Journal of Information Systems Education

Volume 35
Issue 2
Spring 2024

Teaching Case

Teaching Business Students Logistic Regression in R With the Aid of ChatGPT

Chen Zhong and J.B. (Joo Baek) Kim

Recommended Citation: Zhong, C., & Kim, J. B. (2024). Teaching Case: Teaching Business Students Logistic Regression in R With the Aid of ChatGPT. *Journal of Information Systems Education*, 35(2), 138-143. https://doi.org/10.62273/DYLI2468

Article Link: https://jise.org/Volume35/n2/JISE2024v35n2pp138-143.html

Received: May 15, 2023
First Decision: July 24, 2023
Accepted: August 30, 2023
Published: June 15, 2024

ISSN: 2574-3872 (Online) 1055-3096 (Print)

Journal of Information Systems Education, 35(2), 138-143, Spring 2024
https://doi.org/10.62273/DYLI2468

Teaching Case
Teaching Business Students Logistic Regression in R With the Aid of ChatGPT

Chen Zhong
J.B. (Joo Baek) Kim
Sykes College of Business
University of Tampa
Tampa, FL 33606, USA
czhong@ut.edu, jkim@ut.edu

ABSTRACT

Data Analytics has emerged as an essential skill for business students, and several tools are available to support their learning in this area. Due to the students' lack of programming skills and the perceived complexity of R, many business analytics courses employ no-code analytical software like IBM SPSS Modeler. Nonetheless, generative Artificial Intelligence (AI) services such as ChatGPT can bridge the gap for students lacking programming skills. This teaching case demonstrates how students can use ChatGPT to generate R code for logistic regression analysis of a telecommunication company's customer churn based on the Cross-Industry Standard Process Data Mining approach. ChatGPT enables students to implement the analysis method in R with a focus on building business solutions, freeing them from technical details. Teaching business students to use ChatGPT to implement data analysis is effective in helping them understand data, analytics models, and data interpretation. Moreover, this teaching case provides an opportunity for students to understand how to work with Artificial Intelligence in Data Analytics tasks.

Keywords: Data analytics, Generative AI in teaching, Logistic regression analysis, Natural language programming

1. CASE SUMMARY

Customer churn (or customer attrition) is one of the significant issues for many businesses. The cost of customer churn is estimated at $168 billion annually in the U.S. (Amori, 2023). Telecommunication companies are no exception. Predicting customer churn accurately is critical to preparing to take proper actions to maintain the businesses. Hence, many telecommunication companies spend considerable time and effort to build a model that predicts customer churn by analyzing the customers' demographic and behavioral attributes.

Various tools can be used to analyze large datasets, including analytical programming languages, such as Python and R, and analytical software packages, such as SAS Enterprise Mining and IBM SPSS Modeler. R is one of the most popular analytical languages used in data science, but it requires extensive time and effort to learn. As generative Artificial Intelligence (AI) services have become publicly available, an alternative way to use R without extensively learning the analytical programming language has emerged. This teaching case will guide how to use the R programming language to predict the telecommunication company's customer churn using an AI-powered tool, ChatGPT.

The primary goal of this case is to equip students with the skills and techniques necessary to perform logistic regression data analysis in R. These skills include the ability to understand the analysis goal, explore the structure of data, process it effectively, build logistic regression models to identify relationships between the input and the target variables, evaluate the models to determine their effectiveness and build business strategies based on the analysis outcomes. The assignment involves using R to manipulate, visualize, and perform statistical analyses on the dataset, ultimately gaining a deep understanding of logistic regression analysis. Through this project, students will also develop critical thinking and problem-solving skills, which are vital in the field of Data Analytics. Upon completion of the assignment, students will be able to interpret model results and draw meaningful conclusions from them, giving them the necessary tools and knowledge to succeed in Data Analytics. In addition to developing analytics skills, this case guides students through the data analysis process by breaking complex tasks into smaller components, each addressed through individual prompts. The inquiry text used to instruct AI to perform a specific task is called a *prompt*, and the ability to create clear and effective prompts is essential for improving the efficiency of the process.

2. DATA MINING FRAMEWORK

The data analysis process of this case is organized based on the widely used Cross-Industry Standard Process Data Mining (CRISP-DM) cycle framework, which consists of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Chapman et al., 2000). CRISP-DM is widely used due to its advantages in solving business problems in many industries (Mariscal et al., 2010). This exercise includes tasks that aim to deepen the understanding of each data mining phase.

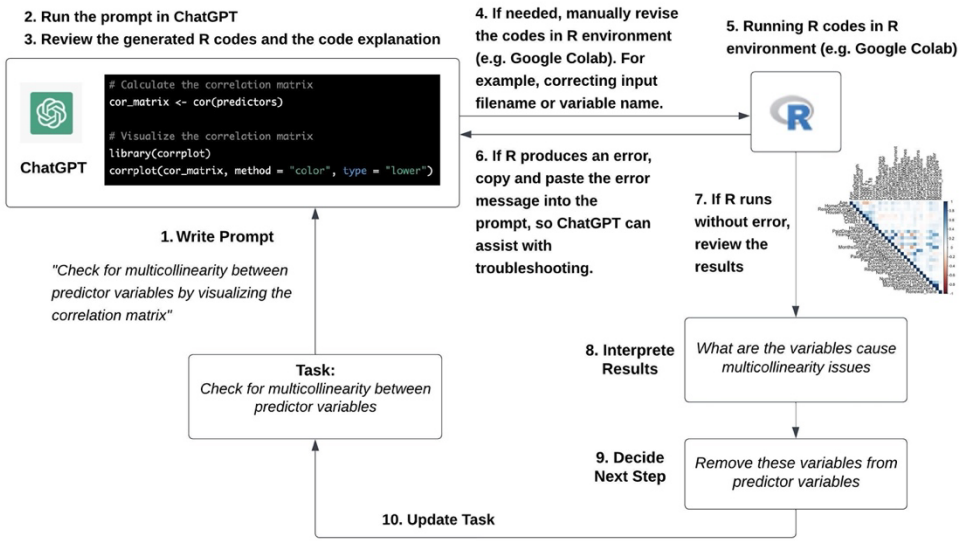


Figure 1. The Workflow of the ChatGPT-Aided R Data Analytics Method

2.1 Business Understanding

In the initial phase, students will review the business problem and a dataset containing a telecommunication company's customer usage and churn status data. They will identify the business purpose of understanding why customers cancel subscriptions and establish evaluation criteria to assess whether the model can accurately predict subscription cancelations.

2.2 Data Understanding and Preparation

During the data understanding phase, students will explore the dataset to gain insights into its quality and relevance, identify any missing or erroneous data, and select the appropriate data for analysis. The CrowdAnalytix telecom churn dataset is adopted to train the students. The dataset was used by the CrowdAnalytix community as part of their prediction competition (https://www.crowdanalytix.com/contests/why-customer-churn). The same dataset is also available on Kaggle.com, which is a large open-source Data Analytics community (https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset). The dataset contains a telecommunication company's customer usage and their churn status data, which indicates whether a customer canceled their subscription.

In the data preparation phase, students would normally clean and prepare the data to ensure its quality and proper format for analysis. However, the CrowdAnalytix telecom churn dataset has already been cleaned and prepared for analysis. The dataset comprises 3,333 observations and 20 variables, consisting of both numerical and categorical factors. For instance, "International.plan" is a character variable that indicates whether the customer has an international calling plan or not, and "Number.vmail.messages" is an integer variable that denotes the number of voicemail messages the customer received. The predictive variable in this dataset is "Churn," which is a character variable that indicates whether the customer canceled their subscription or not. The full description of the variables is provided in the Appendix.

2.3 Modeling, Evaluation, and Deployment

In the modeling phase, students will use data mining models for classification analysis of the data. The evaluation phase assesses the model's performance based on criteria such as classification accuracy, false positives, or false negatives. Finally, in the deployment phase, the results are interpreted, and the analysis results are implemented into the business context.

Supervised classification is a common data mining technique that utilizes labeled data to train algorithms for predicting outcomes on new, unlabeled data. The dataset provided for this assignment is labeled, indicating customer churn, in other words, whether a customer cancels a subscription or not. The objective of data analysis is to predict whether customers would cancel a subscription based on their demographic and behavioral attributes. This task is a binary classification problem, with the aim of finding the best fit that separates the two classes of data. Logistic regression is a commonly used binary classification model that employs a logistic function to model the relationship between the dependent and independent variables. The model estimates the coefficients of the independent variables, which determine the strength and direction of the relationship between the independent variables and the dependent variable. The logistic regression model assumes linearity of independent variables, i.e., the absence of multicollinearity between them. Outliers or influential observations can have a significant impact on the model's estimates and should be identified and addressed.

3. CASE PROCEDURE

3.1 AI-Aided R Coding

This case presents a novel approach for performing Data Analytics using R scripts generated by ChatGPT, which enables students without prior experience in R to utilize the scripts for data mining. The workflow of this ChatGPT-aided R Data Analytics method is illustrated in Figure 1. To begin, the student generates a command prompt in natural language (e.g., English) for a specific data analysis task, and inputs it into

ChatGPT. Based on the provided prompt, ChatGPT generates R scripts and accompanies them with explanations for each section. Students subsequently review these scripts, making modifications as required. Once refined, they run the adjusted code in the R environment to obtain the results. If R produces an error, students can copy and paste the error message into the prompt so ChatGPT can assist with troubleshooting. Once results are obtained, the student can focus on interpreting them, determining the next analysis task, and generating another prompt to obtain the R scripts needed to perform subsequent tasks.

3.2 Environment Setup

The students are instructed to use Google Colab, a cloud-based programming service that enables them to access the R environment directly through a browser (https://colab.research.google.com/#create=true&language=r). Additionally, students need a ChatGPT account to access ChatGPT service (https://chat.openai.com/) through their browser. Students are provided with instructions on how to run prompts in ChatGPT. Moreover, students should be advised to thoroughly examine the results from ChatGPT, encompassing the generated R scripts and their accompanying explanations. Based on the specific task, they might need to make minor script adjustments.

3.3 Phase 1: Business Understanding

As mentioned previously, the dataset used for the exercise is from an anonymized telecommunication company where customer' churn is the primary concern of business. Customer churn, defined as the rate at which customers end their business relationship with a company, has significant financial implications. A high churn rate not only inflates costs but also signals potential loss of revenue growth and brand promotion opportunities. Customer acquisition often incurs higher costs than customer retention due to factors such as marketing expenditures, resource allocation, and time required for onboarding a new client. The students are provided with a brief introduction to the company's business situation (as stated in Section 1) and the goal of the data analysis in this phase.

Question 1.1: What is customer churn? Why should we concern and analyze customer churn?

3.4 Phase 2: Data Understanding and Preparation

As described in the previous section, data is composed of 20 variables. In this step, students perform data exploration and processing by following the instructions and answering the questions.

- **Step 1:** Load the data into R: Use the following ChatGPT prompt to get the R script: ***Load dataset named "churn-bigml.csv" in csv format to R. Call the dataset "churn."***
- **Step 2:** After loading the data, check if the data has been successfully loaded and describe the dataset, using the following ChatGPT prompt: ***After reading the "churn-bigml.csv" file, describe the table and variables in R.***

Question 2.1: How many variables are included in the dataset? Are there any categorical data in the dataset?

- **Step 3:** There are categorical variables in the above step. We first convert the categorical variable "Churn" to a numeric variable using the prompt: ***Convert the Churn variable to 0 ("FALSE") or 1 ("TRUE").***
- **Step 4:** Check for missing or infinite values in the dataset, using the ChatGPT prompt: ***Check if there is any missing value in each column of the dataset.***

Question 2.2: Are there any missing values?

- **Step 5:** Explore distributions of the numeric columns, using the prompts***: How to visualize and get a histogram of the numeric data in R? Here are the numeric column names: "Account.length", "Number.vmail.messages", "Total.day.minutes", "Total.day.calls", "Total.day.charge", "Total.eve.minutes", "Total.eve.calls", "Total.eve.charge", "Total.night.minutes", "Total.night.calls", "Total.night.charge", "Total.intl.minutes", "Total.intl.calls", "Total.intl.charge", "Customer.service.calls".***
- **Step 6:** Identify potential relationships between numeric variables, using the prompts: ***How do you get boxplots for the columns that are numeric in the data (display the column name vertically)? Visualize the correlation matrix among columns in a heatmap.***

Question 2.3: Are there highly correlated pairs of variables? Explain the criteria you used to determine the highly correlated variables.

- **Step 7:** Explore non-numeric variables, using the prompts: ***How to visualize and get an overview of data of non-numeric columns in R? Here are the non-numeric column names: "State", "International.plan", "Voice.mail.plan", "Churn".***

Question 2.4: What is the proportion of customers who have churned in the dataset?

3.5 Phase 3: Modeling

Perform modeling and evaluation by following the instructions.

Question 3.1: Which variable is the dependent variable? What is/are the independent variables?

- **Step 1:** We need to filter out the attributes that we don't need from the analysis. The variables "State" and "Area.code" are identified as not relevant to the customer's churn so they need to be removed. Use the following prompt: ***Remove the "State" variable and "Area.code" variable.***
- **Step 2:** Check all columns in the dataset using the prompt: ***Print all the columns of the dataset.***

Question 3.2: How many predictor variables in the model?

- **Step 3:** Split the Dataset into Training and Test Dataset: To fit a regression model and test its performance, the dataset needs to be split up into training and testing datasets. Since the original dataset is somewhat imbalanced (i.e., customers who have churned are very small compared to those who have not), we need to balance (or weight) the training dataset: Use the following prompt: ***The Churn variable has been converted to numeric variable (0 or 1). Oversample the dataset and increase the number of data samples with Churn=1 by a factor of 5, and then split the above dataset into 60% training data and 40% testing data.***

Question 3.3: What is the proportion of customers who have churned vs. not-churned in the training dataset?

- **Step 4:** Use the training dataset to fit a logistic regression model, using the prompt: ***Fit a logistic regression model in R using the training data and testing data.***

- **Step 5:** Because the logistic regression model is susceptible to multicollinearity among the independent variables (i.e., one independent variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy), we need to assess multicollinearity. One way is to assess the correlation matrix to identify predictor variables with high VIF (Variance Inflation Factor) values (i.e., greater than 10) and remove them from the model. To find the variables with multicollinearity issues, Use the following prompt: ***How to calculate VIF values in the above logistic regression model?***

Question 3.4: Are there variables causing multicollinearity issues? (More specifically, are there any variables that show high multicollinearity (VIF > 10)?) If so, what are the variables?

- **Step 6:** Considering the VIF values identified above and the relationships among the variables, use the prompt to remove these variables: ***Remove the "Total.day.charge", "Total.eve.charge", "Total.night.charge", "Total.intl.charge", columns from the dataset.***
- **Step 7:** Fit the logistic regression model with the variables removed, using the prompt: ***Fit a logistic regression model with train.***
- **Step 8:** Check the logistic model assumptions, such as no multicollinearity and linearity, using the prompt: ***Check the logistic model assumptions by creating diagnostic plots.***

Question 3.5: Are the model assumptions validated?

- **Step 9:** Examine the model coefficients and their associated standard errors, p-values, and confidence intervals, using the prompt***: Examine the model coefficients and their associated standard errors, p-values, and confidence intervals.***

Question 3.6: Report the model coefficients, p-values, and confidence intervals of each predictor variable.

3.6 Phase 4 and 5: Evaluation and Interpretation

In this phase, students are asked to evaluate the model fit using metrics like the model p-value, pseudo R-squared value, and classification accuracy.

- **Step 1:** Get p-value of the logistic regression model, using the prompt: ***How to calculate the significance of the above logistic regression model in R?***

Question 4.1: What is the logistic regression model's p-value? Interpret this p-value.

- **Step 2:** Get pseudo R-squared value of the logistic regression model, using the prompt: ***Print out the pseudo R^2 (Cox and Snell's R^2) when using logistic regression in R.***

Question 4.2: What are the pseudo R-squared (Cox & Snell, Kegelkerke) values of the logistic regression model? Interpret these pseudo R-square values.

- **Step 3:** Get a confusion matrix of the logistic regression for the training and the testing data, using the prompt: ***Get the confusion matrix of the logistic regression model in R.***

Question 4.3: Report the confusion matrix of the logistic regression model.

- **Step 4:** Calculate sensitivity and specificity of the logistic regression classification outcome based on the confusion matrix, using the prompt: ***Calculate the accuracy, sensitivity, and specificity of the logistic regression classification outcome based on the confusion matrix.***

Question 4.4: What is the overall accuracy, sensitivity, and specificity of the logistic regression model for training and testing dataset?

- **Step 5:** Git a Logistic Regression Model and compare it with the Linear Regression Model, using the prompts: ***Get a ROC plot for the training and testing set.***

Question 4.5: Is the logistic regression model a good fit? Interpret the ROC plot.

- **Step 6:** Calculate the propensity score of each customer, using the prompt: ***Calculate the propensity scores for each record in the testing data and show the highest 10 values.***

Question 4.6: Which customer has the highest propensity score and what is the value (i.e., who is most probable to churn)?

3.7 Phase 6: Deploying

In this phase, students are asked to build a business strategy based on the findings from the logistic regression analysis. Based on the logistic regression model outcome, students can formulate the equation to estimate the probability that a customer would churn. Furthermore, students can discuss the ways of utilizing the logistic regression model for various business areas, such as marketing, customer relationships, customer lifetime value calculation, and so on. Group/class discussions can be done on the ways of utilizing the logistic regression model for various business areas, such as marketing, customer relationship, customer lifetime value calculation, and so on.

- **Discussion Point 1:** How does a telecommunication company utilize the logistic regression analysis result?
- **Discussion Point 2:** What are the benefits of using the logistic regression analysis outcome for decision-making compared to traditional decision-making based on human insights?

4. CONCLUSION

In conclusion, this teaching case provides students with an extensive overview of data analysis using the CRISP-DM cycle framework, exposing them to its six essential phases (business understanding, data understanding, data preparation, logistic regression modeling, evaluation, and deployment). The case emphasizes the practical application of logistic regression modeling within this framework and enables students to evaluate and interpret the results effectively. Significantly, it bridges the gap for those lacking expertise in R programming through the utilization of ChatGPT's capabilities. This AI-driven approach allows students to master R for data analysis and visualization, preparing them to address complex business problems involving large datasets. As students' progress beyond this case, the application and refinement of these learned skills will empower them to unlock the capabilities of Data Analytics when tackling a broad range of data analysis challenges.

5. REFERENCES

Amori, M. (2023). How To Address Customer Churn With AI-Driven Data Analysis. *Forbes Technology Council*, https://www.forbes.com/sites/forbestechcouncil/2023/01/30/how-to-address-customer-churn-with-ai-driven-data-analysis/

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0 Step-by-Step Data Mining Guide. Technical Report, CRISP-DM. https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72

Mariscal, G., Marbán, Ó, & Fernández, C. (2010). A Survey of Data Mining and Knowledge Discovery Process Models and Methodologies. *The Knowledge Engineering Review*, 25(2), 137-166. https://doi.org/10.1017/S0269888910000032

AUTHOR BIOGRAPHIES

Chen Zhong is an assistant professor of cybersecurity in the John H. Sykes College of Business at the University of Tampa. She received her doctoral degree in Information Sciences and Technology from Pennsylvania State University. Her current research interest includes cybersecurity analytics, intrusion detection, Artificial Intelligence, and cybersecurity education. Dr. Zhong has published in the refereed journals and conference proceedings, such as *Computers & Security*, *IEEE Systems Journal*, *IEEE Conference on Cognitive and Computational Aspects of Situation Management* (CogSIMA).

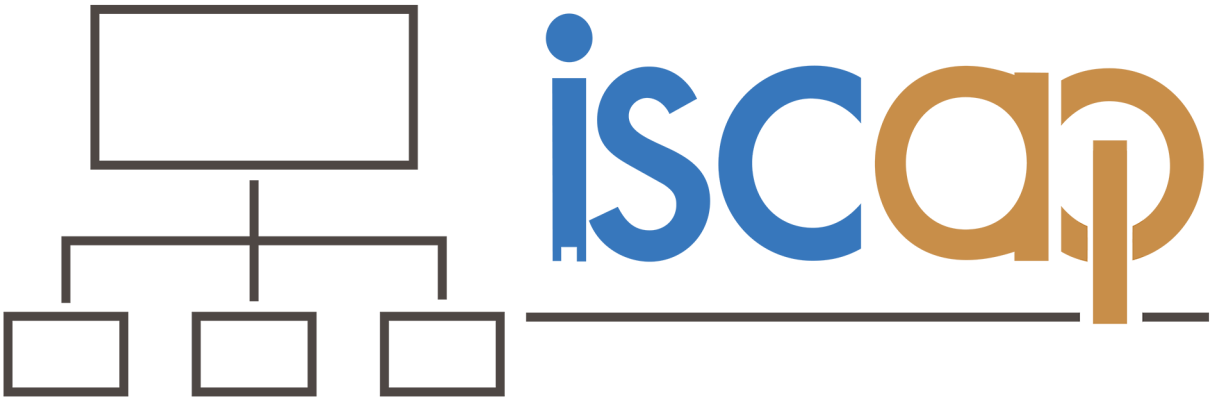
J.B. (Joo Baek) Kim is an assistant professor in the John H. Sykes College of Business at the University of Tampa in Tampa, Florida. Before he joined UT, he worked at Worcester Polytechnic Institute in Worcester, Massachusetts. He received his Ph.D. from Louisiana State University. His work/interest encompasses serious games in business contexts, business simulation games, gamification, and online user reviews and interactions. His research work has been published in peer-reviewed journals such as *Information Technology and People* and *Asia Pacific Journal of Information Systems*, and in various conference proceedings.

APPENDIX

Dataset Description

Variable	Description
State	A character variable indicating the state in which the customer resides.
Account.length	An integer variable indicating the number of days the customer has been with the company.
Area.code	An integer variable indicating the area code of the customer's phone number.
International.plan	A character variable indicating whether the customer has an international calling plan or not.
Voice.mail.plan	A character variable indicating whether the customer has a voicemail plan or not.
Number.vmail.messages	An integer variable indicating the number of voicemail messages the customer has received.
Total.day.minutes	A numeric variable indicating the total number of minutes the customer used during the day.
Total.day.calls	An integer variable indicating the total number of calls the customer made during the day.
Total.day.charge	A numeric variable indicating the total charge for the day's usage.
Total.eve.minutes	A numeric variable indicating the total number of minutes the customer used during the evening.
Total.eve.calls	An integer variable indicating the total number of calls the customer made during the evening.
Total.eve.charge	A numeric variable indicating the total charge for the evening's usage.
Total.night.minutes	A numeric variable indicating the total number of minutes the customer used during the night.
Total.night.calls	An integer variable indicating the total number of calls the customer made during the night.
Total.night.charge	A numeric variable indicating the total charge for the night's usage.
Total.intl.minutes	A numeric variable indicating the total number of international minutes the customer used.
Total.intl.calls	An integer variable indicating the total number of international calls the customer made.
Total.intl.charge	A numeric variable indicating the total charge for the international usage.
Customer.service.calls	An integer variable indicating the number of calls made by the customer to the customer service department.
Churn	A character variable indicating whether the customer cancelled the subscription or not.

INFORMATION SYSTEMS & COMPUTING ACADEMIC PROFESSIONALS



STATEMENT OF PEER REVIEW INTEGRITY

All papers published in the ***Journal of Information Systems Education*** have undergone rigorous peer review. This includes an initial editor screening and double-blind refereeing by three or more expert referees.

ISSN: 2574-3872 (Online) 1055-3096 (Print)