Machine Learning Engineer Nano Degree

Capstone Proposal

Customer Segmentation – Arvato Financial Solution

Gavle Namdev

Domain Background

Arvato is an internationally active services company that develops and implements innovative solutions for business customers from around the world. These include SCM solutions, financial services and IT services, which are continuously developed with a focus on innovations in automation and data/analytics. Globally renowned companies from a wide variety of industries – from telecommunications providers and energy providers to banks and insurance companies, e-commerce, IT and Internet providers – rely on Arvato's portfolio of solutions.[1]

Arvato uses Machine Learning techniques to extract the behavior patterns of customers and provides valueable insights from the patterns to Mail-order companies in order to make business decision. These insights are very useful for customer centric marketing.

Problem Statement

Arvato helps its clients to able to identify customer's payment bahavier, predict payment behavier, recommend credit score and calculate credit score by analyzing demographice attributes.

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

We will solve the problem by dividing into parts

First approach is, unsupervised learning techniques are used to describe the relationship between the demographics of the company's existing customers and the general population of Germany in order to create customer segment.

And second approach is, 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

Benchmark model would be a logistic regression model since it is easy to train and test within less amount of time. The performance of this model will be considered as a baseline for further steps, where different algorithms can be used to compare the performance with this benchmark to decide whether to proceed with an algorithm or not. If this model proves deficient, may be other appropriate model will be selected as benchmark model.

Evaluation Metrics

For customer segmentation, K-means algorithm used. Number of clusters will be hyper parameter and it will be selected based on the squared error i.e the distance between all the clusters

For supervised algorithim, training data will be split into train and validation set. Model is trained on training split and will be evaluated on validation split. Evaluation metrics for classification can be used accuracy, F1 score, recall, precision, area under the receiver operating curve(AUROC)

Project Design

Identify the segment using unsupervised learning techniques

- Data Cleaning and Visualisation: find missing values and fix them and categorical values are encoded into numerical values. Finally normalize values using standard scalers
- 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, 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