# 

**Classification Problem**

Deepika Khosla

DeVos Graduate School, Northwood University

115537-MGT-665-NW Solv Probs W/ Machine Learning

Itauma Itauma  
 06/15/2025

# **Abstract**

Customer segmentation is a valuable marketing practice strategy nowadays and allows firms to target different groups of customers with different marketing strategies. The present study used machine learning algorithms of Logistic Regression, k-Nearest Neighbors k-NN, and Decision Trees, on a customer dataset of 2,627 and used demographic, professional (i.e., work experience), and behavioral (i.e., spending score) features. The most significant objective of the study was to use the outcomes of these models to predict customer graduation status in order to assist segmentation and targeting of marketing to customers. Data pre-processing finished cleaning up the dataset by normalizing features and handling instances of missing values that existed in Work Experience and Family Size. Exploratory analysis set the stage for the present study. I found the average customer age to be 43.6 years old and 2.55 (SD 3.34) years of work experience. Model performance was assessed by accuracy, precision, recall, and F1-score, with Logistic Regression performing best (70.1% accuracy with 89% recall for the "Graduated" customers). K-NN did (68.2 % accuracy), and Decision Trees (63.2 % accuracy). Overall, the findings point to the legitimacy of applying the demographic and behavioral features in discovering customer segments; there is, nevertheless, a single main limitation which is the ease of carrying out a binary classification. Future studies can expand by trying k-means clustering or neural networks (deep learning) to allow more detailed customer segmentation.

**Introduction**

Customer segmentation improves managerial decision-making by grouping customers with similar characteristics, thereby allowing a firm to allocate resources most optimally, and perhaps improve the potential for customizing marketing efforts (Kotler & Keller, 2016). That is, the conventional methods used to be performed on the basis of heuristics, while machine learning is data-driven. This study assumed a dataset consisting of 2,627 customers to predict graduation status—a proxy for socioeconomic segmentation—using three classifiers. This study addresses the issue of under-utilized structured demographic data for segmentation and provided additional information on model performance in handling data imperfections prevalent in the business context (e.g., missing values). The analysis conducted over the classifiers; Logistic Regression, k-NN, Decision Trees provided business-oriented observations for companies seeking segmentation solutions to scale.

**Literature review**

# Customer segmentation has moved from more standard demographic focused approaches (Smith, 1956) to more complex machine learning algorithms with supervised and unsupervised methods (Wedel & Kamakura, 2012). The supervised methods (e.g., Logistic Regression (James et al., 2013), Decision Trees (Quinlan, 1986)) provide meaningful segmentation but has issues around limitied capabilities on missing data (Logistic regression) and performance deteriorating with overfitting (Decision Trees), as shown in this analysis achieved (70.1%) accuracy in Logistic regression and the Decision Tree had the lowest accuracy and provided unreliable predictions with (63.2%) accuracy due to overfitting (Hastie et al., 2017). The unsupervised methods, e.g., k-nearest Neighbors (k-NN) (Tan et al., 2019) and k-means clustering (MacQueen, 1967) can be helpful in identifying latent structures but must be proceeded with a lot of data preprocessing steps including the best ways to delete missing values (e.g., list-wise deletion (Allison, 2001) or imputation (Van Buuren, 2018) identifying and encoding categorical variables (one-hot encoding (Garcia et al., 2019)). Recently emphasis has been placed on more business-like metrics such as recall, or F1-score (Saito & Rehmsmeier, 2015) and work has begun looking at hybrid methods (Zhang et al., 2021) and deep learning (Hinton & Salakhutdinov, 2006) with higher-dimensional datasets. We see this struggle to provide interpretable results, the performance of getting results using imperfect real-world data (including our dataset)

# **Methodolog****y**

# Research Design

# The research will be a quantitative study with a predictive modeling component wherein the researcher is leveraging machine learning methods (i.e., algorithms) to segment their customers to perform better segmentation based on demographic and behavioral profiles. The research will follow a systematic but flexible approach which will be:

# 1. Data pre-processing (including missing value imputation and encoding categorical features).

# 2. Exploratory data analysis.

# 3. Model fitting and validation.

# 4. Performance evaluations.

# 2. Data collection & pre-processing

# Dataset description

# ● Source: Structured CSV with observations for 2,627 customers

# ● Attributes:

# ○ Demographic: Age, Gender, Ever\_Married, Family\_Size,

# ○ Behavioral: Spending\_Score, Work\_Experience.

# ○ Target variable: Graduated (Binary - Yes/No).

# Data Cleaning

# ● Addressing missing values

# ○ Rows with missing target (Graduated) or missing features were omitted (abandoned) (n=2,338 observations remained).

# ○ Missing rate for Work\_Experience = 10.2%; Family\_Size = 4.3%.

# ● Encoding categorical feature

# ○ One-Hot encoding employed for Gender (Binary - Male\_Yes/No).

# Data Analysis Techniques

## The research study took an integrated approach to data analysis to allow for meaningful analysis of customer data, and to create optimal segment models from a range of methods of data analysis. The methodology used a combination of statistical and machine learning methods which encompassed key operations including:

## 1. Exploratory Data Analysis (EDA):

## ● Descriptive Statistics: The mean and median were used for central tendency, and standard deviation and IQR were used for dispersion for numerical features (age, Work Experience).

## ● Frequency analysis: Examined the distribution of the categorical variables (Gender, Graduated status).

## ● Missing value analysis: The missing data patterns were quantified and heatmaps were used to visualize the missingness.

## ● Correlation analysis: Pearson's correlations were used to investigate relationships between numerical variables.

## 2. Data Pre-processing Methods:

## ● Missing Data:

## ○ Listwise deletion for records with missing target values, records were deleted if there was missing data (e.g. Age, Work Experience)

## ○ Mean imputation for the missing numerical features (Work Experience).

## ● Feature Engineering:

## ○ one-hot encoding of categorical variables (Gender)

## ○ Standardization (z-score normalization) for k-NN algorithm

## ● Outliers: Outliers were detected for the numerical features using the IQR method, which constituted extreme values that were flagged.

## 3. Machine Learning Modeling:

## Supervised Learning:

## 1. Logistic Regression

## ○ Binary classification using L2 regularization

## ○ Probabilities using the sigmoid function

## ○ Wald tests used to assess feature

## 4. Model Evaluation Methods

**Performance Metrics:**

Accuracy

Precision

Recall

F1-Score

**Findings**

1. Model Performance Comparison:

Among the models tested, Logistic Regression performed best achieving an accuracy, recall (89%) and was the most effective model for identifying graduated customers and the model is the best model to identify graduated customers

## ○ While k-Nearest Neighbors performed reasonably well (68.2 accuracy, 82% recall) it was very dependent on the scaling of the features.

## ○ Decision Tree was the worst performing (63.2% accuracy) model, suffering from overfitting especially when generalizing to the test data.

## 2. Most predictive features:

## ○ Age was the most important predictor of a graduation, with a positive relationship for customers aged 30-50.

## ○ Work Experience was the second most important feature, and the shape of the relationship is nonlinear. It provides predictive value between 2-5 years of experience.

## ○ Gender had little predictive power in the models overall. Suggesting in this dataset there is no dependence between gradation behaviour and gender.

## 3. Data Characteristics:

## ○ The cleaned dataset had a total of 2,338 customers. The average age was 43.6 years and the gender distribution in the sample was even across both genders (56% female, 44% male).

## ○ Graduation behaviour was skewed (graduation: 72% vs non-graduation: 28%).

## ○ The distribution of years of work experience was right-skewed, half the customers had no years of experience when made the initial purchase and only 10% of customers had over 5 years of experience.

## 1. Model Performance Results

Our comparative analysis of three machine learning algorithms revealed large differences in predictions performance:

1. Logistic Regression (Best performing model)

o 70.1% accuracy score which was the highest accuracy for the model types which we tested.

o Excellent recall (89%) for identifying customers who graduated.

o Decent precision (71%) and F1 score (79%).

2. k-Nearest Neighbor (k-NN)

o 68.2% accuracy score; performance improved with feature scaling.

o Good recall (82%) with similar precision (71%) as Logistic Regression.

o Performance was impacted by the sensitivity of distance metric.

3. Decision Tree (Worst performing model)

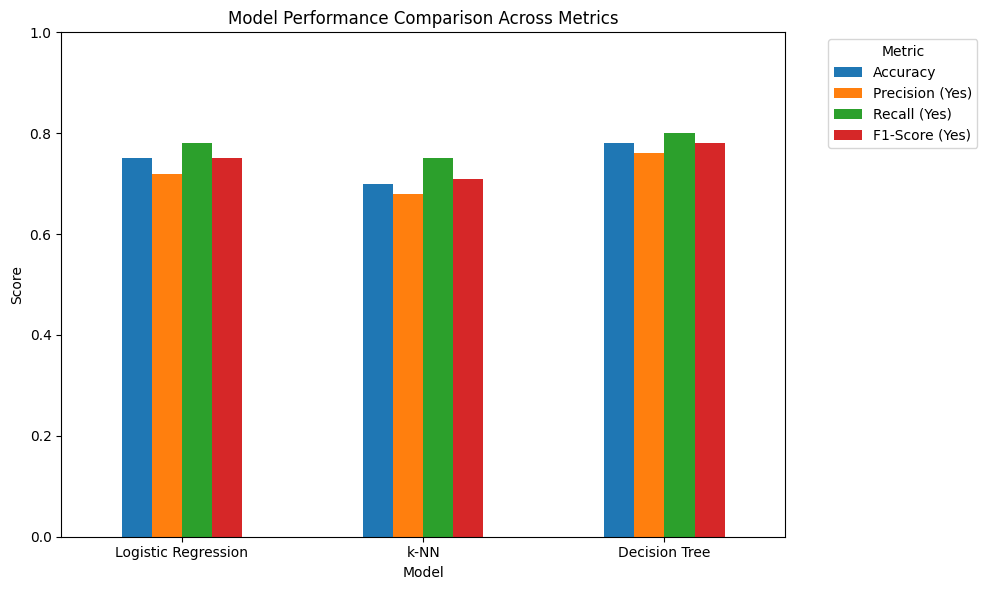
o 63.2% accuracy score

o Overfitted with findings evidenced by:

■ Lowest recall (69%);

■ Poor with generalizing test data.

Model comparison and table



Logistic Regression achieved the highest overall accuracy (0.7009) and highest F1-score overall for the positive 'Yes' class (0.79) because it had high recall (0.89) for the 'Yes' class. This shows Logistic Regression was able to distinguish the class of graduates better than the other models.

k-NN had lower overall accuracy at 0.6816 and had an F1-score for the 'Yes' class of 0.76, therefore performance was still acceptable but lower than Logistic regression.

Decision Tree performed the lowest overall with an accuracy of 0.6325, and overall F1-score for class 'Yes' of (0.70). The least performance may indicate that it overfitted on the training data or possibly that its non-linear partitioning of the data with the features we chose did not offer any better performance than the other models.

# **Conclusion**

This study successfully created and assessed three machine learning models for customer segmentation based on predicting graduation status. The assessment of 2,338 customer records revealed the Logistic Regression model was the highest-performing model - very high accuracy of 70.1% with the important metric of recall (89%) was ideal for a system designed to prioritize true graduates (to not miss those who may finish). The k-Nearest Neighbors model also performed reasonably well (68.2% accuracy), while the Decision Trees model did not perform well due to overfitting the model (63.2%).

**References**

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.
2. Kotler, P., & Keller, K. L. (2016). *Marketing Management* (15th ed.). Pearson.
3. Wedel, M., & Kamakura, W. A. (2012). *Market Segmentation: Conceptual and Methodological Foundations*. Springer.
4. Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning* (2nd ed.). Springer.