# Telco Company Churn Prediction

## Abstract

Based on a dataset containing data about services contracted by customers from a Telco Company, as well the type of the contract, tenure as a customer, charges from the services contracted and historical data which states if the customer left the Telco company, changing it for another company (Churn), we will evaluate the hypothesis if it is possible to predict if a customer, based on these parameters will churn and what is the probability of it happens.

## Step 1 and 2 – Data Loading and Preparation

As a preparation for a classification problem (churn / not churn), the dataset was changed to treat the dummy variables.

* 1. The dummy variables that were Yes/No are below and just the first value (Yes) were kept.

['gender','Dependents', 'PhoneService','InternetService','DeviceProtection','TechSupport','CableService','PaperlessBilling','Churn']

The dummy variables that could assume more than 2 values are below and all possible values were kept in different columns/features.

['Contract','PaymentMethod']

The dataset columns after the dummy transformation became as below.

|  |  |
| --- | --- |
| SeniorCitizen | Predictor Feature |
| tenure | Predictor Feature |
| MonthlyCharges | Predictor Feature |
| TotalCharges | Predictor Feature |
| gender\_Male | Predictor Feature |
| Dependents\_Yes | Predictor Feature |
| PhoneService\_Yes | Predictor Feature |
| InternetService\_Yes | Predictor Feature |
| DeviceProtection\_Yes | Predictor Feature |
| TechSupport\_Yes | Predictor Feature |
| CableService\_Yes | Predictor Feature |
| PaperlessBilling\_Yes | Predictor Feature |
| Churn\_Yes | Predicted Variable |
| Contract\_Month-to-month | Predictor Feature |
| Contract\_One year | Predictor Feature |
| Contract\_Two year | Predictor Feature |
| PaymentMethod\_Bank transfer (automatic) | Predictor Feature |
| PaymentMethod\_Credit card (automatic) | Predictor Feature |
| PaymentMethod\_Electronic check | Predictor Feature |
| PaymentMethod\_Mailed check | Predictor Feature |

* 1. Also the variable TotalCharge had some missing values in some lines, so these lines were removed. After that the dataset, which originally had 7043 rows, became with 7032.

## Step 2 – Logistic Regression Analysis over the whole dataset

The Logistic regression analysis was carried out over the whole dataset (before any features reduction) in order to evaluate if the dataset was adequate (statistically significant) for this classification problem.

### Results: Generalized linear model

=======================================================================================

Model: GLM AIC: 5928.4743

Link Function: logit BIC: -56239.1260

Dependent Variable: y Log-Likelihood: -2946.2

Date: 2020-07-12 21:54 LL-Null: -4071.7

No. Observations: 7032 Deviance: 5892.5

Df Model: 17 Pearson chi2: 8.07e+03

Df Residuals: 7014 Scale: 1.0000

Method: IRLS

---------------------------------------------------------------------------------------

Coef. Std.Err. z P>|z| [0.025 0.975]

---------------------------------------------------------------------------------------

SeniorCitizen 0.2518 0.0832 3.0262 0.0025 0.0887 0.4150

tenure -0.0604 0.0062 -9.7673 0.0000 -0.0725 -0.0482

MonthlyCharges 0.0280 0.0035 7.9927 0.0000 0.0211 0.0348

TotalCharges 0.0003 0.0001 4.1193 0.0000 0.0002 0.0004

gender\_Male -0.0096 0.0643 -0.1489 0.8817 -0.1356 0.1165

Dependents\_Yes -0.1743 0.0808 -2.1566 0.0310 -0.3326 -0.0159

PhoneService\_Yes -1.0730 0.1617 -6.6355 0.0000 -1.3899 -0.7560

InternetService\_Yes -0.0619 0.1820 -0.3403 0.7337 -0.4187 0.2948

DeviceProtection\_Yes -0.1780 0.0803 -2.2150 0.0268 -0.3354 -0.0205

TechSupport\_Yes -0.5762 0.0843 -6.8337 0.0000 -0.7414 -0.4109

CableService\_Yes -0.1463 0.0931 -1.5725 0.1158 -0.3287 0.0361

PaperlessBilling\_Yes 0.3841 0.0738 5.2077 0.0000 0.2395 0.5286

Contract\_Month-to-month -0.0660 0.1151 -0.5732 0.5665 -0.2916 0.1596

Contract\_One year -0.8018 0.1345 -5.9637 0.0000 -1.0654 -0.5383

Contract\_Two year -1.5384 0.1750 -8.7917 0.0000 -1.8813 -1.1954

PaymentMethod\_Bank transfer (automatic) -0.6469 0.1128 -5.7378 0.0000 -0.8679 -0.4260

PaymentMethod\_Credit card (automatic) -0.7341 0.1130 -6.4975 0.0000 -0.9556 -0.5127

PaymentMethod\_Electronic check -0.2988 0.1005 -2.9725 0.0030 -0.4958 -0.1018

PaymentMethod\_Mailed check -0.7263 0.0953 -7.6198 0.0000 -0.9131 -0.5395

By Analyzing the dataset summary above it is possible to verify that some features have P-Value > 0.05, which means that these features are not statistically significant. In other words, these variables do not impact consistently the predicted outcome and if they so, it is just by chance. More precisely, the other variables that are indeed statistically significant impact the outcome in at least 95% of the predictions.

The features that are not statistically significant can be removed from the training and test datasets since they have no relevant impact on the results.

Considering the Telco problem, the features gender\_Male, InternetService\_Yes, CableService\_Yes, Contract\_Month-to-month are not determinant for a customer churning or not.

Other relevant information from the analysis is the feature Coeficient

1. Sign of the Coeficient

Example – SeniorCitizen = 0.2518

A positive sign means that, all else being equal, senior citizens were **more** likely to have churned than non-senior citizens.

Example – tenure = -0.0604

A negative sign means that, all else being equal, the more time a customer has relationship with the Telco, **less** likely to churn than a new customer

1. Magnitude

If everything is a very similar magnitude, a larger pos/neg coefficient means larger effect, all things being equal.

As an example, “SeniorCitizen” has more than 3 times the effect than “tenure” on the prediction result.

Another interesting example is that verifying the Coefficients of the types of the contracts as below, we will see that a customer with a 2 years contract will have about 150 times more impact on a customer to not churn, than if the customer had a month-to-month contract. It makes sense, since a 2 years contract bind a customer to a Telco cause of high contract breaking fees.

Contract\_Month-to-month -0.0660

Contract\_One year -0.8018

Contract\_Two year -1.5384

### Odds Ratio

Odds Ratio = (Probability of an Event / Probability of Non Event). For the features below with Odds Ratio (OR) > 1, it means that these variables, as they increase, **increases** the odds of observing the predicted outcome as 1 (churn). For those features with OR < 1, as they increase, **decreases** the odds of observing the predicted outcome as 1 (churn).

In the example below for "Tenure": An increase in 1 unit (month) of tenure changes the odds of churn vs. not churn by a factor of 0.94. So it means, that more time the customer has contracted the Telco services, the chance of churn reduces.

On the other hand, once a customer become a Senior Citizen, it increases the chance of churn by the factor 1.28 (or in 28%)

2.5% 97.5% OR

SeniorCitizen 1.092789 1.514318 1.286402

tenure 0.930094 0.952899 0.941428

MonthlyCharges 1.021334 1.035439 1.028362

TotalCharges 1.000152 1.000427 1.000289

gender\_Male 0.873183 1.123518 0.990473

Dependents\_Yes 0.717046 0.984235 0.840084

PhoneService\_Yes 0.249096 0.469518 0.341987

InternetService\_Yes 0.657897 1.342903 0.939942

DeviceProtection\_Yes 0.715029 0.979720 0.836976

TechSupport\_Yes 0.476424 0.663032 0.562036

CableService\_Yes 0.719824 1.036723 0.863862

PaperlessBilling\_Yes 1.270659 1.696633 1.468279

Contract\_Month-to-month 0.747063 1.173079 0.936143

Contract\_One year 0.344601 0.583730 0.448502

Contract\_Two year 0.152391 0.302582 0.214734

PaymentMethod\_Bank transfer (automatic) 0.419822 0.653148 0.523647

PaymentMethod\_Credit card (automatic) 0.384589 0.598889 0.479923

PaymentMethod\_Electronic check 0.609078 0.903231 0.741713

PaymentMethod\_Mailed check 0.401260 0.583041 0.483685

## Intermediate Step – Remove Non Statistically Significant Features

After the Logistic Regression analysis, the features that were not statistically significant were removed from the dataset as below.

The following columns have been eliminated using Logistic Binomial Model

['gender\_Male', 'InternetService\_Yes', 'CableService\_Yes', 'Contract\_Month-to-month']

Number of Columnns KEPT after eliminating features using Logistic Binomial Model : 15

Number of Columnns REMOVED using Logistic Binomial Model : 4

## Step 4 and 5 – Split the Dataset, Train and Select the Best Model

The dataset was split with 30% of the rows for training.

By analysing the best model to predict using this dataset, the following accuracy were calculated for the many different models requested.

**Linear SVM** model was chosen as the best one.

Classifier\_Name ... AccuracyScore

2 Linear SVM ... 0.794653

9 XGBoost ... 0.790102

0 Logistic Regression ... 0.789534

7 Random Forest ... 0.783845

4 MPL Neural Net ... 0.782139

5 Decision Tree ... 0.781570

8 Bagging ClassifierAdaBoost ... 0.762230

3 RBF SVM ... 0.754266

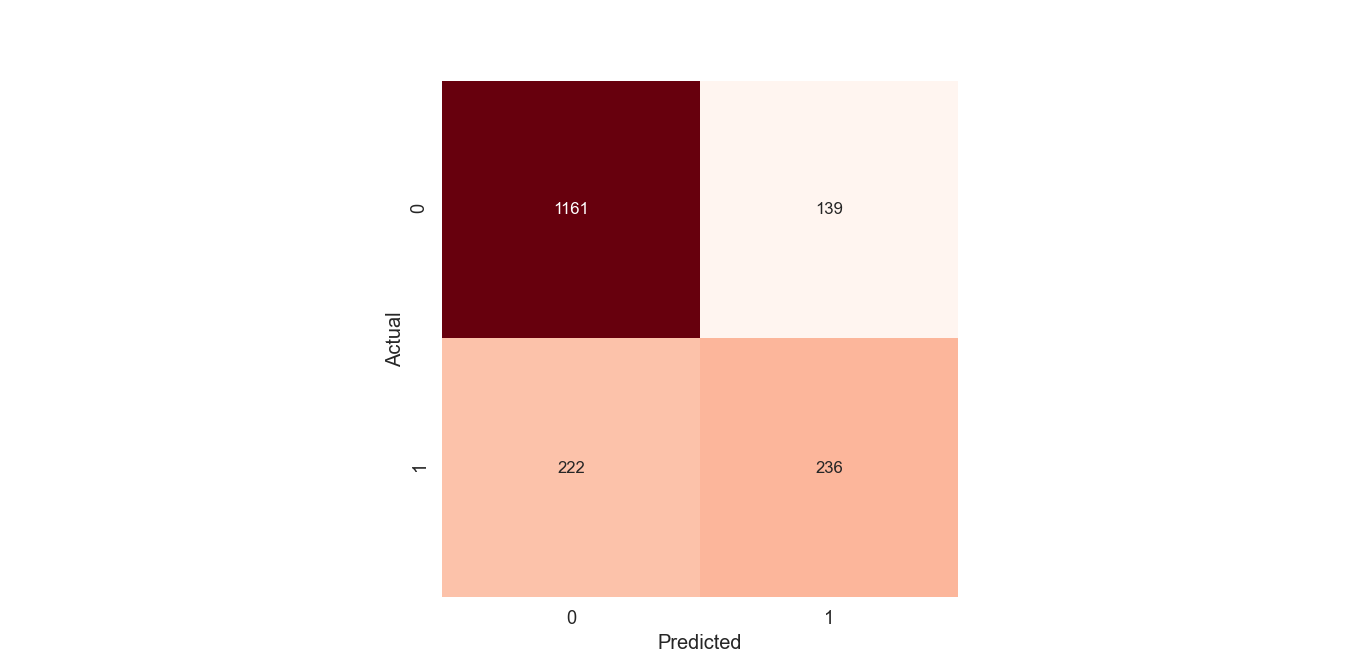
1 Nearest Neighbors ... 0.747440

6 Naive Bayes (GaussianNB) ... 0.741183

## Step 6 - Assess the predictive performance of Linear SVM selected

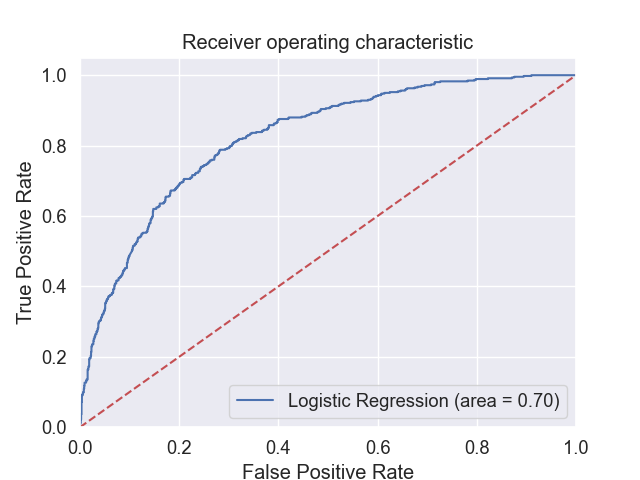
### Confusion Matrix

As we can see on the Confusion Matrix, the model made very good predictions for Customers that will not Churn. On the other hand, it is interesting to notice that when predicting customers that will Churn, the model was a little better than a guess, almost predicting wrongly 50% of the customers that indeed would churn. I interpreted that the model is quite confident in predicting that a client will not churn, but not so confident that a client will churn.



### ROC graph

By the Receiver Operator Characteristic (ROC) curve below it can be verified that the curve passes on the True Positive band of the graph, showing that, as overall, the model performed well (A random classifier would give points lying along the diagonal)



### Classification Report

The classification report shows the precision (Accuracy) of the prediction (Precision = TP/(TP + FP)), being TP = True Positive and FP = False Positive.

As it can be seen below and commented before when analysing the Confusion Matrix, the model is much more precising when predicting that a customer will not churn (0.84 precision) and the performance was not so good when predicting that a client will churn (0.63 precision).

The f1-score shows the percent of positive predictions were correct for each expected outcome.

precision recall f1-score support

0 0.84 0.89 0.87 1300

1 0.63 0.52 0.57 458

accuracy 0.79 1758

macro avg 0.73 0.70 0.72 1758

weighted avg 0.78 0.79 0.79 1758

## Step 7 and 8 - Predict whether a customer will churn or not (i.e., predict class membership) and the probability of it.

Since some features were removed from the dataset, the following data were used as input. The data informed that were not relevant for the prediction model were dismissed.

{'SeniorCitizen': 0, 'tenure': 32, 'MonthlyCharges': 64.75, 'TotalCharges': 2283.3, 'Dependents\_Yes': 0, 'PhoneService\_Yes': 1, 'DeviceProtection\_Yes': 0, 'TechSupport\_Yes': 0, 'PaperlessBilling\_Yes': 1, 'Contract\_One year': 0, 'Contract\_Two year': 0, 'PaymentMethod\_Bank transfer (automatic)': 0, 'PaymentMethod\_Credit card (automatic)': 1, 'PaymentMethod\_Electronic check': 0, 'PaymentMethod\_Mailed check': 0}

### Prediction

Test customer profile below to see if the customer will change Telco operator (Churn)

PREDICTION - The customer with the profile above will NOT CHURN

### Probability of Prediction

The probability that this prediction is right is : 0.7669991603276238

# Conclusion

## Dataset and Model Performance

The dataset and a classification model can be used by a Telco company with an accuracy of 79.4% to predict if a customer will churn or not.

### Business Perspective for Avoiding Churning

Nevertheless, given the low level of precision for predicting that a customer specifically will churn, for sack of campaign investments to revert potential lost of customer, it would be better to start investing on those customers predicted as churn who have more probability of doing so (the csv file with the predictions and the probabilities of predictions is attached in the zip file delivered). For instance, 146 out 375 customers predicted to churn have prediction probability greater than 70%.

# Bonus

By setting the variable REDUCE\_FEATURES\_BEST\_MODEL = True in the program, the features reducing will be executed by the best reducing model, according to its accuracy. The result is as below.

Model\_Name Columns\_Kept Model\_Dataset\_Accuracy

0 LinearSVC 15 0.793365

1 Logistic Binomial Model 15 0.792891

2 SelectKBest / Mutual Info Classification 9 0.78436

3 Logistic Regression 9 0.747867