Raquel Ocasio

D209 – Data Mining I, Task 2

October 24, 2023

Western Governors University

Part A1

Which variables can help estimate the tenure of a customer?

Part A2

The goal of the data analysis is to be able to estimate the number of months the customer will stay with the provider by using Lasso regression.

Part B1

Lasso regression identifies the most important features or predictors in a dataset by setting some of the coefficients of less important features to zero. This is useful in reducing the dimensionality of the model when there is a large number of features. The expected outcome is a number that represents how many months the customer stays with the provider.

Part B2

One assumption of Lasso regression is that the relationship between the target variable and the predictor variables is linear.

Part B3

The glmnet library provides access to methods for implementing Lasso regression and related calculations.

Part C1

One of the data preprocessing goals is ensure that there a no missing values.

Part C2

Variable name	*Variable description*	*Variable type*
Outage_sec_perweek	The average number of seconds per week of service outages in the customer's neighborhood.	Numeric
Email	The number of emails sent to the customer in the last year.	Numeric
Contacts	The number of times the customer contact technical support.	Numeric
Yearly_equip_failure	The number of times the customer's equipment failed	Numeric

Variable name	*Variable description*	*Variable type*
	in the past year and had to be replaced or reset.	
MonthlyCharge	The monthly amount that is charged to the customer.	Numeric
Bandwidth_GB_Year	The yearly amount of data used by the customer in GB.	Numeric

Part C3

The data was prepared for analysis by detecting and treating duplicate values, missing values, and outlier values.

Duplicate values are detected using the sum(duplicated()) functions. No duplicate values were detected. The code segment is starts at line number 11.

Missing values are detected using the colSums(is.na()) functions. No missing values were detected. The code segment is starts at line number 14.

Outliers are detected using a function that uses the mean(), sd(), and sum() functions to calculate the z-score and count how many z-scores have a value greater than three or less than negative three. The code segment is on lines 18-47.

Part C4

See attached file.

Part D1

See attached files. (Zach, 2022)

Part D2

The Lasso regression technique was used to analyze the data. Lasso regression works by starting with a linear regression, then penalizing less significant features by reducing their coefficient to zero.

Intermediate calculations were performed for the Total Sum of Squares (sst), Sum of Squared Errors (sse), R-Squared (rsq), and the Mean Squared Error (MSE). Screenshots of the intermediate calculations are below.

```
> #find SST and SSE
> sst <- sum((y - mean(y))^2)
> sst
[1] 6991656
> sse <- sum((y_predicted - y)^2)
> sse
[1] 90665.8
>
> #find R-Squared
> rsq <- 1 - sse/sst
> rsq
[1] 0.9870323
>
> # calculate MSE
> MSE <- sum((y - y_predicted)^2)/10000
> MSE
[1] 9.06658
```

Part D3

See attached code file.

Part E1

The accuracy of the model is determined by the R-Squared value. In this case, the R-Squared value of 0.9870323 indicates that the model can explain 98.70% of the variation in the response values of the training data. (Zach, 2020)

The MSE of the model is 9.06658. A lower MSE indicates that the model's predictions are closer to the actual values, which is a sign of better predictive accuracy. (Zach, 2020a)

Part E2

Results: When using the Outage_sec_perweek, Email, Contacts, Yearly_equip_failure, MonthlyCharge, and Bandwidth_GB_Year variables, the classification model correctly predicted the correct Tenure result 98.70% of the time.

Implications: While the model can predict reliably, it can possibly be improved further by including more explanatory variables.

Part E3

One limitation of the data analysis is, when two or more variables are highly correlated, Lasso regression will arbitrarily select which variables to remove. (GeeksforGeeks, 2023)

Part E4

It is recommended that the company use the prediction model when determining how long the customer will remain with the company. It is also recommended that the company try variations of the model by adding or removing variables to achieve a more accurate result.

Part F

The demonstration can be viewed at https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=d457079e-533e-4825-a715-b0a600176deb

Part G

GeeksforGeeks. (2023, January 10). Lasso vs Ridge vs Elastic Net ML. https://www.geeksforgeeks.org/lasso-vs-ridge-vs-elastic-net-ml/

Zach. (2020a, April 6). How to calculate MSE in R. Statology. https://www.statology.org/how-to-calculate-mse-in-r/

Zach. (2020, November 13). Lasso regression in R (Step-by-Step). Statology. https://www.statology.org/lasso-regression-in-r/

Zach. (2022, April 12). *How to Split Data into Training & Test Sets in R (3 Methods)*. Statology. https://www.statology.org/train-test-split-r/

Part H

None used.