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: Nukala Ayyappa Bharthwaj

Batch Id: DSWDMCOS 21012022

Topic: Lasso Ridge Regression

1. Business Problem

1.1. Objective

1.2. Constraints (if any)

2. Work on each feature of the dataset to create a data dictionary as displayed in the below image:

Name of Feature	Description	Type	Relevance
ID	Customer ID	Quantitative, Nominal	Irrelevant, ID does not provide useful information

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:

3. Data Pre-processing

3.1 Data Cleaning, Feature Engineering, etc.

3.2 Outlier Imputation

4. Exploratory Data Analysis (EDA):

4.1. Summary

4.2. Univariate analysis

4.3. Bivariate analysis

5. Model Building

5.1 Build the model on the scaled data (try multiple options)

5.2 Perform Lasso and Ridge Regression Algorithm

5.3 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

6. **Briefly explain the model output in the documentation. 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.

Sol:

Business Objective: To predict the profits of the company with other factors by using Lasso-Ridge regression model.

Constraints: Lack of analysis of the company data.

Data Types: All the given data is in numeric format except the states column and the complete data can be used for the analysis.

Data Cleaning: Since the states column is in Non-Numeric format the same is converted to numeric data for doing the analysis.

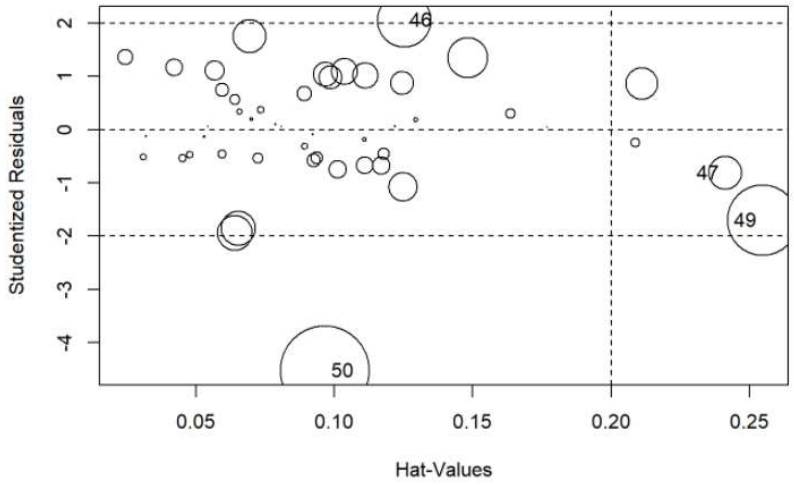
Exploratory Data Analysis: the normal distribution of the all variable of the given data is checked by using box plot, histograms and scatter plots. After observing all the plots the regression analysis on the is made.

Multiple Linear Regression: after observing the scatter plot of all the variables in the given data a multiple linear regression model is made by taking the output variable as profit of the company.

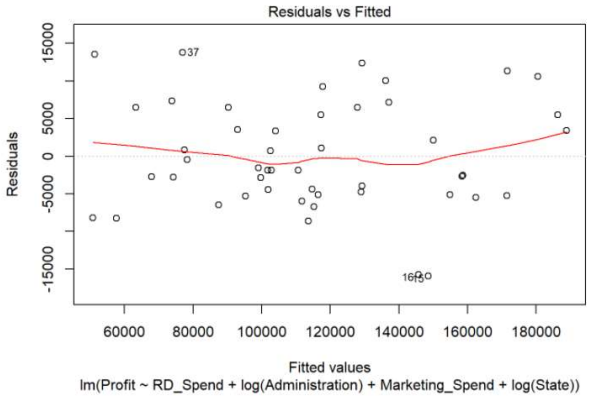
The R^2 value for the basic Multi-linear regression model without applying any transformations on the data is 0.9496.

Influential plot is made to know the observations in the given data which is effecting the accuracy

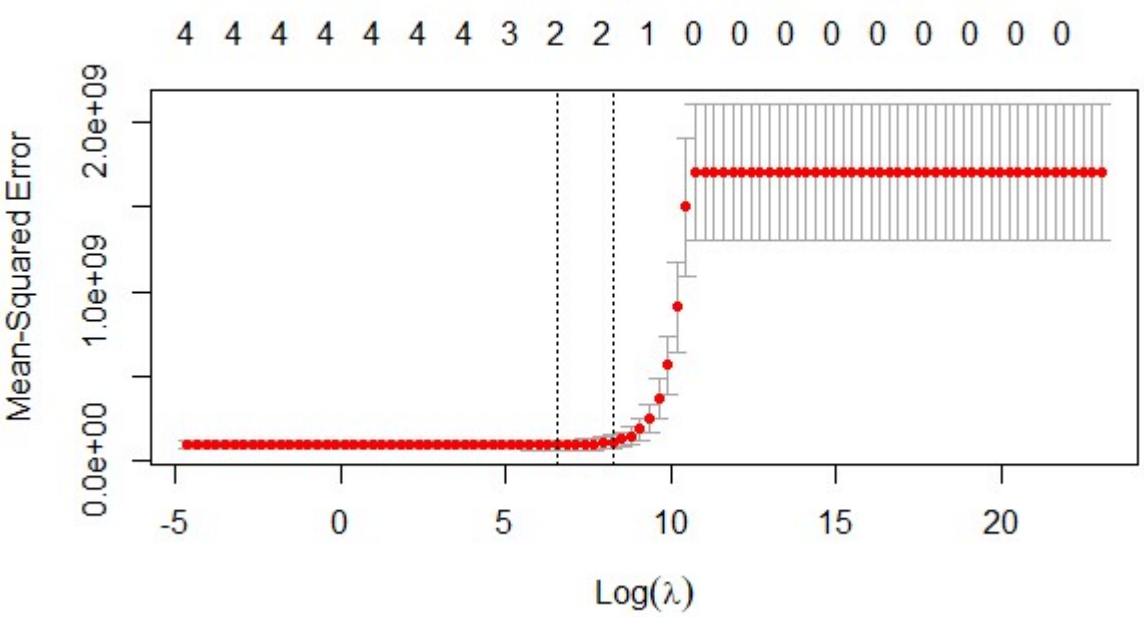
of the complete analysis, the influential observations in the given data are 49,50 and that are found by using the following influential plot.

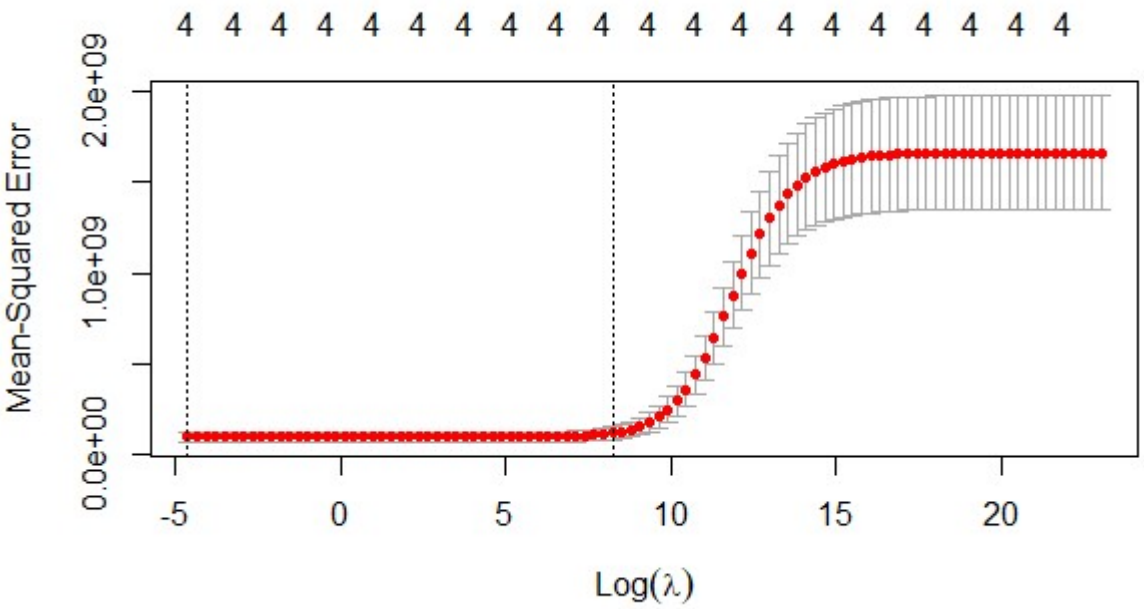


After removing the influential observation in the given data further multiple regression methods are applied and R^2 values are increased further.
After applying Logarithmic transformations the R^2 value is 0.9451 and for the exponential transformations R^2 value is 0.9557 and for the quadratic model is 0.9567.
So, finally quadratic model is giving the best results so the same is used for building the final model to predict the profits of the company.
The following residual plot is done to know the accuracy of the model.



Lasso-Ridge Regression: After applying the multi linear regression models then I applied Lasso and Ridge model for checking the further accuracy in the model. The R^2 value for the Lasso model is 0.9321 and for the Ridge model it is increased slightly and R^2 value is 0.9624. so finally Ridge model can be used for the prediction. The probability graphs for the Lasso and Ridge models are as follows:





	R.D.Spend	Administration	Marketing.Spend	State	Profit
1	165349.20	136897.80	471784.10	New York	192261.83
2	162597.70	151377.59	443898.53	California	191792.06
3	153441.51	101145.55	407934.54	Florida	191050.39
4	144372.41	118671.85	383199.62	New York	182901.99
5	142107.34	91391.77	366168.42	Florida	166187.94
6	131876.90	99814.71	362861.36	New York	156991.12
7	134615.46	147198.87	127716.82	California	156122.51
8	130298.13	145530.06	323876.68	Florida	155752.60
9	120542.52	148718.95	311613.29	New York	152211.77
10	123334.88	108679.17	304981.62	California	149759.96

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.

Sol:

Business Objective: To predict the Price of the car with other factors by using Lasso-Ridge regression model.

Constraints: Lack of analysis of the car sales data of the company

Data Types: The first column in the given data is I'd which is not useful for doing the analysis and the remaining data is used for doing the analysis.

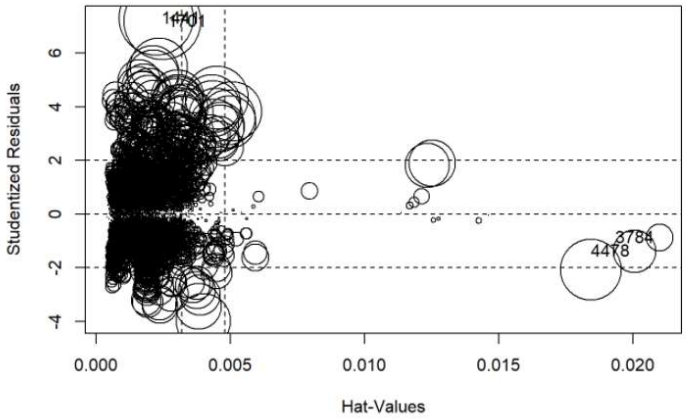
Data Cleaning: Since some of the columns in the given data is no-numeric the same is converted into numeric data so that they can be used for doing the analysis.

Exploratory Data Analysis: the normal distribution of the all variable of the given data is checked by using box plot, histograms and scatter plots. After observing all the plots the regression analysis on the is made.

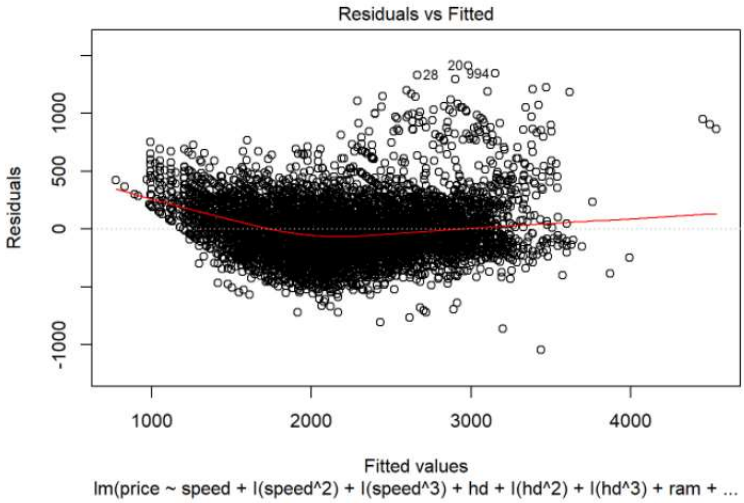
Multiple Linear Regression: after observing the scatter plot of all the variables in the given data a multiple linear regression model is made by taking the output variable as price of the computer

part.

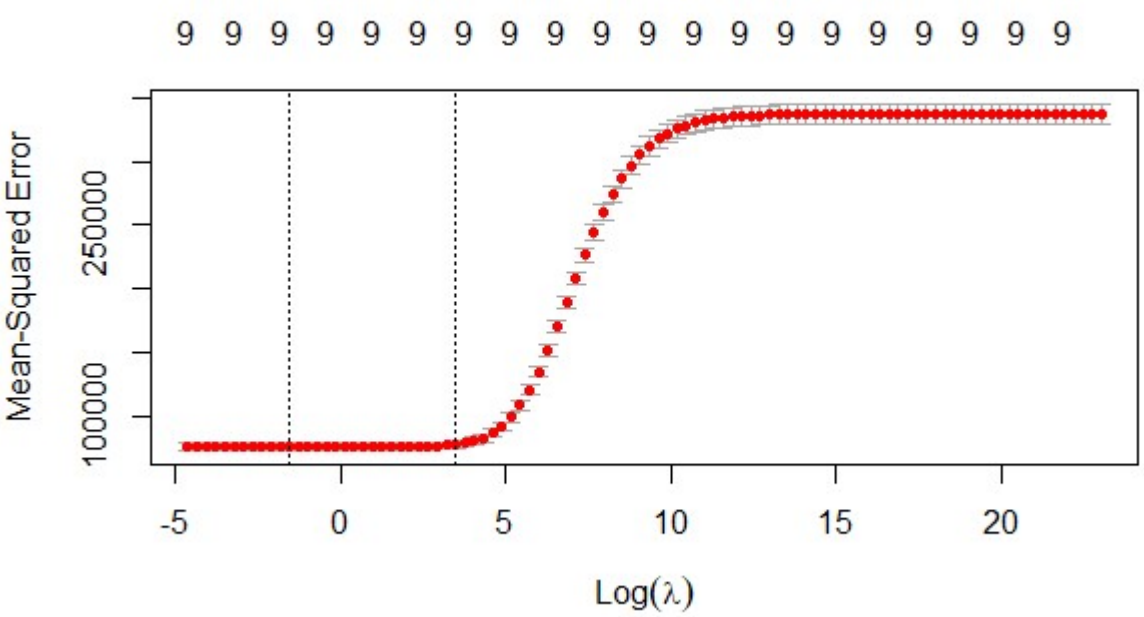
The R^2 value for the basic Multi-linear regression model without applying any transformations on the data is 0.7752.
Influential plot is made to know the observations in the given data which is effecting the accuracy of the complete analysis, the influential observations in the given data are 1441,1701 and that are found by using the following influential plot.

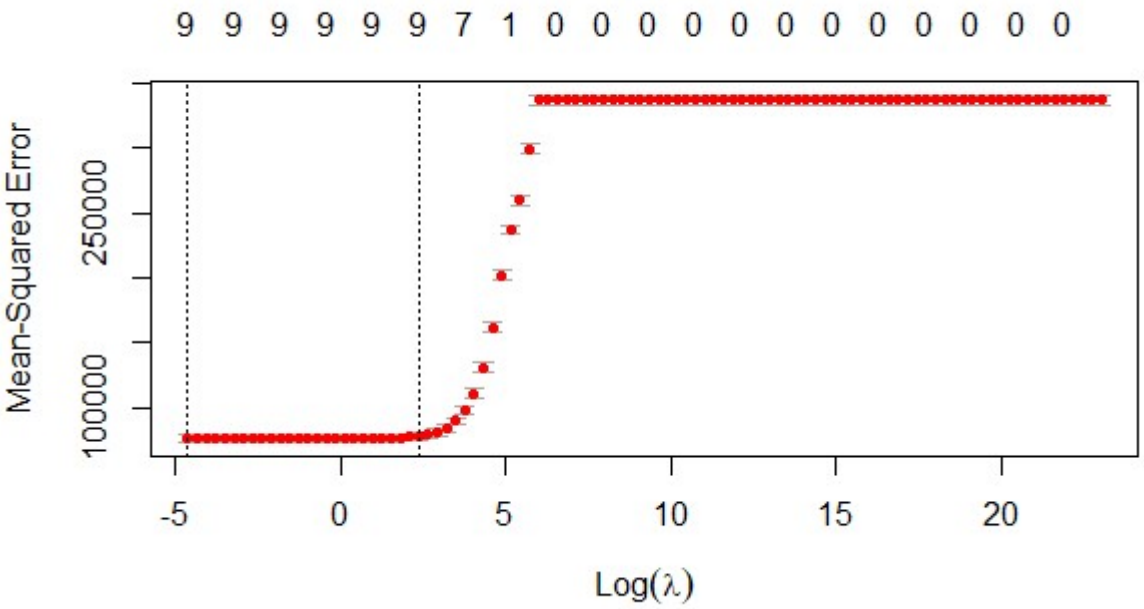


After removing the influential observation in the given data further multiple regression methods are applied and R^2 values are increased further.
After applying Logarithmic transformations the R^2 value is 0.7441and for the exponential transformations R^2 value is 0.7833 and for the quadratic model is 0.8049.
So, finally quadratic model is giving the best results so the same is used for building the final model to predict the profits of the company.
The following residual plot is done to know the accuracy of the model.



Lasso-Ridge Regression: After applying the multi linear regression models then I applied Lasso and Ridge model for checking the further accuracy in the model. The R^2 value for the Lasso model is 0.7666 and for the Ridge model it is increased slightly and R^2 value is 0.8051. so finally Ridge model can be used for the prediction. The probability graphs for the Lasso and Ridge models are as follows:





	X	price	speed	hd	ram	screen	cd	multi	premium	ads	trend
1	1	1499	25	80	4	14	no	no	yes	94	1
2	2	1795	33	85	2	14	no	no	yes	94	1
3	3	1595	25	170	4	15	no	no	yes	94	1
4	4	1849	25	170	8	14	no	no	no	94	1
5	5	3295	33	340	16	14	no	no	yes	94	1
6	6	3695	66	340	16	14	no	no	yes	94	1
7	7	1720	25	170	4	14	yes	no	yes	94	1
8	8	1995	50	85	2	14	no	no	yes	94	1
9	9	2225	50	210	8	14	no	no	yes	94	1
10	10	2575	50	210	4	15	no	no	yes	94	1
11	11	2195	33	170	8	15	no	no	yes	94	1

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

Sol:

Business Objective: To predict the Price of the computer parts with other factors by using Lasso-Ridge regression model.

Constraints: Lack of analysis of the sales data of the company

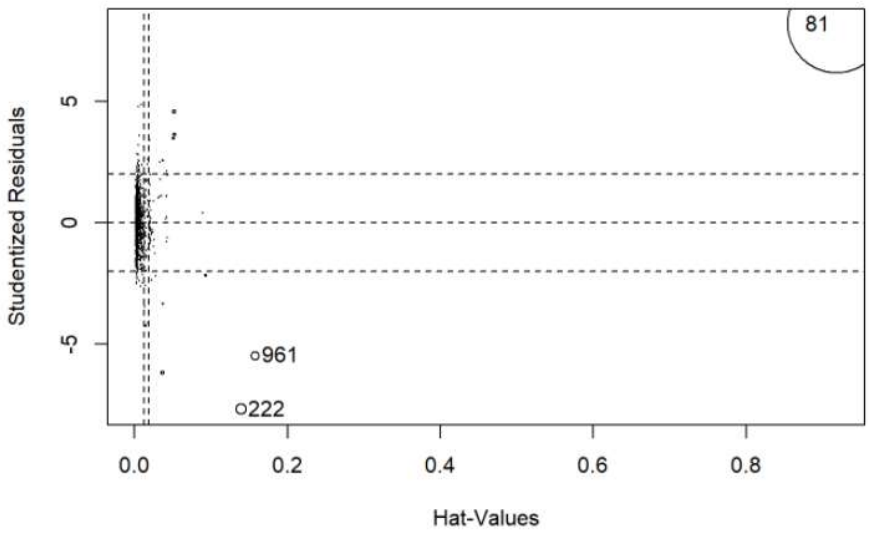
Data Types: The first column in the given data is I'd which is not useful for doing the analysis and the remaining data is used for doing the analysis.

Data Cleaning: Since some of the columns in the given data is non-numeric the same is converted into numeric data so that they can be used for doing the analysis.

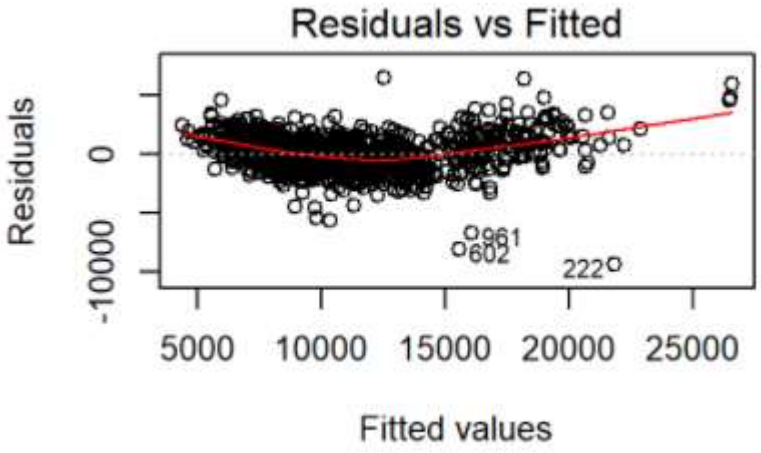
Exploratory Data Analysis: the normal distribution of the all variable of the given data is checked by using box plot, histograms and scatter plots. After observing all the plots the regression analysis on the data is made.

Multiple Linear Regression: after observing the scatter plot of all the variables in the given data a multiple linear regression model is made by taking the output variable as price of the computer.

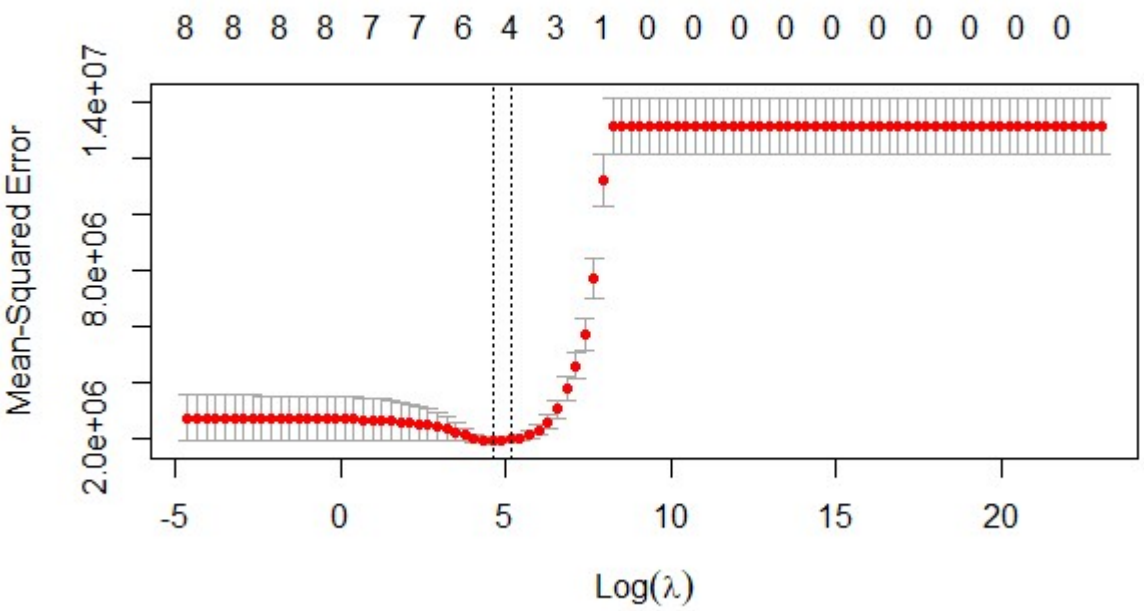
The R^2 value for the basic Multi-linear regression model without applying any transformations on the data is 0.863.
Influential plot is made to know the observations in the given data which is effecting the accuracy of the complete analysis, the influential observations in the given data are 1441,1701 and that are found by using the following influential plot.

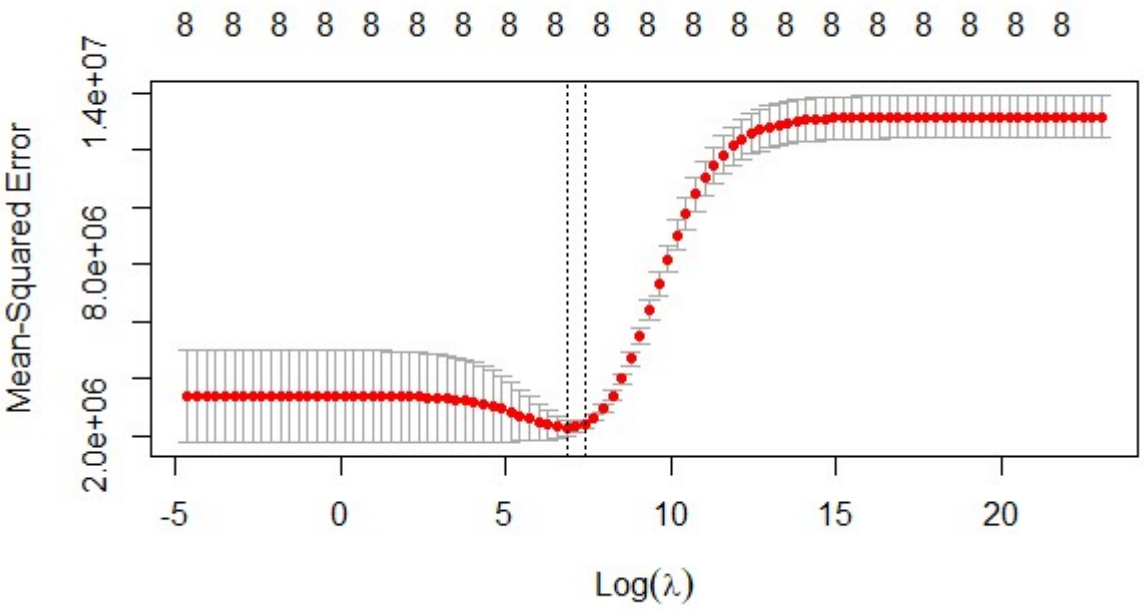


After observing the influential plot it is noted that 81st observation is effecting the complete analysis so the same observation is removed and then the basic regression model is done then the R^2 values is observed as 0.9291 so the same model is used for the future prediction in the data . The residual plot for the model is as follows.



Lasso-Ridge Regression: After applying the multi linear regression models then I applied Lasso and Ridge model for checking the further accuracy in the model. The R^2 value for the Lasso model is 0.8012 and for the Ridge model it is increased slightly and R^2 value is 0.9326. so finally Ridge model can be used for the prediction. The probability graphs for the Lasso and Ridge models are as follows:





	Id	Model	Price	Age_08_04	Mfg_Month	Mfg_Year	KM	Fuel_Type	HP	Met_Color	Color	Automatic	cc	Doors	Cylinder
1	1	TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors	13500	23	10	2002	46986	Diesel	90	1	Blue	0	2000	3	4
2	2	TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors	13750	23	10	2002	72937	Diesel	90	1	Silver	0	2000	3	4
3	3	TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors	13950	24	9	2002	41711	Diesel	90	1	Blue	0	2000	3	4
4	4	TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors	14950	26	7	2002	48000	Diesel	90	0	Black	0	2000	3	4
5	5	TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors	13750	30	3	2002	38500	Diesel	90	0	Black	0	2000	3	4
6	6	TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors	12950	32	1	2002	61000	Diesel	90	0	White	0	2000	3	4
7	7	TOYOTA Corolla 2.0 D4D 90 3DR TERRA 2/3-Doors	16900	27	6	2002	94612	Diesel	90	1	Grey	0	2000	3	4
8	8	TOYOTA Corolla 2.0 D4D 90 3DR TERRA 2/3-Doors	18600	30	3	2002	75889	Diesel	90	1	Grey	0	2000	3	4
9	9	TOYOTA Corolla 1800 T SPORT VVT I 2/3-Doors	21500	27	6	2002	19700	Petrol	192	0	Red	0	1800	3	4
10	10	TOYOTA Corolla 1.9 D HATCHB TERRA 2/3-Doors	12950	23	10	2002	71138	Diesel	69	0	Blue	0	1900	3	4
11	11	TOYOTA Corolla 1.8 VVTL-i T-Sport 3-Drs 2/3-Doors	20950	25	8	2002	31461	Petrol	192	0	Silver	0	1800	3	4
12	12	TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT BNS 2/3-Doors	19950	22	11	2002	43610	Petrol	192	0	Red	0	1800	3	4
13	13	TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT 2/3-Doors	19600	25	8	2002	32189	Petrol	192	0	Red	0	1800	3	4
14	14	TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT 2/3-Doors	21500	31	2	2002	23000	Petrol	192	1	Black	0	1800	3	4
15	15	TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT 2/3-Doors	22500	32	1	2002	34131	Petrol	192	1	Grey	0	1800	3	4
16	16	TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT 2/3-Doors	22000	28	5	2002	18739	Petrol	192	0	Grey	0	1800	3	4
17	17	TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT 2/3-Doors	22750	30	3	2002	34000	Petrol	192	1	Grey	0	1800	3	4

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: -

Business Objective: To predict the Life Expectancy of the country with other factors by using Lasso-Ridge regression model.

Constraints: Lack of analysis of the life Expectancy data of different countries.

Data Types: All the data given is used for doing the analysis on the data.

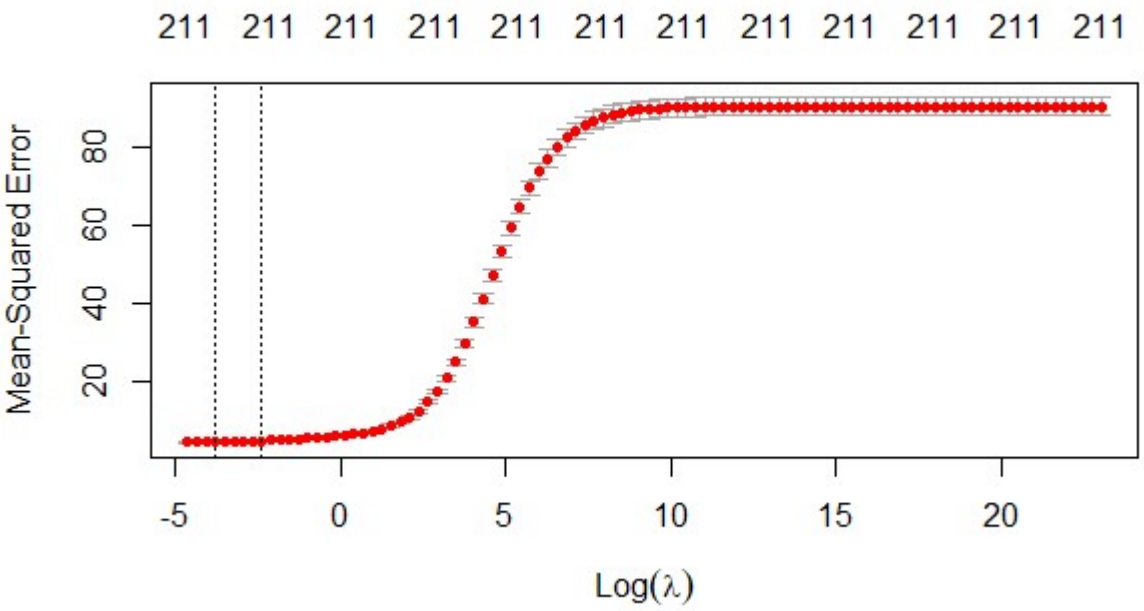
Data Cleaning: Since some of the columns in the given data is non-numeric the same is converted into numeric data so that they can be used for doing the analysis.

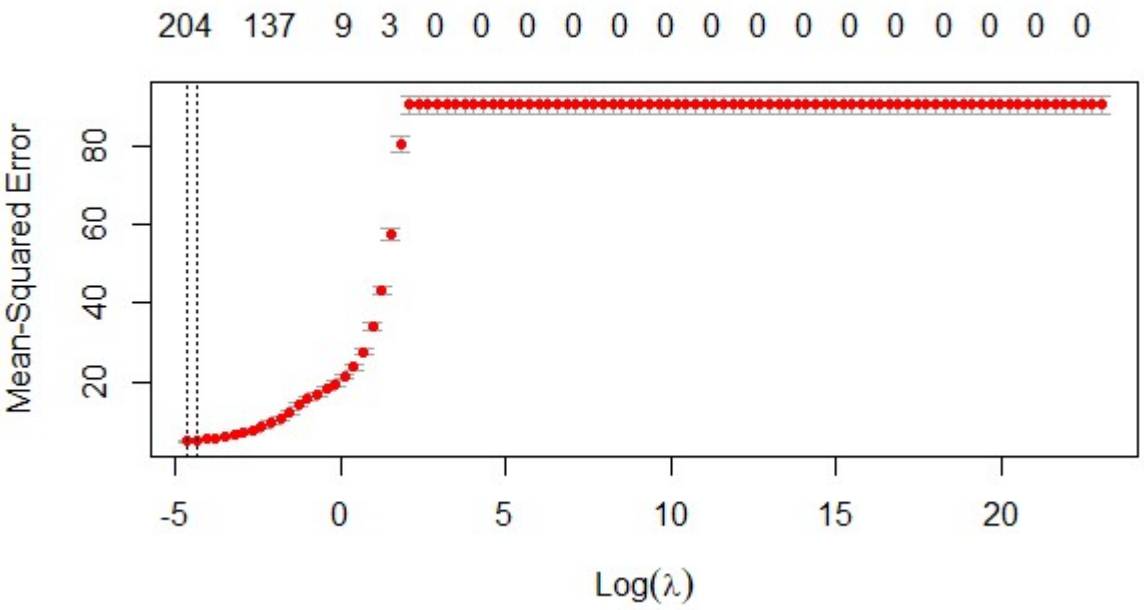
Exploratory Data Analysis: the normal distribution of the all variable of the given data is checked by using box plot, histograms and scatter plots. After observing all the plots the regression analysis on the data is made.

Multiple Linear Regression: after observing the scatter plot of all the variables in the given data a multiple linear regression model is made by taking the output variable as Life Expectancy of the country.

The R^2 value for the basic Multi-linear regression model without applying any transformations on the data is 0.98.

Lasso-Ridge Regression: After applying the linear regression models then I applied Lasso and Ridge model for checking the further accuracy in the model. The R^2 value for the Lasso model is 0.9608 and for the Ridge model it is increased slightly and R^2 value is 0.9540. So finally Multi linear regression can be used for the prediction. The probability graphs for the Lasso and Ridge models are as follows:





Country	Year	Status	Life_expectancy	Adult_Mortality	infant_deaths	Alcohol	percentage_expendi	Hepatitis_B	Measles	BMI	under_five_deaths	Polio
Afghanistan	2015	Developing	65	263	62	0.01	71.27962362	65	1154	19.1	83	6
Afghanistan	2014	Developing	59.9	271	64	0.01	73.52358168	62	492	18.6	86	58
Afghanistan	2013	Developing	59.9	268	66	0.01	73.21924272	64	430	18.1	89	62
Afghanistan	2012	Developing	59.5	272	69	0.01	78.1842153	67	2787	17.6	93	67
Afghanistan	2011	Developing	59.2	275	71	0.01	7.097108703	68	3013	17.2	97	68
Afghanistan	2010	Developing	58.8	279	74	0.01	79.67936736	66	1989	16.7	102	66
Afghanistan	2009	Developing	58.6	281	77	0.01	56.76221682	63	2861	16.2	106	63
Afghanistan	2008	Developing	58.1	287	80	0.03	25.87392536	64	1599	15.7	110	64
Afghanistan	2007	Developing	57.5	295	82	0.02	10.91015598	63	1141	15.2	113	63
Afghanistan	2006	Developing	57.3	295	84	0.03	17.17151751	64	1990	14.7	116	58
Afghanistan	2005	Developing	57.3	291	85	0.02	1.388647732	66	1296	14.2	118	8
Afghanistan	2004	Developing	57	293	87	0.02	15.29606643	67	466	13.8	120	5
Afghanistan	2003	Developing	56.7	295	87	0.01	11.08905273	65	798	13.4	122	41
Afghanistan	2002	Developing	56.2	3	88	0.01	16.88735091	64	2486	13	122	36
Afghanistan	2001	Developing	55.3	316	88	0.01	10.5747282	63	8762	12.6	122	35
Afghanistan	2000	Developing	54.8	321	88	0.01	10.42496	62	6532	12.2	122	24
Albania	2015	Developing	77.8	74	0	4.6	364.9752287	99	0	58	0	99
Albania	2014	Developing	77.5	8	0	4.51	428.7490668	98	0	57.2	1	98
Albania	2013	Developing	77.2	84	0	4.76	430.8769785	99	0	56.5	1	99
Albania	2012	Developing	76.9	86	0	5.14	412.4433563	99	9	55.8	1	99
Albania	2011	Developing	76.6	88	0	5.37	437.0621	99	28	55.1	1	99
Albania	2010	Developing	76.2	91	1	5.28	41.82275719	99	10	54.3	1	99
Albania	2009	Developing	76.1	91	1	5.79	348.0559517	98	0	53.5	1	98
Albania	2008	Developing	75.3	1	1	5.61	36.62206845	99	0	52.6	1	99