Tonie Cooper

Master Data Analytics

[Tcoo158@wgu.edu](mailto:Tcoo158@wgu.edu)

D209 Task 2 Assessment

A1) What customers are at a high risk of churn using the ridge regression analysis?

A2) Stakeholders will have confidence when making decisions on what customers they should focus their assets on to reduce the churn rate.

B1) Ridge Regression is a technique used for analyzing multiple regression data that are affected by multicollinearity. Multicollinearity is the existence of near-linear relationships among independent variables. In ridge regression we start off by standardizing the variables, subtracting the mean the dividing by the standard deviation. In Python we perform L2 regularization which adds a factor of sum of squares of coefficients. This minimizes the sum of squares error by applying a penalty to the coefficients. Doing this results in predicted values to be far away from the actual values (Kargin 2021).

The expected outcome from this analysis is to have a low mean squared error that will follow close to the predicted values and with only having a limited number of variables to add to the model I believe the R2 value will be low as well.

B2) An assumption of ridge regression is that there’s no outliers. Outliers will create false and misleading results that will hurt the decision-making process.

B3) Packages for Python:

Pandas

Numpy

Scikit-learn

Matplotlib

Seaborn

Matplotlib, Pandas and Numpy are standard imports that provide statistical packages for reading, scoring and visualization. Seaborn contains descriptive graphs, matrix, and plots. Scikit-learn allows for splitting, fitting, predicting, and applying the metrics

C1) One data preprocessing goal is to change the qualitative binary data to a dummy variable 1/0

C2) Continuous predictor variables:

Children

Income

Outage\_sec\_perweek

Email

Contacts

Yearly\_equip\_failure

Tenure

MonthlyCharge

Bandwidth\_GB\_Year

Categorical predictor variables:

Techie

Contract

Tablet

InternetService

Phone

Multiple

OnlineSecurity

OnlineBackup

DeviceProtection

TechSupport

StreamingTV

StreamingMovies

C3) Steps to prepare the dataset:

Backup data to my computer – to prevent total loss of data

Read the csv file into Python – data needed to run analysis

Describe the data – find what data needs to be changed to a quantitative dummy variable to allow for calculations. This also allows for better understanding of the data available.

Name dataset as churn\_df and the dataframe as df

Search for misspellings, missing data and correct that data – ensure data are alike to prevent unknown errors and unintentional outliers. This will also allow us to utilize mean, median and mode to fill in missing data where applicable.

Use descriptive statistics to locate outliers by using histograms and boxplots – outliers can disrupt and cause a false calculation. I’ll remove them to have a more precise calculation on this analysis.

Remove less meaningful data – RR will slow down when too much data is introduced to it so removing data that won’t play a role in this analysis is crucial. Removing less meaningful data also reduces the chances of errors accruing.

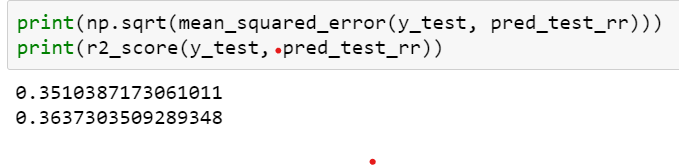
Extract cleaned dataset for use in the ridge regression model – extract dataset to allow for usage in the RR algorithm.

Additionally, when cleaning the dataset dummy variables will be created to change qualitative data to quantitative data (1/0) to allow for calculations to take place and the data entered the RR analysis.

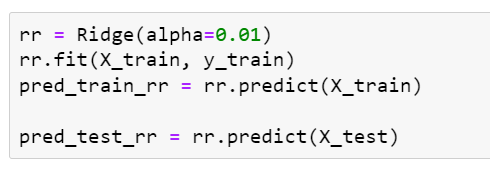
Lastly, discrete variables, shown as survey items, will have their names changed from “Items” to a name that represents what the Item is such as Response, Timely Fixes, Options, etc. This allows for better readability and analysis of the customers opinion on what they deem most important in a telecommunication service.

D1) See Attachment

D2) For this prediction model I utilized the ridge regression method. With this model I was able to achieve a RMSE of 0.3510 and a an R2 value of 0.3637



I chose to have an alpha level of 0.01 when calculating the ridge regression. I then fit the training model and calculated the predicted training and test values.



D3) See Attachment

E1) I had an R2, accuracy, value of 0.3637 which is very low possibly due to the number of independent variables used for this analysis. With R2 being from 0 to 100% a 36% R2 tells us that variance is more spread out. As we get closer to 100% that would tell us the data points are perfectly fitted to the regression line. Per Frost 2022 a low R2 value doesn’t mean there’s necessarily a problem. Some areas have a greater level of unexplainable variance and human behavior tends to be harder to predict. Additionally, I had a MSE of 0.3510 which is in line with my prediction of having a low MSE that will follow close to the prediction line. A low MSE means that I’m closer to finding the line of best fit. The lower the MSE the better the forecast will be. (NCSS 2019).

E2) The independent variables utilized for this model are 'Outage\_sec\_perweek', 'MonthlyCharge', 'Bandwidth\_GB\_Year'. These were running against the dependent variable of ‘DummyChurn’. With this regression model I ended with an R2 value of 36.37% and a MSE of 35.10%. This followed my expected outcome due to the low number of variables used and the similarity between the variables as shown through multiple different analysis with this data set. R2 is based on a scale from 0 to 100%. With a score of 36.37% this tells us that the variance of the data points is more spread out. However, this isn’t necessarily a problem as discussed above. Now, the MSE being rated at 35.10% tells us how the points are fitted to the regression line. A lower value such as the one produced means that the data points are near the line of best fit and a lower percentage is better for forecasting.

E3) A limitation of the ridge regression analysis is that it will decrease the complexity of the model, however, it will not decrease the number of variables used. This model should not be used for large variable reduction (pawangfg 2022).

E4) With a low MSE and R2 score as stated above it’s recommended that the client understands the accuracy of this and should analyze the items that correlate with the customers leaving. They should attempt to reduce those grievances and suggest to more customers the available services that the company provides. It is in the best interest of client to outages by utilizing high quality products and maintaining customers by letting them know of the current deals that are available to them within their monthly price range.

F1) <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=aa799bac-47c4-4e50-a73c-ae9a01373eda>

G1) Code Citations:

Brownlee, J. (2020, October 10). *How to develop ridge regression models in Python*. Machine Learning Mastery. Retrieved May 19, 2022, from https://machinelearningmastery.com/ridge-regression-with-python/

H1) Citation:

NCSS. (2019, April 1). *Chapter 335 ridge regression - ncss-wpengine.netdna-ssl.com*. ncss-wpengine. Retrieved May 19, 2022, from https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge\_Regression.pdf

Grace-Martin, K. (2022, May 9). *Measures of model fit for linear regression models*. The Analysis Factor. Retrieved May 19, 2022, from <https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/>

Kargın, K. (2021, April 17). *Ridge regression fundamentals and modeling in Python*. Medium. Retrieved May 19, 2022, from https://keremkargin.medium.com/ridge-regression-fundamentals-and-modeling-in-python-bb56f4301f62

pawangfg. (2022, February 11). *Lasso vs Ridge vs elastic net: ML*. GeeksforGeeks. Retrieved May 19, 2022, from <https://www.geeksforgeeks.org/lasso-vs-ridge-vs-elastic-net-ml/#:~:text=Limitation%20of%20Ridge%20Regression%3A%20Ridge,not%20good%20for%20feature%20reduction>.

Frost, J., Yunitasari, N., Schwant, H. C., Levitt, A. (2022, February 27). *How to interpret R-squared in regression analysis*. Statistics By Jim. Retrieved May 19, 2022, from https://statisticsbyjim.com/regression/interpret-r-squared-regression/#:~:text=R%2Dsquared%20evaluates%20the%20scatter,multiple%20determination%20for%20multiple%20regression.