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:

Xencode = λ·ȳ + (1 - λ)·ȳlevel (1)

where λ is a hyperparameter in [0,1], but can be set to 0.5 by default, ȳ is the overall value of the target or feature for encoding, and ȳlevel 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 x1 and feature encode x1 by x2. Fill in your answers in Dataset 1 in the x1\_TE and x1\_x2\_FE columns.

1. **(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 x1 in Dataset 1 using the leave-one-out method in the x1\_LOOTE column of Dataset 1.
2. 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, x1\_x2\_FE, and x1\_LOOTE transformations learned in Questions 1 and 2.
3. 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 x1 in Dataset 3 (see Worksheet) in the x1\_IMP column, and create a binary missing marker in the x1\_MISSING column.

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

1. What is the obvious drawback to feature extraction techniques that do not apply a penalized approach to generate so-called sparse extracted features?
2. 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?

|  | **x1** | **x2** | **x3** | **y** |
| --- | --- | --- | --- | --- |
| **x1** | 1 | -0.20 | 0.80 | 0.03 |
| **x2** | -0.20 | 1 | -0.07 | -0.11 |
| **x3** | 0.80 | -0.07 | 1 | 0.7 |
| **y** | 0.03 | -0.11 | 0.7 | 1 |

**Dataset A**

|  | **x1** | **x2** | **x3** | **y** |
| --- | --- | --- | --- | --- |
| **x1** | 1 | 0.52 | -0.68 | -0.66 |
| **x2** | 0.52 | 1 | -0.90 | 0.71 |
| **x3** | -0.68 | -0.90 | 1 | 0.88 |
| **y** | -0.66 | 0.71 | 0.88 | 1 |

**Dataset B**

1. Given a training dataset with 2 input columns, x1 and x2, 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).

1. Does one-hot encoding tend to improve or worsen problems associated with the curse of dimensionality?
2. Given Equations 2 and 3 for non-linear additive noise models (NANM), the correlation matrix and the residual correlation matrix below, draw the causal graph for x1, x2, and y.

x2 = g(x1) + 𝛜x1 (2)

y = g(x2) + 𝛜x2 (3)

|  | **x1** | **x2** | **y** |
| --- | --- | --- | --- |
| **x1** | 1 | -0.87 | 0.07 |
| **x2** | -0.87 | 1 | 0.76 |
| **y** | 0.07 | 0.76 | 1 |

Correlation Matrix

|  | **𝛜x1** | **𝛜x2** | **x1** | **x2** |
| --- | --- | --- | --- | --- |
| **𝛜x1** | 1 | 0.32 | -0.06 | 0.25 |
| **𝛜x2** | 0.32 | 1 | -0.17 | 0.04 |
| **x1** | -0.06 | -0.17 | 1 | -0.87 |
| **x2** | 0.25 | 0.04 | -0.87 | 1 |

Residual Correlation Matrix

1. Determine the number of purchases (N\_PURCHASE) and average purchase amount (AVE\_PURCHASE\_AMT) for Dataset 5 and join this information to the demographic information in Dataset 6 (see Worksheet) to create two new features for predictive modeling.

**Worksheet**

| **x1** | **x2** | **y** | **x1\_TE** | **x1\_x2\_FE** | **x1\_LOOTE** |
| --- | --- | --- | --- | --- | --- |
| A | 1 | 0 |  |  |  |
| B | 6 | 1 |  |  |  |
| C | 11 | 1 |  |  |  |
| C | 14 | 1 |  |  |  |
| B | 4 | 0 |  |  |  |
| A | -1 | 0 |  |  |  |
| A | 2 | 1 |  |  |  |
| A | 1 | 0 |  |  |  |
| B | 7 | 0 |  |  |  |
| B | 12 | 1 |  |  |  |
| A | 0.5 | 0 |  |  |  |

**Dataset 1**

| **x1** | **x1\_TE** | **x1\_x2\_FE** | **x1\_LOOTE** |
| --- | --- | --- | --- |
| C |  |  |  |
| A |  |  |  |
| A |  |  |  |
| A |  |  |  |
| B |  |  |  |
| B |  |  |  |
| A |  |  |  |

**Dataset 2**

| **x1** | **x1\_IMP** | **x1\_MISSING** |
| --- | --- | --- |
| 7 |  |  |
|  |  |  |
| 8 |  |  |
|  |  |  |
| 5 |  |  |
|  |  |  |
| 5 |  |  |
| 6 |  |  |
| 2 |  |  |
| 7 |  |  |

**Dataset 3**

| **x1** | **x2** | **PCA1** |
| --- | --- | --- |
| 6 | 0 |  |
| 4.5 | 2 |  |
| 9 | -1 |  |
| 3.5 | 3 |  |
| 4 | 2 |  |
| 9 | -2 |  |
| 4 | -1 |  |
| 7 | -2 |  |
| 1 | 2 |  |
| 8 | -1 |  |

**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 |  |  |
| jh65432 | 30-39 | 20024 |  |  |
| jk900h | 50-59 | 20005 |  |  |

**Dataset 6**