution.

6.3 Goodness of Fit for Forward Stepwise Linear Regression Model

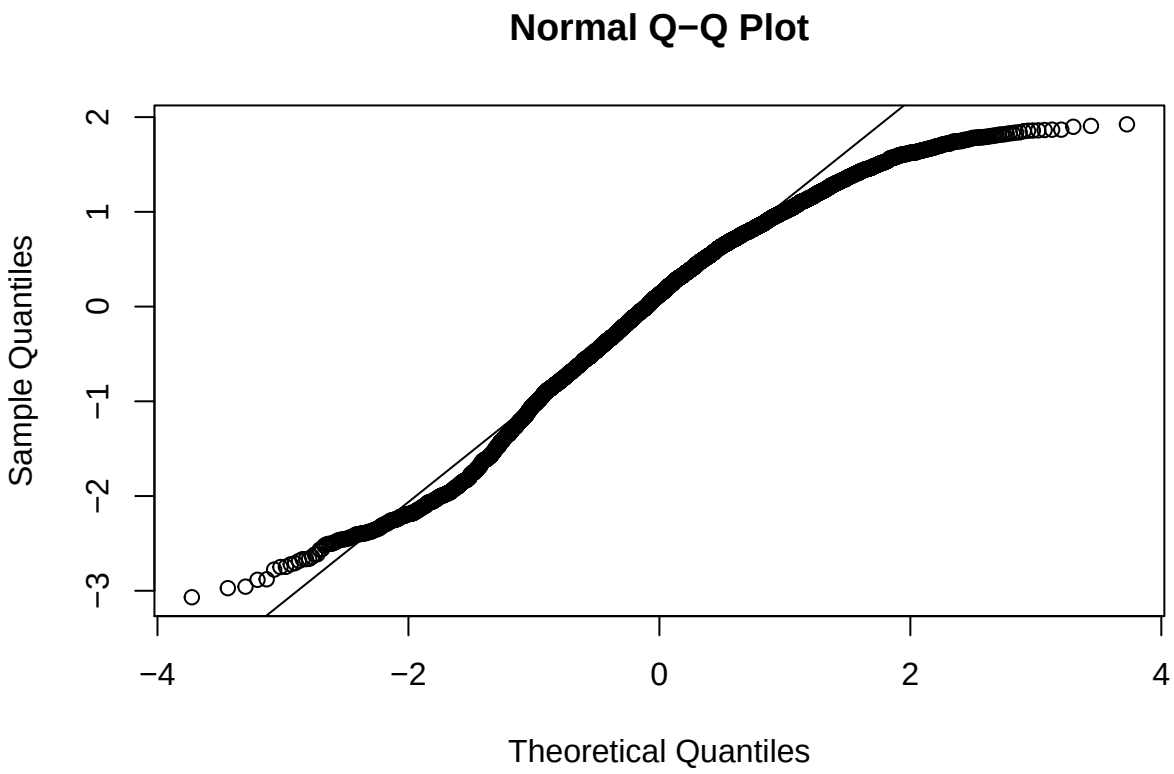


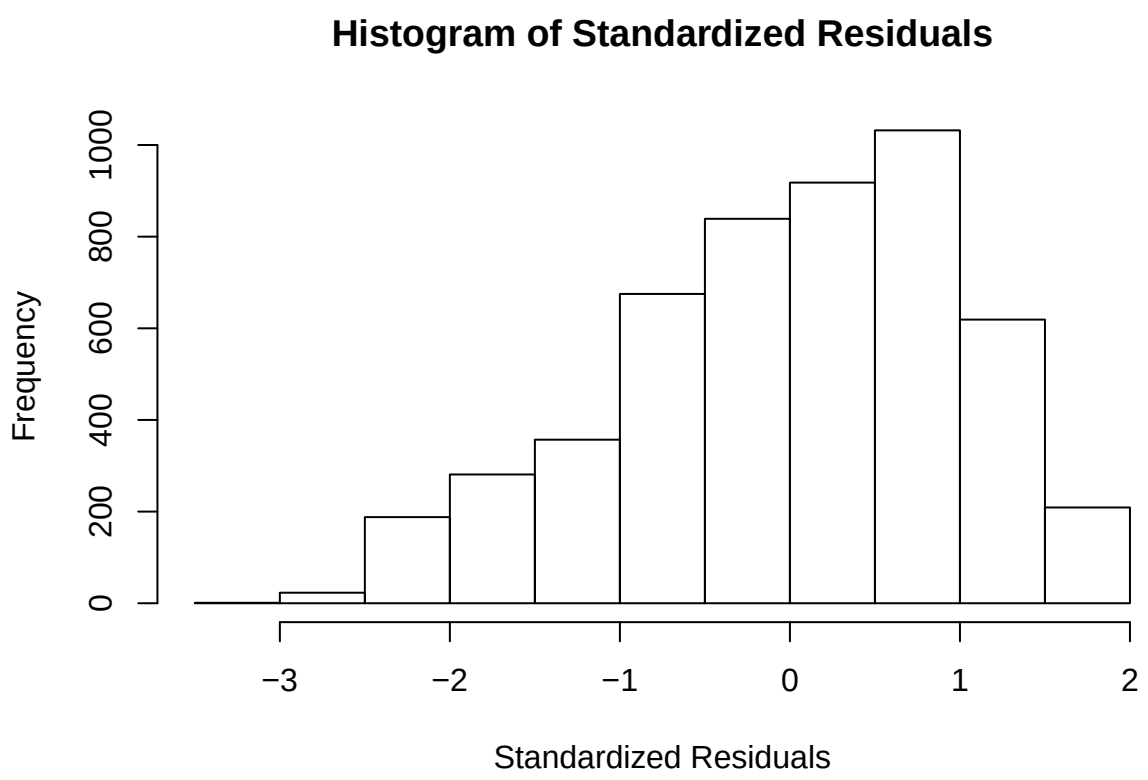
This plot of the the standardized residuals vs the fitted values shows that the independence assumption likely holds, but the constant variance assumption is likely violated because there is a downward trend in the standardized residuals as the fitted values increase.





The two plots above show that the linearity assumption does not hold; the response values (CLTV) are dispersed widely when plotted against both Tenure Months and Total Charges. Note that CLTV values are so large here because of the Box-Cox Transformation.


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−4
−2
0
2
4
−3
−2
−1
0
1
2



The QQ-Plot and histogram of standardized residuals above show that the normality assumption may hold, but that the data has heavy tails. In addition, the histogram of standardized residuals shows a slight leftward skew.

6.4 Clustering Algorithm

Gower distance

For each variable type, a particular distance metric that works well for that type is used and scaled to fall between 0 and 1. Then, a linear combination using user-specified weights (most simply an average) is calculated to create the final distance matrix. The metrics used for each data type are described below:

- quantitative (interval): range-normalized Manhattan distance
- ordinal: variable is first ranked, then Manhattan distance is used with a special adjustment for ties
- categorical: variables of k categories are first converted into k binary columns and then the Dice coefficient is used

Clustering algorithm

We used the partitioning around medoids (PAM) algorithm.

Partitioning around medoids is an iterative clustering procedure with the following steps:

Choose k random entities to become the medoids Assign every entity to its closest medoid (using our custom distance matrix in this case) For each cluster, identify the observation that would yield the lowest average distance if it were to be re-assigned as the medoid. If so, make this observation the new medoid. If at least one medoid has changed, return to step 2. Otherwise, end the algorithm.

Selecting number of clusters

We used silhouette width, an internal validation metric which is an aggregated measure of how similar an observation is to its own cluster compared its closest neighboring cluster. The metric can range from -1 to 1, where higher values are better. After calculating silhouette width for clusters ranging from 2 to 10 for the PAM algorithm, we see that 4 clusters yields the highest value.

Source: https://dpmartin42.github.io/posts/r/cluster-mixed-types

6.5 Gradient Boosted Models

We did not go over boosted models in class, but they can be sometimes useful when dealing with complex non-linear patterns in data. Boosted regression trees partition data at each split and determine the residuals for each partition. The children of each branch will be fitted to those residuals and successive splits aim to find partitions that further reduce error. However, this class of models is black-box, so we lose interpretability. In addition, these models can sometimes severely overfit to the training data, which could explain some of the poor prediction results we got from the boosted tree in our analysis.

Source: StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/.