Capstone Project Proposal.

Project Name: Create a Customer Segmentation Report for Arvato Financial Solutions

1. Domain Background

Arvato is a financial solutions company with solutions such as credit management, fraud management, debt collection management, payment solutions, and other financial services.

Like any other company, Arvato uses analytics to help them understand their customers better and also find new customers based on the existing customers they have.

Prediction of customer churn has been done in papers such as Xia, G. E., & Jin, W. D. (2008). Model of customer churn prediction on support vector machine. Systems Engineering-Theory & Practice, 28(1), 71-77. and Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. Expert Systems with Applications, 39(1), 1414-1425., where they used SVM to predict customer churn in telecommunications and later an improvement of the same through use of K-means clustering, SVM to improve churn prediction.

Since we will be predicting customer conversion, the model will take a similar direction as the customer churn prediction.

2. Problem Statement

In order to serve their clients better, find new customers, and also create better products, Arvato needs to create customer segments from its existing clientele.

The problem is a classification problem, that will make use of classification algorithms to identify customers from the general demographic population.

This project will therefore help Arvato have a better view of their different clientele types, to help them grow their business as well as leverage proper revenue generation through marketing and advertising to target customers.

3. Datasets and Inputs

In this project, I will analyze demographic data characteristics for customers of a mail-order sales company in Germany, comparing it against demographics information for the general population. I will use unsupervised learning techniques to perform customer segmentation, identifying the parts of the population that best describe the core customer base of the company.

I will be utilizing AWS Sagemaker to carry out Explorative Data Analysis and use Supervised/Unsupervised learning to come up with appropriate customer segments as well as also a model to help predict a customers' segment based on certain characteristics.

Specifically, I will a dataset that has target customer groups from a mail campaign. I will use this dataset to create a model to predict if a customer is likely to respond to the campaign or not.

Then, I will apply the model on a third dataset with demographic information for targets of a marketing campaign for the company, and use it to predict which individuals are most likely to convert into becoming customers for the company.

The data that you will use has been provided by Bertelsmann Arvato Analytics.

I will also submit my solution to a Kaggle competition for ranking on Kaggle.

4. **Solution Statement**

The goal will be to create a model that will help Arvato determine who is likely to become a customer through their marketing campaigns.

5. **Benchmark Model**

The benchmark model will be a model that will be able to predict a customer's conversion accurately.

I will be using KNN as the benchmark model for this project.

6. **Evaluation Metrics**

Since there are no 'definite metrics' to check how well an unsupervised algorithm performs, I will use the silhouette/elbow method to determine the number of clusters accurately and then also visualize the results to ensure that the model has identified clusters distinctively.

As for the supervised model, I will use the confusion matrix to assess how well my classifier is able to accurately classify random people in various categories.

I will also use intuition, based on domain knowledge on the same.

7. **Project Design**

I will begin the project by:

a. Loading dependencies and libraries on sagemaker.
b. Loading the data
c. Data preprocessing (data cleaning/ missing values imputation, data normalization so that the dataset is balanced)
d. Exploring the data to understand the various underlying characteristics, statistical analysis, graphical analysis, and visualizations.
e. Creating a model using supervised models. In particular, I will use LightGBM and XGBoost for classification. Since I will be using a couple of models, I intend to eventually have a customized ensemble model for the project.
f. Visualize the cluster groups
g. Using the model to create a prediction system to help determine how likely a target group of people are likely to convert to an Arvato customer.
h. Submit my solution to Kaggle for ranking.
i. Write my findings through a blog post on medium.