6414 Group Project - Telecom Customer Churn Modeling

Jared Babcock, Rishi Bubna, Marta Bras

2019-12-04

Contents

1 Abstract

The telecommunications industry is facing a challenging context, with the emergence of new technologies and the increasing levels of competition resulting in unprecedent levels of customer attrition and price comptetion 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 cruccial 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 this model to also predict which customers are likely to churn in the future.

Additionally, we use modelling to predict each customer's lifetime value (CLV), 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 CLV), 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 competiton 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 modelling to understand how different groups churn and how different factors influence churn rate. We believe 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 CLV. 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 them and which are not.In our analysis, we use modelling to predict the CLV for each customer.

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

We believe that using analytical modelling 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. Building a predictive model for churn rate that best identifies which customers are likely to churn.
2. 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.
3. 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. We also hypothesize that it is possible to predict CLV 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 3). 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 table below.

```
## Warning in table_info$align_vector[column] <-
## unlist(lapply(table_info$align_vector_origin[column], : number of items to
## replace is not a multiple of replacement length
```

Table 3: Methods

method	response	implementation	transformations
Linear Regression	CLTV value	1. Splitting the data int training/testing datasets, 2. Run Full model, 3. Check for model assumptions 4. Transform the variables and remove outliers, 5. Variable selection models, 6. Performance measures in the testing dataset	1. Numerical variable to bins(categorical variable)
Logistic Regression	Churn(Y/N)	1. Splitting the data int training/testing datasets, 2. Run Full model, 3. Aggregating data for categorical variables, 4. Check for model assumptions 5. Transform the variables and remove outliers, 6. Variable selection models, 7. Performance measures in the testing dataset	1. Numerical variable to bins(categorical variable)

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.

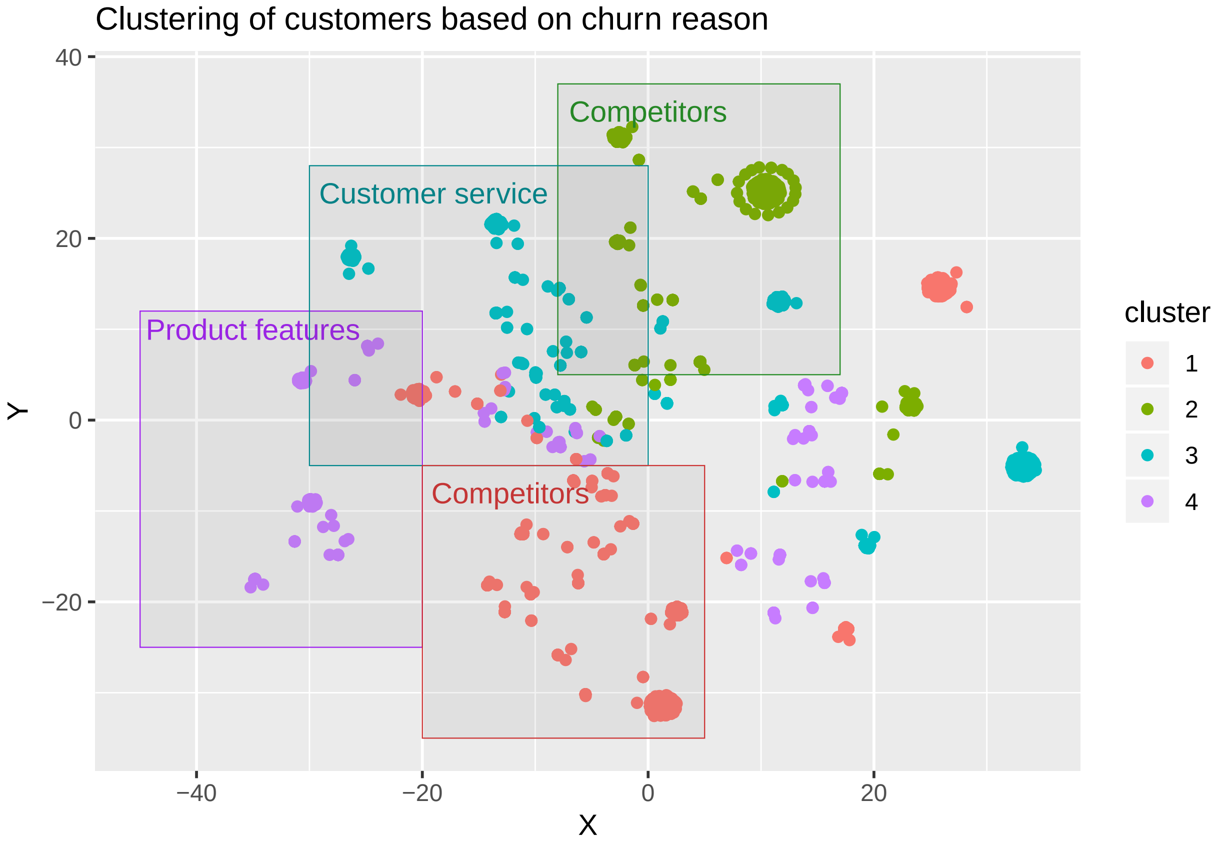


Table 4: Clustering segments

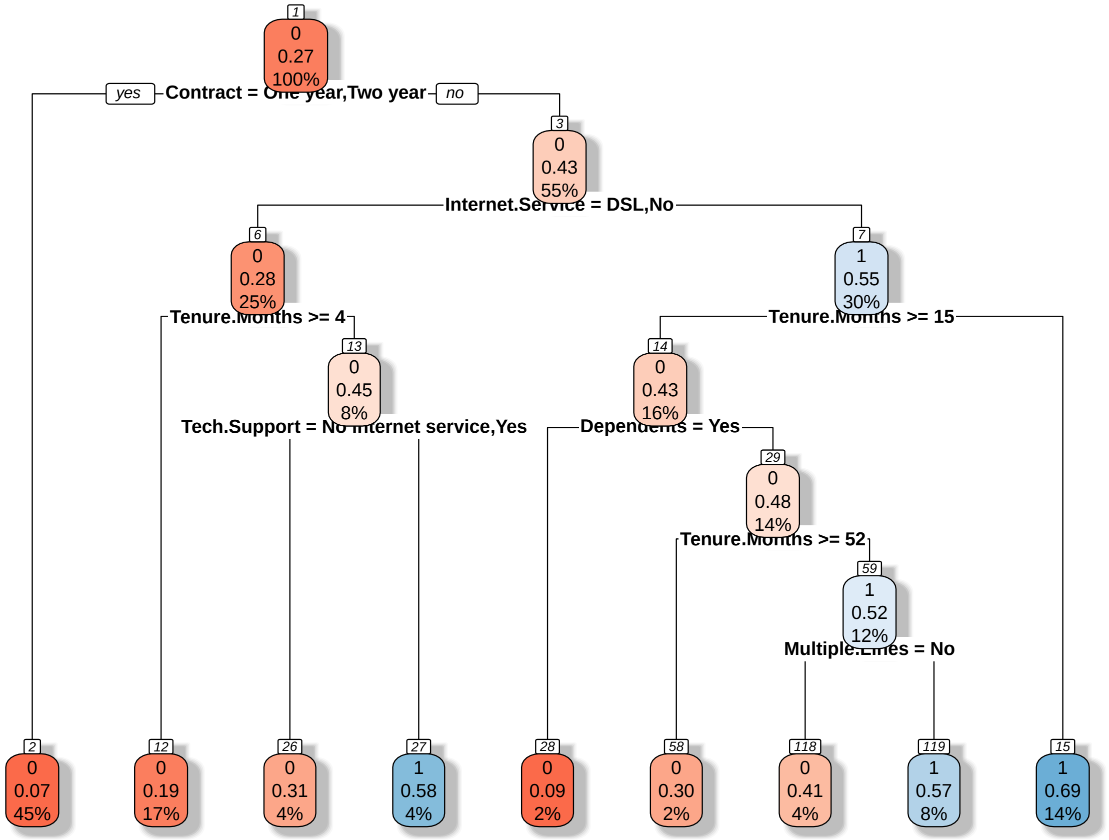
	Gender (F/M)	Age group	Partner	Dependents	Main Service	Main Reasons
cluster1	100% M	92%young	86%No	95%No	62% Fiber	49% competitors, 24% customer service
cluster2	100% F	80%young	100%No	96%No	69% Fiber	54% competitors, 23% product features
cluster3	88% F	86%young	67%Yes	91%No	72% Fiber	63% customer service,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



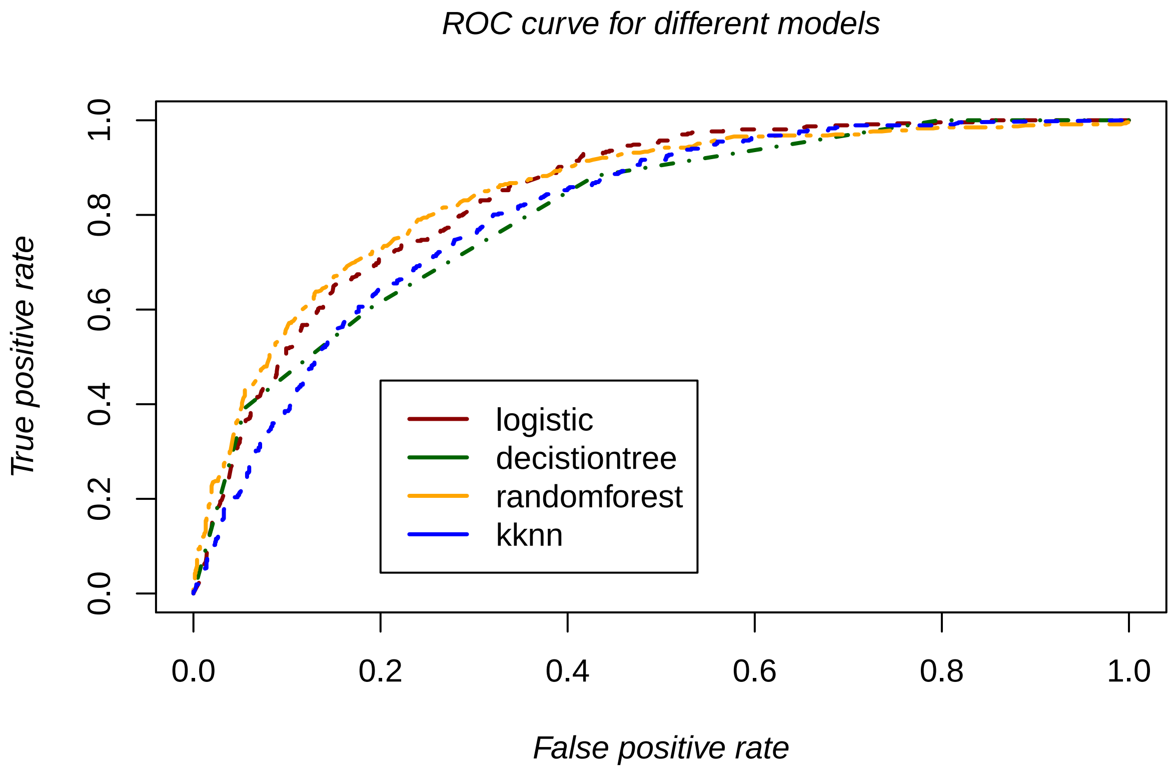
The most important insigths are:

- 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 chose the best model for prediction, we computed the ROC curve as it is one of the most important evaluation metrics for checking any classification model's performance. ROC is a probability curve and 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) and not churn as not churn (true negative). Looking at the plot above, we can see that random forest has a higher under the curve for false positive rate 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 CLV. 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 0.2%. 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 5: 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

N

Table 6: Results 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

4.3 CLTV - Linear Regression Model

Table 7: 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.

Table 8: 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 9: 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 the out-of-the-box gradient boosted regression model.

4.4 Customer segmentation

5 Discussion

5.1 Subject Matter Implications

5.2 Limitations and next steps

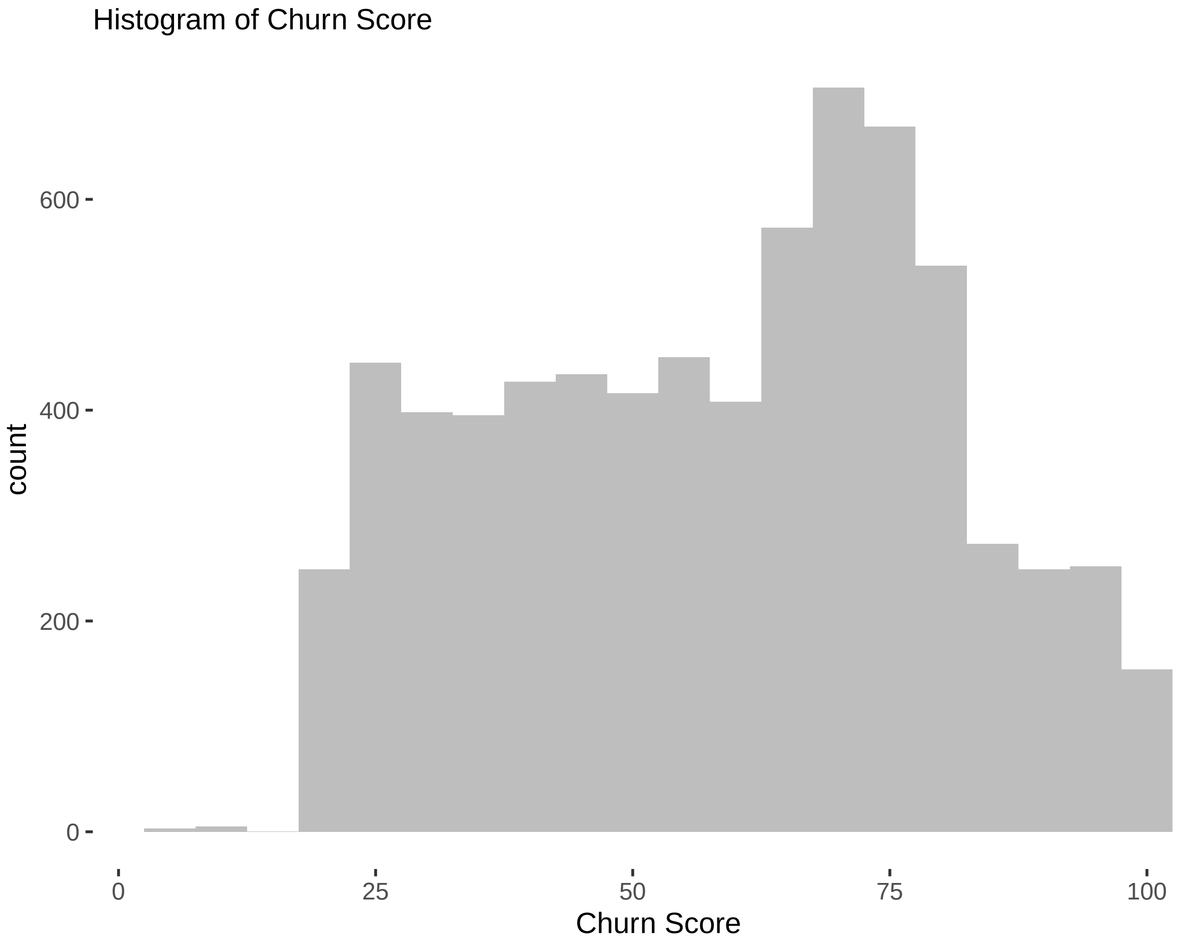
5.2.1 Limitations:

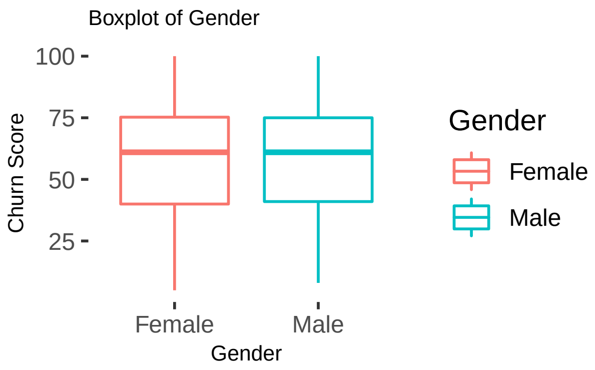
- Not all the assumptions were met when doing the goodness of fit for logisitc regression. Namely, there were discrepancies from a normal distribution.
- The time dedicated to prune the parameters in the random forest and kknn algorithms was limited.

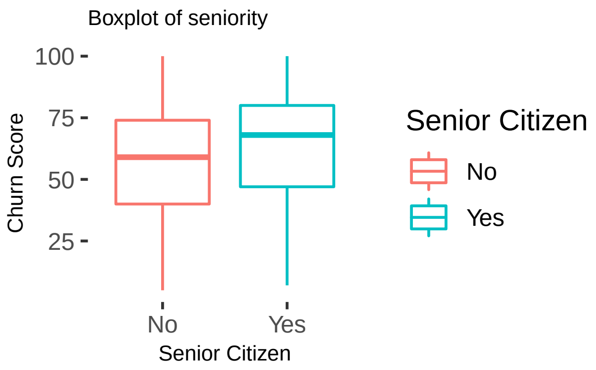
5.3 Next steps:

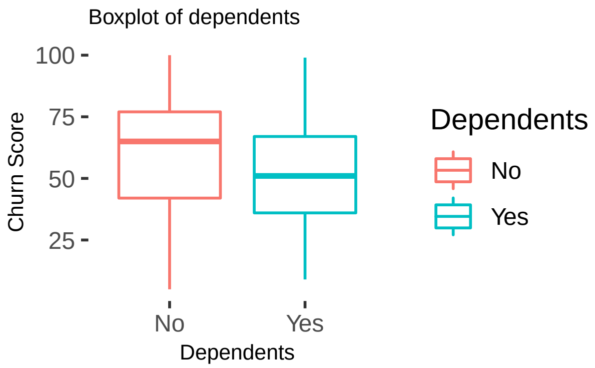
- Improve goodness of fit for logistic model.
- Test different tunning techniques for the models.

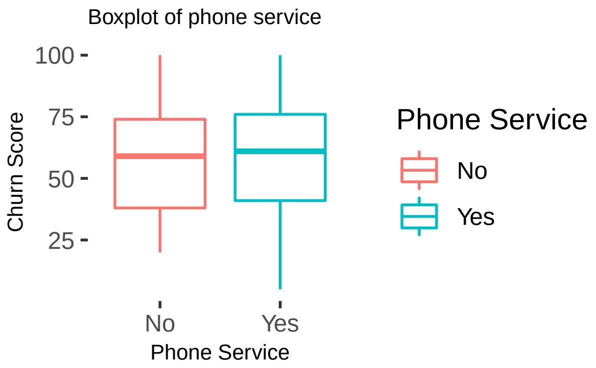
6 Attachments











6.1 Attachment I - Churn Rate

6.1.1 1. Full model

6.1.2 2. Model fit

6.1.3 3. Variable transformation

6.1.4 4. Re-running the model

6.1.5 5. Variable selection

6.1.6 6. Model selection