Customer classification using machine learning

Shahid Gulzar Padder

Eötvös Loránd University, Budapest, Hungary
xpxskk@inf.elte.hu

Abstract. Machine learning techniques are the backbone of any data mining task. From data preprocessing to frequent pattern mining domain knowledge is really important to understand any dataset. Various data preprocessing techniques were deployed to make data feasible and understandable for machine learning algorithms to process it fast and in a right way. For clustering, after processing the data feature scaling followed by dimentionality reduction techniques were deployed. Classification was evaluated on three different models i.e; Logistic regression, KNN and Naive Bayes. A comparative study of different metrics was done.

Keywords: Data mining · Data analysis · Supervised learning · Unsupervised learning · K-means · Logistic regression

1 Introduction

Customer classification is one of the important aspect when it comes to gaint-firms/companies. Analyzing the customer data is really a challenging task as are databases are multidimensional which include numerous transaction and account records[1]. 'Contemporary marketing strategies' consider customers as principal resource to any enterprise. Thus it has become really important for any enterprise or firm to gather database in such a way that they are able to attract new customers and retain customers with great value. In order to achieve this target most of the enterprises collect and huge database which further can be analyzed and thus helps in developing new business tactics. Any enterprise cannot target all customers equally in terms of providing incentives/offers thus segregation of customers is really important to decide whether a customer is good or bad. Current study is focusing on clustering and classification of such customers using various unsupervised and supervised ML techniques.

Machine learning [2] is type of artificial intelligence which enables a particular system learn from data instead of explicit programming. In this study there is classification and clustering of good and bad customers based on provided data set. However, there is need of various steps in order to address this type of problem. The primary step towards any data mining problem is to understand the data which is very much important because sometimes not all data is sufficient/good in order to carry out next steps. Data prepossessing has really

helped to improve the overall performance of the models. In order to find importance and affecting criteria correlation heat maps were handy. On the other hand outliers can really affect the clustering as well as classification. Removal of the outliers in few attributes resulted in drastically predicting the bad customers. For **clustering Hierarchical agglomerative clustering** and **K-means** techniques were deployed to cluster the customers.textbfLogistic regression, KNN and Naive Bayes classifiers with hyper parameter tuning were implemented to classify the customers and a brief comparison of models was also done.

In section 2 a brief review of various techniques used in the current study is done. Also few related works are presented. Afterwards, the provided dataset is explored in which preprocessing, outliers and various techniques are performed in order to cluster the data in first part of section 3. Then training the models for different scenarios is done second part of section 3 and the results are either depicted graphically or in comparative tables in section 4. Lastly, a brief conclusion of the study with possible future improvements is discussed.

2 Background

On one hand Machine learning is guiding a computer program or any algorithm to improve progressively upon a given task. On the other hand if we view it on research basis machine learning can be depicted as mathematical and theoretical modelling of working process[3]. Machine learning can further be classified into three subcategories:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

In supervised learning input is in the form of both data and its labels during the training process which enables a machine to learn from the given input and thus producing an efficient model. However, in unsupervised learning there are no labels are provided and we have to apply any algorithm in order to cluster the input data according to the similarity measure. Moreover, we can also estimate the density of the data which is randomised. Reinforcement learning is the last type of machine learning in which agent interacts with the environment and based upon **SARSA**, **Q learning** or any other RL algorithm, thus learns from the errors and rewards.[4]

2.1 Unsupervised Learning

Clustering is one of the important and fundamental concept in data analysis and data mining. It is the process of aggregating similar objects into groups. It helps in framing a complete analytical solution. It becomes really easier for any analyst to recognise groups of similar objects. The two approaches used in this paper are:

- Hierarchical agglomerative clustering
- k-Means clustering

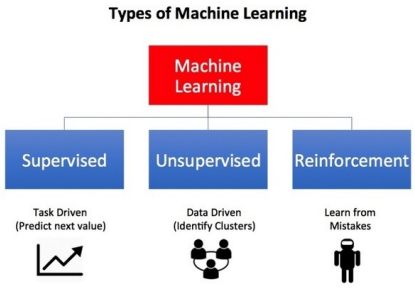

Fig. 1. Types of machine learning.

Hierarchical clustering is based on the distance d or similarity s measure. There are three types of linkage criterion:

– **Single linkage**

$$l(A,B) = min\left\{d(a,b)|a\epsilon A, b\epsilon B\right\} \tag{1}$$

– **Complete linkage**

$$l(A,B) = max\left\{d(a,b)|a\epsilon A, b\epsilon B\right\} \tag{2}$$

– **Average linkage**

$$l(A,B) = \frac{1}{|A||B|}\sum_{a\epsilon A}\sum_{b\epsilon B} d(a,b) \tag{3}$$

where A and B two points in diffused data and l is the linkage criterion. Heirarchical clustering starts by considering each point as a individual cluster. Then it looks for the closest and similar clusters and merges them together. This iterative process is continued until all clusters are merged together.

k-Means Clustering k-Means is another important clustering technique which falls under unsupervised learning category. It is a partitioning algorithm.[5]. k-Means partitions a given data into k mutually exclusive clusters. It is usually suitable for large sets of data.[6]. It treats each observation in any dataset as a object with a location in space. It devises a method in which objects in one cluster are as close as possible and far away from other clusters. Fig. 2 depicts the k-Means algorithm in simple language in order to understand its each step. k-Means clustering produces k clusters from a set of n objects, such that the squared error objective function is minimised.

$$E = \sum_{i=1}^{k}\sum p\epsilon C^i|p - m_i|^2 \tag{4}$$

The effectiveness of k-Means depends on accurate estimation of k clusters for each spectral data[7]

Algorithm 1 k-means algorithm

1: Specify the number k of clusters to assign.
2: Randomly initialize k centroids.
3: **repeat**
4: **expectation:** Assign each point to its closest centroid.
5: **maximization:** Compute the new centroid (mean) of each cluster.
6: **until** The centroid positions do not change.

Fig. 2. k-Means algorithm.

2.2 Supervised Learning

K-nearest Neighbours (KNN)

K-nearest Neighbour is one of the easiest classification algorithm to deploy but one the other hand it is a also known as lazy algorithm [8]. K is the hyper-parameter for this algorithm which indicates number of nearest neighbours to be considered for classifying a new instance. Fig. 3 shows the working of the KNN algorithm with nearest neighbours set to 3 and 6 respectively.

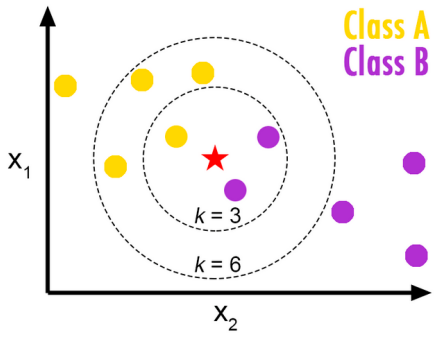

Fig. 3. K-nearest neighbour example.

Naive Bayes Classifier

Naive Bayes classifier is a classification algorithm based on Bayes' theorem. It is used for multi-class classification and binary class classification problems. Bayes' theorem gives a way to calculate probability of a section of data belonging to a particular class. The theorem is stated as below:

$$P(y/X) = \frac{P(X/y) \times P(Y)}{P(X)} \tag{5}$$

where P(X) is evidence of X, P(y) is the prior probability of y and P(y/X) is the probability of the class given the provided data, y is the class label and X is a dependent feature vector of size N.[4]

Logistic regression

Logistic regression is based on linear regression model which utilises a complex cost function σ which is called a sigmoid function. It is also known as logistic function thus giving rise to the name logistic regression. Logistic regression maps the class labels to probabilities. In other words if sigmoid function is applied on any real number will give ew value in the closed interval [0,1]. The sigmoid function is given below:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

Logistic regression uses a decision boundary. It is a fixed threshold value and based on this value the algorithm gives a class label to the probabilities calculated with the help of sigmoid (σ).

2.3 Related Works

- G. Chicco in [9] has used various clustering algorithms to classify customers according to the electricity consumption and clustered customers according to the different patterns. Author has used fuzzy k-means, along with hierarchical clustering and CCA dimension reduction tecnique.
- T.K Das in [10] used classification techniques to classify customers who can respond to the new company offers based upon their past purchasing history and trends. Author has used Naive Bayes, KNN and SVM to classify the customer data.
- Authors in [11] have used fuzzy k-means classification technique on the climatic data as an aid to faorest mapping in USA. Here they discuss how it could improved with GIS, k-means classification and spatial sampling.

3 Experiments

The main goal of this study is to cluster the customers according to the provided data. Firstly data preprocessing is really important. From finding of outliers to encoding of data according to its type was really important. Trying different feature subsets in order to improve clustering was really a challenging task. Classification of data and prediction of labels was carried out through different models and the results are compared.

3.1 Data specification

The dataset provided is a customer data in which various nominal and ordinal attributes are provided regarding the customers e.g, credit history, purpose of

credit, credit amount in euros, purpose of credit and some nominal features like purpose of credit, personal status etc. We have to predict the customer as good or bad according to the provided data.

3.2 Data preprocessing

– This step was challenging because it is the primary guiding principal for any data mining task. The provided data didn't had any null value thus the first step was easier. Checking frequency of the features was really important to understand the dataset. Three features highlighted in the description were checked followed by the description of the data which gave the various metrics e.g minimum value, maximum value, outliers, mean etc. Fig. 4 shows some of the ouliers that were processed in order to get better clustering and classification. It is an obvious fact that outliers effect the classification as well the clustering of the data

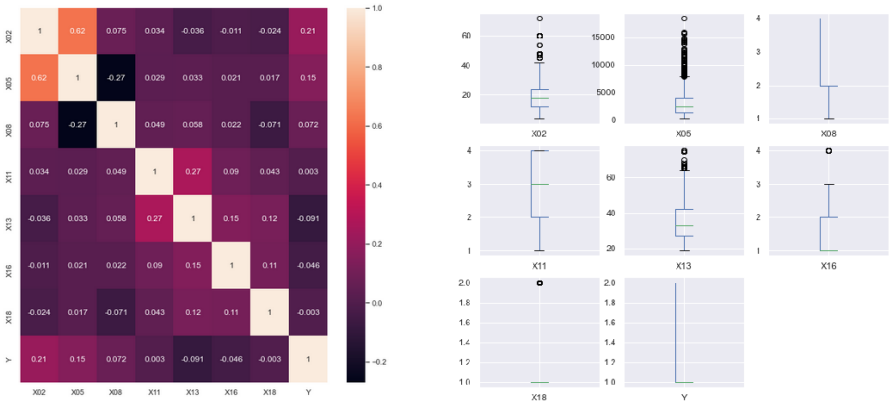

Fig. 4. Heat map and box plot for correlation and outlier detection.

– Encoding is really important for a model to understand the data but we need to be careful while encoding any data as it can end up in terrible classification or data loss. We again need to understand the the data type which is proven by the domain knowledge. For example any ordinal value should be taken care of because it might lead to data loss. Nominal values should not be label encoded as it is categorical data without any numerical value. A proper segregation of the data was done and accordingly the encoding was done. Again describing the dataframe was performed and correlation between each attribute were checked. Features with a weak correlation after evaluating the model were gain dropped. After getting a satisfactory behaviour of features feature scaling was the next step.

- Feature scaling is helpful to provide a common scale for numerical columns without losing information and range information. It helps gradient descent to converge faster and moreover KNN and k-Means use euclidean distance to calculate distance between two points.In this paper I tried both normalization as well as standardization but promising results were provided by standardization technique. The ranges for this methods for scaling are between [-1,1].
- Since our dataset after the above experiments has still more features. Either we need to visualize it in multilple dimensions or we use any dimensionality reduction technique. In the current paper Principal Component Analysis was used to reduce the dimensionality. Reducing the the number of variables in a data of-course leads to reduction in accuracy but a little trade for accuracy leads to easily understandable data and also easier for machine learning algorithms to process.
- After dimensionality reduction is done the next step was to use k-Means algorithm to cluster the processed dataset. We have only two labels in our dataset so the the number of desired clusters was kept **2**. It was ran ten times independently for different centroids so that final model with lowest sum of squared error is chosen. Maximum number of iteration were **300**. While evaluating with external evaluation methods a Precision of **79 percent** for label 1 was achieved and for label 2 precision of 67 percent was achieved and accuracy of **72 percent** was achieved. It was not that good but various approaches like **binning**, **hyper-parameter optimisation**, **correlation matrix approach** was done but the results were not improved. Fig. 4 shows the results

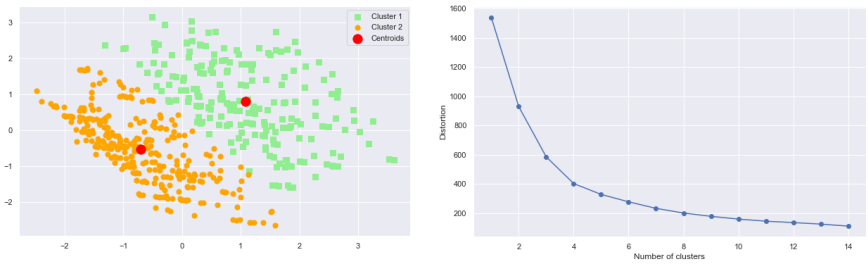

Fig. 5. Clustering and internal evaluation using k-Means algorithm.

- Similarly, for Hierarchical agglomerative clustering different approaches were tried including single, complete, average and ward linkage but results from ward linkage were promising. Again, the results were not that promising for label 2 after implementing various techniques and approaches.
- Last step was to split the data into training and test sets in order to perform supervised learning or classification part. In this paper a comparative

analysis of three different classifiers has been done to check the classification report of each one and thus optimising the hyper parameters to get better precision, accuracy, recall and other report metrics.

4 Results

After applying the three different classifiers the best performing was logistic regression for both the labels. For label 1 it was able to achieve 72 % precision and for label 2 the precision was 70 % . Overall accuracy was 72 %. F1- score for label 1 was 82 % and for label 2 it was around 27 %. On the other hand KNN For label 1 was able to achieve 76 % precision and for label 2 the precision was 38 % . Overall accuracy was 72 %. F1- score for label 1 was 83 % and for label 2 it was around 19 %. Finally Naive bayes classifier for label 1 it was able to achieve 75 % precision and for label 2 the precision was 50 % . Overall accuracy was 74 %. F1- score for label 1 was 85 % and for label 2 it was around 13 %. All the results are tabulated below in **Table 1**

4.1 Frequent pattern mining

Frequent pattern pattern mining is a type of analytical process which finds frequent pattern, relations or causal structure from a dataset. In this process there are rules to predict presence of one specific item based on the occurence of other items in any dataset. Various algorithms such as **Apriori, Eclat** are used to find some association between the the data which occurs frequently. In this study I haven't deployed the same algorithms from analysing the data, I have found two frequent patterns.

- Not all but most of the customers who have more than 11000 EUR credit amount are labelled as bad.
- Most of the customers above the age of 36 are liable to pay maintenance for two people.

These two patterns were not based on the algorithm but on some data understanding and some analysis. But we can deploy frequent pattern mining to classify customers upon their various features which can help in detection of bad customers before offering any credit to them from banks or any other enterprises. Moreover, enterprises can handle valuable or prestigious customers using frequent pattern mining techniques.

Table 1. Summary of the results for different classifiers

Model	Pr.label 1	Pr.label 2	Accuracy	F1-label 1	F1-label 2
Logistic Regression	72 %	70 %	72 %	82 %	27 %
Naive Bayes	75 %	50 %	74 %	85 %	13 %
KNN	76 %	38 %	72 %	83 %	19 %

The confusion matrix was also also used such that different metrics can be extracted from it. We know that for binary classification it has four entries. True positives(TP) and true negatives(TN) are along the diagonal which hold true prediction while as the off diagonal has False positives(FP) and false negatives(FN) which hold wrong predictions done by the classifier. Accuracy, Precision, Recall and F1-score are calculated from the above values.

Table 2. Results of the confusion matrix for different classsifiers

Model	TP	FN	FP	TN
Logistic Regression	86	3	34	7
KNN	108	8	35	5
Naive Bayes	113	3	37	3

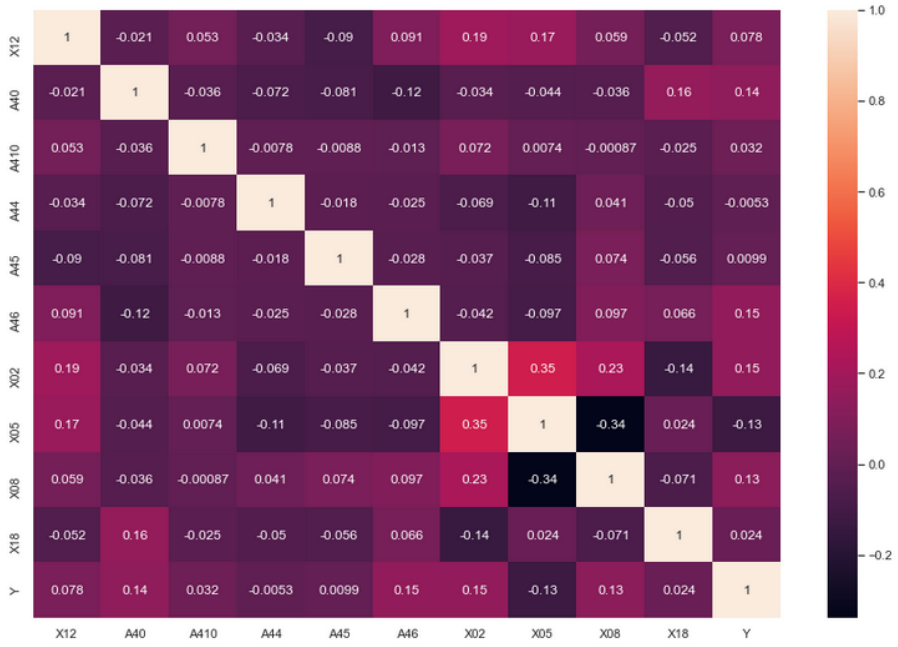

Fig. 6. Heat map showing correlation between various features.

5 Conclusion

Various Data analytic techniques were deployed during the data preprocessing. Feature extraction to frequent pattern mining it was a challenging task. Clustering of dataset was somehow satisfactory while in future goal will be to try different advanced and better algorithms. On the other hand logistic regression performed better than the other two classifiers. In future various other algorithms like SVM and other deep learning algorithms can be tried to check variuos metrics on the same dataset.

References

1. N.-C. Hsieh, "An integrated data mining and behavioral scoring model for analyzing bank customers," *Expert systems with applications*, vol. 27, no. 4, pp. 623–633, 2004.
2. E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.
3. H. Heidenreich, "What are the types of machine learning?" *https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f*, 2018.
4. T. Ahmad and F. Zakarya, "Supervised learning methods for skin segmentation classification," *Central-European Journal of New Technologies in Research, Education and Practice*, 2020.
5. S. Adhau, R. Moharil, and P. Adhau, "K-means clustering technique applied to availability of micro hydro power," *Sustainable Energy Technologies and Assessments*, vol. 8, pp. 191–201, 2014.
6. I. G. Costa, F. d. A. de Carvalho, and M. C. de Souto, "Comparative analysis of clustering methods for gene expression time course data," *Genetics and Molecular Biology*, vol. 27, no. 4, pp. 623–631, 2004.
7. K. Koonsanit, C. Jaruskulchai, and A. Eiumnoh, "Determination of the initialization number of clusters in k-means clustering application using co-occurrence statistics techniques for multispectral satellite imagery," *International Journal of Information and Electronics Engineering*, vol. 2, no. 5, pp. 785–789, 2012.
8. P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," in *Emerging technology in modelling and graphics*. Springer, 2020, pp. 99–111.
9. G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Transactions on power systems*, vol. 21, no. 2, pp. 933–940, 2006.
10. T. Das, "A customer classification prediction model based on machine learning techniques," in *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE, 2015, pp. 321–326.
11. P. A. Burrough, J. P. Wilson, P. F. Van Gaans, and A. J. Hansen, "Fuzzy k-means classification of topo-climatic data as an aid to forest mapping in the greater yellowstone area, usa," *Landscape ecology*, vol. 16, no. 6, pp. 523–546, 2001.