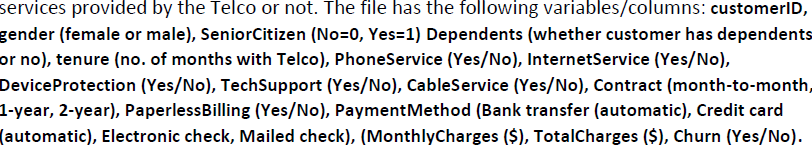
Problem Statement:

In the modern world a big challenge for the telecom companies to keep the customer continue in the business, To find out whether a customer will continue in the business or cancel the subscription with the company.

Variables:



Dependent and Independent Variable:

We are analyzing the churn of a customer; it is dependent variable that depends on all other variables

Complexity In the data set:

There are some complexities in the data set

1. Nonnumeric data: some variables are not in the numeric form such as Senior citizen, even the final variable churn was not in the numeric form
2. Missing values: in some rows there are some missing values in the data set.

Model for all data set:

Results: Generalized linear model

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

Model: GLM AIC: 5938.5096

Link Function: logit BIC: -56235.9489

Dependent Variable: Churn Log-Likelihood: -2952.3

Date: 2020-07-12 21:02 LL-Null: -4071.7

No. Observations: 7032 Deviance: 5904.5

Df Model: 16 Pearson chi2: 8.47e+03

Df Residuals: 7015 Scale: 1.0000

Method: IRLS

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

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

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

SeniorCitizen 0.2377 0.0832 2.8574 0.0043 0.0747 0.4008

tenure -0.0687 0.0057 -12.0011 0.0000 -0.0799 -0.0575

MonthlyCharges 0.0290 0.0035 8.2691 0.0000 0.0221 0.0359

genderdummies -0.0427 0.0635 -0.6728 0.5011 -0.1671 0.0817

Dependentsdummi -0.1981 0.0804 -2.4631 0.0138 -0.3557 -0.0405

PhoneServicedmi -1.3943 0.1327 -10.5065 0.0000 -1.6544 -1.1342

internetservice -0.3605 0.1599 -2.2539 0.0242 -0.6739 -0.0470

DeviceProtection -0.1981 0.0805 -2.4617 0.0138 -0.3559 -0.0404

TechSupport -0.5847 0.0846 -6.9078 0.0000 -0.7505 -0.4188

CableService -0.1822 0.0931 -1.9579 0.0502 -0.3646 0.0002

oneyear -0.7627 0.1058 -7.2104 0.0000 -0.9700 -0.5554

twoyear -1.5239 0.1752 -8.6973 0.0000 -1.8673 -1.1805

PaperlessBilling 0.3506 0.0730 4.8021 0.0000 0.2075 0.4937

Credit card (automatic) -0.1951 0.1082 -1.8031 0.0714 -0.4071 0.0170

Electronic check 0.2377 0.0875 2.7172 0.0066 0.0662 0.4091

Mailed check -0.2616 0.1000 -2.6157 0.0089 -0.4577 -0.0656

TotalCharges 0.0004 0.0001 5.7204 0.0000 0.0002 0.0005

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

From the logit model of regression on over all of the data, the P value for the most of the variables is less the 0.025 the critical value of model, excepting the Gender, Internet service, cable service and, credit card automatic.

ODDS Ratio for the model:

The odds ratio define the association between the two variables, dependent and independent. We have churn of an customer is the dependent variable that depends on other variables.

2.5% 97.5% OR

SeniorCitizen 1.077516 1.492952 1.268337

tenure 0.923229 0.944169 0.933640

MonthlyCharges 1.022374 1.036527 1.029426

genderdummies 0.846152 1.085106 0.958209

Dependentsdummi 0.700666 0.960340 0.820291

PhoneServicedmi 0.191202 0.321675 0.248002

internetservice 0.509694 0.954080 0.697344

DeviceProtection 0.700565 0.960425 0.820269

TechSupport 0.472107 0.657853 0.557294

CableService 0.694455 1.000196 0.833421

oneyear 0.379080 0.573858 0.466410

twoyear 0.154544 0.307138 0.217868

PaperlessBilling 1.230591 1.638319 1.419894

Credit card (automatic) 0.665594 1.017118 0.822793

Electronic check 1.068475 1.505443 1.268278

Mailed check 0.632746 0.936512 0.769789

TotalCharges 1.000247 1.000504 1.000376

In the odds ratio it is clear indication that some attribute have a big weightage over the churn of a customer.

Selection of model:

The following table below the explanation of the result for a pipline pf different model will define the accuracy of the models.

We run a pipeline to select the best model for the further analysis on the data set, we get two model with highest accuracy score.

1. Linear Svm with an accuracy of 0.792378, followed by the Logistic regression with 0.790102

We opted Logistic regression for the further analysis, in the linear SVM, we can not perform the probability estimates. While Logistic regression allow us to perform the probability.

<https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

the above two links from the sklearn python define the major difference, and my answer why I opted logistic regression instead of the LinearSvm even the accuracy of the SVM was higher than the Logistic.

Classifier Accuracy Score

0 **Logististic Regression 0.790102**

1 Nearest Neighbors 0.753129

2 **Linear SVM 0.792378**

3 RBF SVM 0.748009

4 Neural Net 0.755973

5 Naive Bayes 0.732651

6 Decision Tree 0.781570

7 Majority Vote 0.794653

8 Random Forest 0.788965

9 Bagging 0.781001

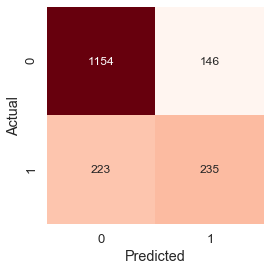
10 AdaBoost 0.785552

11 XGBoost 0.771900

**Confusion matrix:**

A confusion matrix defines the results of a classification model, on a data set

<matplotlib.axes.\_subplots.AxesSubplot at 0x1e3e8642190>



precision recall f1-score support

0 0.84 **0.89** 0.86 1300

1 0.62 **0.51** 0.56 458

accuracy **0.79** 1758

macro avg 0.73 0.70 0.71 1758

weighted avg 0.78 0.79 0.78 1758

True Positive rate: when a customer is not cancelling their subscription and we also get the same result is (1154 out of 1300) 0.89 or **89%,** while the value of false positive is 0.11 or **11%**

True negative rate: the total number of true negatives was 235 out of 458 that is almost **51%,**

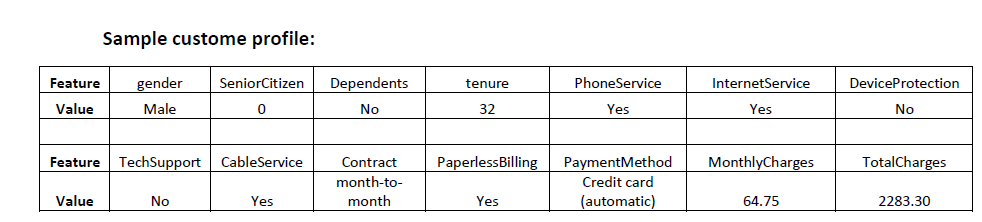
Accuracy of overall model: The total accuracy of the model is 0.79 **or 79%**



Finding of Roc and confusion matrix:

In the confusion matrix we have almost 51% of the true negative success, that define that on those circumstances that are considerable put a consumer to cancel or change their subscription, it is a high rate, we need to put some efforts to put this rate down, otherwise it will put some extra burden on the overall revenue and the profit for the company.

Prediction Of the model and accuracy :



On these values we find out that a customer is not going to change or cancel their subscription.

And the probability of the success was:

