Customer Churns

Gabrielle Avila, Vincent Chau, Chris Cheng , Joseph Del Val, Hassan Osman

Background

- A public dataset released by IBM detailing the customer retention of a telecommunications company
- A ‘churn’ column reveals whether or not a customer ended their subscription to the company
  - ‘Yes’ if they ended their service, ‘No’ if they remained (at the time of data collection)
- Has a variety of other columns with details about the customer’s subscription

DataSource

- Originally this dataset was on IBM's website, however it was later removed for some unknown reason.
- However, it remains on Kaggle.com free for download.

Research Questions

- What causes consumers to Churn from telecommunications services?
  - We performed univariate screening on all variables that could've lead consumers to churn from their services, concluding that some were more significant than others
- What can be done to help telecommunications companies keep their Churn rates low, leading to more consumers?
  - We looked at several potential predictors and conducted regression analysis to see if they had an overall impact at the Churn rate

CustomerID

- Each customer is given a customerID
- This is unique for every customer and consists of four numbers and five letters.
- This is completely random and unique for each customer, so it doesn't tell us anything.

customerID
5575-GNVDE
7795-CFOCW
9237-HQITU
1452-KIOVK
9763-GRSKD
7469-LKBCI
8091-TTVAX
0280-XJGEX
3655-SNQYZ

gender/SeniorCitizen/

```
> summary(Telco$gender)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.5048  1.0000  1.0000
> summary(Telco$SeniorCitizen)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.1621  0.0000  1.0000
```

- **Gender** is self explanatory. ‘Male’ for males and ‘Female’ for females. 3555 males and 3488 females, for a total of 7043.
- **SeniorCitizen** is a binary column, with a 1 if a customer is a senior citizen and 0 otherwise. 1142 seniors and 5901 non-seniors. (i.e. B(0.1621468, 7043))

Partner/Dependents

- **Partner** is whether or not the customer has a partner (3402 Yes, 3641 No)
- **Dependents** is whether or not the customer has dependents (children) (2110 Yes, 4933 No)

```
> summary(Telco$Partner)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   0.483   1.000   1.000
> summary(Telco$Dependents)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.2996  1.0000  1.0000
```

Tenure

- The “Tenure” column denotes how many months this customer has stayed with the company.

```
> summary(Telco$tenure)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    9.00   29.00   32.37   55.00   72.00
```

PhoneService/MultipleLines/InternetService

PhoneService: Does the customer pay for phone service? Yes/no

MultipleLines: If they have phone service, do they have Multiple lines? (Yes / No / No Phone Service)

InternetService: Does the customer have internet service? If so, what kind? (No / DSL / Fiber Optic)

```
> summary(Telco$PhoneService)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  1.0000  1.0000  0.9032  1.0000  1.0000
```

OnlineSecurity/OnlineBackup/DeviceProtection

These are all dependent on whether or not the customer has internet service. As a result, their values all consist of (Yes / No / No Internet Service)

OnlineSecurity: Does the customer pay for an extra online internet security plan?

OnlineBackup: Does the customer pay for an optional online backup cloud service?

DeviceProtection: Does the customer pay for device protection?

```
> summary(Telco$OnlineSecurity)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.2867  1.0000  1.0000
> summary(Telco$OnlineBackup)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.3449  1.0000  1.0000
> summary(Telco$DeviceProtection)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.3439  1.0000  1.0000
```

TechSupport/StreamingTV/StreamingMovies

These values are dependent on internet service in the same way as the last three.

TechSupport: Does the customer pay extra for tech support?

StreamingTV: Does the customer pay extra to stream TV shows?

StreamingMovies: Does the customer pay extra to stream movies?

```
> summary(Telco$TechSupport)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.2902  1.0000  1.0000
> summary(Telco$StreamingTV)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.3844  1.0000  1.0000
> summary(Telco$StreamingMovies)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.3879  1.0000  1.0000
```

Contract/PaperlessBilling

Contract: This variable denotes the length of the customer's contract with the Telecommunications company, with values "month-to-month", "one year", and "two years".

PaperlessBilling: A simple yes/no column noting if the customer opts for paperless billing, as opposed to Getting their bill in the mail.

```
> summary(Telco$PaperlessBilling)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.5922  1.0000  1.0000
```

PaymentMethod/MonthlyCharges

PaymentMethod: Details the payment method the customer is using to pay their bills (credit card, electronic check, mailed check, bank transfer, etc.)

MonthlyCharges: This is how much the customer is currently being charged per month.

```
> summary(Telco$MonthlyCharges)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.25   35.50   70.35   64.76   89.85  118.75
```

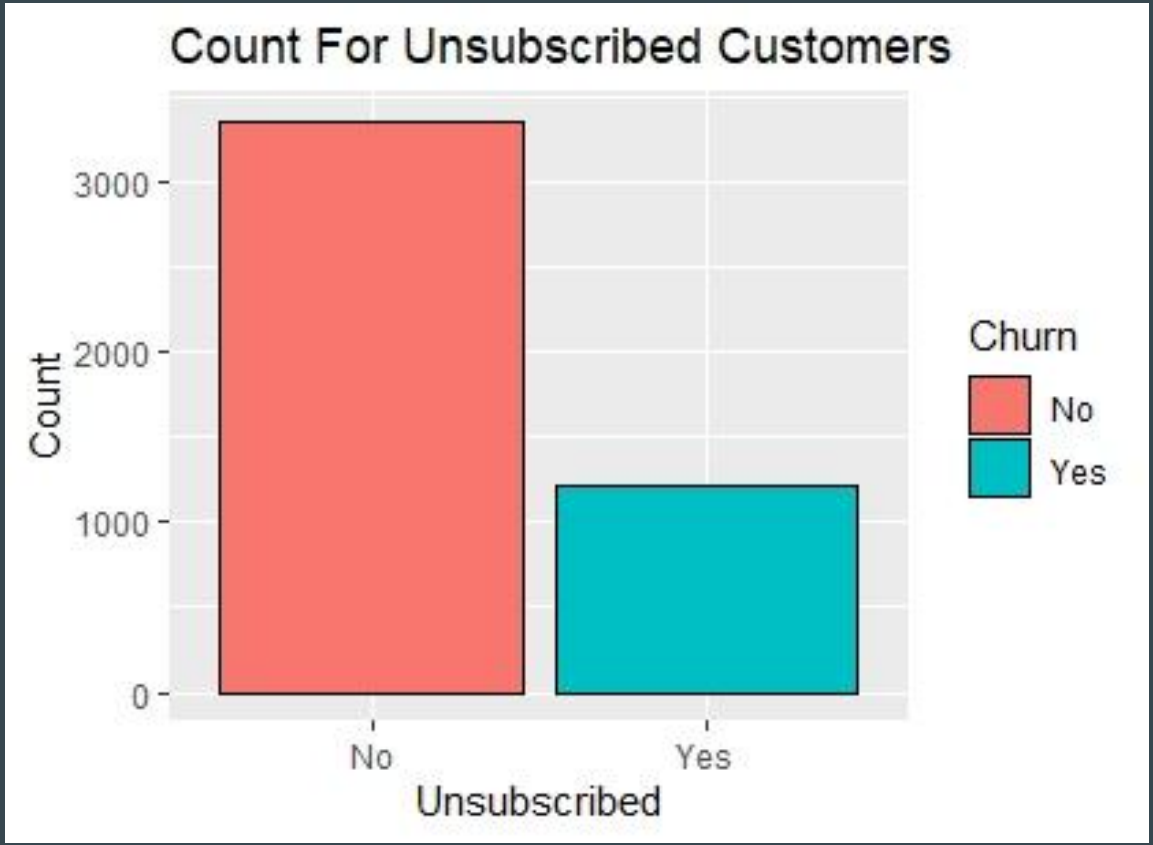
TotalCharges/Churn

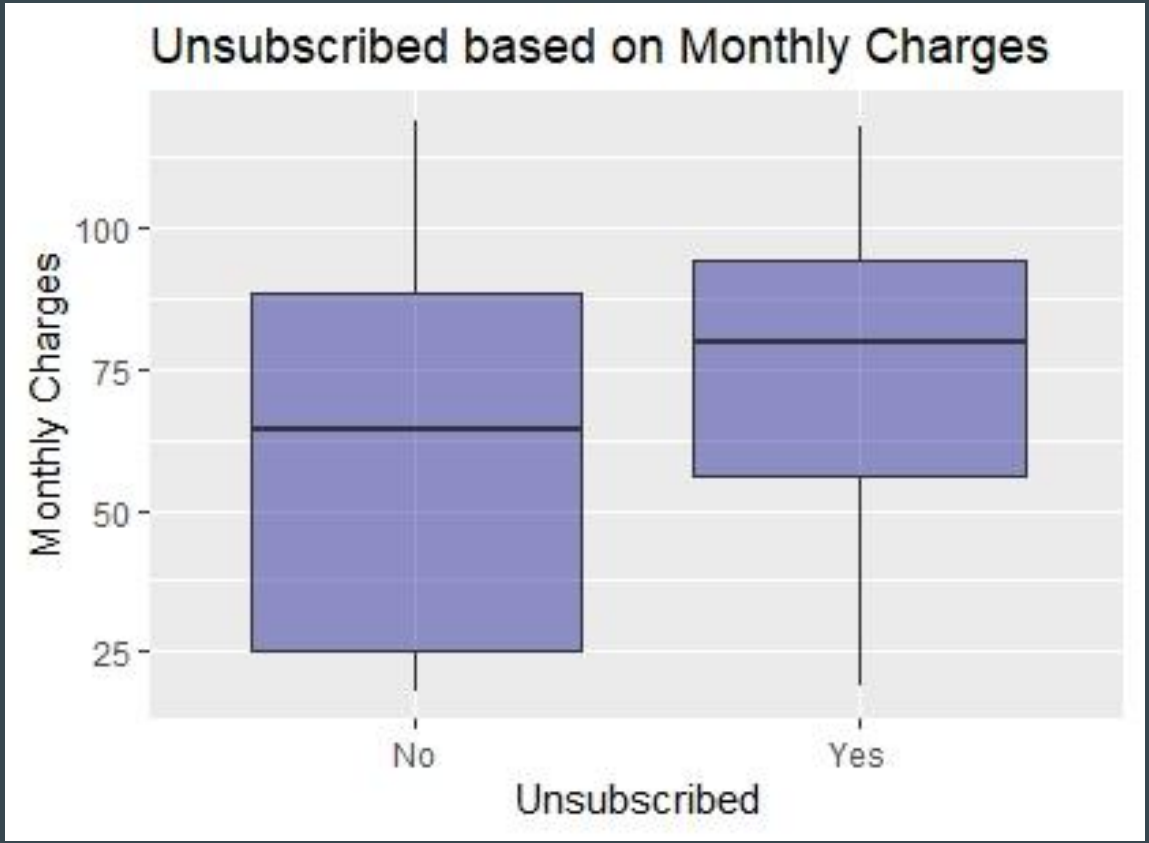
TotalCharges: This is the sum total the customer has been charged thus far.

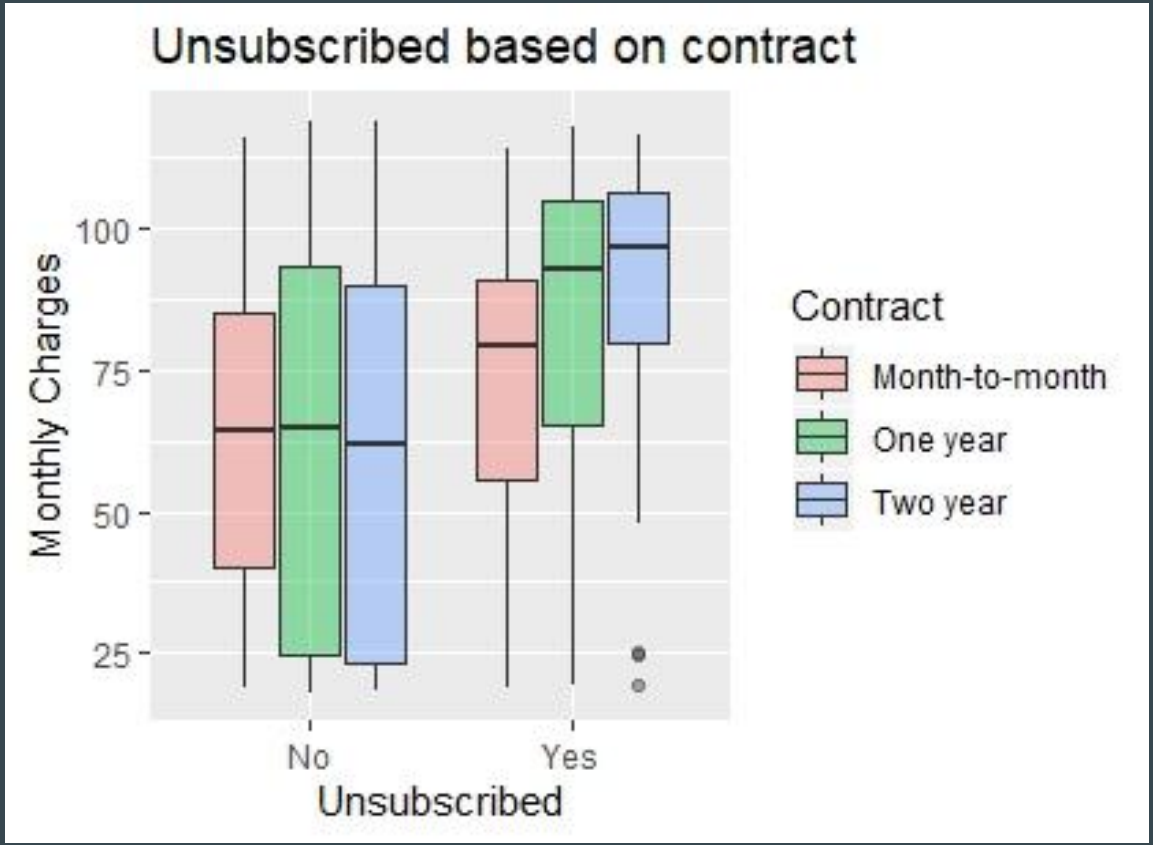
Churn: Whether or not a customer churned.

(1 meaning they did, 0 meaning they didn't)

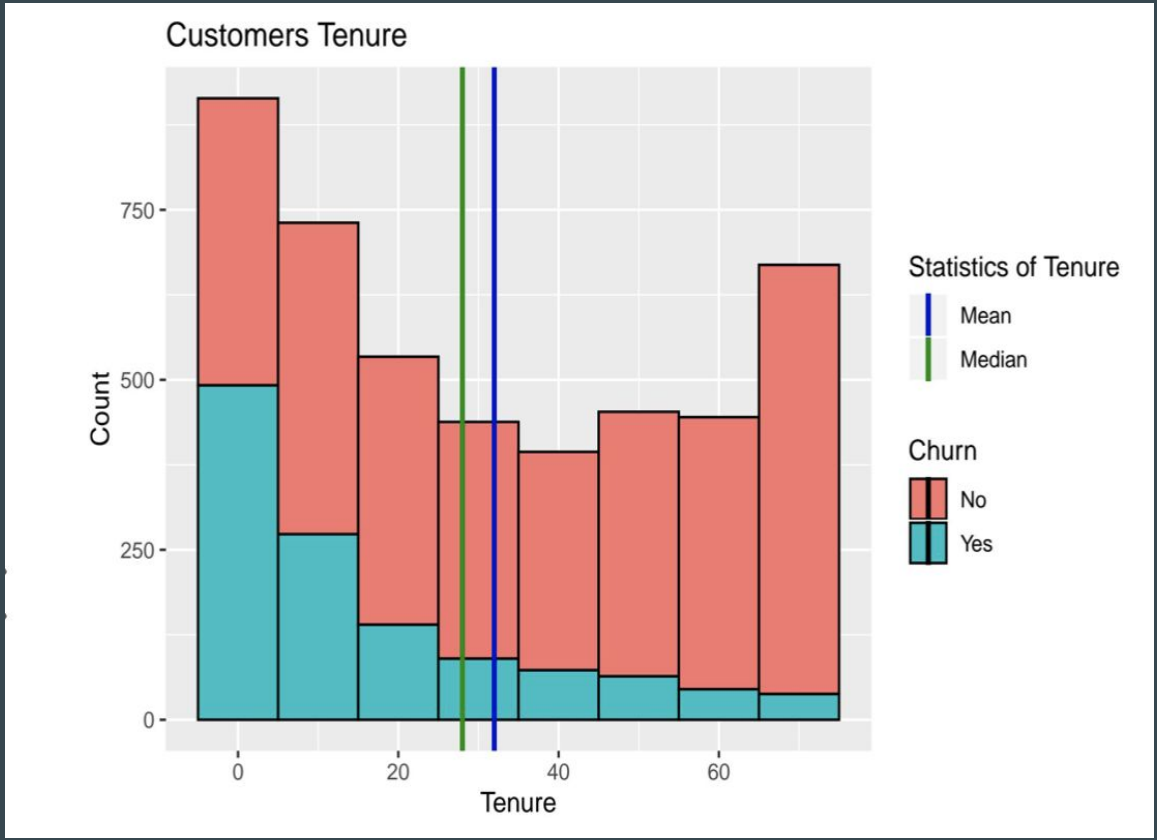
```
> summary(Telco$TotalCharges)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   18.8   401.4  1397.5  2283.3  3794.7  8684.8      11
> summary(Telco$Churn)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.2654  1.0000  1.0000
```


Count For Unsubscribed Customers
3000
2000
1000
0
Count
No
Yes
Unsubscribed
Churn
No
Yes


Unsubscribed based on Monthly Charges
Monthly Charges
100
75
50
25
No
Yes
Unsubscribed


Unsubscribed based on contract
Monthly Charges
100
75
50
25
No
Yes
Unsubscribed
Contract
Month-to-month
One year
Two year


Loyalty Based On Contract Length
Density
0.04
0.03
0.02
0.01
0.00
0
20
40
60
Length Of Subscribtion
Contract
Month-to-month
One year
Two year


Customers Tenure
750
500
250
0
Count
0
20
40
60
Tenure
Statistics of Tenure
Mean
Median
Churn
No
Yes

Logistic Regression w/ Backwards Selection

Our goal was to predict which customers customers churned, and to learn what factors were important in determining this outcome.

Using R's built-in one hot encoding on the categorical variables and ignoring our CustomerID column, we began backwards selection.

Performing Backwards Selection

After finding the P-Values for each of the variables, we began trimming the variable with the highest p-value >0.05, and then refitting the model and analyzing the p-values again.

Our first trim was OnlineBackup with a p-value of 0.9629.

Subsequent iterations took out Gender, Partner, OnlineSecurity, Dependents, and TechSupport.

Variable	P-Value
gender	0.85
SeniorCitizen	0.005
partner	0.612
Dependents	0.3227
tenure	< 2.2e-16
PhoneService	4.5475e-13
MultipleLines	0.01012
InternetService	0.04251
OnlineSecurity	0.6391
OnlineBackup	0.9629
DeviceProtection	0.1393
TechSupport	0.3194
StreamingTV	0.1136
StreamingMovies	0.09704
Contract	6.125e-13
PaperlessBilling	5.417e-05
PaymentMethod	0.005133
MonthlyCharges	0.1975
TotalCharges	8.104e-06

Final Model

Our final model was left with the following variables, each with p-values <0.05. One (Phone Service) was rounded to zero by R.

The resulting model had an R2 value of 0.2963907 and an AIC of 3761.1

Using 10-Fold Cross Validation, we found that the AUC for this model was 0.8460413, which makes this a pretty accurate predictor!

Variable	P-Value
SeniorCitizen	0.002204
tenure	< 2.2e-16
PhoneService	0
MultipleLines	3.361e-09
InternetService	< 2.2e-16
DeviceProtection	0.000313
StreamingTV	3.233e-08
StreamingMovies	5.315e-09
Contract	3.138e-14
PaperlessBilling	4.148e-05
PaymentMethod	0.003476
MonthlyCharges	3.954e-10
TotalCharges	4.843e-06

Conclusions and Interpretations

- Although our backwards selection is limited by its relative inability to account for collinearity, we can still some make some cautious conclusions about the roles of some of these variables.
- Extra packages like tech support and online security seem relatively unimportant, in addition to some of the demographic info, such as gender, partner, and dependents.
- What appears to be most important is if they have Phone and/or Internet Service, and the length of time they have stayed with this company.
    - Lacking a phone or internet makes a customer more likely to cancel
    - Longer-term subscribers are much less likely to cancel than new subscribers