**Internet churn rate**

**D206 Data Cleaning: Performance Assessment**

**By Josue Gonzalez**

A. **what are the factors that contribute to churn rate? Can we determine which features help identify?**

B.  The data set is 10,000 customer records of a popular telecommunications company. The dependent variable in question is whether or not each customer has continued or discontinued. service within the last month. This column is titled "Churn."

Independent variables or predictors that may lead to identifying a relationship with the

dependent-dent variable of "Churn" within the dataset includes 1. Services that each customer signed up for (for example, multiple phone lines, technical support add-ons, or streaming media) 2. Customer account information (tenure with the company, payment methods, bandwidth usage, monthly charge) 3. Customer demographics (gender, marital status, income, etc.).

4. Finally, eight independent variables represent responses to customer-perceived importance of company services and features. Monthly charges are one of the biggest factors that I believe determine if a customer could have a high churn rate. Along with survey responses that will help us identify if customer service is good and satisfactory to the customer The data is both numerical (as in the yearly GB bandwidth usage, customer annual income) and categorical (a "Yes" or "No" for Churn; customer job).

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

1.  The plan is to create a copy of the data set on my own computer to be able to manipulate the data without changing the main set.

2. Read the data set into Python using pandas’ read\_csv command.

3. Evaluate the data as a whole to better understand the input data.

4. Rename the dataset as a variable churn\_df and any other subsequent slices of the data frame as “df”.

5. Examine potential misspellings, missing data fields, and nondescript variable names.

6. find outliers using histograms.

7. Imputing records missing data with a meaningful central tendency (mean, median, or mode) or simply removing the outliers that are way out of the standard deviations. (GfG, 2023)

2.  There is a lot of missing data in the set as well as unnamed columns and some columns named with not-so-descriptive names. this approach puts the data in better working condition without needing to involve methods of initial data collection or querying the data-gatherers on reasons for missing information. Using the documentation from panda’s and matplotlib along with some help from stack overflow when getting stuck with coding issues.

3. I will be using the Python Programming language as I have background knowledge of it having used it for engineering(mechanical) work for the last three years, so I have used math potlib. Python is also good for data science since it has a lot of “out of the box” (Poulson, 2016, section 2) tools needed for these types of tasks. Along with its clean style of code writing and readable syntax are other reasons for using Python. I have not used Jupyter Notebooks before, so I took it as a challenge to complete this assignment on that instead of Visual Studio code. So far, I am enjoying the single document markdown format, and being able to display the code and visualizations right on a browser as well as the ability to take the “notebook” on the go and work on it from any browser is great. At the very start, we will load in all our needed packages through the pip install feature. These will be what we will use for this assignment, NumPy - to work with arrays Pandas - to load datasets Matplotlib - to plot charts Scikit-learn - for machine learning model classes SciPy - for mathematical problems, specifically for linear algebra transformations.

4.

A screenshot of a computer

Description automatically generated

A white background with red and black text

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

 A screenshot of a computer

Description automatically generated

A screenshot of a computer

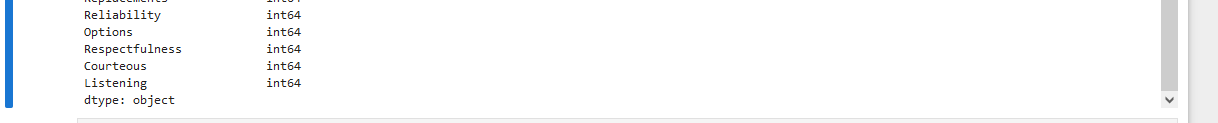
Description automatically generated

A white background with black lines

Description automatically generated

A white background with black dots

Description automatically generated



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A white background with black dots

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A white background with black dots

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

D.

1.  Cleaning Findings: There were missing data with meaningful variable fields including children, Age, Income, Tenure, and Bandwidth\_GB\_Year. Given the mean and variance of these variables, it seemed reasonable to impute missing values with median values. Many categorical (such as whether or not the customer was "Techie") & non-numeric (columns for customer ID numbers& related customer transaction IDs) data were not included in the analysis given they seemed less meaningful to interpretation and decision-making. The anomalies discovered were not significant & were mitigated as follows.

2.  mitigated missing values with imputation using median values. The monthly Charge variable shows outliers so left alone. This did not seem significant to this analysis but could warrant a deeper dive if we can collect customer feedback on why they ended their service.

3.  Cleaned dataset to leave remaining variables describing customer tenure, monthly service charge, yearly bandwidth usage & responses to survey. From there we can deduce how the churn rate would be affected by these variables.

4.  see the code above and <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=83134ec8-0d6b-4542-8442-b13400374591#>

5.  clean data(see ‘churn\_cleaned.csv).

6.   Limitations given the telecom company data set is just a small proportion of their data and may be incomplete with any updated data they may have. In this scenario, we pretty much initiated and gathered the data. So, we cannot reach out to the staff that organized & gathered this information to ask them why certain NAs are there, why are fields such as age or yearly bandwidth usage missing information. That might be relevant to answering questions about customer retention or churn. In a real-world scenario, you would be able to communicate with the data department and ask them questions about the data. And discover why the fields are empty and if it’s appropriate to fill them.

7.  while we did some imputation that may provide a path to move forward and give decision-makers reasonable answers. One way to fix this missing data issue is to have better data acquisition procedures, follow-ups, and feedback collection methods as well as change our feedback avenues instead of doing surveys after a call with an operator we could attach an incentive on the bill for a small discount if they provide better feedback about their service and satisfaction with customer support.

E.

1.  Principal Components: The principal components, and what I determined to be most important, in this dataset include survey responses to:

Timely Responses

Timely Solutions

Timely Replacements

Respectful Response

2.  Intuition about customer service suggests that feedback from user surveys offers the most important component when analyzing churn rate. Also, since survey results were the easiest to select as numeric predictors of whether or not a user would leave the company, I included the 8 responses as variables for the PCA. And, of course, users’ tenure with the company as well as monthly charges. These seem like significant numeric data points for analysis. I used a scree plot & extracted the eigenvalues for visualization of where the “elbow was bending". The elbow bent at about 3 but kept an eigenvalue above 1 until the tenth component. Then, the fewest number of components were selected based on the 86%explanation.at 7 components using the Numpy cumsum method. A rotation & loadings were created which gave us the most important variables of the dataset.

3.  The loadings suggest the variables involved in timely action regarding customer satisfaction (Responses, Fixes, Replacements & Respectfulness) should be given greater emphasis and hopefully help reduce the churn rate from the large number of 27% & "increase the retention period of customers" we get here by” reduction of noise in the data, feature selection” (bigabid). This is a more intuitive result but now the decision makers in the company have a reasonable verification of what the data is telling us.

G.  GfG. (2023, July 31). *Change column names and row indexes in pandas DataFrame*. GeeksforGeeks. https://www.geeksforgeeks.org/python-change-column-names-and-row-indexes-in-pandas-dataframe/

GfG. (2023a, April 22). *Python: Ways to rotate a list*. GeeksforGeeks. https://www.geeksforgeeks.org/python-ways-to-rotate-a-list/

Poulson, B. (2016). Data Science Foundations: Data Mining LinkedIn Learn-

ing. https://www.linkedin.com/learning/data-science-foundations-data-mining/anomaly-

detection-in-python?u=2045532

H.  sources

GfG. (2023, July 31). *Change column names and row indexes in pandas DataFrame*. GeeksforGeeks. https://www.geeksforgeeks.org/python-change-column-names-and-row-indexes-in-pandas-dataframe/

Poulson, B. (2016). Data Science Foundations: Data Mining LinkedIn Learn-

ing. https://www.linkedin.com/learning/data-science-foundations-data-mining/anomaly-

detection-in-python?u=2045532

*What is Principal Component Analysis (PCA) & How to use it?*. Bigabid. (2023, February 8). https://www.bigabid.com/what-is-pca-and-how-can-i-use-it/

Citations