Predictive Churn Rates

Jeff Nguyen

17DEC2018

SDS358

The University of Texas at Austin

Statistics + Data Sciences

Introduction

Objectives:

- Logistic regression will be used to determine: “How do the factors of Total Charges, Streaming Movies, Tech Support, Phone Service, Tenure, Gender, Multiple Lines, and Internet Service determine whether or not a customer churns?”

- Due to Charter Communication policies company data cannot be used for this project, however analogous data is fortuitously available on Kaggle. By using logistic regression, the relationship between variables can be described allowing for greater insight as to how individual explanatory variables relate to Churn. This insight can possibly allow for a Line of Business to create strategies to retain customers that intend to cancel services and improve existing services to ensure happy customers.

Hypotheses:

- What is the best predictor of Churn: Total Charges, Streaming Movies, Tech Support, Phone Service, Tenure, Gender, Multiple Lines, or Internet Service?

Methods

Sample:

- Each Subject is represented by a Customer ID
- The unmodified data set has 7043 unique observations, a sample of n = 385 was taken from that set where 8 outliers were removed. This was done prevent all coefficients being significant due to CLT.
- Variable Units:
    - Total Charges: Quantitative, Continuous, Dollars
    - Streaming Movies, Tech Support, Phone Service, Tenure, Multiple Lines, Internet Service, Churn: Categorical, Ordinal, Dichotomous

Analysis Method: Logistic Regression

- glm() and lrm() functions were run on RStudio 1.1.463

Descriptives

Response Variable	Yes	No
Churn	133	267

Categorical Explanatory Variables:	Yes	No
Streaming Movies	205	195
Tech Support	125	275
Phone Service	357	43
Gender	214(M)	186(F)
Multiple Lines	200	157
Internet Service	224 (Fiber Optic)	176 (DSL)

Quantitative Explanatory Variables	Center (Median)	Spread (IQR)
Tenure (Months)	33	47
Total Charges ($)	2289.70	3955.70

Results

Results table:

Variable	Coefficient	Test Statistic	P-value
StreamingMoviesYes	0.5150	1.64	0.1011
TechSupportYes	-1.2255	-3.14	0.0017
PhoneServiceYes	-0.4161	-0.75	0.4503
tenure	-0.0555	-7.42	<0.0001
genderMale	0.1769	0.63	0.5261
InternetServiceFiber optic	2.2821	6.08	<0.0001

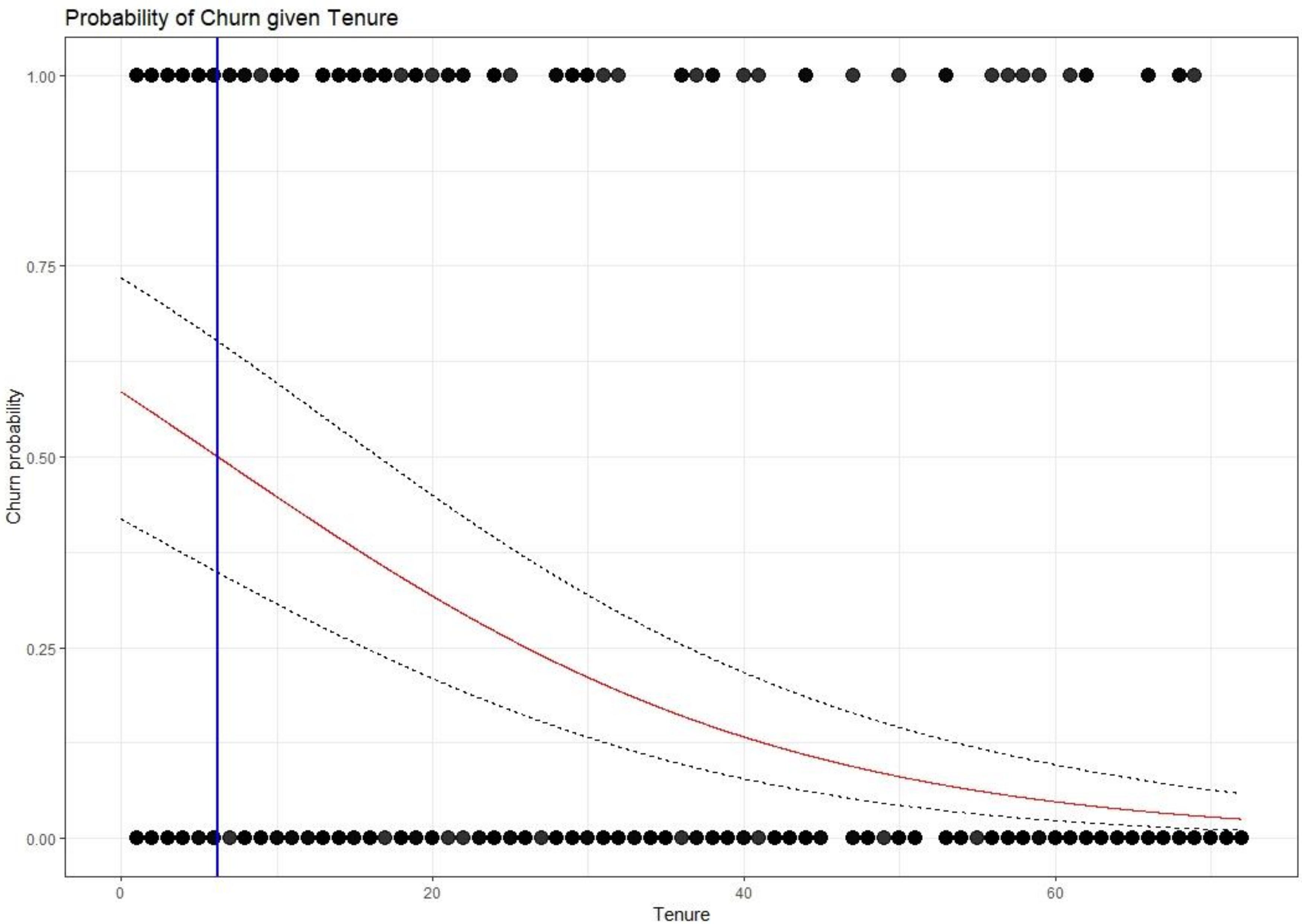
$LR\ \chi^2_6 = 165.51$

$Nagelkerke\ R^2 = 0.481$

```
> exp(g_percent_churn$coef)
        (Intercept)       StreamingMoviesYes           TechSupportYes          PhoneServiceYes                   tenure               genderMale InternetServiceFiber optic
          0.7249666                1.6736172                0.2936108                0.6596384                0.9460499                1.1935096                  9.7976780
>
```

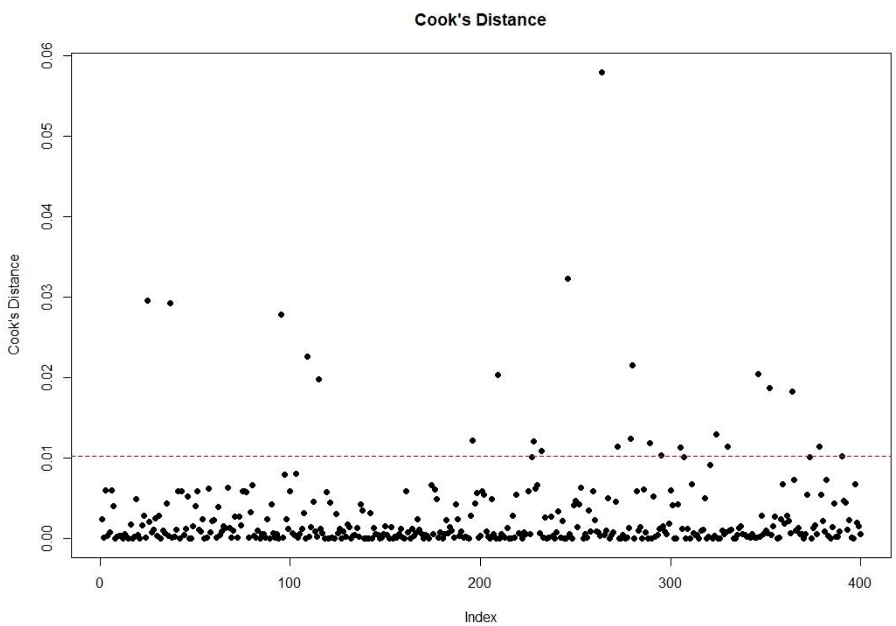
Final Model

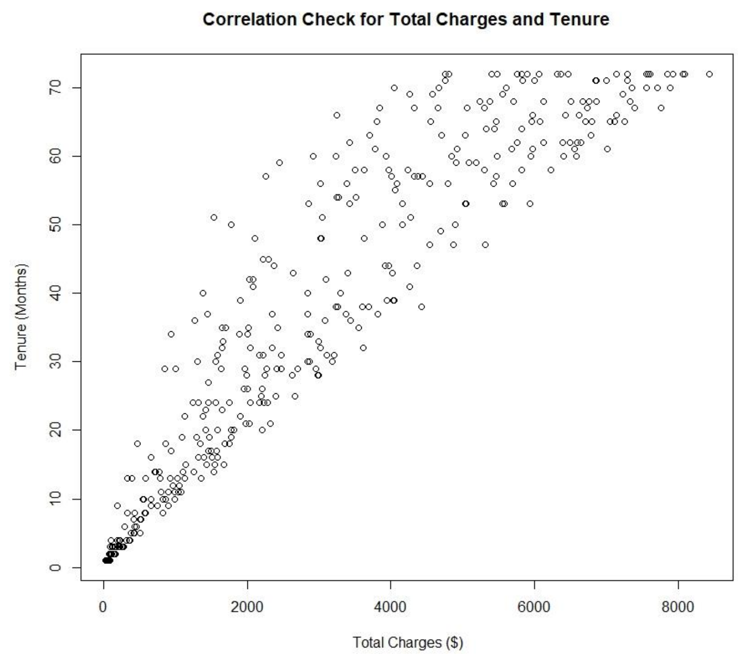
$$\widehat{Churn} = -0.3216 + 0.5150(\boldsymbol{StreamingMovieYes}) - 1.2255(\boldsymbol{TechSupportYes}) - 0.4161(\boldsymbol{PhoneServiceYes}) - 0.0555(\boldsymbol{tenure}) + 0.1769(\boldsymbol{genderMale}) + 2.2821\,(\boldsymbol{InternetServiceFiber\ optic})$$


Probability of Churn given Tenure
1.00
0.75
0.50
0.25
0.00
Churn probability
0
20
40
60
Tenure

Assumptions

Assumptions:






```
> vif(percent_churn)
         TotalCharges         StreamingMoviesYes             TechSupportYes            PhoneServiceYes
            22.646862                   1.442782                   1.059578                   1.608372
               tenure                 genderMale InternetServiceFiber optic
            19.181139                   1.034970                   1.733263
>
```

Discussion

Interpretation:

The overall model was significant with LR X^2(6) = 165.51, P < 0.000.1. When investigating the coefficients: TechSupport=Yes (beta = -1.2255, z = -3.14, p = 0.0017); Tenure (beta = -0.0555, z = -7.42, p < 0.0001); and InternetService=Fiber optic (beta = 2.2821, z = 6.08, p < 0.0001) had a significant impact on Churn. StreamingMovies=Yes (beta = 0.5150, z = 1.64, p < 0.1011), PhoneService=Yes (beta = -0.4161, z = -0.75, p < 0.4503), and gener=Male (beta = 0.1769, z = 0.63, p = 0.5261) were not significant. When looking at odds ratios for the significant coefficients customers with TechSupport=Yes had an OR = 0.293, meaning customers were less likely to churn with Tech Support; customers that had long tenure with the telcom company had an OR = 0.946, meaning customers were less likely to churn when they stay with the company longer; and Customers that had Fiber Optic internet Service had an OR = 9.798, these customers were 9.798 times more likely to churn.

When the probability of P(Churn) = 0.5 the value for tenure was 6.21 month. The model suggests that 50% of customers leave the given telcom company after 6.21 months (with a possible 95% CI range of 0 months to 16.21 months.

Limitations:

This model does not consider seasonality. Certain trends in Churn may shift during certain times of the year.

Implications:

This model may be able to predict Churn over a short time frame, however for data that spans over the course of several years, a time-series model will more accurately model Churn while considering seasonality.

References:

Js. (2018, August 10). Telecom Churn Dataset (IBM Watson Analytics). Retrieved from https://www.kaggle.com/zagarsuren/telecom-churn-dataset-ibm-watson-analytics