1. PROJECT OVERVIEW

Arvato is an international company that develops and implements innovative data driven solutions for business customers from around the world. These include supply chain solutions, financial services and Avarto systems. In this project we will be working with Avarto finances data to provide a strategy for a mail order sales company to expand their customer base.

The research question we pursue is ‘How effectively can a mail order sales company expands its customer base?’. To answer this question, we can use two strategies. First one is a customer segmentation analysis with an unsupervised learning model. Second one is a supervised learning model to predict the probability of individuals turning into customers. In this project, I used both methods to provide an answer to our research question.

This report consists of six sections. Project overview is followed by a methods section in which data sets and data analysis procedures were explained. In the third section, the analysis conducted for preprocessing is presented. In forth and fifth sections, the results for customer segmentation and supervised learning models discussed. In the last section, a summary provided and limitations and possible improvements were discussed.

2. METHOD

2.1.Data sets

There are four data files associated with this project:

- ***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).

Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood.

The "CUSTOMERS" file contains three extra columns ('CUSTOMER_GROUP', 'ONLINE_PURCHASE', and 'PRODUCT_GROUP'), which provide broad information about the customers depicted in the file. The original "MAILOUT" file included one additional column, "RESPONSE", which indicated whether each recipient became a customer of the company or not. This column enables us to run a supervised learning model.

There are also two supplementary files which includes additional information about features, meaning and their coding.

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

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

2.2.Data Analysis

This project consists of two main parts: customer segmentation and supervised learning.

2.2.1. Customer segmentation

It is the process of dividing customers into groups based on some demographic variables to provide distinct groups enabling different market strategies. In this project, customer segmentation conducted in three steps. First, I examined the customer and general population distributions based on some demographics. Second, I conducted a principal component analysis (PCA) for dimension reduction. Lastly, I used kMeans clustering, an unsupervised learning model, to create customer clusters. Lastly, I compared the percentage of individuals in each cluster to understand which clusters are more likely to have future customers.

2.2.2. Supervised Learning

It is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. Depending on the output, one can use different methods. In our data, we have a binary classified output: responded or not responded. The performance of the methods also varies between data sets. To select the best performing model, I tried a couple of methods.

2.2.2.1.Dealing with unbalanced responses

When the class weights are not balanced, it becomes tricky to have a proper ml model to predict the smaller class. There are two approaches to deal with this problem: balancing class distribution with sampling techniques and using cost sensitive models. In this project, I compared sampling methods and cost sensitive training methods.

Sampling methods:

Sampling methods can be classified in two groups: oversampling and under sampling techniques. Oversampling methods duplicate cases in the minority class or synthesize new cases from the examples in the minority class. Some oversampling techniques are Random Oversampling, Synthetic Minority Oversampling Technique (SMOTE), Borderline-SMOTE, Borderline Oversampling with SVM, Adaptive Synthetic Sampling (ADASYN). In this project I used random oversampling, SMOTE and ADASYN techniques. Under sampling methods, on the other hand, delete or select a subset of examples from the majority class. Some under-sampling techniques are Random Under-sampling, Condensed Nearest Neighbor Rule (CNN), Near Miss Under-sampling, and Tomek Links Under-sampling. Under sampling techniques are better for the large data sets having large enough minority class. Hence, in this project I did not use a under sampling technique as my minority class size is not big enough. You can find detailed information about sampling techniques here (https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/)

Prediction Methods:

The most used technique to predict binary classifications is logistic regression. In this project, I used standard logistic regression with sampled data and class balanced logistic regression for imbalanced data.

There are also other prediction techniques works better with imbalanced data. In this project, I used random forest classifier (RFC) and X gradient boosting classifier (XGBoost). RFC is a tree base algorithm which creates decision trees form random samples, get prediction from each tree and select the best performing one. RFC is a robust and accurate technique, but it is slow. Gradient boosting classifier trains many models sequentially. It is a numerical optimization algorithm where each model minimizes the loss function. XGBoost is an advanced and more efficient implementation of Gradient Boosting. In this project, I used XGBoost method as it is more efficient than gradient boosting.

2.2.2.2.Evaluation Metrics

One needs a metric to evaluate performance of a model. The commonly used metrics are accuracy, precision, recall, f1 score and AUC. They were explained in table 1.

Table 1. Evaluation Metrics

Evaluation Metric	Equation
Accuracy: is the ratio of correct predictions to total number of predictions.	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$
Precision: is the ratio of correctly predictied positive values to total predicted positive observations.	$Presicion = \frac{TP}{TP + FP}$
Recall: is the ratio of correctly predicted positive observations to the all observations whose actual values are positive	$Recall = \frac{TP}{TP + FN}$
F1 score: Weighted average of precision and recall	$F1 = \frac{2 * Recall * Precision}{Recall + Presicion}$
AUC: It stands for Area under the ROC curve which measures the entire two-dimensional area under the entire ROC curve. Hence it is a better metric if data is highly unbalances and the ones, we would like to catch is the smaller group.	

Note:
TP: True positives, the ones correctly predicted as positive values
TN: True negative, the ones correctly predicted as negative values
FP: False positive, the ones incorrectly predicted as positives whose actual value is negative
FN: false negative, the ones incorrectly predicted as negatives whose actual value is positive

While accuracy is a handy and commonly used metric to evaluate an ML model, it may create misleading results when the class weights are not balanced. Hence it is always good to check the other metrics. In this project, we have very unbalanced class distribution. Hence, I used AUC as main criteria for model selection.

3. DATA PREPROCESSING

3.1.Understanding the data sets

In this part, I summarized basic data cleaning and preprocessing steps. I have used 6 different data sets which are explained detailly in Data set section. The main data sets were general population data and the customer data. A small preview of each data set is shown on Figure 1.

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4
0	910215	-1	NaN	NaN	NaN	NaN	NaN	NaN
1	910220	-1	9.0	0.0	NaN	NaN	NaN	NaN
2	910225	-1	9.0	17.0	NaN	NaN	NaN	NaN
3	910226	2	1.0	13.0	NaN	NaN	NaN	NaN
4	910241	-1	1.0	20.0	NaN	NaN	NaN	NaN

(a) General population data preview

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4
0	9626	2.0	1.0	10.0	NaN	NaN	NaN	NaN
1	9628	NaN	9.0	11.0	NaN	NaN	NaN	NaN
2	143872	NaN	1.0	6.0	NaN	NaN	NaN	NaN
3	143873	1.0	1.0	8.0	NaN	NaN	NaN	NaN
4	143874	NaN	1.0	20.0	NaN	NaN	NaN	NaN

(b) Customer data preview

Figure 1. Main data previews

General population data consists of 891,221 rows and 366 columns while customer data consists of 191,652 rows and 369 columns. Before diving into the data cleaning, we need to understand each variables and the meanings of the codes. For this purpose, I examined two supplementary data files: Information level data and attributes data. The previews of each data are shown at figure 2. Information levels include the column names in attribute column and description of these columns while attribute data includes column names and value meanings of each variable.

As a first step of understanding the data files. I checked whether we have information for all columns in general population and customer data. For this purpose, I checked whether there are variables in main data files that are not included in the attribute data. I encounter that there are 8 columns in the general population data that we have no information about attribute levels or encodings. In those 8 variables, 4 of them included in the information levels data. Those variables had been coded 1 to 10 however we have no information about the meanings of these codes. Moreover, between 76 to 99 % of the observations were in a single category (category number

10). As these variables does not provide much variance, I dropped them too. At the end, we had 272 columns for both customer and general population data.

	Information level	Attribute	Description	Additional notes
0	NaN	AGER_TYP	best-ager typology	in cooperation with Kantar TNS; the informatio...
1	Person	ALTERSKATEGORIE_GROB	age through prename analysis	modelled on millions of first name-age-referen...
2	NaN	ANREDE_KZ	gender	NaN
3	NaN	CJT_GESAMTTYP	Customer-Journey-Typology relating to the pref...	relating to the preferred information, marketi...
4	NaN	FINANZ_MINIMALIST	financial typology: low financial interest	Gfk-Typology based on a representative househo...

(a) Information level data preview

	Attribute	Description	Value	Meaning
0	AGER_TYP	best-ager typology	-1	unknown
1	NaN	NaN	0	no classification possible
2	NaN	NaN	1	passive elderly
3	NaN	NaN	2	cultural elderly
4	NaN	NaN	3	experience-driven elderly
5	ALTERSKATEGORIE_GROB	age classification through prename analysis	-1, 0	unknown
6	NaN	NaN	1	< 30 years
7	NaN	NaN	2	30 - 45 years
8	NaN	NaN	3	46 - 60 years
9	NaN	NaN	4	> 60 years

(b)Attributes data preview

Figure 2. Supplementary data previews

3.1.1. Understanding the Variables

Variables types are important as each type requires different approaches. For this reason, to understand the variables, I checked the data types. There are two columns having data type integer, 267 columns having float, and 3 columns having object data type. Binary coded object data types were recoded as 0 ,1 and the remaining ones converted to the dummy variables. Float data types, on the other hand required further investigation. The procedure for handling float data is as follows:

- Binary coded values recoded as 0 and 1.
- For the variables having more than two category, category meaning was checked. The variables who are ordered categorical or continuous were kept as is. The ones that are in the nominal scale, were converted to the dummy variables.
- There were some variables converted to the other ones like same measures having smaller/larger number of categories were dropped.

3.2. Handling Missing Data

Attributes data has missing/unknown/not possible category codes specific to each variable. Before working with missing data, first all these values were converted to numpy NAN values. The column-wise missing value distributions for general population and customer data was shown at figure 3.

As shown at figure 3, the median missing value percentage is 10% for general population while for customer data it is 27%. We also observe that there are columns having extreme amount (more than 50%) of missing data. Therefore, in data cleaning, I deleted the columns having more than 30% missing values

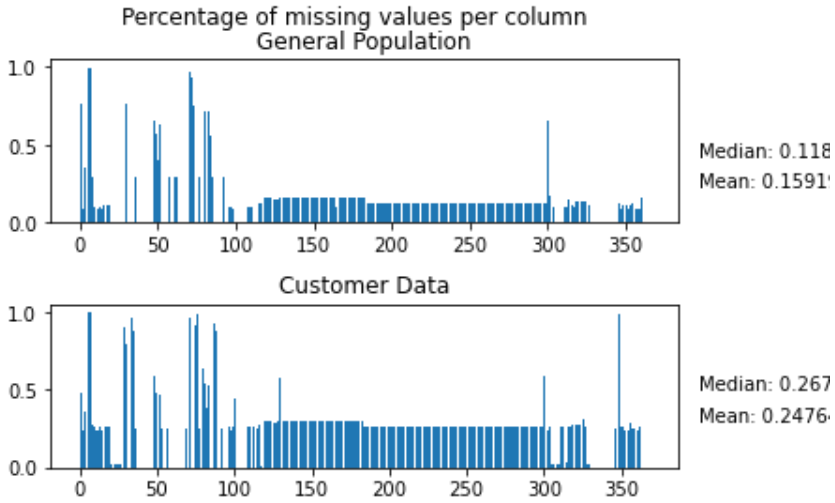


Figure 3. Colum-wise missing value distribution

The figure 4 shows row-wise missing value distributions sorted ascending for general population and customer data. The median missing value percentage is 0.7 % for general population while for customer data it is 0.06%. However, the graphs show that there are rows having extreme missing data. To protect data integrity, I dropped rows having more than 30% of missing value.

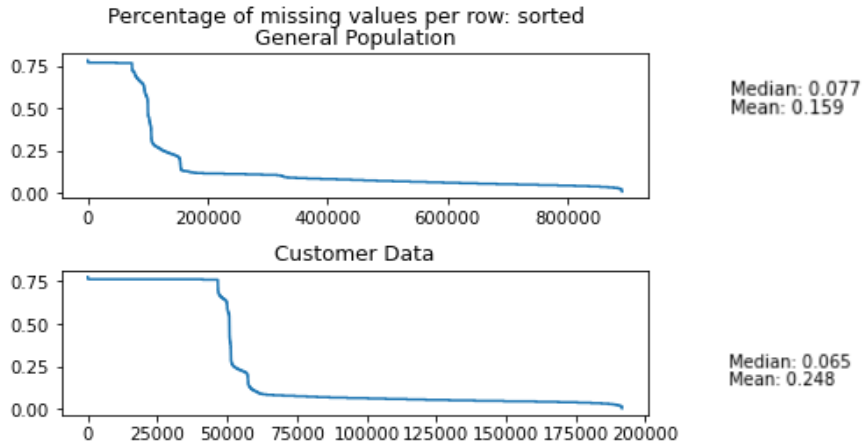


Figure 4. Row-wise missing value distribution

After dropping extreme missing columns and rows, I imputed median values for the remaining missing values. At the end of the data cleaning, imputation and one hot encodings, general population data consists of 784380 rows and 413 columns while customer data consists of 140310 rows and 404 columns.

4. CUSTOMER SEGMENTATION

4.1.Descriptive Analysis

Before clustering, I created some explanatory graphics to understand the distribution of demographics in general population and customer population. The customer and general population distribution according to gender, age and social status variables are shown in figure 5.

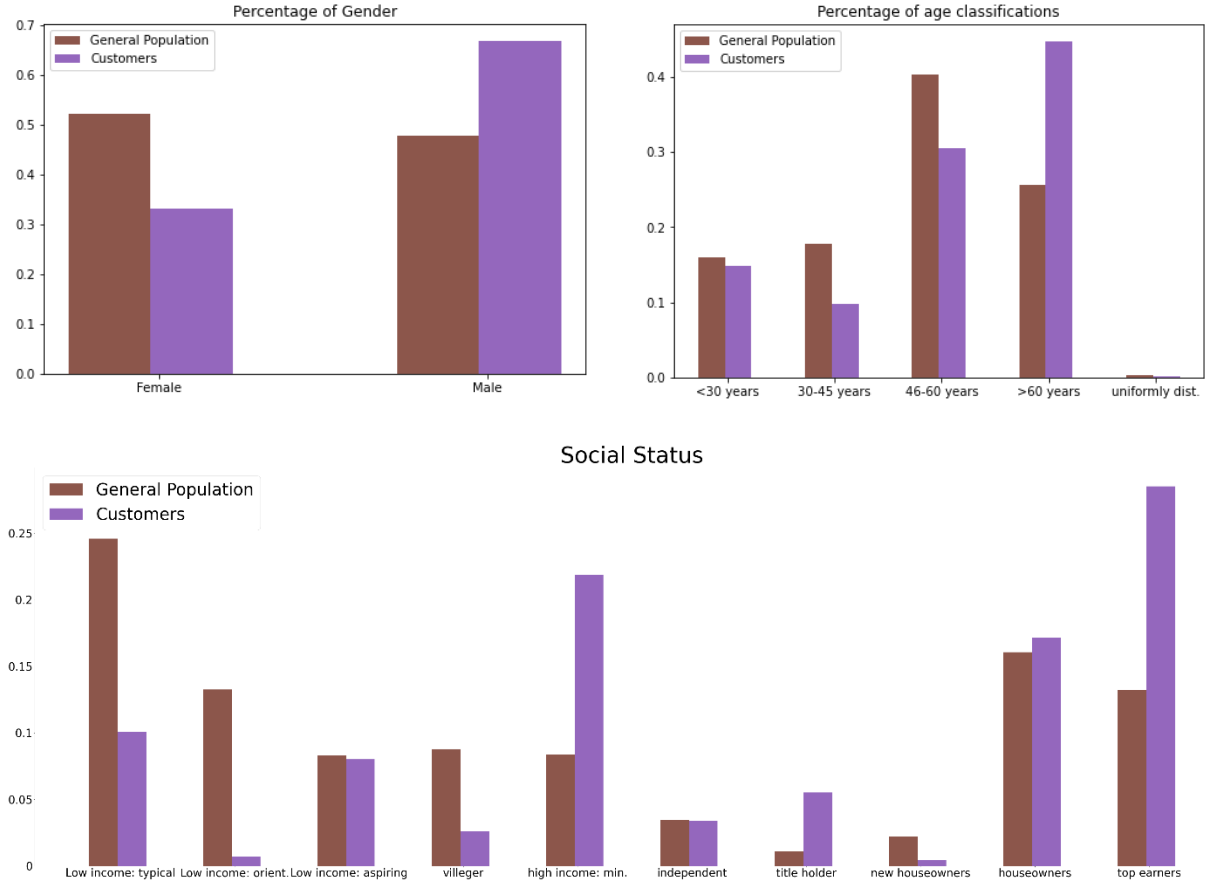


Figure 5. Distribution of demographics

As shown in figure 5, the percentage of males is higher in customer group than general population. Similarly, the percentage of people older than 60 years, high income and top earners is higher in costumer group than the general population.

4.2.Principle Component Analysis

To simplify data and reduce dimensions, PCA analysis was conducted. The number of components versus explained variance is shown at figure 6. I selected 250 components as it is the point having more than 90% variance explained.

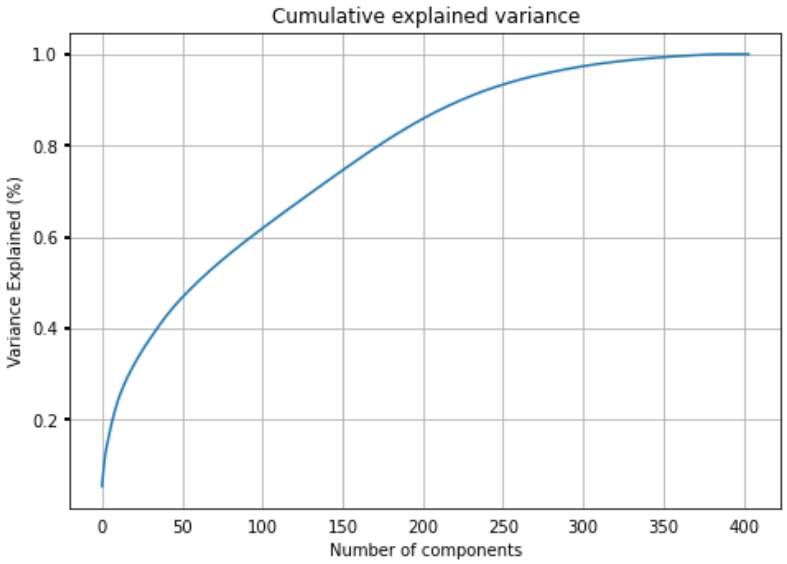


Figure 6. Cumulative explained variance vs number of components

4.3.Cluster Analysis

kMeans clustering was used as unsupervised learning model. To decide optimal number of cluster, I tried cluster size 2 to 70. The resulting model scores was shown in figure 7. From figure 7, I decided to use 10 clusters.

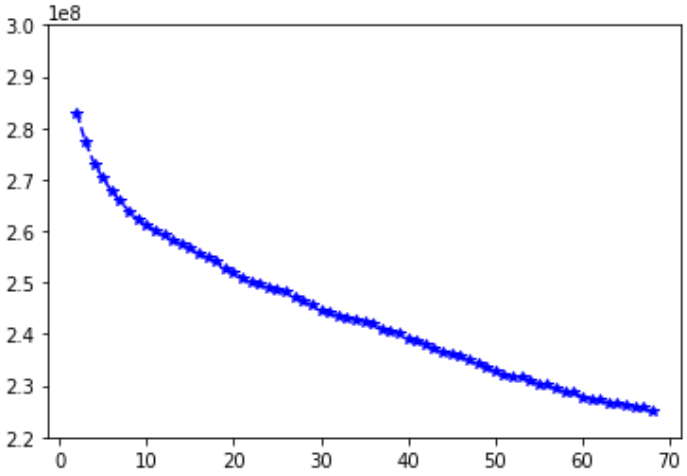


Figure 7. Cluster size versus model score

The percentages of each cluster in general population and customer population are shown in figure 8. As shown in the figure, the cluster percentages are higher in first, second, third and 10th cluster in customer population. For the remaining clusters, the percentages of customer population are lower than general population.

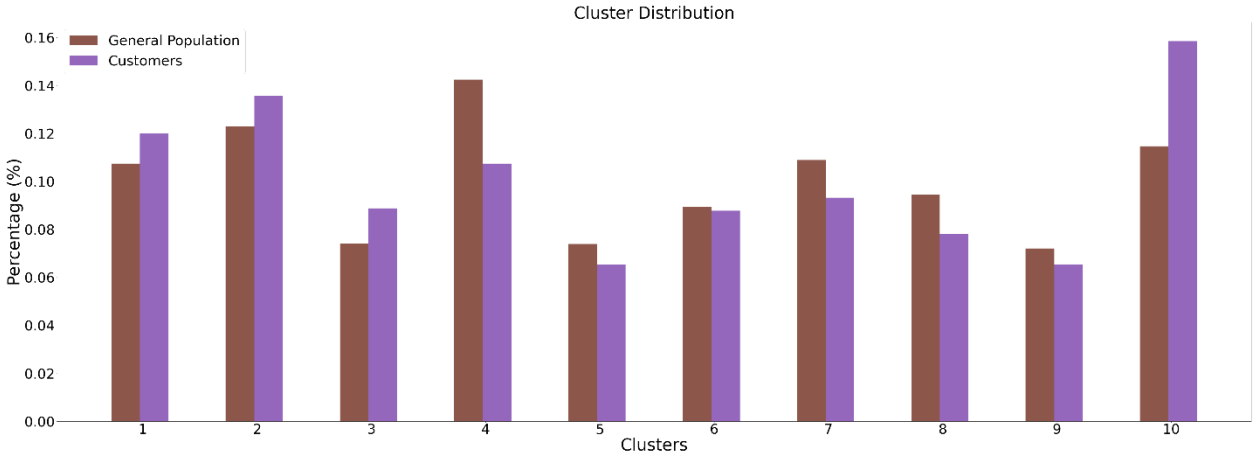


Figure 8. Cluster distribution: general population vs customer population

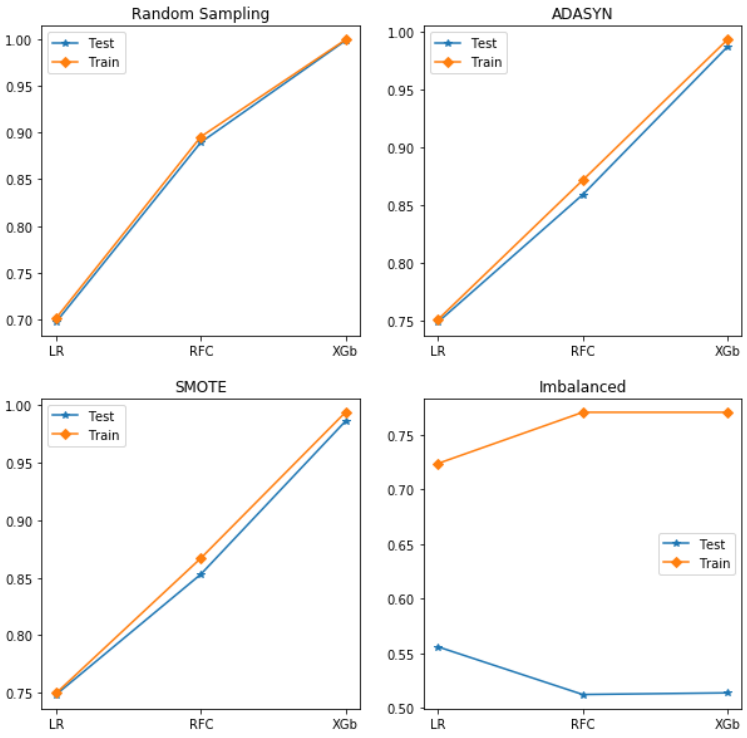
5. SUPERVISED LEARNING MODEL

For supervised learning model, we are given two separate data sets as training and testing data. For the Kaggle competition, testing data does not have a response column. To be able to test my models and I divided training data into test and train splits. I cleaned both training and testing data as explained in the data preprocessing section. Initially, training data consisted of 42962 rows and 367 columns. After cleaning, there were 34987 rows and 413 columns. Test data 42833 rows and 366 columns. After cleaning, there were 34980 rows and 413 columns.

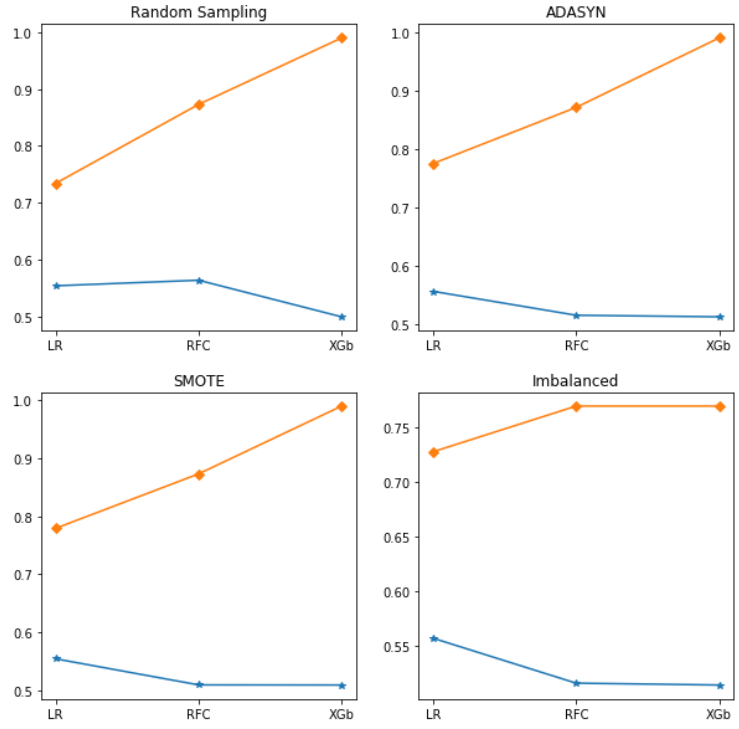
As we have an imbalanced class distribution, I used two different approaches to this situation: sampling and using cost sensitive training algorithms. For sampling strategy, I used 3 sampling methods : random over sampling, SMOTE and ADSYN and compared their performances with imbalanced data using three prediction methods: LR, RFC, XGboost. For each sampling method, I used two different perspectives to test the model integrity. First, I resampled the whole data and then train and the test the model. The results for resampled test and train sets are shown in figure 9.

Figure 9 (a) shows AUC values for each sampling method under different estimators. The results how that sampling techniques are outperformed the imbalanced data. For estimation methods, XGboost seem to be performed better.

My second approach to testing sampling methods was sampling training data testing model performance on imbalanced data. The reason I conducted this analysis is in the real analysis, I was supposed to use imbalanced data. The results of this analysis were reported at Figure 9 (b). Figure 9 (b) shows that sampling methods performed similarly with the algorithms trained with imbalanced data under testing with imbalanced data. Moreover, the AUC results were around 0.50. to increase model auc scores I performed parameter search and cross validation on RFC and XGboost models.



(a) testing with sampled data



(b) Testing with imbalanced data

Figure 9. ROCAUC values for testing with sampled data

For cross validation I used stratified shuffle split method with 5 fold. For RFC method, I tuned six parameters resulting fitting 5 folds for each of 360 candidates, total 1800 fits. For XGboost, I tuned five parameters resulting fitting 5 folds for each of 240 candidates, total 1200 fits. The results of the cross validation with grid search improved RFC's AUC score as 0.60 and XGboost's AUC score as 0.57. I continued to work with best RFC method as it is the one provided higher AUC score.

6. SUMMARY and FINAL NOTES

In this project, I tried to answer the problem 'How effectively can a mail order sales company expands its customer base?'. Within this problem, I used two perspectives: customer segmentation and supervised learning.

There were two major main challenges in this project. First one is understanding the data files, meanings of the variables and preprocessing. As the types of the variables affects how we behave them like keep as is, convert dummies, convert binary coding etc., I needed to understand the nature of the variables. To handle this problem, I divided variables according to types and number of categories, then dig deeper to understand their nature.

The second and most challenging part of this project was having highly imbalanced response categories. To find an approach works better in this situation has its own challenges. To overcome this problem I researched the possible ways to handle imbalance in data and used three different methods. Although, I managed to improve the performances of the predictions by grid search and validated them by grid search, there is still a room to improvement in this perspective as the final AUC score was 0.60.

Note: Under the terms and conditions, I only provided small screenshots of the data, instead of the providing the full data set.

License, terms and conditions:

In addition to Udacity's Terms of Use and other policies, your downloading and use of the AZ Direct GmbH data solely for use in the Unsupervised Learning and Bertelsmann Capstone projects are governed by the following additional terms and conditions. The big takeaways:

You agree to AZ Direct GmbH's General Terms provided below and that you only have the right to download and use the AZ Direct GmbH data solely to complete the data mining task which is part of the Unsupervised Learning and Bertelsmann Capstone projects for the Udacity Data Science Nanodegree program.

You are prohibited from using the AZ Direct GmbH data in any other context.

You are also required and hereby represent and warrant that you will delete any and all data you downloaded within 2 weeks after your completion of the Unsupervised Learning and Bertelsmann Capstone projects and the program.

If you do not agree to these additional terms, you will not be allowed to access the data for this project.

The full terms are provided in the workspace below. You will then be asked in the next workspace to agree to these terms before gaining access to the project, which you may also choose to download if you would like to read in full the terms.

These same exact terms are provided in the next workspace, where you will be asked to accept the terms prior to gaining access to the data.

The detailed terms and conditions can be found in the terms file in this repository.