Predict and Prevent Customer Churn Using Data Science

Paste the appropriate copyright/license statement here. ACM now supports three different publication options:

• ACM copyright: ACM holds the copyright on the work. This is the historical approach.

• License: The author(s) retain copyright, but ACM receives an exclusive publication license.

• Open Access: The author(s) wish to pay for the work to be open access. The additional fee must be paid to ACM.

This text field is large enough to hold the appropriate release statement assuming it is single-spaced in Verdana 7 point font. Please do not change the size of this text box.

Each submission will be assigned a unique DOI string to be included here.

# Abstract

Jaimie Cain

Bellevue Univerisy

Bellevue, NE 68005, USA

author1@anotherco.edu

Manish Shekar

Bellevue Univerisy

Bellevue, NE 68005 author2@author.ac.uk

Jedediah Brough

Bellevue University

Bellevue, NE 68005  
author4@hci.anotherco.com

Vidyasagar Kummarikunta

Bellevue University

Bellevue, NE 68005

author5@yetanotherco.ca

Julia Cuellar

Bellevue University

Bellevue, NE 68005

author6@author.fr

Purpose of the project is to study customer’s behavior and data to predict and possible prevent customer churn using data science techniques. Customer churn occurs when customers or subscribers stop doing business with a company or service. Also known as customer attrition or turnover, it is a critical metric due to the cost savings for a company to retain a customer in comparison to acquiring new ones. Awareness about customer churn helps an organization define customer retention processes, project goal success rates and identify strategies for improvement.

As part of the project, we will research and explore various data science tools and techniques to include: data prep, modeling and discovery. We will use patterns obtained from customer’s data to identify a customer’s potential to churn and explore ways to provide guidance via visualizations and reporting. This information would then be available to the businesses to help define retention strategies. Some common customer churn causes that we will explore are: poor customer service, poor onboarding process, lack of brand loyalty and an inability to maintain product quality.

# Author Keywords

Customer churn, Prediction models, Logistic regression, etc.

# ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous; See<http://acm.org/about/class/1998> for the full list of ACM classifiers. This section is required.

# Introduction

One of the key aspects of a successful business model is how to establish and maintain a loyal customer base. If a company can keep their customers in a subscription-based revenue system, they can maintain a strong financial foundation and increase their return on investments. The cost savings can then be used to increase product technology and to create a more divers marketing strategy to provide support to decrease customer churn. Through a stable customer base the company can establish a strong customer review system and increase the likelihood of customer referrals as a source of increased product subscriptions. Some of the more common issues that can increase customer churn are constantly changing product technology, the low cost for customers to switch companies, a large selection of competitors in a saturated market, and decrease oversight of corporations. The company can directly contribute to increased customer churn by providing underperforming products, increasing costs passed onto the customer, providing poor customer service and nto taking advantage of targeted marketing to existing customers. If a company is able to take advantage of the large amount of data provided by the customers through customer service interactions, survey and online reviews and the base demographical data and usage they can begin to analyze where the deficiencies are and predict what customers are at more risk for attrition.

# Data Science Methods Used for Data Analysis

# Good Utilization of the Side Bar

**Preparation**: Do not change the text box size or position. Do copy text box to other pages. You may change the surrounding box to be visible or invisible, up to you.

**Materials:** This cannot appear higher or lower on the page because of pagination and specific headers added during the indexing and pagination process. A 0.75 inch rule is beneficial to break this apart from the body text. The text in this text box should remain the same size as the Body Text: 8.5 Verdana or Arial (with use of **bold** and *italics* to highlight points)

**Images & Figures:** Images and figures can be placed in this section. They should be captioned in the manner of other images and figures.

Business Understanding

Creating a model to help companies predict a customer’s likelihood to churn, will require analyzing and creating usable data from a large customer data set. Customer demographic data, contractual agreements, online reviews and product usage are some data sets to be considered. Using the core CRISP-DM workflow the data will need to flow through several steps to evolve from a large data set to usable data. The first step, **Business Understanding**, is to outline the problem and brainstorm for possible causes that are consistently associated with customer churn. Once the data is pulled from the appropriate sources, whether that be social media, online reviews, customer satisfaction surveys, product usage and base demographical data it will need to be reviewed to start exploring any discernable trends within the data, **Data Understanding**. This process is typically completed by data visualization using histograms, pie charts and correlation matrix, which will help identify trends within the data and provide the groundwork for the next step, **Data Preparation**. Data will need to be split into a training data set, validation data set and a test data set. The data will be normalized to help identify any outliers or redundant values that would skew the data analysis and the data will need to be balanced so the sampling is consistent with the larger data set. The next process, **Modeling**, will be to fit and train the logistic regression model and then predict for the test data set. Once prediction output is ready, it will go through the **Evaluation** step using a confusion matrix and receive an accuracy score, F1 score, precision score and a recall score. Based on how the scores are returned (~50% or greater) the user will know if the data set is reliable to use and the predictive models are working correctly. The last part of evaluation is to review the data to see what columns are not helping prediction and remove them from the feature list. This modeling and predictive method are then re-applied to the trimmed feature data set to ensure it provides the same accuracy in prediction as the larger data set. Model is now ready to be evaluated for **Deployment**. The scripting will need to be evaluated for its impact on the system regarding usage and the timeliness in pulling predictive values on the larger data sets. Process flow below shows how we plan to utilize CRISP-DM and intricate data science concepts to achieve customer churn prediction.

1111111111111111111111111111111111111111111

# Figures and Tables

The examples on this and following pages should help you get a feel for how screen-shots and other figures should be placed in the template. *Be sure to make images large enough so the important details are legible and clear.* Your document may use color figures, which are included in the page limit; the figures must be usable when printed in black and white.



**Figure 1**: Insert a caption below each figure. We suggest selecting the image and then using Insert, Caption. Then if you use Insert, Cross-Reference then your Figure numbering referencing will be consistent. Make sure you use the Caption style for text formatting.

If you aren’t familiar with Word’s handling of pictures, we offer one tip: the “format picture” dialog is the key to controlling position of pictures and the flow of text around them. You access these controls by selecting your picture, then choosing “Picture…” from the “Format” menu. As for the “picture” tab in that dialog, we recommend using Photoshop, Preview, or other graphics software to scale images, rather than scaling them after you have placed them in Word.



Figure : In this image, the cats are tesselated to save space. You, too, can save space by placing images in the sidebar. Images should have captions and be within the boundaries of the text box on Page 2. Photo CC-BY jofish on Flickr.

| **Objects** | **Caption –  pre-2002** | **Caption – 2003 and beyond** |
| --- | --- | --- |
| Tables | Above | Below |
| Captions | Below | Below |

**Table 1**: Tables should have the caption below. Use Table – SIGCHI style. Use 0.75 rules/borders at 75% grey for your tables, align decimals or center text in the cells. For improved accessibility, header rows of tables should be marked. In Word, right click a header row, and select Table Properties | Row | Repeat as header at the top of each page. Avoid spurious decimal points.



Figure : You can make figures as wide as you need, up to a maximum of the full width of both columns. To achieve this, select the figure and the caption, choose “columns” under the “Format” menu, pick the “One” (single column) icon at the top of the dialog, and make sure you are making the change only for “selected text” (at the bottom of the dialog). Image CC-BY-ND ayman on Flickr

Page 5 shows a treatment of large figures, too big to fit inside a single column of text. All figures should include alt text for improved accessibility. In Word, right click the figure, and select Format Picture | Alt Text).

11111111111111111111111111111111111111111111

# Case Study Examples

|  | **First** | **Second** |
| --- | --- | --- |
| child | 22 | 44 |
| adult | 22 | 16 |
| Gene | 22 | 11 |
| Cliff | 34 | 22 |

Table : A narrow table in the margin

So long as you don’t type outside the right margin, it’s okay to put annotations over here on the right, too. Remember to use the annotation text style.

Elder Research:

Elder Research worked with a banking company to predict and reduce customer churn. The banking company wanted to reduce customer churn by at least 10% using internal data. Elder Research was able to aggregate the data and reduce the number of factors to about 30. After testing various models including regressions, neural networks and random forests, Elder Research was able to create a model that was 20% more accurate than the existing method of predicting customer churn and its indicators.

Comtec:

Comtec partnered with a telecommunications company to find indicators of churn and lower churn rates for the company. The company was interested in identifying factors leading to high churn rates. Comtec used regression models to identify indicators of churn and to predict which customers had a high churn risk. The resulting model had between 60%-80% accuracy. Key factors including a delay in service were identified as leading to high churn risk.

# What is the deliverable for this project?

1. Pre-trained (supervised) model to predict customer churn. We will try multiple algorithms and parameter tunings to get best outcome. We will also consider integrating multiple models for best results.
2. NLP (Natural Language Processing) model to understand customer sentiment overall and for each product (positive, negative, neutral).
3. Process integration with marketing & sales screens to provide real time customer insights, early warnings, and custom guidance for retention.
4. Executive dashboards to understand customer behavior and attributes trend over time for churning customers.
5. Executive dashboards to analyze customer churn reasons (i.e. product issues, bad customer service, bad onboarding, budding competition, etc.) and predict trends over time for customer churn.

# Conclusion

1. It’s cheaper: 70% of companies say that it is cheaper to retain existing customers than to acquire new ones, while others have suggested that cost of acquiring new customers can be as much as seven time.
2. Its faster: Again, it’s often much easier to sell to an existing customer than it is to sell to a new one. This is because all barriers to purchase have already been overcome. Existing customers may just need a slight nudge to increase your share of their wallet, and this will speed up your sales process.
3. It better positions your business: One of the best things about high levels of customer retention is that you’re able to build a more nuanced view of who your customers really are. On one hand, this data is vital to your customer retention efforts and gives you the tools to create personalized and targeted promotions to maximize your chances of success.
4. Retention = Acquisition: Customer loyalty is priceless and can even result in further customer acquisition for your company. After all, word of mouth advertising is not only free, but possible one of the most credible forms of advertising.
5. Closing thoughts: Marketers don’t tend to focus on customer retention enough because loyalty and engagement are often not seen as strictly “measurable”, but this doesn’t mean they aren’t important. Customer acquisition is critical in the early stages of a start-up business, but once you’ve built a customer base, as soon as you’ve got one customer, retention should be on your mind. The more customers you have, the more important retention is.

# Acknowledgements

Thanks to work from home situation that inspired us to pursue MSDS degree. Also, thanks to churning customers everywhere, without whom we won’t have this problem and any training data.

# References

1. @\_CHINOSAUR. 2014. VENUE IS TOO COLD. #BINGO #CHI2016. Tweet. (1 May, 2014). Retrieved February 2, 2014 from https://twitter.com/\_CHINOSAUR/status/461864317415989248
2. ACM. How to Classify Works Using ACM’s Computing Classification System. 2014. Retrieved August 22, 2014 from http://www.acm.org/class/how\_to\_use.html.
3. Ronald E. Anderson. 1992. Social impacts of computing: Codes of professional ethics. *Soc Sci Comput Rev* 10, 2: 453-469.
4. Anna Cavender, Shari Trewin, Vicki Hanson. 2014. Accessible Writing Guide. Retrieved August 22, 2014 from http://www.sigaccess.org/welcome-to-sigaccess/resources/accessible-writing-guide/
5. Morton L. Heilig. 1962. Sensorama Simulator, U.S. Patent 3,050,870, Filed January 10, 1961, issued August 28, 1962.
6. Jofish Kaye and Paul Dourish. 2014. Special issue on science fiction and ubiquitous computing. *Personal Ubiquitous Comput*. 18, 4 (April 2014), 765-766. http://dx.doi.org/10.1007/s00779-014-0773-4
7. Scott R. Klemmer, Michael Thomsen, Ethan Phelps-Goodman, Robert Lee, and James A. Landay. 2002. Where do web sites come from?: capturing and interacting with design history. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '02), 1-8. http://doi.acm.org/10.1145/503376.503378
8. Psy. 2012. Gangnam Style. Video. (15 July 2012.). Retrieved August 22, 2014 from https://www.youtube.com/watch?v=9bZkp7q19f0
9. Marilyn Schwartz. 1995. *Guidelines for Bias-Free Writing.* Indiana University Press, Bloomington, IN.
10. Ivan E. Sutherland. 1963. *Sketchpad, a Man-Machine Graphical Communication System*. Ph.D Dissertation. Massachusetts Institute of Technology (MIT), Cambridge, MA.
11. Langdon Winner. 1999. Do artifacts have politics? In *The Social Shaping of Technology* (2nd. ed.), Donald MacKenzie and Judy Wajcman (Eds.). Open University Press, Buckingham, UK, 28-40.