# Lasso & Ridge Regression (Module - 8)

**Instructions**

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: Nithin Dsouza**

**Batch Id: 05012021-10AM**

**Topic: Lasso Ridge Regression**

1. **Business Problem**
   1. **Objective**
   2. **Constraints (if any)**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.**

**Using R and Python codes perform:**

1. **Data Pre-processing**

**3.1 Data Cleaning, Feature Engineering, etc.**

**3.2 Outlier Imputation**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary**
   2. **Univariate analysis**
   3. **Bivariate analysis**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options)**
   2. **Perform Lasso and Ridge Regression Algorithm**
   3. **Train and Test the data and compare RMSE values tabulate R-Squared values, RMSE for different models in documentation and provide your explanation on it**
   4. **Briefly explain the model output in the documentation.**



1. **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

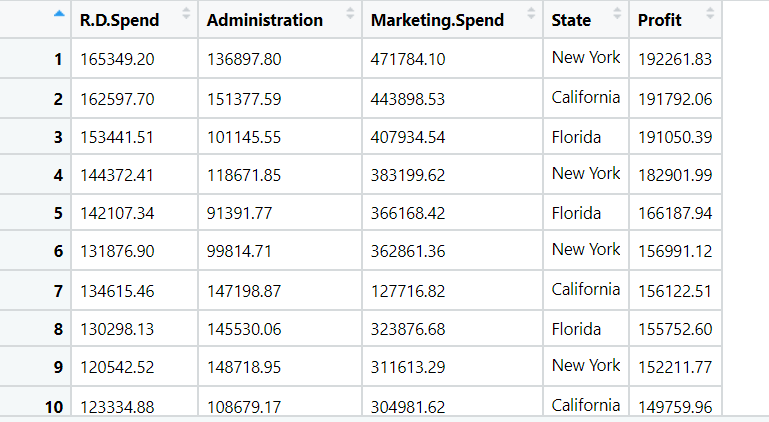
# Note:

The assignment should be submitted in the following format:

* R code
* Python code
* Code Modularization should be maintained
* Documentation of the model building (elaborating on steps mentioned above)

**Problem Statement: -**

An Analytics Company has been tasked by a crucial job of finding out what factors does affect a startup company and will it be profitable to do so or not. For this, they have collected some historical data and would like to applying supervised predictive learning algorithm such as Lasso Ridge Regression on it and provide brief insights about their data. Predict Profit, given different attributes for various startup companies.



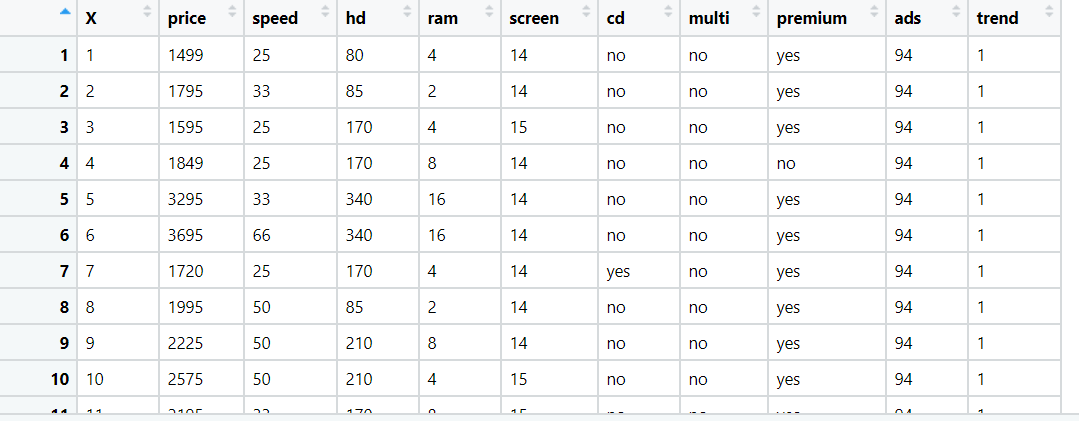
|  |  |  |  |
| --- | --- | --- | --- |
| Name of Feature | Description | Type | Relevance |
| R&D Spend | Research and development spend | Continuous, Ratio | It provides useful information |
| Administration | Administration Spend | Continuous, Ratio | It provides useful information |
| Marketing Spend | Marketing Spend | Continuous, Ratio | It provides useful information |
| State | State Name | Discrete, Nominal | It is not useful in analysis |
| Profit | Profit of company | Continuous, Ratio | It provides useful information |

**Business Objective**: Maximize accuracy between independent input variables and dependent output variable

**Business Constraint**: Minimize linearity problem

**Problem Statement: -**

Officeworks, is a leading retail store in Australia, with numerous outlets around the country, the manager would like to improve their customer experience by providing them online predictive prices about their gadgets/ Laptops if they wants to sell them. To improve this experience the manager would like us to build a model which is sustainable and accurate enough, to get the objective achieved. Apply Lasso Ridge Regression model on the dataset and predict Price, given other attributes and tabulate R squared ,RMSE and correlation values.



**Objective:** Maximize the accuracy in estimating relationship between 2 or more independent variable and one dependent variable.

**Constraints:** Limited to linear Relationships.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Description** | **Type** | **Relevance** |
| X | Index Values | Nominal | Irrelevant, Doesn’t Provides useful information. |
| price | Price of gadget | Continuous, Ratio | Relevant, Provides useful information. |
| speed | Gadget speed | Continuous, Ratio | Relevant, Provides useful information. |
| hd | Scale of HD | Discrete, count | Relevant, Provides useful information. |
| ram | RAM of device | Discrete, count | Relevant, Provides useful information. |
| screen | Screen size | Discrete, count | Relevant, Provides useful information. |
| cd | CD supported | Categorical, Ordinal | Relevant, Provides useful information. |
| multi | Multimedia supported | Categorical, Ordinal | Relevant, Provides useful information. |
| premium | Premium device | Categorical, Ordinal | Relevant, Provides useful information. |
| ads | ads | Discrete, count | Relevant, Provides useful information. |
| trend | Trend | Discrete, count | Relevant, Provides useful information. |

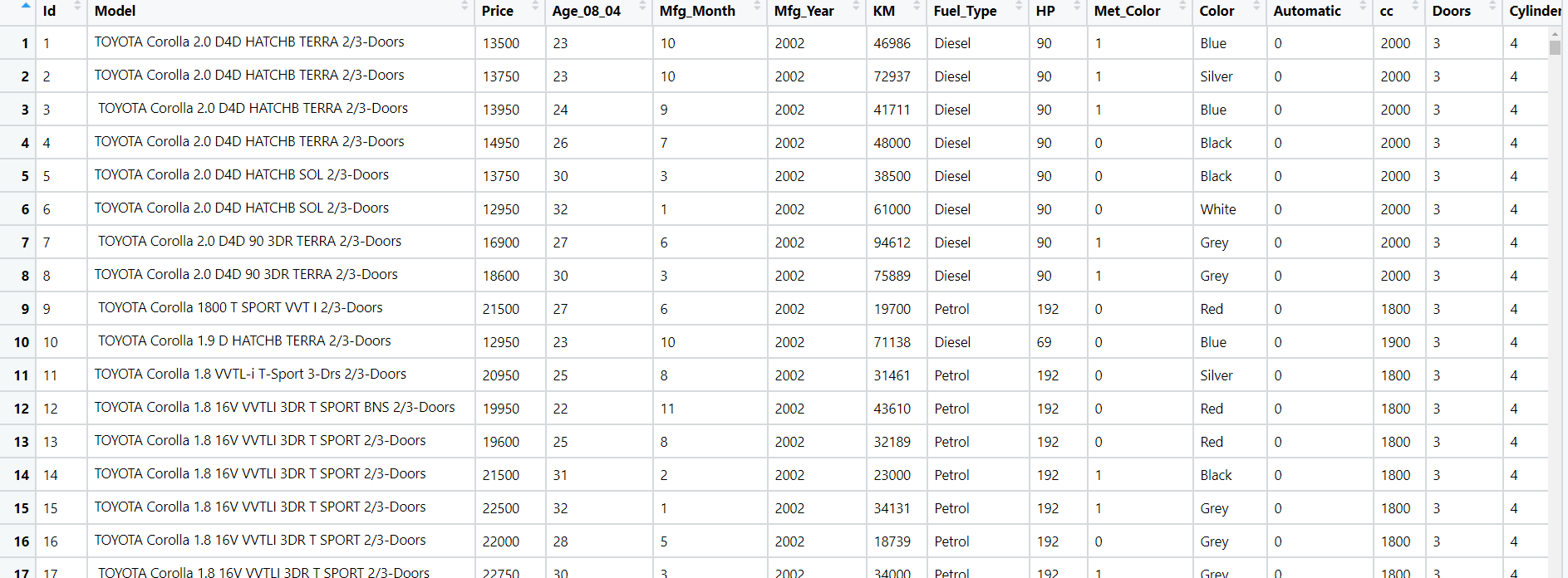


**Problem Statement: -**

An online car sales platform would like to improve its customer base and their experience by providing them an easy way to buy and sell cars. For this, they would like to have an automated model which can predict the price of the car if user inputs the required factors. Help the business achieve the objective by applying Lasso and Ridge regression model on it.

Please use the below columns for the analysis purpose.

Price, Age\_08\_04, KM, HP, cc, Doors , Gears, Quarterly\_Tax, Weight



**Objective:** Maximize the accuracy in estimating relationship between 2 or more independent variable and one dependent variable

**Constraints:** Limited to linear Relationships.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Description** | **Type** | **Relevance** |
| Price | Price of the car | Continuous, Ratio | It provides useful information |
| Age\_08\_04 | Age of a person | Continuous, Ratio | It is not useful |
| KM | Total Travelled Distance | Continuous, Ratio | It provides useful information |
| HP | Horse Power | Continuous, Ratio | It is a useful information |
| Cc | Cubic Capacity | Continuous, Ratio | Useful information |
| Doors | Number of doors in car | Discrete, Count | It is useful information |
| Gears | Number of gears | Discrete, Count | It provides useful information |
| Quarterly Tax | Insurance Tax for car | Continuous, Ratio | It provides useful information |
| Weight | Total Weight of car | Continuous, Ratio | It provides useful information |

**Problem Statement: -**

Data of various countries and the factors affecting their Life expectancy has been recorded over past few decades. An analytics firm would like to know how it varies country wise and what other factors are influential in model building. Use your skills to analyze the data and build a Lasso and Ridge Regression model and also summarize the output of the model.

Snapshot the dataset is given below: -

**Objective:** Maximize the accuracy in estimating relationship between 2 or more independent variable and one dependent variable

**Constraints:** Limited to linear Relationships.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of feature** | **Description** | **Type** | **Relevance** |
| Country | Name of country | Discrete, Nominal | Not useful |
| Year | Year | Continuous, Ratio | It is useful |
| Status | Status of country | Discrete, Nominal | Not useful |
| Life\_expectancy | Average Life | Continuous, Ratio | It is useful |
| Adult\_Mortality | Average Adults | Continuous, Ratio | It is useful |
| infant\_deaths | Death Rate | Continuous, Ratio | It is useful |
| Alcohol | Alcohol percentage | Continuous, Ratio | It is useful |
| percentage\_expenditure | percentage expenditure | Continuous, Ratio | It is useful |
| Hepatitis\_B | Disease | Continuous, Ratio | It is useful |
| Measles | Disease | Continuous, Ratio | It is useful |
| BMI | Body mass index | Continuous, Ratio | It is useful |
| under\_five\_deaths | Death rate | Discrete, Count | It is useful |
| Polio | Disease | Continuous, Ratio | It is useful |
| Total\_expenditure | Total expense | Continuous, Ratio | It is useful |
| Diphtheria | Disease | Continuous, Ratio | It is useful |
| HIV\_AIDS | Disease | Continuous, Ratio | It is useful |
| GDP | Gross Democratic Product | Continuous, Ratio | It is useful |
| Population | Population Rate | Discrete, Count | It is useful |
| Thinness | State of body | Continuous, Ratio | It is useful |
| thinness\_yr | State of body | Continuous, Ratio | It is useful |
| Income\_composition | Income Rate | Continuous, Ratio | It is useful |
| Schooling | Education at school | Continuous, Ratio | It is useful |

A screenshot of a cell phone

Description automatically generated