Understanding and Predicting Customer Churns

Jason Ji, Steven Si

Background

A telecommunications company is concerned about the number of customers leaving their landline business for cable competitor.

Using the dataset provided, our goal is to help the company understand what kind of customers are leaving and build a model to predict employee churns.

Dataset Description

Source: https://zhang-datasets.s3.us-east-2.amazonaws.com/telcoChurn.csv

Important columns:

- Churn (whether a customer has left)
- Tenure (how long they've been a customer)
- Services that each customer has signed up for (phone, multiple lines, internet, online ...)
- Customer account information (contract, payment method, paperless billing ...)
- Demographic info about customers (gender, age range, and if they have partners and dependents)

````
```{r}
#Loading data
churn <- read.csv('https://zhang-datasets.s3.us-east-2.amazonaws.com/telcoChurn.csv')
summary(churn)
```
````

```
  customerID           gender          SeniorCitizen      Partner            Dependents            tenure       PhoneService       MultipleLines
 Length:7043        Length:7043        Min.   :0.0000   Length:7043        Length:7043        Min.   : 0.00   Length:7043        Length:7043
 Class :character   Class :character   1st Qu.:0.0000   Class :character   Class :character   1st Qu.: 9.00   Class :character   Class :character
 Mode  :character   Mode  :character   Median :0.0000   Mode  :character   Mode  :character   Median :29.00   Mode  :character   Mode  :character
                                       Mean   :0.1621                                         Mean   :32.37
                                       3rd Qu.:0.0000                                         3rd Qu.:55.00
                                       Max.   :1.0000                                         Max.   :72.00

 InternetService    OnlineSecurity     OnlineBackup       DeviceProtection   TechSupport        StreamingTV        StreamingMovies      Contract
 Length:7043        Length:7043        Length:7043        Length:7043        Length:7043        Length:7043        Length:7043        Length:7043
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character

 PaperlessBilling   PaymentMethod      MonthlyCharges    TotalCharges       Churn
 Length:7043        Length:7043        Min.   : 18.25   Min.   :  18.8   Length:7043
 Class :character   Class :character   1st Qu.: 35.50   1st Qu.: 401.4   Class :character
 Mode  :character   Mode  :character   Median : 70.35   Median :1397.5   Mode  :character
                                       Mean   : 64.76   Mean   :2283.3
                                       3rd Qu.: 89.85   3rd Qu.:3794.7
                                       Max.   :118.75   Max.   :8684.8
                                                        NA's   :11
```

Data Preprocessing

Removed columns that are highly correlated.

- Attribute <PhoneService> shows whether an account has phone service or not. Because another attribute, <MultipleLines> shows whether an account has multiple lines if it has phone service. So by looking at <MultipleLines>, we will know the value for <PhoneService>.

Removed customer id.

```
# drop customer id and phone service column
churn_new = churn[-c(1,7)]
```

Inferences – Generalized Linear Model

m = glm(factor(Churn)~., data=churn_new, family=binomial)

summary(m)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.337e+00	1.439e+00	0.929	0.35276	
genderMale	-2.183e-02	6.480e-02	-0.337	0.73619	
SeniorCitizen	2.168e-01	8.453e-02	2.564	0.01033	*
PartnerYes	-3.840e-04	7.783e-02	-0.005	0.99606	
DependentsYes	-1.485e-01	8.973e-02	-1.655	0.09796	.
tenure	-6.059e-02	6.236e-03	-9.716	< 2e-16	***
MultipleLinesNo phone service	-1.715e-01	6.487e-01	-0.264	0.79153	
MultipleLinesYes	4.484e-01	1.773e-01	2.530	0.01142	*
InternetServiceFiber optic	1.747e+00	7.981e-01	2.190	0.02855	*
InternetServiceNo	-1.786e+00	8.073e-01	-2.213	0.02691	*
OnlineSecurityNo internet service	NA	NA	NA	NA	
OnlineSecurityYes	-2.054e-01	1.787e-01	-1.150	0.25031	
OnlineBackupNo internet service	NA	NA	NA	NA	
OnlineBackupYes	2.604e-02	1.754e-01	0.148	0.88197	
DeviceProtectionNo internet service	NA	NA	NA	NA	
DeviceProtectionYes	1.474e-01	1.764e-01	0.836	0.40339	
TechSupportNo internet service	NA	NA	NA	NA	
TechSupportYes	-1.805e-01	1.806e-01	-0.999	0.31759	
StreamingTVNo internet service	NA	NA	NA	NA	
StreamingTVYes	5.905e-01	3.263e-01	1.810	0.07035	.
StreamingMoviesNo internet service	NA	NA	NA	NA	
StreamingMoviesYes	5.993e-01	3.267e-01	1.834	0.06658	.
ContractOne year	-6.608e-01	1.076e-01	-6.142	8.15e-10	***
ContractTwo year	-1.357e+00	1.764e-01	-7.691	1.46e-14	***
PaperlessBillingYes	3.424e-01	7.450e-02	4.596	4.31e-06	***
PaymentMethodCredit card (automatic)	-8.779e-02	1.141e-01	-0.770	0.44156	
PaymentMethodElectronic check	3.045e-01	9.450e-02	3.222	0.00127	**
PaymentMethodMailed check	-5.759e-02	1.149e-01	-0.501	0.61627	
MonthlyCharges	-4.034e-02	3.176e-02	-1.270	0.20392	
TotalCharges	3.289e-04	7.063e-05	4.657	3.20e-06	***

Inferences – Generalized Linear Model

m = glm(factor(Churn)~., data=churn_new, family=binomial)

summary(m)

```
                                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)                          1.337e+00  1.439e+00   0.929  0.35276    
genderMale                          -2.183e-02  6.480e-02  -0.337  0.73619    
SeniorCitizen                        2.168e-01  8.453e-02   2.564  0.01033 *  
PartnerYes                          -3.840e-04  7.783e-02  -0.005  0.99606    
DependentsYes                       -1.485e-01  8.973e-02  -1.655  0.09796 .  
tenure                              -6.059e-02  6.236e-03  -9.716  < 2e-16 ***
MultipleLinesNo phone service       -1.715e-01  6.487e-01  -0.264  0.79153    
MultipleLinesYes                     4.484e-01  1.773e-01   2.530  0.01142 *  
InternetServiceFiber optic           1.747e+00  7.981e-01   2.190  0.02855 *  
InternetServiceNo                   -1.786e+00  8.073e-01  -2.213  0.02691 *  
OnlineSecurityNo internet service           NA         NA      NA       NA    
OnlineSecurityYes                   -2.054e-01  1.787e-01  -1.150  0.25031    
OnlineBackupNo internet service             NA         NA      NA       NA    
OnlineBackupYes                      2.604e-02  1.754e-01   0.148  0.88197    
DeviceProtectionNo internet service         NA         NA      NA       NA    
DeviceProtectionYes                  1.474e-01  1.764e-01   0.836  0.40339    
TechSupportNo internet service              NA         NA      NA       NA    
TechSupportYes                      -1.805e-01  1.806e-01  -0.999  0.31759    
StreamingTVNo internet service              NA         NA      NA       NA    
StreamingTVYes                       5.905e-01  3.263e-01   1.810  0.07035 .  
StreamingMoviesNo internet service          NA         NA      NA       NA    
StreamingMoviesYes                   5.993e-01  3.267e-01   1.834  0.06658 .  
ContractOne year                    -6.608e-01  1.076e-01  -6.142 8.15e-10 ***
ContractTwo year                    -1.357e+00  1.764e-01  -7.691 1.46e-14 ***
PaperlessBillingYes                  3.424e-01  7.450e-02   4.596 4.31e-06 ***
PaymentMethodCredit card (automatic) -8.779e-02  1.141e-01  -0.770  0.44156    
PaymentMethodElectronic check        3.045e-01  9.450e-02   3.222  0.00127 ** 
PaymentMethodMailed check           -5.759e-02  1.149e-01  -0.501  0.61627    
MonthlyCharges                      -4.034e-02  3.176e-02  -1.270  0.20392    
TotalCharges                         3.289e-04  7.063e-05   4.657 3.20e-06 ***
```

Inferences

````
```{r}
exp(m$coefficients)
```

                (Intercept)                         genderMale                        SeniorCitizen                   PartnerYes
                  3.8066718                          0.9784039                            1.2420647                    0.9996161
              DependentsYes                             tenure            MultipleLinesNo phone service             MultipleLinesYes
                  0.8620105                          0.9412113                            0.8424274                    1.5657978
   InternetServiceFiber optic                  InternetServiceNo    OnlineSecurityNo internet service            OnlineSecurityYes
                  5.7400901                          0.1675800                                   NA                    0.8143052
OnlineBackupNo internet service                    OnlineBackupYes  DeviceProtectionNo internet service          DeviceProtectionYes
                         NA                          1.0263839                                   NA                    1.1587884
 TechSupportNo internet service                     TechSupportYes        StreamingTVNo internet service               StreamingTVYes
                         NA                          0.8348553                                   NA                    1.8049040
StreamingMoviesNo internet service              StreamingMoviesYes                     ContractOne year             ContractTwo year
                         NA                          1.8208360                            0.5164405                    0.2574046
         PaperlessBillingYes PaymentMethodCredit card (automatic)        PaymentMethodElectronic check    PaymentMethodMailed check
                  1.4082582                          0.9159515                            1.3559025                    0.9440396
             MonthlyCharges                       TotalCharges
                  0.9604594                          1.0003290
````

SeniorCitizen: odds of churn for a senior citizen increases by 24%

tenure: if the year of being a customer increases by 1 unit, the odds of churn decreases by 6%

MultipleLinesYes: if customer signs up for multiple lines as opposed to a single line, the odds of churn increases by 56%

InternetServiceFiber optic: if customer signs up for fiber optic internet service as opposed to DSL, the odds of churn increases by 474%

InternetServiceNo: if customer has no internet service as opposed to DSL, the odds of churn decreases by 83%

ContractOne year: if customer signs up for a one year contract as opposed to month-to-month contract, the odds of churn decreases by 48%.

ContractTwo year: if customer signs up for a two year contract as opposed to month-to-month contract, the odds of churn decreases by 74%.

PaperlessBillingYes: if customer signs up for paperless billing as opposed to no paperless billing, the odds of churn increases by 41%

PaymentMethodElectronic check: if customer chooses to pay by electronic check as opposed to automatic bank transfer, the odds of churn increases by 36%

TotalCharges: if customer's total charge increases by 1 unit, the odds of churn increases by 0.03%.

Retention Plan – Decrease Churn & Increase Revenue

New Strategy Targeting Senior Citizens

- Customer segments that have higher churn rate
- Ex. creating internet packages that are more attractive to seniors.

Increase Tenure

- make sure customer start with our landline product and never switch to cable business.

Sell Longer Contract

- market and sell longer 2 year contract, such as by offering discount, as they could drastically decrease odds of churn (74%).

Paperless billing & Electronic Check Payment

- actually increase churn rate (electronic/online payment makes it convenient for customers to stop the payment and switch?)
- make no paperless billing and automatic bank transfer as the default payment method to decrease churn rate.

DSL/Single Line?

- Preferred over multiple lines and other types of internet services
- however, might decrease the overall sales/revenues based on how the services are charged
- additional research to identify the best approach to decrease churn and improve revenues at the same time.

Prediction Model - Data Preprocessing

- Omitted N/A observations
- Transformed outcome variable into binary 0 and 1
- Split Training and Testing datasets
- Further splitting the x and y within training and testing sets

Prediction Model - Deep Neural Networks

Network Architecture:

- Input layer: units = 256, activation = relu
- Dropout 0.2
- Hidden layer one: units = 256, activation = relu
- Dropout 0.2
- Hidden layer two: units = 64, activation = relu
- Dropout 0.3
- Output layer: unit = 1, activation = sigmoid

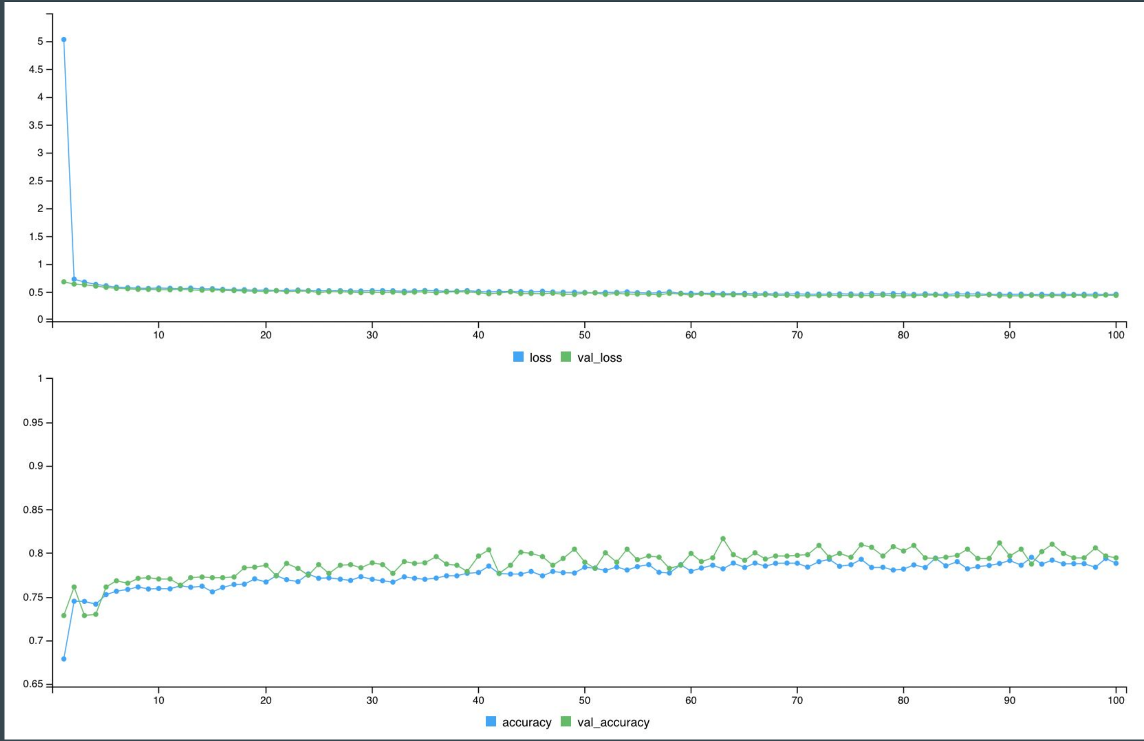
Standardized Hyperparameters:

- Binary cross-entropy loss
- Adam optimizer
- Accuracy as metrics
- Batch Size = 32

Prediction Model - Deep Neural Networks (Cont.)

Tuned Hyperparameters

1. Epoch = 65 (based on right graph)
2. Threshold = 0.25 (based on accuracy)



Prediction Model - Evaluation

- Given a customer left, our model will predict the customer to be leaving correct around 79% of the time
- Given a customer didn't leave, our model will predict the customer to stay correctly around 75% of the time

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 764  79
         1 261 303

               Accuracy : 0.7584
                 95% CI : (0.7351, 0.7805)
    No Information Rate : 0.7285
    P-Value [Acc > NIR] : 0.005985

                  Kappa : 0.4685

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7932
            Specificity : 0.7454
         Pos Pred Value : 0.5372
         Neg Pred Value : 0.9063
              Precision : 0.5372
                 Recall : 0.7932
                     F1 : 0.6406
             Prevalence : 0.2715
         Detection Rate : 0.2154
   Detection Prevalence : 0.4009
      Balanced Accuracy : 0.7693

       'Positive' Class : 1
```

Prediction Model - Findings

- Business Objectives: given the trade-off between minimizing false negative and false positive, the company would choose to minimize cases where the model predict the customer to stay while he is actually leaving (false negative). As a result, the priority is to have high sensitivity value

- Model Evaluation: In this case, our data has a outcome variable imbalance problem. Meaning that even tho y = 1 (churn) is our positive response, our train & test data contains disproportionately small number of them, causing our model not being able to most effectively identify them.

Prediction Model - Recommendations

- From a model building perspective, in order to build a more accurate model, a balanced dataset needs to be provided from our data gathering team inside our company
- Deep learning models are data-hungry. A larger scale of data will also help promote the model accuracy

```
print("The number of rows with y = 0 is:")
sum(y == 0)
print("The number of rows with y = 1 is:")
sum(y == 1)
```

```
[1] "The number of rows with y = 0 is:"
[1] 5163
[1] "The number of rows with y = 1 is:"
[1] 1869
```