Adventures in Feature Engineering
Feature engi-whaaat?:

“Feature engineering is the act of extracting features from raw data and transforming them into formats that are suitable for the machine learning model.”

-Alice Zheng
Agenda

• Overview
• Categorical Data
• Numeric Transforms
• Quantization/Binning
• Text
• Q/A

Squiggly Letters

\[ f(x) = \text{we'll put notation here (usually)} \]

Eye candy

pretty pictures and other visuals will go here (usually)
Categorical data
Dummy & Effect Encoding

- **Dummy Encoding**
  - Converts a two-category feature into two distinct binary features

- **Effect Encoding**
  - Reference population (e.g. unknown) is coded as -1 allowing the model to handle the effect of male and female separately.

**Dummy Encoding**

\[
f(x) = \begin{cases} 
0 & \text{if } x = \text{male} \\
1 & \text{if } x = \text{female} 
\end{cases}
\]

**Effect Encoding**

\[
f(x) = \begin{cases} 
-1 & \text{if } x = \text{unknown} \\
0 & \text{if } x = \text{female} \\
1 & \text{if } x = \text{male} 
\end{cases}
\]
Additional Thoughts

• If missing or unknown values are common within the data, consider effect-encoding rather than dummy encoding.
• With dummy variables, the reference value (e.g. male) is inherent in the model (creating difficulty with unknown values)
One-hot Encoding

• Converts categorical attributes to binary state indicators

• Consider grouping low-volume categories into a larger bin (to avoid overfitting the data)
Frequency Encoding

- Record linkage example: We’re attempting to probabilistically link users together across disparate data sets (e.g. social media and a CRM)
- Matching on last name is a feature, but not all matches are equal
  - Smith = Smith is not the same as Korvink = Korvink
- A frequency encoded value of the name likelihood can inform the model of the name’s degree of scarcity

### Raw Data

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Principal Diagnosis</th>
<th>Frequency (occurrence rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>COPD</td>
<td>.03</td>
</tr>
<tr>
<td>2</td>
<td>AMI</td>
<td>.02</td>
</tr>
<tr>
<td>3</td>
<td>Pneumonia</td>
<td>.04</td>
</tr>
<tr>
<td>4</td>
<td>Von Hippel-Lindau Disease</td>
<td>.000003</td>
</tr>
</tbody>
</table>

### Frequency-encoded data

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>COPD</th>
<th>AMI</th>
<th>Pneumonia</th>
<th>VHL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.03</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>.02</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>.04</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.000003</td>
</tr>
</tbody>
</table>
Target Mean Encoding

- Each attribute of a categorical feature is assigned the mean/median value of the target (dependent variable)
- Provides prior knowledge to the model
- Akin to P(C) in Naïve Bayes

### Raw Data

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Principal Diagnosis</th>
<th>Complication Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>COPD</td>
<td>.001</td>
</tr>
<tr>
<td>2</td>
<td>AMI</td>
<td>.003</td>
</tr>
<tr>
<td>3</td>
<td>Pneumonia</td>
<td>.002</td>
</tr>
<tr>
<td>4</td>
<td>COPD</td>
<td>.001</td>
</tr>
</tbody>
</table>

### Target mean encoded data

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>COPD</th>
<th>AMI</th>
<th>Pneumonia</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>.003</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>.002</td>
</tr>
<tr>
<td>4</td>
<td>.001</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Chi-Square ($\chi^2$)

- Provides a scalar measure of proportional similarity between two frequency distributions
- Example: How similar are my patients to other hospitals like me?
- Sensitive to low volume (careful if counts are < 5)

### Example: Clinical Cohort Comparison

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pneumonia (PN)</td>
<td>237</td>
<td>8,361</td>
<td>~270</td>
</tr>
<tr>
<td>Heart Failure (HF)</td>
<td>341</td>
<td>9,487</td>
<td>~307</td>
</tr>
<tr>
<td>Total</td>
<td>578</td>
<td>17,848</td>
<td></td>
</tr>
</tbody>
</table>

\[
\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}
\]

\[
\chi^2 = \frac{(237-270)^2}{270} + \frac{(341-307)^2}{307} = \sim 8.03
\]
Reshaping numeric data
Min/Max Scaling

- A technique used to normalize continuous data
- Utilizes the observed minimum and maximum values to transform $x$ into a $x'$ between 0 and 1
  - Where 0 represents the lowest observed value and 1 represents the highest observed value, maintaining the relative distances between each $x$
- Caution: be sure that the minimum and maximum range is stable across the training data and any future data that will be collected

$$x' = \frac{x - \text{min}(x)}{\text{max}(x) - \text{min}(x)}$$
Standardization

• A valuable preprocessing step that can aid in:
  • More accurate results for distance based models
  • Improved computational performance
  • Improved stochastic model training
  • Easier interpretation of trained model parameters

• Results in a distribution of identical shape to the original data-set, but with the following statistical characteristics:
  • A mean value of 0 (centered) – Step 1
  • A standard deviation of 1 (variance scaled) – Step 2

Notation

\[ x' = \frac{x - \bar{x}}{\sigma} \]

Visual

\( \sigma = 80 \)

\( \sigma = 1 \)
Prior Weight Smoothing

- Imposes bias on a probability estimation when information is known or assumed outside of the data being observed
  - A global population on a sub-strata
  - Historical data on a more recent sample
  - A known or assumed additional distribution based on domain knowledge
- E.g. When calculating the mortality rate for a physician with low volume...
  - As a rate calculated off of a small sample size has low confidence, it may be necessary to smooth the physician’s mortality rate towards a more reasonable mean probability based on that physician’s specialty

Notation

$$\hat{p} = \frac{|X_{iy}|}{|X_y|}$$

$$\hat{p} = \frac{|X_{iy}| + \alpha}{|X_y| + \alpha + \beta}$$

Visual

- Raw Estimated Probability
- Prior Observed Probability
- Smoothed Estimated Probability
Additional Thoughts

Two Bernoulli probability density functions with the same mode may have different kurtosis

- More confidence is attributed to probabilities estimated from larger samples
- Higher kurtosis PDFs will have a more narrow range of critical values

Helpful in avoiding overfit when data volume is low or it is known that the sample is not fully representative of reality

\[
kurt(P) = \frac{1}{|P|} \sum_{i=1}^{|P|} (p - \mu)^4 \frac{1}{\sigma^4}
\]

\[
g(p) = \frac{1}{\beta(\alpha, \beta)} p^{|x_{iy}|}(1 - p)^{(|x_y| - |x_{iy}|)}, p \in (0,1)
\]

\[
\argmax(\beta(\alpha, \beta)) = \frac{\alpha}{\alpha + \beta}
\]

\[
\alpha = |X_i| * w_p
\]

\[
\beta = (|X| - |X_i|) * w_p
\]
Laplace Smoothing

- Under special circumstances, Laplace smoothing and prior weight smoothing are identical in function and implementation.
- Can be used to impose bias; however, in statistical modeling and NLP it is more commonly employed as a technique to avoid cascading zeros.
  - A word that has not been previously observed in a training corpus, should not result in a document containing that word to have a zero probability of existing.
  - Oftentimes a probability of zero is more accurately expressed as a probability close to zero.

\[ \hat{\theta}_i = \frac{x_i + \alpha}{N + \alpha d} \quad (i = 1, \ldots, d) \]
Binarization

- Converts continuous data into a binary value
- The threshold value (a = alpha) should meaningfully partition the data

Notation

\[ f(x) = \begin{cases} 
0 & \text{if } a < x \\
1 & \text{if } a > x 
\end{cases} \]

Visual

Is Medicare Eligible

0  a = 65  100
Additional Thoughts

• Examples:
  • A user likes a song if number of plays > a
  • A hospital is a top performer if its readmissions rate is in the top decile

• Considerations:
  • An inferential test can also be used to binarize the data

\[
f(z) = \begin{cases} 
0 & \text{if } a < z \\ 
1 & \text{if } a > z 
\end{cases}
\]
Log Transformations

- Converts data that is multiplicative into data that is additive
- A corrective measure for high-magnitude and skewed distributions
- Usually good for distributions that start at zero
- Other bases (e.g. 10) are often applied
- Use $\ln(x + 1)$ if data range is between 0 and 1

Natural log (base e)

$$\ln(x) = \sim 2.71828^x = e^x$$

Examples

- $\ln(54.6) \approx 4 \quad \therefore \quad e^4 \approx 54.6$
- $\ln(2,980.9) \approx 8 \quad \therefore \quad e^8 \approx 2,980.9$
- $\ln(162,754.8) \approx 12 \quad \therefore \quad e^{12} \approx 162,754.8$
Squares

- Intentionally exaggerate the differences in the data
- Can handle a non-linear (i.e. quadratic) relationship between $x$ and $y$
- Handles negatives values

\[ x \]

\[ x^2 \]
Additional Thoughts

• Example: The relationship (or slope) between for > 60 age range and COPD is different than the relationship between 30-59 and COPD

• Rather than one-hot encoding the age ranges where the change of overfitting increases, the full data set could be used through a non-linear (i.e. squared) feature.
Box Cox

- A broader power function
- Use for positive data variants exist for negative data
- Optimize lambda $\lambda$ for best fit
- $\lambda$ should be between -5 and 5
- $\lambda > 1$ exaggerates the data
- $\lambda < 1$ shrinks the data

**Notation**

$$\tilde{x} \begin{cases} x^\lambda - 1 & \text{if } \lambda \neq 0 \\ \frac{\lambda}{\ln(x)} & \text{if } \lambda = 0 \end{cases}$$
Odds Ratio

- Used to quantify the degree to which two binary variables are associated
- Converts two binary or categorical attributes into a continuous feature that describes their association
- Answers: How do the odds of A being true change if B is true?
- An odds ratio greater than 1 indicates that the likelihood of A is greater when B is present; however, it does not address causality
- For Example... What is the likelihood that the name ‘Tyson Neil’ should be inverted?

\[
\begin{align*}
O_E &= \frac{E_E}{E_N} \\
O_C &= \frac{C_E}{C_N} \\
OR &= \frac{O_E}{O_C} \\
P &= \frac{OR}{(OR + 1)}
\end{align*}
\]

\[
\begin{align*}
O_E &= 1.4e^{-4} \\
O_C &= 6.6e^{-4} \\
OR &= 1.22 \\
P &= \frac{0.0698}{0.0698 + 1}
\end{align*}
\]

\[
\begin{align*}
O_E &= 1.144e^{-4} \\
O_C &= 3.775e^{-5} \\
OR &= 17.48 \\
P &= \frac{0.0698}{0.0698 + 1}
\end{align*}
\]

\[
\begin{align*}
O_E &= 1.22 \\
O_C &= 17.48 \\
OR &= 0.0698 \\
P &= \frac{0.0698}{0.0698 + 1}
\end{align*}
\]

\[
P^{-1} = 93.475\%
\]

This example is based on a methodology utilized by the Census Bureau.
Stacking

- When the output of one or more models is used as the input for another model
- This is often used as a way of reducing dimensionality and/or adding enrichment not easily observed in the raw data
- A mechanism for resolving non-linear or multicollinearity issues by modeling a feature according to its shape as a form of preprocessing
- A technique for generating a composite feature based on multiple features that are related to each other

Which of these sentences is more likely?

The stroke patient received a blood thinner.
The stroke patient received a flu shot.

With stacking:
Interaction Features

- Interaction features can offer a more concise and rich feature that can
  - Offer higher predictive power
  - Aid in convergence during stochastic training

- When a single feature cannot offer the differentiation required to be a valuable input to a model, interaction features can be used
  - Some models can intuit these relationships
  - However, many statistical models assume independence between features

- Common methods employed to create:
  - Product between two features
  - Logical AND between two features
  - Complex methodologies or models to generate a synthesized rich feature
Dimensionality Reduction

- Converts a high dimensional vector into a smaller vector, condensing the relevant information needed to distinguish the relationships between each record
  - A good mechanism for forcing generalization in a model when overfit is an issue
- Within feature engineering, it is typically employed as a form of stacking
- It can also be used as a mechanism for visualizing complex data spaces
- Examples:
  - Principal Component Analysis (PCA)
  - T-distributed Stochastic Neighbor Embedding (t-SNE)
  - Text encoding models: Word-to-Vec, FastText, etc.

Notation

\[ KL(P || Q) = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) \]

(A t-SNE example)

Visual

\[ X = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9] \]

\[ Y = [y_1, y_2, y_3] \]
Additional Thoughts

Notation associated with a T-SNE implementation:

\[ g(x_i, x_j) = e^{-\frac{||x_i - x_j||^2}{2\sigma_i^2}} \]

\[ p_{j|i} = \frac{g(x_i, x_j)}{\sum_{k \neq i} g(x_i, x_k)} \]

\[ p_{i|j} = \frac{p_{j|i} + p_{i|j}}{2N} \]

\[ t(y_i, y_j) = (1 + ||y_i - y_j||)^{-1} \]

\[ q_{ij} = \frac{t(y_i, y_j)}{\sum_{k \neq i} t(y_i, y_k)} \]

\[ KL(P||Q) = \sum_{i \neq j} p_{ij} \log \left( \frac{p_{ij}}{q_{ij}} \right) \]

Minimized through gradient descent by making small incremental adjustments.
Quantization/Binning

1          2         3         4
Fixed Width Binning

• Continuous data is divided into equal-sized bins
• Each data point is represented by its binned value
• Association to a larger group (or bin) may result in greater explanatory power (e.g. Age)

Labeling bins

\[
bin \text{ number} = \text{floor}\left(\frac{x}{\text{bin width}}\right)
\]

Age Example (range 0 – 120)
Additional Thoughts

• Custom Bins: Ranges maybe custom designed (e.g. natural clinical grouping)

• Exponential Binning: If data spans multiple magnitudes (e.g. hospital cost), consider binning on powers (e.g. income), consider power-based binning (0-9, 10-99, 100-999, 1000-9999). Simply take the log of the value to find bin assignment
Quantile Binning

- Evenly divides data into quantiles (e.g. percentiles, deciles, quartiles, etc.)
- Ensures that there are no gaps within the data

\[
bin\ number = \text{ceiling}\left(\frac{x}{\text{quantile}}\right)
\]

Age Example (range 0 – 120)
Text Edit Distances

- **Levenshtein**
  - Insertion, deletion, substitution

- **Damerau-Levenshtein**
  - Insertion, deletion, substitution and transpositions

- **Jaro-Winkler**
  - Emphasizes shared prefix
  - Good for short character strings
  - A variant of Jaro distance
  - Looks at transpositions only
  - Preference of prefix match up to 4 characters

Notation

\[
\text{dist} = e_{\text{dist}}(\text{string1}, \text{string2})
\]

\[
\text{similarity} = 1 - \left( \frac{\text{dist}}{\max(|\text{string1}|, |\text{string2}|)} \right)
\]

Example: **Michael** vs **Micheal**

<table>
<thead>
<tr>
<th>Type</th>
<th>Distance</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levenshtein</td>
<td>2</td>
<td>~.71</td>
</tr>
<tr>
<td>Damerau-L</td>
<td>1</td>
<td>~.85</td>
</tr>
<tr>
<td>Jaro-Winkler</td>
<td>~.03</td>
<td>~.97</td>
</tr>
</tbody>
</table>

*Similarity for JW is 1-Distance*
Phonetic Similarity

- **Soundex**
  - Converts names to four-character phonetic representations
  - Difference will be a value between 1-4

- **Metaphone**
  - Improves upon Soundex (limited to English)

- **Double Metaphone 2 & 3**
  - Further improvements to account for spelling nuances as well as non-English words

Example: **Michael** vs **Micheal**

### Formulas

\[
\begin{align*}
  e_i &= \text{encode}(\text{string}_i) \\
  \text{dist} &= e_\text{dist}(e_1, e_2) \\
  \text{similarity} &= 1 - \left( \frac{\text{dist}}{\max(|e_1|, |e_2|)} \right)
\end{align*}
\]

<table>
<thead>
<tr>
<th>Type</th>
<th>Phonetic Encoding String 1</th>
<th>Phonetic Encoding String 2</th>
<th>Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soundex</td>
<td>M240</td>
<td>M240</td>
<td>1</td>
</tr>
<tr>
<td>Metaphone</td>
<td>MXL</td>
<td>MXL</td>
<td>1</td>
</tr>
<tr>
<td>Double Metaphone 2</td>
<td>MKL</td>
<td>MXL</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Jaccard Similarity

- Evaluates the similarity between two sets
- Numerator = the intersection of distinct elements in the set
- Denominator = the distinct union of elements in the set
- Provides a value between 0 and 1 (1 being identical and 0 being disjoint sets)

Formula:

\[ J(A, B) = \frac{|A \cap B|}{|A \cup B|} \]

Example: Patient Resource Utilization
Condition - TYPE 2 DIABETES MELLITUS

<table>
<thead>
<tr>
<th></th>
<th>Patient 1</th>
<th>Patient 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER R&amp;B</td>
<td>ER R&amp;B</td>
<td></td>
</tr>
<tr>
<td>METABOLIC PANEL</td>
<td>METABOLIC PANEL</td>
<td></td>
</tr>
<tr>
<td>GLUCOSE TEST</td>
<td>URINALYSIS</td>
<td></td>
</tr>
<tr>
<td>URINALYSIS</td>
<td>COMPLETE CBC</td>
<td></td>
</tr>
</tbody>
</table>
Spatial Distances

- Used to quantify the distance between two elements in coordinate or vector space

- Euclidean
  - Based on the Pythagorean theorem
  - Related to l2 norms

- Manhattan
  - Related to l1 norms

- Cosine Similarity
  - Magnitude agnostic
  - Requires points to be in positive space

Notation

\[ ed(p,q) = \sqrt{\sum_i (p_i - q_i)^2} \]
\[ md(p,q) = \sum_i |p_i - q_i| \]
\[ S_c(p,q) = \cos \theta = \frac{\sum_i p_i q_i}{\sqrt{\sum_i p_i^2} \sqrt{\sum_i q_i^2}} \]

Visual

\[ md(p,q) = 3 + 1 = 4 \]

\[ ed(p,q) = \sqrt{3^2 + 1^2} = \sqrt{10} \]

\[ S_c(p,q) = \cos \theta = \frac{2}{\sqrt{10} \cdot 2} = \frac{1}{\sqrt{10}} \]
Distribution Distances

- Used to quantify the distance (or inverse similarity) between distributions
  - The examples below are options that can be used for non-normal distributions

- Kernel density estimation
  - Integrals at overlapping ranges can be computed to determine similarity

- Wasserstein’s metric (Earth mover’s distance)
  - The minimum cost required to move the area of one probability distribution so as to be equal to another

- Mann Whitney U test
  - Assesses the likelihood that two independent distributions were drawn from the same population
Kernel density estimation

\[ D_Z(x) = \frac{1}{n} \sum_z^n D_{K_z}(x) \]

\[ I_Z(x) = \frac{1}{n} \sum_z^n I_{K_z}(x) \]

Wasserstein’s metric

\[ \text{EMD}(P, Q) = \frac{\sum_i^m \sum_j^n f_{i,j}d_{i,j}}{\sum_i^m \sum_j^n f_{i,j}} \]

Optimized via a minimum-cost flow linear programming algorithm.

Mann Whitney U test

\[ U_x = \text{RankSum}(x) - \frac{n_x(n_x + 1)}{2} \]

\[ U_{\text{stat}} = \min(U_1, U_2) \]

\[ \sigma_U = \sqrt{\frac{n_1n_2}{12}((n + 1) - \sum_{i=1}^k t_i^3 - t_i)} \]

\[ \mu_U = \frac{n_1n_2}{2} \]

\[ Z_U = \frac{U_{\text{stat}} - \mu_U}{\sigma_U} \]
Feature *Engineering vs. Feature Selection*

- Feature engineering is the transformation of features into more meaningful features, considering the type of model that is being employed and the nature of the data.
- Features selection is deciding which features should be utilized by the model that is employed.
- Feature engineering can sometimes function similarly to feature selection when applying dimensionality reduction.
- Both techniques should be informed by knowledge the model that is being employed.
Feature Selection: A Brief Digression

Goal is largely to increase parsimony. Methods include:

• **Filter**: The process of excluding low-value features without requiring the model implementation (e.g. $R^2$ between x and y)

• **Wrapper**: An iterative approach using model results to exclude low-value features

• **Embedded**: Feature selection is inherent in the model choice (e.g. Ridge & Lasso regression)
Iterative Cycle that Informs Each Phase

1. Clearly Define Business Problem
2. Data Extraction & Munging
3. Analyze Data
4. Feature Engineering
5. Feature Selection
6. Model Selection
7. Train & Evaluate Model
8. Document & Deploy
Book Recommendation

- Neatly organizes common feature engineering concepts
- Inspiration for much of this talk