D214 – Data Analytics Graduate Capstone

Executive Summary

Western Governors University

January 5, 2025

Kevin Rupe
Student ID: XXXXXX
Email: krupe6@wgu.edu
Phone: (864) 704-2340

PROBLEM STATEMENT AND HYPOTHESIS

In our ever-evolving world of technology with the exponentially expanding Internet of Things (IoT), as a society in the United States we have more devices online than ever before. Having quality, fast, and reliable internet service is paramount to business customers who depend and rely on them in order for their businesses to operate efficiently (Torkildson, n.d.). When businesses are offline it hampers their ability to run their business effectively.

For Internet Service Providers (ISPs) that are 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. Afterall, if the ISP struggles to provide excellent service to their customers, then in the end, their customers will find another ISP who can. Therefore, this has a negative impact on the ISPs revenue.

Whenever one of these devices goes offline, an alarm case is created. The hypothesis of this research is that there will be at least one variable from the alarm case that has significant correlation to the alarm clearing (i.e., service affecting issue being resolved).

DATA ANALYSIS SUMMARY

The data was collected by extracting alarm case data directly from the Snowflake database using an advanced SQL query into a CSV file. Null records and duplicate rows were eliminated from the dataset prior to extracting the data. Once the data was extracted it was then imported into Jupyter Notebook where the data had to be prepared prior to analysis. Non-predictor variables were dropped from the dataset that were not needed for this Multiple Linear Regression (MLR) analysis.

Outliers from CUSTOMER_COMMENTS were treated using median imputation. Predictor variables were converted from categorical to numerical using One-Hot Encoding, Ordinal Encoding, and also using dummy variables. To reduce dimensionality in the dataset, several variable names were combined by grouping them by updating the name. Multicollinearity was eliminated from the remaining variables, and all non-statistically significant variables were dropped from the final MLR model.

OUTLINE OF THE FINDINGS

- Adjusted R-squared and R-squared values are only ~4.5% which shows that not much of the variability in the alarm clearing can be attributed to the independent variables.

- The CSOC Queues have the strongest negative effect on the alarm clearing (-0.7127).

- The Outage Queues have the strongest positive effect on the alarm clearing (0.1439).

- Large Omnibus and Jarque-Bera tests show that there is a non-normality in the residuals.

- F-statistic of 306.1 with a p-value of 0.00 assures us this model is statistically significant.

- Positive correlation to alarm clearing:
  - Primary Customer Condition
  - Service Status
  - Service Level: Enterprise
  - Assigned Queue: Outage
  - Open Tier 2: LAN, Switch
  - Open Tier 3: Outage Data, Customer Down, Customer Degraded

- Negative correlation to alarm clearing:
  - Service POD Number
  - Origin: Customer Netcool
  - Customer Comments
  - Assigned Queue: CSOC
  - Open Tier 2: Ethernet
  - Open Tier 3: Proactive Multiple Underlay, Proactive HA, Proactive 4G Wireless, Jitter, Latency, Packet Loss, SD-WAN HA

OLS Regression Results

Dep. Variable:	ALARM_Clear	R-squared:	0.045
Model:	OLS	Adj. R-squared:	0.044
Method:	Least Squares	F-statistic:	306.1
Date:	Fri, 27 Dec 2024	Prob (F-statistic):	0.00
Time:	23:40:07	Log-Likelihood:	-50842.
No. Observations:	124738	AIC:	1.017e+05
Df Residuals:	124718	BIC:	1.019e+05
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.7776	0.004	195.782	0.000	0.770	0.785
CUST_COND_NUMERIC	0.0177	0.001	20.821	0.000	0.016	0.019
SERVICE_POD_NUMBER	-0.0031	0.000	-10.942	0.000	-0.004	-0.003
SVCST_OOS	0.0652	0.003	25.933	0.000	0.060	0.070
ORIGIN_Cust_Netcool	-0.1032	0.003	-29.658	0.000	-0.110	-0.096
OT2_LAN	0.0227	0.006	3.553	0.000	0.010	0.035
OT3_Pro_Multi	-0.0761	0.005	-16.017	0.000	-0.085	-0.067
SVCLVL_Enterprise	0.0229	0.004	5.397	0.000	0.015	0.031
OT3_Outage_Data	0.0584	0.004	13.068	0.000	0.050	0.067
OT3_Pro_HA	-0.0405	0.006	-6.560	0.000	-0.053	-0.028
CUSTOMER_COMMENTS	-0.0082	0.003	-3.247	0.001	-0.013	-0.003
OT3_Cust_Down	0.1021	0.006	15.800	0.000	0.089	0.115
OT3_Pro_4GW	-0.0720	0.006	-12.360	0.000	-0.083	-0.061
OT2_ETH	-0.1003	0.013	-7.989	0.000	-0.125	-0.076
OT2_SWITCH	0.0673	0.015	4.501	0.000	0.038	0.097
QUEUE_Outage	0.1439	0.007	22.037	0.000	0.131	0.157
OT3_JL_PL	-0.1593	0.008	-21.029	0.000	-0.174	-0.144
OT3_SDWAN_HA	-0.1681	0.009	-17.817	0.000	-0.187	-0.150
QUEUE_CSOC	-0.7127	0.018	-40.035	0.000	-0.748	-0.678
OT3_Cust_Deg	0.1249	0.024	5.113	0.000	0.077	0.173

Omnibus:	34012.934	Durbin-Watson:	1.768
Prob(Omnibus):	0.000	Jarque-Bera (JB):	69491.997
Skew:	-1.721	Prob(JB):	0.00
Kurtosis:	4.236	Cond. No.	164.

EXPLANATION OF THE LIMITATIONS OF THE TECHNIQUES AND TOOLS

The limitation of needing to convert variables to numerical is that it adds dimensionality to the dataset. This adds complexity and potentially more issues such as multicollinearity which then needs to be addressed also (Geeks for Geeks, 2023). Statsmodel's *variance_inflation_factor* (VIF) function has a drawback in that when using many variables, it often will provide only a portion of the multicollinearity in the data. To counteract this disadvantage, only a small portion of variables should be removed at a time, and then check again. This process is tedious, but is a crucial step in removing the correct variables that have multicollinearity.

SUMMARY OF PROPOSED ACTIONS

Quite a few variables in this dataset were multicollinear which are adding complexity to the data while not adding any benefit. There were also a handful of variables that had no statistical significance to the alarm clearing. The recommended course of action is for the ISP to simplify the level of detail that the alarm cases are given.

A second proposed direction of approach on a future study would be to create another MLR model looking at the "Closing" codes instead of the "Open" codes. Perhaps this could improve the model's explanatory power since more knowledge is understood at the end of the case than at the beginning.

Lastly, another direction to take would be to focus the MLR model on the service status instead of the alarm clearing. Service status variable is historical and static, whereas the alarm clearing object on the case dynamically changes when the alarm itself clears or activates. This would be useful information for the ISP that might explain more about why some products, queues, origins, etc., tend to correlate to "Out of Service" events.

EXPECTED BENEFITS OF THE STUDY

Predictive Analytics can greatly help an ISP become more proactive in its approach to network optimization than strictly being reactionary (Shillingsburg, 2024). By minimizing downtime, the ISP can provide better service to its customers, which in turn has a positive correlation to its revenue. Being able to accurately predict the future alarms on customers' network devices would aid the ISP in streamlining processes, enhancing their decision making, and allocating resources. All of which would drive the company towards providing a better customer experience.

SOURCES

Geeks for Geeks (May 06, 2023). Introduction to Dimensionality Reduction. Retrieved December 27, 2024, from https://www.geeksforgeeks.org/dimensionality-reduction/.

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.