**Churn data cleaning**

**Nhi Le**

**D206 Data Cleaning**

**Western Governor University**

**January 9, 2023**

**Part I: Research Question**

1. **Question or Decision**

Through the churn data, one question we need to answer is: What services had the highest churn rates?

1. **Required Variables**

‘CaseOrder’ runs from 1 to 10,000 presenting each line of the data continuously. For example: 1, 2, 3, 4, 5,…

‘Customer\_id’ is a unique ID for each customer. For example: ‘K409198’, ‘S120509’.

‘Interaction’ is also unique IDs generated for customer transactions, technical support, and sign-ups. Example: ‘aa90260b-4141-4a24-8e36-b04ce1f4f77b’.

For each customer, there are demographic variables on billing statement including: ‘City’, ‘State’, ‘County’, ‘Zip’, ‘Lat’, ‘Lng’, ‘Population’, ‘Area’, ‘TimeZone’, ‘Job’, ‘Children’, ‘Age’, ‘Education’, ‘Employment’, ‘Income’, ‘Marital’, ‘Gender’. ‘Children’, ‘Age’, and ‘Income’ are quantitative variables. The others in this variable group are qualitative variables. Example: ‘City’ – ‘Del Mar’, ‘State’ – ‘CA’, ‘County’ – ‘San Diego’, ‘Zip’ – ‘92014’, ‘Lat’ – ‘32.96687’, ‘Lng’ – ‘-117.24798’, ‘Population’ – ‘133863’, ‘Area’ – ‘Suburban’, ‘TimeZone’ – ‘America/Los\_Angeles’, ‘Job’ – ‘Solicitor’, ‘Children’ – ‘1’, ‘Age’ – ‘48’, ‘Education’ – ‘Doctorate Degree’, ‘Employment’ – ‘Retired’, ‘Income’ – ‘18925.23’, ‘Marital’ – ‘Married’, ‘Gender’ – ‘Male’.

‘Churn’ is categorical variable as the values are ‘Yes’ or ‘No’. It recorded whether the customers canceled service last month.

‘Outage\_sec\_perweek’ is numeral variable to show system outages in the customer’s neighborhood’s average of seconds per week. Example: ’15.20619’.

‘Email’ is numeral variable to record the number of marketing or correspondence emails sent. Example: 1, 2, 3, 4, 5…,23.

‘Contacts’ is numeral variable for how many times customer contacted technical support. Example: 0, 1…,7.

‘Yearly\_equip\_failure’ is numeral variable to show the number of time customer’s equipment failed and needed to reset or replaced last year. Example: 0, 1, 2, 3, 4, 6.

‘Techie’ has Yes/No value. This categorical variable reflects if the customer thinks that they are good at technology.

‘Contract’ is categorical variable on what kind of contract customer has ‘Month-to-month’, ‘One Year’, or ‘Two Year’.

‘Tablet’ is categorical variable answering if the customer has a table. The values are ‘Yes’ or ‘No’.

‘Port\_modem’, ‘InternetService’, ‘Phone’, ‘Multiple’, ‘OnlineSecurity’, ‘OnlineBackup’, ‘DeviceProtection’, ‘TechSupport’, ‘StreamingTV’, ‘StreamingMovies’ are services that the company provides. The values of these variables are ‘Yes’ or ‘No’ to reflect if the customer signed up for.

‘PaymentMethod’ shows how the customer paid: ‘Electronic Check’, ‘Mailed Check’, ‘Bank Transfer(automatic)’ or ‘Credit Card(automatic)’, which is categorical variable.

‘Tenure’ is numerical variable to record how many months the customer has been with the provider. For example: ’17.08723’.

‘MonthlyCharge’ is the monthly charge for the customer, example: ‘120.2495’. ‘Bandwidth\_GB\_year’ is the average amount of GB used in a year for each customer, example: ‘2164.579’. These are numerical variables.

Lastly, 8 categorical variables reflect customer’s satisfaction ratings on a scale of 1 to 8 (1 = most important, 8 = least important): ‘item1’ – Timely response, ‘item2’ – Timely fixes, ‘item3’ – Timely replacements, ‘item4’ – Reliability, ‘item5’ – Options, ‘item6’ – Respectful response, ‘item7’ – Courteous exchange, ‘Item8’ – Evidence of active listening. The values are 1, 2…,8.

**Part II: Data-Cleaning Plan**

**C1. Plan to find anomalies**

Here is my approach:

* Import the data set to Python using Pandas’ read\_csv.
* Using .info() to get the information about the data set including variables, count of non-null values for each variable, and variable’s type.
* Examine and delete duplicate using .duplicated()
* Use .isnull().sum() to get the total number of null values for each variable.
* Use ordinal encoding to transform categorical variables to numeral values.
* Use central tendency (mean, median, and mode) to imputing missing data
* Detect outliers for quantitative variables using histogram or boxplot.

**C2. Justification of approach**

The data set has a lot of NA values, which includes both categorical and numerical values. By using imputation for missing data, it will be more effective and efficient than trying to collect those missing data. It saves not only time, but also labor. Therefore, we can finish this project fast. Even imputation is a guessing method that replaces missing data with estimated value, we save the data from getting thousands of data lines deleted. It helps to possibly keep the same shape of the distribution and avoid standard deviation dramatically altered (Kang, 2013, pp. 402-406). I will use the course textbook and reliable sources to complete this project.

**C3. Justification of tools**

I will be using Python since I know a little about this programming language. Python is easy to set up and has many different packages and extensive libraries, which makes it easy to use. Python packages are pre-coded that help us to do complex tasks without writing lengthy lines of codes by ourselves (Larose, 2019, p. 11). For example, to create a boxplot to detect outliners, I will just need to import the Seaborn library, then write 1 line of code. As I know Python is widely used in Data Science, I can find tips and help in online open sources.

Here are the libraries and packages I will use:

* NumPy is used for working with arrays.
* Pandas is used for working with a data set. Example, we can use .read.csv() to load the data set to Python. Or I can use .info() to get the information of the data set.
* Matplotlib is a comprehensive library for creating visualizations. I will use matplotlib.pyplot submodule for creating histograms. With the bar charts, they will help me detect outliners, get the shape of data distributions, and fix missing data accordingly.
* Seaborn is also used to create boxplots.
* Scikit-learn is used for machine learning, especially for Principal Components Analysis.

**C4. Provide the code**

I attached the D206\_NL.ipynb file along with the submission for all the code. I also include the screenshots of all the executed codes and the results.

A screenshot of a computer

Description automatically generated

Table

Description automatically generated

Graphical user interface, application, table

Description automatically generated with medium confidence



A picture containing shape

Description automatically generatedGraphical user interface

Description automatically generated

Diagram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart

Description automatically generated

**Part III: Data Cleaning**

**D1. Cleaning findings**

* I did not find any duplicated value since the results showing after running .duplicate() is ’False 10,000’
* There is a big amount of missing data: ‘Children’ variable had 2495 null values. ‘Age’ variable missed 2475 values. ‘Income’ variable missed 2490 values. ‘Techie’ variable had 2477 null values. ‘Phone’ variable had 1026 null values. ‘TechSupport’ had 991 null values. ‘Tenure’ missed 991 values. ‘Bandwidth\_GB\_Year’ missed 1021 values.
* Firstly, I found an outlier for ‘Yearly\_equip\_failure’ variable. There was only 1 customer that had their equipment fail and reset 6 times last year. This value falls outside the two inner fences on the boxplot. Secondly, there are a few outliers for ‘Email’ variables, which have values less than 4 or more than 20. Only 1 customer had 23 emails sent. Thirdly, ‘MonthlyCharge’ and ‘Contacts’ boxplots show that some values fall outside the two inner fences. However, I think these are the results of natural variation, not by human mistakes or system errors. Lastly, the ‘Outage\_sec\_perweek’ variable has some outliers since it contains a few negative values. Besides that, there are also a lot of values that are outside of the two inner fences on the box plot.

**D2. Justification of mitigation methods**

* For the missing data, I will use univariate imputation (Mean, Median, Mode) method. I use this method since the data is missed completely at random. This is the most common method, which is easy to implement, and takes little time to fill in all the missing data (Madaan, 2022). For the numeral variable ‘Children’, ‘Income’, ‘Tenure’ and ‘Bandwidth\_GB\_Year’, I will replace missing data with the median value since they have skewed or bi-modal distribution shapes. For numeral variable ‘Age’, I will use the mean value to fill in the missing data since it has normal distribution shape.
* ‘Techie’, ‘Phone’, ‘TechSupport’ are categorical variables that have Yes or No value. I will use Ordinal Encoding method to transform categorical values to numerical values: 0 represents ‘No’ and 1 represents ‘Yes’. Then I will find the mode value of each variable and use it to fill in the missing data.
* For the ‘Outage\_sec\_perweek’ variable, I will need to replace negative values with positive values. Seconds or time is not supposed to be negative. On the other hand, I will not remove any outlier from the data set. Since we do not know for sure if the data is completely precise or who collected this dataset. It is a safe action to retain the outliers. Removing extreme values due to their extremeness can distort the results of the research question (Frost, 2021). In this case, I think instead of excluding the outliers, we should analyze them to see if they have any impact to the research question

**D3. Summary of the outcomes**

* After I filled in all the missing data with mean and median values, I could notice the changes of distribution shapes for those numeral variables, which was expected. All the missing data was replaced. To verify, I used df.isnull().sum(). All the results are 0.
* For outliers, the negative values of ‘Outage\_sec\_perweek’ were replaced with positive values. Since there were only 11 negative values, the change did not impact the distribution shape significantly.

Before imputation:

Chart

Description automatically generated

After imputation:

Chart

Description automatically generated

This is the verification step after replacing all the null values:

A picture containing graphical user interface

Description automatically generated

Shape

Description automatically generated

**D4. Mitigation code**

I attached screenshots of the code and the results. I also include them in the D206\_NL.ipynb file.

* Treating missing numeral data:

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

* Treating missing categorical data:

Text

Description automatically generated

* Treating outliers for ‘Outage\_sec\_perweek’

Chart, histogram

Description automatically generated

**D5. Clean data**

The copy of cleaned data set was attached with this submission as ‘df\_clean.csv’.



**D6. Limitations**

The biggest limitation of imputing missing data with mean/ median/ mode value is that it could possibly distort data or distribution of data (Mohaan, 2022). I do not see any limitation of replacing negative values of ‘Outage\_sec\_perweek’ with positive values. Since natural variation causes outliers, we will not remove the outliers. This decision will not help in creating a better fitting model or statistically significant results (Frost, 2021).

**D7. Impact of the limitations**

The limitations are not major and will impact much on answering the research question. However, we will still need to be careful. For example, the mean value for ‘Age’ is 50 and we replaced 7525 missing values as 50. When analyzing if the Age factor affects churn, we need to keep in mind that. Another example is about ‘Phone’ variable that missed 8974 values. We filled in the missing data with the mode – ‘Yes’. As we do not have the accurate number of customers (it is just a guessed number after the imputation) that have Phone Service, we will likely come up with an inaccurate churn rate. Since we decided to retain the outliers, it would not give us absolutely accurate results for our research question. However, it is still the best method in this case since the data is possibly precise.

**E1. Principal Components**

Here are the variables in the data set that I will use for the analysis: 'Children', 'Age', 'Income', 'Outage\_sec\_perweek', 'Email', 'Contacts', 'Yearly\_equip\_failure', 'Tenure', 'MonthlyCharge', 'Bandwidth\_GB\_Year', 'item1', 'item2', 'item3', 'item4', 'item5', 'item6', 'item7', 'item8'

Graphical user interface, application

Description automatically generated

This is the PCA loadings:

Graphical user interface, application, table

Description automatically generated

Table

Description automatically generated

**E2. Criteria used**

I will identify how many and which principal components to keep using scree plot/ eigenvalues. I will pick those principal components that have eigen values greater than 1. I will pick ‘PC1’, ‘PC2’, ‘PC3’, ‘PC4’, ‘PC5’. ‘PC6’ eigen value is only slightly greater than 1, so I will decide to exclude it since it does not have strong correlation with other components in the data set.

Chart, line chart

Description automatically generated

**E3. Benefits**

By determining principal components and utilizing PCA, we can determine the correlation among variables, then answer the research question with stronger details. For example, 'item1', 'item2', 'item3', 'item6', 'item7', 'item8' have strong relationship to contribute to ‘PC1’. They are ‘Timely Response’, ‘Timely fixes’, ‘Timely Replacements’, ‘Respectful response’, ‘Courteous exchange’, and ‘Evidence of active listening’. By determining how one factor affects others in this group, we can come to the conclusions of how to improve customers’ experience and to reduce churn. When we actively listen to customers, we can give customers respectful response and courteous exchange. By doing that, customers will get fast responses, fixes, and timely replacements.

**Part IV. Supporting Documents**

**F.  Video**

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=692f0319-9a2a-4b59-a0c3-af8800581619>

**G.  Sources for third-party code**

No sources were used.

**H.  Sources**

Kang, H. (2013). The prevention and handling of the missing data. Korean Journal of Anesthesiology, 64(5), 402. https://doi.org/10.4097/kjae.2013.64.5.402

Larose, C. D. & Larose, D. T. (2019). Data Science: Using Python and R. John Wiley & Sons, Inc.

Mohaan, M. (2022, April 30). Handling missing data: Mean, Median, Mode [web log]. Retrieved January 8, 2023, from https://www.naukri.com/learning/articles/handling-missing-data-mean-median-mode/.

Frost, J. (2021, April 5). Guidelines for removing and handling outliers in data. Statistics By Jim. Retrieved January 9, 2023, from https://statisticsbyjim.com/basics/remove-outliers/