Udacity Data Science Capstone Proposal

Creating a Customer Segmentation Report for Arvato Financial Services

Domain Background:

Arvato Group, an internationally acclaimed service conglomerate, operates across the global landscape, crafting tailor-made solutions for a wide spectrum of business operations in over 40 countries[1]. The group's extensive range of services encompasses an array of vital areas, spanning from supply chain solutions through Arvato to financial services offered by Riverty, and comprehensive IT solutions delivered by Arvato Systems. These services cater to renowned companies spanning diverse industries, such as telecommunications, energy, finance, e-commerce, and IT. Additionally, Arvato is a wholly owned entity of Bertelsmann. This vast scope of expertise underpins Arvato's commitment to delivering tailored solutions for businesses across the globe.

Problem Statement:

The central challenge is to enhance the efficiency of customer acquisition for the German mail-order company. This entails the development of a predictive model capable of discerning prospective customers through their demographic profiles. Furthermore, the task involves assessing the likelihood of individuals with particular demographic attributes transitioning into future customers. The ultimate aim is to address this challenge with precision and unwavering confidence.

Datasets and Inputs:

The project utilizes four datasets. First, "Udacity_AZDIAS_052018.csv" contains demographics data for the general German population, with 891,211 persons and 366 features. Second, "Udacity_CUSTOMERS_052018.csv" comprises demographics data for customers of the mail-order company, including 191,652 persons and 369 features. Third, "Udacity_MAILOUT_052018_TRAIN.csv" provides data on individuals targeted by a marketing campaign, encompassing 42,982 persons and 367 features. Finally, "Udacity_MAILOUT_052018_TEST.csv" contains demographics data for individuals targeted by the same marketing campaign, with 42,833 persons and 366 features.

Solution Statement:

The proposed solution involves a multi-step approach. Initially, data exploration and preprocessing are performed, including cleaning and imputing missing data. Feature engineering is executed to reduce dimensionality using Principal Component Analysis (PCA). Subsequently, unsupervised learning, specifically K-Means clustering, is applied to segment customers based on PCA attributes. In the final step, supervised learning techniques, including Logistic Regression, Random Forest Classifier, and Support Vector Classifier are employed to predict potential customers from the German population dataset by considering customer segments.

Benchmark Model:

Using Random Forest Classifier (RFC) as the benchmark model is a sensible choice in the context of our problem statement. RFC is a robust and widely used machine learning algorithm for binary classification tasks. It provides an established reference point for assessing the performance of alternative algorithms. The benchmark model's measurable metrics, including accuracy, precision, recall, and F-score, allow for a thorough comparison. By quantifying the RFC's ability to predict new customers based on demographic data, we can objectively evaluate its performance.

Random Forest Classifier is particularly relevant to the problem domain as it can handle complex relationships within the data and provide insights into feature importance. By measuring its predictive accuracy and robustness, we can

determine its suitability as a benchmark model for customer acquisition prediction. This benchmark model will help us gauge the effectiveness of our proposed solution and guide us in optimizing the accuracy of our final predictive model.

Evaluation Metrics:

In evaluating the performance of the benchmark model, which is the Random Forest Classifier (RFC), and the solution model, we will primarily focus on two critical evaluation metrics: accuracy and F1-score. Accuracy measures the proportion of correctly predicted instances in a classification problem, providing an overall assessment of the model's correctness. It can be calculated as (True Positives + True Negatives) divided by the total number of instances. The F1-score, on the other hand, balances precision and recall, taking into account false positives and false negatives. It is calculated as 2 * (Precision * Recall) divided by (Precision + Recall), where Precision is the number of true positives divided by true positives and false positives, and Recall is the number of true positives divided by true positives and false negatives. These metrics offer a clear and quantifiable way to assess the effectiveness of the Random Forest Classifier benchmark model and the solution model in predicting new customers for the German mail-order company.

Project Design:

1. Data Collection and Understanding:
Gather and explore the provided datasets, including the general population demographics and customer data.
Perform initial data analysis to understand the structure and characteristics of the data.
Identify missing values, outliers, and data quality issues.

2. Data Preprocessing:
Handle missing data by either imputing values or removing features with a significant number of missing entries.
Assess and address outliers and anomalies that might impact model performance.
Convert non-numeric features into numeric formats, such as one-hot encoding or label encoding.

3. Data Reduction and Feature Engineering:
Apply dimensionality reduction techniques like Principal Component Analysis (PCA) to reduce the number of features while preserving essential information.
Perform feature selection and engineer new features that could improve predictive power.

4. Data Splitting:
Split the dataset into training and testing sets for model evaluation.

5. Benchmark Model (Random Forest Classifier):
Implement a Random Forest Classifier as the benchmark model due to its suitability for classification tasks and capacity to handle a large number of features.
Train the model on the training dataset and evaluate it using metrics like accuracy and F1-score.

6. Alternative Models (e.g., Logistic Regression, SVC):
Explore alternative supervised learning algorithms like Logistic Regression and SVC to identify the best-performing model.
Compare their performance with the benchmark Random Forest Classifier.

7. Model Tuning and Optimization:
Fine-tune the selected model by adjusting hyperparameters, such as the number of trees in the Random Forest or regularization strength in Logistic Regression.

8. Evaluation Metrics:
Assess model performance using accuracy and F1-score, which quantify the accuracy and balance between precision and recall.

Compare benchmark and alternative models based on these metrics.

9. Solution Deployment:
Deploy the best-performing model to predict potential new customers based on demographic data.

10. Documentation and Reporting:
Document the entire process, including data preprocessing, model selection, and evaluation metrics.
Create a report summarizing the findings, methodologies, and results to communicate with stakeholders.
This workflow outlines the step-by-step process for tackling the problem of predicting potential customers for the German mail-order company, aligning with the domain background and the project's goals. The iterative nature of model selection, tuning, and evaluation ensures that the solution meets the requirements of high accuracy and confidence in identifying future customers.

[1]https://www.bertelsmann.com/divisions/arvato/#st-1