A red and white logo

Description automatically generated

**College of Professional Studies, Northeastern University**  
**ALY6020**

**Predictive Analytics**  
**Professor Ajit Appari**  
**October 8th, 2024**

**INTRODUCTION:**

In this analysis, we examine a dataset to comprehend subscription patterns and pinpoint elements that are causing a magazine company's subscriptions to drop. The project is broken up into four sections, each of which focuses on a distinct facet of model construction and assessment.

We preprocessed the data and used the required data cleansing procedures to make sure the dataset was ready for analysis. To forecast subscription behaviors, Part 2 involved developing a Logistic Regression model. This gave rise to a baseline for comprehending how various characteristics affect the probability of renewing a membership.

We were able to evaluate the Quadratic Discriminant Analysis (QDA) model's performance in comparison to the Logistic Regression model. We examined key variables and their business ramifications, evaluating both models on accuracy, precision, and recall.

**DATA ANALYSIS:**

• Several important steps were engaged in the exploratory data analysis (EDA) process to understand the dataset and get it ready for modeling. The dataset, which contained a few variables about subscriber habits and customer demographics, was first imported along with the required libraries.

• After that, to get a sense of the structure and kind of data we were dealing with, we performed a quick evaluation of the data by looking over its shape and the first few rows. To make sure that no important data was lost during the analysis, we next looked over the dataset for missing values and carried out the necessary imputations where needed.

• Using box plots and histograms, among other descriptive statistics and visualizations, we investigated the distribution of numerical features. This made it easier to spot any outliers and comprehend each feature's underlying distribution. To evaluate the representation of categorical variables in the dataset, frequency counts and visualizations such as bar charts were employed for analysis.

• To look at the links between numerical variables, we also performed a correlation analysis. This gave us information about multicollinearity problems that might have an impact on the performance of the model. Categorical variables were formatted into a machine learning algorithm format using one-hot encoding.

• Overall, the EDA process was crucial in identifying key trends and patterns in the data, enabling informed decisions for model selection and feature engineering. This thorough examination laid the groundwork for building robust predictive models for subscription behavior analysis.

• Visualizations:

1) Income Distribution Histogram:

A graph showing the amount of income

Description automatically generated

2) Scatter plot of Income vs Amount Spent on wines:

A graph with a red line

Description automatically generated

3) Box plot for Income distribution by Education level:

A chart with colorful squares

Description automatically generated

• Part 2:

Implementing a Logistic Regression model to predict customer subscription behavior based on various features. It begins by preparing the dataset, excluding non-predictive columns like 'ID' and 'Dt\_Customer,' and encoding categorical variables. The data is split into training and testing sets (80/20 split) using stratified sampling to maintain class distribution. The model is initialized with balanced class weights to address potential class imbalance. After fitting the model, predictions are made on the test set. The confusion matrix and classification report indicate an overall accuracy of 82%, with a precision of 45% for class 1 and a recall of 79%.

ROC curve:

A graph with a red line

Description automatically generated

• PART 3:

We implemented a Quadratic Discriminant Analysis (QDA) model to predict customer subscription behavior using a dataset of customer features. Initially, it defines the features and target variable by dropping non-predictive columns such as 'ID' and 'Dt\_Customer'. Categorical variables are one-hot encoded to ensure compatibility with the QDA model. The dataset is then split into training and testing sets using an 80/20 ratio while preserving class distribution through stratified sampling.

The QDA model is initialized and fitted to the training data, followed by predictions on the test set. The evaluation of the model is performed using a confusion matrix and a classification report. The results show an overall accuracy of 82%, with a precision of 43% for the positive class (subscribers) and a recall of 55%. The weighted average precision, recall, and F1-score are 84%, 82%, and 83%, respectively. These metrics indicate the model's performance in distinguishing between customers who subscribed and those who did not, highlighting the challenges in accurately predicting the positive class within this dataset.

ROC curve:

A graph of a curve

Description automatically generated

• PART 4:

Some valid reasons why the Logistic Regression model may be considered better than the Quadratic Discriminant Analysis (QDA) model based on the metrics provided:

Accuracy: Both models achieved the same accuracy of 0.82, indicating that they classify approximately 82% of the instances correctly. However, accuracy alone does not provide a full picture, especially when dealing with class imbalance.

Precision: The Logistic Regression model slightly outperforms the QDA model in terms of precision for the positive class (Class 1). Logistic Regression has a precision of 0.45 compared to 0.43 for QDA. This means that when the model predicts Class 1, it is correct about 45% of the time with Logistic Regression, compared to 43% with QDA.

Recall: Logistic Regression has a significantly higher recall of 0.79 compared to QDA’s 0.55. This indicates that Logistic Regression is better at identifying actual positive cases compared to QDA.

**CONCLUSION:**

Logistic Regression is the preferred model for this dataset because it demonstrates better precision and significantly better recall than the Quadratic Discriminant Analysis model. This suggests that it is more effective at identifying positive cases while maintaining a similar accuracy level.

**REFERENCES:**

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statisticallearning with applications in R. Springer.

ISLP. (n.d.). Classification lab. In Introduction to Statistical Learning with Applications in R. Retrieved October 8, 2024, from https://islp.readthedocs.io/en/latest/labs/Ch04-classification-lab.html.

GeeksforGeeks. (2021). Quadratic discriminant analysis. Retrieved October 8, 2024, from https://www.geeksforgeeks.org/quadratic-discriminant-analysis/.