D206 Assessment

Menachem Korn

1. **Question of Interest**

Can it be determined which factors have most effect whether a customer will churn or not?

1. **Variables  in the Dataset**

**CaseOrder**

• Data Type: Quantitative

• This column is a place holder to ensure the original order will be kept.

• Example: 1

**Customer\_id**

• Data Type: Qualitative

• This column contains a unique ID for the customer.

• Example: K409198

**Interaction**

• Data Type: Qualitative

• A unique ID which relates to technical interactions,

                        including sign-up, support and transactions.

• Example: aa90260b-4141-4a24-8e36-b04ce1f4f77b

**UID**

• Data Type: Qualitative

• A unique ID which relates to technical interactions,

                        including sign-up, support and transactions.

• Example: e885b299883d4f9fb18e39c75155d990

**City**

• Data Type: Qualitative

• The city included on the billing address of the customer.

• Example: Point Baker

**State**

• Data Type: Qualitative

• The state included on the billing address of the customer.

• Example: AK

**County**

• Data Type: Qualitative

• The county included on the billing address of the customer.

• Example: Prince of Wales-Hyder

**Zip**

• Data Type: Qualitative

• The zipcode included on the billing address of the customer.

• Example: 99927

**Lat**

• Data Type: Quantitative

• The latitude based on the billing address of the customer.

• Example: 56.251

**Lng**

• Data Type: Quantitative

• The longitude based on the billing address of the customer.

• Example: -133.37571

**Population**

• Data Type: Quantitative

• The population within a one-mile radius of the customer, retrieved from census data.

• Example: 38

**Area**

• Data Type: Qualitative

• Categorized area type retrieved from census data. Can be one of the following: rural, urban and suburban.

• Example: Urban

**TimeZone**

• Data Type: Qualitative

• Time zone of the customer based on sign-up info (not billing).

• Example: America/Sitka

**Job**

• Data Type: Qualitative

• Profession of the customer, based on sign-up info.

• Example: Environmental health practitioner

**Children**

• Data Type: Quantitative

• The amount of children the customer has, based on sign-up info.

• Example: 0

**Age**

• Data Type: Quantitative

• Age of the customer, based on sign-up info.

• Example: 68

**Income**

• Data Type: Quantitative

• Annual income of the customer, based on sign-up info.

• Example: 28561.99

**Marital**

• Data Type: Qualitative

• Marital Status of the customer, based on sign-up info.

• Example: Widowed

**Gender**

• Data Type: Qualitative

• Gender of the customer, based on sign-up info.

• Example: Male

**Churn**

• Data Type: Qualitative

• Whether the customer has churned or not. This is the target variable for future model predictions. Answer can be yes or no.

• Example: No

**Outage\_sec\_perweek**

• Data Type: Quantitative

• The average time in seconds the customer's neighberhood experienced system outages.

• Example: 7.978322947

**Email**

• Data Type: Quantitative

• The amount of emails that were sent to customer in the previous year.

• Example: 10

**Contacts**

• Data Type: Quantitative

• The amount of times the customer reached out to customer support. For new customers, this column reflects the amount of times customers with simillar profiles have reached out to customer support.

• Example: 0

**Yearly\_equip\_failure**

• Data Type: Quantitative

• How many times did the customer's equipment fail durng the previous year, and it was either replaced or reset. For new customers, this column reflects the amount of times customers with simillar profiles have experienced equipment failures.

• Example: 1

**Techie**

• Data Type: Qualitative

• Does the customer consider themselves technically savvy. This question was asked on the sign-up questionnaire. Answer can be yes or no.

• Example: No

**Contract**

• Data Type: Qualitative

• How long is the customer's contract for. Can be either monthly, annualy or for two years.

• Example: One year

**Port\_modem**

• Data Type: Qualitative

• Does the customer have a portable modem or not. Answer can be yes or no.

• Example: Yes

**Tablet**

• Data Type: Qualitative

• Does the customer own a tablet or not. Answer can be yes or no.

• Example: Yes

**InternetService**

• Data Type: Qualitative

• The internet service provider of the customer.

• Example: Fiber Optic

**Phone**

• Data Type: Qualitative

• Does the customer own a phone or not. Answer can be yes or no.

• Example: Yes

**Multiple**

• Data Type: Qualitative

• Does the customer have multiple phone lines or not. Answer can be yes or no.

• Example: No

**OnlineSecurity**

• Data Type: Qualitative

• Does the customer have the online security add-on or not. Answer can be yes or no.

• Example: Yes

**OnlineBackup**

• Data Type: Qualitative

• Does the customer have the online backup add-on or not. Answer can be yes or no.

• Example: Yes

**DeviceProtection**

• Data Type: Qualitative

• Does the customer have the device protection add-on or not. Answer can be yes or no.

• Example: No

**TechSupport**

• Data Type: Qualitative

• Does the customer have the technical support add-on or not. Answer can be yes or no.

• Example: No

**StreamingTV**

• Data Type: Qualitative

• Does the customer have streaming TV or not. Answer can be yes or no.

• Example: No

**StreamingMovies**

• Data Type: Qualitative

• Does the customer have streaming movies or not. Answer can be yes or no.

• Example: Yes

**PaperlessBilling**

• Data Type: Qualitative

• Does the customer have paperless billing or not. Answer can be yes or no.

• Example: Yes

**PaymentMethod**

• Data Type: Qualitative

• How does the customer pay for the service. Answer can be: check (mailed, electronic), or automatic (bank trasfer, credit card).

• Example: Credit Card (automatic)

**Tenure**

• Data Type: Quantitative

• The amount of time in months the customer has retained service.

• Example: 6.795512947

**MonthlyCharge**

• Data Type: Quantitative

• How much does the customer pay per month on average. For new customers, this column reflects the average spend value for customers with the same profile.

• Example: 172.455519

**Bandwidth\_GB\_Year**

• Data Type: Quantitative

• How many GB of data does the customer use per year on average. For new customers who haven't yet concluded a year of service, this column reflects the either the data usage approximated from their use up to this point, or average data usage for regular customers within the same demographic profile.

• Example: 904.5361102

The following variables are answers to a survey where the customer identified how important different factors are for them. Answers can be I a range from 1 to 8 with 1 being the most important issue for the customer, and 8 the least.

**Item1**

• Data Type: Qualitative

• How important is a timely response to the customer. Answers can be a number from 1 through 8, 1 being highly important and 8 being not so important.

• Example: 5

**Item2**

• Data Type: Qualitative

• How important are timely fixes to the customer. Answers can be a number from 1 through 8, 1 being highly important and 8 being not so important.

• Example: 5

**Item3**

• Data Type: Qualitative

• How important are timely replacements to the customer. Answers can be a number from 1 through 8, 1 being highly important and 8 being not so important.

• Example: 5

**Item4**

• Data Type: Qualitative

• How important is reliability to the customer. Answers can be a number from 1 through 8, 1 being highly important and 8 being not so important.

• Example: 3

**Item5**

• Data Type: Qualitative

• How important are different options to the customer. Answers can be a number from 1 through 8, 1 being highly important and 8 being not so important.

• Example: 4

**Item6**

• Data Type: Qualitative

• How important is a respectful response to the customer. Answers can be a number from 1 through 8, 1 being highly important and 8 being not so important.

• Example: 4

**Item7**

• Data Type: Qualitative

• How important is a courteous exchange to the customer. Answers can be a number from 1 through 8, 1 being highly important and 8 being not so important.

• Example: 3

**Item8**

• Data Type: Qualitative

•  How important is evidence of active listening to the customer. Answers can be a number from 1 through 8, 1 being highly important and 8 being not so important.

• Example: 4

**C1. Detecting Data Quality Issues – Discussion**

**Duplicates -** Duplicates will be detected using the pandas built-in method, df.duplicated.

**Missing Values -** Missing Values will be detected using df.info() and df.isna().sum().

**Outliers –** Outliers will be detected using df.column.std() to calculate the z-score of the datapoints thus identifying which values are considered outliers.

**Re-expression of categorical variables –** Columns which are expressed inaccurately and have the wrong data type will be detected using the df.info() method.

**C2. Detecting Data Quality Issues – Justification**

**Duplicates –** The df.duplicated method returns a True or False value for each row if duplicates are found in a given a subset of data. The returned value can later be used to filter the dataframe to only show non-duplicate data.

**Missing Values –** The df.info() method returns a Non-Null count for all columns in the data whilst df.isna().sum() counts null values per column.

**Outliers –** df.column.std() to calculate the z-score of the datapoints thus identify which values are considered outliers. An outlier is considered any data point with an absolute z-score of greater than 3 – *wgu courseware lesson 6, Z-Scores.*

**Re-expression of categorical variables –** The df.info() method shows the datatypes of each column in the dataset. The datatype can later be re-expressed after reviewing the data dictionary to determine which columns should be considered categorical.

**C4. Detection Code – *See code attached***

**Duplicates –**

df[~df.duplicated(subset = ['City', 'State', 'County' , 'Zip', 'Lat', 'Lng', 'Gender','Population'] ,keep = 'first')]

**Missing Values –**

df.info()

df.isna().sum()

**Outliers –**

df[abs(df.Population - df.Population.mean())/df.Population.std() > 3]

df[abs(df.Children - df.Children.mean())/df.Children.std() > 3]

df[abs(df.Age - df.Age.mean())/df.Age.std() > 3]

df[abs(df.Income - df.Income.mean())/df.Income.std() > 3]

df[abs(df.Outage\_sec\_perweek - df.Outage\_sec\_perweek.mean())/df.Outage\_sec\_perweek.std() > 3]

df[abs(df.Email - df.Email.mean())/df.Email.std() > 3]

df[abs(df.Contacts - df.Contacts.mean())/df.Contacts.std() > 3]

df[abs(df.Yearly\_equip\_failure - df.Yearly\_equip\_failure.mean())/df.Yearly\_equip\_failure.std() > 3]

df[abs(df.Tenure - df.Tenure.mean())/df.Tenure.std() > 3]

df[abs(df.Bandwidth\_GB\_Year - df.Bandwidth\_GB\_Year.mean())/df.Bandwidth\_GB\_Year.std() > 3]

sns.set\_style('dark')

fig, axes = plt.subplots(nrows=6, ncols=2, figsize=(15, 20))

for i in range(11):

df.boxplot(numerics[i], ax=axes[i//2, i%2])

**Re-expression of categorical variables –**

df.info()