Analyzing Customer Data to Understand Retention Rate

Springboard Capstone

Brad Boldenow

2/7/2017

The client for this project is Telco, a (fictitious) telecommunications company. Telco is concerned with the level of churn throughout their customer base. According to Investopedia, the definition of churn rate is, “the percentage of subscribers to a service who discontinue their subscriptions to that service within a given time period.”

Given today’s subscription based economy, churn rate has become a metric of interest for companies large and small. The ability to retain customers over periods of time equates to consistent and predictable revenue forecasts. It is widely known that it costs nearly 10 times as much money to retain a customer than sign a new deal.

Because of this, Telco has reached out to me in need of some predictive analytical insights. Fortunately, Telco understands the importance of using data as an asset and has kept robust and full data on its current and past customer base. It is now my job to use this customer data to better understand and improve Telco’s customer retention.

I will be using logistic regression modeling to understand which variables influence the probability of churning and use this information to create a predictive model. Telco will then be able to use this model to identify customers at high risk of churning prior to them opting out of their contract. This information can be used for a variety of programs. Two specific programs that will be utilized are:

1. The Telco customer service representative dashboard will now include the information gained from the model. When a rep is pulling up customer profile’s they will be able to see whether the customer is at high risk of churning. If the customer is at high risk, the representative will present the customer with a special retention promotion.
2. A digital advertising campaign will be created that targets predicted churners online. The advertisements will include promotional deals such as free streaming TV if they click on the ad and renew their contract today.

The original dataset used for this project includes 7,043 observations (customers) and 21 descriptive variables. Please see [the appendix](#Appendix) for a full list of variables provided. The variables include demographic information (gender, senior citizen) and information regarding the customers services/plan (Tenure, Phone Service, Internet Service, Charges). The target variable being addressed is Churn, which states whether or not that customer churned.

This dataset does pose potential limitations to the analysis. For example, there are no time stamps for these customers. Although the tenure is provided, we have no insights into the time frame of when the customers were receiving services. This information could have played a role in the analysis to predict any seasonality trends or trends related to outside events.

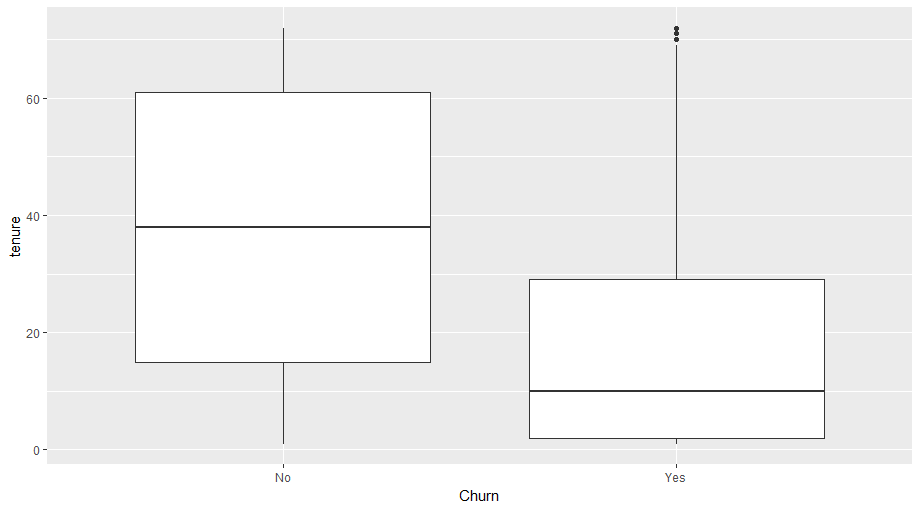
Another potential limitation involves the lack of demographic information provided. The bulk of the variables in this dataset describe the services and contracts for the customers. Not enough information is provided for us to analyze if any demographic trends are influential on the probability of churning.

As mentioned earlier, Telco has done a superb job in recording complete records for their customers. Because of this, the dataset being analyzed only has 11 missing values. These missing values all fall under the total charges variable. After analyzing these 11 observations with missing values, there is no clear formulation as to why they are not recorded. I concluded it was mostly due to human error. Because it was only 11 observations that had missing values, I removed them to give the dataset a new total of 7,032 complete observations.

After omitting these observations, the following additional data cleaning and wrangling tasks were performed:

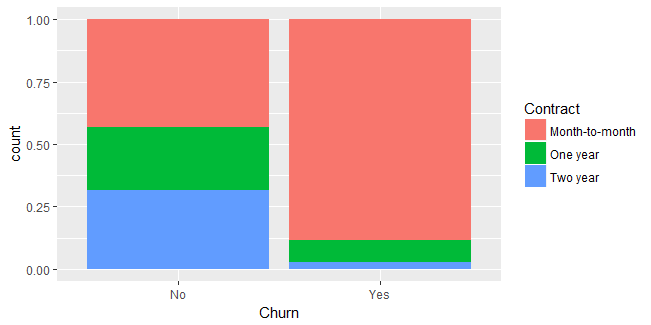
* The senior citizen variable, which has values of 0 (not a senior citizen) and 1 (senior citizen) was originally classified as an integer. We will need to transform this variable to a factor with 2 levels.
* Next we will add a churn\_yes column with the values 1 if the customer churned, and 0 if the customer did not churn. This variable will be used in our analysis of the predictive model.
* For each of the services offered by Telco (phone, multiple lines, internet, online security, online back-up, device protection, tech support, streaming tv, and streaming movies) we will add a binary column with the value 1 if the customer has the corresponding service and a 0 if they do not. This column will be used to analyze the total number of services for each customer.
* Using these newly created binary variables, we will add two new calculated columns. The first column will be a sum of all the possible services. The second calculated column will only sum the binary values for phone and internet service. These are the two core services offered by Telco and will want to be looked at separately.

Because the target variable of interest is Churn, we will start our exploration there. Of the 7,032 observations, 1,869 (27%) churned and 5,163 (73%) did not. When comparing summary statistics from these two groups, one notable difference was in regards to tenure. For those who churned the average tenure is only 18 while the mean tenure for those who did not is 38, representing a 111% difference.



(figure 1: Box plot depicting tenure for both levels of churn)

Another discrepancy between the two subsets lies within the proportion of contract types. Of the observations that did churn, 89% had month-to-month contracts, 8% had one year contracts, and 3% had two year contracts. Compare these statistics with those who did not churn; 43% had month-to-month contracts, 25% had one year contracts, and 32% had two year contracts. The graph below visually depicts the differences in these proportions.



(figure 2: Bar chart depicting proportion of contract types for churners and non-churners.)

It is important to note when examining these figures that the subset of observations who did not churn is much larger than those who did (5,163 and 1,869 respectively).

Next, we used a correlation table to analyze the correlation between the different variables of the dataset (note: in order to do this we created binary variables (1, 0) for the services and other categorical variables).

In order to minimize the level of multicollinearity in our model, we analyzed the correlation between the different independent (predictor) variables. The independent variables with high levels of correlation include tenure and total charges (r = .83), total charges and monthly charges (r = .65), internet service and monthly charges (r = .76), sum (number) of services with monthly charges (r = .85) and total charges (r = .78), and tenure and contract bin (r = .67)

Using this correlation table, we can also analyze the variables that seem highly correlated with the dependent variable, churn. The independent variables that have high correlation values are contract bin (r = -.40), tenure (r = -.35), internet service (r = .23), sum (number) of core services (r = .21), and monthly charges (r = .20).

Another measure I looked at was the information gain for each of the predictor variables. I wanted to analyze these statistics in order to further gauge which variables are influential in terms of whether or not a customer will churn.

In order to efficiently collect the information gain for the variables used, I used a software called Weka (<http://www.cs.waikato.ac.nz/ml/weka/>). Weka is a software that aids in many data mining and machine learning tasks.

The variables with the highest level of information gain include Contract, Tenure, Online Security, and Tech Support.

This information is useful because it suggests that specific services may be more significant in predicting churn than others. Originally, I assumed the number of total services each customer has would be more significant that individual services. I will continue to analyze this trend after testing different models.

A baseline comparison for our model will be 73%. This figure represents the most common class for the target variable, churn. Of the 7,032 observations, 5,163 (73%) did not churn. If we were to classify all observations as not churn, we would expect to see an accuracy rate of 73%.

The first logistic regression model ran included all original variables from the raw dataset. The full summary of this model can be seen in the [appendix.](#Model1) From this model we see an overall accuracy rate of 81%. The specificity is 10% and the sensitivity is 56%. The positive predicted rate is 85% and the negative predicted rate is 66%.

From this model, we can see many variables are insignificant in predicting churn, which aligns with our EDA findings. We ran two additional logistic regression models removing variables that were not significant. These models both produced worse results than the original for the training data. The full calls for these can be found in [summary 2](#Model2) and [summary 3](#Model3) sections of the appendix.

Now that we have ran these three models with the training data, the true test will be to see how they perform on the new data (test set).

The first model created had an accuracy rate of 80.6% with the test data, the second model had an accuracy rating of 80.9% and the third had a rating of 80.3%. The full calls for these models can be found in the following sections of the appendix: [Test Model Summary 1](#TestModel1), [Test Model Summary 2](#TestModel2), [Test Model Summary 3.](#TestModel3)

Due to the consistency of the results for each model on the new data, we will choose the second logistic regression model due to its simplicity in comparison to the first. The final metrics for the second model on the test data are:

* Accuracy rate of 81% (11% improvement over baseline)
* Specificity rate of 10%
* Sensitivity rate of 55%
* Positive predicted value rate of 67%
* Negative predicted value rate of 85%

For further research, there are several things I would consider. First would be to test decision trees and random forests. Another would be to add a ranking based on each observations probability of churning and their value to Telco (which would be formulated using tenure, monthly charges, and total charges). This ranking could be used for numerous initiatives. One last area for further research would be to optimize the threshold used based on Telco’s business problem. The threshold used to define the probability of churning was .5. If Telco were to lower the threshold they would most likely see more customers classified as churners. This would be useful if they were running an inexpensive advertising campaign. If Telco wanted to use this information to run a high cost promotional campaign, they may want to raise the threshold to only classify those with the highest probabilities of churning.

**Appendix**

|  |  |  |
| --- | --- | --- |
| Variable Name | Example Data | Source |
| customerID | 7590-VHVEG | Telco Data |
| Gender | Female | Telco Data |
| SeniorCitizen | 0 (no) 1 (yes) | Telco Data |
| Partner | Yes, No | Telco Data |
| Dependents | Yes, No | Telco Data |
| Tenure | 12 | Telco Data |
| PhoneService | Yes, No | Telco Data |
| MultipleLines | Yes, No | Telco Data |
| InternetService | Yes, No | Telco Data |
| OnlineSecurity | Yes, No | Telco Data |
| OnlineBackup | Yes, No | Telco Data |
| DeviceProtection | Yes, No | Telco Data |
| TechSupport | Yes, No | Telco Data |
| StreamingTV | Yes, No | Telco Data |
| Streaming Movies | Yes, No | Telco Data |
| Contract | Month-to-month | Telco Data |
| PaperlessBilling | Yes, No | Telco Data |
| PaymentMethod | Mailed Check | Telco Data |
| MonthlyCharges | 49.95 | Telco Data |
| TotalCharges | 1889.50 | Telco Data |
| Churn | Yes, No | Telco Data |
| Churn\_yes | 1 (yes) 0 (no) | Manually Created |
| Phoneservice\_yes | 1 (yes) 0 (no) | Manually Created |
| Multiplelines\_yes | 1 (yes) 0 (no) | Manually Created |
| Internetservice\_yes | 1 (yes) 0 (no) | Manually Created |
| Onlinesecurity\_yes | 1 (yes) 0 (no) | Manually Created |
| Onlinebackup\_yes | 1 (yes) 0 (no) | Manually Created |
| Deviceprotection\_yes | 1 (yes) 0 (no) | Manually Created |
| Techsupport\_yes | 1 (yes) 0 (no) | Manually Created |
| Streamingtv\_yes | 1 (yes) 0 (no) | Manually Created |
| Streamingmovies\_yes | 1 (yes) 0 (no) | Manually Created |
| Contract\_bin | 1, 2, 3 | Manually Created |
| Sum\_services | 1, 2, 3, 4, 5, 6, 7 | Manually Created |
| Sum\_core | 1, 2 | Manually Created |
| Month\_bin | 1, 2, 3, 4, 5, 6 | Manually Created |
| Tenure\_bin | 1, 2, 3, 4, 5, 6, 7, 8 | Manually Created |

Train Model Summary 1:

Call:

glm(formula = Churn ~ gender + SeniorCitizen + Partner + Dependents +

tenure + PhoneService + MultipleLines + InternetService +

OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport +

StreamingTV + StreamingMovies + Contract + PaperlessBilling +

PaymentMethod + MonthlyCharges + TotalCharges, family = binomial,

data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.8444 -0.6795 -0.2688 0.7558 3.2411

Coefficients: (7 not defined because of singularities)

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.340e+00 1.010e+00 1.326 0.184711

genderMale 1.134e-02 8.053e-02 0.141 0.888054

SeniorCitizen1 1.570e-01 1.057e-01 1.485 0.137583

PartnerYes 1.032e-02 9.718e-02 0.106 0.915412

DependentsYes -1.909e-01 1.122e-01 -1.702 0.088844 .

tenure -5.803e-02 7.741e-03 -7.496 6.59e-14 \*\*\*

PhoneServiceYes 2.892e-01 8.036e-01 0.360 0.718943

MultipleLinesNo phone service NA NA NA NA

MultipleLinesYes 4.427e-01 2.210e-01 2.003 0.045158 \*

InternetServiceFiber optic 1.773e+00 9.925e-01 1.786 0.074038 .

InternetServiceNo -2.006e+00 1.003e+00 -2.000 0.045527 \*

OnlineSecurityNo internet service NA NA NA NA

OnlineSecurityYes -2.007e-01 2.234e-01 -0.898 0.368962

OnlineBackupNo internet service NA NA NA NA

OnlineBackupYes -5.819e-03 2.186e-01 -0.027 0.978763

DeviceProtectionNo internet service NA NA NA NA

DeviceProtectionYes 2.126e-01 2.181e-01 0.975 0.329804

TechSupportNo internet service NA NA NA NA

TechSupportYes -1.997e-01 2.242e-01 -0.891 0.373030

StreamingTVNo internet service NA NA NA NA

StreamingTVYes 4.744e-01 4.043e-01 1.173 0.240675

StreamingMoviesNo internet service NA NA NA NA

StreamingMoviesYes 6.231e-01 4.063e-01 1.534 0.125104

ContractOne year -8.745e-01 1.366e-01 -6.401 1.54e-10 \*\*\*

ContractTwo year -1.588e+00 2.292e-01 -6.929 4.24e-12 \*\*\*

PaperlessBillingYes 3.415e-01 9.246e-02 3.694 0.000221 \*\*\*

PaymentMethodCredit card (automatic) 1.146e-02 1.402e-01 0.082 0.934810

PaymentMethodElectronic check 2.840e-01 1.171e-01 2.424 0.015344 \*

PaymentMethodMailed check -8.312e-02 1.423e-01 -0.584 0.559160

MonthlyCharges -4.398e-02 3.940e-02 -1.116 0.264385

TotalCharges 3.315e-04 8.844e-05 3.748 0.000178 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5293.6 on 4570 degrees of freedom

Residual deviance: 3769.7 on 4547 degrees of freedom

AIC: 3817.7

Number of Fisher Scoring iterations: 6

Test Model Summary 1:

Call:

glm(formula = Churn ~ gender + SeniorCitizen + Partner + Dependents +

tenure + PhoneService + MultipleLines + InternetService +

OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport +

StreamingTV + StreamingMovies + Contract + PaperlessBilling +

PaymentMethod + MonthlyCharges + TotalCharges, family = binomial,

data = test)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.0443 -0.6660 -0.3021 0.6759 3.3938

Coefficients: (7 not defined because of singularities)

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.7636458 1.4056527 0.543 0.586945

genderMale -0.0687689 0.1106582 -0.621 0.534302

SeniorCitizen1 0.3128356 0.1427690 2.191 0.028437 \*

PartnerYes -0.0114484 0.1313760 -0.087 0.930558

DependentsYes -0.0866615 0.1511721 -0.573 0.566466

tenure -0.0640936 0.0106381 -6.025 1.69e-09 \*\*\*

PhoneServiceYes -0.0935319 1.1207278 -0.083 0.933489

MultipleLinesNo phone service NA NA NA NA

MultipleLinesYes 0.4477429 0.3020994 1.482 0.138312

InternetServiceFiber optic 1.7159281 1.3686616 1.254 0.209941

InternetServiceNo -1.3064752 1.3846878 -0.944 0.345417

OnlineSecurityNo internet service NA NA NA NA

OnlineSecurityYes -0.2120974 0.3034636 -0.699 0.484601

OnlineBackupNo internet service NA NA NA NA

OnlineBackupYes 0.0876184 0.2987397 0.293 0.769298

DeviceProtectionNo internet service NA NA NA NA

DeviceProtectionYes 0.0034288 0.3055929 0.011 0.991048

TechSupportNo internet service NA NA NA NA

TechSupportYes -0.1584071 0.3116235 -0.508 0.611223

StreamingTVNo internet service NA NA NA NA

StreamingTVYes 0.7930683 0.5633524 1.408 0.159200

StreamingMoviesNo internet service NA NA NA NA

StreamingMoviesYes 0.5228875 0.5599238 0.934 0.350379

ContractOne year -0.2802341 0.1794957 -1.561 0.118469

ContractTwo year -0.9709189 0.2809843 -3.455 0.000549 \*\*\*

PaperlessBillingYes 0.3451185 0.1275387 2.706 0.006810 \*\*

PaymentMethodCredit card (automatic) -0.2894026 0.1996104 -1.450 0.147104

PaymentMethodElectronic check 0.3341456 0.1614008 2.070 0.038426 \*

PaymentMethodMailed check -0.0183452 0.1963869 -0.093 0.925575

MonthlyCharges -0.0319056 0.0546630 -0.584 0.559436

TotalCharges 0.0003068 0.0001187 2.585 0.009727 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2849.7 on 2460 degrees of freedom

Residual deviance: 2029.8 on 2437 degrees of freedom

AIC: 2077.8

Number of Fisher Scoring iterations: 6

Train Model Summary 2:

Call:

glm(formula = Churn ~ tenure + MultipleLines + InternetService +

Contract + PaperlessBilling + PaymentMethod + TotalCharges,

family = binomial, data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7101 -0.6901 -0.2849 0.7860 3.2487

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -6.513e-01 1.494e-01 -4.358 1.31e-05 \*\*\*

tenure -5.719e-02 7.303e-03 -7.831 4.83e-15 \*\*\*

MultipleLinesNo phone service 6.527e-01 1.561e-01 4.180 2.91e-05 \*\*\*

MultipleLinesYes 2.570e-01 9.879e-02 2.602 0.009273 \*\*

InternetServiceFiber optic 8.738e-01 1.170e-01 7.466 8.28e-14 \*\*\*

InternetServiceNo -6.351e-01 1.628e-01 -3.901 9.59e-05 \*\*\*

ContractOne year -9.938e-01 1.333e-01 -7.457 8.84e-14 \*\*\*

ContractTwo year -1.813e+00 2.254e-01 -8.043 8.78e-16 \*\*\*

PaperlessBillingYes 3.850e-01 9.088e-02 4.237 2.27e-05 \*\*\*

PaymentMethodCredit card (automatic) 1.233e-02 1.388e-01 0.089 0.929254

PaymentMethodElectronic check 3.672e-01 1.155e-01 3.180 0.001475 \*\*

PaymentMethodMailed check -9.356e-02 1.403e-01 -0.667 0.504975

TotalCharges 2.831e-04 7.861e-05 3.601 0.000318 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5293.6 on 4570 degrees of freedom

Residual deviance: 3824.6 on 4558 degrees of freedom

AIC: 3850.6

Number of Fisher Scoring iterations: 6

Test Model Summary 2:

Call:

glm(formula = Churn ~ tenure + MultipleLines + InternetService +

Contract + PaperlessBilling + PaymentMethod + TotalCharges,

family = binomial, data = test)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7636 -0.6896 -0.3086 0.7400 3.5254

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.9474376 0.2105474 -4.500 6.80e-06 \*\*\*

tenure -0.0689096 0.0101246 -6.806 1.00e-11 \*\*\*

MultipleLinesNo phone service 0.9246229 0.2332814 3.964 7.38e-05 \*\*\*

MultipleLinesYes 0.3608107 0.1308393 2.758 0.005822 \*\*

InternetServiceFiber optic 1.1149781 0.1626860 6.854 7.20e-12 \*\*\*

InternetServiceNo -0.3387835 0.2212130 -1.531 0.125651

ContractOne year -0.3045575 0.1737685 -1.753 0.079660 .

ContractTwo year -1.1780708 0.2746578 -4.289 1.79e-05 \*\*\*

PaperlessBillingYes 0.4230626 0.1248767 3.388 0.000704 \*\*\*

PaymentMethodCredit card (automatic) -0.3072248 0.1969294 -1.560 0.118742

PaymentMethodElectronic check 0.4095223 0.1586899 2.581 0.009862 \*\*

PaymentMethodMailed check -0.0794386 0.1934455 -0.411 0.681328

TotalCharges 0.0003611 0.0001058 3.412 0.000645 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2849.7 on 2460 degrees of freedom

Residual deviance: 2068.9 on 2448 degrees of freedom

AIC: 2094.9

Number of Fisher Scoring iterations: 6

Train Model Summary 3:

Call:

glm(formula = Churn ~ tenure + MultipleLines + InternetService +

Contract + PaperlessBilling + TotalCharges, family = binomial,

data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.6531 -0.6905 -0.2881 0.8007 3.2675

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -5.662e-01 1.107e-01 -5.112 3.19e-07 \*\*\*

tenure -5.688e-02 7.250e-03 -7.846 4.30e-15 \*\*\*

MultipleLinesNo phone service 6.849e-01 1.552e-01 4.412 1.02e-05 \*\*\*

MultipleLinesYes 2.765e-01 9.821e-02 2.815 0.004876 \*\*

InternetServiceFiber optic 9.782e-01 1.141e-01 8.574 < 2e-16 \*\*\*

InternetServiceNo -7.357e-01 1.599e-01 -4.601 4.20e-06 \*\*\*

ContractOne year -1.046e+00 1.325e-01 -7.896 2.89e-15 \*\*\*

ContractTwo year -1.903e+00 2.241e-01 -8.492 < 2e-16 \*\*\*

PaperlessBillingYes 4.172e-01 9.021e-02 4.624 3.76e-06 \*\*\*

TotalCharges 2.740e-04 7.857e-05 3.487 0.000488 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5293.6 on 4570 degrees of freedom

Residual deviance: 3847.0 on 4561 degrees of freedom

AIC: 3867

Number of Fisher Scoring iterations: 6

Test Model Summary 3:

Call:

glm(formula = Churn ~ tenure + MultipleLines + InternetService +

Contract + PaperlessBilling, family = binomial, data = test)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7434 -0.6894 -0.3394 0.7406 3.1814

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.998029 0.152368 -6.550 5.75e-11 \*\*\*

tenure -0.039825 0.003548 -11.226 < 2e-16 \*\*\*

MultipleLinesNo phone service 0.844065 0.221246 3.815 0.000136 \*\*\*

MultipleLinesYes 0.448611 0.127729 3.512 0.000444 \*\*\*

InternetServiceFiber optic 1.408847 0.148302 9.500 < 2e-16 \*\*\*

InternetServiceNo -0.569456 0.211756 -2.689 0.007162 \*\*

ContractOne year -0.301932 0.169810 -1.778 0.075394 .

ContractTwo year -1.137085 0.266633 -4.265 2.00e-05 \*\*\*

PaperlessBillingYes 0.453332 0.123103 3.683 0.000231 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2849.7 on 2460 degrees of freedom

Residual deviance: 2106.1 on 2452 degrees of freedom

AIC: 2124.1

Number of Fisher Scoring iterations: 6