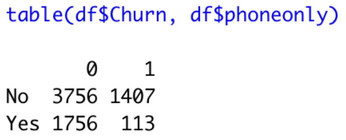
ISM 6137 - Statistical Data Mining
Assignment 7
Sahil Shah – 19895141

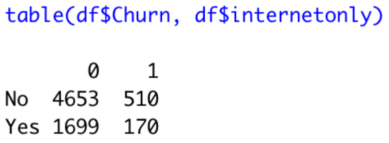
1. Feature engineering & data partitioning

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)

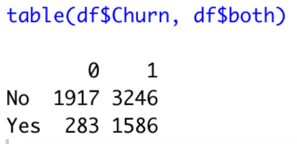
B. Examined tables of the different target cases to understand balance in the data set



```
table(df$Churn, df$phoneonly)

        0    1
No   3756 1407
Yes  1756  113
```



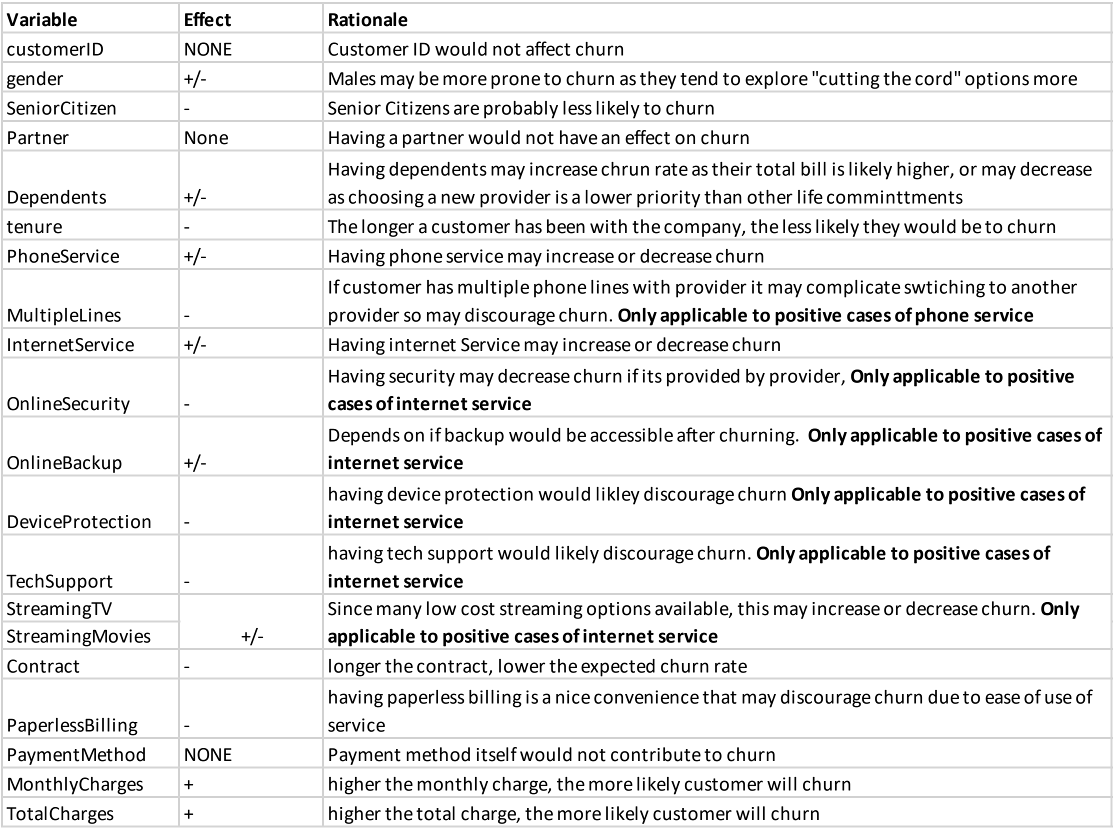
```
table(df$Churn, df$internetonly)

        0    1
No   4653  510
Yes  1699  170
```



```
table(df$Churn, df$both)

        0    1
No   1917 3246
Yes   283 1586
```

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”

2. Predictor Table



Variable	Effect	Rationale
customerID	NONE	Customer ID would not affect churn
gender	+/-	Males may be more prone to churn as they tend to explore "cutting the cord" options more
SeniorCitizen	-	Senior Citizens are probably less likely to churn
Partner	None	Having a partner would not have an effect on churn
Dependents	+/-	Having dependents may increase chrun rate as their total bill is likely higher, or may decrease as choosing a new provider is a lower priority than other life comminttments
tenure	-	The longer a customer has been with the company, the less likely they would be to churn
PhoneService	+/-	Having phone service may increase or decrease churn
MultipleLines	-	If customer has multiple phone lines with provider it may complicate swtiching to another provider so may discourage churn. **Only applicable to positive cases of phone service**
InternetService	+/-	Having internet Service may increase or decrease churn
OnlineSecurity	-	Having security may decrease churn if its provided by provider, **Only applicable to positive cases of internet service**
OnlineBackup	+/-	Depends on if backup would be accessible after churning. **Only applicable to positive cases of internet service**
DeviceProtection	-	having device protection would likley discourage churn **Only applicable to positive cases of internet service**
TechSupport	-	having tech support would likely discourage churn. **Only applicable to positive cases of internet service**
StreamingTV	+/-	Since many low cost streaming options available, this may increase or decrease churn. **Only applicable to positive cases of internet service**
StreamingMovies		
Contract	-	longer the contract, lower the expected churn rate
PaperlessBilling	-	having paperless billing is a nice convenience that may discourage churn due to ease of use of service
PaymentMethod	NONE	Payment method itself would not contribute to churn
MonthlyCharges	+	higher the monthly charge, the more likely customer will churn
TotalCharges	+	higher the total charge, the more likely customer will churn

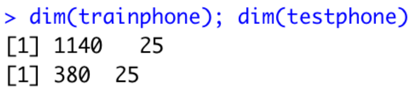
Phone Only, Internet Only, Both and Churn are not included as they are either a target case or the target variable itself. This table can be used for all three cases. **Cases where certain variables are not applicable** (i.e. online backup is not applicable to the phone only case) **are noted in the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

3. Models and Output

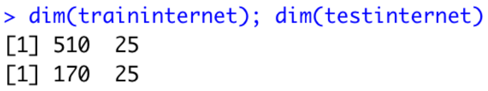
Built three models, one for each case. The models are shown below along with stargazer output

```
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)
```

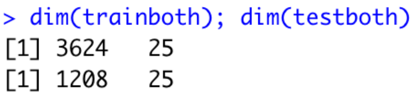
Train-test Split Dimensions for each subset:



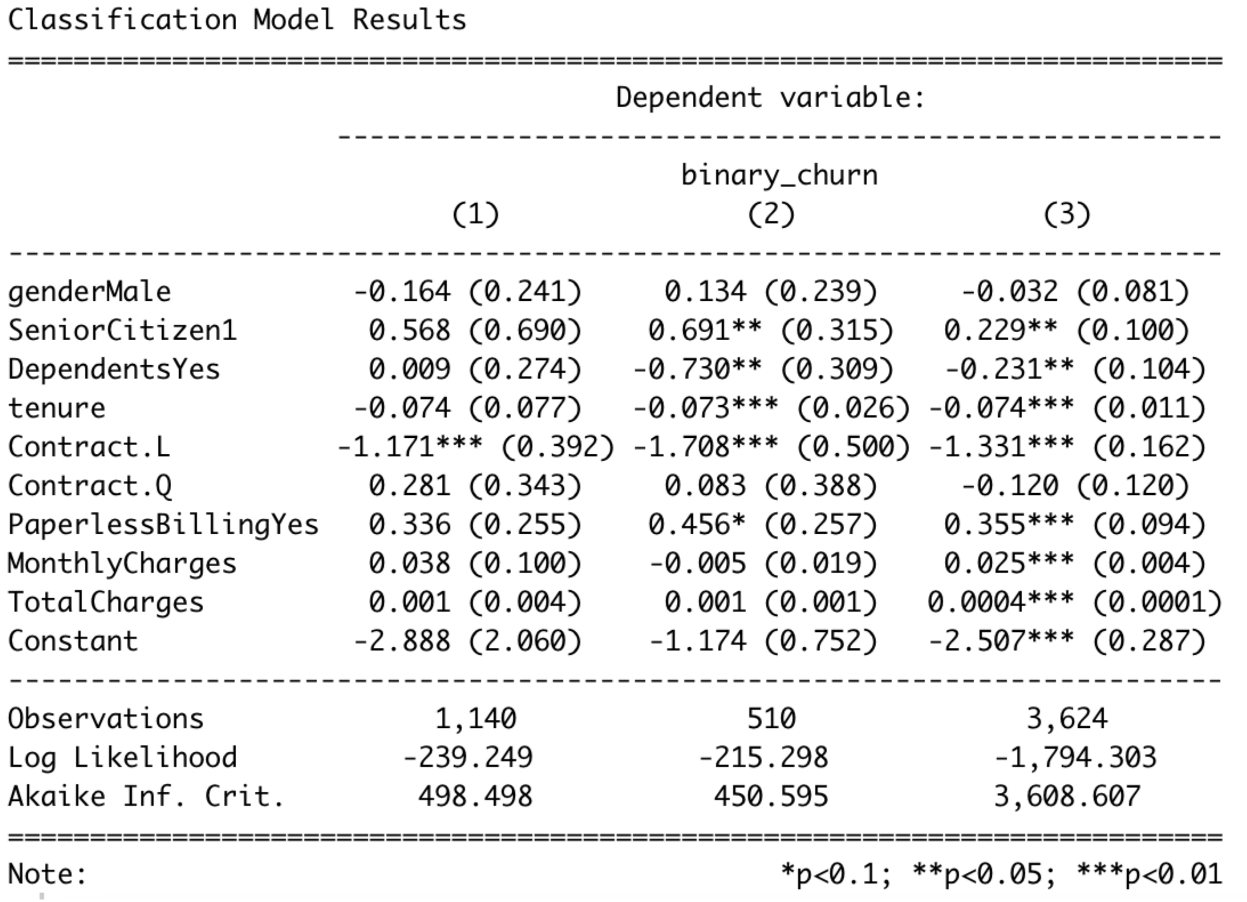
```
> dim(trainphone); dim(testphone)
[1] 1140   25
[1] 380  25
```



```
> dim(traininternet); dim(testinternet)
[1] 510  25
[1] 170  25
```



```
> dim(trainboth); dim(testboth)
[1] 3624   25
[1] 1208   25
```



```
Classification Model Results
=========================================================================
                                      Dependent variable:
                    -----------------------------------------------------
                                         binary_churn
                          (1)               (2)               (3)
-------------------------------------------------------------------------
genderMale          -0.164 (0.241)     0.134 (0.239)     -0.032 (0.081)
SeniorCitizen1       0.568 (0.690)    0.691** (0.315)    0.229** (0.100)
DependentsYes        0.009 (0.274)   -0.730** (0.309)   -0.231** (0.104)
tenure              -0.074 (0.077)  -0.073*** (0.026) -0.074*** (0.011)
Contract.L         -1.171*** (0.392) -1.708*** (0.500) -1.331*** (0.162)
Contract.Q           0.281 (0.343)     0.083 (0.388)     -0.120 (0.120)
PaperlessBillingYes  0.336 (0.255)    0.456* (0.257)    0.355*** (0.094)
MonthlyCharges       0.038 (0.100)    -0.005 (0.019)    0.025*** (0.004)
TotalCharges         0.001 (0.004)     0.001 (0.001)   0.0004*** (0.0001)
Constant            -2.888 (2.060)    -1.174 (0.752)   -2.507*** (0.287)
-------------------------------------------------------------------------
Observations            1,140              510               3,624
Log Likelihood         -239.249          -215.298          -1,794.303
Akaike Inf. Crit.       498.498           450.595           3,608.607
=========================================================================
Note:                                         *p<0.1; **p<0.05; ***p<0.01
```

4. 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

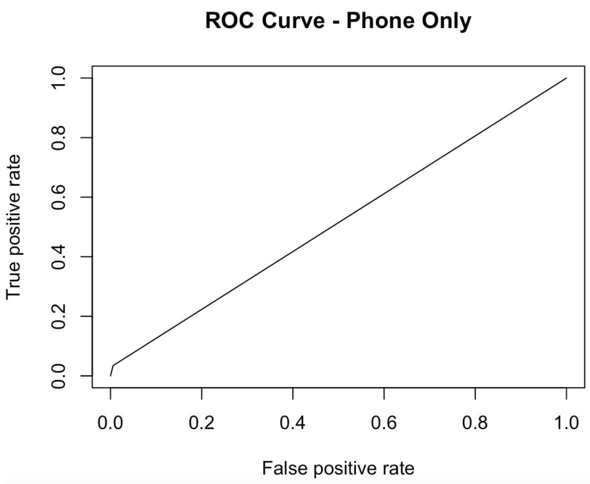
BOTH

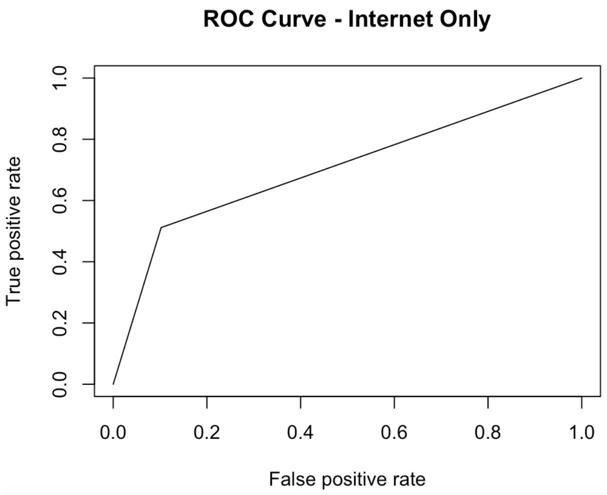
Predictor name	Type	β Coefficient Value	Exp(β) Value	Marginal Effect Interpretation
Contract.L	Ordinal factor	-1.331	0.26	For each level change in the contract length(from Month to month to 1 year, 1 year ->2 years), the odds of churn decreases by 74%
Paperless Billing	Binary	0.355	1.42	If the customer 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

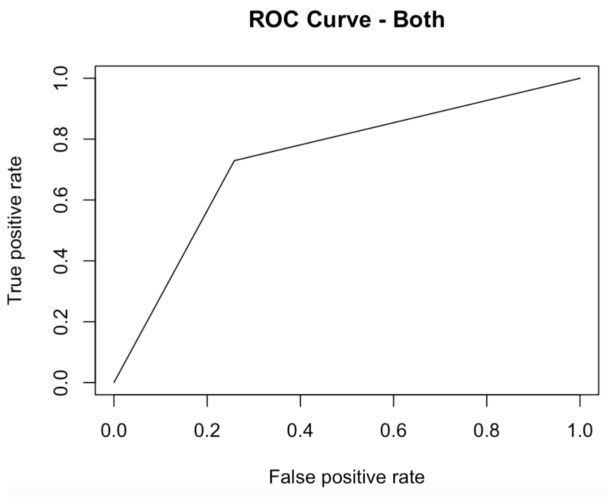
5. Recall, Precision, F-1 Score & AUC values for each model

	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


ROC Curve - Phone Only
True positive rate
False positive rate


ROC Curve - Internet Only
True positive rate
False positive rate


ROC Curve - Both
True positive rate
False positive rate