**Assignment 7**

**Mudasir Wazir**

1. **Data Processing, preparation, and cleaning.**

* The data contains 7043 observations and 21 variables across all three types of subscribers.
* There are 11 missing values for Total Charges which are removed.
* All categorical variables are converted to factors and the response variable - Churn is converted and releveled to 1 and 0 form Yes and No.
* The data is unbalanced with 5163 subscribers not churning and 1869 churning. Thus, accuracy may not be a good metric of comparison as it favors the majority class. Model evaluation metrics that are sensitive to class imbalance, such as the F1 score, precision-recall curves, or AUC-ROC should be considered.
* Checking for distribution of various categorical variables across Churn classes.
  + Gender seems evenly distributed suggesting that gender might not be a strong predictor of churn on its own.
  + There are proportionally more senior citizens in the churned group compared to the non-churned group, suggesting that being a senior citizen may be associated with a higher likelihood of churn.
  + There is a higher churn rate among those with paperless billing.
  + Customers without a partner have a higher rate of churn, which might suggest single customers are less tied to a particular service and have fewer barriers to switching.
* We also check for any high correlation between the numeric variables to check for multicollinearity and find high correlation between tenure and total charges as well as monthly charges and total charges. Thus, we may exclude them from the analysis.

**Data Partitioning:**

* We need 3 models on three separate groups of subscribers i.e. Telephone only, Internet only and customers with both services.
* Subsets of the data are created using subset function and logical conditions based on 'PhoneService' and 'InternetService'.
* This results in 3 subsets containing:
  + Telephone only: 1520 observations.
  + Internet only: 680 observations.
  + Both services: 4832 observations.

1. **Hypothesized Predictors**

**Mutual Predictor Table**

(Predictors included in all 3 groups)

|  |  |
| --- | --- |
| **Predictor** | **Rationale** |
| **Tenure** | Longer tenure may indicate more loyalty and satisfaction, leading to a decreased likelihood of churn. |
| **Contract** | Longer-term contracts can often have penalties for early termination, deterring churn. |
| **Paperless Billing** | It could increase churn if these customers are not tech-savvy or do not regularly use online services, making the process inconvenient. Alternatively, it could decrease churn if it is associated with cost savings or if customers prefer the environmental benefits and convenience of electronic statements. |
| **Monthly Charges** | Higher monthly charges might lead to increased churn due to cost sensitivity. |
| **Dependents** | Having Dependents may lower the churn probability as they maybe using the services as well. |
| **Payment Method** | Customers value ease and reliability in payment processes. Inconvenient payment methods might lead to frustration and an increased likelihood of switching to a provider with more user-friendly options. |
| **Senior Citizen** | This demographic may have specific needs or service usage patterns that are distinct from younger customers. |
| **Partner** | Customers with a partner may prefer plans that offer better rates for multiple users, affecting churn. |

**Dropped Predictors excluded in all 3 groups:**

* **Gender:** No specific effect due to gender alone. It’s unlikely to have a direct impact on churn unless linked to other factors.
* **Total charges**: As stated in section 1, it is highly correlated with tenure and monthly charges.

**Telephone-Only Predictor Table**

(Predictors in addition to mutual predictors in Table 1)

|  |  |
| --- | --- |
| **Predictor** | **Rationale** |
| **Multiple Lines** | Customers with multiple lines may be less likely to churn due to higher reliance on service. |

**Dropped:**

* **InternetService:** Since this group does not use internet services, this variable is irrelevant.
* **OnlineSecurity, DeviceProtection, OnlineBackup, TechSupport, and Streaming**: These are features associated with internet services.

**Internet- Only Predictor Table**

(Predictors in addition to mutual predictors in Table 1)

|  |  |
| --- | --- |
| **Predictor** | **Rationale** |
| **TechSupport** | Adequate support can reduce churn by resolving connectivity and technical issues efficiently. |
| **OnlineSecurity** | Customers value their online safety; lack of security services might increase churn. |
| **DeviceProtection** | Customers who have invested in device protection may perceive a higher value from their service provider. This can increase their satisfaction and reduce the likelihood of churn. |
| **OnlineBackup** | Online backup services provide customers with peace of mind regarding data loss. Customers using such services might be less likely to churn |
| **StreamingTV,**  **StreamingMovies** | These services typically represent engagement with the provider’s additional offerings. Customers who use their internet or bundled services for entertainment purposes, like streaming TV shows or movies, might be more satisfied and see more value in their subscription. This could lead to lower churn rates, especially if the content offered is exclusive or of high quality.  We may use only 1 of these due to high correlation. |

**Dropped:**

* **PhoneService, MultipleLines:** As these features relate to telephone services, they do not apply to internet-only customers.
* **InternetService:** As it only contains one type (DSL) for this group.

**Both Services Group** will use all predictors except **Gender** and **Total Charges.** **Phone service variable** will also not be used as this is true for all.

1. **Modelling**
   * Each group is split into a training set and test set on a 75:25 ratio.
   * A logit model is tested for each group. We check for multicollinearity and drop variables that have a high GVIF in a VIF test. (typically, >5)
     1. Streaming services are dropped as they are correlated with monthly charges. These services usually increase monthly charges.
     2. Fabric Optics may also increase monthly charges due to higher charges than DSL. But we keep it in the Both Services group as a predictor to understand its individual effect on churn.

* Mode1 1: Telephone Only Group

model\_phone <- glm(Churn ~ tenure + Contract + PaymentMethod + PaperlessBilling +

MonthlyCharges + SeniorCitizen +

Partner + Dependents, family = binomial(link = "logit"), data = phsplit$train)

* Model 2: Internet Only Group

model\_internet = glm(Churn ~ tenure + Contract + PaymentMethod + PaperlessBilling +

MonthlyCharges + SeniorCitizen +

Partner + Dependents + OnlineSecurity + OnlineBackup + DeviceProtection

+ TechSupport , family = binomial(link = "logit"), data = internetsplit$train)

* Model 3: Both Services Group

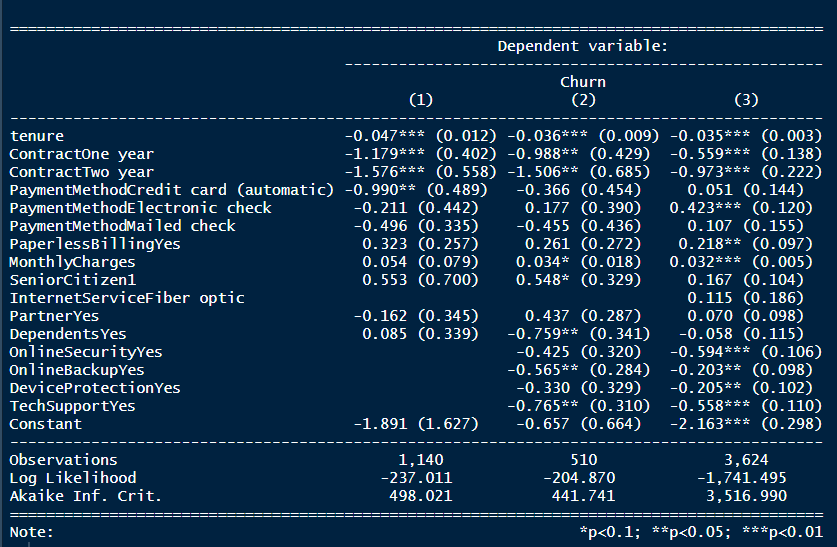
model\_both = glm(Churn ~ tenure + Contract + PaymentMethod + PaperlessBilling +

MonthlyCharges + SeniorCitizen + InternetService +

Partner + Dependents + OnlineSecurity + OnlineBackup + DeviceProtection

+ TechSupport , family = binomial(link = "logit"), data = bothsplit$train)

**Figure 1. Stargazer Output of all three model for the three groups**



1. **Top Predictors Using Marginal Effects**

We first check for the predictors which are statistically significant and then, based on their marginal effects, select the top 3 with the highest absolute values.

**Telephone Only**: Marginal effects of top three predictors:



* **ContractTwo\_year** (-0.094643897): Customers with a two-year contract are, on average, about 9.46 percentage points less likely to churn compared to customers on a month-to-month contract. This suggests that longer-term contracts are associated with lower churn rates, possibly due to the increased commitment and potential penalties for early termination.
* **ContractOne\_year** (-0.070781643): Customers with a one-year contract have, on average, a 7.08 percentage point lower likelihood of churning compared to the customers on a month-to-month contract. While this is still a reduction in churn probability, it is less pronounced than for customers with a two-year contract, which could indicate that the lesser duration of the contract provides a smaller, but still significant, deterrent to churn.
* **PaymentMethodCredit\_card\_automatic** (-0.059458734): Customers who use automatic credit card payments are on average about 5.95 percentage points less likely to churn than the bank transfer payment method group. This suggests that the convenience and possibly the perception of security associated with automatic credit card payments can contribute to lower churn rates.

**Internet Only**: Marginal effects of top three predictors:



* **ContractTwo\_year** (-0.197330628): Customers who have a two-year contract are, on average, approximately 19.73 percentage points less likely to churn compared to customers on a month-to-month contract. This substantial decrease suggests that long-term contracts are strongly associated with customer retention.
* **ContractOne\_year** (-0.129531631): Customers with a one-year contract have, on average, a 12.95 percentage point lower likelihood of churning compared to the customers on a month-to-month contract. This effect is significant, indicating that a one-year commitment can markedly reduce churn probability, although it is less impactful than a two-year contract.
* **TechSupportYes** (-0.089076271): Customers who have technical support are on average about 8.91 percentage points less likely to churn than those without this service. This suggests that access to technical support improves customer satisfaction or mitigates issues that could otherwise lead to churn.

**Both Services:** Marginal effects of top three predictors:



* **ContractTwo\_year** (-0.155227296): Having a two-year contract reduces the probability of churn by about 15.52 percentage points, on average compared to the customers on a month-to-month contract.
* **OnlineSecurityYes** (-0.094702899): Access to online security services reduces the probability of churn by approximately 9.47 percentage points. This implies that customers who value or utilize online security features offered by the service are less likely to churn, possibly because these features add to the perceived value of the service or address security concerns.
* **ContractOne\_year** (-0.089219591): Customers with a one-year contract have an approximately 8.92 percentage point lower probability of churning compared to those on a month-to-month plan.

1. **Performance Metrics**

**A screenshot of a graph

Description automatically generated**

**Phone Only Group:**

* Accuracy: Very high at approximately 92.37%, but this metric can be misleading in the context of an unbalanced dataset because the model could be predicting the majority class well while failing to accurately predict the minority class (which is often of more interest).
* Recall: Perfect at 100%, indicating that the model predicted all churn cases correctly. However, the AUC value suggests that this might be an overestimation due to class imbalance.
* Precision: High at approximately 92.37%, suggesting that when the model predicts churn, it is correct most of the time.
* F1 Score: Also high, at approximately 96.03%, but should be interpreted with caution because of the perfect recall.
* AUC: The AUC is 0.5, which is equivalent to random chance. This implies that the model has no discrimination capacity between the churned and non-churned customers.

**Internet Only Group:**

* Accuracy: At approximately 77.06%, it seems relatively good, but again, it might not reflect the model's performance on the minority class.
* Recall: About 85.83%, which is fairly high and indicates a good ability to identify actual cases of churn.
* Precision: At approximately 83.86%, this suggests that the predictions of churn are reasonably reliable.
* F1 Score: With approximately 84.82%, this balance between precision and recall is good, showing a better model performance compared to just looking at accuracy.
* AUC: At approximately 0.684, the AUC indicates a moderate ability of the model to differentiate between the churned and non-churned customers.

**Both Services Group:**

* Accuracy: At approximately 76.07%, it is relatively lower than the Phone Only group but still decent. However, it may not tell us much about the model's performance on the minority class.
* Recall: At about 82.89%, this group has a lower recall than the Internet Only group, indicating it missed some churn cases.
* Precision: Slightly lower at approximately 82.19%, indicating the predictions made by the model are less often correct compared to the Internet Only group.
* F1 Score: At approximately 82.54%, the balance between precision and recall is good but not as high as the Internet Only group.
* AUC: The highest among the three groups at approximately 0.721, indicating a good ability to discriminate between churned and non-churned customers.

Considering the unbalanced nature of the data, the AUC and F1 Score are likely the most reliable metrics here. The AUC for the Phone Only group indicates that the model might not be performing well in distinguishing churn, despite the high accuracy, recall, and F1 Score. For the Internet Only and Both Services groups, the AUCs suggest a better predictive performance.