Working paper

Customer Churn - Prevention Model – Prediction Model

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..

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.

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 repurchases 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.
- **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

Once all the customer retention strategies have been implemented, it is necessary to know which of these customers are potential casualties, so that the loyalty teams can execute their strategies and thus avoid this announced loss. This requires a two-pronged approach:

- The first one is focused on identifying which of these customers are already definitively low in the organization, using the customer's purchase behavior for this marking, since currently there is no such classification.
- As a next step, once we have identified which customers are marked as churn, we proceed to obtain this input, in order to train supervised algorithms that allow us to predict the leakage of these customers and select the one that gives the best result. Developed with the recommendation engine are generated based on the normalization of the customer-product data and a reference dummy dataset.

doi:10.20944/preprints202106.0113.v1

2.1.1. Customer Churn - Mark

Experiments will be carried out to mark customers as lost, using the information we have on purchases, frequency, seasonality and amounts of the transactions carried out. This is applied with the RFM model, of which three components will be used:

- **Recency**: This is the time that has passed since your last purchase. This is equal to the duration between a customer's first purchase and their last purchase. (Thus, if they have only made 1 purchase, the recency is 0.)
- **Frequency**: This represents the number of repeats purchases the customer has made. This means, it is the count of periods in days in which the customer made a purchase. Therefore, it is the count of days in which the customer made a purchase.
- **T:** Represents the age of the customer in units of time in days This is equal to the duration between a customer's first purchase and the end of the period.

Based on the definition of RFM the following hypothesis for marking Churn is applied:

- **Recency:** By its very approach to customer seniority analysis, anything less than 1, can be marked as churn.
- **Frequency**: It will be considered that the customer has a purchase frequency of less than 1 in a month.
- **T:** It is observed that at level 400 onwards the values are purple, this means that it has been practically still in the last year, since the measurement is in days.

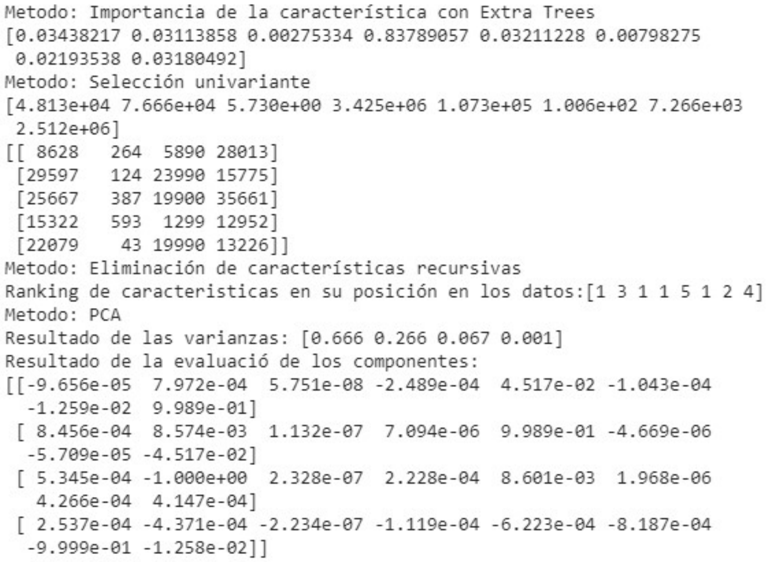
By applying this hypothesis, it was obtained that 18% of the historical customer database is churn. be seen, the result is effectively aligned with the initial definition, that all customers would have the same products, in this case, the five best sellers.

2.1.2. Variable selection – First Strategy

In these stages, with the customers marked as lost, we proceed to identify the relevant information variables, which allow us to perform a series of experiments and thus identify the best artificial intelligence algorithm for the prediction of customer loss.

Before starting to run the AI algorithms, it is necessary to find the relevant information characteristics of the entire dataset. For this purpose, different prioritization methods are applied, being a total of four (4) approaches that allow to obtain with greater assertiveness the most appropriate variables to be used in the supervised learning algorithms. The techniques used for this feature prioritization are:

- Feature importance with Extra Trees
- Univariate selection
- Recursive feature elimination
- PCA



```
Metodo: Importancia de la característica con Extra Trees
[0.03438217 0.03113858 0.00275334 0.83789057 0.03211228 0.00798275
 0.02193538 0.03180492]
Metodo: Selección univariante
[4.813e+04 7.666e+04 5.730e+00 3.425e+06 1.073e+05 1.006e+02 7.266e+03
 2.512e+06]
[[ 8628   264  5890 28013]
 [29597   124 23990 15775]
 [25667   387 19900 35661]
 [15322   593  1299 12952]
 [22079    43 19990 13226]]
Metodo: Eliminación de características recursivas
Ranking de caracteristicas en su posición en los datos:[1 3 1 1 5 1 2 4]
Metodo: PCA
Resultado de las varianzas: [0.666 0.266 0.067 0.001]
Resultado de la evaluació de los componentes:
[[-9.656e-05  7.972e-04  5.751e-08 -2.489e-04  4.517e-02 -1.043e-04
  -1.259e-02  9.989e-01]
 [ 8.456e-04  8.574e-03  1.132e-07  7.094e-06  9.989e-01 -4.669e-06
  -5.709e-05 -4.517e-02]
 [ 5.345e-04 -1.000e+00  2.328e-07  2.228e-04  8.601e-03  1.968e-06
   4.266e-04  4.147e-04]
 [ 2.537e-04 -4.371e-04 -2.234e-07 -1.119e-04 -6.223e-04 -8.187e-04
  -9.999e-01 -1.258e-02]]
```

Figure 1. Variables evaluation results - Variables selection

As a result of the values generated by the four methods, the four (4) most relevant variables are prioritized, according to the number of times they were selected and the hierarchical position in which they were proposed:

- **Feature Importance with Extra Trees**: identified the following features as relevant: DATE, seller_id, product_id, Price, customer_zip_code_prefix. The last three variables had the same value so it does not give a result of only four characteristics.
- **Univariate selection:** Identified the following features as relevant: order_status, customer_state, customer_city, seller_id.
- **Recursive Feature Elimination:** Identified the following variables as relevant: seller_id, order_status, DATE, customer_state
- **PCA:** Identified the following characteristics as relevant: customer_zip_code_prefix, customer_city, Price, product_id

The characteristics selected as most relevant in their order are:

1. **seller_id:** Appears in three results.

2. **order_status:** Appears in two results, in a ranking of 1st and 2nd.

3. **DATE:** Appears in two results, in a ranking of 1st and 3rd.

4. **customer_zip_code_prefix:** Appears in two results, in a ranking of 1st and 4th.

5. **product_id:** Appears in two results, in a ranking of 3rd and 4th.

6. **price:** Appears in two results, in a ranking of 3rd and 4th.

7. **customer_city:** Appears in two results, in a ranking of 3rd and 2nd.

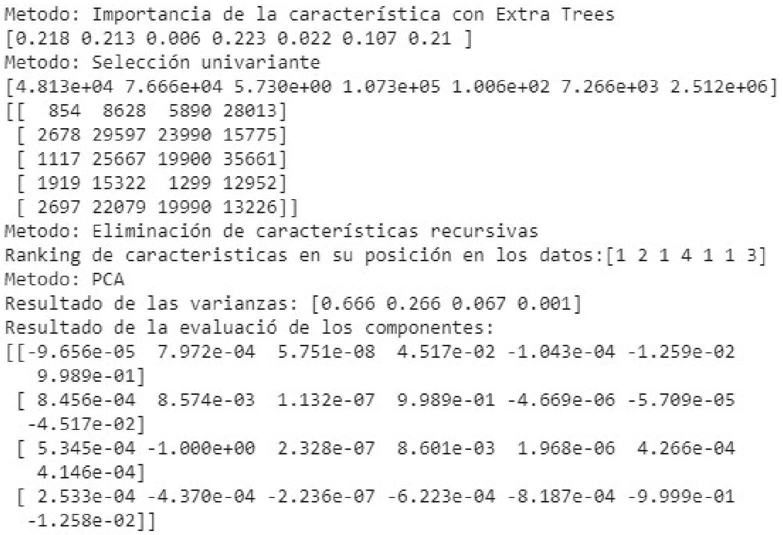
8. **customer_state:** Appears in two results, in a ranking of 2nd and 4th. previous results are generated by specialized model, it is necessary to include quality metrics in the results to define which is the best recommender. The metrics are precision, recall and rsme.

As a result, a ranking is made with the criteria initially described, and it is concluded that the four key fields for this experiment are seller_id, order_status, DATE, customer_zip_code_prefix, customer_zip_code_prefix, and customer_zip_code_prefix.

2.1.3. Variable selection – Second strategy

In this experiment, we arbitrarily chose to eliminate the temporal information in order to have a reference result of the previous experiment, and thus identify the changes in the decision of the prioritized variables.

As a result, the following results are shown below:



```
Metodo: Importancia de la característica con Extra Trees
[0.218 0.213 0.006 0.223 0.022 0.107 0.21 ]
Metodo: Selección univariante
[4.813e+04 7.666e+04 5.730e+00 1.073e+05 1.006e+02 7.266e+03 2.512e+06]
[[  854  8628  5890 28013]
 [ 2678 29597 23990 15775]
 [ 1117 25667 19900 35661]
 [ 1919 15322  1299 12952]
 [ 2697 22079 19990 13226]]
Metodo: Eliminación de características recursivas
Ranking de caracteristicas en su posición en los datos:[1 2 1 4 1 1 3]
Metodo: PCA
Resultado de las varianzas: [0.666 0.266 0.067 0.001]
Resultado de la evaluació de los componentes:
[[-9.656e-05  7.972e-04  5.751e-08  4.517e-02 -1.043e-04 -1.259e-02
   9.989e-01]
 [ 8.456e-04  8.574e-03  1.132e-07  9.989e-01 -4.669e-06 -5.709e-05
  -4.517e-02]
 [ 5.345e-04 -1.000e+00  2.328e-07  8.601e-03  1.968e-06  4.266e-04
   4.146e-04]
 [ 2.533e-04 -4.370e-04 -2.236e-07 -6.223e-04 -8.187e-04 -9.999e-01
  -1.258e-02]]
```

Figure 2. Variables evaluation results (No DATE) - Variable selection

As a result of the values generated by the four methods, the four (4) most relevant variables are prioritized, according to the number of times they were selected and the hierarchical position in which they were proposed:

- **Feature Importance with Extra Trees:** identified the following features as relevant: price, seller_id, product_id, customer_state, customer_zip_code_prefix. The last three variables had the same value that is why a result of 4 values is not given.
- **Univariate selection:** Identified the following features as relevant: order_status, customer_state, customer_city, product_id.
- **Recursive Feature Elimination:** Identified the following features as relevant: seller_id, order_status, customer_state, customer_city
- **PCA:** Identified the following features as relevant: customer_zip_code_prefix, customer_city, Price, product_id

As a result, the most relevant characteristics are, in their order:product_id: Appears in three results

1. customer_state: Appears in three results
2. seller_id: Appears in two results, in a ranking of 1st and 2nd.
3. order_status: Appears in two results, in a ranking of 1st and 2nd.
4. customer_zip_code_prefix: Appears in two results, in a ranking of 1st and 2nd.
5. price: Appears in two results, in a ranking of 1st and 3rd.
6. customer_city: Appears in two results, ranked 3rd and 2nd.

As a result, a ranking is made with the criteria initially described, and it is concluded that the four key fields for this experiment are product_id, customer_state, seller_id, order_status.

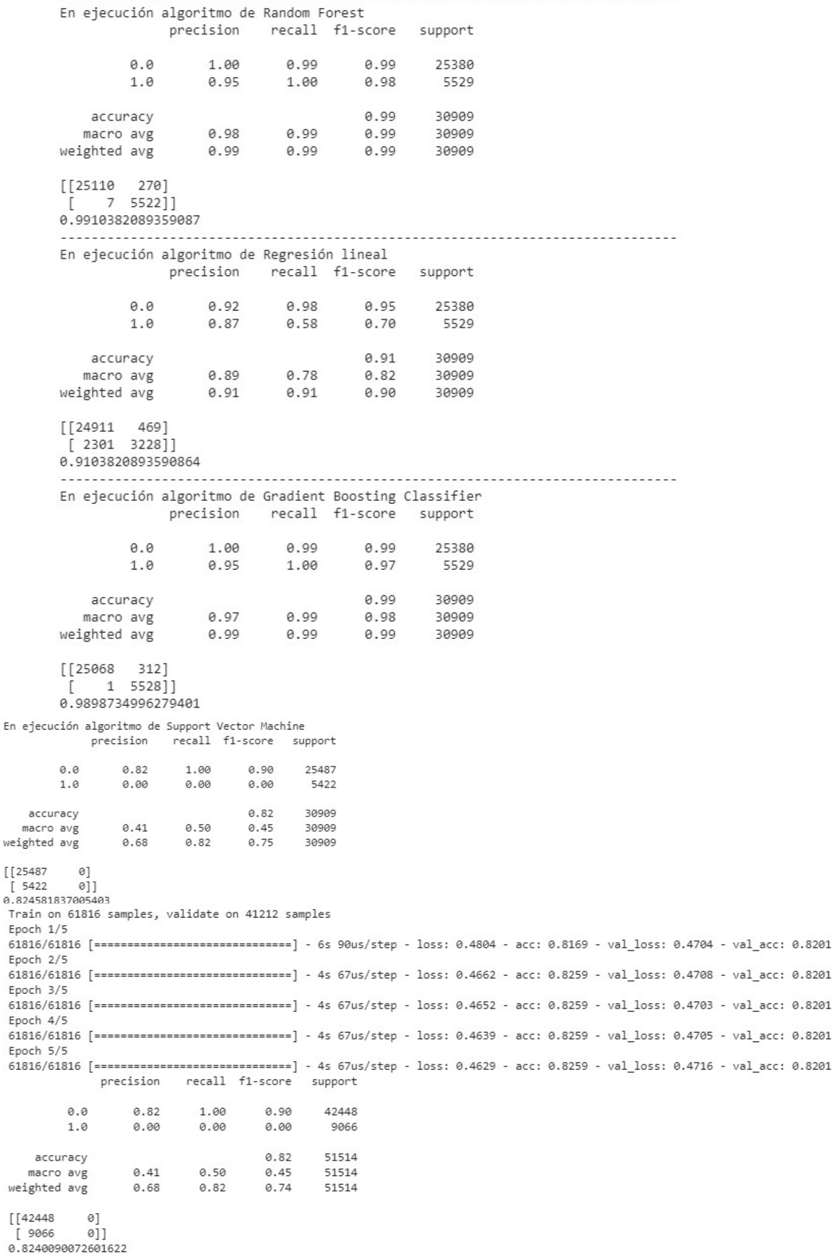
At the end of these two experiments, two subsets of information to be used in the supervised learning algorithms are concluded:

- **Group 1:** seller_id, order_status, DATE, customer_zip_code_prefix.
- **Group 2 (control 1):** product_id, customer_state, seller_id, order_status
- **Group 3 (control 2):** no prioritization

2.1.4. Algorithms application

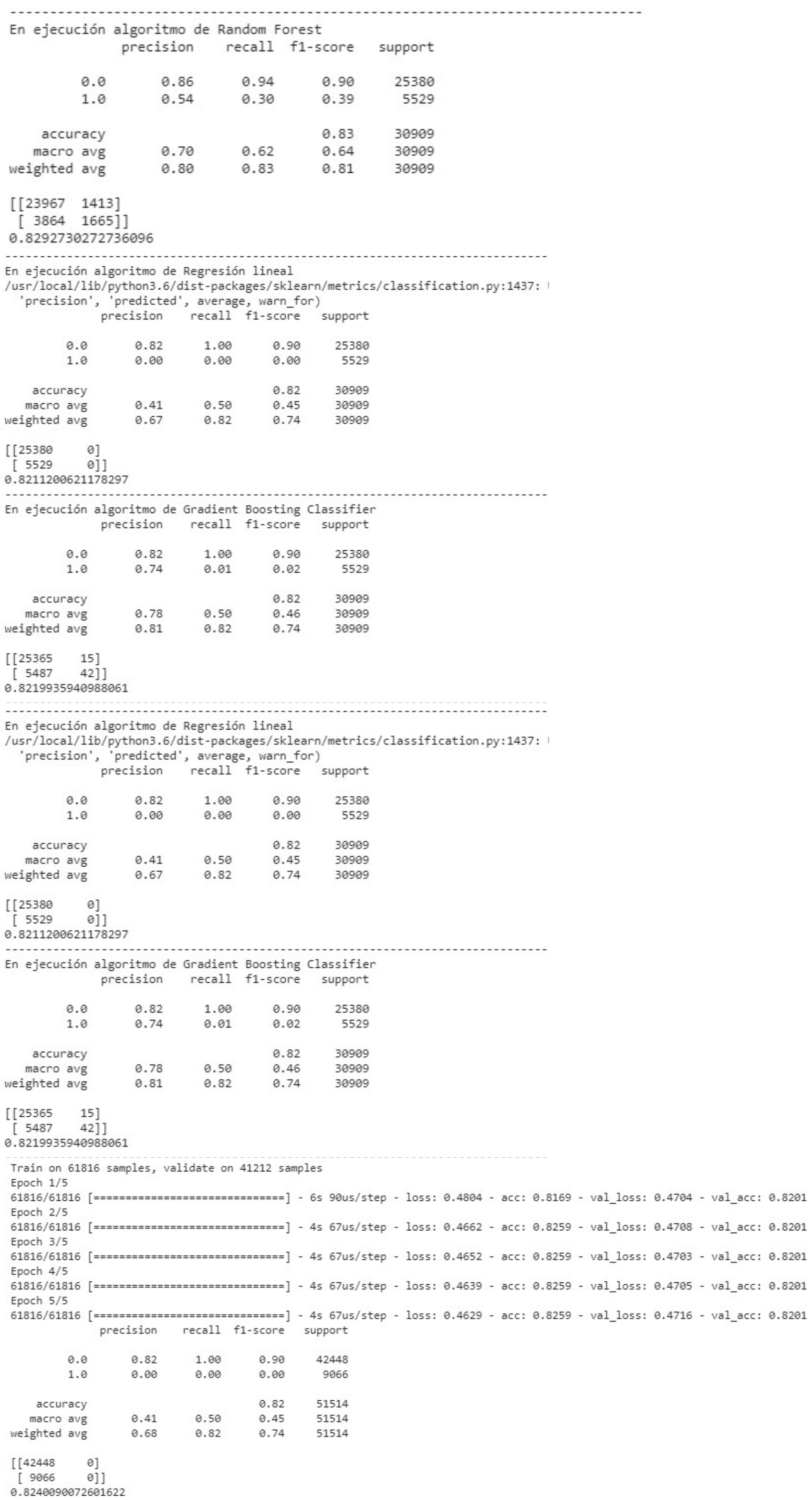
In this stage, experiments will be carried out with each of the algorithms described below, according to the data sets defined in the previous stage. As a scope of these algorithms, the following will be implemented:

- Random Forest
- Linear Regression
- Gradient Boosting Classifier
- Support Vector Machine
- Neural Networks



```
En ejecución algoritmo de Random Forest
              precision    recall  f1-score   support

         0.0       1.00      0.99      0.99     25380
         1.0       0.95      1.00      0.98      5529

    accuracy                           0.99     30909
   macro avg       0.98      0.99      0.99     30909
weighted avg       0.99      0.99      0.99     30909

[[25110   270]
 [    7  5522]]
0.9910382089359087
--------------------------------------------------------------------------------
En ejecución algoritmo de Regresión lineal
              precision    recall  f1-score   support

         0.0       0.92      0.98      0.95     25380
         1.0       0.87      0.58      0.70      5529

    accuracy                           0.91     30909
   macro avg       0.89      0.78      0.82     30909
weighted avg       0.91      0.91      0.90     30909

[[24911   469]
 [ 2301  3228]]
0.9103820893590864
--------------------------------------------------------------------------------
En ejecución algoritmo de Gradient Boosting Classifier
              precision    recall  f1-score   support

         0.0       1.00      0.99      0.99     25380
         1.0       0.95      1.00      0.97      5529

    accuracy                           0.99     30909
   macro avg       0.97      0.99      0.98     30909
weighted avg       0.99      0.99      0.99     30909

[[25068   312]
 [    1  5528]]
0.9898734996279401
En ejecución algoritmo de Support Vector Machine
              precision    recall  f1-score   support

         0.0       0.82      1.00      0.90     25487
         1.0       0.00      0.00      0.00      5422

    accuracy                           0.82     30909
   macro avg       0.41      0.50      0.45     30909
weighted avg       0.68      0.82      0.75     30909

[[25487     0]
 [ 5422     0]]
0.824581837005403
 Train on 61816 samples, validate on 41212 samples
 Epoch 1/5
 61816/61816 [==============================] - 6s 90us/step - loss: 0.4804 - acc: 0.8169 - val_loss: 0.4704 - val_acc: 0.8201
 Epoch 2/5
 61816/61816 [==============================] - 4s 67us/step - loss: 0.4662 - acc: 0.8259 - val_loss: 0.4708 - val_acc: 0.8201
 Epoch 3/5
 61816/61816 [==============================] - 4s 67us/step - loss: 0.4652 - acc: 0.8259 - val_loss: 0.4703 - val_acc: 0.8201
 Epoch 4/5
 61816/61816 [==============================] - 4s 67us/step - loss: 0.4639 - acc: 0.8259 - val_loss: 0.4705 - val_acc: 0.8201
 Epoch 5/5
 61816/61816 [==============================] - 4s 67us/step - loss: 0.4629 - acc: 0.8259 - val_loss: 0.4716 - val_acc: 0.8201
              precision    recall  f1-score   support

         0.0       0.82      1.00      0.90     42448
         1.0       0.00      0.00      0.00      9066

    accuracy                           0.82     51514
   macro avg       0.41      0.50      0.45     51514
weighted avg       0.68      0.82      0.74     51514

[[42448     0]
 [ 9066     0]]
0.8240090072601622
```

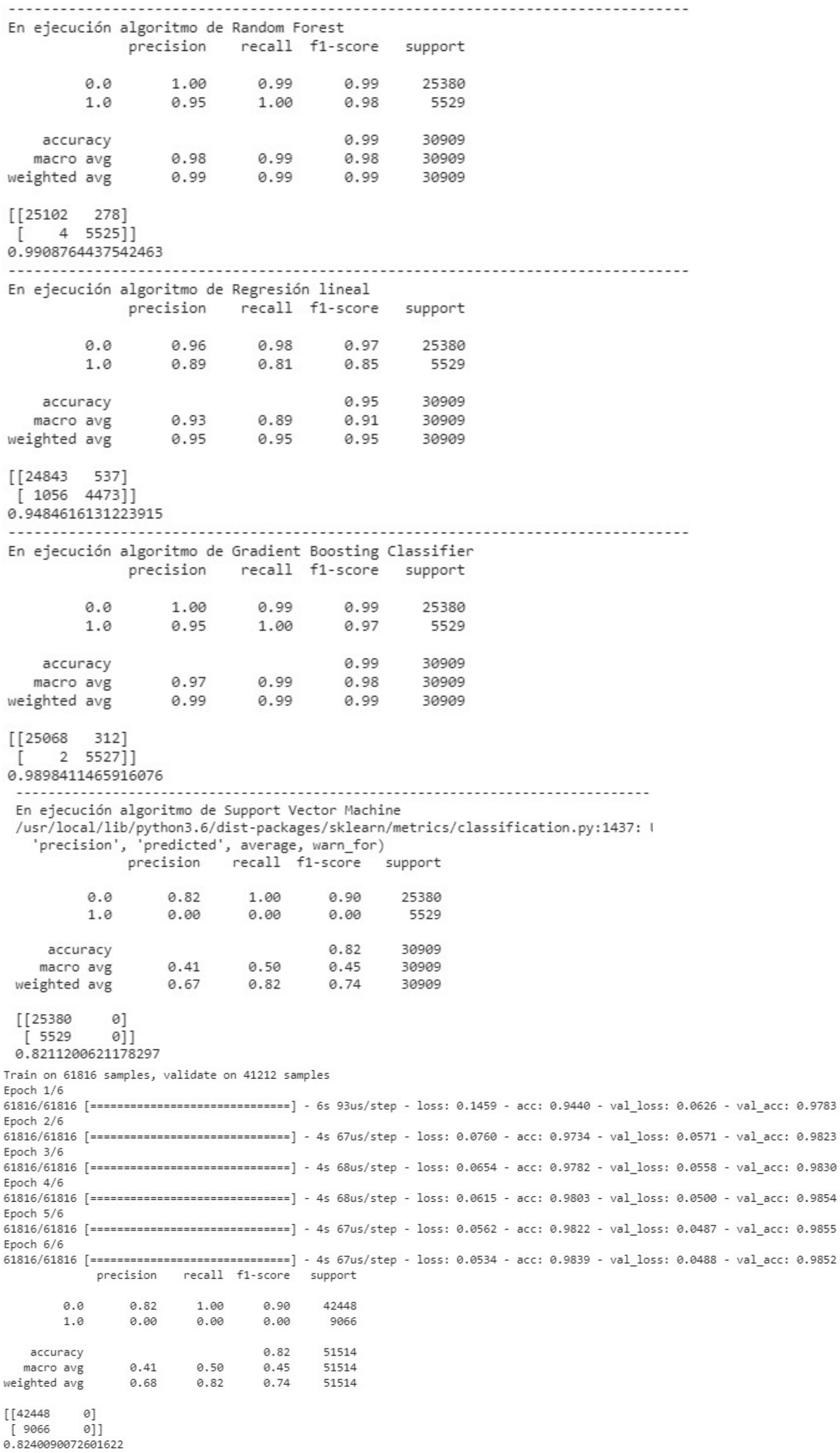
Figure 3. Algorithm training and validation results with Group 1 data



```
--------------------------------------------------------------------------------
En ejecución algoritmo de Random Forest
              precision    recall  f1-score   support

         0.0       0.86      0.94      0.90     25380
         1.0       0.54      0.30      0.39      5529

    accuracy                           0.83     30909
   macro avg       0.70      0.62      0.64     30909
weighted avg       0.80      0.83      0.81     30909

[[23967  1413]
 [ 3864  1665]]
0.8292730272736096
--------------------------------------------------------------------------------
En ejecución algoritmo de Regresión lineal
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1437: 
  'precision', 'predicted', average, warn_for)
              precision    recall  f1-score   support

         0.0       0.82      1.00      0.90     25380
         1.0       0.00      0.00      0.00      5529

    accuracy                           0.82     30909
   macro avg       0.41      0.50      0.45     30909
weighted avg       0.67      0.82      0.74     30909

[[25380     0]
 [ 5529     0]]
0.8211200621178297
--------------------------------------------------------------------------------
En ejecución algoritmo de Gradient Boosting Classifier
              precision    recall  f1-score   support

         0.0       0.82      1.00      0.90     25380
         1.0       0.74      0.01      0.02      5529

    accuracy                           0.82     30909
   macro avg       0.78      0.50      0.46     30909
weighted avg       0.81      0.82      0.74     30909

[[25365    15]
 [ 5487    42]]
0.8219935940988061
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
En ejecución algoritmo de Regresión lineal
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1437: 
  'precision', 'predicted', average, warn_for)
              precision    recall  f1-score   support

         0.0       0.82      1.00      0.90     25380
         1.0       0.00      0.00      0.00      5529

    accuracy                           0.82     30909
   macro avg       0.41      0.50      0.45     30909
weighted avg       0.67      0.82      0.74     30909

[[25380     0]
 [ 5529     0]]
0.8211200621178297
--------------------------------------------------------------------------------
En ejecución algoritmo de Gradient Boosting Classifier
              precision    recall  f1-score   support

         0.0       0.82      1.00      0.90     25380
         1.0       0.74      0.01      0.02      5529

    accuracy                           0.82     30909
   macro avg       0.78      0.50      0.46     30909
weighted avg       0.81      0.82      0.74     30909

[[25365    15]
 [ 5487    42]]
0.8219935940988061
--------------------------------------------------------------------------------
 Train on 61816 samples, validate on 41212 samples
 Epoch 1/5
 61816/61816 [==============================] - 6s 90us/step - loss: 0.4804 - acc: 0.8169 - val_loss: 0.4704 - val_acc: 0.8201
 Epoch 2/5
 61816/61816 [==============================] - 4s 67us/step - loss: 0.4662 - acc: 0.8259 - val_loss: 0.4708 - val_acc: 0.8201
 Epoch 3/5
 61816/61816 [==============================] - 4s 67us/step - loss: 0.4652 - acc: 0.8259 - val_loss: 0.4703 - val_acc: 0.8201
 Epoch 4/5
 61816/61816 [==============================] - 4s 67us/step - loss: 0.4639 - acc: 0.8259 - val_loss: 0.4705 - val_acc: 0.8201
 Epoch 5/5
 61816/61816 [==============================] - 4s 67us/step - loss: 0.4629 - acc: 0.8259 - val_loss: 0.4716 - val_acc: 0.8201
              precision    recall  f1-score   support

         0.0       0.82      1.00      0.90     42448
         1.0       0.00      0.00      0.00      9066

    accuracy                           0.82     51514
   macro avg       0.41      0.50      0.45     51514
weighted avg       0.68      0.82      0.74     51514

 [[42448     0]
  [ 9066     0]]
 0.8240090072601622
```

Figure 4. Algorithm training and validation results with Group 2 data

doi:10.20944/preprints202106.0113.v1



```
-------------------------------------------------------------------------------
En ejecución algoritmo de Random Forest
              precision    recall  f1-score   support

         0.0       1.00      0.99      0.99     25380
         1.0       0.95      1.00      0.98      5529

    accuracy                           0.99     30909
   macro avg       0.98      0.99      0.98     30909
weighted avg       0.99      0.99      0.99     30909

[[25102   278]
 [    4  5525]]
0.9908764437542463
-------------------------------------------------------------------------------
En ejecución algoritmo de Regresión lineal
              precision    recall  f1-score   support

         0.0       0.96      0.98      0.97     25380
         1.0       0.89      0.81      0.85      5529

    accuracy                           0.95     30909
   macro avg       0.93      0.89      0.91     30909
weighted avg       0.95      0.95      0.95     30909

[[24843   537]
 [ 1056  4473]]
0.9484616131223915
-------------------------------------------------------------------------------
En ejecución algoritmo de Gradient Boosting Classifier
              precision    recall  f1-score   support

         0.0       1.00      0.99      0.99     25380
         1.0       0.95      1.00      0.97      5529

    accuracy                           0.99     30909
   macro avg       0.97      0.99      0.98     30909
weighted avg       0.99      0.99      0.99     30909

[[25068   312]
 [    2  5527]]
0.9898411465916076
 -------------------------------------------------------------------------------
 En ejecución algoritmo de Support Vector Machine
 /usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1437:
   'precision', 'predicted', average, warn_for)
              precision    recall  f1-score   support

         0.0       0.82      1.00      0.90     25380
         1.0       0.00      0.00      0.00      5529

    accuracy                           0.82     30909
   macro avg       0.41      0.50      0.45     30909
weighted avg       0.67      0.82      0.74     30909

 [[25380     0]
  [ 5529     0]]
 0.8211200621178297
Train on 61816 samples, validate on 41212 samples
Epoch 1/6
61816/61816 [==============================] - 6s 93us/step - loss: 0.1459 - acc: 0.9440 - val_loss: 0.0626 - val_acc: 0.9783
Epoch 2/6
61816/61816 [==============================] - 4s 67us/step - loss: 0.0760 - acc: 0.9734 - val_loss: 0.0571 - val_acc: 0.9823
Epoch 3/6
61816/61816 [==============================] - 4s 68us/step - loss: 0.0654 - acc: 0.9782 - val_loss: 0.0558 - val_acc: 0.9830
Epoch 4/6
61816/61816 [==============================] - 4s 68us/step - loss: 0.0615 - acc: 0.9803 - val_loss: 0.0500 - val_acc: 0.9854
Epoch 5/6
61816/61816 [==============================] - 4s 67us/step - loss: 0.0562 - acc: 0.9822 - val_loss: 0.0487 - val_acc: 0.9855
Epoch 6/6
61816/61816 [==============================] - 4s 67us/step - loss: 0.0534 - acc: 0.9839 - val_loss: 0.0488 - val_acc: 0.9852
              precision    recall  f1-score   support

         0.0       0.82      1.00      0.90     42448
         1.0       0.00      0.00      0.00      9066

    accuracy                           0.82     51514
   macro avg       0.41      0.50      0.45     51514
weighted avg       0.68      0.82      0.74     51514

[[42448     0]
 [ 9066     0]]
0.8240090072601622
```

Figure 5. Algorithm training and validation results with Group 3 data

It can be noted that the most effective methods for this type of information are the RANDOM FOREST and GRADIENT BOOSTING CLASSIFIER algorithms, since they have offered the best results, with group 1, group 2 (control 1) and group 3 (control 2).

doi:10.20944/preprints202106.0113.v1

Table 1. The results of algorithms applied to Group 1.

Group 1	Random Forest	Linear Reggresion	Gradient Boosting Classifier	Support Vector Machine	Neuroal Network
Recall	0,99	0,91	0,99	0,68	0,41
Precision	0,99	0,91	0,99	0,82	0,5
F1-Score	0,99	0,9	0,99	0,75	0,45
Consolidated	**0,991**	**0,9103**	**0,9898**	**0,8245**	**0,824**

Table 2. The results of algorithms applied to Group 2.

Group 2	Random Forest	Linear Reggresion	Gradient Boosting Classifier	Support Vector Machine	Neuroal Network
Recall	0,8	0,67	0,81	0,68	0,68
Precision	0,83	0,82	0,82	0,82	0,82
F1-Score	0,81	0,74	0,74	0,75	0,74
Consolidated	**0,8297**	**0,8211**	**0,8219**	**0,8245**	**0,824**

Table 3. The results of algorithms applied to Group 3.

Group 3	Random Forest	Linear Reggresion	Gradient Boosting Classifier	Support Vector Machine	Neuroal Network
Recall	0,99	0,95	0,99	0,67	0,68
Precision	0,99	0,95	0,99	0,82	0,82
F1-Score	0,99	0,95	0,99	0,74	0,74
Consolidated	**0,9908**	**0,9898**	**0,9898**	**0,8211**	**0,824**

The results in summary the algorithms with the best results are:

- Table 1: Random Forest - Gradient Boosting Classifier
- Table 2: Random Forest - Gradient Boosting Classifier
- Table 3: Random Forest - Gradient Boosting Classifier

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.

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

- One of the most common mistakes in the data management industry is to assume that applying artificial intelligence algorithms is enough to solve any problem we have, and this is a totally false hypothesis. As part of this idea it was essential to understand the problem of churn, not only from the result of the problem, which is the customer leakage, and how to avoid it through algorithms that allow to predict this situation.
- A hybrid solution allows not only to generate a working tool for organizations, but to provide added value from the first moment, since it integrates technology, algorithmic, statistics, and makes it available to organizations, in order to be part of the business strategies based on data.
- The ensemble methods generated the best churn prediction values, unlike others found in the literature such as SVM. This can be understood, due to the binary-required classification in this solution, number of features and the amount of data available for training and testing development.
- For historical data models, which do not have customer churn marking, the RFM data model allows to generate the marking based on customer purchase behavior. Its scope could be extended, since it could not only be marking churn, but also proactively detecting whether customers are about to be churned or not.
- As a premise for the evolution of this degree work approach, it is necessary to define a robust business process that ensures the chaining of the three stages of customer prevention, in such a way that it can be formalized and trained at different levels of the organization, so that the entire solution is used properly and has a continuous improvement.
- From the Customer Experience point of view, it is essential that a pleasant and easy-to-use visual environment is developed in the solution. With existing technologies, this environment can be created through business intelligence platforms such as PowerBI or Tableau, the latter being the front end of the solution. With this product, you can directly impact the sales force and business, becoming a key tool to support them in preventing customer churn and improving customer loyalty.
- The solution must be designed for mobile devices, so that any salesperson can have all their knowledge tools prepared and ready on their cell phone.
- The use of the cloud as a fundamental part of the evolution of this project, can be applied from different views, one of these is the deployment of a web service with the information of Machine learning trained, which allows through the characteristics of a customer, that a salesperson, financial, marketing professional or similar, can predict whether or not a customer is likely to be churn, and thus proactively initiate the intervention of this.
- The areas of opportunity from the data point of view are practically infinite, what is necessary to advance is how this area can learn from the business and support

it to generate new tools to improve the final customer experience and the internal processes of any organization.

- There are current technologies that allow to merge programming languages such as Python in databases, which would allow an improvement in processing times, and a native integration between the two main points of paper, data and the application of artificial intelligence algorithms.

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.