

---

# As The World Churns: Customer Data, Business Models, and Predicting Customer Trends and Behaviors

**Rahul Gupta**

rgupta@my365.bellevue.edu

**Tushar Muley**

tmuley@my365.bellevue.edu

**Michael Koffie**

mmkoffie@my365.bellevue.edu

**Edris Safari**

esafari@my365.bellevue.edu

**Brandon May**

brmay@my365.bellevue.edu

## Abstract

A standard business task in customer-relationship management is to estimate the likelihood that an individual customer will perform an action. The term propensity modeling is used to describe the task of taking action with the model's goal to narrow activities

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org) or Publications Dept., ACM, Inc., fax +1 (212) 869-0481.

or actions the individual will perform. Prediction in machine learning can go beyond the customer-relationship management. The model can calculate the likely outcomes based on historical data. The patterns found in historical data forecast the likelihood of related events occurring in the future. Customer churning means defection of customers to other competitors and frequent acquisition and then the loss of customers in a short period of time resulting in lost resources.

Data collection methods and data warehousing practices along with data mining techniques are critical components of a prediction model. In these models, multiple independent variables contribute to the value of a single dependent variable. Various papers present the applications of predictive data analysis and present

the algorithms instrumental in ascertaining accurate results.

Our goal here is to do a critical analysis of theory behind predictive analysis and current methods to analyze the future direction of data science in the growing field of customer behavior, specifically customer churning in financial industries.

### **Author Keywords**

Customer churn; predictive analytics; k-means; tableau; decision tree analysis; data science.

### **ACM Classification Keywords**

<https://web.archive.org/web/20090126131933/http://acm.org:80/about/class/ccs98-html>

### **Introduction**

As our economy and corporations begin to operate in a global context, there have been increasing efforts to retain customers. Frequent acquisition and loss of customers is defined as customer churn and has been a particular area of focus in data science, especially for “high-value customers” [1]. It is important to businesses as “it is directly tied to firm profitability” [4]. The costs of keeping a customer are usually less than the costs of recruiting new customers [21]. This is why it is becoming increasingly important to use data science techniques and advanced analytics to predict which customers are vulnerable to leaving. It can be difficult to differentiate between customers who will respond to interventions and those who will not and some advocate looking at individual behaviors and probabilities of defection in order to segment that customer population [4]. In addition, excessive customer turnover can be a sign of potentially

fraudulent activity. This is complicated by the fact that technology can serve two purposes: to become closer to customers as well as alienate them [21].

There is also the risk of customer churn on customers that were won back after churning originally. This makes the situation even more complex to further analyze [16]. There are multiple fields to study in this; some organizations use predictive modeling by studying customer behavior while others focus on more traditional demographics [14].

Traditionally, the data science technique of k-means clustering is used to determine risk of customer churn [23]. However, there are multiple other methods that may be valid.

The risk can be more than financial; in certain insurance industries, customer churn can signify loss of critical healthcare coverage and can significantly impact a person’s health. In fact, data science technique and predictive analytics are being applied to treat cancer and impact healthcare outcomes [5,17,18,20]. Churning has also been applied to employee turnover and its effect on business operations [6]. Therefore, it benefits us all both economically and personally to obtain further insight into customer churn, its prediction, and its avoidance (if at possible). This project aims to critically evaluate the current state of customer churn and customer behavior in the financial and insurance industry, propose a data science framework and algorithm to ascertain customer churn, and reflect on the future direction of this field.

## Why Is This Data Science?

### “Data is the new oil.” (Clive Humby, 2006)

Data Science transforms raw data into useful information. Industries require data to help them make careful decisions and is used in almost every industry including health, finance, and banking industries to name a few. Companies use the data to analyze their marketing strategies and create better advertising. The industry needs data scientists to help them make smarter decisions that make financial sense [11,28].

Let us understand the importance of data science in our lives. Getting a ride with Uber is easy. We simply open the app, set your pick-up and drop-off location, book a taxi, get picked up and pay with your phone. Each time you book a taxi through Uber, you will receive an estimated fare and the time it takes to travel the route. How can these applications display all of the information they do? The answer is data science. Using data science predictive analytics, Uber can determine the pick-up, drop-off location and arrival time with ease.

Technology giants such as Facebook, Amazon and Google are constantly working in the field of machine learning and data science. Data science encompasses processes such as purging, processing, and analyzing data. A data scientist collects data from multiple sources, e.g. from surveys and physical data plots. Then data is passed through strict algorithms to extract important information from the data and create a record. This record could also be used to parse algorithms to make more sense [22].

According to DOMO research, "More than 2.5 billion bytes of data are created every day and will only grow from there, and by 2020 an estimated 1.7 MB of data will be generated per second for every human being on Earth."

Predicting customer attrition and churn is a shining example of the data science process including determining opportunities within a business, generating a hypothesis, selecting and finding applicable data, analyzing that data, and then generating conclusions and potential courses of action from that project. Customer churn and attrition presents yet another opportunity for data science to flex its capabilities in a modern world.

## Deliverables

The main end point of our analysis is to hypothesize different models and techniques to reduce potential customer churn by 50%. Our goal has been divided into three different stages:

- The short-term goal is to reduce churn by 20% in 1 year.
- The medium-term goal is to reduce churn an additional 20% in 2-4 years.
- The long-term goal is to reduce churn by an additional 10% for a total 50% reduction in 5 years.

To accomplish these tasks, we would build various machine learning algorithms to understand root causes of churn and identify at risk customers including:

- Neural Networks
- Decision Tree
- K-Means
- Random Forest
- Logistic Regression

This would be ideally augmented with designing dashboards and visualizations on the fly to analyze customer data with Tableau to explain current progress

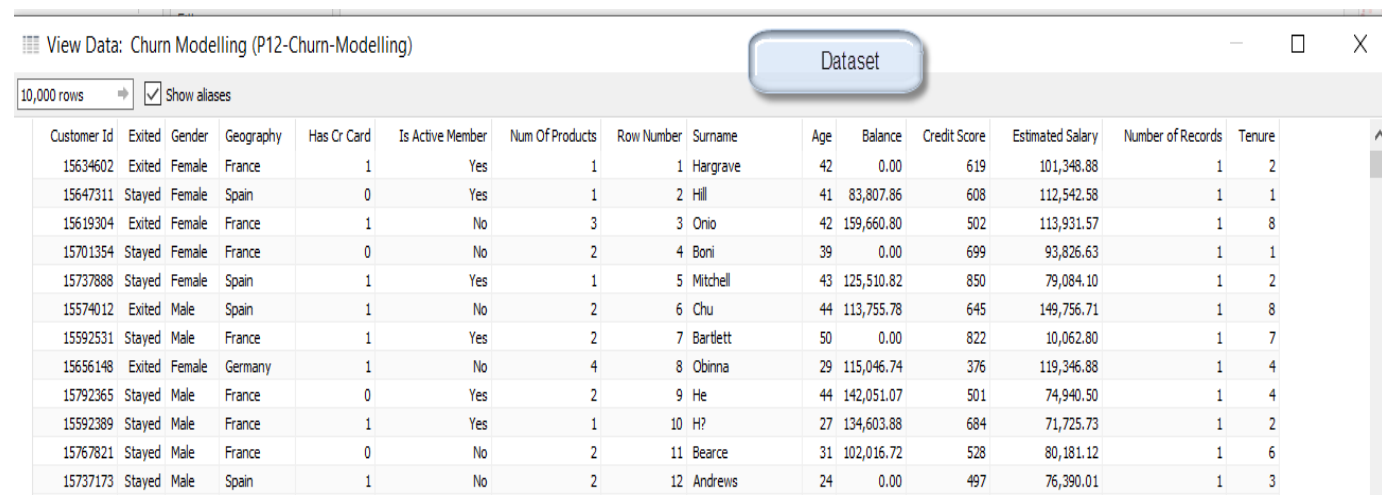
and trends in an understandable way to key stakeholders.

Once we would identify optimal algorithms using one or a combination of these techniques, we would focus resources on customers that are at high risk of leaving in an attempt to retain their business. This would require adopting customer focus strategies to retain and reduce the rate of customer churn.

Over time, the process would be refined and progress would be measured quarterly.

## Data Mining

To accomplish this task, we decided to apply a regression logistics algorithm, which is commonly used in predicting binary outcomes. In the case of churning, we want to predict if a new customer or existing customers is likely to stay or exit. We proposed examining a dataset composed of 10,000 records. The 'exited' column in this dataset is regarded as the dependent variable which is the subject of this analysis, and the rest of the variables are the independent variables. A sample data set is described in Figure 1.



Customer Id	Exited	Gender	Geography	Has Cr Card	Is Active Member	Num Of Products	Row Number	Surname	Age	Balance	Credit Score	Estimated Salary	Number of Records	Tenure
15634602	Exited	Female	France	1	Yes	1	1	Hargrave	42	0.00	619	101,348.88	1	2
15647311	Stayed	Female	Spain	0	Yes	1	2	Hill	41	83,807.86	608	112,542.58	1	1
15619304	Exited	Female	France	1	No	3	3	Onio	42	159,660.80	502	113,931.57	1	8
15701354	Stayed	Female	France	0	No	2	4	Boni	39	0.00	699	93,826.63	1	1
15737888	Stayed	Female	Spain	1	Yes	1	5	Mitchell	43	125,510.82	850	79,084.10	1	2
15574012	Exited	Male	Spain	1	No	2	6	Chu	44	113,755.78	645	149,756.71	1	8
15592531	Stayed	Male	France	1	Yes	2	7	Bartlett	50	0.00	822	10,062.80	1	7
15656148	Exited	Female	Germany	1	No	4	8	Obinna	29	115,046.74	376	119,346.88	1	4
15792365	Stayed	Male	France	0	Yes	2	9	He	44	142,051.07	501	74,940.50	1	4
15592389	Stayed	Male	France	1	Yes	1	10	H?	27	134,603.88	684	71,725.73	1	2
15767821	Stayed	Male	France	0	No	2	11	Bearce	31	102,016.72	528	80,181.12	1	6
15737173	Stayed	Male	Spain	1	No	2	12	Andrews	24	0.00	497	76,390.01	1	3

Figure 1: A sample data set used to analyze customer churning segmented into different variables.

After performing data mining in software like Tableau, we built a logistic regression model in Gretl and performed 5 backward eliminations. We made dummy variables 'Spain', 'Germany' and 'France' from the 'Geography' set and 'Male' and 'Female' variables from

'Gender' set. We included 'Female', 'Spain' and 'Germany' in the model along with the other independent variables in the dataset. Tableau showed correlation with the variable "Has Credit Card" and "Is Active" at the 20% stayed/exited reference line.



BWElimination Number	Variable eliminated	Variable P-Value	Model's Adjusted R-squared before/after removal	Adjusted R-Squared Difference
1	Spain	0.6181	0.150787/ 0.150961	0.000174
2	HasCrCard	0.4489	0.150961/ 0.151102	0.000141
3	EstimatedSalary	0.3091	0.151102/ 0.151197	0.000095
4	Tenure	0.0873	0.151197/ 0.151106	-0.000091

Figure 3 Figure 4: This table shows the summary of the results of the elimination that Gretl recommended.

As shown in elimination 4 'Tenure' was removed, but not by recommendation from Gretl, but because we wanted to see the impact of removal to test the p-value threshold. It showed that the Adjusted R-Squared was not significantly impacted, so we reincluded 'Tenure' in

the model. After transforming the 'Balance' variable to  $\text{Log}_{10}(\text{Balance} + 1)$  for better uniformity, we got the result shown below.

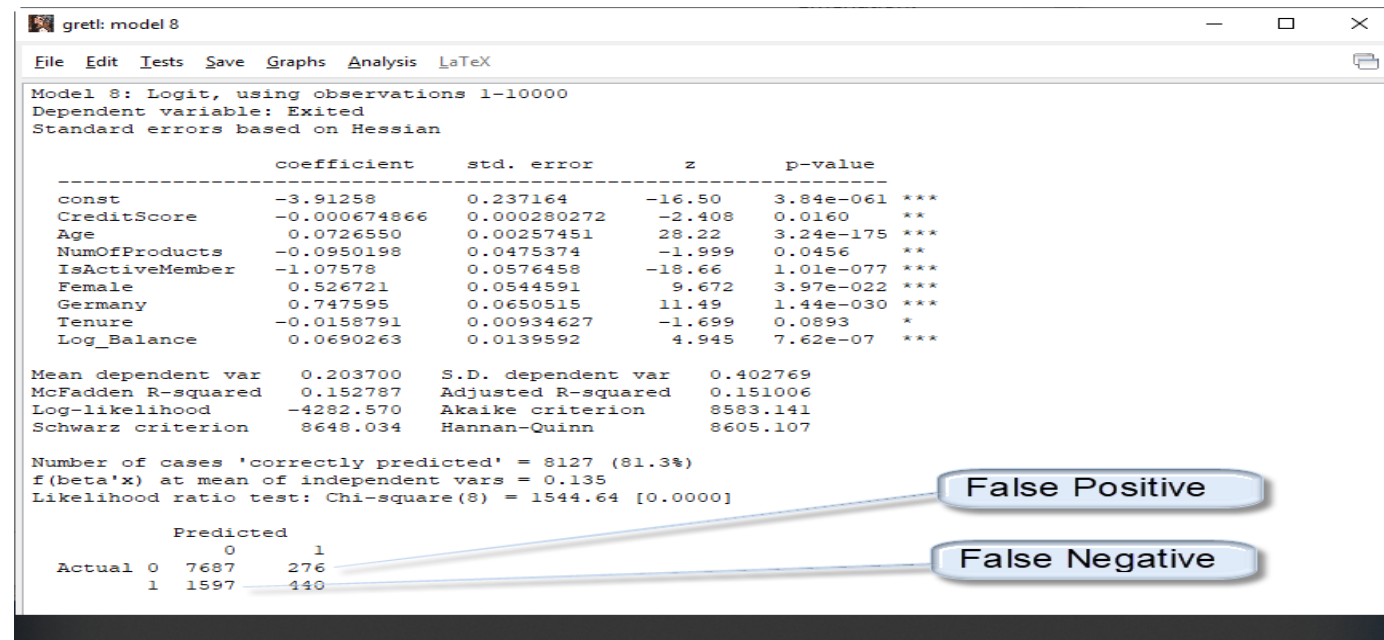


Figure 5: This screen capture image describes the calculation of our accuracy and error rates.

From the confusion Matrix, the accuracy and error rates were calculated as shown below:

Accuracy Rate = Correct/Total =  $(7687+440)/10000 = 81.27\%$

Error Rate = Wrong/Total =  $(276+1597)/10000 = 18.73\%$

Based on the current analysis, we deem the accuracy rate to be unsatisfactory and would look to maximize

this rate by introducing new variables from our data mining.

## Conclusions

Customer churn for any company is costly, but it is especially expensive in customer service areas. The financial and insurance sector has a large amount of competition and with new digital-only institutions coming into the arena the amount of competition has only grown. The biggest reason sighted for leaving a bank was "poor service" and high fees. High fees are intrinsic and dependent upon each company's profitability. Each institution attempts to show its value

by offering the most economic service for the customer while remaining profitable. Poor service was the driver for 56% of the individuals that changed banks [13]. Financial institutions and insurance companies struggle with customer churn as the institutions do not usually see the customer leaving before they have closed all of their accounts and have left. Normally, they never get a chance to try and attempt to retain the customer.

There are challenges to predicting customer attrition at a higher rate. The most obvious issue is in the data being used. The data usually require cleansing or preparation. The most common method to analyze data in this context, decision tree analysis, usually lacks flexibility. Decision trees are based on expectations and if the data arrives with unexpected inputs, the model has a lower accuracy rate [19]. This is just one example showing how many models suffer from incorrectly selected variables, which can have a big influence on the type of model being run. The logit-model is one example where selecting the correct variables has a big influence on the success rating of the model.

Many companies have implemented different techniques to manage customer attrition like neuro-fuzzy, k-means, spatio-temporal, linear regression, decision tree, logit-model (logistics regression) or some combination of two or more techniques [2,3,29,30]. These are all used to increase the ability to predict with the highest accuracy which customer will defect to other institutions. One item of particular interest is regarding the type of data being used. Many of the models used some type of tenure information or social behavioral input. Due to the variations in the type of customer data customer churn is still hard to predict

with a high rate of accuracy. Most human decisions are emotional and usually follow some type of adverse experience with a service the institution provides. Some sources advocate a behavioral analysis technique to predict which customers will respond positively to attempts to retain them as customers to better concentrate resources [4].

Overall, the various methods used to study customer churn in the financial industries and other companies all have pros and cons and one model is not necessarily superior to all of the other models. The key step is making sure that the data you are selecting is appropriate based on the model you choose. This field is diverse and has many different advancements and is another example of data science having a significant impact on everyday business transactions.

### **Acknowledgements**

We would like to thank our DSC 500 colleagues, each other, and our professor Shankar Parajulee for encouragement, lively discussion, and generating further interest and debate in the growing field of data science.

### **References**

1. Abbasimehr, H., Setak, M., & Soroosh, J. (2013). A framework for identification of high-value customers by including social network-based variables for churn prediction using neuro-fuzzy techniques. *International Journal of Production Research*, 51(4), 1279–1294. <https://doi-org.ezproxy.bellevue.edu/10.1080/00207543.2012.707342>



2. Al-Shboul, B., Faris, H., & Ghatasheh, N. (2015). Initializing Genetic Programming Using Fuzzy Clustering and Its Application in Churn Prediction in the Telecom Industry. *Malaysian Journal of Computer Science*, 28(3), 213–220. <https://doi-org.ezproxy.bellevue.edu/10.22452/mjcs.vol28no3.3>
3. Amin, Adnan & Anwar, Sajid & Adnan, Awais & Nawaz, Muhammad & Aloufi, Khalid & Hussain, Amir & Huang, Kaizhu. (2016). Customer Churn Prediction in Telecommunication Sector using Rough Set Approach. *Neurocomputing*. 10.1016/j.neucom.2016.12.009. <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6>
4. Ascarza, E. (2018). Retention Futility: Targeting High-Risk Customers Might be Ineffective. *Journal of Marketing Research*, 55(1), 80–98. doi: 10.1509/jmr.16.0163
5. Brugnoli-Ensin, I., & Mulligan, J. (2018). Instability in Insurance Coverage: The Impacts of Churn in Rhode Island, 2014-2017. *Rhode Island Medical Journal*.
6. Della Torre, E., Zatzick, C. D., Sikora, D., & Solari, L. (2018). Workforce churning, human capital disruption, and organisational performance in different technological contexts. *Human Resource Management Journal*, 28(1), 112–127. <https://doi-org.ezproxy.bellevue.edu/10.1111/1748-8583.12167>
7. Farquhar, J. D. (2005). Retaining customers in UK financial services: The retailers' tale. *The Service Industries Journal*, 25(8), 1029–1044. doi: 10.1080/02642060500237478
8. Ferreira, P., Telang, R., & Matos, M. G. D. (2019). Effect of Friends' Churn on Consumer Behavior in Mobile Networks. *Journal of Management Information Systems*, 36(2), 355–390. doi: 10.1080/07421222.2019.1598683
9. Flores-Méndez, M. R., Postigo-Boix, M., Melús-Moreno, J. L., & Stiller, B. (2018). A model for the mobile market based on customers profile to analyze the churning process. *Wireless Networks (10220038)*, 24(2), 409–422. <https://doi-org.ezproxy.bellevue.edu/10.1007/s11276-016-1334-8>
10. Gunthera, C-C., Tvetea, I., Aasa, K., Sandnesb, G., & Rorganc O. (2014). Modelling and predicting customer churn from an insurance company. *Scandinavian Actuarial Journal* 2014 Vol. 2014, No. 1, 58–71
11. Imarticus.org. (2018, October 8). Why is Data Science So Famous? - Imarticus Learning. Retrieved from <https://imarticus.org/why-is-data-science-so-famous/>.
12. Jennings, A., & Stratagree. (2015, December 25). The 4 D's of Customer Attrition. Retrieved from <https://thefinancialbrand.com/55772/banking-customer-attrition-analysis/>.
13. Kaemingk, D. (2018, August 29). Reducing customer churn for banks and financial institutions. Retrieved from

- <https://www.qualtrics.com/blog/customer-churn-banking/>.
14. Kaya, E., Dong, X., Suhara, Y., Balcisoy, S., Bozkaya, B., & Pentland, A. "S. (2018). Behavioral attributes and financial churn prediction. *EPJ Data Science*, 7(1). doi: 10.1140/epjds/s13688-018-0165-5
  15. Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, 2(1). doi: 10.1186/s40854-016-0029-6
  16. Kumar, V., Leszkiewicz, A., & Herbst, A. (2018). Are you Back for Good or Still Shopping Around? Investigating Customers Repeat Churn Behavior. *Journal of Marketing Research*, 55(2), 208–225. doi: 10.1509/jmr.16.0623
  17. Menden, M. P., Iorio, F., Garnett, M., Mcdermott, U., Benes, C. H., Ballester, P. J., & Saez-Rodriguez, J. (2013). Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS ONE*, 8(4). doi: 10.1371/journal.pone.0061318
  18. Mousavirad, S. J., & Ebrahimpour-Komleh, H. (n.d.). A Comparative Study on Medical Diagnosis Using Predictive Data Mining. *Data Mining and Analysis in the Engineering Field Advances in Data Mining and Database Management*, 327–360. doi: 10.4018/978-1-4666-6086-1.ch017
  19. Nayab, N. (2019). A Review of Decision Tree Disadvantages <https://www.brighthubpm.com/project-planning/106005-disadvantages-to-using-decision-trees/>
  20. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, 375(13), 1216–1219. doi: 10.1056/nejmp1606181
  21. Panaggio, T. (2015, May). The Customer Comes First - Always. *USA Today*, pp. 52–54.
  22. Quora. (2017, October 25). Why Data Science Is Such A Hot Career Right Now. Retrieved from <https://www.forbes.com/sites/quora/2017/10/25/why-data-science-is-such-a-hot-career-right-now/#72bf6ee9106b>.
  23. Rachid, A. D., Abdellah, A., Belaid, B., & Rachid, L. (2018). Clustering Prediction Techniques in Defining and Predicting Customers Defection: The Case of E-Commerce Context. *International Journal of Electrical and Computer Engineering*, 8(4), 2367–2383.
  24. Roman, O. (2019). Churn prediction. <https://towardsdatascience.com/churn-prediction-770d6cb582a5>
  25. Saha, J. M. (2011). Business sustainability amidst global churning - paradoxes and dilemmas. *GMJ*, 5(1-2), 19–24.

26. Syahida Binti, M. Z., & Ameer, R. (2010). Turnaround prediction of distressed companies: Evidence from Malaysia. *Journal of Financial Reporting and Accounting*, 8(2), 143-159. doi:<http://dx.doi.org.ezproxy.bellevue.edu/10.1108/19852511011088398>.
27. THE FINANCIAL BRAND FORUM 2020 — Discover the big ideas disrupting banking and explore the latest trends redefining the future of financial marketing at the FORUM 2020. The world's most elite conference on marketing. (2017, November 2). Plug Those Leaks: Stop Attrition From Stalling Your Growth Strategy. Retrieved from <https://thefinancialbrand.com/68371/banking-customer-acquisition-attrition-growth-strategy/>
28. Thompson, R. (2017, February 27). Understanding Data Science and Why It's So Important. Retrieved from <https://blog.alex.com/know-data-science-important/>.
29. Vijaya, J., & Sivasankar, E. (2018). Computing efficient features using rough set theory combined with ensemble classification techniques to improve the customer churn prediction in telecommunication sector. *Computing*, 100(8), 839-860. <https://doi-org.ezproxy.bellevue.edu/10.1007/s00607-018-0633-6>.
30. Zoric, A. B. (2016). Predicting Customer Churn in Banking Industry using Neural Networks. *Interdisciplinary Description of Complex Systems*, 14(2), 116-124. doi: 10.7906/indecs.14.2.1