Title: **Course Project**



Course: Analytics for Business (A220A0752)

Group: Ayushma Rayamajhi

Pablo Garcia Hurtado

Rodrigo da Matta

Sergey Zakrytnoy

ANALYSIS OF SALES OF A RETAIL STORE ON BLACK FRIDAY

# INTRODUCTION

Marketing analysts often investigate differences between groups of people. Do a gender subscribe to a service at a higher rate? Which demographic area can be more receptive to such product? Does the product entice more youngsters or middle agers? The answers to these questions aid businesses to understand the market, to target customers effectively, and to evaluate the outcome of marketing activities.

In this study, we examine the kinds of comparisons that often arise in marketing, with a dataset of the transactions made in a retail store during Black Friday. From that, we illustrate a consumer segmentation project.

# OBJECTIVE

The objective of this study can be summarized in three points:

* Identify consumer buying patterns in order to create consumer segmentation;
* Support on creating and delivering specialized marketing strategies so as to increase sales;
* Screening products with a higher turnover.

## MOTIVATION

According to Carpenter, Bauer, & Erdogan (2010), greater personalization and customization is one of the major challenges of a changing world. Thus, it is crucial to understand customer behavior and categorize customers based on their demography and buying behavior. This is the so-called customer segmentation, or marketing clustering, that allows companies to better tailor their marketing efforts to various audience subsets in terms of promotional, marketing and product development strategies.

For this resort, companies already realized that Data Science can help to equip themselves for these new challenges: “*58% of enterprises are tackling the most challenging marketing problems with AI and machine learning first, prioritizing personalized customer care, new product development*.” (Capgemini, 2019).

# EXPECTED OUTCOMES

The expected outcomes of this study are two, namely:

* Clustering of the customers based on the available variables to better understand customer purchase behavior against different groups, i.e., consumer segmentation;
* Classification of dataset to make data-based prediction for direct marketing processes.

# DATA

The used dataset comprises 12 individual variables, both numerical as well as categorical data, such as: gender, age, occupation, city, years of residence, marital status, product category, purchase value, and so on.

The original dataset includes 500.000 observations, each row representing a purchase of one item.

We performed several procedures to find descriptive summaries by groups, and then visualize the data in several ways …

# METHODOLOGY

## CLUSTERING

Businesses often view segmentation as discovering groups in the data in order to derive new insight about customers. If one uses clustering to discover groups, it is consequential the desire to predict or classify future responses into those groups.

There are hundreds of available methods for clustering, but basically, they subdivide in two categories: (1) model-based clustering methods, and (2) distance-based clustering methods.

Model-based clustering methods view the data as a mixture of groups sampled from different distributions. The assumption is that original distribution and group membership is unknown. Model-based clustering methods attempt to model the data in such a way that the observed variance can be best represented by a small number of groups with specific distribution characteristics (different means and standard deviations).

Distance-based clustering methods on other hand, attempt to find groups by minimizing the distance between members within a group, while maximizing the distance of members from other groups. There are many algorithms in this category. To name a few:

* k-means
* Hierarchical clustering
* k-means++
* Bisectional k-means
* K-medoids
* K-modes

In many applications with large numbers of observations, k-means is the preferred algorithm due to its simplicity, linear time complexity with respect to the number of objects, and facility of use (there are many libraries and functions available for many programming languages. Nevertheless, it has some disadvantages (Wierzchoń & Kłopotek, 2018):

1. The result depends on the order, in which the data is processed and on the value range (scale) of individual components of the vector, describing the objects.
2. It is a greedy algorithm the result of which depends on initial conditions.
3. It is sensitive to the presence of abnormal observations (outliers).
4. The number of clusters must be known in advance.
5. It can be used only for analysis of numerical data.

The use of alternative methods to solve those issues has been increasing in various domains. Sheikha, Ghanbarpourb, and Gholamiangonabadib (2019) used the k-means++ algorithm for customer segmentation and behavior analysis among Fintech companies. The crucial idea of the k-means++ method is the preselection of observations taking care of the fact that if the dataset contains at least k outliers, then the choice of the most distant observations as proposed by Gonzalez (1985) will result in malformed initial centers. Hence, the algorithm uses a probabilistic variant of the selection rule.

IBM Corporation implemented a procedure referred to as Two Step Clustering (SPSS, 2001). It consists of running a partitioning procedure, like k-means, followed by a hierarchical procedure. The procedure has been used in a wide variety of application areas like identifying and characterizing potential electric vehicle adopters (Mohamed, Higgins, Ferguson, & Kanaroglou, 2016). The choice of the number of clusters is not crucial because the primary aim of the first step is to reduce the size of the dataset by retaining only one representative members of each cluster. Therefore, it mitigates the concern of deciding how many cluster to start with.

## CLUSTER ANALYSIS

The objective with the cluster analysis is to find patterns related to the users during the Black Friday. As the exploratory analysis have shown, there are some strong variables such as city category and product category that mark an initial pattern.

We have decided to engineer new features based on the original data, for instance, we want to know how much the preferred categories weight is in the total basket, by preferred we mean how much is the user willing to pay for a product compare to the other categories.

|  |  |
| --- | --- |
|  |  |
| **Figure 1** – |  |

As we can see, the number of products and the purchase value have similar behavior, that represent have of the total purchase. With this in mind, we want to validate if the category preference determines a pattern in the user shopping attitude.

Correspondingly, we first need to normalize our data and specify the distance method to use.

We ran several combinations of normalization, distance and set of variables; the more relevant results based on the Silhouette index are the following:

|  |  |  |  |
| --- | --- | --- | --- |
| All data |  |  |  |
| No purchase amount variables |  |  |  |
| No preferred categories |  |  |  |
| No gender, marital status, no ratio preference |  |  |  |

**NOTE:** one cluster must be added to all charts.

**Figure 2** –

The Silhouette index is a measure of similarity among points in a cluster compared to the other clusters. The index has a range between -1 and 1, the higher the value, the better the match to a cluster. The usefulness of the Silhouette method to evaluate the clusters is the flexibility to use any distance.

Based on the results, the better cluster method is to not use gender, marital status, ratio preference and second category preference, keeping age, city, amount of purchase and preferred category. Accordingly, the Silhouette index scores 0.425 using cosine as a distance method. It is important to notice that the second preferable category was weighting strongly the clustering though with lower silhouette score of 0.238.

## CLUSTER ASSIGNMENT

## CENTROIDS





 

**Table 1** -

In essence, the user types can be describe by strong variables and minor attributes that will give a further horizon for the store to take advantage of.

|  |  |  |
| --- | --- | --- |
| **User** | **Strong Attributes** | **Strategy** |
| Type 1 | * Lives in city B * Its preferred product category is 1 * Likes to buy large quantities and more than half (59%) is represented by product category 1 * In addition to quantities, also is willing to spend a higher amount than the rest of users.   Minor attributes:   * Commonly a male young adult between 18 to 35 * General preferences for a complementary product | * Postal code marketing segmentation * Probably roundabouts high class city districts * Direct announcement of new top tier brands products, particularly of category 1, and accessories * Have a coverage in young adult male communities * Probably a good credit limit offer will be increase customer satisfaction and loyalty * Prime service (delivery) |
| Type 2 | * Lives in city C * It has a preference for products of category 8 and a small tendency to other type of products * Its preferred category covers more than 60% of its purchase * It likes to buy small quantities and does not spend as much   Minor attributes:   * Could be a woman or a man older than 26 and probably married | * Postal code marketing segmentation * Direct advertising of category 8 products * Average priced products marketing will cause an interest * Campaigns should be gender neutral and oriented to a more mature target but still young, probably small families * Offer regular delivery service |
| Type 3 | * Also lives in city C * Definitely prefers product category 1 * Also favors small quantities * However, is willing to spend more on its preferable category than the rest, more than 69% of purchase value   Minor attributes:   * Probably a man older than 26 * General preferences for a complementary product | * Postal code marketing segmentation * Direct advertising of category 1 products * Could be mixed with not expensive accessories * Reach young adult males communities |
| Type 4 | * Lives in city C as well * Is incline towards products of category 5 * Likes bargains, is not willing to spend that much, almost 25% of Type 1 user purchase * Small quantities, however, most of the purchase is from category 5, more than 71% | * Postal code marketing segmentation * Offer discounts and outlet promotions of product belonging to category 5 * Could be complementary to offer other categories as well * Gender neutral and all around ages campaigns |
| Type 5 | * Lives in city A * It does not have a major preference category, however is willing to higher prices * And also, large quantities   Minor attributes:   * It is most likely to be a man between 26 to 45 years old * Would rather buy products of category 1 but it is open to other alternatives | * Postal code marketing segmentation * Probably roundabouts high class city districts * Direct announcement of new top tier brands products of categories 1, 5 or 8 * Offer prime services * Acknowledge mature man events (yacht, golf, cars) |
| Type 6 | * Lives in city B * This user is willing to pay slightly more than the cheapest products but is not fond to buy expensive products   Minor attributes:   * Could be a woman or man older than 26 years * Would rather buy products of category 5 and 8 and it is open to other alternatives | * Postal code marketing segmentation * Offer time limit discount * Small credit could increase user purchases * Regular service * Gender neutral campaigns focus on product categories 5 and 8 * Have a slightly mature tone in online communications |

**Table 2** -

Cluster analysis goes beyond interpreting customer behavior, it can give a full road map of the customer journey during its multiple interactions with the company. Besides understanding each group, it can evolve to a brand loyalty goal; customers can change styles, tastes and their willing to spend in valuable things. The closer the firm interpret attitudes, the more satisfied their customers will be forming strong and long-lasting relationship.

## CLASSIFICATION

With customer segmentation being the main focus of our research, we also perform classification analysis to predict what product category a customer is most likely to be interested in. Such information can be used in internal recommendation system to take full advantage of the long-tail phenomenon, i.e. opportunity to reach out to the most outnumbered customer groups with personal offers.

In the course of the analysis, we use the variables that were derived in the clustering part, i.e. general information about a customer such as Gender, Marital status, City and Age, as well as the extracted features related to their consumer behavior, including number of products in the basket, preferred product categories etc. The complete set of variables in the processed normalized data is presented in Table 3 below. Please note, that categorical variables are already transformed into dummies.

|  |  |
| --- | --- |
| 0 Gender  1 Marital\_Status  2 oneProduct  3 prefCount  4 totCount  5 preftotCount  6 prefPurchase  7 totPurchase  8 preftotPurchase  9 Age\_0\_17  10 Age\_18\_25  11 Age\_26\_35  12 Age\_36\_45  13 Age\_46\_50  14 Age\_51\_55  15 Age\_55gt  16 City\_Category\_A | 17 City\_Category\_B  18 City\_Category\_C  19 CityYears\_0  20 CityYears\_1  21 CityYears\_2  22 CityYears\_3  23 CityYears\_4gt  24 prefCat1\_1  25 prefCat1\_5  26 prefCat1\_8  27 prefCat1\_19  28 prefCat2\_1  29 prefCat2\_5  30 prefCat2\_8  31 prefCat2\_19 |

**Table 3 –**

During the research, we have experimented with multiple models and target variables seeking business insight from the data. Those models include KNN (K-Nearest-Neighbors), Decision Trees, Random Forest, Artificial Neural Networks with fully connected dense layers. Finally, we ended up solving the following problem:

Based on the general information available about the customer and current products in the basket (assume known second preferred product category), predict the product category that he/she is most likely to buy. As mentioned earlier, such model would serve as predictive driver of user-based recommendation system for real-time personalized offers in e-store.

Our hypothesis was that the social profile of a customer alongside with information on his personal buying patterns lays the foundation for reliable prediction of preferred product category that can be used for marketing purposes.

To perform such classification, we separate variables prefCat1\_X (id 24 – 27) from the rest of the dataset. Then, we split the predictors and the labels into randomly shuffled train and test subsets in proportion of 60%/40%.

Among the approaches listed earlier, Decision Tree classifier proved most accurate and robust. Decision trees, or classification trees and regression trees, predict responses to data. To predict a response, we follow the decisions in the tree from the root (beginning) node down to a leaf node. The leaf node contains the response. Classification trees give responses that are nominal, such as 'true' or 'false'[[1]](#footnote-1), which is convenient and efficient when working with categorical/discrete valued variables.

In our case, we managed to fit a Decision Tree classifier that has the capacity to predict preferred product category among the most popular ones with accuracy of 79-81% on test subset (variance due to random split). This is better than most of the predictive risk models that are being used in industry, which makes us confirm the tested hypothesis and prove the applicability of the approach in practice.

Looking into the model, we can compare the importance of different predictors in the decision-making process. The estimate of predictor importance for tree is the sum of changes in the risk due to splits on every predictor divided by the number of branch nodes[[2]](#footnote-2).



**Figure 3 –**

We can see that almost all of the 28 predictors are involved in the classification process, i.e. have their representation in nodes and logical connections between them. We also conclude that due to global unification of purchase processes both online and in large centers, such features as age of the customer and his/her city of residence bear relatively low importance in their consumer patterns. Instead, the predictive process becomes recursive and in a way autoregressive, relying primarily on the other products in the basket and the observed shipping habits of a customer.

# RESULTS AND DISCUSSION

*Lorem Ipsum*

# CONCLUSION

*Lorem Ipsum*

# REFERENCES

Carpenter, M., Bauer, T., Erdogan, B. (2010). Management Principles 1.1. Mountain View, CA: Creative Commons.

Dagdoug, M. (2018, July 25). Black Friday. Retrieved from: <https://www.kaggle.com/mehdidag/black-friday>

Gonzalez, T. F. (1985). Clustering to minimize the maximum intercluster distance. Theoretical Computer Science, 38, 293-306. doi:10.1016/0304-3975(85)90224-5

Mohamed, M., Higgins, C., Ferguson, M., & Kanaroglou, P. (2016). Identifying and characterizing potential electric vehicle adopters in Canada: A two-stage modelling approach. Transport Policy, 52, 100-112. doi:10.1016/j.tranpol.2016.07.006

SPSS. (2001). White paper - technical report: The SPSS Two Step Cluster Component - A scalable component enabling more efﬁcient customer segmentation.

Turning AI into concrete value: The successful implementers' toolkit. (2019, January 28). Retrieved from https://www.capgemini.com/resources/turning-ai-into-concrete-value/

Wierzchoń, S. T., & Kłopotek, M. (2018). Modern algorithms of cluster analysis. Cham: Springer.

1. <https://in.mathworks.com/help/stats/decision-trees.html> [↑](#footnote-ref-1)
2. <https://in.mathworks.com/help/stats/compactclassificationtree.predictorimportance.html> [↑](#footnote-ref-2)