MQM Stats Final Project: Identifying At-risk Customers

Team43: Yvonne Xie / Mingyu Gu / Rachel Zhong / Nisha "Mint" Tantivess / Aritra "Auri" Shome

Introduction

MQ&M is a leading company in the telecom industry. Its marketing department wants to identify customers who are at risk of leaving their services. The dataset that we have provided collects attributes of MQ&M's customers, including their demographic information (e.g. gender, age, if they have partners and dependents), account information (e.g. how long they've been a customer, contract, payment method, monthly and total charges), services that each customer signed up for (e.g. phone, Internet, tech support, online backup), and finally a column called *Churn* that describes whether a customer has left in the past month or not. There are 7,043 rows in the dataset, with each row representing a record for a single customer. There are 21 columns, which include both continuous and categorical variables. We retrieved the data from Kaggle.

Business Context

We intend to use this dataset to identify at risk customers in the telecommunication industry – customers who are likely to cancel their service with their current telecom company. The U.S. Telecom industry is highly competitive. As the market matured, it is hard for companies to gain access to new untapped customers. Therefore, they compete aggressively to gain their peers' market shares in their customer segments. In addition, it is found that acquiring a new customer can cost five times as much than keeping an existing customer. Therefore, we would like to build a statistical model from the dataset to understand what will most likely cause a customer to cancel their service and predict whether a customer will leave or not. From there, MQ&M can identify their at-risk customers and come up with strategies targeting these customers to keep them from leaving.

Data Cleaning

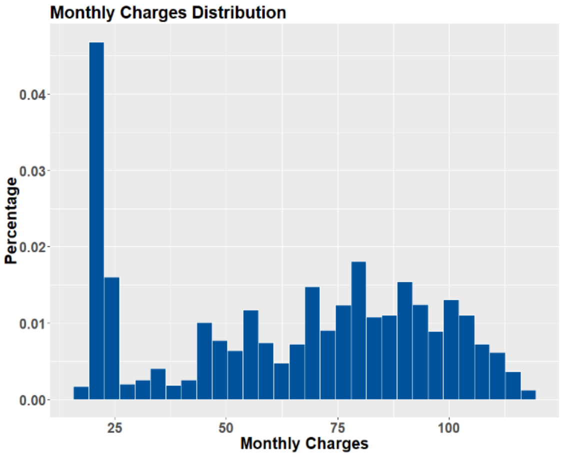
The first step in our data cleaning process involves checking for missing values. We found that *TotalCharges* is the only column with missing values, with 11 NAs. Instead of deleting them immediately, we investigated the 11 records with missing *TotalCharges* and found that these are new customers who have just signed up for their service. They have 0 as *tenure* and have not churned out yet obviously. Therefore, we chose to replace their *TotalCharges* with their *MonthlyCharges*. After the replacement, we run the summary of data again to confirm that there are no missing values for *TotalCharges* now. We also confirm that *customerID* is the primary key, which means that there is no duplicate of customer record in the dataset.

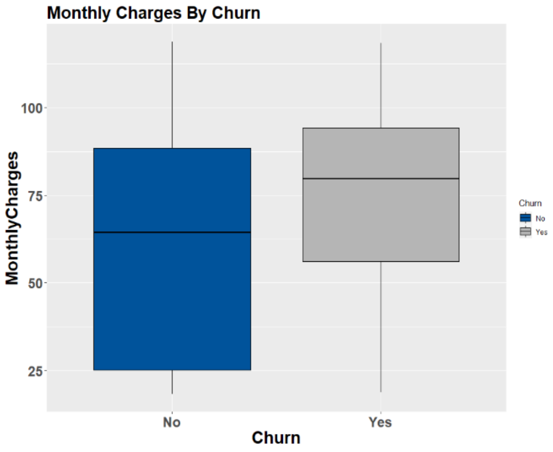
Exploratory Analysis

First, we took a quick look at the dataset and found that 73.46% of the customers have not churned out yet. For the next step, we want to examine how continuous variables (*MonthlyCharges*, *TotalCharges*, and *tenure*) affect customer churn rate. Below are the results and interpretations of our exploratory analysis:

Continuous Variable #1: Monthly Charges

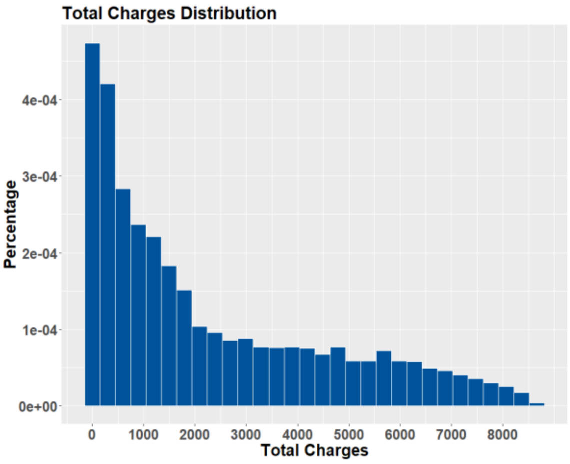
The histogram on the left shows that the most popular monthly plan is around $24. From the boxplot on the right, we can see that customers who have churned paid significantly more than those who did not.

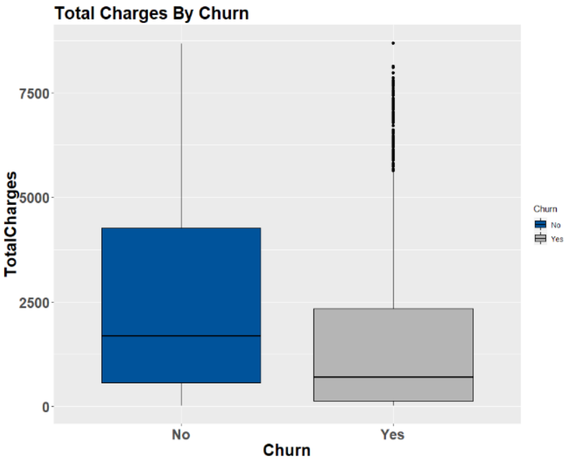




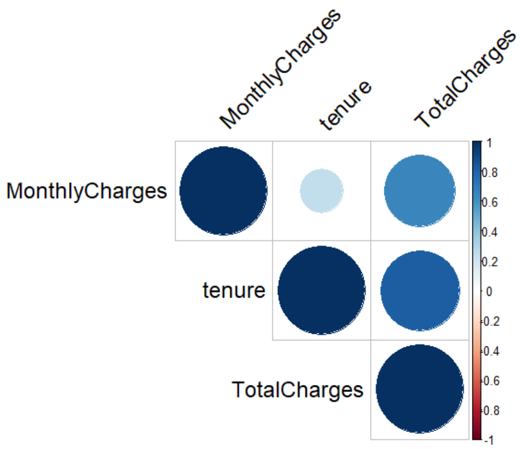
Continuous Variable #2: Total Charges

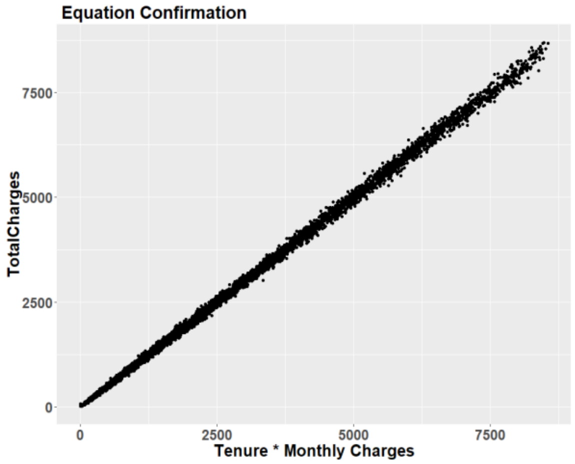
We plotted *TotalCharges* against *Churn* (Yes or No) and found out that customers who have churned out have lower total charges compared to those that did not. *TotalCharges* is determined by two factors – *MonthlyCharges* and *tenure*. We assume customers who have not churned out have higher *TotalCharges* because they stay in the contract for longer period of time (and therefore longer tenure).





To determine the relationship amongst the three variables, we ran a correlation matrix:



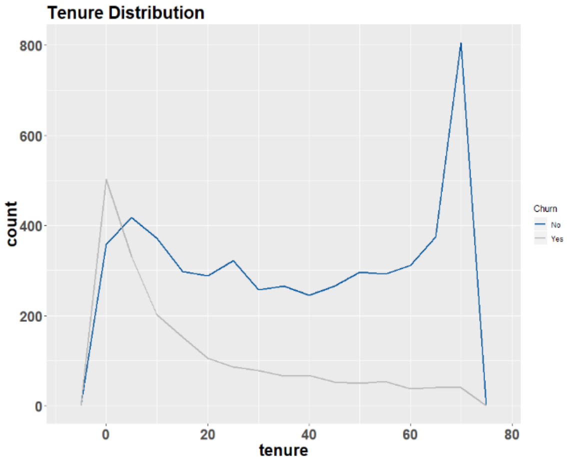


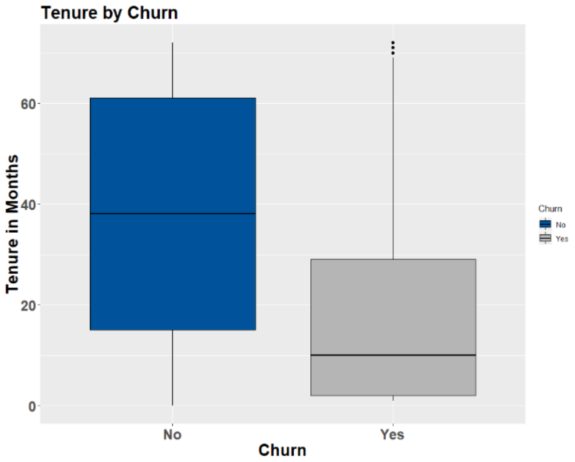
The graph indicates that the correlation (positive) is highest between *tenure* and *TotalCharges*, followed by *MonthlyCharges* and *TotalCharges*. There is no correlation between *MonthlyCharges* and *Tenure*. Therefore, we hypothesize that:

$$TotalCharges = Tenure \times\ MonthlyCharges$$

Thus, from the plot we can confirm that the equation applies to most of the data.

Continuous Variable #3: Tenure

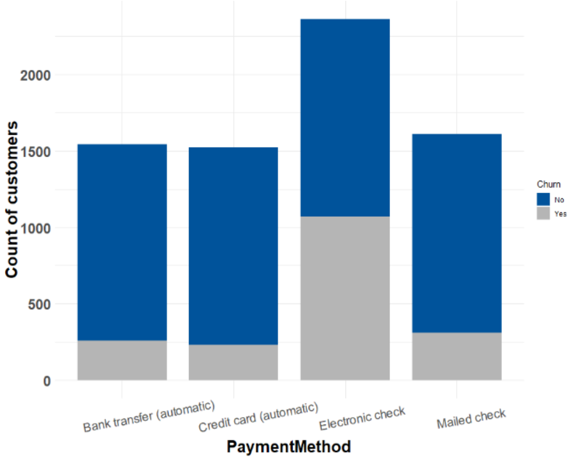


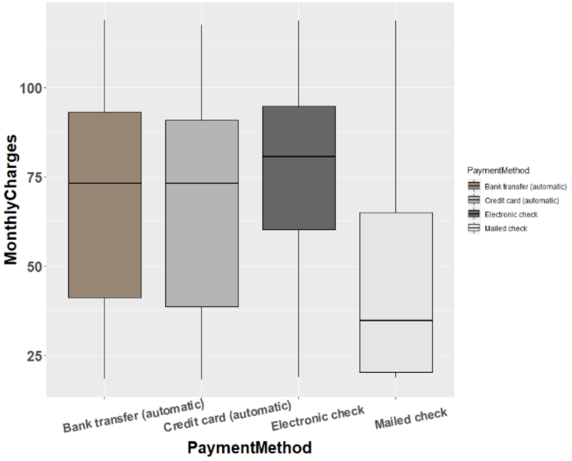


This plot shows again that customers who are of longer tenure are less likely to churn.

Next, we look at the count of customers and respective churn rates across categorical dimensions Payment Method, Contract Type and Internet Service, comparing these dimensions against monthly charges to see whether the pattern we discovered holds true across all dimensions.

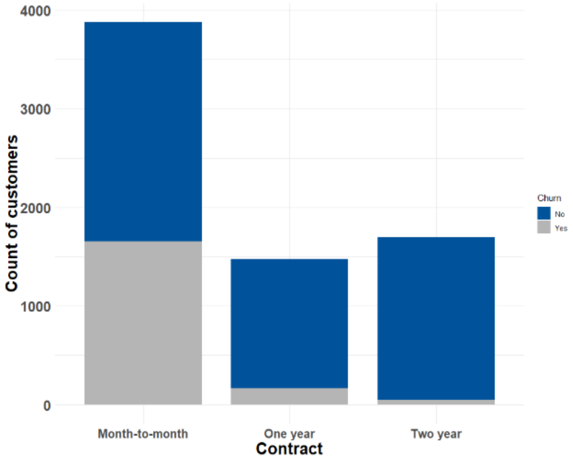
Categorical Variable #1: Payment Method:

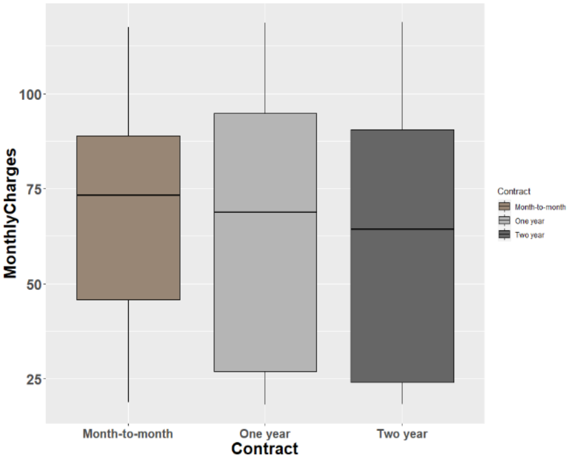




It is worrying that the payment method with the highest customer count (electronic check) also has the highest churn rates.

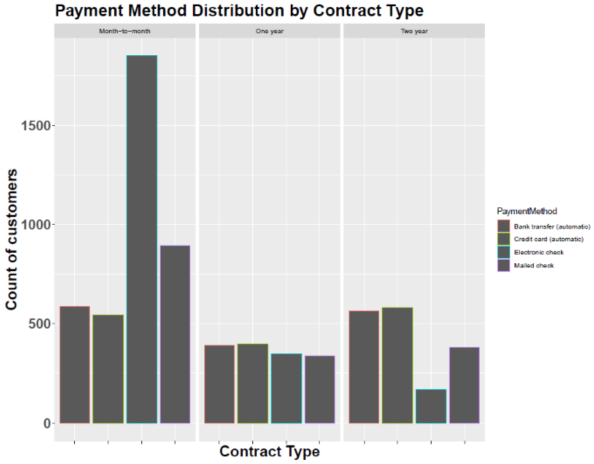
Categorical Variable #2: Contract Type

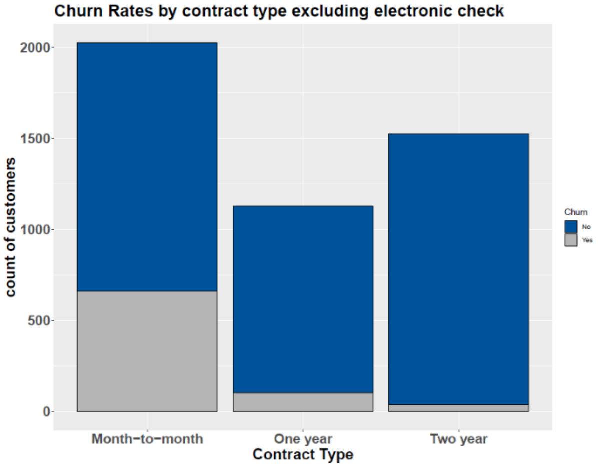




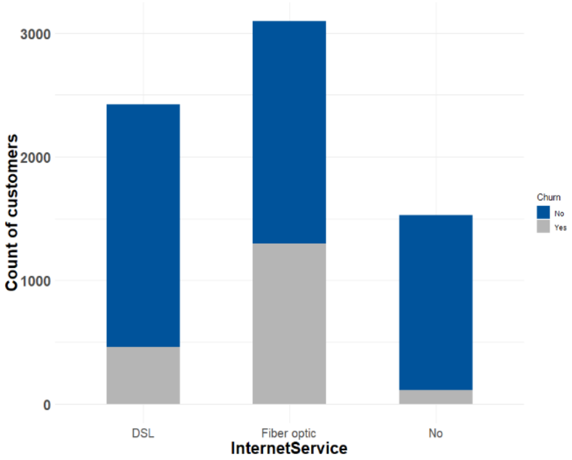
The average *MonthlyCharges* across all contract types do not have a huge discrepancy. However, customers with month-to-month contract type have significantly higher churn rate.

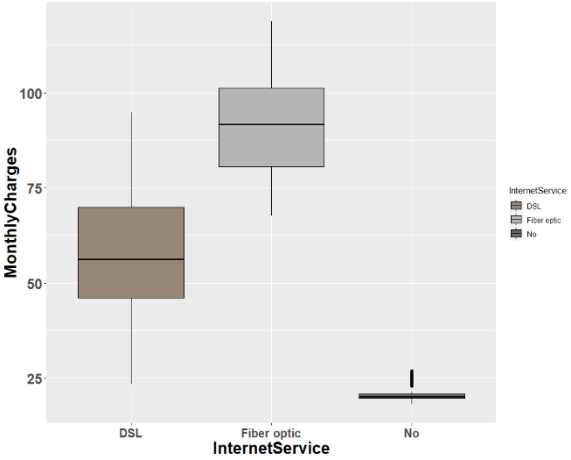
We combined two categorical variables and explored the relationship between *PaymentMethod* and *Contract*. Here, we can see that the most popular payment method for month to month customers is through electronic check. This could be the factor that drives up churn rates for month to month customers. To confirm, we created a graph on churn rate by contract type, excluding electronic check as the payment option. Despite excluding electronic check, the churn rate for month to month customers are still significantly higher than other contract types.



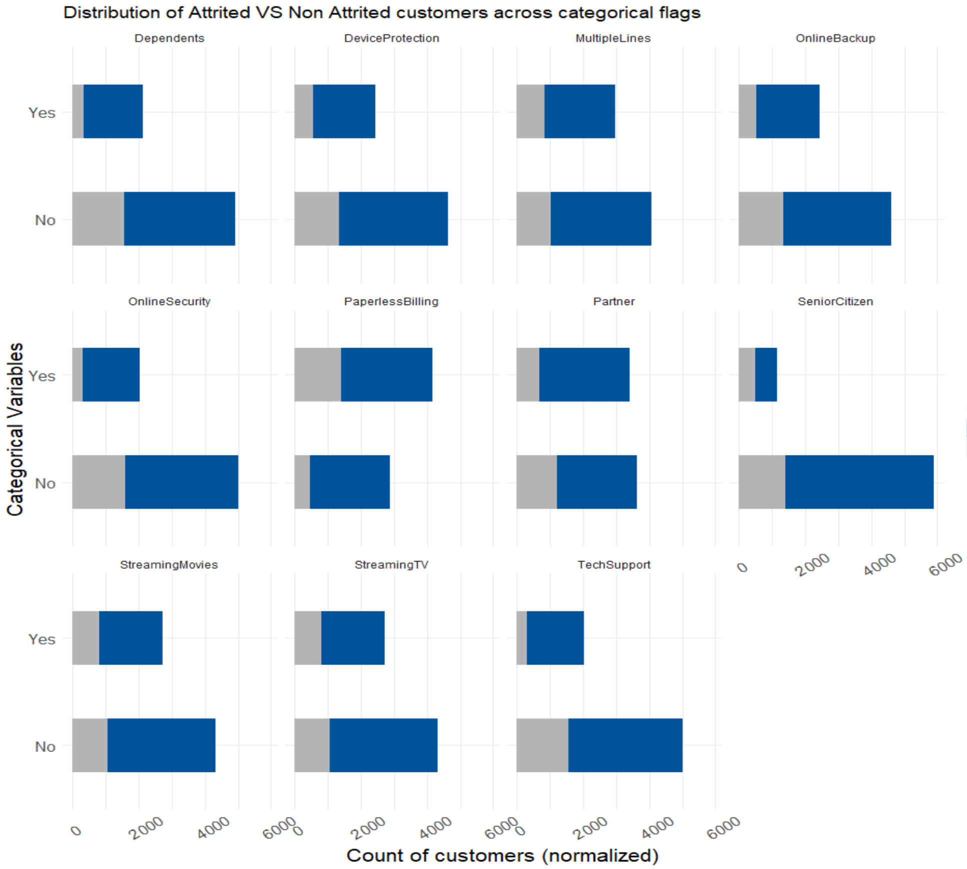


Categorical Variable #3: Internet Service:





Customers who have signed up for Fiber Optics have a very high churn rate. *MonthlyCharges* for customers who use Fiber Optics are also significantly higher than customers who sign up for other Internet services. After this, we took a snapshot of the churn rates across all variables in order to identify those that did not have a significant impact.



This graph demonstrates that variables such as *Dependents*, *Device Protection*, *Online Backup*, *Online Security*, *Paperless Billing*, *Partner* and *Tech Support* play a huge role in determining whether a customer is likely to churn or not.

Variables that are particularly interesting are *TechSupport*, *PaperlessBilling* and *OnlineSecurity*. 50% of customers have not enrolled in *TechSupport* and *OnlineSecurity* and are much more likely to churn compared to customers who have enrolled. Customers who have enrolled in *PaperlessBilling* are also more likely to churn.

Modeling

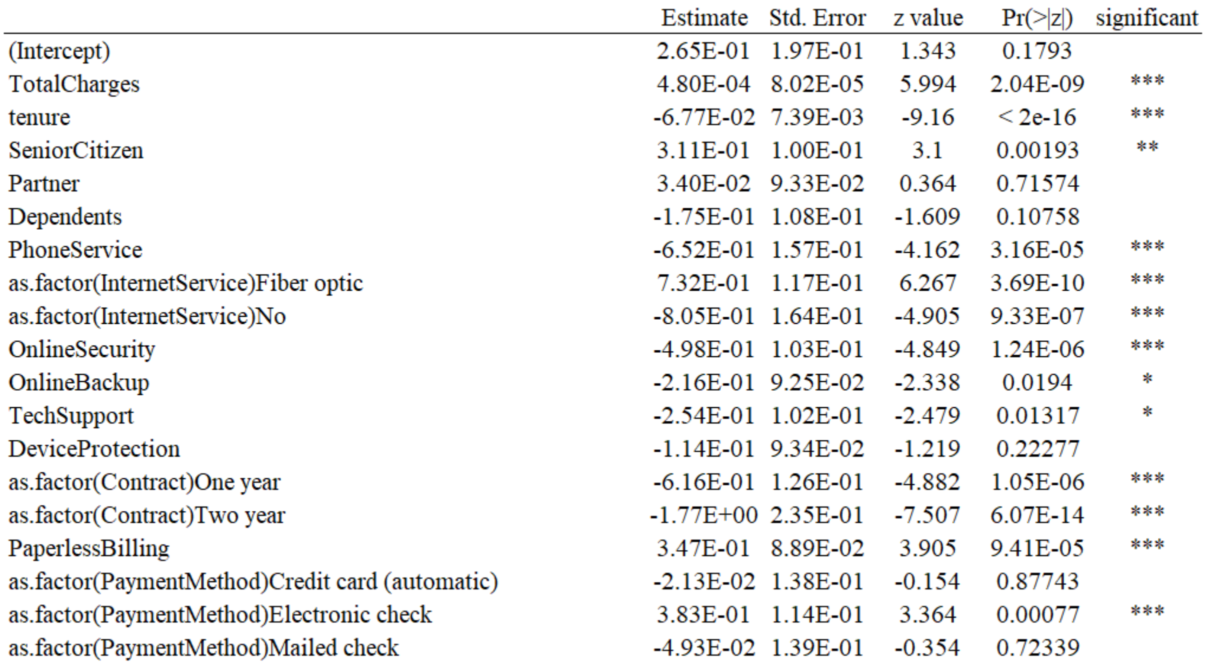
In order to prepare the data to run the models we conducted the following steps:

1. Factored all variables that had “Yes” or “No” fields as “1” and “0” respectively
2. Dropped variables that our exploratory analysis indicated as having no significant effect on Churn (*gender*, *MultipleLines*, *StreamingMovies*, *StreamingTV*)

After this, we split the data set by 70% (training) and 30% (testing) so that we could test the accuracy of our model later.

Base Model

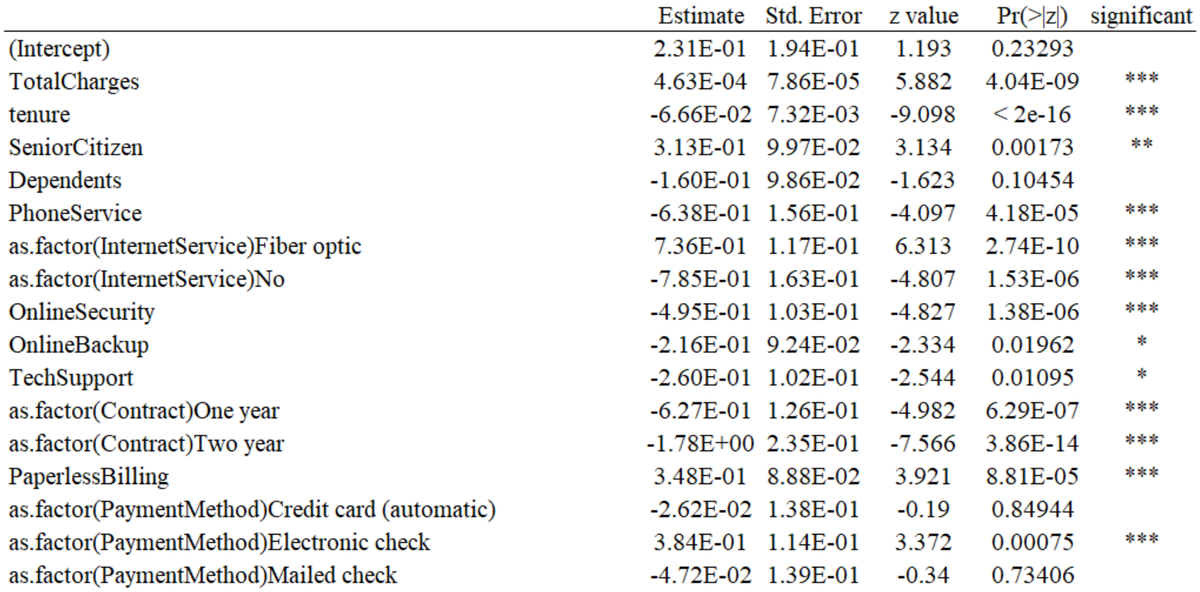
For the base model we inputted in all variables except for those that have already been identified as having no significant effect on Churn.



	Estimate	Std. Error	z value	Pr(>|z|)	significant
(Intercept)	2.65E-01	1.97E-01	1.343	0.1793	
TotalCharges	4.80E-04	8.02E-05	5.994	2.04E-09	***
tenure	-6.77E-02	7.39E-03	-9.16	< 2e-16	***
SeniorCitizen	3.11E-01	1.00E-01	3.1	0.00193	**
Partner	3.40E-02	9.33E-02	0.364	0.71574	
Dependents	-1.75E-01	1.08E-01	-1.609	0.10758	
PhoneService	-6.52E-01	1.57E-01	-4.162	3.16E-05	***
as.factor(InternetService)Fiber optic	7.32E-01	1.17E-01	6.267	3.69E-10	***
as.factor(InternetService)No	-8.05E-01	1.64E-01	-4.905	9.33E-07	***
OnlineSecurity	-4.98E-01	1.03E-01	-4.849	1.24E-06	***
OnlineBackup	-2.16E-01	9.25E-02	-2.338	0.0194	*
TechSupport	-2.54E-01	1.02E-01	-2.479	0.01317	*
DeviceProtection	-1.14E-01	9.34E-02	-1.219	0.22277	
as.factor(Contract)One year	-6.16E-01	1.26E-01	-4.882	1.05E-06	***
as.factor(Contract)Two year	-1.77E+00	2.35E-01	-7.507	6.07E-14	***
PaperlessBilling	3.47E-01	8.89E-02	3.905	9.41E-05	***
as.factor(PaymentMethod)Credit card (automatic)	-2.13E-02	1.38E-01	-0.154	0.87743	
as.factor(PaymentMethod)Electronic check	3.83E-01	1.14E-01	3.364	0.00077	***
as.factor(PaymentMethod)Mailed check	-4.93E-02	1.39E-01	-0.354	0.72339	

Modeling Approach 1: Stepwise on Base Model

After the base model, we progressed to do a blind stepwise model using the base model above.



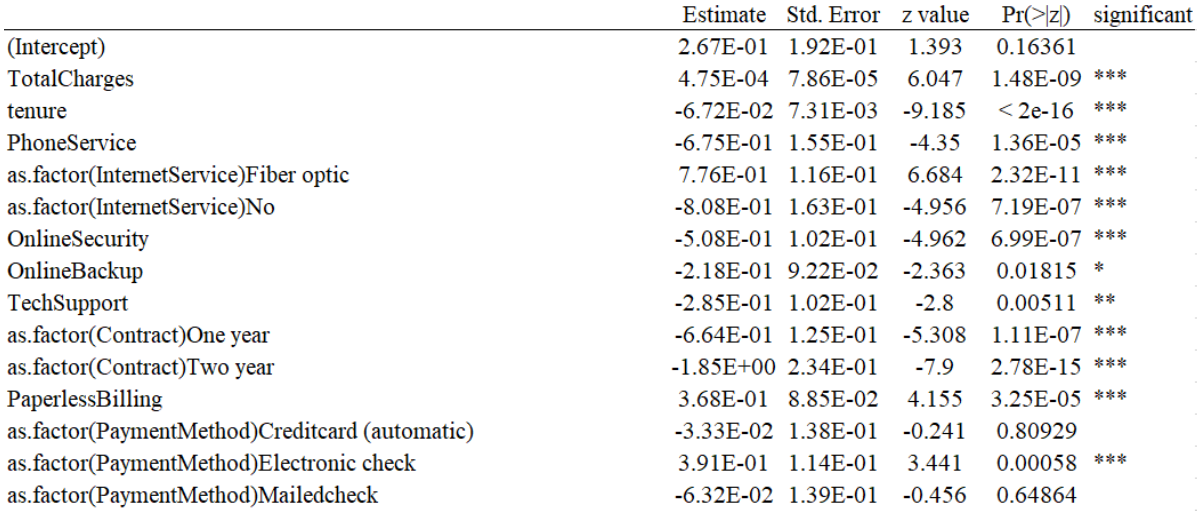
	Estimate	Std. Error	z value	Pr(>|z|)	significant
(Intercept)	2.31E-01	1.94E-01	1.193	0.23293	
TotalCharges	4.63E-04	7.86E-05	5.882	4.04E-09	***
tenure	-6.66E-02	7.32E-03	-9.098	< 2e-16	***
SeniorCitizen	3.13E-01	9.97E-02	3.134	0.00173	**
Dependents	-1.60E-01	9.86E-02	-1.623	0.10454	
PhoneService	-6.38E-01	1.56E-01	-4.097	4.18E-05	***
as.factor(InternetService)Fiber optic	7.36E-01	1.17E-01	6.313	2.74E-10	***
as.factor(InternetService)No	-7.85E-01	1.63E-01	-4.807	1.53E-06	***
OnlineSecurity	-4.95E-01	1.03E-01	-4.827	1.38E-06	***
OnlineBackup	-2.16E-01	9.24E-02	-2.334	0.01962	*
TechSupport	-2.60E-01	1.02E-01	-2.544	0.01095	*
as.factor(Contract)One year	-6.27E-01	1.26E-01	-4.982	6.29E-07	***
as.factor(Contract)Two year	-1.78E+00	2.35E-01	-7.566	3.86E-14	***
PaperlessBilling	3.48E-01	8.88E-02	3.921	8.81E-05	***
as.factor(PaymentMethod)Credit card (automatic)	-2.62E-02	1.38E-01	-0.19	0.84944	
as.factor(PaymentMethod)Electronic check	3.84E-01	1.14E-01	3.372	0.00075	***
as.factor(PaymentMethod)Mailed check	-4.72E-02	1.39E-01	-0.34	0.73406	

Overall, it was a good model that resulted in an accuracy of 81% after tested on the test dataset. To improve the model, we explored some manual iterations.

Modeling Approach 2: Manual Model Iterations

We manually dropped variables based on two factors: p-value, which tests the significance of the predictor, and VIF, which measures collinearity between predictors.

First, we dropped variables of higher p-values across multiple iterations step by step, which are *Partners* (0.799), *DeviceProtection* (0.956), *Dependents* (0.079) and *SeniorCitizen* (0.120).Then we checked the VIF values of rest of the variables, which all turned out to be under 5. Therefore, it means that the variables left are not highly correlated, so we do not have to drop any of them at this moment. After dropping the four variables above based on p-value, we finalized the model with the outputs below:



	Estimate	Std. Error	z value	Pr(>|z|)	significant
(Intercept)	2.67E-01	1.92E-01	1.393	0.16361	
TotalCharges	4.75E-04	7.86E-05	6.047	1.48E-09	***
tenure	-6.72E-02	7.31E-03	-9.185	< 2e-16	***
PhoneService	-6.75E-01	1.55E-01	-4.35	1.36E-05	***
as.factor(InternetService)Fiber optic	7.76E-01	1.16E-01	6.684	2.32E-11	***
as.factor(InternetService)No	-8.08E-01	1.63E-01	-4.956	7.19E-07	***
OnlineSecurity	-5.08E-01	1.02E-01	-4.962	6.99E-07	***
OnlineBackup	-2.18E-01	9.22E-02	-2.363	0.01815	*
TechSupport	-2.85E-01	1.02E-01	-2.8	0.00511	**
as.factor(Contract)One year	-6.64E-01	1.25E-01	-5.308	1.11E-07	***
as.factor(Contract)Two year	-1.85E+00	2.34E-01	-7.9	2.78E-15	***
PaperlessBilling	3.68E-01	8.85E-02	4.155	3.25E-05	***
as.factor(PaymentMethod)Creditcard (automatic)	-3.33E-02	1.38E-01	-0.241	0.80929	
as.factor(PaymentMethod)Electronic check	3.91E-01	1.14E-01	3.441	0.00058	***
as.factor(PaymentMethod)Mailedcheck	-6.32E-02	1.39E-01	-0.456	0.64864	

Exploring Interactions

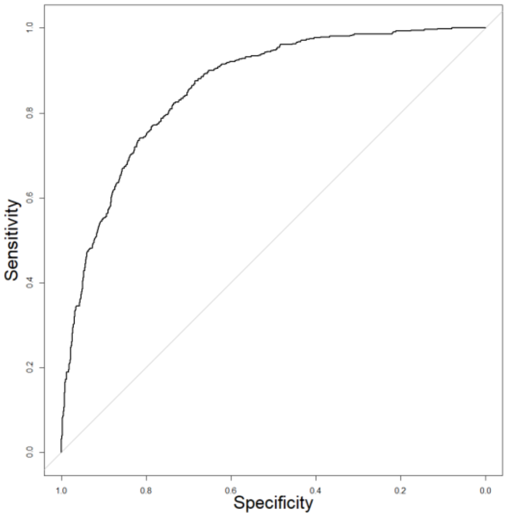
- Do loyal customers with Fiber optic Internet Service have lower risk of churning?
- How does Payment method interact with Total Charges?
- What about internet type opted for and charges?

In order to answer these questions, we iterated over more than five different models exploring the interaction effects of internet service, total charges, tenure, payment method etc.

However, the outputs of these couldn't be interpreted as they have very high p-value in all the iterations with multiple permuted variable selection. Hence, we decided to drop this exploration and stick to the earlier model finalized

Choosing the Threshold

It is a trade-off between True Positive and True Negative. The model we finalized had an accuracy of 82% with the optimal cut-off which maximize the accuracy of model in terms of choosing the value of the cut off at which the model predicts the churn status of the maximum number. We got optimal cut-off at a little over 50%.



However, with this cut-off we found that the True positive rate or the Sensitivity of the model is relatively low at <60% with a higher specificity (true negative) at close to 90%. In order to mitigate this, we reduce the threshold to flag more customers as "at-risk" than the earlier threshold (little more than 50%) thereby increasing the True Positive rate. **AUC** indicates the rate of successful classification by the logistic model, and our model has a high value at 0.85. Looking at the TP VS TN graph we see the sweet spot lies where sensitivity and specificity are both between 75%-80%. The final threshold was chosen to be **35%** which gave a higher sensitivity of roughly 75% and a specificity of 85% with accuracy at ~80%. Below is the confusion matrix with 0.35 as the threshold.

Confusion Matrix	Predicted Churn Yes	Predicted Churn No
Actual Churn Yes	415	145
Actual Churn No	293	1259

Insights and Recommendations

From the final model we have we concluded that factors that would lead customers to churn out include: Fiber optic as the option for Internet Service, Month-to-Month as contract type, Electronic check and Paperless billing as payment method, and high total charges. While the others seem plausible, the hypothesis behind Paperless billing customers we presume, is easier for them to cancel. These are customers who handle their phone plans and billing on their cellphone which offers more flexibility in opting out of MQ&M's service.

We also identify factors that would help retain current customers, which include Two-year contract, One-year contract type, and phone service. Based on the findings, we suggest that the marketing team can split customers based on the risk:

35%-50% as ***low-risk customers*** – Target them with E-mails/SMS reminding of MQ&M other offerings and try to improve their engagement;

51%-75% as ***at-risk customers*** – Target them with bundled discounts and offers to improve their stickiness with our organization;

76%+ as ***high-risk customers*** – Run the extra mile for them, give them a call and offer personalized discounts based on their usage patter;

These offers can be prioritized based on their *TotalCharges* as well, which is a signal of their potential Life Time Value to the company. In addition, the sales team could also push more customers to switch to one- or two-year plans instead of the month-to-month plan and do not sell phone service and Internet service as a bundle because Internet service is more likely to cause customers to churn. At last, the Product or Pricing team should put the Fiber Optic Internet service on their priority list to improve.