EAS506 - Statistical Data Mining I Homework 1 – Question 2

Aniket Kamlender Maheshwari 09/17/21

Abstract

This report summaries the steps taken to perform linear regression on a clean and transformed cereal dataset.

Content

1	Introduction4
2	Method4
2.1	Initialization Steps4
2.2	Plotting a Co-Relation Plot5
2.3	Create Linear Regression Models6
	2.3.1 Fitting a linear regression model with all the features6
	2.3.2 Fitting linear regression model with updated features6
	2.3.3 Fitting linear regression model with features that are negatively correlated to 'rating'7
	2.3.4 Fitting linear regression model with features that are positively correlated to 'rating'7
	2.3.5 Interactions with all features8
	2.3.6 using ":" with most significantly impacting features to
	our response8

1 Introduction

The Cereal data frame has 73 rows and 14 columns. The data come from the 1993 ASA Statistical Graphics Exposition and are taken from the mandatory F&DA food label. The data have been normalized here to a portion of one American cup.

This report summarizes the process undertaken to build a linear regression model on cereal dataset that was cleaned previously.

2 Method

2.1 Initialization Steps

- Clear the memory
- Install and load all required libraries.
- Analyze the dataset.

2.2 Plotting a Co-relation plot

COLLEGUOU PIOL

	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	weight	rating	_ 1
calories	1.00		0.50	0.30		0.22	0.54		0.27	0.73	-0.64	[1
protein	0.11	1.00	0.22	0.06	0.55		-0.25	0.57	0.06	0.25	0.44	- 0.8
fat	0.50	0.22	1.00	0.00	0.09	-0.32	0.27	0.24	-0.04	0.22	-0.40	- 0.6
sodium	0.30	0.06	0.00	1.00	-0.10	0.35	0.01	-0.04	0.33	0.31	-0.40	- 0.4
fiber	-0.14	0.55	0.09	-0.10	1.00	-0.32	-0.07	0.92	-0.03	0.32	0.51	- 0.2
carbo	0.22	-0.01	-0.32	0.35	-0.32	1.00	-0.51	-0.31	0.26	0.14	0.16	- 0
sugars	0.54	-0.25	0.27		-0.07	-0.51	1.00	0.08	0.07	0.46	-0.76	0.2
potass	0.06	0.57	0.24	-0.04	0.92	-0.31	0.08	1.00	0.01	0.46	0.31	0.4
vitamins	0.27	0.06	-0.04	0.33	-0.03	0.26	0.07	0.01	1.00	0.32	-0.22	0.6
weight	0.73	0.25	0.22	0.31	0.32	0.14	0.46	0.46	0.32	1.00	-0.32	0.8
rating	-0.64	0.44	-0.40	-0.40	0.51	0.16	-0.76	0.31	-0.22	-0.32	1.00	-0.0

Q.1) Which predictors appear to have a significant relationship to the response 'rating'?

Ans 1) \sim 'calories' and 'sugars' are highly negatively co-related to our response 'rating' feature.

 $[\]sim$ 'protein', 'fiber' and 'potass' are positively co-related to our response 'rating' feature.

2.3 Create Linear Regression Models

- A number of Multiple-Linear-Regression Models were created to predict the rating.
 - ~ Fitting a linear regression model with all the features.
 - ~ Fitting linear regression model with updated features.
 - ~ Fitting linear regression model with features that are negatively correlated to 'rating' (our response)
 - ~ Fitting linear regression model with features that are positively correlated to 'rating' (our response)
 - ~ Interactions with all features.
 - ~ using ":" with most significantly impacting features to our response.

2.3.1 Fitting a linear regression model with all the features.

- Used all the numeric and integer features in the data frame to fit a model.
- Residual standard error: 3.058e-07 on 62 degrees of freedom.
- I noted that that 'weight' feature is not important, so we'll fit next model without that feature.

2.3.2 Fitting a linear regression model with all the features.

- Used all features except the weight feature.
- Residual standard error: 3.064e-07 on 63 degrees of freedom

2.3.3 Fitting linear regression model with features that are positively correlated to 'rating' (our response)

- Used only features that are co-related to 'ratings' feature negatively.
- features used were calories + sugars + fat +sodium
- Residual standard error: 6.25 on 68 degrees of freedom
- This model suggests that calories and fat is not important feature which is surprising because in co-relation plot 'calories' had -0.64 co relation with 'rating' feature.

2.3.4 Fitting linear regression model with features that are negatively correlated to 'rating' (our response)

- Used only features that are co-related to 'ratings' feature positively.
- features used were : protein + fiber + potass
- Residual standard error: 9.358 on 69 degrees of freedom

Q.2) What does the coefficient variable for "sugar" suggest?

Ans 2) Co-efficient variable for "sugar" were:

Eastimate	Std. Error	t-value	Pr(> t)
-7.249e-01	3.397e-08	-2.134e+07	<2e-16 ***

Signif. codes: 0 '*** 0.001 '** 0.01 '* 0.05 '. 0.1 ' 1

This suggests that sugar is an important feature and is highly co-related to our response feature 'rating'.

Q.3) Use the * and : symbols to fit models with interactions. Are there any interactions that are significant?

Ans 3)

2.3.5 Interactions with all features.

- Used all the features.
- Residual standard error: 1.347 on 9 degrees of freedom
- Significant relations:
 - a) fat:sodium:carbo:sugars: 0.000666 ***
 - b) calories:protein:fat:sodium:sugars: 0.000997 ***
 - c) calories:fat:sodium:carbo:sugars : 0.000665 ***

Signif. codes: 0 "*** 0.001 "** 0.01 "* 0.05 ". 0.1 " 1

2.3.6 using ":" with most significantly impacting features to our response.

- Used ":" with all positively and negatively impacting features with corelation to our target feature.
- positive features were: calories:sugars
- negative features were: protein:carbo:potass
- Residual standard error: 6.671 on 70 degrees of freedom
- Significant relations:
 - a) calories: sugars : < 2e-16
- b) protein:carbo:potass: < 2e-16
