DNSC 6314
Spring 2022
Assignment 1
Section 10 Group 3: Harshit Aggarwal, Terrell Staten, David Goforth & Claudio Escudero

Target and feature encoding are powerful feature engineering techniques, but do give rise to overfitting in many cases. Questions 1 – 3 below use Datasets 1 and 2 (see Worksheet) and relate to methods for target and feature encoding that mitigate overfitting.

1. **(1 pt.)** Smoothing approaches are used in target and feature encoding to prevent the "Law of Small Numbers" from leading to overfitting in machine learning (ML) models. A basic technique for smoothing balances the overall mean value of the target feature, y, with it's in-category mean as follows in Equation 1:

$$X_{encode} = \lambda \cdot \bar{y} + (1 - \lambda) \cdot \bar{y}_{level} \qquad (1)$$

   where λ is a hyperparameter in [0,1], but can be set to 0.5 by default, $\bar{y}$ is the overall value of the target or feature for encoding, and $\bar{y}_{level}$ is the mean of the target or feature for encoding within the current categorical level of x. Apply Eq. 1 to Dataset 1 to target encode x_1 and feature encode x_1 by x_2. Fill in your answers in Dataset 1 in the x1_TE and x1_x2_FE columns.

2. **(1 pt.)** Leave-one-out target encoding is an effective, but computationally intensive, encoding approach in which a feature takes on the value of the mean of the target for every row in the category, except the current row. Target encode x_1 in Dataset 1 using the leave-one-out method in the x1_LOOTE column of Dataset 1.

3. **(1 pt.)** Training exercises are great for educational purposes, but trivial and irrelevant if they cannot be used in production environments, where target values are often future unknown quantities. Score Dataset 2 (see Worksheet) with the x1_TE and x1_x2_FE using the transformations learned in Questions 1 and 2.

---

4. **(1 pt.)** Imputation of missing values is often necessary to use certain ML training algorithms. Given the drawbacks of mean imputation, creating a new missing marker feature to go along with the imputed feature allows modelers to understand if missingness is meaningful in training data, and presents the options of using the imputed feature, using the missing marker instead, or using both for learning purposes. Mean-impute x_1 in Dataset 3 (see Worksheet) in the x1_IMP column, and create a binary missing marker in the x1_MISSING column with values of 1 and 0.

Feature extraction, e.g. by principal components analysis (PCA), is a common feature engineering approach. Answer questions 5 – 7 relating to feature extraction.

5. **(1 pt.)** What is the obvious drawback to feature extraction techniques that do not apply a penalized approach to generate so-called sparse extracted features?

   The obvious drawback to feature extraction techniques that do not apply a penalized approach is that the new generated features are uninterpretable. For example if a large number of features are combined using PCA the new feature will have numbers which will look like random numbers to the human eye.

6. **(1 pt.)** Given the two correlation matrices below, is Dataset A or B more appropriate to apply feature extraction? Is Dataset A or B more appropriate to apply feature selection? Write your answer in the space below the matrices.

	x_1	x_2	x_3	y
x_1	1	-0.20	0.80	0.03
x_2	-0.20	1	-0.07	-0.11
x_3	0.80	-0.07	1	0.7
y	0.03	-0.11	0.7	1

Dataset A

	x_1	x_2	x_3	y
x_1	1	0.52	-0.68	-0.66
x_2	0.52	1	-0.90	0.71
x_3	-0.68	-0.90	1	0.88
y	-0.66	0.71	0.88	1

Dataset B

Comparing dataset A & B, we can see that there is a high correlation between the input variables (x1, x2, x3) and the output variables (y) in dataset B as compared to correlation in dataset A. Given the high correlation conducting feature extraction using techniques such as PCA would yield a better result on dataset B, while feature selection would suit dataset A.

7. **(1 pt.)** Given a training dataset with 2 input columns, x_1 and x_2, and the output of your favorite ML package for PCA eigenvectors [0.743, -0.345], calculate the score for the non-centered first principal component in the PCA1 column of Dataset 4 (see Worksheet).

---

8. **(1 pt.)** Does one-hot encoding tend to improve or worsen problems associated with the curse of dimensionality? Write your answer below.

One-hot encoding tends to worsen the problems associated with the curse of dimensionality. This is because one-hot encoding can create massive dimensions. This problem is most obvious when you one-hot encode a categorical variable with many levels. These huge dimensions will cause the curse of dimensionality.

---

9. **(1 pt.)** Given Equations 2 and 3 below for non-linear additive noise models (NANMs), the correlation matrix and the residual correlation matrix below, draw the causal graph for x_1, x_2, and y in the space below the matrices.

$$x_2 = g(x_1) + \epsilon_{x1} \qquad (2)$$

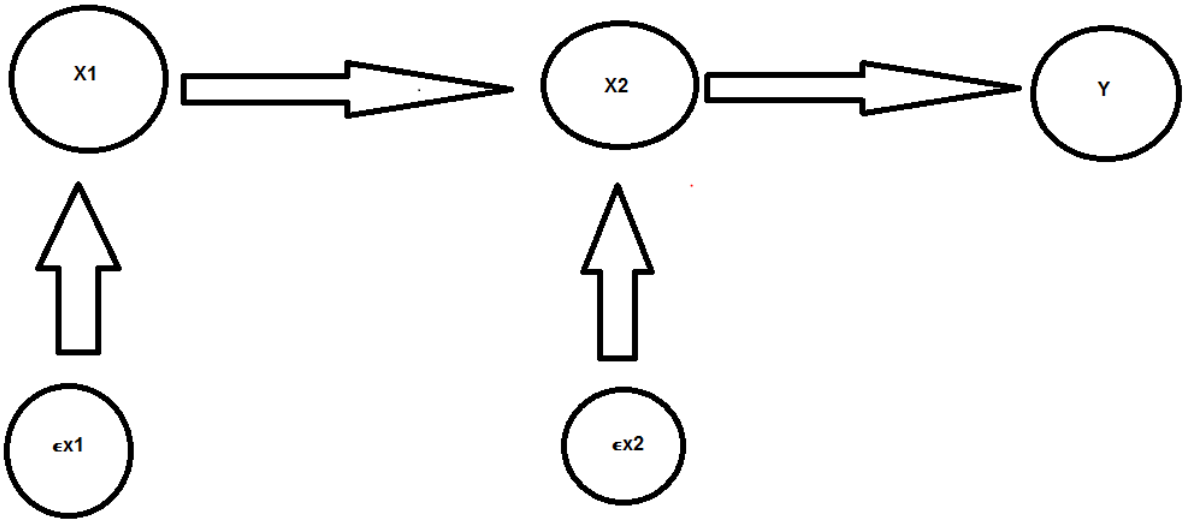
$$y = g(x_2) + \epsilon_{x2} \qquad (3)$$

	x_1	x_2	y
x_1	1	-0.87	0.07
x_2	-0.87	1	0.76
y	0.07	0.76	1

Correlation Matrix

	ϵ_{x1}	ϵ_{x2}	x_1	x_2
ϵ_{x1}	1	0.32	-0.06	0.25
ϵ_{x2}	0.32	1	-0.17	0.04
x_1	-0.06	-0.17	1	-0.87
x_2	0.25	0.04	-0.87	1

Residual Correlation Matrix



10. (1 pt.) Determine the number of purchases (N_PURCHASE) and average purchase amount (AVE_PURCHASE_AMT) by CUST_ID for Dataset 5 (see Worksheet) and join this information to the demographic information in Dataset 6 (see Worksheet) to create two new features for predictive modeling.

Worksheet

x_1	x_2	y	x1_TE	x1_x2_FE	x1_LOOTE
A	1	0	0.327273	2.963636	0.25
B	6	1	0.477273	6.238636	0.33
C	11	1	0.727273	8.863636	1
C	14	1	0.727273	8.863636	1
B	4	0	0.477273	6.238636	0.67
A	-1	0	0.327273	2.963636	0.25
A	2	1	0.327273	2.963636	0
A	1	0	0.327273	2.963636	0.25
B	7	0	0.477273	6.238636	0.67
B	12	1	0.477273	6.238636	0.33
A	0.5	0	0.327273	2.963636	0.25

Dataset 1

x1	x1_TE	x1_x2_FE
C	0.727273	8.863636
A	0.327273	2.963636
A	0.327273	2.963636
A	0.327273	2.963636
B	0.477273	6.238636
B	0.477273	6.238636
A	0.327273	2.963636

Dataset 2

x_1	x1_IMP	x1_MISSING
7	7	0
	5.714286	1
8	8	0
	5.714286	1
5	5	0
	5.714286	1
5	5	0
6	6	0
2	2	0
7	7	0

Dataset 3

x_1	x_2	PCA1
6	0	4.4580
4.5	2	2.6535
9	-1	7.0320
3.5	3	1.5655
4	2	2.2820
9	-2	7.3770
4	-1	3.3170
7	-2	5.8910
1	2	0.0530
8	-1	6.2890

Dataset 4

CUST_ID	AMOUNT
jh65432	14.58
jk900h	27.99
jk900h	99.50
jh789k	101.67
jh65432	14.32
jh789k	17.20
jk900h	13.87
jk900h	140.78
jh789k	84.39
jh789k	88.99
jh789k	1.05
jh65432	9.99
jk900h	27.99

Dataset 5

CUST_ID	AGE_BIN	ZIP	N_PURCHASE	AVE_PURCHASE_AMT
jh789k	40-49	20005	**5**	**58.66**
jh65432	30-39	20024	**3**	**12.96333**
jk900h	50-59	20005	**5**	**62.026**

Dataset 6