Final Project Proposal

•••

Joseph Nelson Farrell & Michael Massone

DS 5220 Supervised Machine Learning

Northeastern University

Professor Steven Morin, PhD

Data Overview

Dataset : Customer Segmentation

Origins : https://www.kaggle.com/datasets/vetrirah/customer?select=Train.csv

Size: (8068, 11)

- Instances: 8068
- Attributes: 11

Description :

- The dataset contains information about customers of an automobile company segmented into 4 classes (target).

The Target Variable

The column name of the target variable is segmentation.

- There are **0** missing values in the target variable column.
- As a result, **0** observations (rows) are dropped.
- Dataset size: (8068, 11)
- The datatype of the target is a string with 4 values:
  - Value: A ~ Proportion: 0.244422
  - Value: B ~ Proportion: 0.230293
  - Value: C ~ Proportion: 0.244175
  - Value: D ~ Proportion: 0.281111
- The categories appear relatively balanced.
- These values correspond to customer segments used in outreach.

The Design Matrix

ID	Unique identifier	Int
Gender	Male, female	String
Ever_Married	Yes, no	String
Age	Customer age	Int
Graduated	Yes, no	String
Profession	Customers profession	String
Work_Experience	In years	Float
Spending_Score	Low, average, high	String
Family_Size	Including customer	Float
Var_1	Anonymized Segment	String

Train/Test Split Outcome

Test Set: 20%

File Name: test_df.csv

Test Set Size: (1614, 11)

Target Proportions:

- Value: A ~ Proportion: 0.244114.
- Value: B ~ Proportion: 0.230483.
- Value: C ~ Proportion: 0.244114.
- Value: D ~ Proportion: 0.281289.

Train Set: 80%

File Name: train_df.csv

Train Set Size: (6454, 11)

Target Proportions:

- Value: A ~ Proportion: 0.244500.
- Value: B ~ Proportion: 0.230245.
- Value: C ~ Proportion: 0.244190.
- Value: D ~ Proportion: 0.281066.

The split has preserved the distribution of the target variable across the train/test sets.

Objective

The company has segmented its existing customers into four distinct classes based on the marketing strategy that works best for them. These strategies include text messaging (A), physical mailers (B), weekly emails (C), and quarterly emails(D).

The marketing team has determined that these strategies have proven to be effective for their corresponding customer segments; however, they tend to have the opposite effect on customers outside those groups, potentially driving them away. For example, customers who have preferred quarterly emails, have been more likely to label promotional emails as spam when the frequency of emails is increased, while those who have preferred weekly emails are less likely to convert when email frequency decreases.

The company aims to generalize this segmentation to new customers by using available information from prospective customers who have signed up for promotional materials. Their goal is to develop a predictive model that can accurately assign new customers to the appropriate marketing strategy class, ensuring that each customer receives the communication style most likely to resonate with them, while avoiding strategies that might alienate them. This classification model must carefully consider customer features to predict which marketing strategy will yield the best results, helping the company optimize its outreach and improve overall customer engagement and retention.

Class: A

Text Messages

Class: B

Physical Mailers

Class: C

Weekly Emails

Class: D

Quarterly Emails

Performance Measures

The goal of this project will be to develop a model that can generalize to unseen data, i.e., new customers. As such, various classification measures will used to evaluate model performance:

- Accuracy: The ratio of correct predictions to total number of predictions.
    - Balanced dataset, where neither false positive and false negatives is significantly worse.
- Precision: The ratio of true positives prediction to total positives predictions.
    - Imbalanced dataset, when it's very important to avoid false positives.
- Recall: The ratio of true positives to actual positives.
    - Imbalanced dataset, when false negatives are more expensive than false positives.
- F1: The harmonic mean of precision and recall.
    - Imbalanced dataset, when both precision and recall are equally important.
- AUROC: The area under the Receiver Operating Characteristic curve (TP rate vs FP rate).
    - Shows overall performance of the of the model across different classification boundaries.

Due to the high cost associated with mislabeling a new customer—not only in terms of financial expense but also the potential of losing a prospective customer or damaging the company's brand—we will prioritize precision as the primary metric for evaluating the success of our model. By focusing on precision, we aim to minimize false positives, ensuring that when a particular marketing strategy is chosen, it is highly likely to be the correct one, reducing the risk of negatively impacting customer relationships.

Confusion Matrix

The confusion matrix is an effective tool for visualizing the performance of a classifier. In binary classification, it clearly displays the counts of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), which allows for a straightforward evaluation of metrics such as precision, recall, and accuracy.

For multiclass classification, the confusion matrix becomes more complex. While TPs for each class are shown explicitly, FPs and FNs are distributed across multiple other classes, making it harder to interpret class-specific performance.

To address this, a one-vs-all confusion matrix can be constructed for each class. This technique treats one class as positive and all other classes as negative, enabling the calculation of precision, recall, and other metrics for each class individually.

Expected

Predicted

Class	A	B	C	D
A	True Positive	Error	Error	Error
B	Error	True Positive	Error	Error
C	Error	Error	True Positive	Error
D	Error	Error	Error	True Positive

Multi-Class Confusion Matrix

Expected

Predicted

Class	A	~A
A	True Positive	False Positive
~A	False Negative	True Negative

One vs All Confusion Matrix

$$\text{Precision}_A = \frac{\text{TP}_A}{\text{TP}_A+\text{FP}_A}$$

Quantifying Success/Failure

- As discussed, precision measures the ratio of true positive predictions to total positive predictions, which emphasizes false positives (a case we would like avoid). Mathematically precision is denoted:
  - $\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}$
- Precision can vary between the various classes, as such:
  - *Precision will be computed for every class individually.*
- This model will then employ the weighted precision across the classes to determine overall success.
- $\text{Weighted Precision} = \frac{1}{N} \sum_{i=1}^{N} w_i \cdot \text{Precision}_i$
  - Where w_i is the proportion of instances of class *i* and *N* is the number of classes.
- The model will be deemed successful and put in production if weighted precision >= 0.80.
- This has been established the as the minimum threshold where the reachout campaigns remain effective.

Assumptions

Correct Labeling:

We are assuming that the customer data provided by the company has been accurately labeled, meaning that each customer is correctly assigned to their respective segment. This assumption is crucial because the quality of the model depends on the accuracy of the labeled data. If the labels are incorrect, the model may learn incorrect patterns and make poor predictions.

Domain Knowledge:

The company has used its internal domain knowledge to annotate and classify customers. This means they have likely relied on their business expertise and understanding of their customers' behaviors to create meaningful customer segments. These segments are assumed to be actionable, meaning that they correspond to different marketing or business strategies, such as different types of outreach campaigns.

Assumptions (continued)

Generalizability:

The model being developed will be based on the assumption that the company's previous classification strategy will apply to new customers in the future. This means that the patterns the model learns from the historical data will generalize well to unseen customers. The goal is to use these segments to predict which type of outreach will work best for new customers, driving more targeted marketing strategies and improving customer engagement.

Feature Relevance:

It is assumed that the features provided in the dataset are relevant and sufficient for customer classification. This means that the existing features are capable of differentiating customers into meaningful segments. No significant or important feature is missing from the dataset, and all the required data points that could impact customer segmentation have been accurately captured and included.

Acknowledgments

- Aurélien Géron - Hands on Machine Learning with Scikit-Learn, Keras, and Tensorflow
- Google Machine Learning Education - Classification: Accuracy, recall, precision, and related metrics
- Steve Morin, PhD