Data Science: Machine Learning Team Project

Devansh Shah

Jennifer Indrupati

Rithika Mothukuri

Part A

A.1

Cross Validation Fold	Decision Tree	Logistic Regression
Fold 1	0.6090	0.6200
Fold 2	0.6280	0.6340
Fold 3	0.6275	0.6425
Fold 4	0.6085	0.6385
Fold 5	0.5965	0.6260
Fold 6	0.6090	0.6265
Fold 7	0.6120	0.6355
Fold 8	0.6050	0.6270
Fold 9	0.6050	0.6330
Fold 10	0.6120	0.6430
Average Error %	0.61125	0.63260
Std. Dev. Error %	0.009770	0.007619

The technique with the higher average accuracy is clearly Logistic Regression model with an accuracy of 63%

A.4

The benefit/cost analysis assigns financial values to different outcomes of the predictive model, specifically in predicting churn. The Python code provided in the appendix loads the churn dataset, performs cross-validation for Decision Tree and Logistic Regression models, and conducts the benefit/cost analysis using defined values for each cell in the confusion matrix. The benefit/cost analysis is implemented by iterating through the defined benefit/cost values and printing out the associated numbers for each cell.

	p	**n**
Y	$100	$-20
N	$-50	$0

For each of the four cells in the confusion matrix, a benefit/cost analysis is conducted based on a business understanding of the costs and benefits of misclassification:

Cell ('p', 'Y'): True Positive - Predicted Positive, Actually Positive
Benefit: $100
This cell represents the case where the model correctly predicts that a customer will respond positively to the coupon offer (p) and they actually do respond (Y). The benefit associated with this cell is $100, indicating the potential revenue or profit gained.

Cell ('p', 'N'): False Positive - Predicted Positive, Actually Negative
Benefit: $-50
This cell represents cases where the model incorrectly predicts that a customer will respond positively to the coupon offer (p), but they actually do not respond (N). Here there is a $50 loss incurred.

Cell ('n', 'Y'): False Negative - Predicted Negative, Actually Positive
Cost: $-20
This cell represents cases where the model incorrectly predicts that a customer will not respond positively to the coupon offer (n), but they actually do respond (Y). The loss faced here is $20.

Cell ('n', 'N'): True Negative - Predicted Negative, Actually Negative
No Cost/Benefit: $0
This cell represents cases where the model correctly predicts that a customer will not respond positively to the coupon offer (n) and they actually do not respond (N). There is no additional benefit or cost associated with this scenario.

A.5

After we studied the "CHURN/LEAVE" node-leaves of the decision tree, several unique churning segments emerged, each representing a subgroup of customers with similar characteristics likely to churn.

Segment 1: High-Value, High-Risk Customers

These are customers with high annual spend and a history of limited usage or complaints. This segment represents customers who contribute significantly to revenue but are at risk of churn.

Strategies which focus on retention strategies, such as loyalty programs can help churn while maximizing revenue.

Segment 2: New Customers with Low Engagement

These are customers who recently joined but showcase low amounts of interaction with the service. Customers in this segment may have signed up out of curiosity but haven't fully utilized the service.

Segment 3: Price-Sensitive Customers

These are customers who are constantly on the lookout for discounts and/or switching providers. These customers usually prioritize cost savings over brand loyalty.

Segment 4: Inactive Customers

These are customers who go on long periods of inactivity or low usage. Customers who may have forgotten about the service or found alternatives fall into this category.

Our recommendation:

We chose to focus resources on *Segment 1: High-Value, High-Risk Customers*. These customers contribute significantly to revenue as compared to the rest, which make them valuable and worthy to implement efforts and resources. By implementing personalized retention strategies tailored to their preferences and addressing any critical points of concern proactively, the likelihood of churn can be reduced.

Additionally, targeting high-value customers ensures that resources are allocated effectively, maximizing the return on investment in retention efforts, with low risk of loss. Techniques like loyalty programs help with producing loyal customers. Addressing churn within this segment can yield significant long-term benefits by making customers advocates for the company’s products and turning them into loyal assets.

Part B

B.1

Coefficient	Value
BETA 0	-2.0067206154422275
BETA 1	0.32989442356674037
BETA 2	0.9178862828888504

B.2

Customer	Probability of Response
Jack	0.3943542838553842
Jill	0.3346689457446288

Based on the probabilities of response calculated using the logistic regression model:

- Jack has a probability of response of approximately 0.394.
- Jill has a probability of response of approximately 0.335.
- Therefore, Jack is more likely to use the coupon compared to Jill. This conclusion is drawn based on their respective probabilities of response, with Jack having a higher probability of responding positively to the coupon offer.

B.3

In rolling out the logistic regression model to predict coupon usage for a large database of customers, determining the optimal cutoff probability is crucial for effectively balancing between false positives and false negatives.

Based on the concept of a confusion matrix, which evaluates the performance of a classification model, we aim to minimize misclassifications and maximize the accuracy of our predictions. The cutoff probability determines the threshold above which a customer is classified as likely to use the coupon, and below which they are classified as unlikely to use it.

The optimal cutoff probability identified for this logistic regression model is 0.478. This probability is determined by analyzing the trade-off between sensitivity (true positive rate) and specificity (true negative rate) in the confusion matrix. By selecting a cutoff probability close to the optimal value, we aim to achieve a balance between capturing as many true positives as possible while minimizing false positives and false negatives.

Therefore, the chosen cutoff probability of 0.478 provides a balanced approach to predicting coupon usage for a large database of customers, optimizing the model's performance and accuracy in identifying potential coupon responders.

Appendix (Python file)

file:///Users/rithika/Downloads/ML%20Group%20Project%20Final.html