Machine Learning Engineer Nanodegree Capstone Project

Customer Segmentation Report for Arvato Financial Services

Qiyue Ren

October 17th, 2021

Table of Contents

I. Definition

Project Overview

In this project, we analyzed the demographics data for the general population of Germany and the customer characteristics of a mail-order company in Germany. And we apply the analysis to predict potential customers for the company. Unsupervised learning algorithms are used to segment customers. Supervised learning methods help identity the target for marketing campaign.

The report consists of the following sections:

1. Definition
2. Data Analysis
3. Methodology
4. Result
5. Conclusion

Domain Background

Arvato is a globally active services company that designs and delivers award-winning customer service and back-office processing services for customers. It helps its clients achieve strategic objectives and delight its customers. [1]

Arvato is helping a mail-order company with efficient customer acquisition. As a typical customer segmentation problem, it will identify the similarities of different groups and build up connections with clients in the groups. The customer segmentation analysis will enable marketing teams to identify customers with high efficiency and accuracy so as to maximize the value of customers.

Machine learning methodologies are a great tool for analyzing customer data and finding insights and patterns. For example, Torizuka [2] used the Random Forest algorithm for segmentation and mentioned that the algorithm recognizes training data with high accuracy even in the presence of noise and outliers. Ezenkwu, Ozuomba and Kalu [3] applied clustering algorithm to discover subtle but tactical patterns or relationships buried within a repository of unlabeled datasets. Artificially intelligent models are powerful tools for decision-makers. We will analyze the attributes of existing clients and employ machine learning models to help identify most likely new clients based on certain characteristics.

Datasets and Inputs

The data are provided by Arvato Company:

- 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).

In addition, there are 2 datasets for attribute description:

- DIAS Attributes — Values 2017.xlsx: Maps the values and meanings to each attribute.
- DIAS Information Levels — Attributes 2017.xlsx: Describe the details of attributes.

Problem Statement

As stated by the Arvato manager, our main goal from a business perspective is: given the attributes and demographic information of existing clients, how can the mail-order company acquire new German clients more efficiently?

In specific, we need to solve the following problems:

- Identify what demographic features are corresponding to customers of the mail-order company;
- Segment customers into proper groups;
- Predict possible customers based on demographic information.

In order to solve these problems, first of all, we will explore the datasets provided by Arvato. Secondly, we will use unsupervised learning models to help analyze the attributes and characteristics of the established customer to create customer segmentation and try to understand the relations between demographics features and existent customers and predict potential customers. Thirdly, we will use the previous analysis to build supervised learning models to predict whether a person will be a potential customer. In order to achieve our goals, the solution is divided into three steps:

- Data analysis: The data will be explored and cleaned. If there are any problems, such as missing values, inappropriate measurement and high correlation, we will fix them and prepare data for further analysis.
- Customer segmentation: We will use K-Mean Algorithm to obtain clusters and assign data into groups based on selected features. In order to speed up this process, Principal Component Analysis (PCA) algorithm will be employed for dimension reduction.
- Customer prediction: Based on the previous segmentation, we will try several models and choose the most accurate one for the prediction: logistic regression (benchmark model), Decision Tree classifier, Gradient Boosting classifier, Random Forest classifier, XGBoost classifier and LGBM classifier.

Metrics

Unsupervised learning algorithms, which are used on unlabeled data, will play an important role in our solution. However, unlike supervised learning algorithms which have clear reference of accuracy, it requires us to define or choose a proper evaluation metric to evaluate the performance of the algorithm.

In PCA, we will make of use Within-cluster sum of square (WSS), which means the sum squares of distance of all points in a cluster from its center.

In classification, we choose Area under Receiver Operating Curve (AUROC) as the metric. Note that the data is highly imbalanced. In the MAILOUT-TRAIN dataset, over 98% of the response was labeled as zero. Under this circumstance, simply using accuracy as a metric might cause over-fitting towards the non-response data while AUROC could address this imbalance and evaluate the mode.

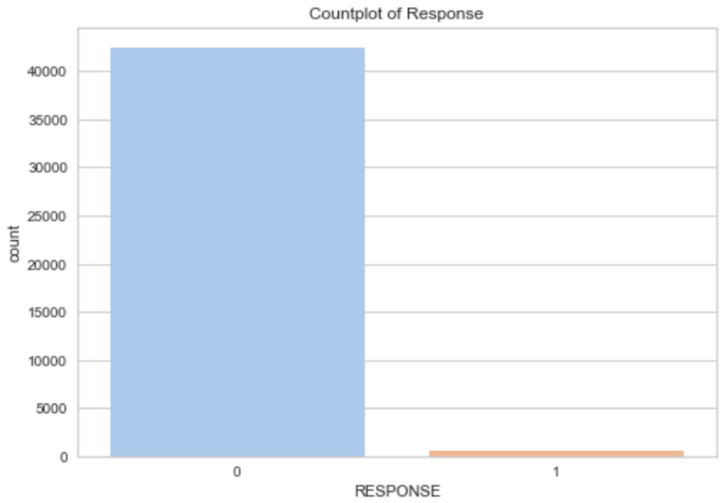


Fig1: Imbalanced Training Data

This is also the evaluation method for the Kaggle competition. To be specific, a ROC, or receiver operating characteristic, is a graphic used to plot the true positive rate (TPR, proportion of actual customers that are labeled as so) against the false positive rate (FPR, proportion of non-customers labeled as customers). The AUC, or area under the curve, summarizes the performance of the model. If a model does not discriminate between classes at all, its curve should be approximately a diagonal line from (0, 0) to (1, 1), earning a score of 0.5. A model that identifies most of the customers first, before starting to make errors, will see its curve start with a steep upward slope towards the upper-left corner before making a shallow slope towards the upper-right. The maximum score possible is 1.0, if all customers are perfectly captured by the model first [4].

II. Analysis

Data Exploration and Visualization

The datasets are checked for mixed types, missing values, correlation and attributes.

Overall information

`AZIDIAS`:

- Range Index: 891221 entries, 0 to 891220
- Columns: 366 entries, LNR to ALTERSKATEGORIE_GROB

- dtypes: float64(267), int64(93), object(6)
- memory usage: 2.4+ GB

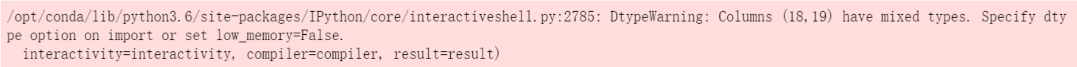
`CUSTOMERS`:

- Range Index: 191652 entries, 0 to 191651
- Columns: 369 entries, LNR to ALTERSKATEGORIE_GROB
- dtypes: float64(267), int64(94), object(8)
- memory usage: 539.5+ MB

We noticed that `CUSTOMERS` has 369 attributes while `AZIDIAS` has 366 attributes. There are 3 attributes unknown to `AZIDIAS`. In order to keep consistence, we will select the common attributes from them.

Mixed types

As loading the data set from csv files, there is a warning message: "typeWarning: Columns (18,19) have mixed types. Specify dtype option on import or set low_memory=False." , which means column 18 and 19 have mixed type data.



```
/opt/conda/lib/python3.6/site-packages/IPython/core/interactiveshell.py:2785: DtypeWarning: Columns (18,19) have mixed types. Specify dty
pe option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

Fig2: Warning Message

By obtaining the unique values of column 18 and 19, we can identify the source of warning is miss valued "X" and "XX".

Missing values

There are missing values in datasets. We selected 10 attributes with the highest missing proportion from `AZDIAS` and `CUSTOMERS`.

AZDIAS(%)		CUSTOMERS(%)	
ALTER_KIND4	99.86	**ALTER_KIND4**	99.88
ALTER_KIND3	99.31	**ALTER_KIND3**	99.33

ALTER_KIN	96.69	**ALTER_KIND2**	97.34
ALTER_KIND1	90.90	**ALTER_KIND1**	93.86
EXTSEL992	73.40	**KK_KUNDENTYP**	58.41
KK_KUNDENTYP	65.60	**EXTSEL992**	44.50
ALTERSKATEGORIE_FEIN	29.50	**KBA05_KW1**	29.21
D19_LETZTER_KAUF_BRANCHE	28.85	**KBA05_MOD1**	29.21
D19_LOTTO	28.85	**KBA05_MOTOR**	29.21
D19_VERSI_ONLINE_QUOTE_12	28.85	**KBA05_MOD4**	29.21

Table1: Attributes with High Missing Value Proportion

The distributions NaN values in column and row are plotted below.

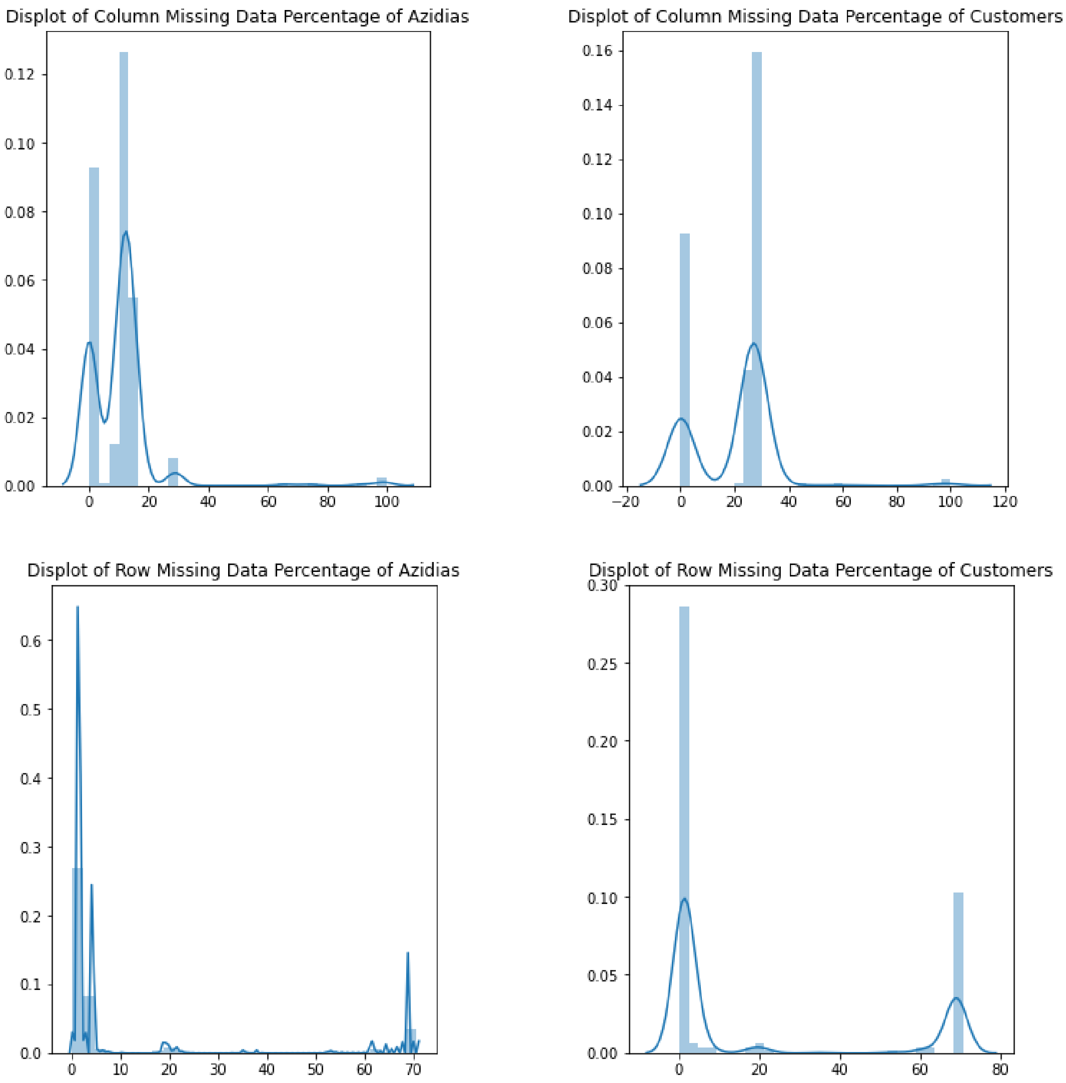


Fig3: Displot of Missing values in Azdias and Customers Datasets

High missing data proportion might cause outlier problems and hurt the clustering result while dropping too many attributes might affect the accuracy of algorithms. We expected 30% would be a proper threshold to clean the datasets.

Correlations

There are 26 correlated features with correlation larger than 0.9:

'ANZ_STATISTISCHE_HAUSHALTE', 'CAMEO_DEUG_2015', 'CAMEO_INTL_2015',
D19_GESAMT_ANZ_24', 'D19_VERSAND_ANZ_12', 'D19_VERSAND_ANZ_24',
D19_VERSAND_DATUM', 'D19_VERSAND_ONLINE_DATUM',
D19_VERSAND_ONLINE_QUOTE_12', 'KBA05_KRSHERST2',
KBA05_KRSHERST3', KBA05_SEG9', 'KBA13_HALTER_66',
'KBA13_HERST_SONST', 'KBA13_KMH_250', LP_FAMILIE_GROB',
'LP_LEBENSPHASE_FEIN', 'LP_LEBENSPHASE_GROB', LP_STATUS_GROB',
'MIN_GEBAEUDEJAHR', 'ORTSGR_KLS9', 'PLZ8_ANTG1', PLZ8_ANTG3',
'PLZ8_BAUMAX', 'PLZ8_GBZ', 'PLZ8_HHZ'

Generally, highly correlated data hurt the performance of machine learning algorithms since they could attract too much attention to certain features and over value the importance of the common features. Therefore, we should identify the highly correlated features and drop them.

Attribute exploration

Due to the different format of dataframe and excel, there are some NaN values in **Attribute** column and **Description** column in `DIAS Attributes - Values 2017` and in **Information level** column in `DIAS Information Levels - Attributes 2017`. We would like to fill these NaN values with their former values for the convenience of later use.

Algorithms and Techniques

Customer segmentation

Unsupervised learning algorithms can group data points based on similar attributes in the dataset. One of the main types of unsupervised models is clustering models.

K-means clustering is a great method for our customer segmentation problem. The algorithm determines K clusters in data and assigns each input to one of K clusters.

There are more than 300 attributes in datasets. However, some dimensions are not as important as others. We want to find the features that are of great importance and could actually help segmenting data. In addition, too many attributes will also take a long time to group data. So, before clustering data, we want to take dimension reduction. PCA or principal component analysis is a strong method to solve this problem. PCA attempts to reduce the number of features within a dataset while retaining the "principal components", which are defined as weighted linear combinations of existing features that are designed to be linearly independent and account for the largest possible variability in the data. [5]

Besides, the Elbow method helps us select the number of clusters to minimize the variation within each cluster.

Given a set of principal data points and the optimal value of K, K-means clustering algorithm will identify the centroids values for clusters and label each input.

Customer prediction

Supervised learning algorithms are used on labeled data. We are aim to identify customers who will response to marketing campaign, which is a binary classification problem.

A typical and simple model for binary classification problem is logistic regression. We also proposed some advanced classifiers:

- Decision Tree classifier uses a decision tree as a predictive model to go from observations about an item represented in the branches to conclusions about the item's target value [6]
- Random Forest classifier is an ensemble algorithm which combines several methods for classification. Then it aggregates different methods to get the final result.
- Gradient Boosting classifier is the classifier of Gradient Boosted Decision Trees (GBDT) which is a generalization of boosting to arbitrary differentiable loss functions. GBDT is an accurate and effective off-the-shelf procedure that can be

used for both regression and classification problems in a variety of areas including Web search ranking and ecology. [7]

- XGBoost classifier is kind of GBDT framework. But it improved the regularization objective function so as to implement the GBDT algorithm with higher efficiency. .
- LightGBM classifier is also a framework that implements the GBDT algorithm. It supports high-efficiency parallel training, and has the advantages of faster training speed, lower memory consumption, better at processing of massive data.

All these models will be trained with default hyperparameters to solve a binary classification problem.

Benchmark

The benchmark model would be the logistic regression model with 1 stands for a potential new customer and 0 stands for not a potential new customer. This simple classification model will establish initial results upon which further improvements can be made to assess the relative improvements.

After fitting the data, the AUROC score was 0.6738 and the training process took 1.9972 seconds.

III. Methodology

Data Preprocessing

The preprocessing step consists of the following steps:

- Address the warning message and mixed type data.
- We replaced ”X” and ”XX” will NaN and converted the data type to float.
- Deal will missing data: Drop columns and rows with high proportion of NaN values. We will set 30% as a threshold. And replace the remaining with proper values, such as median and mode.

- Convert categorical data to numerical data: Employ re-encoding for categorical features. For example, in AZDIAS, OST_WEST_KZ, D19_LETZTER_KAUF_BRANCHE, CAMEO_DEU_2015 and EINGEFUEGT_AM needs to be processed.
- Drop the 26 highly correlated attributes in each dataset.
- Scale for all attributes by MinMax scaler.

Implementation

PCA

We employed PCA to reduce dimensions and select principal components.

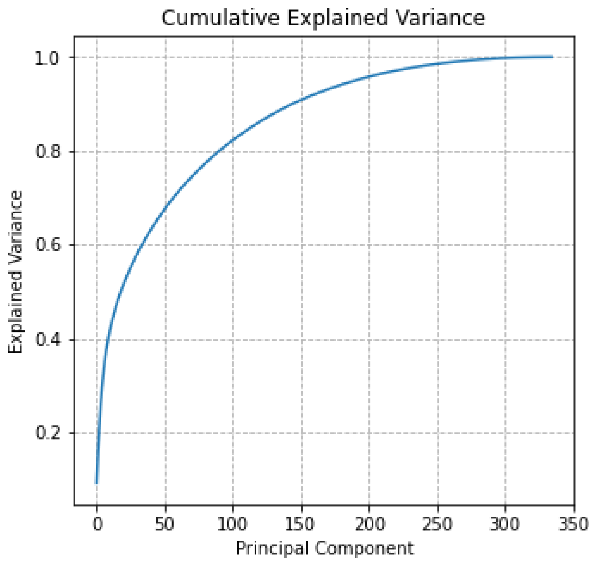


Fig4: Cumulative Explained Variance

As shown above, cumulative explained variance with the number of principal components is showed above. We can see that around 150 components explain 90% of the variance.

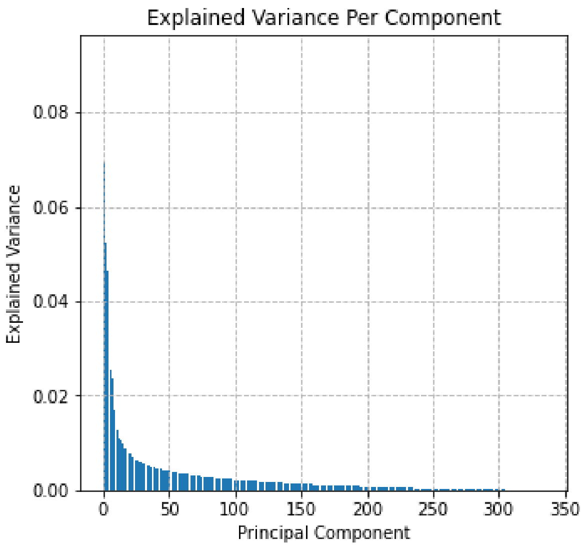


Fig5: Explained Variance Per Component

And after 150 components, the variance explained by each component is less significant. By further calculation, the explained variance for the top 150 principal components is 90.82%. Therefore, we set 150 as the number of principal components.

We also defined a function to help display the features of a specific PCA component. For example, the top 10 weighed features in the seventh component:

HEALTH_TYP, VERS_TYP, SHOPPER_TYP, KOMBIALTER, PRAEGENDE_JUGENDJAHRE, SEMIO_REL, AKT_DAT_KL, D19_KONSUMTYP, HH_EINKOMMEN_SCORE, FINANZ_UNAUFFAELLIGER

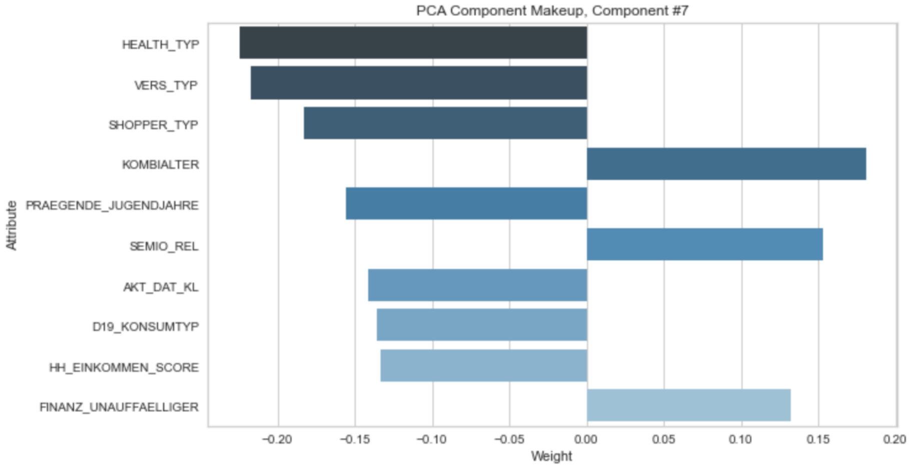


Fig6: The Makeup of Component #7

Elbow Method

Then we used the Elbow method to select the optimal number of clusters.

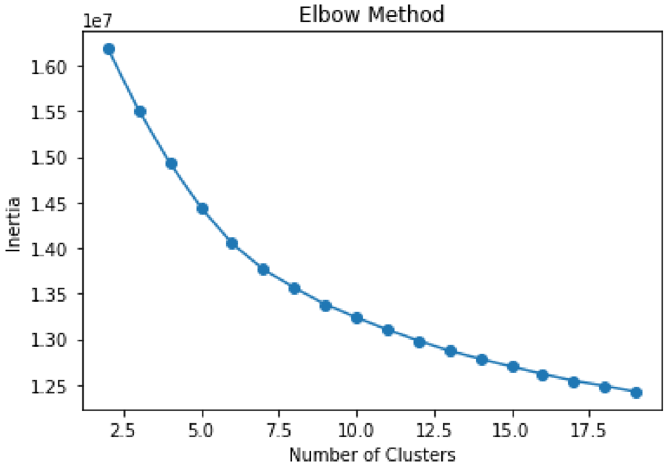


Fig7: Elbow graph

In the plot of Distortion score versus K, we can see that the sum of squared error decreasing. At around 9 clusters, the decreasing speed significantly reduced. This is visible as an elbow.

K-Means Algorithm

Based on PCA and Elbow method, we fitted the PCA transformed data with K-Means model and got the clustering results.

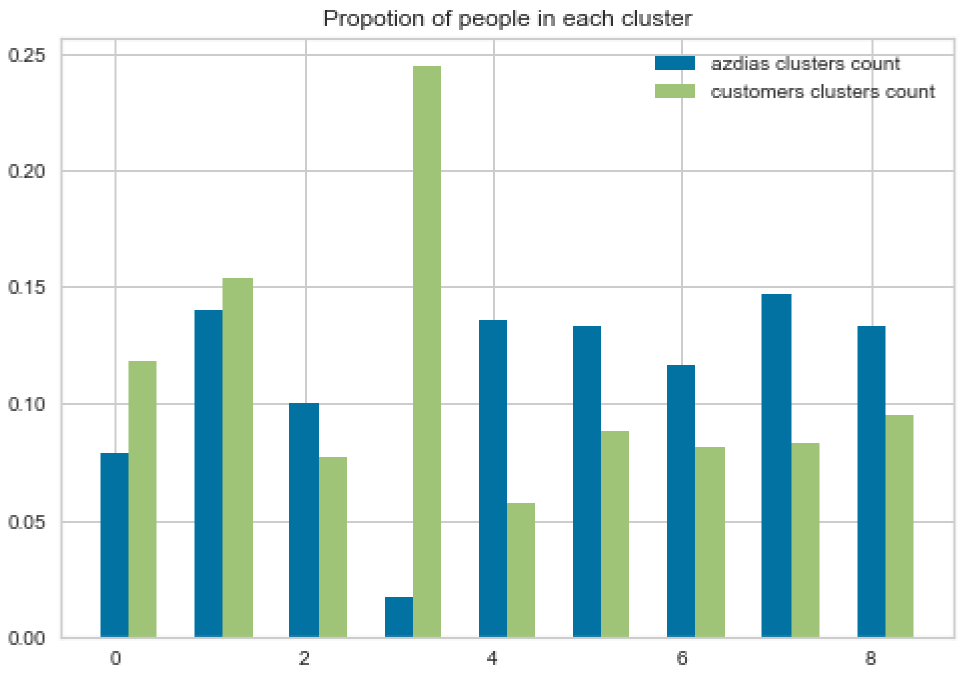


Fig8: Proportion of People in Each Cluster

From the graph, we can see the distributions of Azdias clusters and Customers clusters vary.

Note that if we regard the features of mail-order company's customers to be similar with the country's population, then the distribution of customers clusters should be similar with that of Azdias clusters.

We want to explore the special characteristics of customers. So, we should pay more attention to the mismatches. If there are particular clusters of the population are interested in the company's products, then we should see a mismatch from the one to others. For example, compared to the population, a cluster has a higher proportion in customers. In this circumstance, we could expect it is the characteristics of target customers. Conversely, if a cluster has a lower proportion, we could expect it is not belong to the target customers.

Therefore, we are super interested in clusters which have higher proportions in customers than in Azdias. From the graph above, we can see that cluster 3 in customers have visibly higher proportion. Cluster 4 have higher proportions in Azdias.

Classification

The performance of classifiers is showed below.

	Accuracy	AUROC score	Training Time (seconds)
Benchmark (LogisticRegression)	0.9873	0.6738	1.9972
DecisionTreeClassifier	0.9727	0.4938	3.5619
RandomForestClassifier	0.9873	0.6020	14.5524
GradientBoostingClassifier	0.9868	0.7551	57.9933
XGBClassifier	0.9873	0.6711	10.9956
LGBMClassifier	0.9873	0.7073	2.9873

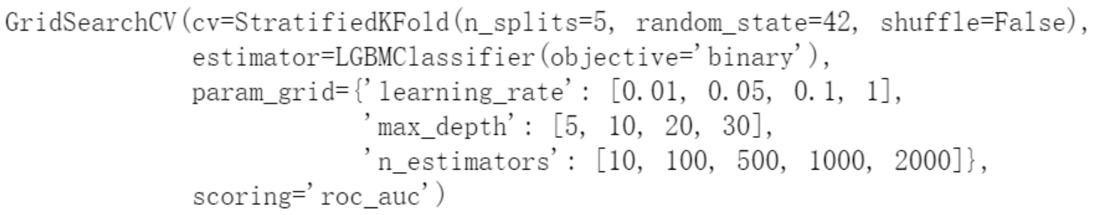
Table2: Classification Results

Among the five training models with the default hyperparameters, GradientBoostingClassifier got the highest AUROC score which is approximately 0.7551. And LGBMClassifier got the second high score, 0.7073. However, training the model with GradientBoostingClassifier took around 58 seconds while LGBMClassifer took 2.98 seconds. We decided to take refinement on the two types of models later.

Refinement

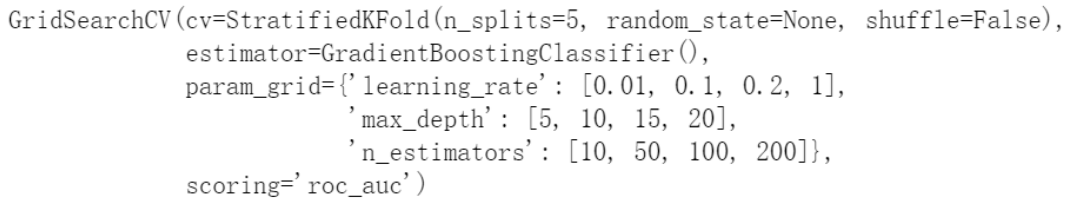
To test the general performance of classifiers on our data, we used the default hyperparameters. And now we would like to refine the well performed classifier.

A set of hyperparameters was selected to tune classifiers with the help of grid search.



```
GridSearchCV(cv=StratifiedKFold(n_splits=5, random_state=42, shuffle=False),
             estimator=LGBMClassifier(objective=’binary’),
             param_grid={’learning_rate’: [0.01, 0.05, 0.1, 1],
                         ’max_depth’: [5, 10, 20, 30],
                         ’n_estimators’: [10, 100, 500, 1000, 2000]},
             scoring=’roc_auc’)
```

Fig9: LGBM Classifier Grid Search Result



```
GridSearchCV(cv=StratifiedKFold(n_splits=5, random_state=None, shuffle=False),
             estimator=GradientBoostingClassifier(),
             param_grid={’learning_rate’: [0.01, 0.1, 0.2, 1],
                         ’max_depth’: [5, 10, 15, 20],
                         ’n_estimators’: [10, 50, 100, 200]},
             scoring=’roc_auc’)
```

Fig10: Gradient Boosting Classifier Grid Search Result

The final LGBMClassifier performed better.

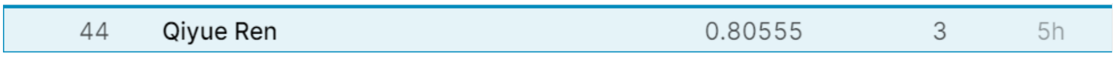
	Initial AUROC score	**Final AUROC score**
GradientBoostingClassifier	0.7551	0.76
LGBMClassifier	0.7073	0.77

Table3: Refinement Result

IV. Results

Model Evaluation and Validation

We have tuned a LGBM classifier which got 0.77 AUROC score. It has the best performance compared to other classifiers. Therefore, we decided to use it to predict which individuals are most likely to respond to a mailout campaign. The prediction on test set was uploaded to Kaggle and got a score of 0.8055.



44 Qiyue Ren 0.80555 3 5h

Fig11: Kaggle Result

Justification

In benchmark model, we use Logistic regression to identify potential customers. Its AUROC value is 0.67, while our final model, tuned LGBM classifier has better performance and got 0.77 scores. The improvement was around 15%. Our final model is significant in searching potential customers.

V. Conclusion

In this project, we analyzed the demographics data for the German population and the customer characteristics of a mail-order company in Germany. It turned out that the major features in the customers include the health conditions of customers, the distance to the city center, the types of cars.

In addition, we made prediction on potential customers. We applied six different supervised learning algorithms to identify which individuals are most likely to response to marketing campaign and become a customer in the future. And the optimal model in our training process is LGBM classifier. Its Kaggle score is 0.80555.

Reflection

There are several aspects that still need improvements:

- Feature engineering: The features in Azdias and Customers datasets are of great number and also complicated. We cannot know each very well. There are some features that are not well defined or classified, which might affect the accuracy of clustering and potential customer prediction.
- Data pre-processing: The Mailout dataset are highly imbalanced. In our project, we choose a proper evaluation metrics to address this problem. If we could re-sample the dataset, there might be more flexible learning algorithms for choice and might perform better on prediction.
- Supervised learning model refinement: Classifiers will have better performance if we use more variated hyperparameters. However, due to the limitation of device and time, only a small portion are tested.

Reference

[1] https://www.arvato.co.uk/

[2] Torizuka, K., Oi, H., Saitoh, H., & Ishizu, S. (2018). Benefit Segmentation of Online Customer Reviews Using Random Forest. 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 487-491.

[3] Chinedu Pascal Ezenkwu, Simeon Ozuomba and Constance kalu, “Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services” International Journal of Advanced Research in Artificial Intelligence(IJARAI), 4(10), 2015. http://dx.doi.org/10.14569/IJARAI.2015.041007

[4]https://www.kaggle.com/c/udacity-arvato-identify-customers/overview/evaluation

[5] https://en.wikipedia.org/wiki/Decision_tree_learning

[6]https://scikit-learn.org/stable/modules/ensemble.html#gradient-tree-boosting