 doi:10.20944/preprints202106.0119.v1

Working paper.

Customer Churn - Prevention Model – Recomendation Engine

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 thesis work 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.

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..

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 June 2021 doi:10.20944/preprints202106.0119.v1

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.

In the system, each order is assigned to a unique customer_id. This means that the same customer will get different identifiers for different orders. The purpose of having a customer_unique_id in the data set is to allow you to identify customers who made re-purchases in the store. Otherwise, you will find that each order has a different customer associated with it. The information it contains is as follows:

- **customer_id**: Key of the order dataset. Each order has a unique customer_id.
- **customer_unique_id:** Unique identifier of a customer**.**
- **customer_zip_code_prefix:** First five digits of the customer's zip code.
- **customer_city:** Name of the customer's city.
- **customer_state**: Customer's state

2.1.2. Geolocation data

This dataset contains information on Brazilian postal codes and their lat/long coordinates. The information it contains is the following:

- **geolocation_zip_code_prefix:** First 5 digits of the zip code.
- **geolocation_lat:** Latitude
- **geolocation_lng:** Longitude
- **geolocation_city:** City name
- **geolocation_state:** State where city is located

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. The information it contains is the following :
- **order_id:** Unique order id.
- **order_item_id:** Sequential number of the items included in the same order.
- product_id: Unique identifier of the product
- **seller_id:** Unique identifier of the seller
- **shipping_limit_date:** Shows the seller's deadline to deliver the product.
- priceitem: Price
- **freight_valueitem:** Value of the freight (If there are more items in the order, this value is distributed among the items)

2.1.4. Payment data

This dataset includes data on order payment options. The information it contains is as follows::

- **order_id:** Unique identifier of an order.
- **payment_sequential:** A customer can pay for an order with more than one payment method. If he does so, a sequence will be created to accommodate all payments.
- **payment_type:** Payment method chosen by the customer.
- **payment_install:** Number of installments chosen by the customer.
- **payment_value:** Value of the transaction.

2.1.5. Order data

This is the central data set. For each order you can find all other information. The information it contains is the following:

- **order_id:** unique identifier of the order**.**
- **customer_id:** Key of the customer dataset. Each order has a unique customer_id**.**
- **order_status:** Reference to the order status (delivered, shipped, etc.).
- **order_purchase_timestamp:** Displays the purchase timestamp.
- **order_approved_at:** Displays the payment approval timestamp.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 June 2021 doi:10.20944/preprints202106.0119.v1

- **order_delivered_carrier_date:** Displays the order release timestamp. When it was handled to the logistics partner.
- **order_delivered_customer_date:** Displays the actual delivery date of the order to the customer.
- **order_estimated_delivery_date:** Shows the estimated delivery date that was informed to the customer at the time of purchase**.**

2.1.6. Product data

This dataset includes data about the products sold by Olist. The information it contains is as follows:

- **product_id:** Unique identifier of the product
- **product_category_name:** Category name in Portuguese.
- **product_name_lenght:** Number of characters extracted from the product name.
- **product_description_lenght:** Number of characters extracted from the product description.
- **product_photos_qty:** Number of published photos of the product.
- **product_weight_g:** Product weight measured in grams.
- **product_length_cm:** Product length measured in centimeters.
- **product_height_cm:** Product height measured in centimeters.
- **product_width_cm:** Width of the product measured in centimeters**.**

2.1.7. Vendor data

This dataset translates the product_category_name into English.
The information it contains is as follows :

- **product_category_name**: Category name in Portuguese.
- **product_category_name_english**: Name of the category in English.

2.1.8. Category name translation data

This dataset contains information on Brazilian postal codes and their lat/long coordinates. The information it contains is the following:

- **geolocation_zip_code_prefix:** First 5 digits of the zip code.
- **geolocation_lat:** Latitude
- **geolocation_lng:** Longitude
- **geolocation_city:** City name
- **geolocation_state:** State where city is located
- 2.1. Transactional data model

2.2. Experiments

The experiments developed with the recommendation engine are generated based on the normalization of the customer-product data and a reference dummy dataset.

- Normalized Dataset: A Customer-Product pivot table is generated in order to identify which products have been consumed by the customers. Once this information is obtained, the data is normalized to make it comparable. This matrix contains customer, product, and purchase history normalization information.
- Reference Dataset: A purchase definition is made for the products associated with the customer, with which a stable frequency can be obtained, since it is arbitrarily defined that everyone buys a product. This Dataset is taken as a contrast to the normalized one to evaluate the results of the algorithm.

All the experiments have been defined with a subset of customers and a limit of five recommended items, so that the results can be compared and evaluated with the same criteria.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 June 2021 doi:10.20944/preprints202106.0119.v1

The test customers are:

- **871766c5855e863f6eccc05f988b23cb**
- **eb28e67c4c0b83846050ddfb8a35d051**
- **3818d81c6709e39d06b2738a8d3a2474**

2.1.1. Content-based recommendation engine

This data tool is not related to the user, but to the products that are most purchased by all users, so its result should be consistent with the items with the highest number of sales among customers.

customer_id	product_id	score	rank
8	985c412b0ac92ed9d8a76bbeab...	1.0	1
8	2bc88b31190908684ebece09e5...	1.0	2
8	a8b89693eabf7221621a71285e...	1.0	3
8	4563095e06df1fa67de2eade86...	1.0	4
8	6bd90496e4292446cceb64b93d...	1.0	5

[160 rows x 4 columns]

customer_id	product_id	score	rank
e	985c412b0ac92ed9d8a76bbeab...	1.0	1
e	2bc88b31190908684ebece09e5...	1.0	2
e	a8b89693eabf7221621a71285e...	1.0	3
e	4563095e06df1fa67de2eade86...	1.0	4
e	6bd90496e4292446cceb64b93d...	1.0	5

[160 rows x 4 columns]

customer_id	product_id	score	rank
3	985c412b0ac92ed9d8a76bbeab...	1.0	1
3	2bc88b31190908684ebece09e5...	1.0	2
3	a8b89693eabf7221621a71285e...	1.0	3
3	4563095e06df1fa67de2eade86...	1.0	4
3	6bd90496e4292446cceb64b93d...	1.0	5

[160 rows x 4 columns]

Figure 1. Normalized Dataset - Content-based

As can be seen, the result is effectively aligned with the initial definition, that all customers would have the same products, in this case, the five best sellers.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 June 2021 doi:10.20944/preprints202106.0119.v1

customer_id	product_id	score	rank
8	a92930c327948861c015c919a0...	1.0	1
8	9cc0259ca653fa86df45978b37...	1.0	2
8	a659cb33082b851fb87a33af8f...	1.0	3
8	afeeea6271148ee1bb15173b81...	1.0	4
8	0f784f8f15179b9e101beb8579...	1.0	5

[160 rows x 4 columns]

customer_id	product_id	score	rank
e	a92930c327948861c015c919a0...	1.0	1
e	9cc0259ca653fa86df45978b37...	1.0	2
e	a659cb33082b851fb87a33af8f...	1.0	3
e	afeeea6271148ee1bb15173b81...	1.0	4
e	0f784f8f15179b9e101beb8579...	1.0	5

[160 rows x 4 columns]

customer_id	product_id	score	rank
3	a92930c327948861c015c919a0...	1.0	1
3	9cc0259ca653fa86df45978b37...	1.0	2
3	a659cb33082b851fb87a33af8f...	1.0	3
3	afeeea6271148ee1bb15173b81...	1.0	4
3	0f784f8f15179b9e101beb8579...	1.0	5

[160 rows x 4 columns]

Figure 2. Reference Dataset - Content-based

In the case of the reference dataset, the same results are maintained, all customers have the same products, the five best-selling products

2.1.2. Collaborative recommendation engine

This data tool is related to the user and the products they have purchased in their history. The approach presented is to identify how similar customers are, based on the products they have already purchased.

This approach, unlike the previous one, presents two approaches to measure the similarity between users, one of these is the cosine distance and the other, the Pearson correlation. The value resulting from this measurement shows that the closer it is to one (1), the more similar the customers are, and the closer it is to zero (0), the less similar they are.

customer_id	product_id	score	rank
8	ba80c9f47a84d1e08465f72e22...	0.040320448875427246	1
8	3ae28b124972bb81eddcb644cd...	0.040320448875427246	2
8	a00d11a2119bd70d658fc7cdcf...	0.040320448875427246	3
8	7c1e2b3fa0233e46fb3bcdcb99...	0.040320448875427246	4

[128 rows x 4 columns]

customer_id	product_id	score	rank
e	3ae28b124972bb81eddcb644cd...	0.040320448875427246	1
e	a00d11a2119bd70d658fc7cdcf...	0.040320448875427246	2
e	7c1e2b3fa0233e46fb3bcdcb99...	0.040320448875427246	3

[96 rows x 4 columns]

customer_id	product_id	score	rank
3	7650dd3b2dc10798a8cbccc78d...	0.040320448875427246	1
3	ba80c9f47a84d1e08465f72e22...	0.040320448875427246	2
3	3ae28b124972bb81eddcb644cd...	0.040320448875427246	3
3	a00d11a2119bd70d658fc7cdcf...	0.040320448875427246	4
3	7c1e2b3fa0233e46fb3bcdcb99...	0.040320448875427246	5

[160 rows x 4 columns]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 June 2021 doi:10.20944/preprints202106.0119.v1

Figure 3. Normalized dataset - Collaborative engine recommendations - Cosine distance

customer_id	product_id	score	rank
0	985c412b0ac92ed9d8a76bbeab...	1.0	1
0	2bc88b31190908684ebece09e5...	1.0	2
0	a8b89693eabf7221621a71285e...	1.0	3
0	4563095e06df1fa67de2eade86...	1.0	4
0	6bd90496e4292446cceb64b93d...	1.0	5

[160 rows x 4 columns]

customer_id	product_id	score	rank
e	2bc88b31190908684ebece09e5...	1.0	1
e	a8b89693eabf7221621a71285e...	1.0	2
e	4563095e06df1fa67de2eade86...	1.0	3
e	6bd90496e4292446cceb64b93d...	1.0	4

[128 rows x 4 columns]

customer_id	product_id	score	rank
3	b0528299d65ab35e3ed853f6a8...	1.0	1
3	985c412b0ac92ed9d8a76bbeab...	1.0	2
3	2bc88b31190908684ebece09e5...	1.0	3
3	a8b89693eabf7221621a71285e...	1.0	4
3	4563095e06df1fa67de2eade86...	1.0	5
3	6bd90496e4292446cceb64b93d...	1.0	6

[192 rows x 4 columns]

Figure 4. Normalized dataset - Collaborative engine recommendations – Pearson correlation

customer_id	product_id	score	rank
8	cfd6c873a8d86ecd3a2cc3b96f...	0.0	1
8	6cc859e89d080218ff4416539f...	0.0	2
8	1522589c64efd46731d3522568...	0.0	3
8	f35927953ed82e19d06ad3aac2...	0.0	4
8	ff2c1ec09b1bb340e84f0d6b21...	0.0	5

[160 rows x 4 columns]

customer_id	product_id	score	rank
e	6cc859e89d080218ff4416539f...	0.0	1
e	1522589c64efd46731d3522568...	0.0	2
e	f35927953ed82e19d06ad3aac2...	0.0	3
e	ff2c1ec09b1bb340e84f0d6b21...	0.0	4

[128 rows x 4 columns]

customer_id	product_id	score	rank
3	fb829a6572df5239767c1ccde5...	0.0	1
3	cfd6c873a8d86ecd3a2cc3b96f...	0.0	2
3	6cc859e89d080218ff4416539f...	0.0	3
3	1522589c64efd46731d3522568...	0.0	4
3	f35927953ed82e19d06ad3aac2...	0.0	5
3	ff2c1ec09b1bb340e84f0d6b21...	0.0	6

[192 rows x 4 columns]

Figure 5. Reference dataset - Collaborative engine recommendations - Cosine distance

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 June 2021 doi:10.20944/preprints202106.0119.v1

customer_id	product_id	score	rank
8	cfd6c873a8d86ecd3a2cc3b96f...	0.0	1
8	6cc859e89d080218ff4416539f...	0.0	2
8	1522589c64efd46731d3522568...	0.0	3
8	f35927953ed82e19d06ad3aac2...	0.0	4
8	ff2c1ec09b1bb340e84f0d6b21...	0.0	5

[160 rows x 4 columns]

customer_id	product_id	score	rank
e	6cc859e89d080218ff4416539f...	0.0	1
e	1522589c64efd46731d3522568...	0.0	2
e	f35927953ed82e19d06ad3aac2...	0.0	3
e	ff2c1ec09b1bb340e84f0d6b21...	0.0	4

[128 rows x 4 columns]

customer_id	product_id	score	rank
3	fb829a6572df5239767c1ccde5...	0.0	1
3	cfd6c873a8d86ecd3a2cc3b96f...	0.0	2
3	6cc859e89d080218ff4416539f...	0.0	3
3	1522589c64efd46731d3522568...	0.0	4
3	f35927953ed82e19d06ad3aac2...	0.0	5
3	ff2c1ec09b1bb340e84f0d6b21...	0.0	6

[192 rows x 4 columns]

Figure 6. Reference dataset - Collaborative engine recommendations – Pearson correlation

The previous results are generated by specialized model, it is necessary to include quality metrics in the results to define which is the best recommender. The metrics are precision, recall and rsme.

Before generating a analyze, it is necessary to contextualize the metrics used during this evaluation, such as:

- Accuracy: Its function is to analyze whether the results obtained have really been used by users of information, an example of this is, if 10 items are recommended and of these, the customer only buys 3, we can say that we have an accuracy of 30%, which indicates that it is very good value and the model has great impact.

- Recall: It is analyzed if the products purchased by the customer is related to the recommended ones, an example of this is, if a customer buys 10 items and among the recommended ones there were 2 of them, then the recall would be 20%.

- RSME: Measures the error of the recommended products, the lower the value of this indicator, the better the results.

Table 1. The results of the metrics applied for each of the models are described below.

	Normalized		**Reference**	
	Cosine	Person	Cosine	Person
Recall	0.033	0.0045	0.0024	0
Precision	0.029	0	0.001	0
RSME	0.297	0.361	0.99	1

3. Methods (required)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 June 2021 doi:10.20944/preprints202106.0119.v1

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.

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 content-based recommendation engine fulfills its objective of recommending the best-selling products to customers.
- When analyzing the results obtained, the recommendations are strong, the same recommendation is always given to all customers selected from the chosen Dataset.
- It is necessary to clarify that, if we compare the recommended products between the normalized Dataset and tests, they are totally different, and this is correct, since their recommendation is different, due to the same nature of the data.
- It can be detailed that the best model to be used is the Cosine distance model, with the normalized Dataset, since it presents the lowest squared error and the best precision and recall values.
- It is also suggested that for future experiments, these tests could be extended, since the current ones have not been good.

References

1. Rose R. y Pulizzi J. (2017). Killing Marketing: How Innovative Businesses Are Turning Marketing Cost Into Profit. (chap. 1-5).
2. Davis, J. (2017). Measuring Marketing. Part 4: Customers Metrics
3. Meerman, S. (2015). The New Rules of Marketing and PR: How to Use Social Media, Online Video, Mobile Applications, Blogs, News Releases, and Viral Marketing to Reach Buyers Directly, growing your business: how marketing and pr drive sales (cap.11)
4. Reibstein D. , Bendle N., Farris P¨., Pfeifer P. (2015). Marketing Metrics: The Manager's Guide to Measuring Marketing Performance, Third Edition, 5.1 Customers, Recency, And Retention (chap. 5.1).
5. Laursen G. (2011). Business Analytics for Sales and Marketing Managers: How to Compete in the Information Age. Case Study of a Retention Strategy (chap.9).
6. Guo-en and Wei-dong, n. s. (2008). Model of customer churn prediction on support vector machine.
7. Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian, Yong Shii, n. s. (2011). Credit card churn forecasting by logistic regression and decision tree
8. A. Keramati a, R. Jafari-Marandi a, M. Aliannejadi b, I. Ahmadianc, M. Mozzafari a, U. Abbasi a (2014). Improved churn prediction in telecommunication industry using data mining techniques.
9. A Nie et al (2011). Credit card churn forecasting by logistic regression and decision tree.
10. Farquad MAH, Ravi V, Raju SB (2014). Churn prediction using comprehensible support vector machine: An analytical CRM application.
11. Keramati A, Jafari-Marandi R, Aliannejadi M, Ahmadian I, Mozzafari M, Abbasi (2014). Improved churn prediction in telecommunication industry using data mining techniques.

12. Kuanchin chen, Ya-Han Hu (2015). Predicting customer churn from valuable B2B customers in the logistics industry: a case study.
13. 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.
14. A Philip Spanoudes, Thomson Nguyen (2016). A Fuzzy Rule-Based Learning Algorithm for Customer Churn Prediction.
15. A Adnan Amin, Feras Al-Obeidat, Babar Shah, Awais Adnan, Jonathan Loo, Sajid Anwaren (2019). Customer churn prediction in telecommunication industry using data certainty.