Telecom Churn

1. **Clean, process, and partition data as necessary, using appropriate R code. (1 point)**

#change seniorcitizen var from 0/1 to No/Yes to match rest of binary vars

df <- df %>%

mutate(seniorcitizen = ifelse(seniorcitizen == "0", "No" , "Yes"))

#change appropriate vars to factors

factorcols <- c("gender", "seniorcitizen", "partner", "dependents", "phoneservice", "multiplelines", "internetservice", "onlinesecurity" ,"onlinebackup", "deviceprotection", "techsupport", "Streamingtv", "streamingmovies", "contract", "paperlessbilling", "paymentmethod","churn")

df[factorcols] <- lapply(df[factorcols], factor)

# Check for missing data, 11 rows have NA for totalcharges

colSums(is.na(df))

# Keep only complete cases (11 rows out of 7043 isn't significant)

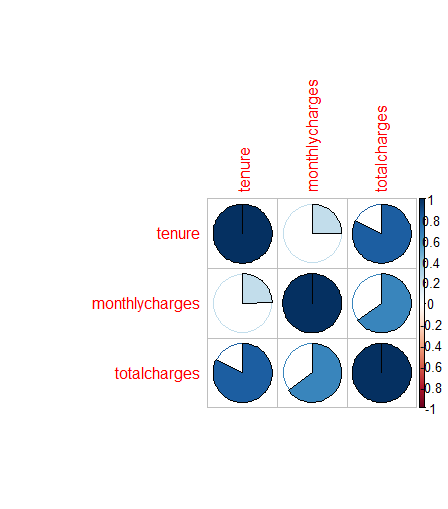
df <- df[complete.cases(df), ]

#test correlation of tenure, monthlycharges, totalcharges

corsub <- df[c("tenure","monthlycharges","totalcharges")]

corrdf <- cor(corsub[1:3])

corrplot(corrdf, method ="pie")



#Partition datasets by service

dfTel <- subset(df, internetservice == 'No')

dfInt <- subset(df, phoneservice == 'No')

dfBoth <- subset(df, phoneservice == 'Yes' & internetservice != 'No')

* The SeniorCitizen variable is changed from a 0/1 to No/Yes to match all of the other binary variables in the dataset.
* Every relevant character variable is transformed into a factor.
* Of the 7043 rows, 11 have the totalcharges variable missing. Since this probably indicates new customers who have not yet made a first payment, these are not very helpful when investigating churn. In addition, 11 out of 7043 is a negligible amount so we can feel ok removing these customers entirely.
* A corrplot shows that tenure and totalcharges are highly correlated, so we should pick one as a predictor to avoid multicollinearity. We’ll choose tenure since totalcharges is also highly correlated with monthlyocharges.
* Lastly, we partition the dataset into 3 sets based upon the values of the phoneservice and internetservice variable.

1. **What predictors do you think contribute to the churn of (i) only telephone customers, (ii) only Internet service customers, and (iii) customers who subscribe to both phone and Internet services? List reasoning for your answer. No points without reasoning. (2 points)**

I. Predictors Affecting Churn in telephone-only customers

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Effect** | **Reasoning** |
| SeniorCitizen | + | Senior citizens are more likely to pass away in any given year, meaning their service will be cancelled. |
| Tenure | - | The longer a customer keeps their service, the more likely they are satisfied and don’t feel the need to change. |
| MultipleLines | - | A customer with multiple lines is likely to feel more tied to their company and there may be a higher perceived barrier to switching. |
| Contract  (One year/  two year) | - | One or two year contracts are likely to reduce churn because customers have less freedom than in month-to-month to cancel whenever they want. |
| Paperless Billing | - | A customer with paperless billing is less likely to see their bills, so they’ll have fewer impetuses to cancel their account. |
| Payment Method  (automatic types) | - | A customer with automatic payments set up is more likely to “set it and forget it”, meaning it’s essentially effortless to keep service going. |
| MonthlyCharges | + | A customer with a higher bill may see more inclined to cancel service for personal finance reasons. |

Notes:

* Although TotalCharges should have a negative effect because of it’s high correlation with Tenure, we’ll leave it out to avoid multicollinearity.

II. Predictors Affecting Churn in internet-only customers

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Effect** | **Reasoning** |
| SeniorCitizen | + | Senior citizens are more likely to pass away in any given year, meaning their service will be cancelled. |
| Tenure | - | The longer a customer keeps their service, the more likely they are satisfied and don’t feel the need to change. |
| OnlineSecurity | - | A customer who is hacked or loses personal information is more likely to want to change internet service providers. |
| TechSupport | - | A customer who receives support from their internet provider likely finds more value and is less likely to sqitch. |
| Contract  (One year/  two year) | - | One or two year contracts are likely to reduce churn because customers have less freedom than in month-to-month to cancel whenever they want. |
| Paperless Billing | - | A customer with paperless billing is less likely to see their bills, so they’ll have fewer impetuses to cancel their account. |
| Payment Method  (automatic types) | - | A customer with automatic payments set up is more likely to “set it and forget it”, meaning it’s essentially effortless to keep service going. |
| MonthlyCharges | + | A customer with a higher bill may see more inclined to cancel service for personal finance reasons. |

Notes:

* Although TotalCharges should have a negative effect because of it’s high correlation with Tenure, we’ll leave it out to avoid multicollinearity.
* internetservice needs to be excluded because internet-only customers only have DSL-type internet
* I don’t believe any of the internet “options” (onlinesecurity, onlinebackup, deviceprotection, techsupport, streamingtv, streamingmovies) will have a significant effect on churn because internet service is a commodity regardless of customer-preferences.

III. Predictors Affecting Churn in customers of both services

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Effect** | **Reasoning** |
| SeniorCitizen | + | Senior citizens are more likely to pass away in any given year, meaning their telephone service will be cancelled. |
| Tenure | - | The longer a customer keeps their service, the more likely they are satisfied and don’t feel the need to change. |
| Contract  (One year/  two year) | - | One or two year contracts are likely to reduce churn because customers have less freedom than in month-to-month to cancel whenever they want. |
| Paperless Billing | - | A customer with paperless billing is less likely to see their bills, so they’ll have fewer impetuses to cancel their account. |
| Payment Method  (automatic types) | - | A customer with automatic payments set up is more likely to “set it and forget it”, meaning it’s essentially effortless to keep service going. |
| MonthlyCharges | + | A customer with a higher bill may see more inclined to cancel service for personal finance reasons. |

Notes:

* These effects should be the intersection of the effects of the telephone-only and internet-only datasets.

1. **Create training and test data sets with a 75:25 split using a random seed of 1024. Train logit models with the variables you identified in (b) and the training data. Combine the three model outputs using stargazer. (1 point)**

set.seed(1024)

sample = sample.split(dfTel, SplitRatio = .75)

dfTelTrain = subset(dfTel, sample == TRUE)

dfTelTest = subset(dfTel, sample == FALSE)

sample = sample.split(dfInt, SplitRatio = .75)

dfIntTrain = subset(dfInt, sample == TRUE)

dfIntTest = subset(dfInt, sample == FALSE)

sample = sample.split(dfBoth, SplitRatio = .75)

dfBothTrain = subset(dfBoth, sample == TRUE)

dfBothTest = subset(dfBoth, sample == FALSE)

TelLOGIT <- glm(churn ~ seniorcitizen + tenure + multiplelines + contract + paperlessbilling + paymentmethod + monthlycharges

, family=binomial (link = "logit"), data=dfTelTrain)

IntLOGIT <- glm(churn ~ seniorcitizen + tenure + onlinesecurity + techsupport + contract + paperlessbilling + paymentmethod + monthlycharges

, family=binomial (link = "logit"), data=dfIntTrain)

BothLOGIT <- glm(churn ~ seniorcitizen + tenure + contract

+ paperlessbilling + paymentmethod + monthlycharges

, family=binomial (link = "logit"), data=dfBothTrain)

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

Dependent variable:

churn

(1) (2) (3)

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

seniorcitizenYes 0.996\* (0.585) 0.704\*\* (0.326) 0.299\*\*\* (0.101)

tenure -0.046\*\*\* (0.013) -0.038\*\*\* (0.009) -0.037\*\*\* (0.003)

multiplelinesYes 0.979 (1.201)

onlinesecurityYes -0.621\*\* (0.313)

techsupportYes -0.506 (0.316)

contractOne year -1.426\*\*\* (0.439) -1.297\*\*\* (0.462) -0.682\*\*\* (0.134)

contractTwo year -1.928\*\*\* (0.599) -2.028\*\* (0.803) -1.628\*\*\* (0.230)

paperlessbillingYes 0.456\* (0.251) 0.451 (0.275) 0.407\*\*\* (0.097)

paymentmethodCredit card (automatic) -0.966\* (0.506) -0.799\* (0.464) -0.080 (0.143)

paymentmethodElectronic check -0.422 (0.483) -0.032 (0.394) 0.413\*\*\* (0.118)

paymentmethodMailed check -0.541 (0.355) -0.654 (0.446) -0.084 (0.152)

monthlycharges -0.175 (0.228) 0.024 (0.015) 0.029\*\*\* (0.003)

Constant 2.733 (4.580) -0.424 (0.610) -2.223\*\*\* (0.229)

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

Observations 1,086 486 3,451

Log Likelihood -232.128 -195.192 -1,727.275

Akaike Inf. Crit. 486.257 414.384 3,474.551

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

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

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 probability. (3 points)**

#Telephone Odds and Probability

cbind( Coeff = coef(TelLOGIT) ,

Odds = exp(coef(TelLOGIT)) ,

Probability = exp(coef(TelLOGIT))/(1+exp(coef(TelLOGIT))) ,

DistancefromEven = abs(0.5 - exp(coef(TelLOGIT))/ (1+exp(coef(TelLOGIT))))

)

Coeff Odds Probability DistancefromEven

(Intercept) 2.73332345 15.3839299 0.9389646 0.43896458

seniorcitizenYes 0.99606262 2.7076000 0.7302837 0.23028374

tenure -0.04587365 0.9551626 0.4885336 0.01146640

multiplelinesYes 0.97881086 2.6612897 0.7268722 0.22687220

contractOne year -1.42569449 0.2403415 0.1937704 0.30622958

contractTwo year -1.92793628 0.1454481 0.1269792 0.37302082

paperlessbillingYes 0.45615104 1.5779887 0.6121007 0.11210070

paymentmethodCredit card (automatic) -0.96555087 0.3807734 0.2757682 0.22423180

paymentmethodElectronic check -0.42190257 0.6557979 0.3960616 0.10393843

paymentmethodMailed check -0.54058282 0.5824087 0.3680520 0.13194799

monthlycharges -0.17527472 0.8392264 0.4562932 0.04370684

#Internet Odds and Probability

cbind( Coeff = coef(IntLOGIT) ,

Odds = exp(coef(IntLOGIT)) ,

Probability = exp(coef(IntLOGIT))/(1+exp(coef(IntLOGIT))) ,

DistancefromEven = abs(0.5 - exp(coef(IntLOGIT))/ (1+exp(coef(dfOmtLOGIT))))

)

Coeff Odds Probability DistancefromEven

(Intercept) -0.42411644 0.6543477 0.3955321 0.104467856

seniorcitizenYes 0.70364832 2.0211129 0.6689962 0.168996154

tenure -0.03830001 0.9624242 0.4904262 0.009573833

onlinesecurityYes -0.62079770 0.5375155 0.3496000 0.150399952

techsupportYes -0.50640081 0.6026608 0.3760376 0.123962362

contractOne year -1.29667547 0.2734393 0.2147251 0.285274938

contractTwo year -2.02750616 0.1316635 0.1163451 0.383654934

paperlessbillingYes 0.45102319 1.5699177 0.6108825 0.110882479

paymentmethodCredit card (automatic) -0.79889012 0.4498279 0.3102630 0.189737018

paymentmethodElectronic check -0.03175586 0.9687431 0.4920617 0.007938299

paymentmethodMailed check -0.65354908 0.5201963 0.3421902 0.157809793

monthlycharges 0.02381015 1.0240959 0.5059523 0.005952257

#Both Odds and Probability

cbind( Coeff = coef(IntLOGIT) ,

Odds = exp(coef(BothLOGIT)) ,

Probability = exp(coef(BothLOGIT))/(1+exp(coef(BothLOGIT))) ,

DistancefromEven = abs(0.5 - exp(coef(BothLOGIT))/ (1+exp(coef(BothLOGIT))))

)

Coeff Odds Probability DistancefromEven

(Intercept) -0.42411644 0.1082428 0.09767069 0.402329310

seniorcitizenYes 0.70364832 1.3490804 0.57430151 0.074301508

tenure -0.03830001 0.9634590 0.49069473 0.009305269

onlinesecurityYes -0.62079770 0.5053992 0.33572437 0.164275633

techsupportYes -0.50640081 0.1962940 0.16408507 0.335914930

contractOne year -1.29667547 1.5029897 0.60047778 0.100477782

contractTwo year -2.02750616 0.9229735 0.47997204 0.020027958

paperlessbillingYes 0.45102319 1.5120190 0.60191384 0.101913836

paymentmethodCredit card (automatic) -0.79889012 0.9193473 0.47898954 0.021010462

paymentmethodElectronic check -0.03175586 1.0293844 0.50723974 0.007239735

paymentmethodMailed check -0.65354908 0.1082428 0.09767069 0.402329310

monthlycharges 0.02381015 1.3490804 0.57430151 0.074301508

i) Telephone-only customers

* Contract: Two year contract = 12.7% probability of churn, One year contract = 19.4% probability of churn
* SeniorCitizen: 73% probability of churn
* PaymentMethod: Credit Card (automatic) payers have 27.6% probability of churn

ii) Internet-only customers

* Contract: Two year contract = 11.6% probability of churn, One year contract = 21.5% probability of churn
* PaymentMethod: Credit Card (automatic) payers have 31% probability of churn
* SeniorCitizen: 66.9% probability of churn

iii) Customers with both services

* Contract: Two year contract = 16.4% probability of churn, One year contract = 33.6%
* PaymentMethod: Electronic check payers have 60.2% probability of churn
* PaperlessBilling: 60% probability of churn

1. **Use TWO metrics to indicate which of these three models in Question 4 has best fit with the training data set and which model has the worst fit? How do you know? (1 point)**

#AIC

AIC(TelLOGIT,IntLOGIT,BothLOGIT)

df AIC

TelLOGIT 11 486.2566

IntLOGIT 12 414.3840

BothLOGIT 10 3474.5509

#PseudoR2

cbind(Telephone=PseudoR2(TelLOGIT, "McFadden"),

Internet=PseudoR2(IntLOGIT, "McFadden"),

Both=PseudoR2(BothLOGIT, "McFadden"))

TelLOGIT IntLOGIT BothLOGIT

McFadden 0.2277562 0.2900314 0.2172361

The best model is the Internet-only model (IntLOGIT), indicated by the lowest AIC and the highest McFadden PseudoR^2 values.

The worst model is the Both-services model (BothLOGIT), indicated by the highest AIC and lowest McFadden PseudoR^2 values.

AIC is an estimator of prediction error, so a lower score is better. The McFadden PsuedoR^2 attempts to replicate an R^2 value, which is the proportion of the variance in the DV that is explained by the IVs, so a higher value is better.

1. **Fit your models using test data, and compute recall, precision, F1-score, and AUC values for each of your three models. Which model worked best for your classification analysis? (2 points)**

#Telephone

predTel <- predict(TelLOGIT, newdata=dfTelTest, type="response")

predTel <- ifelse(predTel > 0.5, 1, 0)

ClassificationError <- mean(predTel != dfTelTest$churn)

# Confusion matrix

CMTel <- table(dfTelTest$churn, predTel)

#Precision

precTel <- CMTel[2,2]/(CMTel[2,2]+CMTel[1,2])

#Recall

recTel <- CMTel[2,2]/(CMTel[2,2]+CMTel[2,1])

#F1-Score

F1Tel <- 2 \* ((recTel \* precTel) / (recTel + precTel))

#AUC

Telpr <- prediction(predTel, dfTelTest$churn)

TelAUC <- performance(Telpr, measure = "auc")

TelAUC <- TelAUC@y.values[[1]]

#-------------------

#Internet

predInt <- predict(IntLOGIT, newdata=dfIntTest, type="response")

predInt <- ifelse(predInt > 0.5, 1, 0)

# Confusion matrix

CMInt <- table(dfIntTest$churn, predInt)

#Precision

precInt <- CMInt[2,2]/(CMInt[2,2]+CMInt[1,2])

#Recall

recInt <- CMInt[2,2]/(CMInt[2,2]+CMInt[2,1])

#F1-Score

F1Int <- 2 \* ((recInt \* precInt) / (recInt + precInt))

#AUC

Intpr <- prediction(predInt, dfIntTest$churn)

IntAUC <- performance(Intpr, measure = "auc")

IntAUC <- IntAUC@y.values[[1]]

#-------------------

#Both

predBoth <- predict(BothLOGIT, newdata=dfBothTest, type="response")

predBoth <- ifelse(predBoth > 0.5, 1, 0)

# Confusion matrix

CMBoth <- table(dfBothTest$churn, predBoth)

#Precision

precBoth <- CMBoth[2,2]/(CMBoth[2,2]+CMBoth[1,2])

#Recall

recBoth <- CMBoth[2,2]/(CMBoth[2,2]+CMBoth[2,1])

#F1-Score

F1Both <- 2 \* ((recBoth \* precBoth) / (recBoth + precBoth))

#AUC

Bothpr <- prediction(predBoth, dfBothTest$churn)

BothAUC <- performance(Bothpr, measure = "auc")

BothAUC <- BothAUC@y.values[[1]]

cbind(precInt,precBoth)

cbind(recInt,recBoth)

cbind(F1Int,F1Both)

cbind(TelAUC,IntAUC,BothAUC)

precInt precBoth

Precision 0.5641026 0.6319797

recInt recBoth

Recall 0.4680851 0.5914489

F1Int F1Both

F1-Score 0.5116279 0.6110429

TelAUC IntAUC BothAUC

AUC 0.5 0.6762194 0.7202036

Precision, Recall, and F1-Score are not calculable for the Telephone-only model because there are no positives in the confusion matrix.

Of the other two models, it is clear that classification ability of the “Both” model is superior to that of the Internet-only model due to Precision, Recall, F1-Score, and AUC that are all higher.