**INSY 662: Data Mining and Visualization (Fall 2023)**

**Individual Assignment 1**

**Due by Monday, September 25, 2023 at 11:59pm on MyCourses**

* This is an individual assignment. The McGill Academic Integrity code applies to this and other assignments!
* You need to submit **TWO FILES to MyCourse**. One, **a Word file with your answers**. You can type or select your answers (highlighted places) directly on this assignment file. Two, you should also submit **your Python code (in .py format)**.
* The goal of this assignment is to understand how to apply **linear regression, ridge regression, and LASSO**.

**You need to download the dataset named “ToyotaCorollas.csv” from MyCourse.**

|  |
| --- |
| *Background:*  A large Toyota dealership (in the Netherlands) offers purchasers of new Toyota cars the option to buy their used car as part of a trade-in. The dealer then sells the used cars for a small profit. To ensure a reasonable profit, the dealer needs to be able to understand what factors influence the price and also predict the price that the dealership will get for the used cars. For that reason, data were collected on all previous sales of used Toyota Corollas at the dealership.  Below are descriptions of the variables:  **ID:** Unique identifier for each record  **Model:** Car model  **Price:** Offer price in Euros  **Age\_08\_04:** Age in months as of August 2004  **KM:** Accumulated kilometers on odometer  **HP:** Horsepower  **Automatic:** Whether a car is an automatic or not  **CC:** Cylinder volume in cubic centimeters  **Doors:** Number of doors  **Cylinders:** Number of cylinders  **Gears:** Number of gear sets  **Weight:** Weight in kilograms |

Task 0. Define the predictors as all variables except “ID,” “Model,” and “Price,” and define the target variable as the variable “Price”.

Task 1. Check multicollinearity for each predictor. Based on the measure for detecting multicollinearity, which of the following predictors should be removed?

No need to remove any predictors

If you choose to remove a predictor, remove it from your predictor dataframe.

Task 2. Next, standardize the predictors.

Task 3. Then, separate the dataset into training (65%) and test (35%). Specify random\_state = 662 when splitting the data for reproducibility purposes.

Task 4. Use the sklearn package and develop a linear regression model based on the training dataset. Use that model to predict the target value of the test dataset and report the MSE below.

MSE (round to the nearest integer) : 1626925

Task 5. Use the sklearn package and develop a ridge regression model (with alpha = 1) based on the training dataset. Use that model to predict the target value of the test dataset and report the MSE below.

MSE (round to the nearest integer) : 1626677

Task 6. Use the sklearn package and develop a LASSO model (with alpha = 1) based on the training dataset. Use that model to predict the target value of the test dataset and report the MSE below.

MSE (round to the nearest integer) : 1627430

Based on the MSE observed so far, Ridge performs the best in predicting the target variable.

Task 7. Develop another ridge regression model and Lasso model with the following alpha value, and report the MSE (round to the nearest integer) of those models accordingly:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Alpha = 10 | MSE of ridge regression: | 1625151 | MSE of LASSO: | 1631231 |
| Alpha = 100 | MSE of ridge regression: | 1661808 | MSE of LASSO: | 1685659 |
| Alpha = 1000 | MSE of ridge regression: | 3219985 | MSE of LASSO: | 3351632 |
| Alpha = 10000 | MSE of ridge regression: | 3219985 | MSE of LASSO: | 3351632 |

At alpha = 10000, report the coefficients of the following variables (If you decided to remove a predictor, put ‘N/A’):

|  |  |  |
| --- | --- | --- |
|  | Ridge | Lasso |
| Age\_08\_04 | -255.49853171 | -2323.58993 |
| KM | 162.00766784 | -721.55495729 |
| HP | 79.02586815 | 434.74989115 |
| Automatic | 7.51148239 | 85.36114513 |
| CC | 41.06424943 | - 3.00538181 |
| Doors | 49.93896511 | 0. |
| Cylinders | N/A | 0 |
| Gears | 11.84737094 | 80.09478236 |
| Weights | 181.0063638 | 910.54914831 |

Compare the MSEs for Ridge and LASSO at alpha = 10000 and linear regression. Which one performs the best in predicting the target variable and why?

Linear regression performs the best in predicting the target variable. That is because The original model was not suffering from high variance

\*\*End of Assignment 1\*\*