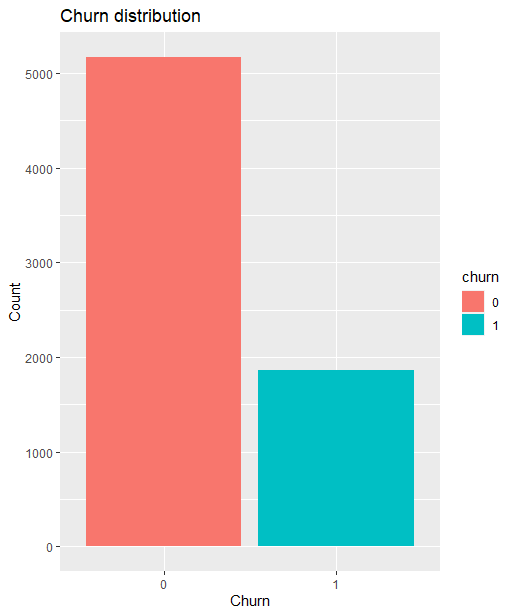
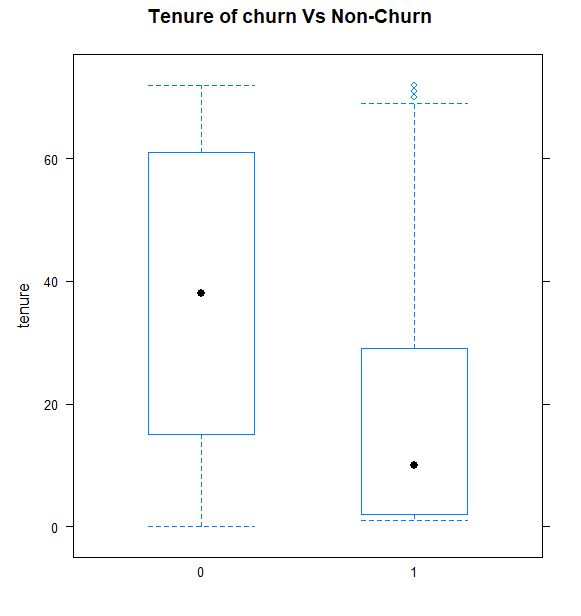
1. Describe the process by which you cleaned, processed, and partitioned data as necessary.

* Converted gender, partner, dependents, phoneservice, paperlessbilling, churn to 1/0.
* Replaced all occurrences of "No phone service" in the "multiplelines" with the value "No".
* Similarly, replaced all occurrences of “No internet service” in columns “onlinesecurity”, “onlinebackup”, “deviceprotection”, “techsupport”, “streamingtv” and “streamingmovies” with the value “No.”
* Converted the above columns to 1 for “Yes” and to 0 for “No.”
* Converted Internet Service, Payment method, Gender, StreamingTv, Phoneservice to factors.
* Handling the NA’s in totalcharge variable: If totalcharges are missing, it is computed as the product of monthlycharges and tenure.
* Created 3 data frames containing only telephone customers, internet customers and customers having both the services.

Chart, box and whisker chart

Description automatically generatedExploratory Data Analysis

Text

Description automatically generated

Chart, histogram

Description automatically generated

This seems to be a highly imbalanced dataset.

The box plot shows that churners have a higher monthly charge than non-churners.

High correlation between tenure and total charges. We can drop total charges and use monthly charges.

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** | **Effect** | **Rationale** |
| *DV: Churn of only Internet Customers* | | |
| SeniorCitizen | +/- | Senior citizens may have lower usage and be less likely to churn |
| Partner | - | Customers with partners may stay for social reasons. |
| Dependents | - | Customers with dependents may stay for family reasons. |
| Tenure | - | Customers with longer tenure may be more loyal to the company. |
| InternetService | - | Customers who have higher-speed Fiber optic may have more specific and demanding requirements. |
| Contract | - | Customers with longer contracts are likely to stay longer. |
| PaperlessBilling | + | Customers who are digitally savvy may switch more frequently |
| MonthlyCharges | + | High monthly charges may prompt customers to switch to cheaper plans |

**Table 1: Predictors for Churn of Only Telephone Customers Table 2: Predictors for Churn of Only Internet Customers**

Explanation: Since we are considering only Internet Service customers, variables related to telephone services such as PhoneService and MultipleLines are excluded. Additionally, variables related to streaming services like StreamingTV and StreamingMovies may not be relevant to all Internet Service customers.

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Effect** | **Rationale** |
| *DV: Churn of only Telephone Customers* | | |
| SeniorCitizen | + | Senior citizens may have lower usage and might not need phone. |
| Partner | - | Customers with partners may stay for social reasons. |
| Dependents | - | Customers with dependents may stay for family reasons. |
| Tenure | - | Customers with longer tenure may be more loyal to the company. |
| PaperlessBilling | + | Customers who are digitally savvy may switch more frequently. |
| MultipleLines | +/- | Additional lines may increase or decrease loyalty and need. |
| MonthlyCharges | + | High monthly charges may prompt customers to switch to cheaper plans |

Explanation: Since we are considering only telephone customers, we can exclude variables related to internet services such as InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies. Similarly, Contract and PaymentMethod are also excluded since they are not relevant to customers with only telephone services.

**Table 3: Predictors for Churn of Customers with Both Phone and Internet Services**

Explanation: Since we are considering customers with both Phone and Internet Services, we need to consider all variables that are relevant to both services. Thus, all variables mentioned in the original description would be included in this table.

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Effect** | **Rationale** |
| *DV: Churn of both the service Customers* | | |
| SeniorCitizen | +/- | Senior citizens may have lower usage and be less likely to churn |
| Partner | - | Customers with partners may stay for social reasons. |
| Dependents | - | Customers with dependents may stay for family reasons. |
| Tenure | - | Customers with longer tenure may be more loyal to the company. |
| MultipleLines | +/- | Additional lines may increase or decrease loyalty and need |
| InternetService | - | Customers who have higher-speed Fiber optic may have more specific and demanding requirements. |
| Contract | - | Customers with longer contracts are likely to stay longer. |
| PaperlessBilling | + | Customers who are digitally savvy may switch more frequently. |
| MonthlyCharges | + | High monthly charges may prompt customers to switch to cheaper plans |

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.

Output of the three trained models

Trained Churn Logit Models

================================================================================

Dependent variable:

-----------------------------------------------------

churn

(1) (2) (3)

--------------------------------------------------------------------------------

seniorcitizen 0.965 (0.637) 0.269 (0.323) 0.216\*\* (0.103)

partner1 -0.243 (0.372) 0.518\* (0.282) -0.074 (0.098)

dependents1 -0.404 (0.367) -0.817\*\* (0.335) -0.001 (0.115)

tenure -0.075\*\*\* (0.012) -0.037\*\*\* (0.008) -0.035\*\*\* (0.003)

paperlessbilling1 0.398 (0.257) 0.357\*\*\* (0.097)

multiplelines -0.637 (1.282) 0.195 (0.216)

internetservice

techsupport -0.858\*\*\* (0.311) -0.507\*\* (0.219)

internetserviceFiber optic 0.395 (0.974)

contractOne year -0.792\*\* (0.372) -0.536\*\*\* (0.137)

contractTwo year -2.035\*\*\* (0.786) -1.195\*\*\* (0.232)

monthlycharges 0.055 (0.234) 0.038\*\* (0.018) 0.021 (0.039)

onlinesecurity -0.486 (0.301) -0.557\*\* (0.219)

onlinebackup -0.391 (0.281) -0.215 (0.214)

deviceprotection -0.270 (0.326) -0.084 (0.216)

streamingmovies 0.018 (0.397)

streamingtv1 0.075 (0.399)

Constant -2.309 (4.666) -0.775 (0.537) -1.487 (1.755)

--------------------------------------------------------------------------------

Observations 1,144 511 3,626

Log Likelihood -236.604 -210.494 -1,750.819

Akaike Inf. Crit. 489.208 444.988 3,535.638

================================================================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

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.

|  |  |  |
| --- | --- | --- |
| Top 3 predictors of churn for telephone customers | Top 3 predictors of churn for internet customers | Top 3 predictors of churn for both service customers |
| Seniorcitizen:  The odds of a senior citizen churning is 2.624 times of the odds of a non-senior citizen having other predictors in the model are invariant. | Dependents:  The odds of a customer churning who has dependents is 0.44 times of the odds of a customer without dependents having other predictors in the model are invariant. | ContractOneYear:  The odds of a customer churning who has contract length of 1 year is 0.42 times of the odds of a customer who has a month-to-month contract, having other predictors in the model are invariant. |
| Multiplelines:  The odds of a customer churning who has multiple lines is 0.5 times of the odds of a customer who do not have multiple lines having other predictors in the model are invariant. | Techsupport:  The odds of a customer churning who has tech support is 0.42 times of the odds of a customer not having tech support, keeping other predictors in the model are invariant. | ContractTwoYear:  The odds of a customer churning who has contract length of 2 year is 0.30 times of the odds of a customer who has a month-to-month contract, having other predictors in the model are invariant. |
| Dependents:  The odds of a customer churning who has dependents is 0.66 times of the odds of a customer without dependents having other predictors in the model are invariant. | ContractOneYear  The odds of a customer churning who has contract length of 1 year is 0.42 times of the odds of a customer who has a month-to-month contract, having other predictors in the model are invariant. | OnlineSecurity:  The odds of a customer churning who has online security is 0.57 times of the odds of a customer who do not have online security, having other predictors in the model are invariant.  TechSupport:  The odds of a customer churning who has tech support is 0.60 times of the odds of a customer not having tech support, keeping other predictors in the model are invariant. |

5. 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.

> print(metrics\_table)

Model Recall Precision F1\_Score AUC Accuracy

1 Model 1 0.02150538 1.0000000 0.04210526 0.8182912 0.9297297

2 Model 2 0.46496815 0.6697248 0.54887218 0.8372412 0.8116170

3 Model 3 0.57702703 0.6640747 0.61749819 0.8121247 0.7726687

The choice of metric to optimize for predicting customer churn depends on the specific business goals and priorities. If the business is primarily concerned with identifying as many customers who are likely to churn as possible (i.e., minimizing false negatives), then Recall would be a suitable metric to optimize.

* Model 1: The recall is 0.02150538, which indicates that only a very small proportion (2.15%) of churn cases for telephone customers are correctly predicted by this model.
* Model 2: The recall is 0.46496815, which suggests that approximately 46.50% of churn cases for internet customers are correctly predicted by this model. This indicates a significantly higher recall compared to Model 1.
* Model 3: The recall is 0.57702703, which indicates that approximately 57.70% of churn cases for customers with both telephone and internet services are correctly predicted by this model. This is the highest recall among the three models, suggesting that Model 3 performs better in terms of capturing churn cases for customers with both types of services.

In summary, Model 3 has the highest recall, followed by Model 2, and Model 1 has the lowest recall among the three models. This implies that Model 3 is the most effective in correctly predicting churn cases, especially for customers with both telephone and internet services.