MLA TEAM PROJECT

Module 6

**Machine Learning Applications**

2023SP 53:716:545:01

Dr. Ram Gopalan

**Group 3**

Suchetana Ghosh

Parth Gupta

Westly Mandoske

Arvind Sarma

# Part A: Telco Churn Analysis

We have submitted one Jupyter notebook named TA\_ProjectA\_APPENDIX\_A2\_A3

**Part A.1: Describe the numbers below in a table:**

| **Cross-validation Fold** | **Decision Tree** | **Logistic Regression** |
| --- | --- | --- |
| **Fold 1** | 0.706 | 0.6 |
| **Fold 2** | 0.701 | 0.632 |
| **Fold 3** | 0.707 | 0.635 |
| **Fold 4** | 0.673 | 0.645 |
| **Fold 5** | 0.681 | 0.628 |
| **Fold 6** | 0.709 | 0.638 |
| **Fold 7** | 0.706 | 0.64 |
| **Fold 8** | 0.686 | 0.633 |
| **Fold 9** | 0.703 | 0.630 |
| **Fold 10** | 0.704 | 0.625 |
| **Average Error %** | **0.70** | **0.57** |
| **Std. Dev. Error %** | **0.01** | **0.01** |

**Part A.4 :**

| **Confusion Matrix** | **p Actual Churn** | **n Actual Retain** |
| --- | --- | --- |
| **Y Predicted Churn** | $0 BENEFIT | $300 BENEFIT |
| **N Predicted Retain** | $400 COST | $200 BENEFIT |

**Write a few sentences supporting the $benefit/cost numbers. How did you come up with these numbers?**

1. We assumed a customer's gross revenue is $300/yr. and that the cost of marketing acquisition/retention is $100. We decided on a basic strategy of only spending money on customers we predicted to retain. Consequently, customers predicted to churn were not targeted for any offer.
   1. In case (Y/n), we get the full benefit of $300/per customer/per year without having paid additional marketing costs.
   2. In case (Y/p), we spend no marketing dollars on the customer, and since they left, we did not earn any revenue.
   3. In case (N/p), we predicted that customers would not leave, but they left. We incur the costs of losing customers for both marketing and also the lost revenue.
   4. In case (N/n), We predicted correctly that the customers would not leave, and they did not leave. We decided to target customers predicted to retain, so we incurred a $100 marketing expense. Each customer retained is a gross revenue of $300/year. Since they are in the system, they will also generate some business. Here we kept the net benefit to $200.

Keeping these cost-benefit numbers, we have calculated the expected net revenue from the confusion matrix. The expected net revenue is $44.34/per customer, or a 14.8% gross margin.

**Part A.5:**

| **Churn Segment** | **RULES for identifying** | **Description in words** |
| --- | --- | --- |
| **Churn segment 1…** | **Entropy, Samples & overall model accuracy** | If the conditions of  House <= 604338 is True, overage <= 98.5 is True, Leftover < = 24.5 is False, Handset price <= 801 is False,  Then it is a Churn |
| **Churn segment 2…** | **Entropy, Samples & overall model accuracy** | If the conditions of  House <= 604338 is False, Income<= 100012 is True , Leftover < = 49.5 is True or False)  Then it is a Churn |

**A.5 The Rules are as stated by the decision tree. WHICH CHURN SEGMENT DO YOU RECOMMEND FOCUSING ON? WHY?**

We recommend churn segment 2, where:

House <= 604338 (false), Income<= 100012 (True), Leftover < = 49.5 (True) where entropy is 0.679 (lowest amounts of the other rules) & samples are 17.6% (which is highest among the different rules of churn). Hence we recommend churn segment 2 as the entropy for that is lowest, and the sample size is highest.

All the churn segments rules:

IF House = TRUE

Segment 1 Rules:

House <= 604338 (True), overage <= 98.5 (True), leftover <= 2.5 (False) - Leave

House <= 604338 (True), overage <= 98.5 (True) , Leftover < = 24.5 (False), Handset price <= 801(False) – Leave

IF House = FALSE

Segment 2 Rules:

House <= 604338 (False), Income<= 100012 (True), Leftover < = 49.5 (True or False) – Leave

House <= 604338 (False), Income<= 100012 (False), overage <=144(True), Leftover < = 26.5 (True) – LEAVE

**Appendices for PART A, place Jupyter notebooks Parts A.2 and A.3.**

**Appendix A.2: Decision Tree cross-validation notebook: construct this notebook by combining ideas from Churn\_Telco and Iris\_practice\_crossval notebooks.**

**Appendix A.3: Logistic regression cross-validation notebook:**

**construct this notebook by combining ideas from Churn\_Telco, WBCD and Iris\_practice\_crossval notebooks.**

# Part B: Simmons Data Analysis

## Part B.1 (2 points): What are the coefficients (BETAs) for the logistic regression model? Answer as below:

| **LR coefficients** | **Value** |
| --- | --- |
| **BETA0 (or constant term)** | -2.00672 |
| **BETA1 (coeff. For X1 )** | 0.32989 |
| **BETA2 (coeff. For X2)** | 0.9178 |

## B.2: Predict Jack & Jill’s probability of response

| **New Customer** | **Probability of Response** |
| --- | --- |
| **Jack** | .394 |
| **Jill** | .335 |

Jack is more likely to respond because the probability of Jack responding is higher than Jill responding.

## B.3: What cutoff probability would you choose?

We would select a 40% cutoff because we are trying to affect the outcome of new customer sales. By offering customers a coupon, we hope they will come to the store and increase footfall. Since we can reasonably expect customers with a greater than 50% probability to use any coupon offered, it is not meaningful to our model if they use it or not. However, if a coupon offer can change the model’s predicted probabilities, the coupon campaign can be understood to have a meaningful indirect effect on gross revenue.

The specific reason for choosing a 40% threshold is to minimize the marketing cost associated with making a coupon offer to additional customers. A 40% threshold is an acceptable trade-off between the risk of sunk costs and potential revenue.