Machine Learning Engineer Nano Degree

Capstone Proposal

Customer Segmentation – Arvato Financial Solution

Gavle Namdev

Domain Background

Arvato financial solutions is a german company. It provides services of invocing, accounting and payment, credit management and debt collection manangement. Arvato uses machine learning techniques to extract the user behavior and tells who is likely going to be a potential customers. Then mail order companies targets those users with advertisements. Like this, mail order companies gets insights from the arvato which results in better financial performance.

Problem Statement

The major job of arvato financial service is to help their clients identify the customers who are likely going to be their potential customer given demographic data. Problem statement is "Given the demographic data, How can a mail order company acquire new customers more efficiently?"

Datasets and Inputs

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).

DIAS Information Levels - Attributes 2017.xlsx : is a top-level list of attributes and descriptions, organized by informational category

DIAS Attributes - Values 2017.xlsx: is a detailed mapping of data values for each feature in alphabetical order.

Solution Statement

For customer segmentation, we will use unsupervised techniques to find the potential customers segments from the population by providing demographic features to the unsupervised model.

After finding out segment of most likely to be customers from the population, we build a machine learning model that predicts whether or not each individual will respond to the campaign and most likely are gonna be new clients of mail order company.

Benchmark Model

I will consider logistic regression as benchmark model as it is the simple classification model. If this model doest not give good results, may be other tree models or boosting models will be taken as benchmark model.

Evaluation Metrics

For unsupervised problem, number of clusters is the hyperparameter for Kmeans algorithms. I will be using squared error metric

For any classification problem, I think accuracy and F1 score(precision and recall) metrics are enough to evalute. I will be adding area under the receiver operating curve(AUROC) as this metric is being asked in kaggle competition.

Project Design

Identify the customer segment using unsupervised learning techniques

- Data Cleaning and Visualisation: find missing values and fix them with imputation techniques and categorical values are encoded into numerical values. Finally normalize values using standard scalers
- Feature Engineering: Based on the visualisation and analysis, new features will be created
- Feature Selection: identify the minimum number of features that would be sufficient to explain dataset. Dimentionality reduction techniques like Principal Component Analysis can be used to complete this task
- Modelling: Apply K-means (unsupervised algorithm) to segment general population and customers into different segments based on the selected features
- Model Tuning: GridSearch algorithm is used to tune the value of number of clusters

Predict whether the mail order company can acquire a customer

- Prepocessing steps are carried out as explained in the above method
- We will train the model using training data and evaluate using validation split in training step. Proposed supervised algorithm are Logistic Regression, SVM, Decision Tree Classifier, Random Forest and XGBoost.
- Finally we test and evalute trained model performance using test data
- To tune hyperparameters, GridSearch algorithm will be used

References

1. Arvato Financial Solutions: https://www.bertelsmann.com/divisions/arvato/#st-1
2. Kaggle Competition: **https://www.kaggle.com/c/udacity-arvato-identify-customers**
3. PCA: **https://en.wikipedia.org/wiki/Principal_component_analysis**
4. XGBoost: **https://xgboost.readthedocs.io/en/latest/**
5. Udacity course materials
6. Arvato Financial Solutions, https://finance.arvato.com/en/about-us/