D206 Data Cleaning

Jacob Polomsky

School of Technology, College of IT, Western Governors University

Data Acquisition - D206

Dr. Eric Straw

May 3, 2024

1. Which factor have the biggest impact on whether a customer churns or not?
2. Data Dictionary

A screenshot of a computer

Description automatically generated

C1. My plan for assessing the data is to check for duplicates, look for null values, and then check for outliers. To check for duplicates I run this code.



This helps to find duplicates values and returns all duplicates values. For Null values I use.



This returns all null values in each column of a dataset. For Outliers I run a boxplot from matplotlib.

A screenshot of a computer program

Description automatically generated

C2 The purpose of checking for duplicates, null, values is to maintain data integrity. Dropping duplicates helps to elimate already used data. This can throw off an analyse by introducing a datapoints that already exist which might be enough to throw of a mean values or if you use a KNN method it might read a data point twice which might be enough to return an entirely different value. Or for some functions like count when you have duplicates it will return the wrong value which can throw off your analysis. Removing null values is important for data for when you run some method they return null values or errors which is not what you want when you want to see the mean of values. Finally checking for outlier are important for a couple of reasons the first reason it that it can be helpful for checking for errors if you have a scale from 1-10 and you see 11 you know that their has been an error and you can go and correct it. The secound reason it is important is to check for outliers is that outliers might be big enough to drastically change some values like with the Bill gates walked into a bar joke, showcasing how mean income increase drastically even though nothing happened for the average person mean to go up. For some outliers it might be better for the analyses just to drop those points.

C3. The Programming Language I am using is Python. You could you either Python or R for this analysis, but I have more experience with Python. I find that Pythons syntax is easier to read and understand what the code is doing. Now python might have less libraries than R, but does have the packages need for data cleaning. The packages I am using are Pandas and numpy for importing and handling the data in Python environment allowing you to check for null values and duplicated values and allowing you easily remove them from the dataset for further analysis. Fancyimpute to run a K-nearest-neighbor machine learning code this allows you make a more accurate guess for null values by looking at similar datapoints allowing you to retain alot of datapoint you might miss. Sklearn package to run a PCA model which helps you to reduce the dimensions of your dataset allowing for faster analysis and reduce noise. Matplotlib and seaborn to use visualization allowing to better a better understanding of your data and easily present your finding to customers or executives.

C.4 See attached code

D1. I found that there was no duplicates in the data set. I found that Children, Age, Income, Techie, Phone, Tenure, Bandwidth\_GB\_Year values have nulls. I also found some outliers where Population is zero which is not possible since at least one person has to live in that location. For the other quantitative values, I like to keep outliers as long as they don’t throw off the average by a huge amount and even if they do you can use median instead and it will give a similar result you were looking for.

D2. The Method I used for null values is that I used KNN from fancyimpute package to find the 5 closes datapoint to describe na values and give them a value. I picked this method for filling NA values because each customer are completely unique; so using method like backfill or frontfill will not be as accurate as looking for customer that have similar datapoints. For outliers with population I used a for loop for all unique zipcodes and found their area mean population to replace those values that have 0. I think that this method was the most accurate way of guessing a rough estimate of population size, because rural cities don’t have as much population as urban cities so I wanted to look at area to find a more accurate prediction for 0 population.

D3. I did not find any duplicates if I did I would have deleted them from the dataset since they might skew the data. This would remove all datapoints that are duplicates. For the Children, Age, Income, Techie, Phone, Tenure, Bandwidth\_GB\_Year I used K nearest nearbor to find the 5 closest data points, this returns a float with the mean of 5 nearest data points. I then converted most of these rows into Integers since they were supposed to be Boolean like values.

A screenshot of a computer

Description automatically generated

. For population I group Area, since I think Area is a best indicator for city size, and used the mean for the city zip and passed it to population.

A screenshot of a computer

Description automatically generated

I Converted Yes and No columns into 1 and 0.

A screenshot of a computer

Description automatically generated

I also removed the comma in the job tab so that we can eliminate close to 100 similar like jobs

A screenshot of a computer

Description automatically generated

D4. (BlueFeet, 2024) see code attached

D5. See CSV attached.

D.6. The disadvantage is that for population and the column with null values we do not know what their true values are at best is an estimated guess based on other data points. If you wanted a data set with only true values this data set will have a lot of mistakes and might be enough to throw off an analysis by causing the predictions to be less accurate or giving you an entire different result. But the advantage with what I did is that I was able to retain a lot of datapoint the might have been lost because you also have to factor when you drop rows with na you not only lose the data point that has the null value but you also loose the rest of the data points that are not null allows you to retain a lot more data and this might cause since you avoid overfitting when you don’t have enough data points.

D.7. We know that the data cleaning process has a lot of variables we guessed on. Now there were not missing values for churn which is good for our question ;so, we don’t have to guess for that. The issue will arise when you start doing machine learning process because we have a lot of missing values that we just took a guess on close to 2500 data points that is about ¼ of our dataset are just guess. These guesses will give an entirely different result than if we had a complete dataset this different result might be the difference between being 75% and 50% accurate costing the company a lot of money on customer churn.

E1. I inputed CaseOrder, Lat, Lng, Population, Children, Age,Income, outage\_sec\_perweek, Email, Contracts,Yearly-eqip\_failure, Tenure, MonthlyCharge, and Bandwith\_GB\_Year into the PCA model to get 14 PCA. These are all quantitative data points that can be used in a PCA.

A screenshot of a computer

Description automatically generated

E2.

A line graph with numbers and symbols

Description automatically generated

So, I am using the Kaiser method which states that you should keep all PC with a higher than 1 eigenvalues. This method would have you keep 7 PC method, I also would like to add that there is the possibility to add 3 more because they are between .99 and 0.95 which is close but are blow the 1 allowing you to keep 10 PC in total, So you would retain PC1, PC2, PC3, PC4, PC5, PC6, PC7. I eigenvalues over

E3. PCA helps to reduce the number of variables by making correlated variables into a set of uncorrelated variables. This helps to reduce noise in the data which can cause overfitting. Also by doing this it helps to reduce the number of variables in the dataset that need to be tested making the process faster and less expensive to run. Like in my PCA analysis I was able to reduce amount of data used in the PCA model from 14 to 7. This will make running analysis a lot faster and less resource intensive. Doing this all while retaining close to 70% of the data. It could also help make visualization easier plot by reducing the number of dimensions that need to be plotted. Making the visualization easier to read when presenting it to customers or

Reference

1. BlueFeet. (2024). *Filling missing values by mean in each group*. Stack Overflow. https://stackoverflow.com/questions/19966018/filling-missing-values-by-mean-in-each-group

‌