Data Analysis Project

Mirnes Salkic

12/12/2017

Using customer behavior to improve customer retention

1. Introduction

The data set, found on the IBM website (link at the bottom of the page), comes from a telecommunications company. I suspect that the name of the company and the year, and time the data set relates to has been kept anonymous due to privacy concerns. The data set is composed of 7043 rows, representing customers, and 21 columns, depicting attributes.

I would like to explore the following:
1. whether the churn rate between male and female differs
2. whether `Churn` is dependent on `PaymentType`
3. which logistic regression model would best fit the data

2. Exploratory Data Analysis

First, we take a look at the data to get an idea of the type of variables we have.

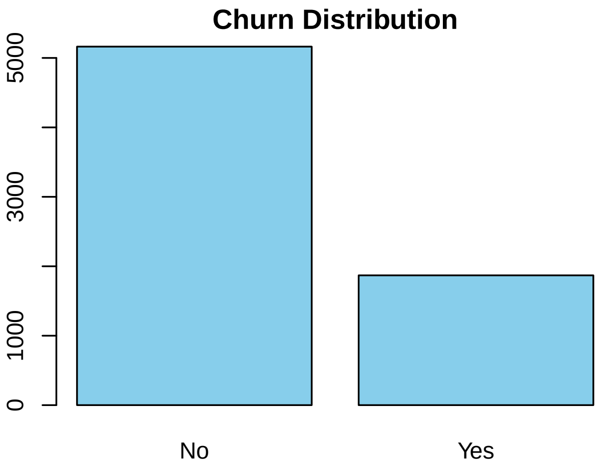
Factor	customer ID	7043 unique values
	gender	"male", "female"
	Contract	"Month-to-month" "One year" "Two year"
	Partner, Dependents, PhoneService, PaperlessBilling, Churn	"Yes", "No"
	OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies	"Yes", "No", "No internet service"
	Multiplelines	"Yes", "No", "No phone service"
	InternetService:	"DSL", "Fiber optic", "No"
	PaymentMethod	"Electronic check", "Mailed check", "Bank transfer", "Credit card"
int	SeniorCitizen	0, 1
	tenure	
num	MonthlyCharges, TotalCharges	

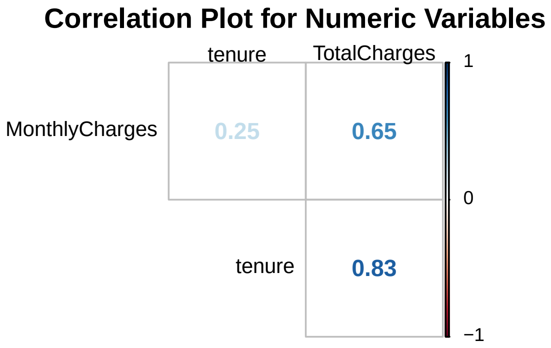
Observations:
1. customerID - is a factor variable with 7043 unique levels and as such will not contribute much to our analysis. Therefore, it will be dropped.
2. SeniorCitizen - this variable will be converted into a factor variable to better reflect its meaning. Its values will be mapped to "Yes" and "No".
3. OnlineSecurity/OnlineBackup/DeviceProtection/TechSupport/StreamingTV/StreamingMovies - "No internet service" will be converted to "No".
4. MultipleLines - the "No phone service" level will be converted to "No".

2.1. Missing Values

Since there are only 11 missing values all part of the variable TotalCharges, I decided to delete the the rows that contain them. After deletion, we end up with 7032 rows and 21 columns.

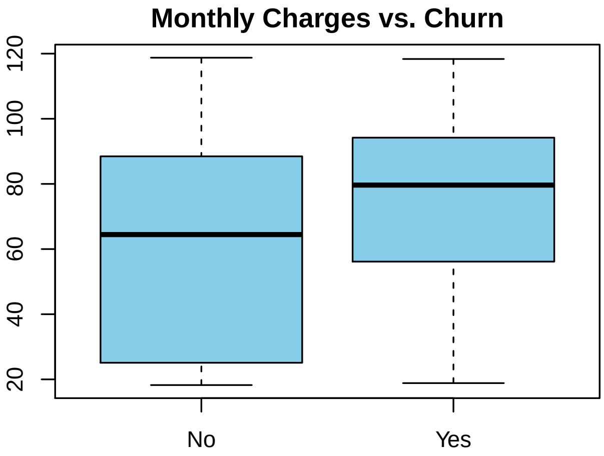
2.2. Churn distribution/Correlation between numerical attributes

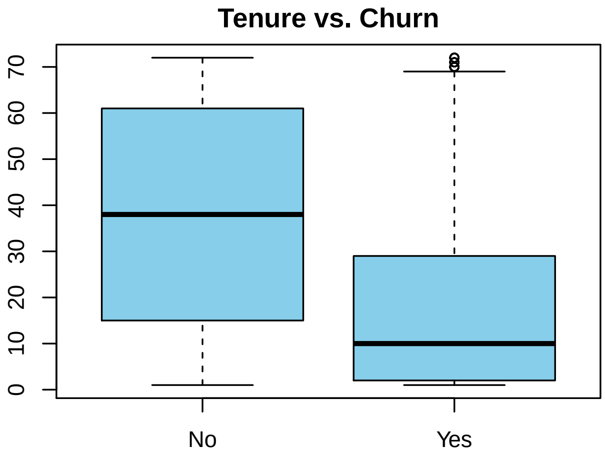




We can see that the Churn varibale is uneavenly distributed. However, this should not affect our model. TotalCharges is correlated with both MonthlyCharges and tenure. As TotalCharges is strongly correlated with both MonthlyCharges and tenure it would be best to drop TotalCharges.

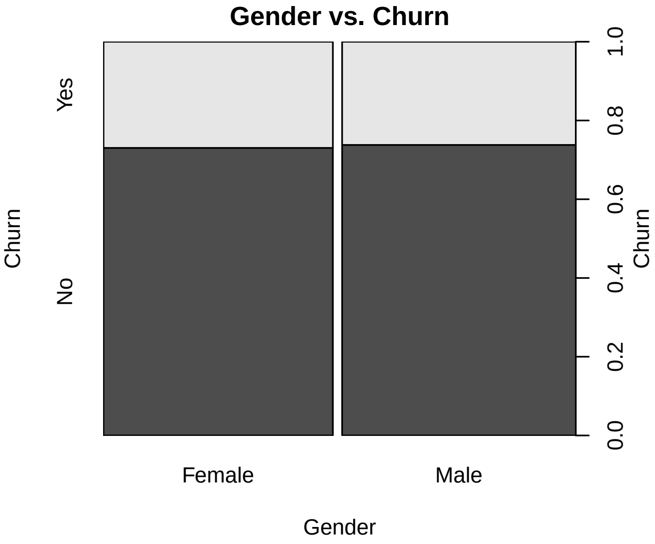
2.3. MonthlyCharges vs. Churn and Tenure vs. Churn

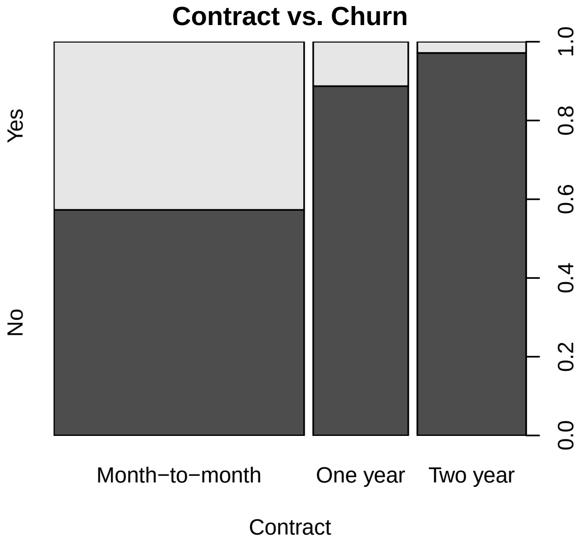




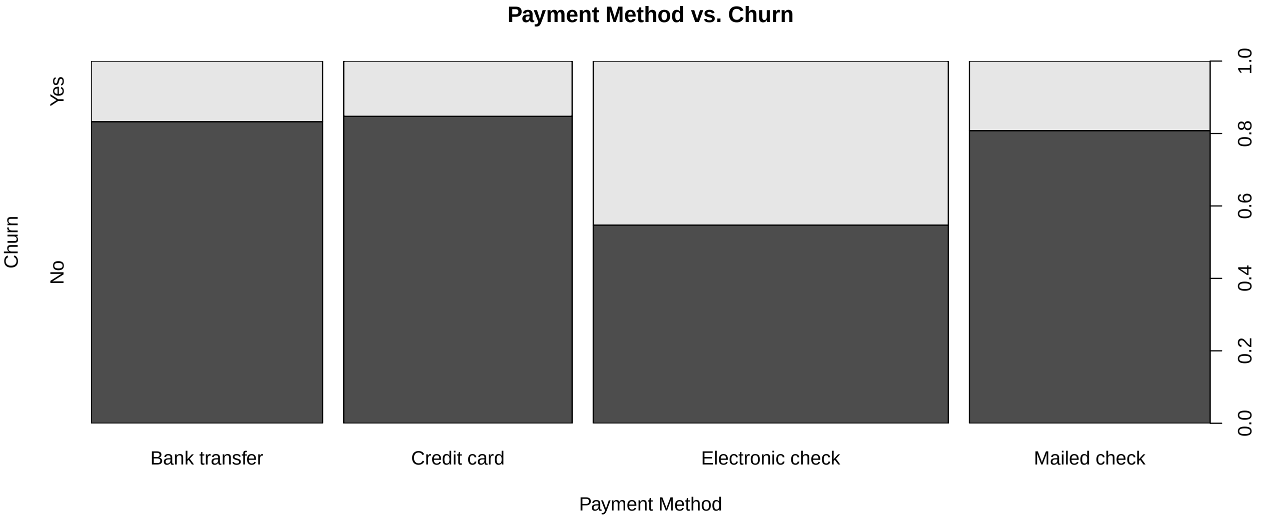
Clients who churn tend to be the ones that pay around 80 dollars per month for services while the ones that do not churn tend to pay on average slightly more than 60 dollars per month. From the “Tenure vs. Churn” boxplots we can see that customers who have a higher tenure tend to be churn less than those with s smaller tenure.

2.4. Gender vs. Churn and Contract vs. Churn





By observing the graph we can see that the percentage of male and female who churn is about the same. I will further use statistical methods to test this observation (section 3). The “Contract vs. Churn” graph reveals that clients who have a one or two year contract are less likely to churn than clients on a month-to-month contract. Therefore, the variable `Contract` seems like a good predictor of `Churn`.



3. Is there a difference between male and female churn rate?

To see whether there is a statistically significant (α=0.05) difference in churn rate between male and female we can set up a hypothesis test.

$$H_0 : p_{\text{female}} - p_{\text{male}} = 0 \qquad H_A : p_{\text{female}} - p_{\text{male}} \neq 0$$

The null hypothesis states that there is not a difference in the proportion of males and females churning while the alternative suggests that there is.

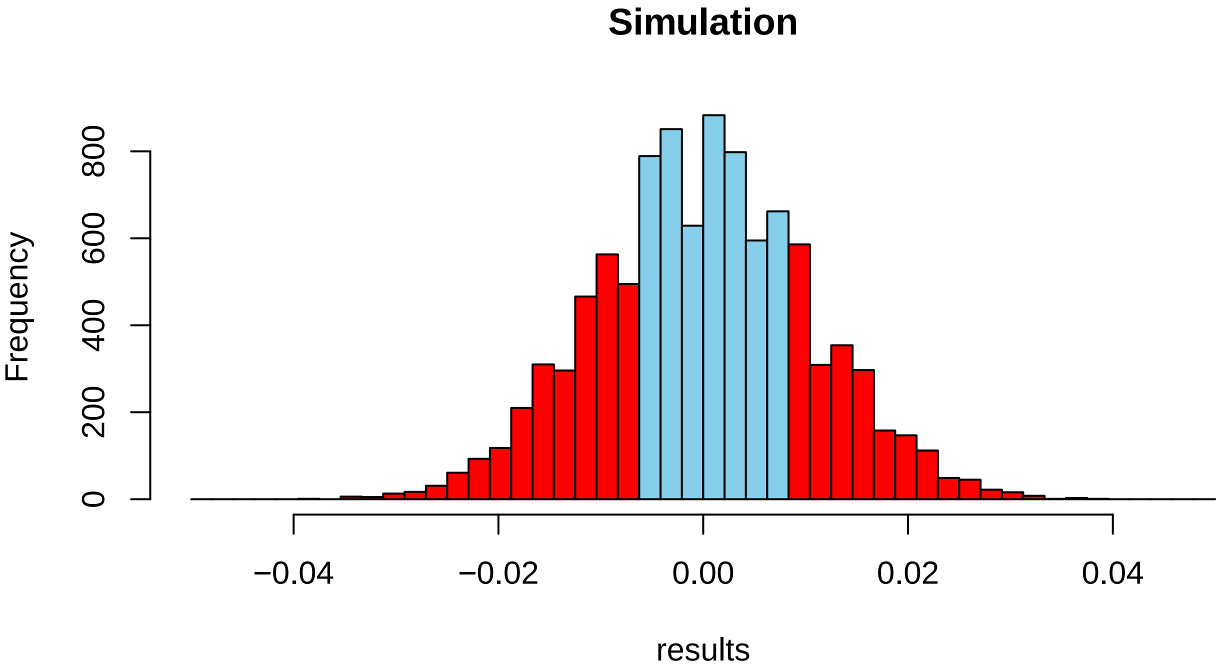
We can easily calculate the point estimate from the data (see Table 2):

$$\hat{p}_{\text{female}} - \hat{p}_{\text{male}} = \frac{939}{3483} - \frac{930}{3549} = 0.0075$$

Table 2: Gender vs. Churn

	No	Yes	Total
Female	2544	939	3483
Male	2619	930	3549
Total	5163	1869	7032

To prepare for the simualtion, I will create a vector that contains 1869 instances of 1's (those that churned) and 5163 of 0's (those that did not churn). In one trial, I would sample 3483 instances (representing female) and calculate the proportion of those that churned. The remaining instances will represent the male for whom the proportion of those who churned will also be calculated. Finaly, the difference between the female and male churn rate will be stored in a variable called `results`. This process would be repeated 10000 times. The distribution of the simulation is shown in a histogram.



Our point estimate was 0.0075, and the p-value for the two-tailed test is 0.4746 (red area of the histogram). We **fail to reject the null hypothesis.** In other words, there is not enough statistically significant evidence to conclude that there is a difference in the female and male churn rate. Given this result we can say that churn is not associated with gender, and therefore it might not be useful to include gender in the logistic regression analysis.

4. Are Churn and PaymentMedhod dependent or independent?

To determine whether `Churn` and `PaymentMethod` are dependent or independent we can set up a hypothesis:

$$H_0 : \text{There is no relationship between payment method and churn (independent)}$$

$$H_A : \text{There is a relationship between payment method and churn (dependent)}$$

This problem can be solved using the Chi-square test of independence. Below, we can see two tables: the observed counts of each payment type and churn and the expected counts of each payment type and churn.

From the table we can calculate the degrees of freedom:

$$df = (2-1) * (4-1) = 3$$

Table 3: Observed

	No	Yes	Total
Bank transfer	1284	258	1542
Credit Card	1289	232	1521
Electronic check	1294	1071	2365
Mailed check	1296	308	1604
Total	5163	1869	7032

Table 4: Expected

	No	Yes	Total
Bank transfer	1132.16	409.84	1542
Credit Card	1116.74	404.26	1521
Electronic check	1736.42	628.58	2365
Mailed check	1177.68	426.32	1604
Total	5163.00	1869.00	7032

Checking the conditions
1) I have to assume that the sample is random (given that there is no information about it)
2) The expected counts are above 5,
3) df>1

The conditions are satisfied. The expected counts are calculated by multiplying the sum of a row by the sum of the column and dividing it by the grand total. Below, I show how the expected counts for the first row and first column is calculated:

$$\begin{aligned} E_{\text{row 1, col 1}} &= \frac{1542 * 5163}{7032} \\ &= 1132.16 \end{aligned}$$

```r
t_payment_churn <- table(df2$PaymentMethod, df2$Churn)
payment_churn <- chisq.test(t_payment_churn)
```

The test statistic X-squared = 645.43 with the associated df=3 gives a p-value of 1.42e-139. The evidence is sufficiently strong to **reject the null hypothesis**. Thus the data provide strong evidence to conclude that churn is associated (dependent) with the type of payment the customer chose to pay for the services provided by the telecommunications company. Therefore, the `PaymentMethod` variable could be a good predictor for our logistic regression model.

5. Modelling

I will build logistic regression models based on 17 predictor variables as `gender` and `TotalCharges` will not be included based on the earlier discussion. I will use stepwise forward and stepwise backward elimination based on AIC which is a metric that penalizes a model by removing variables that do not have significant predictive power.

The model that minimized the AIC score consists of 14 variables:

```r
model = glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +
    InternetService + OnlineSecurity + DeviceProtection + TechSupport +
    StreamingTV + StreamingMovies + Contract + PaperlessBilling +
    PaymentMethod + MonthlyCharges, family = binomial, data = df3)
summary(model)
```

```
##
## Call:
## glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +
##     InternetService + OnlineSecurity + DeviceProtection + TechSupport +
##     StreamingTV + StreamingMovies + Contract + PaperlessBilling +
##     PaymentMethod + MonthlyCharges, family = binomial, data = df3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9842  -0.6702  -0.2957   0.6917   3.1412
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      0.399787   0.256870   1.556 0.119619
## SeniorCitizenYes                 0.217848   0.084369   2.582 0.009820 **
## DependentsYes                   -0.164274   0.081305  -2.020 0.043337 *
## tenure                          -0.034327   0.002256 -15.215  < 2e-16 ***
## MultipleLinesYes                 0.423878   0.088806   4.773 1.81e-06 ***
## InternetServiceFiber optic       1.515092   0.199058   7.611 2.71e-14 ***
## InternetServiceNo               -1.417761   0.173663  -8.164 3.24e-16 ***
## OnlineSecurityYes               -0.238108   0.090440  -2.633 0.008469 **
## DeviceProtectionYes              0.123149   0.083538   1.474 0.140437
## TechSupportYes                  -0.209477   0.091542  -2.288 0.022119 *
## StreamingTVYes                   0.513732   0.097428   5.273 1.34e-07 ***
## StreamingMoviesYes               0.525667   0.096296   5.459 4.79e-08 ***
## ContractOne year                -0.662616   0.106698  -6.210 5.29e-10 ***
## ContractTwo year                -1.334524   0.174501  -7.648 2.05e-14 ***
## PaperlessBillingYes              0.336874   0.074237   4.538 5.68e-06 ***
## PaymentMethodCredit card        -0.087619   0.114012  -0.769 0.442186
## PaymentMethodElectronic check    0.313273   0.094562   3.313 0.000923 ***
## PaymentMethodMailed check       -0.005565   0.113613  -0.049 0.960932
## MonthlyCharges                  -0.024513   0.005777  -4.243 2.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8143.4  on 7031  degrees of freedom
## Residual deviance: 5849.3  on 7013  degrees of freedom
## AIC: 5887.3
##
## Number of Fisher Scoring iterations: 6
```

Some observations

- keeping all other variables constant, the odds ratio of churning for seniors vs. nonseniors is (exp 0.217848) = 1.24. This means that **senior citizens are more likely to churn than nonsenior citizens**. However, we do not know the probabilities so we cannot infer how much likely.
- keeping all other variables constant, the odds ratio of churning for those who have dependents vs. those who do not is (exp -0.164274) = 0.85. This means that **those who have dependents are less likely to churn than those who do not have**.

- keeping all other variables constant, the odds ratio of churning if the `tenure` increase by a week is exp(-0.024513)=0.96. The relationship between `MonthlyCharges` and `Churn` is negative (given the negative sign in from of the coefficient) which means that **increased tenure are associated with lower probability of churn**.
- keeping all other variables constant, the odds ratio of churning for those who have multiple lines vs. those who do not is (exp 0.42387) = 1.5. This means that **those who have multiple lines are more likely to churn than those who do not have**.
- keeping all other variables constant, the odds ratio of churning for those who have Fiber Optic internet service vs. those who have DSL is (exp 1.515092) = 4.6. This means that **those who have Fiber Optic internet are more likely to churn than those who have DSL**.
- keeping all other variables constant, the odds ratio of churning for those who have do not have internet service vs. those who have DSL is (exp -1.417761) = 0.24. This means that **those who have do not have internet are less likely to churn than those who have DSL**.
- keeping all other variables constant, the odds ratio of churning for those who have do not have internet service vs. those who have DSL is exp(-0.238108) = 0.79. This means that **those who have do not have internet are less likely to churn than those who have DSL**.
- keeping all other variables constant, the odds ratio of churning for those who received online support vs.those who did not is (exp -0.238108) = 0.81. This indicates that **those who receive online support are less likely to churn than those who do not**.
- keeping all other variables constant, the odds ratio of churning if the `MonthlyCharges` increase by a dollar is exp(-0.024513)=0.96. The relationship between `MonthlyCharges` and `Churn` is negative which means that **increased monthly charges are associated with lower probability of churn**.

Ideally the company would target all the customers who are very likely to churn in the future. This model could be used to make predictions and determine whether a customer is likely to churn.

Conclusion

We learned that there is no statistically significant difference in churning between male and female customers. Churn is associated (dependent) with the type of payment the customer chose to pay for the services provided by the telecommunications company. The model that best predicts whether a customer will churn or not is:

model = glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines + InternetService + OnlineSecurity + DeviceProtection + TechSupport + StreamingTV + StreamingMovies + Contract + PaperlessBilling + PaymentMethod + MonthlyCharges, family = binomial, data = df3)