

Data Cleaning

What customer factors might influence churn rate?

Vera Lake | D206 | 04/02/2021

# Part I: Research Question

## A.  Describe **one** question or decision that you will address using the data set you chose. The summarized question or decision must be relevant to a realistic organizational need or situation.

The description states a question or decision that can be addressed through analysis of the chosen data set. The question or decision is relevant to a realistic organizational need or situation.

The question chosen is: What customer factors might influence churn rate? This is a common question in businesses interested in customer retention and marketing.

## B.  Describe the variables in the data set and indicate the specific type of data being described. Use examples from the data set that support your claims.

The description includes the variables in the data set and indicates the specific type of data being described and includes examples from the data set to support claims.

Range is from the min and max in the describe table by Pandas.

Since the source of the data is unknown for this project some of the fields’ description are unknown and will not be used in the data discovery or data cleaning phases. Irrelevant will be determined later.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column Id(s) | Column Name(s) | Data Type | Description | Range or Example |
| 0 | Unnamed | Int64 | This serves as the table index | 1-10000 |
| 1 | CaseOrder | Int64 | unknown | 1-10000 |
| 2 | Customer\_Id | Object | Used to identify the customer | K409198 |
| 3 | Interaction | Object | unknown | aa90260b-4141-4a24-8e36-b04ce1f4f77b |
| 4 | City | Object | This is the city where the customer lives | Point Baker |
| 5 | State | Object | This is the state where the customer lives | AK |
| 6 | County | Object | The county where the customer lives | Prince of Wales-Hyder |
| 7 | Zip | Int64 | The zip code of the customer | 601-99929 |
| 8 | Lat | Float64 | The latitude of where the customer lives | 17.966120-  70.640660 |
| 9 | Lng | Float64 | The longitude of where the customer lives | -171.688150 –  -65.667850 |
| 10 | Population | Int64 | The population of the county where the customer lives | 0.000000 - 111850.000000 |
| 11 | Area | Object | What type of area the customer lives in | Urban |
| 12 | Timezone | Object | The timezone of the customer | America/Sitka |
| 13 | Job | Object | The type of job the customer holds | Environmental health practitioner |
| 14 | Children | Float64 | How many children the customer has | NaN,  0.000000 – 10.000000 |
| 15 | Age | Float64 | The age of the customer | 18-89 |
| 16 | Education | Object | Education level of the customer | Master's Degree |
| 17 | Employment | Object | The type of employment the customer has | Part Time |
| 18 | Income | Float64 | The annual income of the customer | 740.66 – 258900.700000 |
| 19 | Marital | Object | The marital status of the customer | Widowed |
| 20 | Gender | Object | The gender of the customer | Male |
| 21 | Churn | Object | Boolean for if the customer has churned | Yes/No |
| 22 | Outgage\_sec\_perweek | Float64 | How many seconds per week on average the customer experiences outages | -1.348571 -  47.049280 |
| 23 | Email | Int64 | Customer’s email | 1-23 |
| 24 | Contacts | Int64 | Unknown | 0-7 |
| 25 | Yearly\_equip\_failure | Int64 | How many times per year there is equipment failure | 0-6 |
| 26 | Techie | Object | Boolean of whether the customer is a techie | Yes/No |
| 27 | Contract | Object | The type of contract the customer is under | One Year |
| 28 | Port\_modem | Object | Boolean of whether the customer has a port modem | Yes/No |
| 29 | Tablet | Object | Unknown | Yes/No |
| 30 | InternetService | Object | Type of internet service the customer has | Fiber Optic |
| 31 | Phone | Object | Customer Phone Number | Yes/No |
| 32 | Multiple | Object | Unknown | Yes/No |
| 33 | OnlineSecurity | Object | Customer opted in to online security | Yes/No |
| 34 | OnlineBackup | Object | Customer opted int to backing up data online | Yes/No |
| 35 | DeviceProtection | Object | Customer opted in to device protection | Yes/No |
| 36 | TechSupport | Object | Customer opted in to technical support | Yes/No |
| 37 | StreamingTV | Object | Boolean for whether the customer streams TV | Yes/No |
| 38 | StreamingMovies | Object | Boolean for whether the customer streams movies | Yes/No |
| 39 | PaperlessBilling | Object | Boolean for whether the customer prefers paperless billing | Yes/No |
| 40 | PaymentMethod | Object | How the custom r pays their bill | Credit Card (automatic) |
| 41 | Tenure | Float64 | The number of months the customer has been subscribed for | 1.000259 -  71.999280 |
| 42 | MonthlyCharge | Float64 | How much the customer pays per month | 77.505230 -  315.878600 |
| 43 | Bandwidth\_GB\_Year | Float64 | The amount of bandwidth used per year | 155.506715 - 7158.982000 |
| 44-51 | item1-item8 | Int64 | unknown | 1-8 |
|  |  |  |  |  |

# Part II: Data-Cleaning Plan

*Note: You may use Python, R, or any other programming language for implementing your coding solutions, manipulating the data, and creating visual representations.*

## C.  Explain the plan for cleaning the data by doing the following:

### 1.  Propose a plan that includes the relevant techniques and specific steps needed to identify anomalies in the data set.

The proposal includes a detailed description of the techniques and steps needed for identifying anomalies in the selected data set.

1. Split the data into train and test data sets
2. Clearly enunciate the project objectives
   1. Learn about current customers. That is, learn the characteristics of those who choose to buy services from the company, as well as those who do not.
   2. Develop a method of identifying those who have churned and why, so that it can be theorized why some customers churn.
3. Translate the objectives into a data science problem
   1. Learn about current customers who do and do not churn
      1. Use exploratory data analysis to express some simple graphic relationships among all the known variables and the Boolean churn variable.
         1. Example: Histogram of churn overlain with age to visualize whether age has a bearing on customer response
      2. Use clustering to determine whether there is natural grouping within the current customers. Then see if the clusters differ with respect to their churn.
         1. Example: younger/more-educated vs older/less-educated.
      3. Use association rules to see whether there are useful relationships among subsets of the records.
         1. Example: “If phone, then churn = no” has a good support and high confidence. This would allow the theorization that a customer has a phone and does not churn then the fact a customer has a phone means they prefer to stay with the company.
   2. Develop a suite of data science models to identify customers who are likely to churn. Since the churn variable is (yes/no) this is considered to be a categorical problem, only classification models can be used.
      1. Develop the best classification model possible, using the following algorithms:
         1. Decision Trees
         2. Random Forests
         3. Naïve Bayes Classification
         4. Neural Networks
         5. Logistic Regression
      2. Evaluate each model based on predetermined model evaluation criteria, such as misclassification costs.
      3. Construct a table of the best models and their costs.
      4. Select and argue the best model.
4. Prepare the data
   1. Check for missing data
      1. Null, NaN values
   2. Change misleading field values
      1. If the graphs seem off, then figure out why and change the values to something more appropriate
   3. Re-express categorical data as numeric data
      1. To provide a categorical field (such as the variable education) as numeric data
      2. Take care to keep the numeric data relative to the weight of the category, i.e. Masters Degree would have a larger numeric value than Bachelors Degree
   4. Standardize the numeric fields
      1. Certain algorithms perform better with numeric fields that are standardized
      2. Positive z-values may be interpreted as representing the number of standard deviations above the mean the data value lies
      3. Negative z-values represent the number of standard deviations below the mean the data value lies
   5. Identify outliers
      1. Use z-values from standardization
      2. Query to find which rows that are outside of a particular condition
         1. Python: data\_train.query(age\_z>3 | age\_z < -3)
            1. Returns a list of outliers

### 2.  Justify your approach for assessing the quality of the data, include:

The justification includes the characteristics of the data being assessed and references the approach used to assess the quality of the data. The justified approach aligns with the selected data set.

•  characteristics of the data being assessed

Example: Outgage\_sec\_perweek has a negative value…how do you have

negative outage…0 would be the minimum

•  the approach used to assess the quality

Some of the data cannot be known what it is without access to someone with more knowledge of the company or the data

### 3.  Justify your selected programming language and any libraries and packages that will support the data-cleaning process.

The justification describes the benefits of using the programming language, including any libraries and packages used to clean the data, and includes specific examples of how these tools are ideal in this scenario as opposed to other available tools.

### 4.  Provide the code you will use to identify the anomalies in the data.  The submission provides the complete and executable code, which could be used to identify anomalies in the data set.

# Part III: Data Cleaning

## D.  Summarize the data-cleaning process by doing the following:

### 1.  Describe the findings, including all anomalies, from the implementation of the data-cleaning plan from part C.

The description accurately includes all of the anomalies found by running the code from part C4.

### 2.  Justify your methods for mitigating each type of discovered anomaly in the data set.

The justification includes the specific mitigation methods for each type of anomaly listed in part D1.

### 3.  Summarize the outcome from the implementation of *each* data-cleaning step.

### The summary details the outcome from the implementation of each data-cleaning step. The summarized expected outcomes are plausible given the interventions.

### 4.  Provide the code used to mitigate anomalies.

### The submission provides complete and executable code that could be used to mitigate the anomalies.

### 5.  Provide a copy of the cleaned data set.

### The submission includes a clean data set created from the raw data. The provided data set includes the complete list of variables from the chosen data set in part A.

### 6.  Summarize the limitations of the data-cleaning process.

### The submission accurately summarizes the limitations of the implemented data-cleaning process.

### 7.  Discuss how the limitations in part D6 affect the analysis of the question or decision from part A.  The submission includes a discussion of the impact of the limitations from part D6. The discussion logically aligns with the question or decision from part A.

## E.  Apply principal component analysis (PCA) to identify the significant features of the data set by doing the following:

### 1.  List the principal components in the data set.

The submission lists all principal components of the data set.

### 2.  Describe how you identified the principal components of the data set.

The description of how the principle components of the data set were identified is accurate and complete.

### 3.  Describe how the organization can benefit from the results of the PCA  The description of how the organization can benefit from the results of the PCA is logical and accurate.

# Part IV. Supporting Documents

## F.  Provide a Panopto recording that demonstrates the warning- and error-free functionality of the code used to support the discovery of anomalies and the data cleaning process and summarizes the programming environment.  The Panopto video recording demonstrates the warning-and error-free functionality of the code used to support the discovery of anomalies and the data cleaning process. An accurate summary of the programming environment is provided in the video.

*Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access", and then choose to log in using the “WGU” option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto’s website.*

*To submit your recording, upload it to the Panopto drop box titled “Data Cleaning – NUM2 \ D206” Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.*

## G.  Reference the web sources used to acquire segments of third-party code to support the application. Be sure the web sources are reliable.

## H.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

## I.  Demonstrate professional communication in the content and presentation of your submission.

# File Restrictions

File name may contain only letters, numbers, spaces, and these symbols: ! - \_ . \* ' ( )  
File size limit: 200 MB  
File types allowed: doc, docx, rtf, xls, xlsx, ppt, pptx, odt, pdf, txt, qt, mov, mpg, avi, mp3, wav, mp4, wma, flv, asf, mpeg, wmv, m4v, svg, tif, tiff, jpeg, jpg, gif, png, zip, rar, tar, 7z