Capstone Proposal for Machine Learning Engineer Nanodegree Program

Domain Background

Acquiring customers more efficiently is an essential part of running a successful company. One of the ways to improve its efficiency is through targeted marketing. For the long time, the process of choosing the most likely customers has largely been done using human intuitions and experiences. In the recent times, with the advancements in data collection techniques, it is now possible to use large amount of data to make more efficient and effective business decisions instead of solely relying on gut feelings. The purpose of this project is to use the given dataset to make the targeted marketing process more efficient using various data analysis techniques.

Problem Statement

From a business perspective, the underlying question is: "How can a mail order company acquire new clients more efficiently?"

The Arvato Financial Solution has provided the data of its existing customers along with the data of population of Germany and has tasked us to identify and analyze the traits that its customers share with the population in such a way that one can identify which individuals in population will be most likely to become customers in future. The goal is to identify such segments of the population which can then be used for targeted marketing campaigns to achieve a higher expected rate of return.

Datasets and Input

The dataset is provided by Arvato Financial Solutions and includes the following files:

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

Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood.

The "CUSTOMERS" file contains three extra columns ('CUSTOMER_GROUP', 'ONLINE_PURCHASE', and 'PRODUCT_GROUP'), which provide broad information about the customers depicted in the file. The original "MAILOUT" file included one additional column, "RESPONSE", which indicated whether or not each recipient became a customer of the company. For the "TRAIN" subset, this column has been retained, but in the "TEST" subset it has been removed.

Solution Statement

In the first part of this project, the unsupervised learning model will be used to tell which segments of population will be most likely to become customers in future. For the supervised learning part, our model will be trained using the training dataset and then will be accessed using the test dataset. Same data pre-processing techniques will be used that were applied in first part. Different machine learning models and techniques will be employed to achieve the best result.

Benchmark Model

The benchmark model will be XGBoost model. The reason I specifically chose this model is because it is one of the top performing algorithms across Kaggle competitions and handles missing values really well without the need for imputation.

Evaluation Metrics

Since the data is highly imbalanced, the accuracy score will not determine good results. Therefore, the Area under the ROC curve (AUC) will be used as a metric as it ranks a randomly chosen positive instance higher than a randomly chosen negative example and is also used in Kaggle competitions.

Project Design

The Bertelsmann/Arvato Project consists of four parts which are broken down as follows.

Data Pre-Processing

“Garbage in, garbage out” is a phrase of particular importance in machine learning world as the quality of inputted data greatly determines the quality of the output. In other words, the quality of machine learning model heavily relies on the data pre-processing techniques used before training. In this project, I will use various data exploration techniques to identify missing or inconsistent data and then will later decide which columns or rows to drop and which missing values require imputation.

Customer Segmentation with Unsupervised Learning

Customer segmentation part will require two steps. In first step, I will use Principle Component Analysis (PCA) to reduce dimensions to the appropriate number based on elbow method then K-Means

clustering algorithm will be used along with data exploration to identify which parts of the general population will be more likely to become customers in future, and which parts will be less so.

Supervised Learning Model

In this part, I will use various machine learning algorithms and match their performance against each other to decide which algorithm best predicts the individuals that are most likely to become customers for the company. Then I will tune the chosen model's hyperparameters to maximize its performance.

Kaggle Competition

Finally, the trained model will be submitted in Kaggle competition where it will give predictions on their public test data and the results will be compared against model of other competitors. ROC AUC will be used as a metric for choosing the best model, and for scoring in Kaggle competition.