**Telco Churn**

1. Describe the process by which you cleaned, processed, and partitioned data as necessary

* Checked for null values but did not remove them since they were all in total charges which I was not using in my models
* Dropped CustomerID ad total charges since I was not going to use them for analysis.
* Recoded the target variable ie Churn as 0 and 1
* Partitioned the dataset into three parts using the following conditions

**Phone = (Phone services == “Yes” & Internet service ==”No”)**

**Internet = (Phone services == “No” & (Internet service ==”DSL” or Internet service ==”Fiber optic”))**

**Both = (Phone services == “Yes” & (Internet service ==”DSL” or Internet service ==”Fiber optic”))**

* Removed unnecessary predictors from the partitioned datasets.
* Converted the categorical variables into factors after partitioning.

1. What predictors do you think contributes to the churn of (i) only telephone customers, (ii) only Internet service customers, and (iii) customers who subscribe to both phone and Internet services? Explain the rationale for your answer.

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Sign** **of effect** | **Rationale** |
| Senior Citizen | +/- | **Telephone,Internet,both**- Senior citizens would have been using either of these services for a long time. So the churn rate among them should be low. Although the churn rate could be higher for internet services as they may find it unnecessary. |
| Tenure | - | **Telephone,internet,both –** Higher tenures indicates that a customer has been loyal which would make the unlikely to discontinue. While shorter tenures could mean a customer can churn. |
| Contract | +/- | **Telephone,internet,both -** A month to month contract indicates that the customer is not looking long term. 1 year and 2 year contracts shows a long term commitment |
| Paperless Billing | +/- | **Telephone,internet,both -** Not sure how this would impact. Kept it for analysis. |
| Payment Method | +/- | **Telephone,internet,both -** Customers with certain payment methods could indicate whether they would leave. |
| Monthly Charges | + | **Telephone,internet,both –** Customers with low monthly charges could indicate that they could leave. |
| Multiple Lines | +/- | **Telephone,both –** Customers with multiple lines indicates that they could be in a family plan and would be less likely to leave. |
| Online Security | +/- | **Internet,both –** Online security is absolutely essential for privacy related concerns. Importantly, to safe guard from cyber attacks so not having security could make customers leave. |
| Online Backup | +/- | **Internet,both –** If there is a data loss then it becomes a issue if the telecom service does not provide backup, so not having it could make customers leave. |
| Device Protection | +/- | **Internet,both –** If the company does not offer device protection then customers are more likely to leave. |
| Tech Support | +/- | **Internet,both –** Tech support is absolutely essential for customers to get assistance with their services, so not having it would make customers more likely to leave |
| Streaming TV | +/- | **Internet, both –** Customers using streaming services could indicate they are more likely to stay, it could also have a negative effect if the streaming service is bad. |
| Dependents | +/- | **Telephone,internet,both –** Customers with dependents would be less likely to leave. |
| InternetService | +/- | **Both -** If a type internet service is bad that could cause customers to leave. |

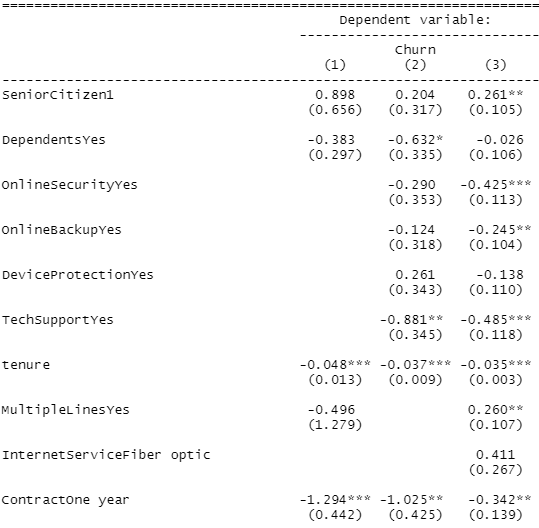
Variables such as CustomerID, Gender are irrelevant. Total Charges and Streaming movies could be explained by other predictors, including them would create high correlation.

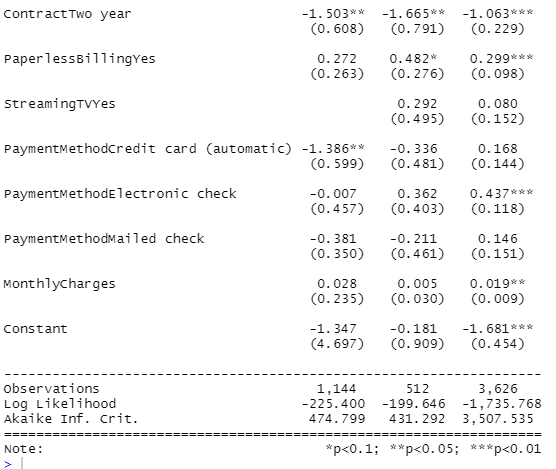
1. Create training and test data sets with a 75:25 split using a random seed of 1024. Use the training data to train three logit models with the variables you identified in Question 2. Combine the outputs of the three modes using stargazer.

**Telephone -** p\_logit <- glm(Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines + Contract + PaperlessBilling + PaymentMethod + MonthlyCharges, family = binomial(link = "logit"),data = train )

**Internet** - i\_logit <- glm(Churn ~ SeniorCitizen + Dependents + OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport + tenure + Contract + PaperlessBilling + StreamingTV + PaymentMethod + MonthlyCharges, family = binomial(link = "logit"),data = train )

**Both** - b\_logit <- glm(Churn ~ SeniorCitizen + Dependents + MultipleLines + InternetService + OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport + tenure + Contract + PaperlessBilling + StreamingTV + PaymentMethod + MonthlyCharges, family = binomial(link = "logit"),data = train )





1. What are the top three predictors of churn of (i) only telephone customers, (ii) only Internet service customers, and (iii) customers who subscribe to both phone and Internet services. Explain using marginal effects how much each predictor contributes to churn occurrence. (3 points)

Top three predictors of churn among customers who use only **telephone** services

|  |  |
| --- | --- |
| **Senior Citizen** | The odds of customers who are senior citizens churning is **2.45** times of the odds of a non- senior citizen given all other predictors are constant |
| **Paperless Billing** | The odds of customers who use paperless billing churning is **1.31** times of the odds of customers who don’t use it given all other predictors are constant |
| **Monthly charges** | If monthly charges increases by 1 dollar, the odds that customer would churn increases by **2.8%** given all other variables are constant |

Top three predictors of churn among customers who use only **internet** services

|  |  |
| --- | --- |
| **Paperless Billing** | The odds of customers who use paperless billing churning is **1.62** times of the odds of customers who don’t use it given all other predictors are constant. |
| **Payment method** | The odds of customers who use electronic payment method churning is **1.44** times of the odds of customers who prefer bank transfer give all other predictors are constant. |
| **Streaming TV** | The odds of customers who use TV streaming churning is **1.34** times of the odds of customers who do not use it keeping all other variables constant. |

Top three predictors of churn among customers who use **both** the services

|  |  |
| --- | --- |
| **Payment method** | The odds of customers who use electronic payment method churning is **1.55** times of the odds of customers who prefer bank transfer give all other predictors are constant. |
| **Internet Service** | The odds of customers who use fiber optic based internet services churning is **1.51** times of the odds of customers who use DSL service keeping all other variables constant. |
| **Paperless Billing** | The odds of customers who use paperless billing churning is **1.35** times of the odds of customers who don’t use it given all other predictors are constant |

1. Fit your models using test data, and compute recall, precision, F1-score, and AUC values for each of your three models. Create a table with these values.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **p\_logit – Telephone only** | **i\_logit – Internet only** | **b\_logit – Both services** |
| **Precision** | 0.25 | 0.48 | 0.55 |
| **Recall** | 0.48 | 0.78 | 0.87 |
| **F1 Score** | 0.33 | 0.59 | 0.67 |
| **AUC value** | 0.68 | 0.72 | 0.74 |