Data Analytics Capstone Topic Approval Form

Student Name: Kevin Rupe

Student ID: XXXXXXXXXXXX

Capstone Project Name: Multiple Linear Regression Analysis on Alarm Cases for Customer Premises Equipment (CPE)

Project Topic: Analyzing influencing factors of alarm resolution on CPE.

☒ **This project does not involve human subjects research and is exempt from WGU IRB review.**

Research Question: What variables, if any, are highly correlated to the alarm being cleared on a related case?

Hypothesis:
Null hypothesis: H_0: There is no significant correlation between any of the independent variables and the alarm being cleared.
Alternate Hypothesis: H_1: There is a significant correlation between at least one of the independent variables and the alarm being cleared.

Context: The contribution of this study to the field of Data Analytics and the MSDA program is to build a predictive model that can predict with accuracy the variables that are highly correlated to the alarm clearing on a service affecting issue. Having quality internet service is paramount to business customers who depend and rely on them for their business to operate efficiently (Torkildson, n.d.). When businesses are "offline," they often can't function. This hampers their ability to run their business and ultimately, this costs them money. This can have a ripple effect on their customers as well. As a Service Provider that is providing internet service to a business, it is imperative that when their customers are experiencing service affecting events that they be repaired and/or services are restored quickly, efficiently, and correctly. Getting their customers back "online" is the most important thing, and time is of the essence. In the world today, there are countless devices that are plugged in to the internet. Businesses are no exception, usually having multiple advanced network configurations deployed on their premises. Whenever one of these devices goes offline, an alarm case is created. By applying a Multiple Linear Regression model, we can gain insights into what factors might have a high correlation to the alarm clearing. These insights would provide extremely useful intel to the Service Provider at helping them streamline processes, enhance their decision making, allocate resources, and even provide predictive insights. All of which would drive the company towards providing a better customer experience.

Data: I will need to collect data on all alarm cases from the 2024 calendar year, to date. Each case has many variables that relate to each specific device that has had an alarm event. The dataset for this analysis has 30 variables, and 124,152 observations. The dataset has variables of Case_Duration, CreatedDate, ClosedDate, City, and State. The predictor variables are broken down below:

Field	Type	Description
Arbitrary_ID	Nominal Categorical	Arbitrarily created ID to mask any sensitive information
Service_POD_Number	Nominal Categorical	Unique group number indicating the service level team that will handle the case
Customer_Comments	Discrete Quantitative	Indicates the number of times the customer interacted with the case
Assigned_Queue	Nominal Categorical	The queue in which the case is assigned
Origin	Nominal Categorical	The originating system of the case
Product	Nominal Categorical	The product of the service affecting issue
Priority	Ordinal Categorical	The priority of the case

Primary_Customer_Condition	Nominal Categorical	The condition of the customer
Service_Level	Ordinal Categorical	The service level of the customer
Customer_Impact	Ordinal Categorical	The impact level of the service affecting issue
Open_Tier_1	Nominal Categorical	First level tier related to the device and the alarm
Open_Tier_2	Nominal Categorical	Second level tier related to the device and the alarm
Open_Tier_3	Nominal Categorical	Third level tier related to the device and the alarm
Alarm_Status	Nominal Categorical	Clear (indicates the alarm has been resolved) Active (indicates the issue is still ongoing)
Issue	Nominal Categorical	Service is Down (indicates the customer has no service) Service is Impaired (indicates the customer has partial service)
Service_Status	Nominal Categorical	Out of Service (indicates that the customer is out of service) In Service (indicates that the customer is not out of service)
Closing_Fault	Nominal Categorical	The fault of the alarm
Closing_Issue	Nominal Categorical	The issue of the alarm
Closing_Resolution	Nominal Categorical	The resolution of the alarm
Action_Code	Nominal Categorical	The action of the alarm
Cause_Code	Nominal Categorical	The cause of the alarm
Class_Item_Code	Nominal Categorical	The class item of the alarm

The dataset is owned by XXXXXXXXXXXXX. Permission to use this dataset has been authorized. The signed authorized release form will be submitted with Task 1.

Limitations: This dataset is limited by the fact that all the variables only point to what is known by the ISP (Internet Service Provider). There are many factors that can cause an alarm that are not easily identifiable. For example, the internet is connected by all sorts of different ISP's, and the issue may be something related to another carrier's network, and therefore might not be fully known. Other issues with alarms could be related to Utility Company power outages, or even natural disasters.

Delimitations: The dataset is delimited by only displaying States from USA, Canada, and other US regions (such as Guam, Puerto Rico, and US Virgin Islands). Because it is important to have no missing data to build a Multiple Linear Regression model, I have removed any row which had missing data in any of the variables.

Data Gathering: Data was gathered directly from a Snowflake database that stores tabular data of alarm cases for enterprise customers' CPE (Customer Premises Equipment). A SQL query was performed to extract the data into a .csv file. The SQL query pulled in the dependent and many independent variables for alarm cases. To extract the most useful data for the organization, certain filters were applied such as only retrieving cases that are currently closed that were opened in the year 2024. The data was also treated by excluding null rows on all variables. The City and State variables were filtered to only display actual city names, removing any names that begin with a number; the State column was filtered to only display actual State names from USA, Canada, and US regions (such as Puerto Rico, Guam, Virgin Islands, etc.).

For Customer_Comments, there appears to be several outliers. These outliers will need to be evaluated by visualizing the box plot chart using Seaborn (Waskom, 2021). Any outliers should be treated using median imputation.

In order to successfully perform a Multiple Linear Regression analysis, I will need to convert several categorical variables using One Hot Encoding, Ordinal Encoding or use Dummy variables (Pandas, 2024). Upon inspection,

these variables will be Origin, Product, Priority, Primary_Customer_Condition, Service_Level, Customer_Impact, Open_Tier_1, Open_Tier_2, Open_Tier_3, Issue, Service_Status, Closing_Fault, Closing_Issue, Closing_Resolution, Action_Code, Cause_Code, and Class_Item_Code.

Data Analytics Tools and Techniques: Given the number of independent variables, multicollinearity may possibly be a problem with this dataset. Therefore, Statsmodels' variance_inflaction_factor function will be performed to check and remove any variables that are multicollinear. An OLS (Ordinary Least Squared) model will be performed before and after removing multicollinearity to view the differences and validity of the model (Seabold, 2010).

Univariate and bivariate graphs will be presented on Task 3 including a Tableau Dashboard.

Justification of Tools/Techniques: Multiple Linear Regression is a technique that analyzes how certain variables are correlated to another. Statsmodels package in Python will help determine how closely the independent variables correlates to the dependent (Seabold, 2010). Seaborn and Matplotlib libraries will provide visual representations of how closely these variables are linearly related (Waskom, 2021). Multiple Linear Regression allows the analyst to predict with some certainty what level of effect the explanatory variables have on the dependent variable (Massaron, 2016).

Project Outcomes: The project will seek to create a multiple linear regression model to discover if there are any correlations in the predictor variables to why an alarm on a device clears (i.e., is resolved). The alternative hypothesis is supported by Shillingsburg (2024) in that predictive analytics can help an ISP (Internet Service Provider) become more proactive in its approach to network optimization than strictly reactionary.

Projected Project End Date: 01/31/2025

Sources:

Massaron, L. & Boschetti, A. (2016). Regression Analysis with Python: Learn the Art of Regression Analysis with Python. Packt Publishing.

Pandas (2024, September 20). Pandas documentation. Retrieved December 10, 2024, from https://pandas.pydata.org/docs/index.html.

Seabold, S. & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Scientific Conference. Retrieved December 10, 2024, from https://www.statsmodels.org/stable/index.html.

Waskom, M. L. (2021). Seaborn: statistical data visualization. Retrieved December 10, 2024, from https://seaborn.pydata.org/index.html.

Shillingsburg, S. (2024, October 30). Predictive Analytics for Network Optimization. Retrieved December 19, 2024, from https://sonar.software/blog/predictive-analytics-for-network-optimization.

Torkildson, A. (2017-2024). Why The Internet is so Important for Modern Business. Retrieved December 19, 2024, from https://www.thinkers360.com/tl/blog/members/why-the-internet-is-so-important-for-modern-business.

Course Instructor Signature/Date:

☒ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Instructor's Approval Status: Approved

Date: 12/19/2024

Reviewed by:

Comments: Click here to enter text.