**Project Classification**

# **Step 1 to 8 – See the Code File**

# **Step 9 – Discussion about the implementation the Logistic Regression using the full dataset**

**Results**

1) Logistic Regression Model Score (full dataset): 0.797

2) Confusion Matrix

A picture containing clock

Description automatically generated

3) Classification Report

|  |
| --- |
| Classification Report - Logistic Regression (full dataset)  precision recall f1-score support  0 0.84 0.90 0.87 5163  1 0.65 0.51 0.57 1869  micro avg 0.80 0.80 0.80 7032  macro avg 0.74 0.71 0.72 7032  weighted avg 0.79 0.80 0.79 7032 |

4) ROC Curve



5) Odds Ratio

|  |  |  |
| --- | --- | --- |
| Variables | Coefficients | Odds Ratio |
| SeniorCitizen | 0.289120 | 1.335252 |
| tenure | -0.078535 | 0.924469 |
| MonthlyCharges | 0.015648 | 1.015771 |
| TotalCharges | 0.000406 | 1.000406 |
| gender\_Male | -0.155258 | 0.856194 |
| Dependents\_Yes | -0.305607 | 0.736676 |
| PhoneService\_Yes | -0.481387 | 0.617926 |
| InternetService\_Yes | 0.074793 | 1.077661 |
| DeviceProtection\_Yes | -0.123173 | 0.884110 |
| TechSupport\_Yes | -0.435762 | 0.646771 |
| CableService\_Yes | 0.079890 | 1.083168 |
| Contract\_One year | -0.264745 | 0.767402 |
| Contract\_Two year | -0.332027 | 0.717468 |
| PaperlessBilling\_Yes | 0.293177 | 1.340680 |
| PaymentMethod\_Credit card (automatic) | -0.178528 | 0.836501 |
| PaymentMethod\_Electronic check | 0.404295 | 1.498246 |

**Discussions**

1) Using the whole dataset to create the model we got an accuracy score almost 80%. So, even using all data, we can note that 20% of the predictions were incorrect.

2)

True Positive = 4653 – it is positive (true) and predicted positive(true)

False Positive = 510 – it is negative (false) and predicted positive(true)

False Negative = 918 – it is positive (true) and predicted negative(false)

True Negative = 951 – it is negative (false) and predicted negative (false)

3) The precision to predict 0 (‘no churn’) is greater than the prediction of ‘churn’ (84% vs. 65%). Similar behaviour in relation to recall (90% vs 51%). It is due to unbalanced relationship between the total of “0’s” and “1’s” in the dataset.

4) ROC Curve: the area of the under de curve (A = 0.71) can be considered as the probability of the choice of a positive instance be greater than a negative one. In general, the greater the area the greater the predictive accuracy. Looking the graphic, we can see that with a rate about 20% of false positive, we have about a rate about 70% of true positive.

5) Odds Ratio is defined as the quotient between the probability of success over probability of failure. So, when the odds ratio is greater than 1 means that the probability of success of that variable is bigger than failure and proportion to the value. The odds ratio below than 1 is the opposite and when it is equal to 0 means that the probability is equal for both cases. The interpretation for each variable is shown in the table below.

|  |  |  |
| --- | --- | --- |
| Variables | Odds Ratio | Interpretation |
| tenure | 0.924 | One-unit increase in tenure, it is expected to see 7.6% decrease in the odds of being in ‘churn’ class (holding other variables) |
| MonthlyCharges | 1.016 | One-unit increase in monthly charges, it is expected see 1.6% increase in the odds of being in ‘churn’ class (holding other variables) |
| TotalCharges | 1.000 | One-unit increase in total charges, it is expected see no change in the odds of being in ‘churn’ class (holding other variables) |
| SeniorCitizen | 1.335 | The odds of getting into an ‘churn’ class for customer senior is 33.5% greater than the odds for customer no senior. |
| gender\_Male | 0.856 | The odds of getting into an ‘churn’ class for males is 14.4% smaller than the odds for females. |
| Dependents\_Yes | 0.737 | The odds of getting into an ‘churn’ class for customer with dependent is 26.3% smaller than the odds for customer without dependent. |
| PhoneService\_Yes | 0.618 | The odds of getting into an ‘churn’ class for customer with phone service is 38.2% smaller than the odds for customer without phone service. |
| InternetService\_Yes | 1.078 | The odds of getting into an ‘churn’ class for customer with internet service is 7.8% greater than the odds for customer without internet service. |
| DeviceProtection\_Yes | 0.884 | The odds of getting into an ‘churn’ class for customer with device protection is 11.6% smaller than the odds for customer without device protection. |
| TechSupport\_Yes | 0.647 | The odds of getting into an ‘churn’ class for customer with tech support is 35.3% smaller than the odds for customer without tech support. |
| CableService\_Yes | 1.083 | The odds of getting into an ‘churn’ class for customer with cable service is 8.3% greater than the odds for customer without cable service. |
| Contract\_One year | 0.767 | The odds of getting into an ‘churn’ class for customer with contract one year is 23.3% smaller than the odds for customer without contract one year. |
| Contract\_Two year | 0.717 | The odds of getting into an ‘churn’ class for customer with contract two year is 28.3% smaller than the odds for customer without contract two year. |
| PaperlessBilling\_Yes | 1.341 | The odds of getting into an ‘churn’ class for customer who pays by paperless billing is 34.1% greater than the odds for customer who does not pay by paperless billing. |
| PaymentMethod\_Credit card (automatic) | 0.837 | The odds of getting into an ‘churn’ class for customer who pays by credit card is 16.3% smaller than the odds for customer who does not pay by credit card. |
| PaymentMethod\_Electronic check | 1.498 | The odds of getting into an ‘churn’ class for customer who pays by electronic check is 49.8% greater than the odds for customer who does not pay by electronic check. |

**Managerial Implications**

The managerial implications can be observed in the “interpretation” column in the table above. The manager should monitor all customers that can have tendence to become a “customer churn” based on the odds ratio. The table can be used as reference to know where the failure points are.

# **Step 10 – Model Selected**

After the pipeline processing, the ranking of the best models is in the table below. It was selected the **Random Forest Model** because it presented the best Accuracy Score among the models.

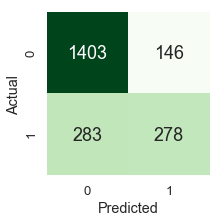
|  |  |
| --- | --- |
| Classifier | Accuracy Score |
| Random Forest | 0.796682 |
| Logistic Regression | 0.790521 |
| AdaBoost | 0.789100 |
| Linear SVM | 0.787678 |
| Neural Net | 0.787204 |
| XGBoost | 0.781043 |
| Decision Tree | 0.778673 |
| Bagging | 0.764929 |
| Nearest Neighbors | 0.759242 |
| RBF SVM | 0.745498 |
| Naive Bayes | 0.735071 |

# **Step 11 – Discussion about the implementation the Best Model**

**Results**

1) Random Forest Model Score: 0.797

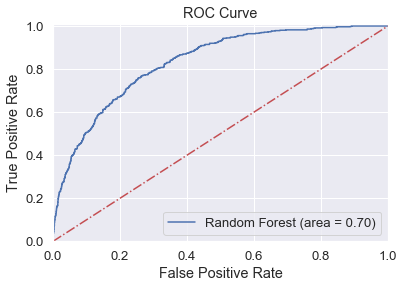
2) Confusion Matrix



3) Classification Report

|  |
| --- |
| Classification Report - Random Forest  precision recall f1-score support  0 0.83 0.91 0.87 1549  1 0.66 0.50 0.56 561  micro avg 0.80 0.80 0.80 2110  macro avg 0.74 0.70 0.72 2110  weighted avg 0.79 0.80 0.79 2110 |

4) ROC Curve



**Discussions**

1) Using the best model and applying the train and test dataset, we got an accuracy score almost 80%. So, it is expected an accuracy around this value when we apply this model in a specific profile customer. The accuracy is the same obtained when we used Logistic Regression with whole dataset.

2)

True Positive = 1403 – it is positive (true) and predicted positive(true)

False Positive = 146 – it is negative (false) and predicted positive(true)

False Negative = 283 – it is positive (true) and predicted negative(false)

True Negative = 278 – it is negative (false) and predicted negative (false)

3) The precision to predict 0 (‘no churn’) is greater than the prediction of ‘churn’ (83% vs. 66%). Similar behaviour in relation to recall (91% vs 50%). Again, the values are practically the same to the values obtained using Logistic Regression with whole dataset.

4) ROC Curve: as the observed values in the Logistic Regression, the values are almost the same for Random Forest.

# **Step 12 – Discussion about the expected outcome for the specific customer**

Sample Customer Profile

|  |  |
| --- | --- |
| Feature | Value |
| Gender | Male |
| SeniorCitizen | 0 |
| Dependents | No |
| Tenure | 32 |
| PhoneService | Yes |
| InternetService | Yes |
| DeviceProtection | No |
| TechSupport | No |
| CableService | Yes |
| Contract | Month-to-month |
| PaperlessBilling | Yes |
| PaymentMethod | Credit card (automatic) |
| MonthlyCharges | 64.75 |
| TotalCharges | 2283.30 |

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

Predicted Class using Random Forest is **'No Churn'** with a probability of **0.843**

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

The predicted accuracy for this customer is about 84% for the class “No Churn”, so it is expected that this customer does not become a “customer churn”. It is interesting to note that both the accuracy of model (80%) and precision on prediction of class “No Churn” (83%) are very close to this result.

**Managerial Strategies**

As the goal is monitoring the customers to avoid them to become a “customer churn”, thus the manager should keep a special attention or strategies in a customer with this profile: low tenure, senior, female, no dependents, no phone service, with internet service, no device protection, no tech support, with cable service, no contract of one or two years, and with payment by billing or electronic check. This profile has the critical characteristics observed before and should be monitored more carefully. If a customer changes his profile to one of these characteristics, it is important to include him in the monitoring process and evaluate what were the reasons that led the customer to make the changes.