ML ND CAPSTONE PROJECT

Customer Segmentation Report for Arvato Financial Services

Amar Singh | ML ND | 13-07-2020

Domain background:

In this world making the business grow faster and better is the toughest job than starting the new business. Finding or you can say targeting the specific customers that can use you product is very important for the company as it reduces the marketing campaign cost and increase its profit.
Arvato Financial Solution does the same thing it provides solution to different company so that they can grow their business quickly. They provide financial solution also.
Arvato company has teamed up with the udacity team to provide the project for ML nanodegree in which we have to find customer segmentation which will respond to the mail beig sent by the company. Bertelsmann was founded in 1835 and provide guidance for business development.

Problem Statement

Statement:

Making A Machine learning model to predict the customer segment of the population of Germany to which the mail must be sent in order to target the maximum response from the customer using selective mails thereby decreasing the cost in campaign and increasing profit in the business. Also that services are provided to the customer segment who really needs them.

Evaluation Metrics

Kmeans should be used to found the no. of cluster to segment the population after analyzing the elbow graph

ROC_AUC will be used by me in the supervised section of the Arvato project this is because the dataset provide is imbalanced and therefore left us with ROC to get us the best accuracy scoring.

Dataset Used:

Source of data set: Provided by Arvato company for completing the project Identifying the customer segmentation

Note: This data set should be used for this project and cannot be available to any other individual

Below are the Data sets given in this project:

AZDIAS Data set: General population of Germany .(891,211 customers x 366 features)

CUSTOMER Data set: Arvato company mailing customers .(191,652 customers x 366 features)

Mailout_train data set: Training data set that were send to the population of germany(42,982 customers x 367 features)

Mailout_test data set: Testing data set that were send to the population of germany((42,982 customers x 366 features))

Attributes 2017.xlsx : In Depth information of the attributes.

Values 2017.xlsx: In depth explanation of the feature attribute

Azdias and customer data set should be used for unsupervised learning of this project while mailout_train data for making model for supervised section and then mailout_test data should be used to test your model finally.

Solution Statement

In order to solve this project following step should be done:

-We’ll use unsupervised learning techniques to perform customer segmentation in order to find the section of the population which must be targeted by company.

- Then, by applying supervised technique on the company mailout train dataset in order to get the model ready for making prediction about the likeliness of the customer to respond to mail campaign.

- Then we’ll test the model made in supervised section to predict the probability of the customer to respond to the mail.

Benchmark Model

Different Classifier will be used like Adaboost Classifier, Gradientboost, logisticregression and Randomforest classifier for modelling of the given data.

GradientBoostClassifier were chosen to be the benchmark model for this type of data set as it has a great historical relevance in Ml industry.

Project Design

Some common steps to be followed throughout the project are listed below:

- **Preprocessing**: Data given should be explored of its null values using graphs and a cleaning function to clean all the data set given in the project .
- **Reducing Dimensionality** : PCA should be used to reduce the dimensionality of the data set from very large features .
- **K Mean Cluster**: Kmeans should be used to further found the no. of cluster to segment the population after analysing the elbow graph.
- **Supervised section** : Apply supervised technique on the company mailout train dataset in order to get the model ready for making prediction about the likeliness of the customer to respond to mail campaign
- **Important features of the data set**
- **Kaggle submisstion :** In order to participate in this competion we have to submit a csv file in which there is user and the corresponding probablity associated with it.

****