Arvato Financial Solutions Project Proposal

Domain Background:

Arvato is a services company that provides financial services, Information Technology (IT) services and Supply Chain Management (SCM) solutions for business customers on a global scale. It develops and implements innovative solutions with a focus on automation and data analytics. Arvato provides its customers with valuable insights from data in order to make business decisions. Customer centric marketing is one of the growing fields. Identifying patterns and customer behavior from the data is providing valuable insights for the companies operating in customer centric marketing. Data Science and Machine Learning are immensely used now a days to fulfil business goals and to satisfy customers.

In this project, Arvato is helping a Mail-order company, which sells organic products in Germany, to understand its customers segments in order to identify next probable customers. The existing customer data and the demographic data of population in Germany are to be studied to understand different customer segments, and then building a system to make predictions on the probability of whether a person will be a customer or not.

Problem Statement:

The problem statement is how can a mail order company acquire new possible customers in an efficient way of marketing given the demographic data of people in Germany.

Datasets & Inputs:

- Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 instances (rows) x 366 features (columns).
- Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 instances (rows) x 369 features (columns).
- Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 instances (rows) x 367 (columns).
- Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 instances (rows) x 366 (columns).

Additionally, 2 metadata files have been provided to give attribute information:

- DIAS Information Levels - Attributes 2017.xlsx: top-level list of attributes and descriptions, organized by informational category
- DIAS Attributes - Values 2017.xlsx: detailed mapping of data values for each feature in alphabetical order

Solution Statement:

First off, my solution would be done using a single AWS Jupyter notebook instance in training and testing the models chosen for this project.

In the first half of the project, the task is to identify any customer segments present in the provided dataset and match these segments with the segments of population present in the general population dataset using unsupervised learning methods.

- First, the dataset will be explored to examine if there are any missing values or mis recorded values in the data and fix them. Also, any categorical features need to be re encoded into numerical features with the help of Label encoders.
- Second, we identify the minimum number of features that would be sufficient to explain the dataset. Since there are 366 features that represent a single person and not all the features will be important in forming the segments. A dimensionality reduction technique can be used here to identify minimum number of features which explain the variation in the dataset.
- Third, we need to divide the customers into different segments based on the selected features with the help of unsupervised learning algorithm. K-means clustering is a good choice for this step as this algorithm tries to assign each data point to a cluster based on the distance from a cluster center.

In the second half of the project, the task is to predict whether the mail order company can acquire a customer through a mail-out campaign or not.

We use a supervised learning algorithm which will be trained and evaluated on the pre-processed training data and then used to make predictions on the test data provided.

Proposed algorithms for supervised learning.

- Logistic Regression – A simple binary classification algorithm
- Decision Tree Classifier – A tree-based algorithm which uses rule-based approach for classification.
- Random Forest Classifier and XG Boost Classifier can also be used since they are derived from the decision trees algorithm

A grid search algorithm can be used for hyperparameters tuning job.

Benchmark Model:

The baseline model fit for this project would be a simple logistic regression model as we can compare its performance with other complicated models, so that we can decide if proceed with it.

Evaluation Metrics:

For the customer segmentation part of this project, we're using the squared mean error value to know how accurate our segmentation is.

As for the supervised learning part, we can evaluate our model using the confusion matrix, the accuracy of our predictions as well as using the AUROC metric which is commonly used in data science competitions like Kaggle.

Project Design:

1. Explorative Data analysis: The data needs to be cleaned, explored for insights, and visualized by its main features to further understand any noticeable patterns in the data.

2. Feature Engineering: Understanding explained variance of features in the dataset and determining the required number of features that can amount for maximum variance in the dataset using a dimensionality reduction technique like PCA. Determining correlations between features will also help in identifying redundant features.

3. Modelling: First step is to identify the customer segments using unsupervised learning algorithms. A KMeans Clustering algorithm will be used to segment the data into desired number of clusters. In the second step, different supervised algorithms will be trained and evaluated in the context of predicting whether a person will be our next customer or not. Algorithms like Logistic Regression, Decision Tree, Random Forests and Gradient Boosted Trees will be used to make predictions and will be evaluated. The previously proposed evaluation metrics will be used to determine the best model in this step.

4. Model Tuning: After evaluating different algorithm's performance on the evaluation data. The algorithm which has a good score will be selected and tuned to improve the performance.

5. Predictions on Test data: Finally, the best model will be used to make predictions on the test data

References:

[1] Arvato-Bertelsmann, "Arvato," Bertelsmann, [Online]. Available: https://www.bertelsmann.com/divisions/arvato/#st-1. [Accessed October 2022].

[2] Bertelsmann, "Company," Bertelsmann, [Online]. Available: https://www.bertelsmann.com/company/. [Accessed October 2022].

[3] AUC (Area under the ROC Curve) [Online]. Available: https://machine-learning.paperspace.com/wiki/auc-area-under-the-roc-curve. [Accessed November 2022]