Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 June 2021 doi:10.20944/preprints202106.0118.v1

Working paper.

Customer Churn - Prevention Model – Unsupervised Classification

Julian Eduardo Grijalba Facundo [1]

[1] julian@here4data.com;juliangrijalba@gmail.com

Abstract: The strategy of any organization is based on the growth of its customer base, and one of its principles is that selling a product to an existing customer is much more profitable than acquiring a new customer. However, this approach has several opportunities for improvement, since it usually has a totally reactive approach, which does not give opportunity to the areas specialized in customer experience and recovery, to give an effective response for that moment, since the customer is gone at the time of the intervention. This happens because usually a diagnostic analysis of customers who have stopped buying products or services in a defined period, commonly three (3) periods or months, is performed. This paper challenges the way to face this problem, and proposes the development of a complete solution, which does not focus exclusively on the prediction of churn, as is usually done in the state of the art research, but to intervene in different interactions that can be carried out with customers. The above focused not only to prevent customer churn, but to generate an added value of continuous improvement in sales processes, increase customer penetration, leading to an improvement in customer experience and consequently, an increase in customer loyalty.

Dataset License: CC BY-NC

Keywords: Churn 1; customer churn; customer segmentation; churn prevention; predictive churn model; recommendation system engine.

1. Summary

The strategy of any organization is based on the growth of its customer base, and one of its principles is that selling a product to an existing customer is much more profitable than acquiring a new customer. It is not surprising that companies pay close attention to the analysis and impact of churn on their business strategies. However, this approach has several opportunities for improvement, since it usually has a totally reactive approach, which does not give opportunity to the areas specialized in customer experience and recovery, to give an effective response for that moment, since the customer is gone at the time of the intervention. This happens because usually a diagnostic analysis of customers who have stopped buying products or services in a defined period, commonly three (3) periods or months, is performed.

The focus of this research is how different concepts and techniques related to the artificial intelligence sub-branch can possibly address a mixed solution, not only from the data perspective but also integrating it with the business approach. This second research support the experiments of including Unsupervised Algorithms with the goal of having a churn classification by customer grouping

© 2021 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 June 2021 doi:10.20944/preprints202106.0118.v1

2. Data Description (required)

2.1. Transactional data model

Data is divided into multiple data sets for better understanding and organization, data flow is visualized in transactional information:

2.1.1. Customer Data

This dataset contains information about the customer and their location. It allows to identify unique customers in the order dataset and to find the order delivery location.

2.1.2. Geolocation data

This dataset contains information on Brazilian postal codes and their lat/long coordinates.

2.1.3. Order item data

This dataset includes data about the items purchased within each order. Example: Order_id = 00143d0f86d6fbd9f9b38ab440ac16f5 has 3 items (same product). Each item has the load calculated according to its measurements and weight.

2.1.4. Payment data

This dataset includes data on order payment options.

2.1.5. Order data

This is the central data set. For each order you can find all other information.

2.1.6. Product data

This dataset includes data about the products sold by Olist.

2.1.7. Vendor data

This dataset translates the product_category_name into English.

2.1.8. Category name translation data

This dataset contains information on Brazilian postal codes and their lat/long coordinates.

More details about the data is described into Customer Churn Prevention – Recommendation System

2.2. Experiments

The experiments has been designed with an unsupervised algorithm (machine learning) with the objective of generating customer segmentation and thus providing the business with new information tools to support the strategies and monitoring of results that may be proposed.

2.1.1. Elbow Method

The first step is to define the number of clusters, we start by performing the elbow method.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 June 2021 doi:10.20944/preprints202106.0118.v1

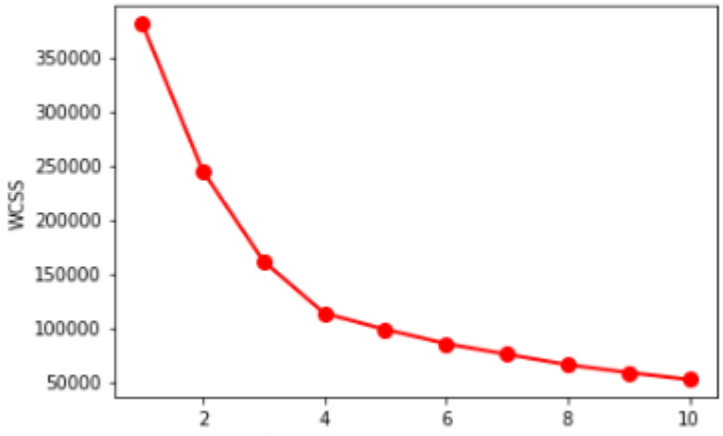

Figure 1. Elbow Method

2.1.1. Silhouette Method

The This method is applied to ensure that the clustering defined as k=4 is the best for the K-means algorithm. With this information, it can be concluded that four (4) clusters are the most suitable for grouping customers.

```
Scoring silueta para 3 Clusters: 0.3563
Scoring silueta para 4 Clusters: 0.3698
Scoring silueta para 5 Clusters: 0.3400
Scoring silueta para 6 Clusters: 0.3467
Scoring silueta para 7 Clusters: 0.3305
Scoring silueta para 8 Clusters: 0.3502
```

Figure 2. Scoring Silhouette

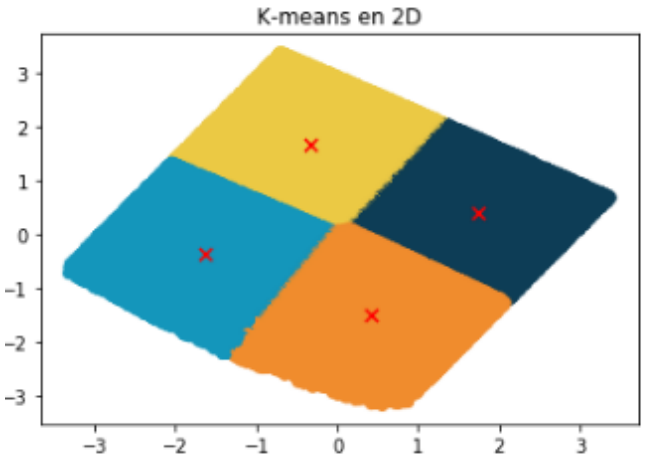

Figure 3. Clusters

A correlation matrix is performed to understand the cohesion between the data and the new segment dimension included.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 June 2021 doi:10.20944/preprints202106.0118.v1

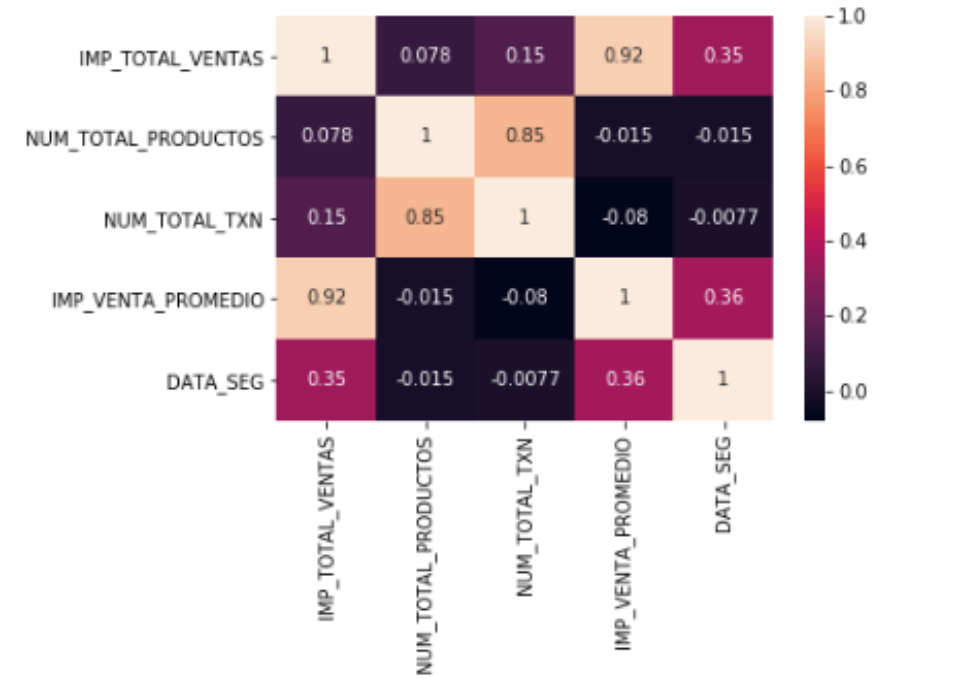

Figure 4. Correlation Matrix

It also describes how the variables with the strongest correlations are grouped.

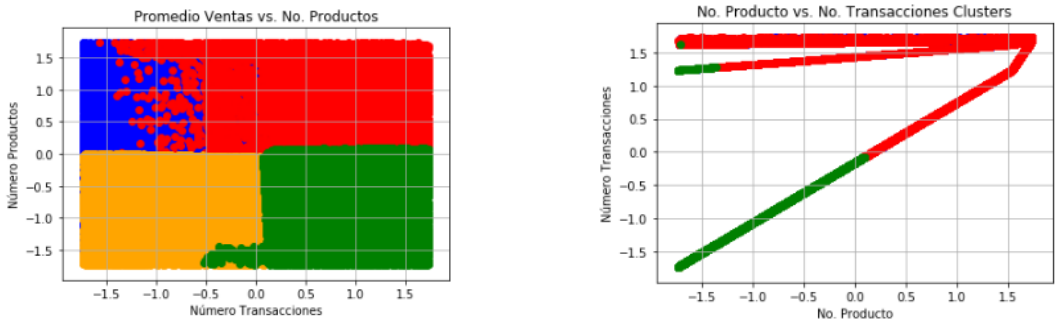

Figure 5. Relation between variables

3. Methods (required)

For research purposes, a public dataset has been used so that it can be improved by other researchers. Similarly, the following parameters were defined for the data set used for this model.

- **Data completeness:** Data with an acceptable amount of null information for analysis, less than or equal to 5% of the total records.
- **Churn Flags:** Should not have churn marking, due to the need to test the marking model.
- **Data volume:** Information greater than 100,000 transactions.
- **Data source:** Real information, not pre-created by software vendors or from courses generated by universities.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 June 2021 doi:10.20944/preprints202106.0118.v1

Upon completion of this review, the public domain data selected is Brazilian E-Commerce Public Dataset by Olist. Retrieved June 10, 19 from https://www.kaggle.com/olistbr/brazilian-ecommerce, and described as "This dataset was generously provided by Olist, the largest department store in the Brazilian markets. Olist connects small businesses throughout Brazil with seamless, single-contract channels. These merchants can sell their products through the Olist Store and ship directly to customers using Olist's logistics partners. See more on our website: www.olist.com."

4. Conclusions

- The use of the unsupervised model allows an important data input to a subsequent model, however, by itself, it does not guarantee a churn classification.
- A form of proactive customer classification will need to be integrated to ensure that Churn marking enables the business to take action.
- False correlations are identified, due to the fact that one metric is derived from another, it is necessary to eliminate them from the model for future experiments.
- The development of a model that unifies recommender systems, unsupervised classification of churn customers and a proactive aspect of tagging could be a new approach to e-commerce research.

References

1. Rose R. y Pulizzi J. (2017). Killing Marketing: How Innovative Businesses Are Turning Marketing Cost Into Profit. (chap. 1-5).
2. Bingquan Huang1(B) , Ying Huang1, Chongcheng Chen2, and M.-T. Kechadi1 (2016). Deep Learning in Customer Churn Prediction: Unsupervised Feature Learning on Abstract Company Independent Feature Vectors.
3. A Adnan Amin, Feras Al-Obeidat, Babar Shah, Awais Adnan, Jonathan Loo, Sajid Anwaren (2019). Customer churn prediction in telecommunication industry using data certainty.