*Formatting Guidelines for Papers Published in Informing Science Institute journals*

A Logistic Regression Model for Customer Churn Classification in a Brazilian Startup

**[Do not enter author information until after the paper is accepted.]**

|  |  |  |
| --- | --- | --- |
| Antonio Carlos da Silva Júnior\* | Universidade Federal do Paraná, Curitiba, Brasil | acsjunior@ufpr.br |

\* Corresponding author

# Abstract

|  |  |
| --- | --- |
| Aim/Purpose | [In a sentence or two, explain the need for this paper. What is the problem?] |
| Background | [Explain in a sentence or two in what way this paper addresses the problem.] |
| Methodology | [Mention for the reader the methods used in the paper.  Briefly describe any research sample.] |
| Contribution | [How does this paper contribute to the body of knowledge?] |
| Findings | [List the paper’s major findings.] |
| Recommendations  for Practitioners | [Enter any recommendations for practitioners.] |
| Recommendations  for Researchers | [Enter recommendations your paper makes for researchers.] |
| Impact on Society | [What are the larger implications of the paper’s findings?] |
| Future Research | [Now that this paper has advanced your findings, what research should follow?] |
| Keywords | Place three or more keywords here, separated by commas |

# Introduction

The importance of the long-term relationship between a company and its customers is widely discussed in the literature. Due to learning effects and decreased service costs, serving customers becomes less costly in each additional year of the relationship (Ganesh et al., 2000). On account of the increased costs to attract customers in a competitive market and the potential cost reduction relating to the long-term relationships, customer retention became essential for the economic survival and success of service companies (Hennig-Thurau, 2004). According to Gallo (2014), the cost to acquire a new customer can be from 5 to 25 times higher than the retention cost of an existing one, depending on the company sector.

Developing strategies for customer retention became a common practice among companies from different sectors and as a consequence, identify early customers prone to leave became a constant desire. At a time of general effort face with the data-driven culture, the customer churn prediction models, predominantly used in the large telecommunication players, became popular tools in various companies, regardless of their magnitude and business area.

The literature demonstrates that customer churn prediction models are widely explored and allow for several outcomes: Botelho & Tostes (2020) fitted a logistic regression model to predict customer churn probability in a large retail company; Vafeiadis et al. (2015) compared different methods and were successful through an SVM polynomial using AdaBoost model for customer churn classification in the telecommunications industry; Kumar & Yadav (2020) proposed a rule-based customer churn prediction model using artificial neural network based and rough set theory on e-commerce customers review data.

Through data provided by Olist, a Brazilian startup that offers complete e-commerce solutions for small retailers, this work presents a prediction model that identifies the Olist's customers prone to leave and allows interpreting the reasons that impact the outcome.

Because of the different existing methods and the particularities of each business model, choose the appropriate algorithm becomes a crucial step of the modeling process. Therefore, based on the approach of Silva Júnior et al. (2020), which used a hybrid multicriteria model for choosing a churn prediction model, the logistic regression was chosen to develop the proposed classifier.

# Material and Methods

## Structuring of the Dataset

The dataset used in this work refers to Olist’s customers and was made anonymized and with quantitative variables standardized with mean 0 and standard deviation 1. Considering that customers are small retailers, they will be called sellers in this article. Due to the database architecture characteristics and the particularities of the company's business, it was necessary an extensive data wrangling process. This process starts with a preliminary diagnosis where is checked if they are in the appropriate format if they answer the questions that motivated the analysis and what is necessary to put them in the ideal format. Then, the occurrence of missing or duplicated data is checked, and finally, a cleaning and transformation process is performed, to obtain an adequate dataset for the study (Kandel et al., 2011).

### Definition of response variable and performance covariates

Initially, sellers with 30 or more calendar days of inactivity were defined as churned, considering accessing or making a sale through the digital platform as an activity. Then, only sellers with at least 90 days of historical data were kept in the dataset. The 90-day period ended on the cutoff date, as shown in table 1, was divided equally into two subperiods and then, for each of the periods, several metrics were calculated such as revenue, average ticket, number of orders, published products, inactivity days, etc. Then, the performance of the sellers was calculated based on these metrics using the equation where and are calculated values for each subperiod.

**Table 1. Definition of cutoff date**

|  |  |
| --- | --- |
| SELLER | CUTOFF DATE |
| Churned | Last activity |
| Active | Analysis date |

The behavior of the performance metrics is explained in table 2, given the nature of the origin equation. At the end of this step, a dataset composed of the response variable and nine performance metrics as covariates were obtained, where each observation corresponds to a seller.

**Table 2. Interpretation of performance metrics**

|  |  |
| --- | --- |
| VALUE | PERFORMANCE |
| 0.5 | Maintained |
| > 0.5 | Increased |
| < 0.5 | Decreased |

### Addition of other variables

Qualitative covariates such as seller stage, agreement plan, and the region of origin as well as other quantitative covariates such as the total number of orders and published products, total revenue, etc. were added resulting in a dataset with 23 covariates.

### Creating binary covariates

When it is necessary to analyze the behavior of the response variable in function of a qualitative covariate with levels, should be created binary covariates (dummies), going to the researcher to decide which of the levels should be the reference (Fávero & Belfiore, 2017). Therefore, all of the qualitative covariates were transformed to binary, resulting in a dataset with 32 variables and 11.131 observations.

## Logistic Regression Model

The dataset used in this work refers to Olist’s customers and was made anonymized and with quantitative variables standardized with mean 0 and standard deviation 1. Considering that customers are small retailers, they will be called sellers in this article. Due to the database architecture characteristics and the particularities of the company's business, it was necessary an extensive data wrangling process. This process starts with a preliminary diagnosis where is checked if they are in the appropriate format if they answer the questions that motivated the analysis and what is necessary to put them in the ideal format. Then, the occurrence of missing or duplicated data is checked, and finally, a cleaning and transformation process is performed, in order to obtain an adequate dataset for the study (Kandel et al., 2011).

### Model fit and covariates selection

Initially, sellers with 30 or more calendar days of inactivity were defined as churned, considering accessing or making a sale through the digital platform as an activity. Then, only sellers with at least 90 days of historical data were kept in the dataset. The 90-day period ended on the cutoff date, as shown in table 1, was divided equally into two subperiods and then, for each of the periods, several metrics were calculated such as revenue, average ticket, number of orders, published products, inactivity days, etc. Then, the performance of the sellers was calculated based on these metrics using the equation where and are calculated values for each subperiod.

# Results and Discussion

## Likelihood Ratio Test

To put the first page of your document in the proper format, copy the content of the first page from this file into your document and fill in the abstract information.

### Author information (added after paper has been accepted)

Once the Editor has accepted your paper, insert the full name, the affiliation (University or Company), City, Country, and email address for each author into the table on the first page. Insert additional rows if there are more than two authors. If there is more than one author, place an asterisk after the corresponding author’s name.

Also include one or two short paragraphs and a head-and-shoulders photo (if available) for each author at the end of the paper.

### What to put in the abstract

The abstract is a brief summary of the contents of the article; it should give enough information to make the reader want to learn more about your research, but it needs to be concise. In each of the boxes on the first page, enter one or two brief sentences appropriate for your paper. Omit citations from the abstract; citations for the paper’s sources do appear in the body of the paper.

If any of the abstract subtopics do not apply to your paper, put “NA” in that box.

## Diagnosis Analysis

### Introduction or background

All papers should begin with an introduction that sets the stage for the discussion. For some disciplines, it is more appropriate to use Background as an alternative first section.

### Body

The body is a collection of multiple sections describing the main content of the paper. You should use up to three levels of headings to categorize content as deemed necessary: Heading 1, Heading 2, and Heading 3.

### Conclusion

This section summarizes the paper, presents challenges, suggests future study, and so on to create a lasting impression of the paper.

### Appendix

If there is an appendix, place it after the References and before the Biography. If there is more than one appendix, add a letter after “Appendix.”

If you have used a questionnaire in a study, include a copy of it as an appendix.

### References

Following the conclusion is a list of all references used in the body of the paper. The current APA formatting guidelines are used to make internal citations within the body as well as provide the complete alphabetic list of reference citations at the end of the paper. (See the Entering References section for more details. We have placed a summary of these guidelines at <https://www.informingscience.org/Uploads/APA_7ed.pdf> ) The References list contains **only** works cited in the paper and **all** works cited in the paper must be listed in the References section.

## Analysis and Interpretation of Parameter Estimates

After the paper is accepted, for each author of the paper, please provide a one or two paragraph biography that describes the author’s background relevant to this article. If you have one, insert a head-and-shoulder photo to the left of the biography of each author or send it separately and we will insert it.

## Prediction Power of the Model

After the paper is accepted, for each author of the paper, please provide a one or two paragraph biography that describes the author’s background relevant to this article. If you have one, insert a head-and-shoulder photo to the left of the biography of each author or send it separately and we will insert it.

# Conclusions

To make it easier to read the paper online, use single column formatting for the paper.

## Page Size

Set the paper size to Letter, which is 8 1/2 by 11 inches.

## Margins

Select Mirror Margins. Top and bottom margins should be 1 inch. Set the inside margin to 1.5 inches and the outside margins to 1 inch. (If you have problems doing this, leave the left and right margins at 1.25 inches.)

## Headers and Footers

Insert page numbers in the footer. We will add the remaining information for the headers, and footers.

## Hyphenation

Hyphenate the text in the document. To turn on hyphenation:

* Select Language on the Layout or Page Layout menu.
* Select Hyphenation.
* Check Automatically hyphenate document.

## Footnotes

**Footnotes should not be used at all**. Insert your note within the body of the paper (if it is important) or omit it. The editors will remove most footnotes and place the material within the text.

# Conclusion

If you have not attached these styles to your paper, use Word’s default paragraph styles for your document, making just the changes indicated below.

## Headings (This is a Heading 2 Style)

**Do not number headings**. Enter the headings with no outline numbers or letters in front of them.

### Paper title

Use the Paper Title style for the title of your paper. It is centered with a border under it. The font for this style is Garamond, 16 point, Bold, and Small Caps.

The paragraph formatting is Centered with a 6 point space after it and an underline border at the bottom. The “Keep with Next” property is selected.

Capitalize the first letter of every major word. Do **not** use all upper case.

### First level headings

Use the **Heading 1** style for the title and for major headings. The font for this style is Garamond, 16 point, Bold, and Small Caps.

The paragraph formatting has a 3 point space before and a 6 point space after it and an underline border at the bottom. The “Keep with Next” property is selected.

Capitalize the first letter of every major word. Do **not** use all upper case.

### Second level headings

Use the **Heading 2** style for second level headings. The font for this heading is Garamond, 14 point, Bold, Italic and Small Caps. The space before the paragraph is 12 point and the space after is 3 point. The “Keep with Next” property is selected.

Capitalize the first letter of every major word in second level headings. Do **not** use all upper case.

### Third level headings (This is a heading 3 style)

Use the **Heading 3** style for third level headings. The font for this heading is Garamond, 12 point, Bold. The space before the paragraph is 12 point and the space after is 3 point. The “Keep with Next” property is selected.

Capitalize only the first word and proper nouns in this heading.

## Text Paragraphs

Use the **Normal** style for paragraphs of text. The paragraph is single-spaced with **no** indentation and has a 6-point space after it. The font for this style is 11 point Garamond. Do **not** put blank lines between paragraphs.

### Other types text of paragraphs

**Forth level.** Three levels of headings are enough for most papers. If you need another level, such as for this paragraph, use the Normal style and place the heading at the beginning of the paragraph in bold font.

**Lists**. Use Word’s automatic bullet or number formats for lists.

**References:** Use a 10 point Garamond font with a hanging indent of 0.25 inches.

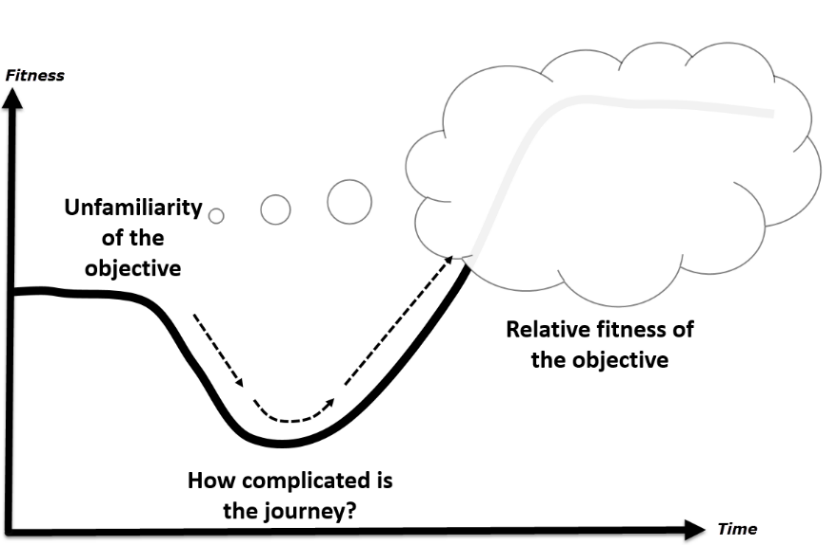
**Other**. Use other formats only when absolutely necessary.

# References

A table is data presented in tabular format with rows and columns. A figure is any other pictorial representation of data such as graphs or drawings. Each figure or table must be numbered and have a brief caption that describes it. Every figure or table **must be referenced** in the body of the paper. Table 1 is an example of a table and Figure 1 is an example of a figure.

**Table 1. Example of a table**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID# | LAST | FIRST | CATALOG # | CATEGORY | QUANTITY | AMOUNT | COMMISSION |
| S00001 | Golden | Rod | M00002 | Multiple | 2 | $250.00 | $12.50 |
| S00001 | Golden | Rod | M00012 | Hiking | 1 | $50.00 | $2.50 |
| S00001 | Golden | Rod | M00028 | Multiple | 1 | $95.00 | $4.75 |
| S00002 | Red | Rose | M00038 | Multiple | 1 | $35.00 | $1.92 |



**Figure 1. Example of a figure**

The caption for tables is placed above the table; the caption for figures is placed below the figure.

All accepted papers need to be reformatted before publication; therefore, **it is important that all figures and tables can be easily resized and/or moved**. Since tables and figures may be moved during the final formatting, do not use “above” or “following” when referring to them; just give the table or figure number. Also, do not use automatic numbering of tables and figures as these can become corrupted when figures or tables have to be rearranged. Tables and figures may be inserted directly into the paper or placed on separate pages at the end of the paper.

There are two ways to submit figures.

1. insert the figure, formatted as a picture that can be resized. It is best if it is inserted directly into the paper with “in line” wrapping.
2. send figures as **PowerPoint slides** in a separate file and, within the paper, indicate where they are to appear,

**Do not** send figures formatted as separate text boxes or groups of images on the page.

# Conclusion

We will publish your paper under a Creative Commons Attribution-NonCommercial 4.0 International License and the author retains the copyright.

By submitting the paper, as author you certify the following:

1. You hold copyright for this submission, and
2. You warrant that you have not infringed on any copyright and assume full liability in case of copyright dispute.

## Copyright Issues for Figures

There are three common sources of figures.

1. Figures you have copied from another source, including a web site. You must contact the holder of the copyright for the image and get permission to use it. Cite the source and add “used with permission.”
2. Figures that you create based on another’s work. You do not need to get permission but include in the citation “adapted from” or “based on” and give the source.
3. Figures that are your original work. Since you hold the copyright for these, there are no copyright issues.

# References

References are to follow the current American Psychological Association (APA) guidelines. We have placed a summary of these guidelines on the web at   
<https://www.informingscience.org/Uploads/APA_7ed.pdf>

List the sources alphabetically at the end of the paper under “References” using a Heading 1 style. Place entries in alphabetical order according to the last name of the first author. Within the text of your paper, cite sources by placing the author’s last name and the date in parentheses.

**Reference list**. In the reference list, when a work has **up to (and including) 20** authors, list all authors (last name followed by initials). Place a comma after the last name of each author and after that author’s initial(s). Place an ampersand (&) before the last author. If there are **more than 20 authors**,provide last names and initials of first 19 authors, insert three ellipsis points, and add the last author’s name. Follow the authors by the date, the title, and the source.

**In-text citations**. If a work has two authors, include both authors in each in-text citation (Boyd & Cohen, 2003). If the work has three or more authors, **in all in-text citations** place only the first author followed by et al. (Gill et al., 2019).

**DOIs and URLs** **in Reference list**. When citing sources from the Web, include the year of publication or the most recent update. End the entry with a DOI if it has one. If it does not have a DOI, end the entry with the URL. **Do not end the path statement or a DOI with a period**. DOIs can be found by going to <https://search.crossref.org/references> and entering the reference entry. DOIs begin with https://doi.org/ followed by numbers and letters that identify the document, for example, https://doi.org/10.28945/2714

# Author

(Leave this blank when submitting for review.)

Include one or two short paragraphs about each author. Please include a head and shoulder photo of each author. You should include this photo and bio when you are asked to upload your final, formatted, camera ready copy. However, if you need to, you can send the photos by email to [Publisher@InformingScience.org](mailto:Publisher@InformingScience.org) as separate attachments and we will insert them for you.