CSC6780 - Data Science; Assignment 8

Jesse Annan || ID: 002708111

April 6, 2023

Question 1

Find data in the attached notebook.

(i) The sums of squared errors

$$\begin{aligned}\textbf{Error[Model 1 Prediction]} &= \frac{1}{2}\sum_{1}^{30}(Target[i] - Model1Prediction[i])^2 \\ &= 16750 \\ \textbf{Error[Model 2 Prediction]} &= \frac{1}{2}\sum_{1}^{30}(Target[i] - Model2Prediction[i])^2 \\ &= 47369\end{aligned}$$

(ii) The R^2 measure

$$\begin{aligned}R^2\ \textbf{[Model 1 Prediction]} &= 1 - \frac{\frac{1}{2}\sum_{1}^{30}(Target[i] - Model1Prediction[i])^2}{\frac{1}{2}\sum_{1}^{30}(Target[i] - Target\_mean)^2} \\ &\approx 0.92844 \\ R^2\ \textbf{[Model 2 Prediction]} &= 1 - \frac{\frac{1}{2}\sum_{1}^{30}(Target[i] - Model2Prediction[i])^2}{\frac{1}{2}\sum_{1}^{30}(Target[i] - Target\_mean)^2} \\ &\approx 0.79762\end{aligned}$$

Question 2

Model:
`HEATING LOAD = -26.030 + 0.0497 x SURFACE AREA + 4.942 x HEIGHT - 0.090 x ROOF AREA + 20.523 x GLAZING AREA`

$$\texttt{ID1:}\quad \texttt{Heating Load} = -26.030 + 0.0497 * 784.0 + 4.942 * 3.5 - 0.090 * 220.5 + 20.523 * 0.25$$
$$15.51755$$

$$\texttt{ID2:}\quad \texttt{Heating Load} = -26.030 + 0.0497 * 710.5 + 4.942 * 3.0 - 0.090 * 210.5 + 20.523 * 0.10$$
$$7.21515$$

$$\texttt{ID3:}\quad \texttt{Heating Load} = -26.030 + 0.0497 * 563.5 + 4.942 * 7.0 - 0.090 * 122.5 + 20.523 * 0.40$$
$$33.75415$$

$$\texttt{ID4:}\quad \texttt{Heating Load} = -26.030 + 0.0497 * 637.0 + 4.942 * 6.0 - 0.090 * 147.0 + 20.523 * 0.60$$
$$34.3647$$

Question 3

`Imputation`

age[3] = 32.7 and socio_economic_band[5] = 'a'

`Normalized Features`

ID	AGE	SOCIO ECONOMIC BAND	SHOP FREQUENCY	SHOP VALUE
1	-0.11111	a	-0.34615	0.42127
2	0.68889	b	-0.46154	-0.07559
3	-1.00000	c	1.23077	-0.95490
4	-0.34667	b	-0.56538	0.76890
5	0.95556	a	-0.75000	-0.06513

$$\texttt{Logistic(X)} = \frac{1}{1+\exp(-X)}$$

$$ID1\_wd = 0.6679 - 0.5795 * -0.11111 + 0 + 2.0499 * -0.34615 + 3.4091 * 0.42127$$

$$ID1 = \texttt{Logistic(ID1\_wd)} \approx 0.811357$$

$$ID2\_wd = 0.6679 - 0.5795 * 0.68889 - 0.1981 + 2.0499 * -0.46154 + 3.4091 * -0.07559$$

$$ID2 = \texttt{Logistic(ID2\_wd)} \approx 0.243571$$

$$ID3\_wd = 0.6679 - 0.5795 - 1 - 0.2318 + 2.0499 * 1.23077 + 3.4091 * -0.95490$$

$$ID3 = \texttt{Logistic(ID3\_wd)} \approx 0.570335$$

$$ID4\_wd = 0.6679 - 0.5795 * -0.34667 - 0.1981 + 2.0499 * -0.56538 + 3.4091 * 0.76890$$

$$ID4 = \texttt{Logistic(ID4\_wd)} \approx 0.894065$$

$$ID5\_wd = 0.6679 - 0.5795 * 0.95556 + 0 + 2.0499 * -0.75000 + 3.4091 * -0.06513$$

$$ID5 = \texttt{Logistic(ID5\_wd)} \approx 0.161744$$

Question 4

(a) Yes, with a reasonable choice of k. Similarity-based predictive modeling approach (KNN) will be a good choice for this data set.

(b)

$$
\begin{aligned}
\texttt{Logistic(X)} &= \frac{1}{1+\exp(-X)} \\
ID1\_wd &= -0.848 * (1) + 1.545 * (0.50) - 1.942 * (0.75) + 1.973 * (0.50^2) \\
&\quad + 2.495 * (0.75^2) + 0.104 * (0.50^3) + 0.095 * (0.75^3) + 3.009 * (0.50 * 0.75) \\
ID1 &= \texttt{Logistic(ID1\_wd)} \approx 0.824356 \\
ID2\_wd &= -0.848 * (1) + 1.545 * (0.10) - 1.942 * (0.75) + 1.973 * (0.10^2) \\
&\quad + 2.495 * (0.75^2) + 0.104 * (0.10^3) + 0.095 * (0.75^3) + 3.009 * (0.10 * 0.75) \\
ID2 &= \texttt{Logistic(ID1\_wd)} \approx 0.386754 \\
ID3\_wd &= -0.848 * (1) + 1.545 * (-0.47) - 1.942 * (-0.39) + 1.973 * (-0.47^2) \\
&\quad + 2.495 * (-0.39^2) + 0.104 * (-0.47^3) + 0.095 * (-0.39^3) + 3.009 * (-0.47 * -0.39) \\
ID3 &= \texttt{Logistic(ID1\_wd)} \approx 0.630339 \\
ID4\_wd &= -0.848 * (1) + 1.545 * (-0.47) - 1.942 * (0.18) + 1.973 * (-0.47^2) \\
&\quad + 2.495 * (0.18^2) + 0.104 * (-0.47^3) + 0.095 * (0.18^3) + 3.009 * (-0.47 * 0.18) \\
ID4 &= \texttt{Logistic(ID1\_wd)} \approx 0.158179
\end{aligned}
$$