Data Cleaning – D206

Tonie Cooper

Student ID: 001124608

[Tcoo158@wgu.edu](mailto:Tcoo158@wgu.edu)

Mentor: Krissy Bryant

Part I: Research Question

1. What customers are at a high risk of churn? What features are most significant to those that churn?
2. The data set contains 10,000 customer records of a telecommunications company. The dependent variable is if the customers have continued or discontinued service with the company. The variable is labeled as “Churn” in the data set. Churn is a qualitative data type, yes or no, but the churn rate calculated will be a float to show what percentage of customers discontinue service with this company.

The independent variables that may have an impact on the Churn rate and will be used for the analysis are:

Item 1 through Item 8. These will be renamed Response, Fixes, Replacement, Reliability, Options, Respectfulness, Courteous and Active\_Listening. They are part of a customer survey where customers list what they deem most important. All eight of these values are quantitative in nature and will be used as an integer to find key items, such as mean, median and variance.

Then, there’s Tenure, MonthlyCharge and Bandwidth. They’re a quantitative data type and a float, as shown in Out[18] of the Churn Data code.

Data types not used but identified in the coding are objects. These will contain all other data types in columns, such as Customer\_ID and City (Python Documentation).

Part II: Data-Cleaning Plan

1C. **Plan for Cleaning**

First and foremost, I’ll make a copy and back-up the data in case any unwanted intrusion happens. This will be done using GitHub’s command line.

Then, enter data into Python using the Pandas import command.

Next, create a name for the dataset being used to answer the research question. The name will be set as “churn\_cd”.

Additionally, I’ll look at the dataset for misspellings, missing data and data that doesn’t fit its category.

Then, I’ll fill in the missing data, dependent on how much is missing, using the mean, median or mode. If a plethora of data is missing, then deleting the row of data may be suitable.

Finally, I’ll remove outliers that are several standard deviations past the mean as this will skew the results (Badr 2019).

2. **Justification of Approach**

The dataset contains several NA inputs, but there’s enough data to create a reasonable mean to fill in those missing fields. As an initial approach to this dataset, I believe this will give a better understanding of it before needing to create a more advanced method of cleaning.

3. **Justification of Programming Language**

I will be using Python to clean the data. I chose this language as I have an interest in exploring this language and diving deeper into machine learning and AI. From my research, this appears to be the more widely used language for the interest I have in data science and deep learning (Yildirim 2021).

Python provides an easy-to-read and intuitive type of syntax. Through research from online and the course material, Jupyter Notebooks appears to be a keyway to run code for Data Scientist’s and I find it easy to navigate and display the results of the code I’ve created (Yildirim 2021).

I’ll need to import several packages within Jupyter which include NumPy for arrays, Pandas to load the CSV file, Matplotlib for chart visualization, SciPy for mathematical equations needed and Seaborne for presentation visualization.

4. **Provide code**

See attachment

Part III

1D. **Results**

Through cleaning the CSV file provided, I found several anomalies and missing data in fields vital to interpreting the churn rate. There was a large amount of missing data in critical fields such as: Children, Age, Income, Bandwidth and Tenure. By using the mean and variance, as shown in the code attachment, I was able to fill the missing data with the median value. Though, I was able to fill these quantitative fields with the median I chose not to fill qualitative fields, such as Techie, as this field didn’t seem relative to the data and is left too open to interpretation. Now, the anomalies present were not significant enough to change the results of the findings. Below I describe the methods used to mitigate these anomalies.

2D. **Justification of Anomaly Mitigation**

First, I found the variable, MonthlyCharge, to contain outliers, as shown in the box plot. I chose to leave those outliers alone. Next, I added the median values to columns containing missing values that could change the outcome if left alone.

3D. **Outcome Summarization**

After cleaning the data set and renaming the Item columns I’m left with: Tenure, MonthlyCharge, Bandwidth, Response, Fixes, Replacement, Reliability, Options, Respectful, Courteous and Active\_Listening. Adding the median value allowed for NA data to now contain a value meaningful to the outcome of my PCA analysis.

Columns that were removed for this analysis did not provide meaningful data that may alter the churn rate. This would include the removal of all qualitative columns and the location of customers. The columns and reasons to remove are:

“Unnamed” column – there’s no title for the column and appear to be a copy of the column, CaseOrder.

CaseOrder – A list of the rows. This value will have no effect on the churn rate.

Customer\_ID – these are specific to each customer’s account and no average can be made of the dataset

Interaction – no average can be made of this object

City, state, zip and county – the location of the customer. With these fields being qualitative, I don’t see a need to assign these fields a number to create a bar graph or heat graph.

Lat and Lng – a more precise location of the customer. I don’t see how that is relevant to the research question, so it will be removed.

Children and Age– too many NA values within column to want to keep

All other columns removed are due to a combination of being a qualitative data type and containing a lot of NA values. Converting qualitative data to quantitative data leaves a lot of room for error. Providing a churn rate based off incorrect data can lead to bad decision-making and an overall bad experience for the client. I want to keep data that are quantitative and provide a reason why they are churning. This will provide a greater output and picture of why customers are terminating service with the company.

4D. **Mitigation Code**

See attachment

5D. **Copy of Cleaned Data**

See attachment

6D. **Limitation Summary**

The limitations of this data set are the sources it derives from. Without a manager or client to ask questions about the missing data I’m left to using variance and averages to fill in the missing data. By using this method, with so much missing data, in can change the results of the research question.

7D. **Limitation Effective**

As stated above, the results of this analysis will be used by managers to present to the client or those that have an interest in the results. When I fill the data using the median value, that does allow for a picture to be formed and an answer to be presented, but those answers aren’t telling the full story. With so much missing data, I’m left to using averages to find the answer. The client should consider making several of the columns a requirement before it can be submitted and stored. This will provide a more accurate churn rate analysis.

1E. **Principal Components**

The principal components that are most important to the dataset are:

Response, Fixes, Replacement and Respect.

2E. **Principal Components Identification**

I identified the principal components by focusing on the customers. The customers can provide great insight on why they’re choosing to leave the service provider. Looking at the survey the customers took, you can find patterns as to why they want to leave or in their eyes what’s most important to them. This will allow the company to shift their focus on specific client needs.

With all that in mind, I focused my analysis on those quantitative columns and used the scree plot to visualize the results as well as the eigenvalues. The eigenvalue starts its bend at 3 components. Next, a rotation and loading were created which would identify what is the most important values of the dataset.

3E. **Organizational Benefit**

Loading the dataset shows that focusing on response, fixes, replacement and respect should be given a higher priority if they want to reduce the churn rate. Focusing on those items will help reduce the churn rate 27%, which in turn will increase the profits of the company.

Part IV

G.

**Third Party Code Sources**

Malik, Usman. “Customer Churn Prediction with Python.” *Learnpython.Com*, 26 Mar. 2019, learnpython.com/blog/python-customer-churn-prediction.

Saeed, Hassan. “How to Rename Pandas DataFrame Column in Python.” *Stackabuse*, 2 Feb. 2021, stackabuse.com/how-to-rename-pandas-dataframe-column-in-python.

Lianne & Justin. “Data Cleaning in Python: The Ultimate Guide.” *Towards Data Science*, 3 Feb. 2020, towardsdatascience.com/data-cleaning-in-python-the-ultimate-guide-2020-c63b88bf0a0d.

H.

**Sources**

Badr, Will. “5 Ways to Detect Outliers/Anomalies That Every Data Scientist Should Know (Python Code).” *Towards Data Science*, 5 Mar. 2019, towardsdatascience.com/5-ways-to-detect-outliers-that-every-data-scientist-should-know-python-code-70a54335a623.

Yildirim, Soner. “The Most Helpful Python Data Cleaning Modules.” *Learnpython.Com*, 22 Apr. 2021, learnpython.com/blog/python-libraries-data-cleaning.

*Built-in Types¶*. Built-in Types - Python 3.9.6 documentation. (n.d.). https://docs.python.org/3/library/stdtypes.html.