MQM Stats Final Project: Churn Prediction for the Telecom Industry

Decision 518Q-C-Team 40

Arwen Wang, Kewei Jiang, Kieu Anh Nguyen, Taru Dharra, Vedant Sahay

I- Introduction

The dataset we have chosen focuses on the Telecom industry. There are 21 columns in the dataset describing the factors which could have an impact on the customer churn (customer attrition). Some of the variables are gender, payment method, tenure, contract, multiple lines etc. The dataset used in this project has been retrieved from Kaggle and is an IBM sample dataset. It is labeled as 'telecom-churn-prediction'. The types of variables included in the dataset are both categorical and numerical.

II- Business Understanding

We attempted to use this data to predict the causes of the customer churn and predict behavior to retain customers in Telecom industry. The telecommunication services market, which includes fixed-network services and mobile services, had a value of around 1.4 trillion U.S. dollars in 2017, and is forecast to grow to almost 1.46 trillion U.S. dollars in size by 2020. With the rapid development of telecommunication industry, the service providers are inclined more towards expansion of the subscriber base. To meet the need of surviving in the competitive environment, the retention of existing customers has become a huge challenge. In the survey done in the Telecom industry, it is stated that the cost of acquiring a new customer is far more that retaining the existing one. Therefore, our data mining solution, including logistic regression model, will address the problem of customer attrition by helping telecom companies better understand customer behavior.

III- Data Understanding

Data Source: https://www.kaggle.com/blastchar/telco-customer-churn

This dataset consists of 7043 rows and 21 columns. We have one dependent categorical variable, 19 independent variables (categorical=16 and numerical=3). Each row represents a single customer, and each column contains customer's information:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

IV- Data cleaning

(1) We see that there are 11 missing values under the column Total Charges, and we decided to replace these missing values with the median of the Total Charges column.

(2) The column 'Senior Citizen' is coded as 1 and 0 and these were coded as 'YES' and 'NO' respectively.

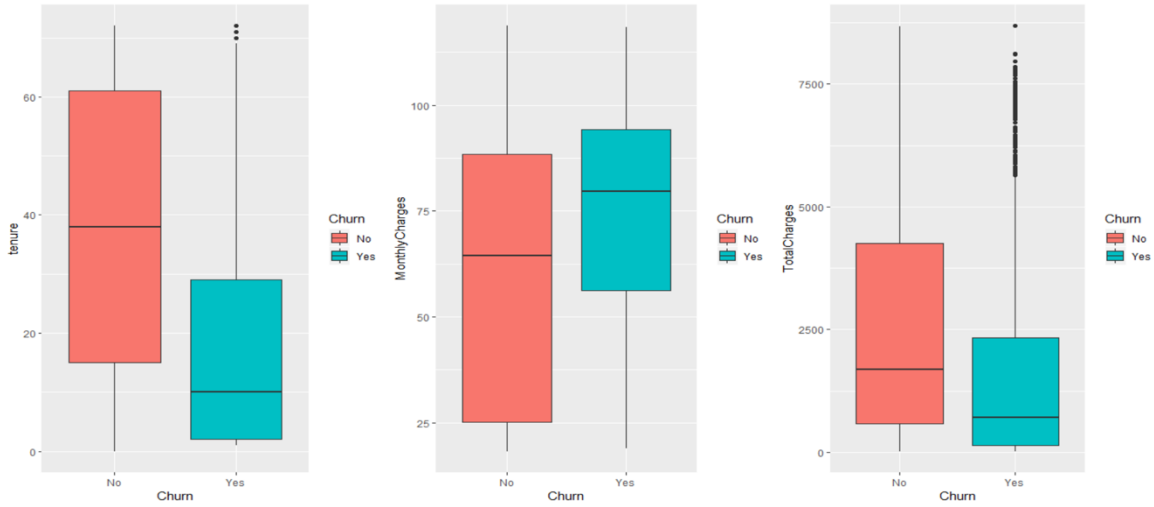
V- Exploratory Data Analysis (EDA)

Considering the fact that the majority of our independent variables are categorical, we decided to use stacked bar plots to look at each one of the variables given to see the relationship it had with our target variable. For numerical variables, we used boxplots to look at their relationship with Churn.

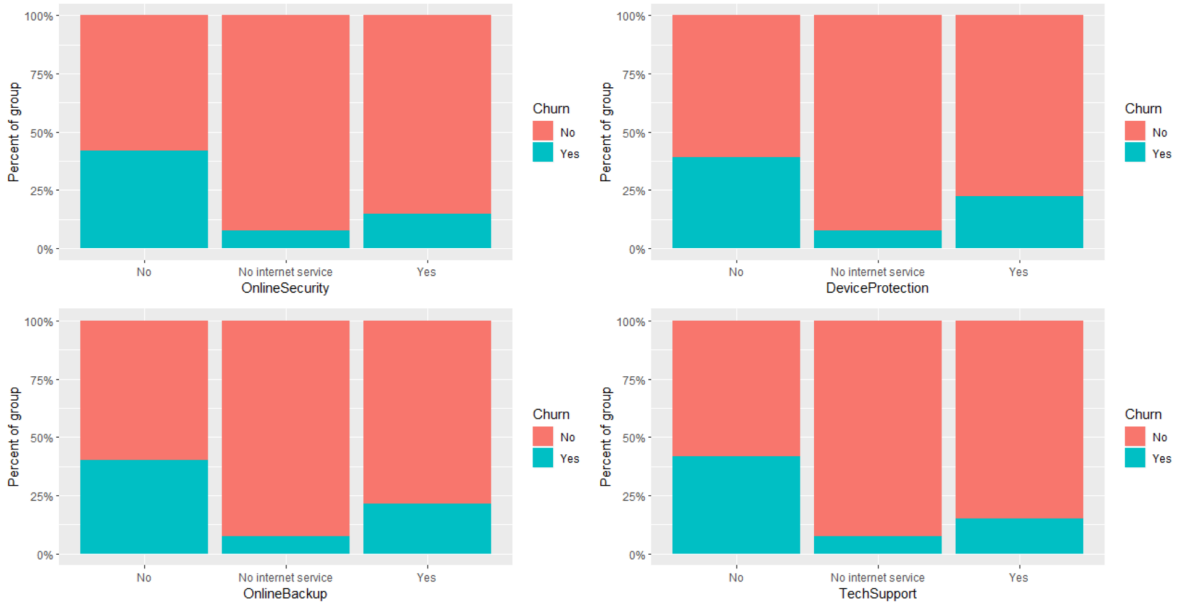
Variables that appeared to have an impact on churn was included in the following analysis and variables that did not appear to have a relationship with Churn are not included. Due to the constraints of this reports, we will not include all of the plots we used to identify the relationship between the independent variables and the dependent variable.

- **Senior Citizens**: We have observed that people who are senior citizens are more likely to churn compared to people who are not senior citizen.
- **Dependents**: On observing the dependents with respect to Churn, we see that customers who have dependents tend to churn less as compared to customers who do not have dependents.
- **Partner**: On observing the above graph, we see that customers who have partner tend to churn less than those who do not have a partner
- **Paperless Billing:** There seems to be a relationship between Paperless Billing and Churn. For people who use paperless billing, there seems to be a bigger portion of customers who churned.
- **Payment Method**: While there doesn't seem to be any difference in terms of portion of people who churned in customers who paid with bank transfer, credit card or mailed check, the proportion of customers who churned for people who paid with an electronic check is particularly high.
- **Products:** Customers are more likely to churn if they are using the Fiber optic internet services than those who are using DSL internet service or those who are not using

internet service.



- **Tenure:** The boxplot above indicates that the median tenure for customers who have left is around 10 months.
- **Monthly Charges**: Customers who churned on average paid more monthly compared to customers who did not churn.
- **Total charges**: Even though people who churned paid more monthly, the total amount they paid the company is, on average, lower than the amount people who didn't churn paid the company. Clearly, customers who did not churn paid more to the company in total.

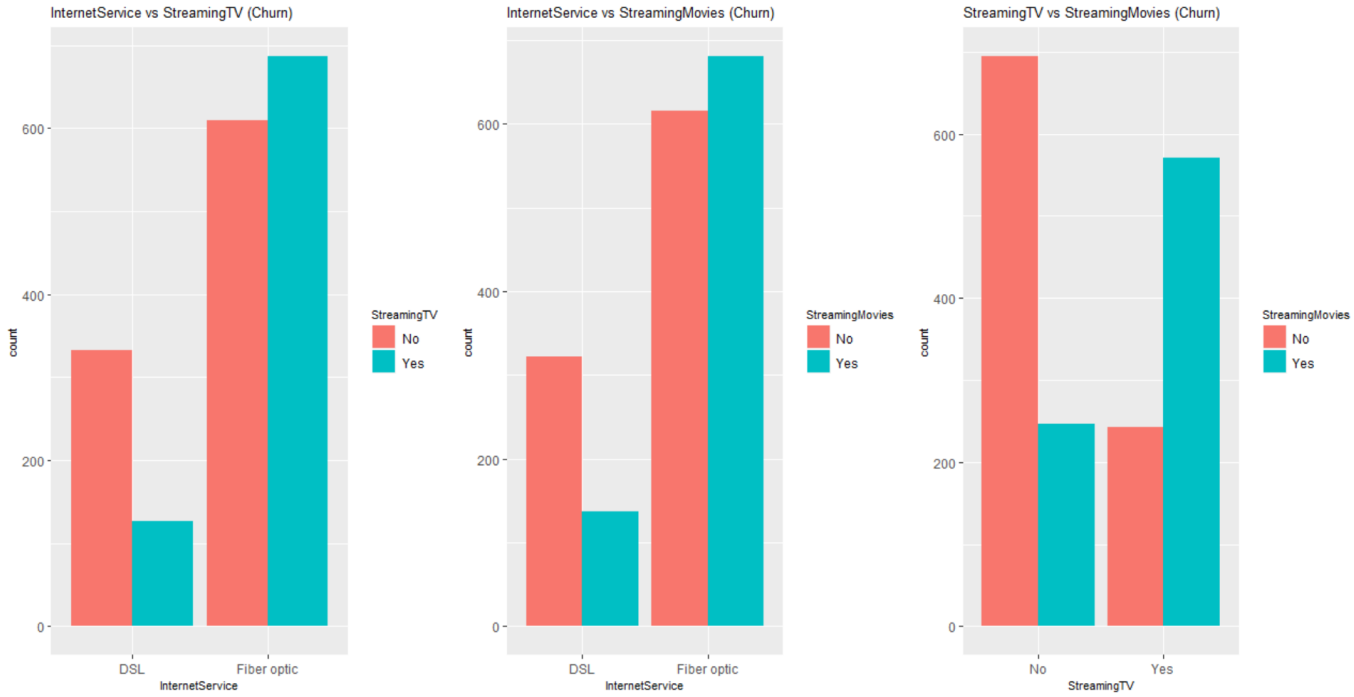


- **Online Security:** It is observed that users who do not have Online Security also are the most likely to churn.
- **Device Protection**: Users who have Device Protection churn a lot more than those users who do not have device protection.

- **Online Backup:** As observed, we see that users without Online Backup churn way more than those who have online backup.
- **Tech Support:** As observed from the plot, we see that users without Tech Support churn more than users who have tech support or who do not have Internet service.
- **Contract**: It is not surprising to find that the longer the contract terms are, the less churn occur, since people are bind with the contract and they have more stable relationships with Telecom. Also, from the bar plot, we can see that the month-to-month contract has the highest churn, and much higher than the one-year and two-year contract groups.
- **Streaming**: According to the bar chart above, we can interpret that the people who subscribe the StreamingMovies and StreamingTV services are less likely to churn, which makes sense because the subscription on extra service can increase the stability.
- **Multiple Lines:** People who have multiple lines are slightly more likely to churn.

VI- Interaction exploration

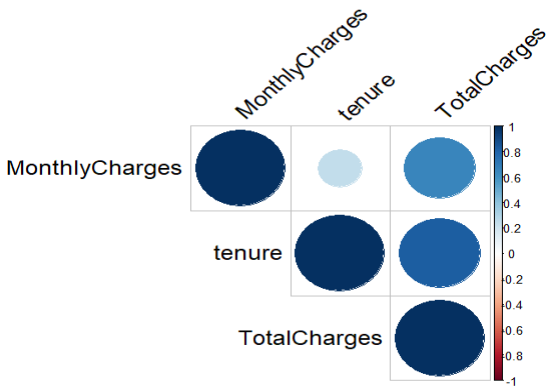
After our exploratory regression analysis, we found some inconsistency between the findings in our exploratory data analysis (EDA) and the regression output with regards to the variables StreamingTV, StreamingMovies and InternetServices. We decided to examine potential interactions between these highly significant variables.



From the plots above, we have observed an interaction between the three variables Streaming Movies, Streaming TV and Internet Service. For people who used DSL, they are more likely to churn if they don't stream movies or TV but the opposite is true for people who use fiber optics. For people who didn't stream movies and also didn't stream TV, they are more likely to churn

compared to people who only streamed movies and not TV. However, the opposite is true for people who do stream TV since people who stream both TV and movies are more likely to churn compared to people who only stream TV and not movies. In light of these findings, we are including a three-way interaction term between InternetService, StreamTV and StreamMovies.

Correlation plot on all continuous variables



It is observed that tenure and TotalCharges are the most correlated variables. MonthlyCharges and TotalCharges are also highly correlated. To avoid the problem of multicollinearity in our model, we decided to eliminate TotalCharges from our models because it is highly correlated with tenure and MonthlyCharges while MonthlyCharges and tenure are not highly correlated.

Modeling:

We will build 4 logistic regression models to predict the churn of a customer in telco company. This is a supervised learning problem because it involves predicting the customer churn using the independent variables provided in the data.

Intuitive:

AIC – 5888.7

$$P = 1 / (1+ e^{-(0.994017 + 0.214341(\text{SeniorCitizen})\text{Yes} + -0.034840\ \text{Tenure} + 1.579894\ (\text{InternetService})\text{Fiber optic} + -1.79\ (\text{InternetService})\ \text{No} + -0.186493(\text{OnlineSecurity})\ \text{Yes} + 0.195368\ (\text{StreamingTV})\ \text{Yes} + 0.337(\text{StreamingMovies})\ \text{Yes} + -0.665188(\text{Contract})\ \text{One Year} + -1.35(\text{Contract})\ \text{Two Year} + 0.346704(\text{PaperlessBilling})\text{Yes} + -0.078602\ (\text{PaymentMethod})\ \text{Credit card (automatic)} + 0.30855\ (\text{PaymentMethod})\ \text{Electronic Check} + -0.037590(\text{PaymentMethod})\ \text{Mailed Check} + -0.034417\text{MonthlyCharges} + -0.168640\ (\text{Dependents})\ \text{Yes} + 0.005597(\text{Partner})\text{Yes} + -0.034840\ \text{Tenure} + 0.172021(\text{DeviceProtection})\ \text{Yes} + -0.170529(\text{OnlineBackup})\ \text{Yes} + 0.463986\ (\text{MultipleLines})\ \text{Yes} + -0.157491(\text{TechSupport})\ \text{Yes} + 0.380456\ (\text{StreamingTV})\text{Yes:}(\text{StreamingMovies})\ \text{Yes} + 0.379186\ (\text{InternetService})\ \text{Fiber optic} : (\text{StreamingTV})\ \text{Yes} + 0.200208\ (\text{InternetService})\ \text{Fiber optic} : (\text{StreamingMovies})\ \text{Yes} + -0.079244\ (\text{InternetService})\ \text{Fiber optic} : (\text{StreamingTV})\ \text{Yes} : (\text{StreamingMovies})\ \text{Yes})}$$

	Churn		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	2.70	0.16 – 46.19	0.493
as.factor(SeniorCitizen)YES	1.24	1.05 – 1.46	**0.012**
as.factor(InternetService)Fiber optic	4.85	1.00 – 23.67	0.051
as.factor(InternetService)No	0.17	0.03 – 0.82	**0.028**
as.factor(OnlineSecurity)Yes	0.83	0.58 – 1.18	0.300
as.factor(StreamingTV)Yes	1.22	0.59 – 2.50	0.596
as.factor(StreamingMovies)Yes	1.40	0.68 – 2.89	0.360
as.factor(Contract)One year	0.51	0.42 – 0.63	**<0.001**
as.factor(Contract)Two year	0.26	0.18 – 0.36	**<0.001**
as.factor(PaperlessBilling)Yes	1.41	1.22 – 1.64	**<0.001**
as.factor(PaymentMethod)Credit card (automatic)	0.92	0.74 – 1.16	0.492
as.factor(PaymentMethod)Electronic check	1.36	1.13 – 1.64	**0.001**
as.factor(PaymentMethod)Mailed check	0.96	0.77 – 1.20	0.742
MonthlyCharges	0.97	0.91 – 1.03	0.282
as.factor(Dependents)Yes	0.84	0.71 – 1.01	0.061
as.factor(Partner)Yes	1.01	0.86 – 1.17	0.943
tenure	0.97	0.96 – 0.97	**<0.001**
as.factor(DeviceProtection)Yes	1.19	0.84 – 1.68	0.334
as.factor(OnlineBackup)Yes	1.06	0.75 – 1.49	0.757
as.factor(MultipleLines)No phone service	0.84	0.23 – 3.04	0.794
as.factor(MultipleLines)Yes	1.59	1.12 – 2.26	**0.009**
as.factor(TechSupport)Yes	0.85	0.60 – 1.22	0.387
as.factor(StreamingTV)Yes:as.factor(StreamingMovies)Yes	1.46	0.84 – 2.56	0.181
as.factor(InternetService)Fiber optic:as.factor(StreamingTV)Yes	1.46	0.93 – 2.29	0.099
as.factor(InternetService)Fiber optic:as.factor(StreamingMovies)Yes	1.22	0.79 – 1.89	0.370
as.factor(InternetService)Fiber optic:as.factor(StreamingTV)Yes:as.factor(StreamingMovies)Yes	0.92	0.47 – 1.80	0.816

In this model, the significant variables are Senior Citizen, Internet Service, Contract, PaperlessBilling, ElectronicCheck, Tenure, and MultipleLines. On evaluation of the output above, these following factors increases the probability of a person churning:

- SeniorCitizen: If a person is a senior citizen, the odds of that person churning is 1.24, making him/her more likely to churn rather than staying
- InternetService: If a person has Fiber Optics, he/she is 4.85 times more likely to churn than if he/she doesn't have Fiber Optics
- Contract: If customers who have month-to-month contracts are 2 times more likely to churn compared with people who have one-year contracts, and 4 times more likely to churn compared with those who have two-year contracts
- PaperlessBilling: People who use Paperless billing are 1.4 times more likely to churn compare to traditional billing
- ElectronicCheck: People who use the electronic check are 1.36 times more likely to churn compare to the traditional payment method
- Tenure: The tenure goes up, the churn reduces
- MultipleLines: People who have multiple lines are 1.59 times more likely to churn compare to people who have single lines

Running a Stepwise selection and Forward selection:

AIC – 5880.94

$$P = 1 / (1 + e^{-(-0.579244 + -0.668828\ (\text{Contract})\ \text{One Year} + -1.355(\text{Contract})\ \text{Two Year} + 0.764811(\text{InternetService})\text{Fiber Optic} + -0.910246\ (\text{InternetService})\ \text{No} + -0.034716\ \text{Tenure} + -0.077197\ (\text{PaymentMethod})\ \text{Credit Card (automatic)} + 0.309901\ (\text{PaymentMethod})\ \text{Electronic Check} + -0.034103(\text{PaymentMethod})\ \text{Mailed Check} + 0.511445(\text{MultipleLines})\ \text{No Phone Service} + 0.293307\ (\text{MultipleLines})\ \text{Yes} + 0.345725(\text{PaperlessBilling})\ \text{Yes} + --0.359087\ (\text{Online Security})\ \text{Yes} + -0.330008(\text{TechSupport})\ \text{Yes} + 0.122405(\text{StreamingMovies})\ \text{Yes} + 0.214781(\text{SeniorCitizen})\ \text{Yes} + -0.115(\text{OnlineBackup})\ \text{Yes} + -0.174633\ (\text{StreamingTV})\ \text{Yes} + -(\text{Dependents})\ \text{Yes} + 0.404545(\text{InternetService})\text{Fiber optic} : (\text{StreamingTV})\ \text{Yes} + 0.341682\ (\text{StreamingMovies})\ \text{Yes} : (\text{StreamingTV})\ \text{Yes})}$$

	Churn		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.56	0.43 – 0.72	**<0.001**
as.factor(Contract)One year	0.51	0.42 – 0.63	**<0.001**
as.factor(Contract)Two year	0.26	0.18 – 0.36	**<0.001**
as.factor(InternetService)Fiber optic	2.15	1.74 – 2.65	**<0.001**
as.factor(InternetService)No	0.40	0.31 – 0.53	**<0.001**
tenure	0.97	0.96 – 0.97	**<0.001**
as.factor(PaymentMethod)Credit card (automatic)	0.93	0.74 – 1.16	0.499
as.factor(PaymentMethod)Electronic check	1.36	1.13 – 1.64	**0.001**
as.factor(PaymentMethod)Mailed check	0.97	0.77 – 1.21	0.765
as.factor(MultipleLines)No phone service	1.67	1.30 – 2.14	**<0.001**
as.factor(MultipleLines)Yes	1.34	1.15 – 1.56	**<0.001**
as.factor(PaperlessBilling)Yes	1.41	1.22 – 1.63	**<0.001**
as.factor(OnlineSecurity)Yes	0.70	0.59 – 0.82	**<0.001**
as.factor(StreamingMovies)Yes	1.13	0.92 – 1.39	0.253
as.factor(TechSupport)Yes	0.72	0.61 – 0.85	**<0.001**
as.factor(StreamingTV)Yes	0.84	0.63 – 1.12	0.239
as.factor(SeniorCitizen)YES	1.24	1.05 – 1.46	**0.011**
as.factor(Dependents)Yes	0.85	0.72 – 1.00	**0.044**
as.factor(OnlineBackup)Yes	0.89	0.77 – 1.04	0.132
as.factor(InternetService)Fiber optic:as.factor(StreamingTV)Yes	1.50	1.10 – 2.03	**0.009**
as.factor(StreamingMovies)Yes:as.factor(StreamingTV)Yes	1.41	1.04 – 1.91	**0.027**

Stepwise Selection

	Churn		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.56	0.43 – 0.72	**<0.001**
as.factor(Contract)One year	0.51	0.42 – 0.63	**<0.001**
as.factor(Contract)Two year	0.26	0.18 – 0.36	**<0.001**
as.factor(InternetService)Fiber optic	2.15	1.74 – 2.65	**<0.001**
as.factor(InternetService)No	0.40	0.31 – 0.53	**<0.001**
tenure	0.97	0.96 – 0.97	**<0.001**
as.factor(PaymentMethod)Credit card (automatic)	0.93	0.74 – 1.16	0.499
as.factor(PaymentMethod)Electronic check	1.36	1.13 – 1.64	**0.001**
as.factor(PaymentMethod)Mailed check	0.97	0.77 – 1.21	0.765
as.factor(MultipleLines)No phone service	1.67	1.30 – 2.14	**<0.001**
as.factor(MultipleLines)Yes	1.34	1.15 – 1.56	**<0.001**
as.factor(PaperlessBilling)Yes	1.41	1.22 – 1.63	**<0.001**
as.factor(OnlineSecurity)Yes	0.70	0.59 – 0.82	**<0.001**
as.factor(StreamingMovies)Yes	1.13	0.92 – 1.39	0.253
as.factor(TechSupport)Yes	0.72	0.61 – 0.85	**<0.001**
as.factor(StreamingTV)Yes	0.84	0.63 – 1.12	0.239
as.factor(SeniorCitizen)YES	1.24	1.05 – 1.46	**0.011**
as.factor(Dependents)Yes	0.85	0.72 – 1.00	**0.044**
as.factor(OnlineBackup)Yes	0.89	0.77 – 1.04	0.132
as.factor(InternetService)Fiber optic:as.factor(StreamingTV)Yes	1.50	1.10 – 2.03	**0.009**
as.factor(StreamingMovies)Yes:as.factor(StreamingTV)Yes	1.41	1.04 – 1.91	**0.027**

Forward Selection

We ran a Stepwise Regression and a Forward Regression. These two models ended up generating the same result with the same variables in the models. For both of these models, some variables that were most significant in predicting churn are:

- Contract: people with Month to month contracts are twice as likely to churn compared to people who have one-year contracts and almost four times as likely to churn compared to people who have two-year contracts
- InternetService: people who have Fiber Optics are twice as likely to churn compared to people who have DSL and people who don't have DSL or Fiber Optics are less likely to churn
- PaymentMethod: people who pay using electronic checks are 1.36 times more likely to churn compared to people who pay using bank transfer
- MultipleLines: people who have no phone service are 1.67 times more likely to churn compared to people who have one line and people who have multiple lines are 1.34 times more likely to churn compared to people who have one line
- PaperlessBilling: people who use paperless billing are 1.41 times more likely to churn compared to people don't use paperless billing
- OnlineSecurity: People who have online security are less likely to churn
- TechSupport: people who have tech support are less likely to churn 0.72

- Senior Citizen: people who are senior citizens are 1.24 times more likely to churn compared to people who are not senior citizens
- Dependents: people who have dependents are less likely to churn
- InternetService*StreamTV:People are 1.5 times more likely to churn if they use Fiber Optics to stream TV
- StreamTV*StreamMovies: People are 1.41 times more likely to churn if they stream both TV and movies

Running a backward selection:

We start by selecting all 20 variables excluding monthly charges because of its collinearity with Streaming TV. Using the in-built function in R step() we remove variables not significant in the model based on AIC value

		Churn	
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.55	0.43 – 0.71	**<0.001**
as.factor(SeniorCitizen)YES	1.24	1.05 – 1.46	**0.011**
as.factor(Dependents)Yes	0.85	0.72 – 0.99	**0.040**
tenure	0.97	0.96 – 0.97	**<0.001**
as.factor(InternetService)Fiber optic	2.15	1.75 – 2.65	**<0.001**
as.factor(InternetService)No	0.41	0.32 – 0.54	**<0.001**
as.factor(OnlineSecurity)Yes	0.70	0.59 – 0.82	**<0.001**
as.factor(StreamingTV)Yes	0.84	0.63 – 1.12	0.241
as.factor(StreamingMovies)Yes	1.13	0.91 – 1.39	0.259
as.factor(Contract)One year	0.51	0.42 – 0.63	**<0.001**
as.factor(Contract)Two year	0.26	0.18 – 0.36	**<0.001**
as.factor(PaperlessBilling)Yes	1.41	1.22 – 1.63	**<0.001**
as.factor(PaymentMethod)Credit card (automatic)	0.92	0.74 – 1.16	0.494
as.factor(PaymentMethod)Electronic check	1.36	1.13 – 1.64	**0.001**
as.factor(PaymentMethod)Mailed check	0.97	0.77 – 1.21	0.765
as.factor(MultipleLines)No phone service	1.66	1.30 – 2.13	**<0.001**
as.factor(MultipleLines)Yes	1.34	1.15 – 1.56	**<0.001**
as.factor(TechSupport)Yes	0.72	0.61 – 0.85	**<0.001**
as.factor(StreamingTV)Yes:as.factor(StreamingMovies)Yes	1.41	1.04 – 1.91	**0.026**
as.factor(InternetService)Fiber optic:as.factor(StreamingTV)Yes	1.49	1.10 – 2.02	**0.011**

$$P = 1 / (1+ e^{-(-0.595935+ 0.214395(SeniorCitizen)Yes + -0.166854 (Dependents)Yes + -0.035545 Tenure + 0.767752 (InternetService)Fiber Optic + -0.881392(InternetService) No + -0.362248(OnlineSecurity)Yes + -0.173549 (StreamingTV) Yes + 0.120919(StreamingMovies) Yes + -0.669913(Contract)One Year + -1.355 (Contract) Two Year + 0.343106(PaperlessBilling)Yes + -0.078057 (PaymentMethod)Credit Card(automatic) + 0.310595 (PaymentMethod)Electronic Check + -0.034061 (PaymentMethod) Mailed Check + 0.50882(MultipleLines) No Phone Service +0.291438 (MultipleLines) Yes + -0.333506 (TechSupport) Yes +0.345445 (StreamingTV) Yes : (StreamingMovies)Yes + 0.397995 (InternetService) Fiber optic : (StreamingTV) Yes)}$$

AIC = 5881.21

In this model, the significant variables are Senior Citizen, Dependents, Tenure, Internet Service, Online Security, Contract, PaperlessBilling, ElectronicCheck, Tenure, MultipleLines, and

Techsupport. On evaluation of the output above, these following factors increases the probability of a person churning:

- Dependents: people who have dependents are less likely to churn
- SeniorCitizen: If a person is a senior citizen, the odds of that person churning is 1.24, making him/her more likely to churn rather than staying
- InternetService: If a person has Fiber Optics, he/she is 2.15 times more likely to churn than if he/she doesn't have Fiber Optics
- OnlineSecurity: People who have online security are less likely to churn
- Contract: If customers who have month-to-month contracts are 2 times more likely to churn compared with people who have one-year contracts, and 4 times more likely to churn compared with those who have two-year contracts
- PaperlessBilling: People who use Paperless billing are 1.41 times more likely to churn compare to traditional billing
- ElectronicCheck: People who use the electronic check are 1.36 times more likely to churn compare to the traditional payment method
- Tenure: The tenure goes up, the churn reduces
- MultipleLines: People who have multiple lines are 1.59 times more likely to churn compare to people who have single lines
- TechSupport: people who have tech support are less likely to churn 0.72
- InternetService*StreamTV:People are 1.49 times more likely to churn if they use Fiber Optics to stream TV
- StreamTV*StreamMovies: People are 1.41 times more likely to churn if they stream both TV and movies

Evaluation:

As per the summary table below, the Forward and Step Wise Selection have the lowest AIC and hence are considered as the best models amongst others.

The significant variables in this model are:

- Contract
- InternetService
- PaymentMethod
- MultipleLines
- PaperlessBilling
- OnlineSecurity
- TechSupport
- Senior Citizen
- Dependents
- Interaction between InternetService and StreamTV
- Interaction between StreamTV and StreamMovies

Model	AIC
Forward	5880.94
Backward	5881.21
Step wise	5880.94
Intuitive	5888.7

Insights:

From the regression results of the models above, we can generate many insights and give suggestions to Telco.

Target Market

As is shown in the regression result, customers who are senior citizens are 1.24 times more likely to churn. Thus, we suggest Telco to allocate the marketing resources towards non-senior citizens more.

Internet Services

The company should improve the internet services to prevent customers from churning. As is indicated in the result of the model, customers who use Fiber optic service is 4.85 times more likely to churn compared with those who do not use this service. And customers who do not have internet services are much less likely to churn. We regard this as a growth opportunity since internet service is a necessity.

Contract Options

We suggest Telco to design more long-time contracts. Customers who have month-to-month contracts are 2 times more likely to churn compared with those who have one-year contracts, and 4 times more likely to churn compared with those who have two-year contracts; thus, we believe if Telco can provide customers with more long-time contract options, more customers will stay with Telco for a longer time.

Payment Methods

Telco should advance paperless billing and electronic check since people who are now using these methods are 1.4 times more likely to churn compared with those who are using traditional payment methods.

Additional Services

As for the additional services, Telco should further the streaming movies and streaming TV services, for example, providing more trending movies and TV shows since customers who are using those services right now are 1.5 times more likely to churn.

Also, the tech support service helps Telco to keep customers with the company, so we suggest Telco to provide tech support to more customers.