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**

Test 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.

Business Objective

Our *client* plans to enter a new market and has concluded that customer behavior in the new market is similar to their existing market. The dataset contains information about current customers including their segmentation. This segmentation is used for targeted outreach and communication and has been remarkably effective.

The company would like to use the same strategy in the new market and as such would like have a automated method to segment new customers.

The goal of this project will be to development a machine learning solution to classify new customers within the 4 segments.

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.

Performance Objectives

Primary Metric - Accuracy

Since our the distribution of classes in our target vector are well balanced (approximately 25% per class) we can use accuracy as a measure of our model's performance. Also, there is no preference for false negative versus false positives in this multi-classification problem. All customers will be contacted regardless of class and no outreach strategy is significantly more costly than the other. If some customers are misclassified, the sales team can manually reclassify as more information about the customer becomes available. In the case of class imbalance we could rely more on recall or precision - depending the the client's needs.

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