Arvato Bertelsmann Customer Analysis

Udacity Machine Learning Engineer Nanodegree Capstone Project

By Jiaxu Luo
2019-08-04

I. Definition

Project Overview

Facebook and Twitter, along with many other “Digital 100” companies (*Business Insider*, 2012), have high valuations due primarily to data assets they are committed to capturing or creating. Increasingly, marketers have to organize and understand data-driven campaigns. Data can be used to build models to discriminate between those to whom we should and should not advertise. Instead of reaching out to all people and targeting them with a marketing campaign, we want to be reaching out to the people we identified as becoming the most likely new customers, then do target advertising. The efficiency in the customer acquisition process is the key to business success.

In the project, a mail-order sales company in Germany is interested in identifying segments of the general population to target with their marketing in order to grow their customer base. Additionally, the company want to identify which individuals are most likely to respond to the marketing campaign and become customers of the it.

Data files

(1) Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891211 persons (rows) x 366 features (columns).

(2) Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191652 persons (rows) x 369 features (columns).

(3) Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42982 persons (rows) x 367 (columns).

(4) Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42833 persons (rows) x 366 (columns).

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

(6) DIAS Information Levels - Attributes 2017.xlsx: a top-level list of attributes and descriptions, organized by informational category (Person, Household, Building, etc.).

A new file “documented_features.csv” was manually created containing the attribute, information level, data type, missing or unknown values from DIAS Attributes - Values 2017.xlsx (5).

Problem Statement

In first part of this project, unsupervised learning method (K-means Clustering) is applied to identify clusters or segments from general population who best match the company’s customer base by creating clusters of customer and general population, and then identify the difference. For the second part, supervised learning techniques (Gradient Boosting Trees for binary classification) to train a model using training data and apply to the test data to predict whether an individual would respond positively to a marketing campaign with probability estimation. The predictions were tuned locally using roc-auc and final scoring was done by uploading to the Kaggle competition.

Metrics

For the customer segmentation, average of sum of squared errors (SSE) within-cluster distances was used to help determine the number of clusters to produce for K-Means algorithm. Then clusters were analyzed and evaluated on how informative they are on the customer base.

For the marketing campaign, the evaluation metric is AUC for the ROC curve, where he ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR). It divides accuracy into sensitivity (true positive rate) and specificity (true negative rate). If only accuracy were to be considered, with a very imbalanced dataset, a classifier that predict the majority class the dominant class, which is not helpful.

II Analysis: Customer Segmentation

Data Exploration

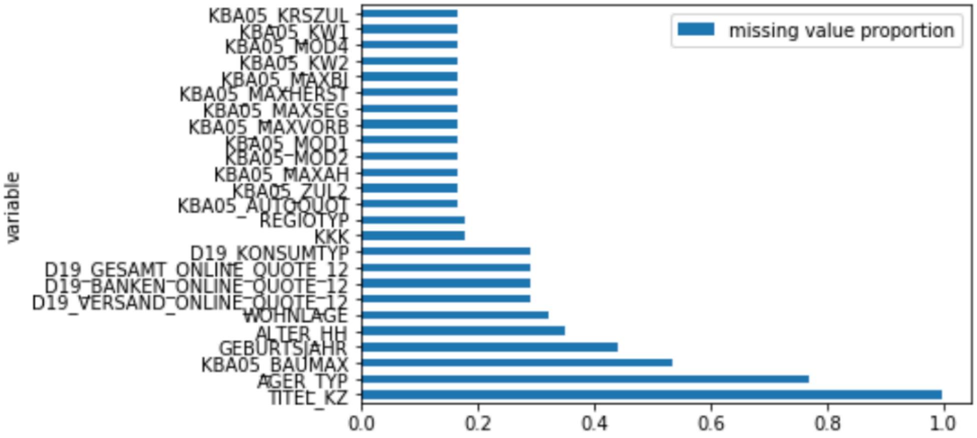
Features missing detailed information

Out of the 314 features in the demographics data, 94 of them were not described in the data dictionary. Because for the first part of this project (Customer Segmentation), how to perform feature engineering depends on data types of variables, disregard the variables that don't have description in the dictionary as there is no way of knowing their data types (Categorical, Ordinal, etc.) and the business interpretation.

Exploratory Visualization

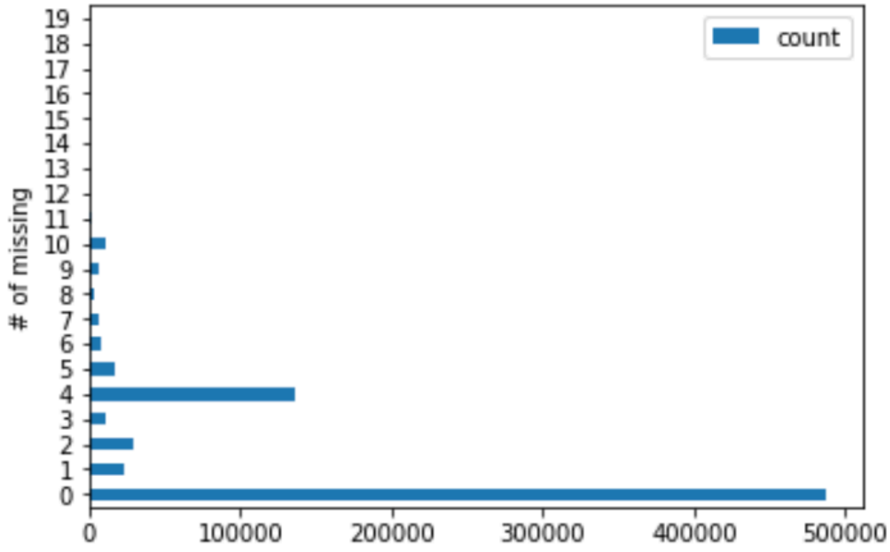
Missing values by column

Description files gave a detailed information of the meaning of each label including how missing value are labeled. We need to map the label of missing value back to NaN and investigate the number of missing value in each column. A majority of the features have missing values amount around 20%, so the threshold is selected as 30% to exclude some of the variables that have too much missing values (5 features would be removed as shown in the chart below listing the variables with most missing values).



Missing values by row

We want to keep as many data as possible while removing rows with too many missing variables. Based on the analysis in the chart below, 10 is a good threshold to exclude rows with more than 10 missing variables. As a result, 737067 out of 891211 rows could be retained.



Algorithms and Techniques

PCA

Principal Component Analysis (PCA) was performed to reduce dimension because of the large number of features and the fact that many of them provide almost the same information (Cumulative variance explained graph would be used to pick the optimal number of component to keep at lease 90% of variance). For the customer segmentation task, a clustering algorithm will be used and they run very slowly with large numbers of features.

K-Means Clustering

For the customer segmentation, average of sum of squared errors (SSE) within-cluster distances was used to help determine the number of clusters to produce for KMeans algorithm. The clusters were then analyzed and evaluated on how informative they are on the customer base.

III. Methodology

Data Preprocessing

Missing Values

(1) Columns with greater than 30% missing were dropped

(2) Rows with greater than 10 missing values were dropped

Feature Selection

(1) Features not found in the data dictionary would be dropped

(2) Categorical variables that have more than two levels are dropped as well to avoid introducing so many features after on-hot encoding.

(3) KBA_ variables were dropped because they caused problems with PCA

Re-encoding and Engineered Features

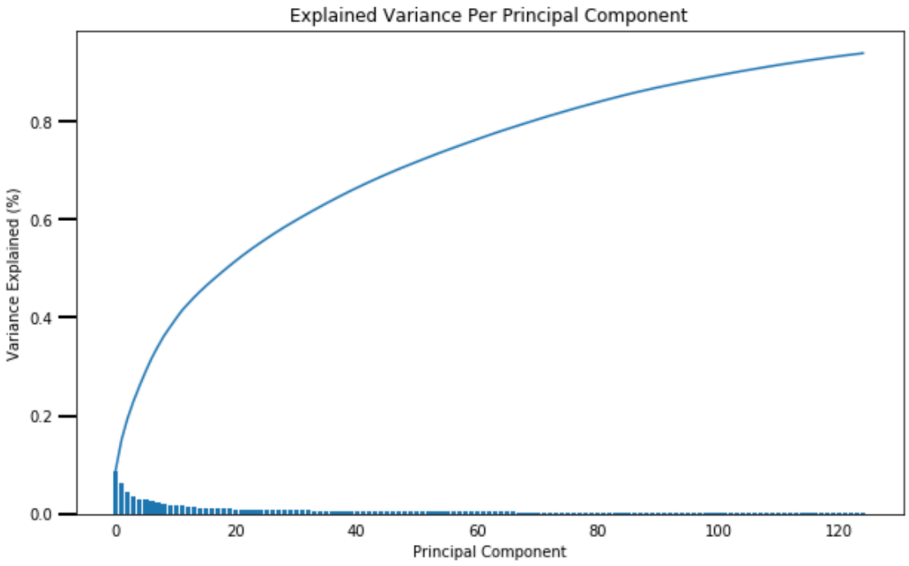
(1) Apply one-hot encoding to categorical data. For numeric and interval data, these features can be kept without changes. Impute the missing value data by “median” and use StandardScaler to scale all the features.

(2) PRAEGENDE_JUGENDJAHRE was split in decade and movement

(3) Columns starting with D19_: 10’s (no transactions) were re-coded as 0’s (no transactions). Only some of the D19 columns had 10’s. Re-coding was done to be consistent with the other D19 columns.

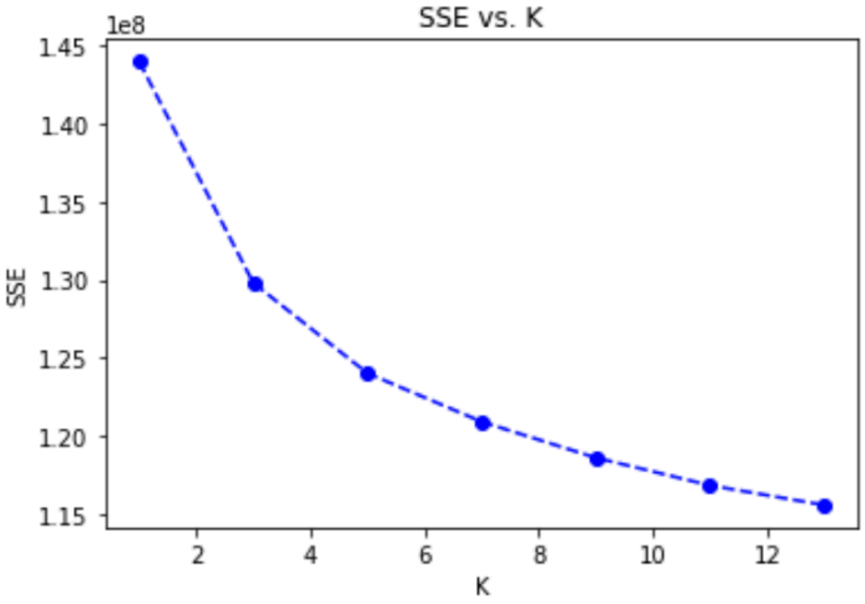
PCA

Cumulative variance explained graph would be used to pick the optimal number of component to keep at lease 90% of variance. Finally, 150 components were selected to keep 90% variance.



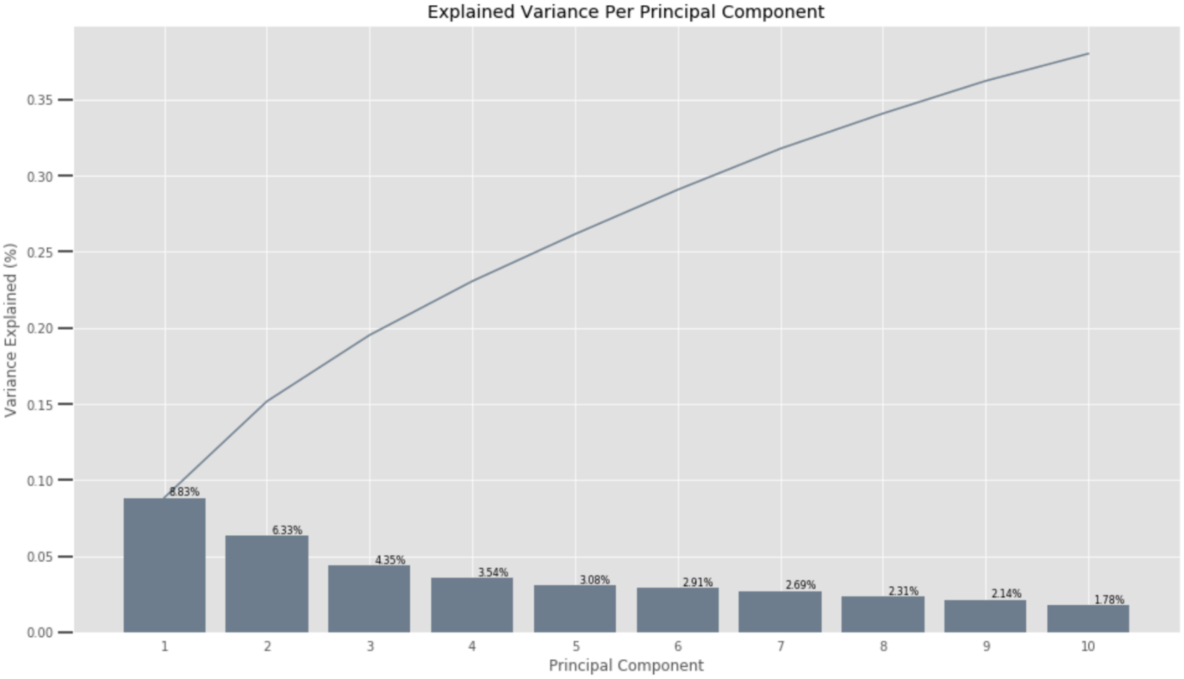
Implementation

Elbow method was used to identify an ideal number of clusters for k-means clustering on the PCA-transformed data. The plot demonstrates that the SSE rapidly decreased for the first 8 clusters and then continued to decrease for higher number of clusters but with lower slope. So, 8 clusters were selected as ideal number for k-means clustering. and the next step is to re-fit a K-Means instance to perform the clustering operation and apply it to our CUSTOMERS dataset.



IV. Results

Model Evaluation and Validation



PCA Interpretation

First principal component: Rich People Index (8.83 % variance)

Lowest:

HH_EINKOMMEN_SCORE	-0.161
CAMEO_DEUG_2015	-0.152
PLZ8_ANTG3	-0.135

Highest:

LP_STATUS_GROB	0.145
KBA05_ANTG1	0.148
MOBI_REGIO	0.152

The most prominent features are LP_STATUS_GROB, KBA05_ANTG1, MOBI_REGIO.
The most prominent negative features are HH_EINKOMMEN_SCORE, CAMEO_DEUG_2015 and PLZ8_ANTG3.

The first principal component is associated with size, wealth and type of family. Due to low mobility (moving patterns), high share of 1–2 family houses in the PLZ8, top earners with high income. It might be some people who are living in a busy city area (like financial district) with stable high-income jobs and tend to live alone or live with partner only.

Second principal component: Expensive Car Index (6.33 % variance)
Lowest:

KONSUMNAEHE	-0.153
MOBI_REGIO	-0.143
KBA13_SITZE_5	-0.139

Highest:

KBA13_SEG_OBEREMITTELKLASSE	0.161
EWDICHTE	0.174
KBA13_HERST_BMW_BENZ	0.182

The most prominent features are KBA13_SEG_OBEREMITTELKLASSE, EWDICHTE and KBA13_HERST_BMW_BENZ.
The most prominent negative features are KONSUMNAEHE, MOBI_REGIO, KBA13_SITZE_5.

The second principal component is primarily affected by car ownership and density of inhabitants and high mobility. This group of people living near consumption cells owns a lot of luxury cars such as BWM and Mercedes increase this component. The building here might be for short-term rent so inhabitants would not stay a long period of time.

Third principal component: Young and wandering Index (4.35 % variance)
Lowest:

ALTERSKATEGORIE_GROB	-0.252
FINANZ_VORSORGER	-0.218
SEMIO_ERL	-0.203

Highest:

SEMIO_PFLICHT	0.233
SEMIO_REL	0.241

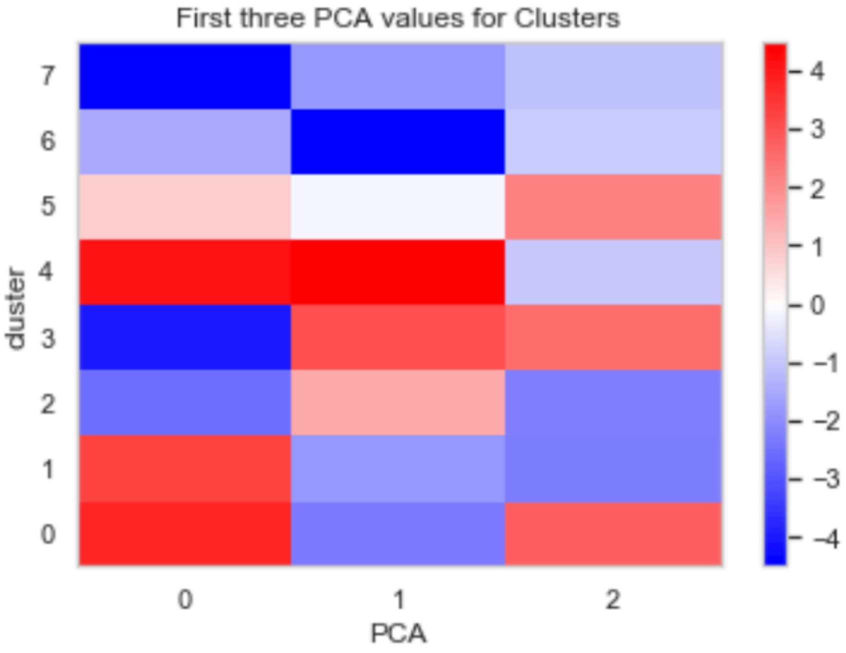
```
    FINANZ_SPARER       0.246
```

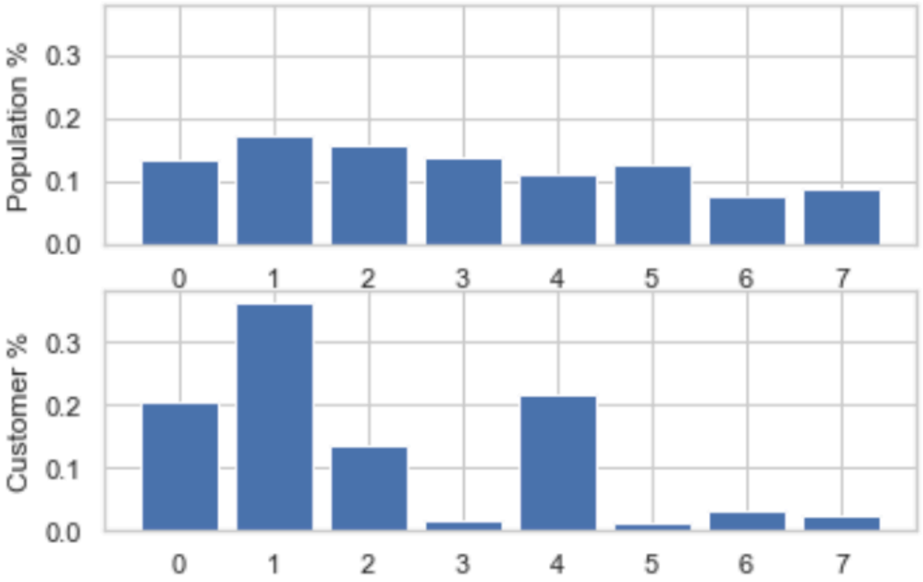
The most prominent features are SEMIO_PFLICHT, SEMIO_REL and FINANZ_SPARER.
The most prominent negative features are ALTERSKATEGORIE_GROB, FINANZ_VORSORGER, SEMIO_ERL.

The third principal component is primarily composed of young people. They are not money savers, has high eventful orientated affinity, and commonly no religious belief.

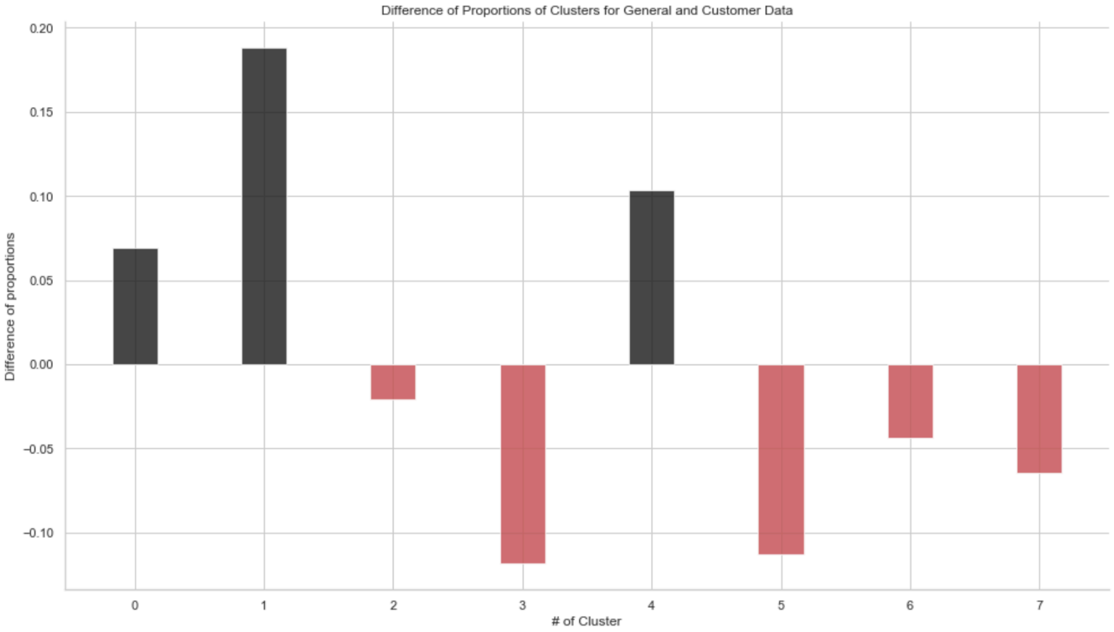
To sum up,
(1) First Component is an indicator of wealth, higher social status, couple or alone living.
(2) Second Component is an indicator of expensive car owner and high population density.
(3) Third Component is an indicator of lack of religiousness, youth, free spending

Justification





Difference in proportion between customers and general audience: positive is overrepresented and negative is underrepresented



Cluster 1: overrepresented

PCA values: 3.27996434, -1.77458916, -2.23528007
The most distinguishing feature of Cluster 1 is the high for first component. This indicates this cluster is people with high social status and income, who live with their couples or alone. It also as a negative value for second component and third component which is an indicator of not having expensive cars, years of working experience and long-term money investor. They might have high interest in financial news or products.

Cluster 3: underrepresented
PCA values: -3.93042388, 3.06903086, 2.54018613
Cluster 3 has a very high negative value for first component and a high value for second component and third component. People fall into this cluster has no high income but afford expensive cars and living near highly populated consumption area. Their income might be spending on the expensive car mortgage and hence no enough savings for financial products.

Cluster 5: also underrepresented
PCA values: 0.87828182, -0.14754324, 2.19030391
For cluster 6, the most distinguishing feature is the high third component value. This cluster is similar to cluster 3 except for the fact that the people in this group have decent income. They are of younger age and free spending. They might not high interest in financial news or enough savings for financial products.

II Analysis: Marketing Prediction

Data Exploration

Here is the second part of the project. Supervised training techniques were used to predict whether a individual would respond positively to a marketing campaign.

Imbalanced data

After some quick investigation against MAILOUT_TRAIN dataset, we can find among 42450 individuals, only 532 people responses to the mail-out campaign, which means the training data is highly imbalanced One of the problems.

Algorithms and Techniques

Xgboost (lightgbm) is optimized gradient-boosting machine learning library. It is selected for this project as the final model for good reasons: (1) Core algorithm is parallelizable: The speed is very fast. As a result, I can afford the time for finer parameters tuning; (2) Consistently outperforms single-algorithm methods: Boosting itself is a sequential technique which works on the principle of an ensemble. It combines a set of weak learners and delivers improved prediction accuracy. As our dataset is highly imbalanced, an ensemble boosting tree should perform well in detecting the small signal of the minority class.

Benchmark

An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. In fact, it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means model has no class separation capacity whatsoever.

III. Methodology

Data Processing

Missing Values

(1) Columns with greater than 30% missing were dropped

(2) For prediction task, to avoid dropping important signal (missing values co-occurrence pattern), all rows are retained.

Feature Selection

(1) Features not found in the data dictionary would not be dropped but treated as feature input in the modeling.

(2) Drop categorical variables with large number of levels and least indication to the response variable.

(3) EINGEFUEGT_AM' (a datetime) was dropped

Re-encoding and Engineered Features

(1) Apply one-hot encoding to categorical data. For numeric and interval data, these features can be kept without changes. Impute the missing value data by “median” for numeric variable and “most frequent” for categorical variable.

(2) PRAEGENDE_JUGENDJAHRE was split in decade and movement

(3) Columns starting with D19_: 10’s (no transactions) were re-coded as 0’s (no transactions). Only some of the D19 columns had 10’s. Re-coding was done to be consistent with the other D19 columns.

Implementation

5-fold cross-validation (to avoid overfitting) with ROC score was applied to the fairly evaluate the performance of each of the model. With imbalanced datasets is that the model is more biased towards the class with higher occurrence; in this case the model would be biased towards Certified. Based on this discovery, we have to split the data based on the distribution of target. Here we can use ‘Stratifi-edKFold’ in the cross validation to split the training data to meet our needs.

Refinement

Hyperparameters Tuning

(1) Try a bunch of different hyperparameters (learning rate, number of ite rations, boosting type, number of leaves)
(2) Fit all of them separately
(3) See how well each performs
(4) Choose the best performing one ('boosting_type': 'dart', 'learning_rat

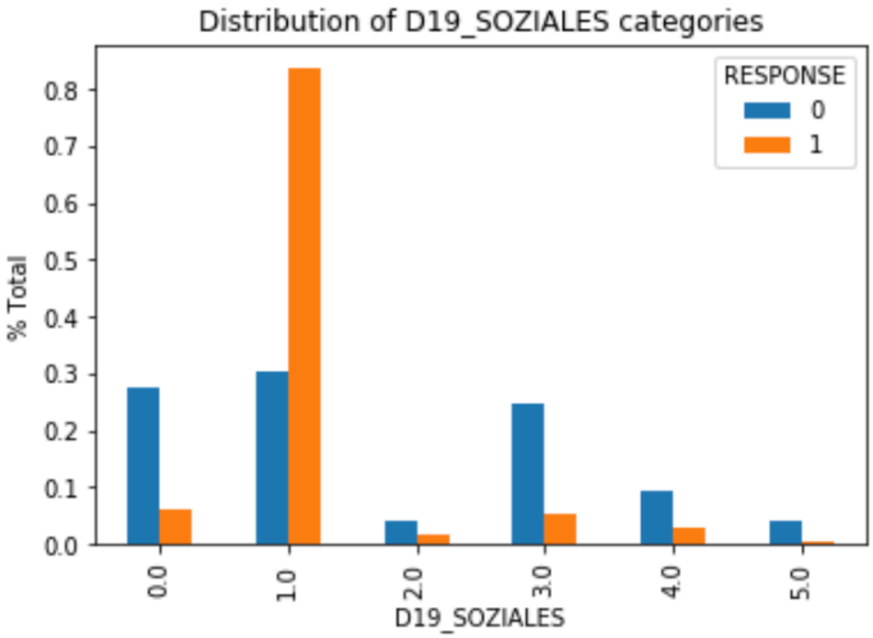
e': 0.01, 'num_iterations': 200, 'num_leaves': 62)
(5) It is essential to use cross-validation (5-fold cross-validation)

As part of the project, half of the mailout data has been provided with included response column. For the competition, the remaining half of the mailout data has had its response column withheld; the competition will be scored based on the predictions on that half of the data. My current rank in the public leaderboard is #9 with the name “Labienus”. My final score in kaggle is 0.80568 with AUC 0.764120, which is an improvement of your previous score of 0.79459 using gradient boosting.

https://www.kaggle.com/c/udacity-arvato-identify-customers/leaderboard

V. Conclusion

Free-Form Visualization



The feature importance plot demonstrates that the most important feature is 'D19_SOZIALES', for which description is not available. Even though there is no information available about D19_SOZIALES specifically, Overall D19_* features are related to frequency of using certain group of products. Thus, the method distinguished two groups with high and low transaction activity. So, people who are involved in this (D19_SOZIALES) kind of transaction are most likely to respond to the market campaign and become customer of the mail-order company.

The next important features are 'EINGESOGENAM_HH_JAHR ' and 'VK_DISTANZ'. Fortunately, the descriptions are not available either. And there are no other similar variables that help interpret their meaning. Would ask the mail-order company for more clarity.

Reflection

For Customer segmentation part, with understanding the difference between the general population and customer base, the mail order company can be much more focus on their target, and then increase conversion rate or lower down their marketing cost. The impact is large.

Improvement

For the marketing prediction part, the model output yield decent result yet still has room for improvement. Feature importance plot is useful to select only best features with highest correlation to the outcome(s). To improve model fitting performance (time or overfitting), less important features can be removed before training the model.