**ISM 6137 - Statistical Data Mining**

**Assignment 7**

**Sahil Shah – 19895141**

1. **Feature engineering & data partitioning**

**A**. First checked for missing values and structure of data. Then created binary columns to account for presence of the three main cases for which to analyze churn (Phone only, Internet Only, Both)

**B**. Examined tables of the different target cases to understand balance in the data set

A picture containing diagram

Description automatically generated Table

Description automatically generated with medium confidence A picture containing text

Description automatically generated

**C**. Converted appropriate variables to factors and then created three subsets of the complete dataset wherein each subset includes only the positive class for “phone\_only”, “internet\_only” and “both”

1. **Predictor Table**



Phone Only, Internet Only, Both and Churn are not included as they are either a target case or the target variable itself. This table can be used for all three cases. **Cases** **where certain variables are not applicable** (i.e. online backup is not applicable to the phone only case) **are noted in the rationale.**

1. **Models and Output**

Built three models, one for each case. The models are shown below along with stargazer output

logitphone <- glm(binary\_churn ~ gender + SeniorCitizen + Dependents + tenure + Contract + PaperlessBilling +MonthlyCharges + TotalCharges , family=binomial (link="logit"), data=trainphone)

logitinternet <- glm(binary\_churn ~ gender + SeniorCitizen + Dependents + tenure + Contract + PaperlessBilling +MonthlyCharges +TotalCharges , family=binomial (link="logit"), data=traininternet)

logitboth <- glm(binary\_churn ~ gender + SeniorCitizen + Dependents + tenure + Contract + PaperlessBilling +MonthlyCharges +TotalCharges , family=binomial (link="logit"), data=trainboth)

Train-test Split Dimensions for each subset:

A picture containing graphical user interface

Description automatically generated A picture containing graphical user interface

Description automatically generated A picture containing text

Description automatically generated

Table

Description automatically generated

1. **Marginal Effects of Top 3 Predictors for each case**

**PHONE ONLY**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor name** | **Type** | **β Coefficient Value** | **Exp(β) Value** | **Marginal Effect Interpretation** |
| Contract.L  (Linear Trend) | Ordinal factor | -1.171 | 0.31 | For each level change in the contract length( from Month to month to 1 year, 1 year ->2 years), the odds of churn decreases by 69% |
| Senior Citizen | Binary | 0.568 | 1.76 | If the customer is a Senior citizen, the odds of churn are 1.76 times higher than non-seniors |
| Paperless Billing | Binary | 0.336 | 1.39 | If the customer has paperless billing, the odds of churning are 1.39 times higher than those with paper billing |

**INTERNET ONLY**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor name** | **Type** | **β Coefficient Value** | **Exp(β) Value** | **Marginal Effect Interpretation** |
| Contract.L | Ordinal factor | -1.708 | 0.18 | For each level change in the contract length( from Month to month to 1 year, 1 year ->2 years), the odds of churn decreases by 82% |
| Dependents | Binary | -0.730 | 0.48 | If customer has dependents, the odds of them churning are 52% less than those without dependents |
| Senior Citizen | Binary | 0.691 | 1.99 | If the customer is a Senior citizen, the odds of churn are 1.99 times higher than non-seniors |

**BOTH**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor name** | **Type** | **β Coefficient Value** | **Exp(β) Value** | **Marginal Effect Interpretation** |
| Contract.L | Ordinal factor | -1.331 | 0.26 | For each level change in the contract length( from Month to month to 1 year, 1 year ->2 years), the odds of churn decreases by 74% |
| Paperless Billing | Binary | 0.355 | 1.42 | If the customer has paperless billing, the odds of churning are 1.42 times higher than those with paper billing |
| Dependents | Binary | -0.231 | 0.79 | If customer has dependents, the odds of them churning are 21% less than those without dependents |

1. **Recall, Precision, F-1 Score & AUC values for each model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Recall** | **Precision** | **F-1 Score** | **AUC** |
| **Phone Only** | 0.994 | 0.925 | 0.958 | 0.514 |
| **Internet Only** | 0.897 | 0.844 | 0.870 | 0.704 |
| **Both** | 0.741 | 0.847 | 0.791 | 0.735 |

AUC Curves for each predictive model

Chart

Description automatically generatedChart

Description automatically generatedChart

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated