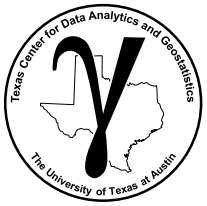


PGE 383 Machine Learning

Feature Selection

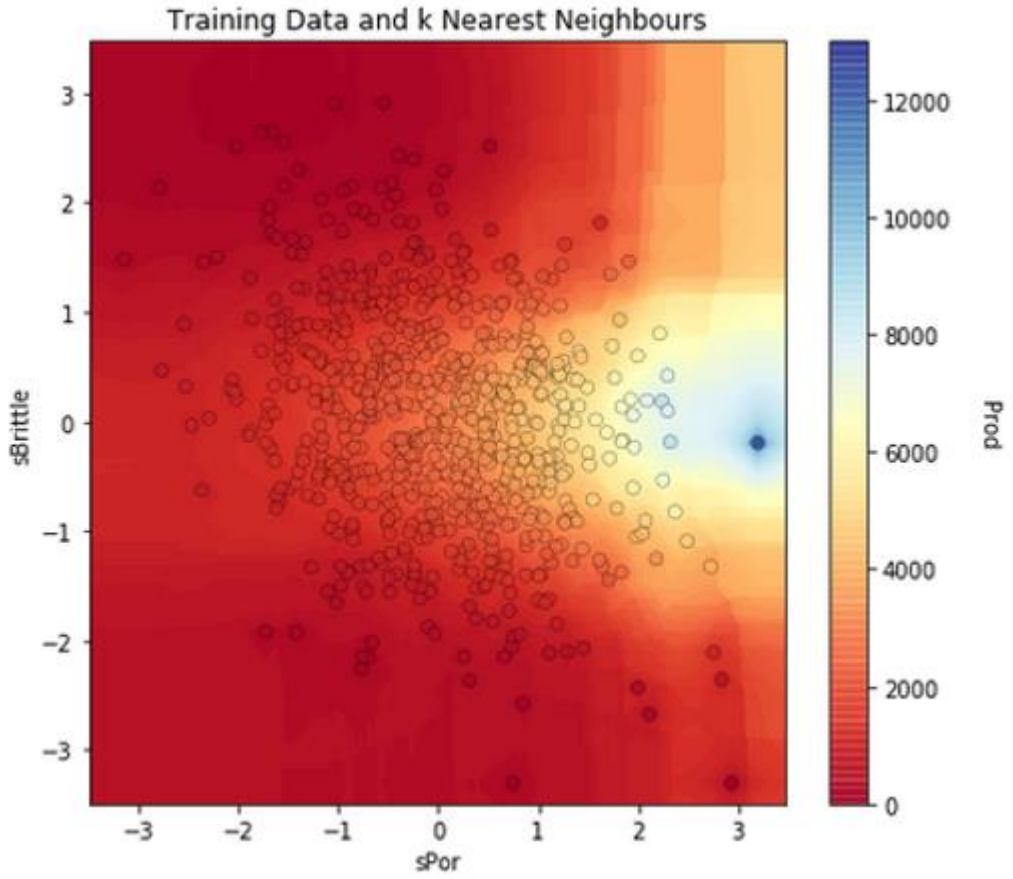
Lecture outline . . .

- **Curse of Dimensionality**
- **Feature Selection**
- **Mutual Information**
- **Shapley Values**
- **Feature Selection Hands-on**

Michael Pyrcz, The University of Texas at Austin

Motivation for Multivariate Methods

- **We build better models when we careful select the most informative set of predictor features.**



PGE 383 Machine Learning

Feature Selection

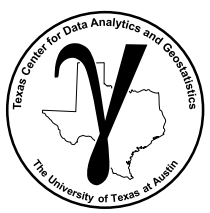
Lecture outline . . .

- **Curse of Dimensionality**

Michael Pyrcz, The University of Texas at Austin

Multivariate

- One of the definitions of Big Data is variety
  - This suggests massively multivariate datasets, many features
- Traditional reservoir modeling workflows were bivariate
  - Facies, then porosity in facies and permeability constrained to porosity
  - The most complicated simulation is permeability accounting for the joint porosity simulated realization
- Unconventionals, and Whole Earth Models
  - Require inclusion many more variables
  - We need to model facies, porosity, geomechanical properties, geophysical properties, total organic carbon, maturity etc.
- When working with Multivariate it is very challenging:
  - Visualize
  - Detect relationships and patterns



Curse of Dimensionality

- The challenges of working with many features (i.e., high dimensional models)
  - visualization
  - sampling
  - coverage
  - distorted space
  - multicolinearity

Feature Space Definition

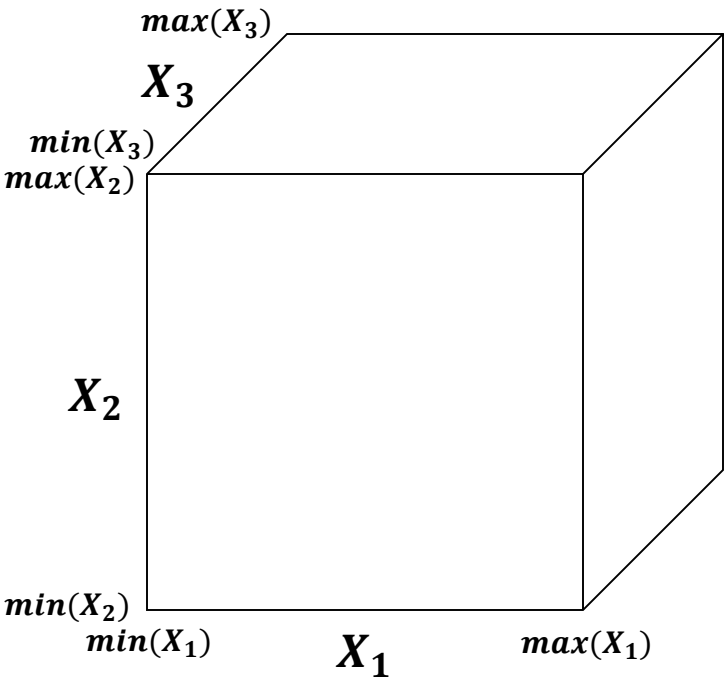
Feature Space

- The m-dimensional space, R^m, over the ranges of each of the features,
- In general our feature space includes all possible cases of our features.

$x_1 \in [min(X_1), max(X_1)]$, ...,$x_1 \in [min(X_m), max(X_m)]$

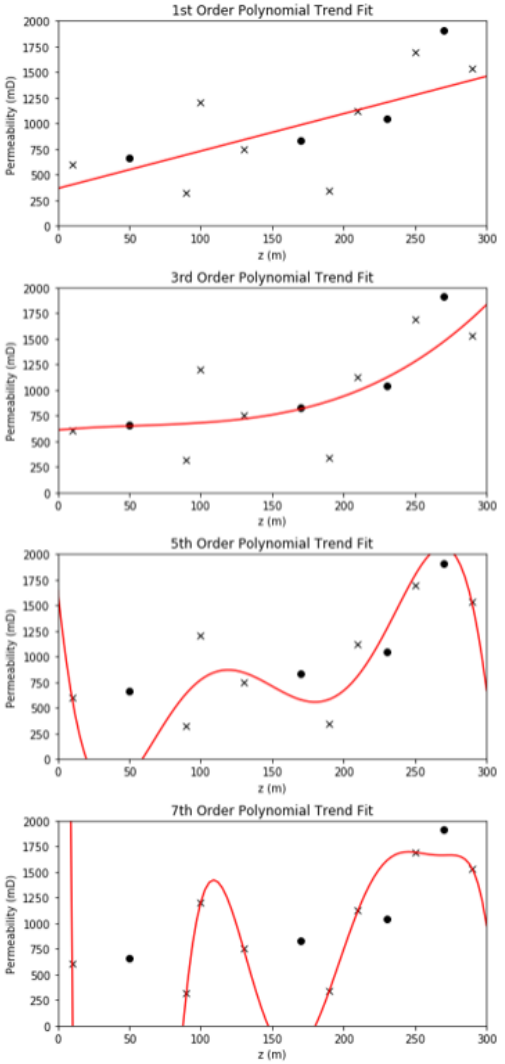
Feature Space in Machine Learning

- Commonly 'feature space' only refers to the predictor features and does not include the response feature(s)
- In this course I will specify predictor feature space.
- Typically we will train our machines to make predictions over the predictor feature space.
- More complicated shapes of predictor feature space are possible, e.g. we could mask/remove subsets with poor data coverage.



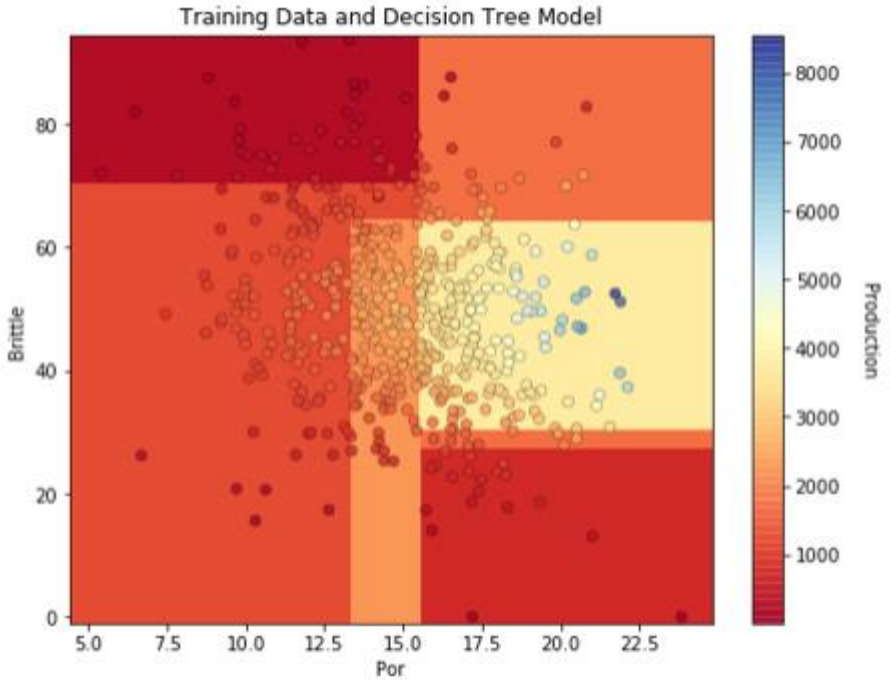
Curse of Dimensionality Visualization

- Consider this simple model:
  - 1 predictor feature
  - 1 response feature
- How's our model performing?
  - Accuracy in training and testing
- Range of Applicability?
  - Are we extrapolating?
- Overfit
  - Is the model defendable given the data?



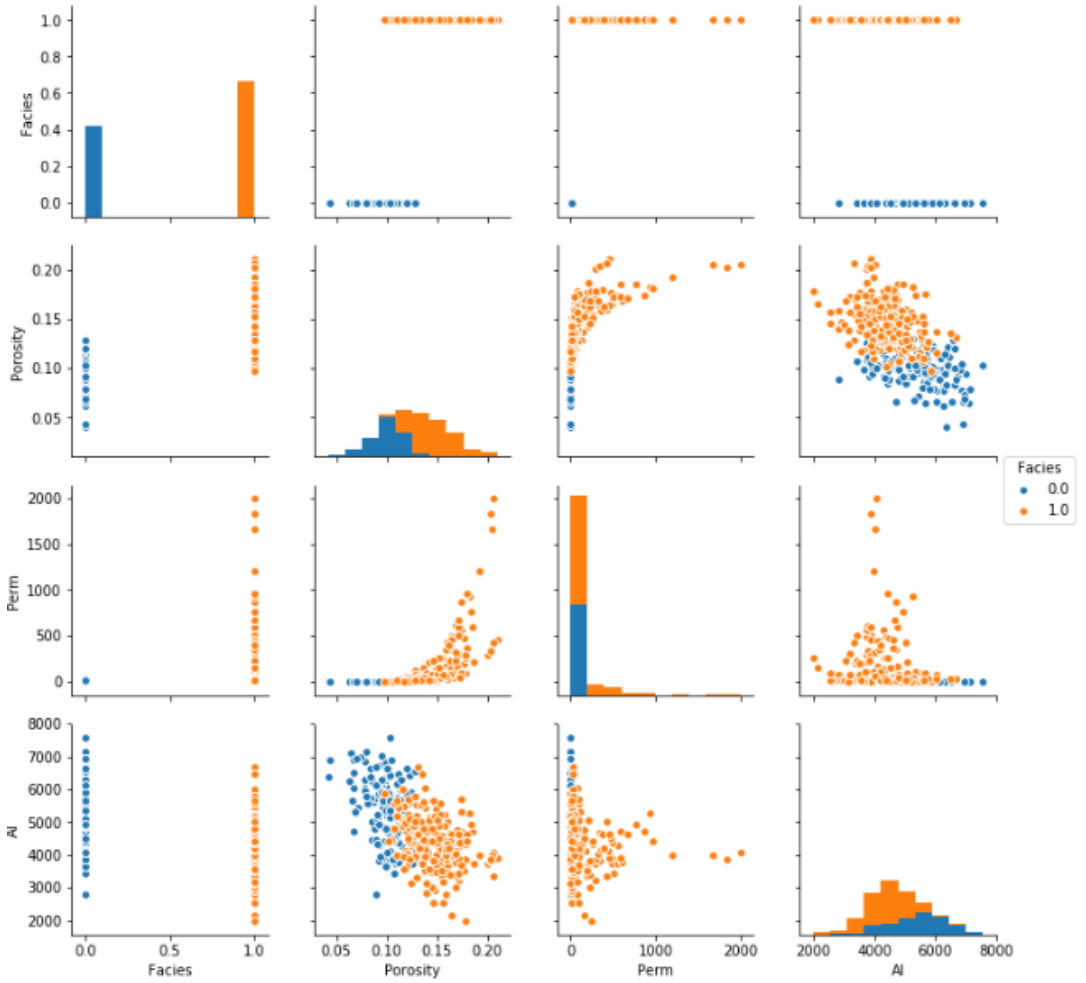
Curse of Dimensionality Visualization

- Consider this simple model:
  - 2 predictor features
  - 1 response feature
- How's our model performing?
  - Accuracy in training and testing
- Range of Applicability?
  - Are we extrapolating?
- Overfit
  - Is the model defendable given the data?



Curse of Dimensionality Visualization

- Consider this simple model:
  - 4 predictor features
  - 1 response feature (not shown)
- What are the relationships between features?
- Are there constraints?

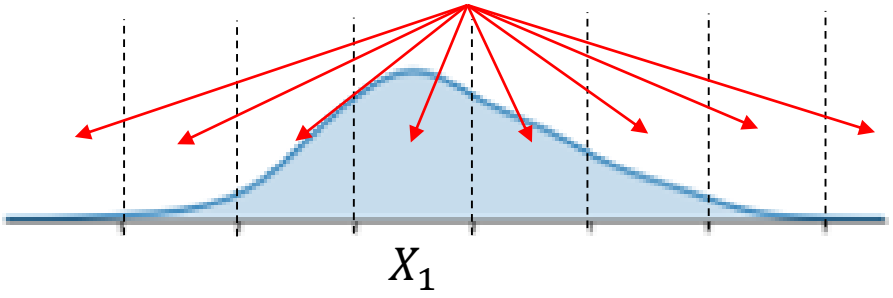


Curse of Dimensionality Sampling

Recall the calculation of a histogram / normalized histogram.

1. We establish ‘bins’, discretize the range of each feature.

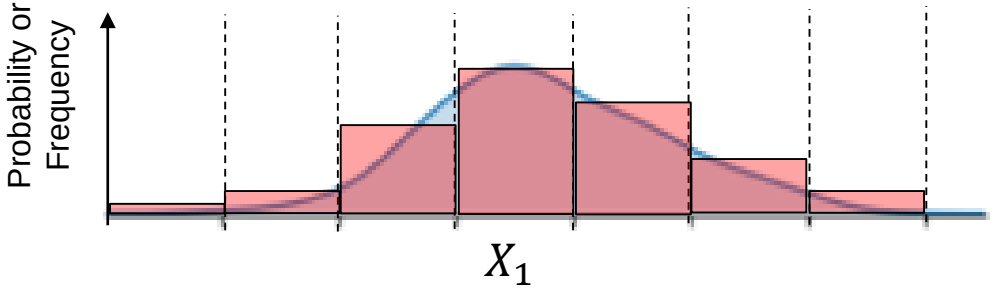
possible outcomes of X_1



Related to combinatorics

2. We calculate probabilities with a ratio and enough samples/bin.

sampling to calculate probability of possible outcomes of X_1



Related to combinatorics and sampling

$$P\left(X_1^i \leq X \leq X_1^{i+1}\right) = \frac{n(X_1^i \leq X \leq X_1^{i+1})}{n}$$

Curse of Dimensionality Sampling

Calculating the Joint Probability

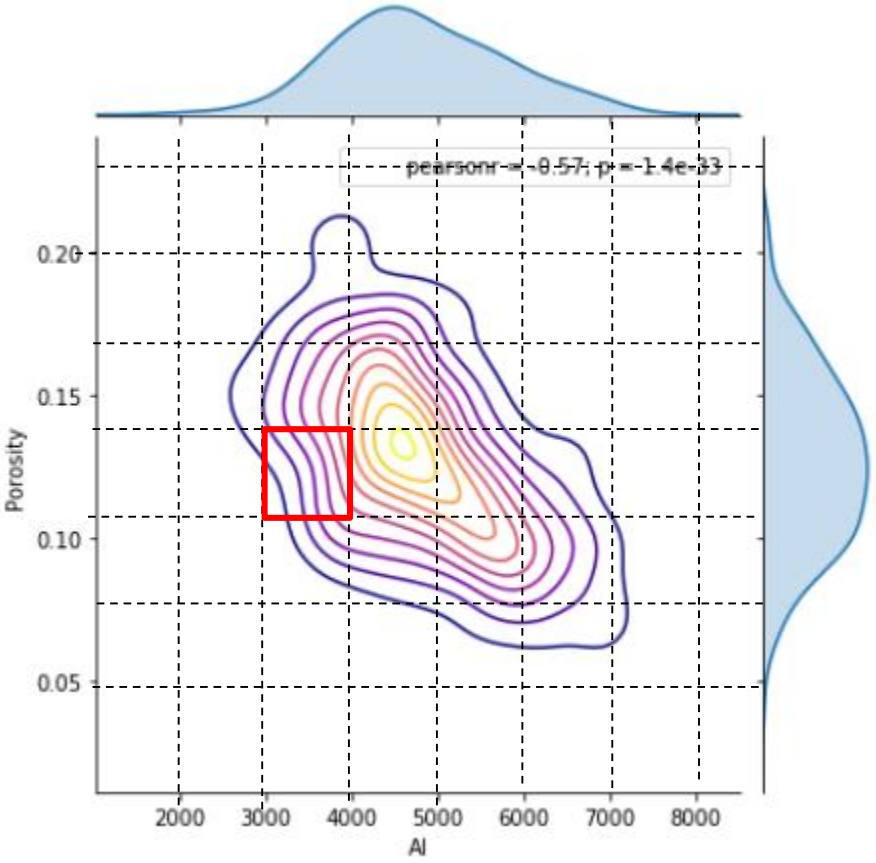
- Consider any joint probability:
  $P(X_1 \in bin_{i_1} \cap, \ldots, \cap X_m \in bin_{i_m})$
- E.g. $P(3k < AI < 4k, 0.11 < \varphi < 0.13) = n(3k < AI < 4k, 0.11 < \varphi < 0.13)/n$

where n is the total number of samples.

- We need enough samples, n , replicates of all possible combinations to go from frequency to probability.
- Where, $n_{s/bin}$, is the nominal number of samples per combination.

$$n = n_{s/bin} \cdot {n_{bins}}^{m}$$

- Note: this is optimistic, as it assumes uniform sampling

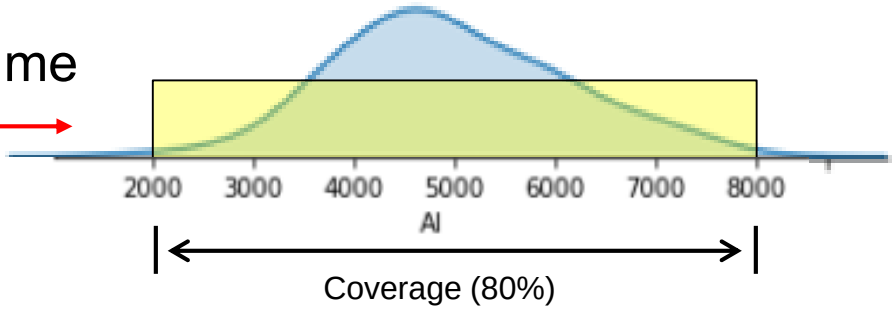


For this example, $n = n_{s/bin} \cdot {n_{bins}}^{m} = 10 \cdot 8^2$, given we need 10 nominally samples per bin, we need 640 data.

Curse of Dimensionality Coverage

Consider coverage:

- The range of the sample values
- The fraction of the possible solution space that is sampled
- Let's return to 1 feature and assume 80% coverage!



- That's pretty good right?
- Remember, we usually, directly sample only $\frac{1}{10^9}$ of the volume of the subsurface
- Yes, the concept of coverage is subjective, how much data to cover? What about gaps? etc.

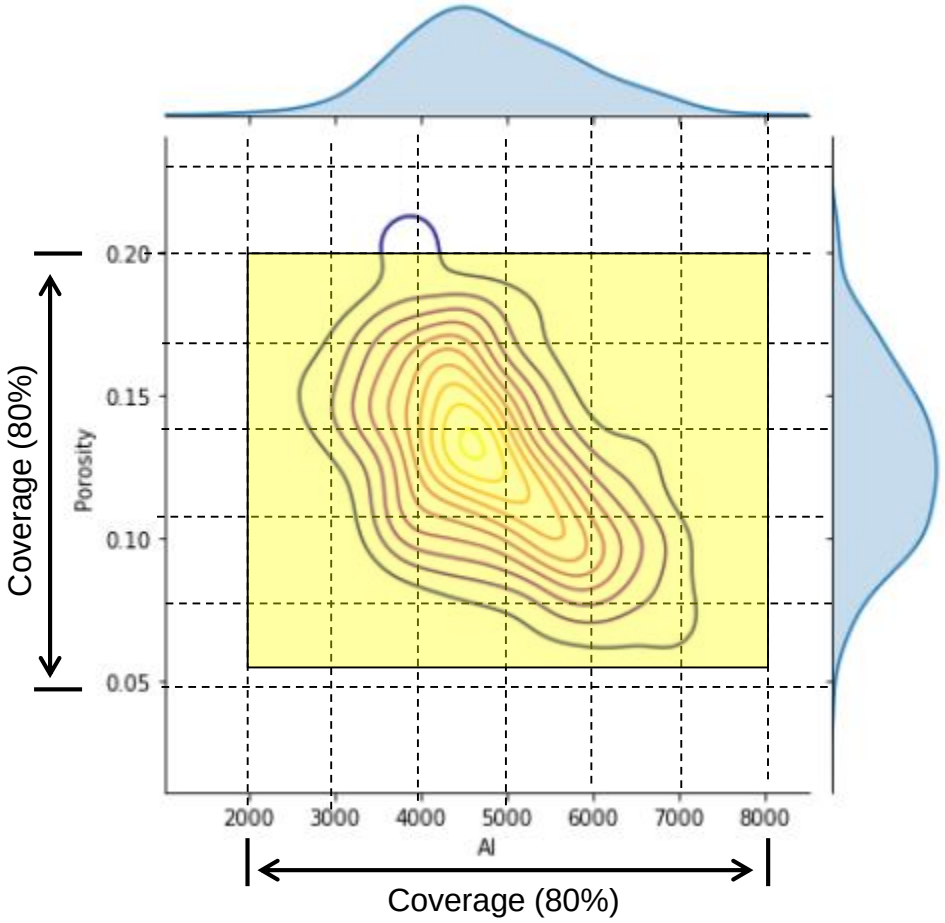
Curse of Dimensionality Coverage

Consider coverage:

- It is common not to have samples that cover the entire predictor feature space
- Consider coverage over each feature, c_1.
- How much of the solution space is covered?

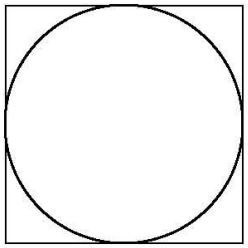
$$c = c_1^m$$

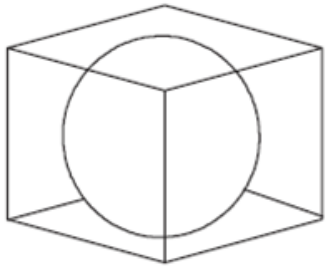
coverage is decreasing, exponential decay with decay constant, $\lambda = 1$, as we increase the number of features, m!



For this example, $c = c_1^m = 0.8^2 = 0.64$

Curse of Dimensionality Distorted Space





Distances in High Dimensional Space

The vastness of hyperdimensional space

- Take the ratio of the volume of an inscribed hypersphere in a hypercube.

$$\frac{\pi^{m/2}}{m2^{m-1}\Gamma(m/2)} \to 0 \text{ as } m \to \infty$$

Recall, $\Gamma(n) = (n-1)!$

Interpretation: High dimensional space is all corners and no 'middle' and all of high dimensional space is far from the middle.

$$\lim_{m\to\infty} E\{dist_{max}(m) - dist_{min}(m)\} \to 0$$

Expectation of max and min distance separation of random points.

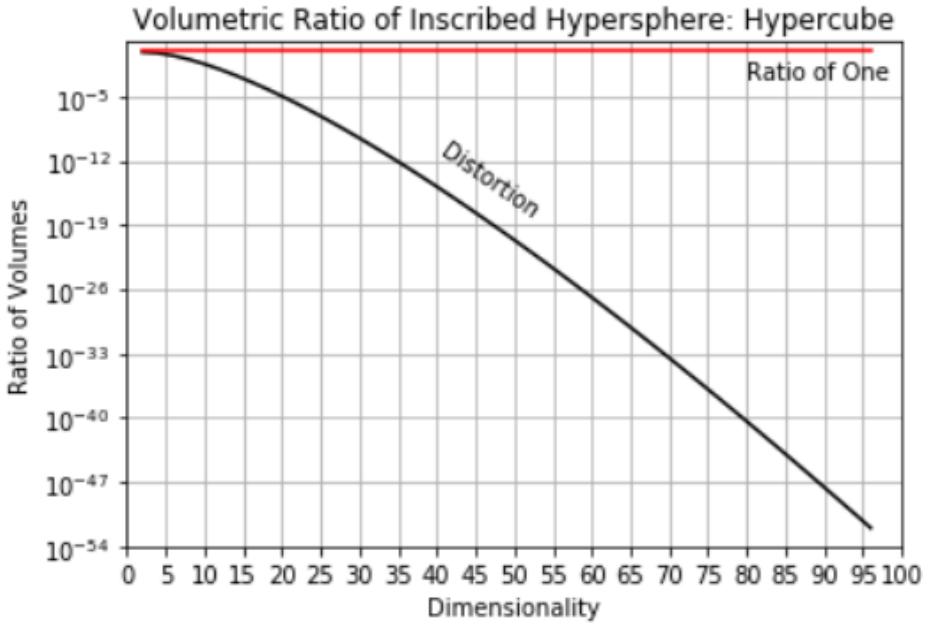
- The limit of the expectation of the range of pairwise distances over random points in hyperdimensional space tends to zero.
  - Distances are almost all the same and Euclidian distance is no longer meaningful

Curse of Dimensionality
Distorted Space

Distances in High Dimensional Space

How many dimensions are we talking about? Here's the ratio of center to corners from the previous equation.

Note, the y axis is in log scale.



Volumetric ratio of an inscribed hypersphere (n-sphere) in a hypercube (n-cube).

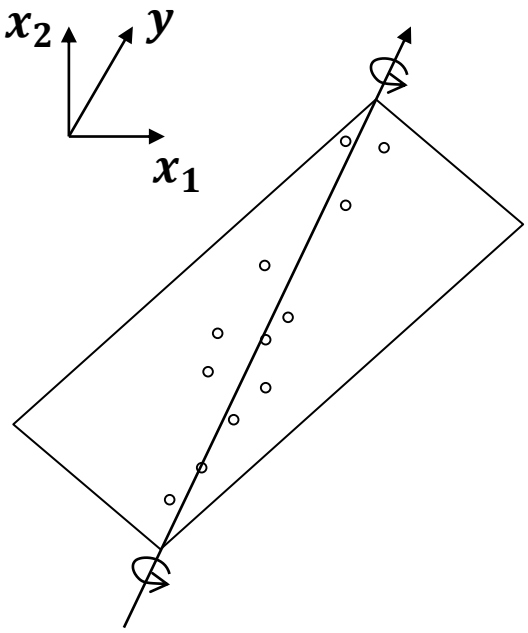
Curse of Dimensionality
Multicolinearity

- "*The existence of such a high degree of correlation between supposedly independent variables being used to estimate a dependent variable that the contribution of each independent variable to variation in the dependent variable cannot be determined*"

  - Merriam-Webster Online Dictionary

- "*In statistics, multicollinearity (also collinearity) is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.*"

  - Wikipedia



It is like fitting a plane to a line!

Curse of Dimensionality

- We get a better model with fewer, informative features than

  'Throwing everything and the kitchen sink into the model!'

- Fewer features for models are simpler, faster, easier to visualize and less likely overfit and results in the **best model!**

Curse of Dimensionality

Working with more features / variables is harder!

1. More difficult to visualize data and model
2. More data are required to infer the joint probabilities
3. Less data coverage of feature space
4. More difficult to interrogate / check the model
5. More likely redundant features resulting in model instability
6. More computational effort, more computational resources and longer run times
7. More complicated model is more likely overfit
8. More professional time for model construction

PGE 383 Machine Learning

Feature Selection

Lecture outline . . .

- **Feature Selection**

Michael Pyrcz, The University of Texas at Austin

What is Feature Ranking?

Feature Ranking:

- Feature ranking is a set of metrics that assign relative importance or value to each feature with respect to information contained for inference and importance in predicting a response feature.
- There are a wide variety of possible methods to accomplish this.
- My recommendation is a **wide-array** approach with multiple metric, while understanding the assumptions and limitations of each metric.

Here's the general types of metrics that we will consider for feature ranking:

1. Visual Inspection of Data Distributions and Scatter Plots
2. Statistical Summaries
3. Model-based
4. Recursive Feature Elimination

What is Feature Ranking?

Expert Knowledge:

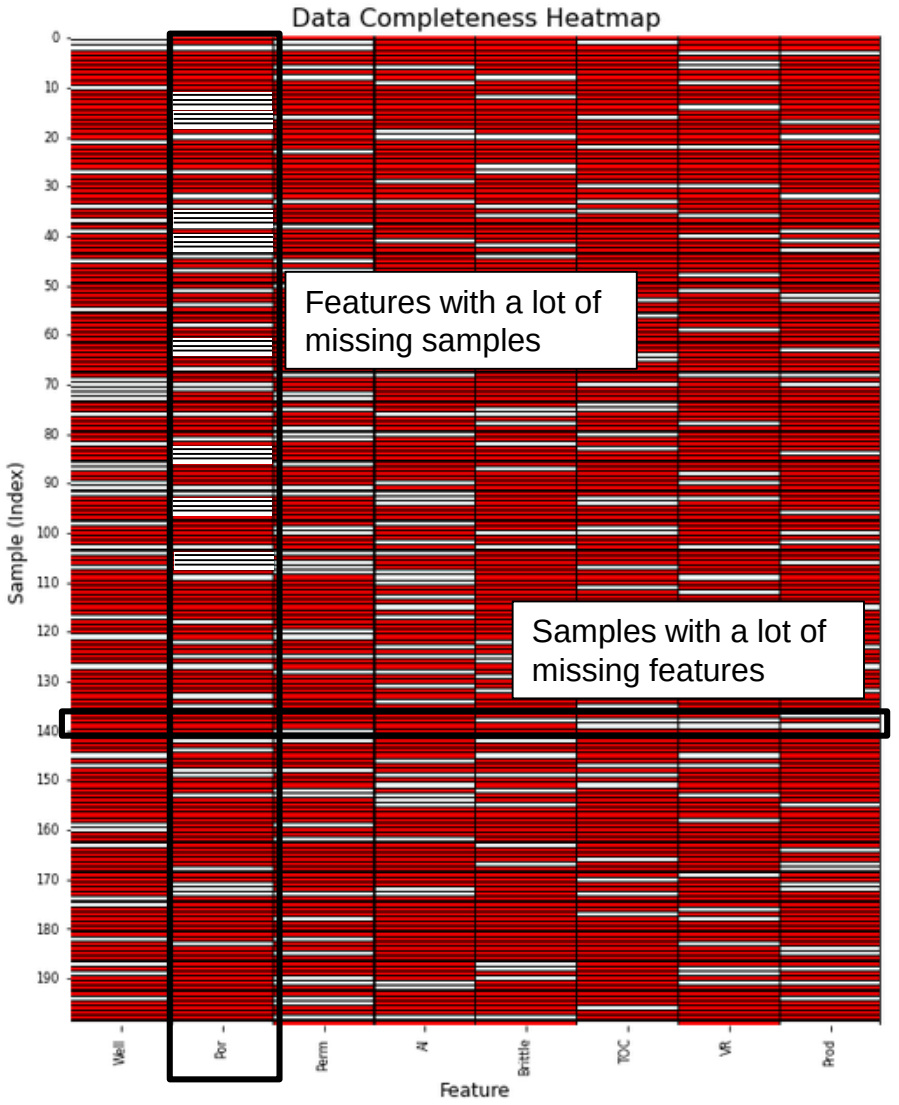
- Also, we should not neglect expert knowledge.
- If additional information is known about physical processes, causation, reliability and availability of features this should be integrated into assigning feature ranks.
- We should be learning as we perform our analysis, testing new hypotheses.

What is Feature Ranking?

Sample Coverage:

- When we include features with a lot of missing values we either:
  - increase the data imputation challenge
  - loose a lot of samples if likewise deletion is used

- We can use a heat map to check data coverage
  - identify features with a lot of missing samples
  - identify samples with a lot of missing features.



Feature Ranking Metrics

Metric - Visual Inspection

- In any multivariate work we should start with the univariate analysis, summary statistics of one variable at a time. The summary statistic ranking method is qualitative, we are asking:
  - are there data issues?
  - do we trust the features? do we trust the features all equally?
  - are there issues that need to be taken care of before we develop any multivariate workflows?
  - coverage, are there a lot of missing values?

Feature Ranking Metrics

Summary statistics are a critical first step in data checking.

	count	mean	std	min	25%	50%	75%	max
Well	200.0	100.500000	57.879185	1.000000	50.750000	100.500000	150.250000	200.000000
Por	200.0	14.991150	2.971176	6.550000	12.912500	15.070000	17.402500	23.550000
Perm	200.0	4.330750	1.731014	1.130000	3.122500	4.035000	5.287500	9.870000
AI	200.0	2.968850	0.566885	1.280000	2.547500	2.955000	3.345000	4.630000
Brittle	200.0	48.161950	14.129455	10.940000	37.755000	49.510000	58.262500	84.330000
TOC	200.0	0.991950	0.478264	0.000000	0.617500	1.030000	1.350000	2.180000
VR	200.0	1.964300	0.300827	0.930000	1.770000	1.960000	2.142500	2.870000
Prod	200.0	3864.407081	1553.277558	839.822063	2686.227611	3604.303507	4752.637556	8590.384044
const	200.0	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000

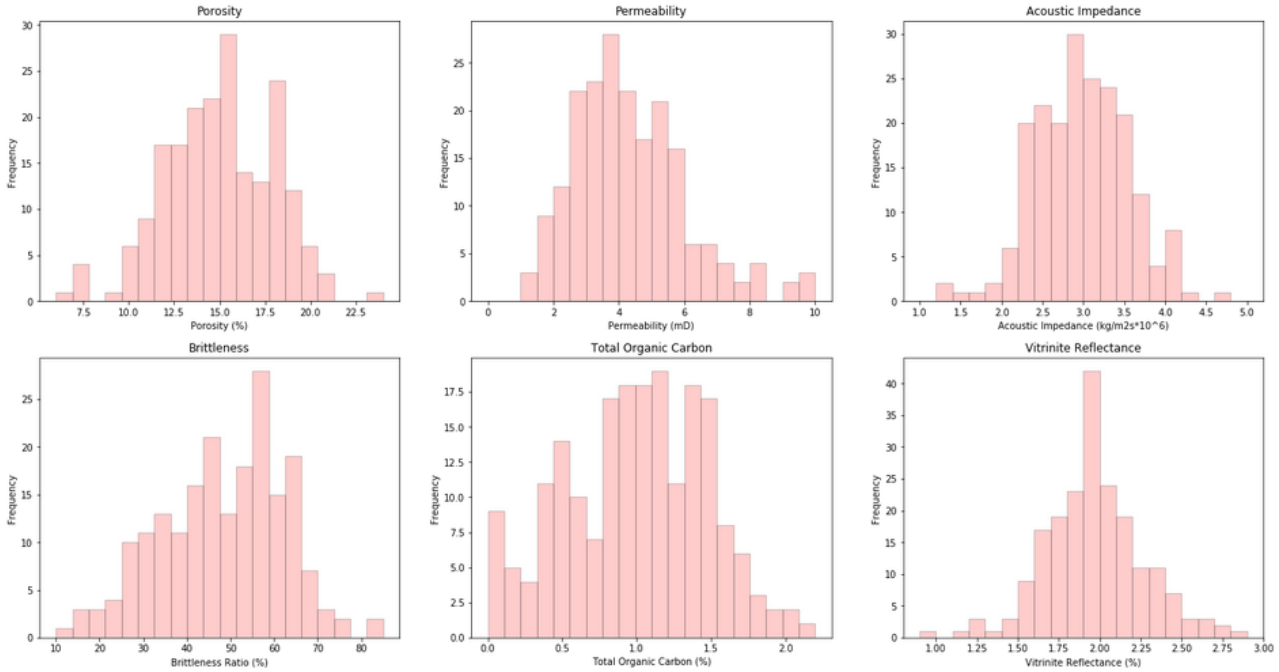
- the number of valid (non-null) values for each feature
- general behaviors such as central tendency, mean, and dispersion, variance.
- issues with negative values, extreme values, and values that are outside the range of plausible values for each property.

Feature Ranking Metrics

Ranking Method - Univariate Distributions

- As with summary statistics, this ranking method is a qualitative check for issues with the data and to assess our confidence with each feature.
- It is better to not include a feature with low confidence of quality as it may be misleading (while adding to model complexity as discussed previously).
- Assess our ability to use methods that have distribution assumptions

Feature Ranking Metrics



The univariate distributions look good:

- there are no obvious outliers
- the permeability is positively skewed as often observed
- the corrected TOC has a small zero truncation spike, but it's reasonable
- some departure from Gaussian form, could transform

Feature Ranking Metrics

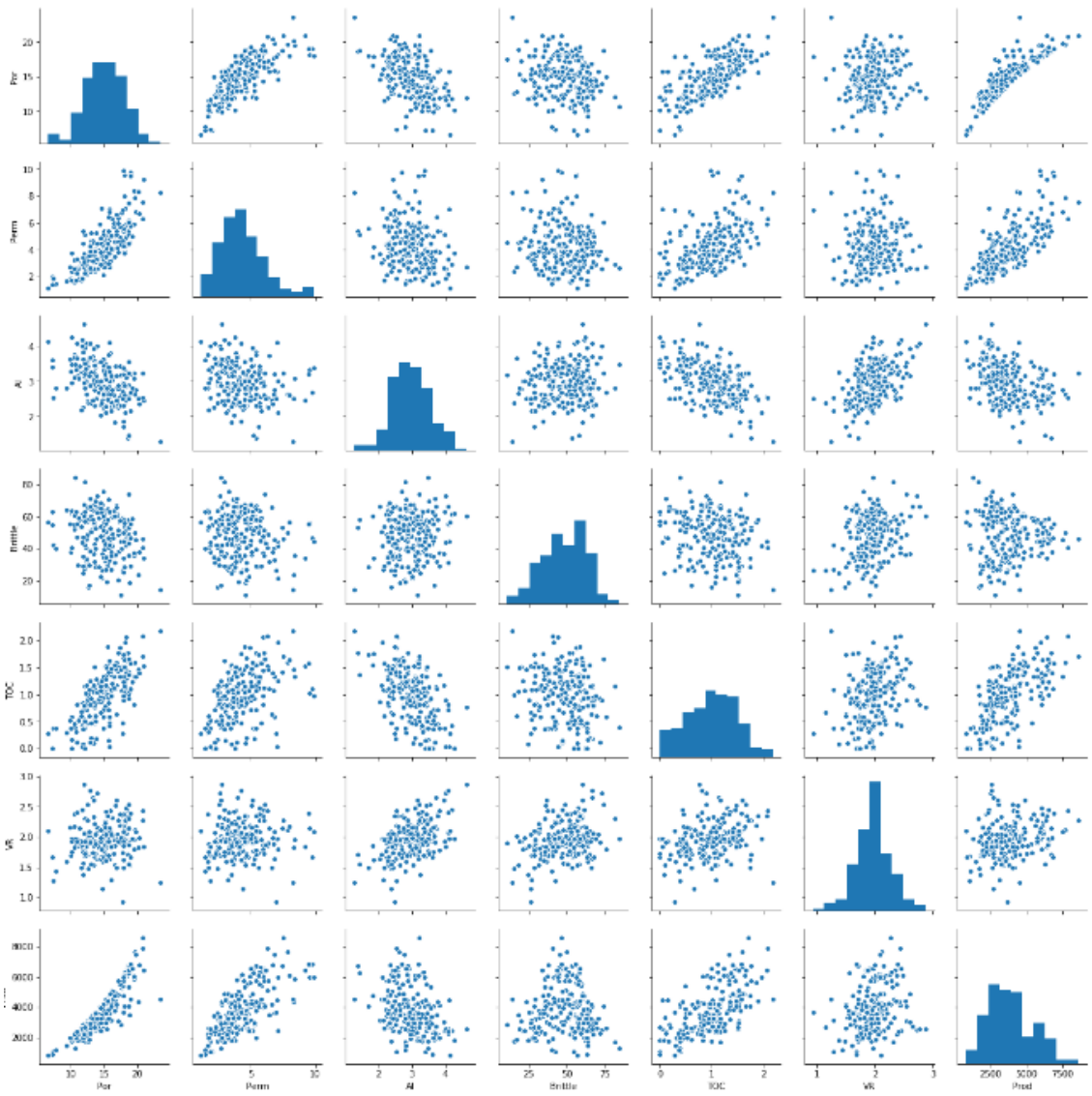
Ranking Method – Bivariate Statistics

- matrix scatter plots are a very efficient method to observe the bivariate relationships between the variables.
- this is another opportunity through data visualization to identify data issues, outliers
- we can assess if we have collinearity, specifically the simpler form between two features at a time
- Bivariate Gaussian is assumed for methods such as correlation and partial correlation

Feature Ranking Metrics

How could we use this plot for variable ranking?

- variables that are closely related to each other.
- linear vs. non-linear relationships
- constraint relationships and heteroscedasticity between variables.



Feature Ranking Metrics

Ranking Method - Bivariate

- bivariate visualization and analysis is not sufficient to understand all the multivariate relationships in the data
- multicollinearity includes strong linear relationships between 2 or more features.
- higher order nonlinear features, outliers and coverage?
- these may be hard to see with only bivariate plots.

Feature Ranking Metrics

Ranking Method - Pairwise Covariance

- Pairwise covariance provides a measure of the strength of the linear relationship between each predictor feature and the response feature.
- We now specify our goal of this study is to predict production, our response variable, from the other available predictor features.
- We are thinking predictively now, not inferentially, we want to estimate the function, $\hat{f}$ to accomplish this

Covariance:

- measures the strength of the linear relationship between features
- sensitive to the dispersion / variance of both the predictor and response

Feature Ranking Metrics

Ranking Method - Pairwise Covariance

- Sensitive to feature variance
- Feature variance is somewhat arbitrary.
  - For example, what is the variance of porosity in fraction vs. percentage or permeaiblity in Darcy vs. milliDarcy. We can show that if we apply a constant multiplier, cc, to a variable, XX, that the variance will change according to this relationship (the proof is based on expectation formulation of variance):

$$\sigma_{cX}^2 = c^2 \sigma_X^2$$

  - By moving from percentage to fraction we decrease the variance of porosity by a factor of 10,000!
  - The variance of each variable is potentially arbitrary, with the exception when all the features are in the same units.

Feature Ranking Metrics

Ranking Method - Pairwise Correlation Coefficient

- Pairwise correlation coefficient provides a measure of the strength of the linear relationship between each predictor feature and the response feature.

- The correlation coefficient:
  - measures the linear relationship
  - removes the sensitivity to the dispersion / variance of both the predictor and response features, by normalizing by the product of the standard deviation of each feature

Feature Ranking Metrics

Ranking Method – Rank Correlation Coefficient

- The rank correlation coefficient applies the rank transform to the data prior to calculating the correlation coefficient. To calculate the rank transform simply replace the data values with the ranks, where n is the maximum value and 1 is the minimum value.
- The rank correlation:
  - measures the monotonic relationship, relaxes the linear assumption
  - removes the sensitivity to the dispersion / variance of both the predictor and response, by normalizing by the product of the standard deviation of each.

Feature Ranking Metrics

Ranking Method – Partial Correlation Coefficient

This is a linear correlation coefficient that controls for the effects all the remaining variables

- $\rho_{XY.Z}$ and is the partial correlation between X and Y after controlling for Z.

1. perform linear, least-squares regression to predict X from $Z_{1,\ldots,m-2}$.
2. calculate the residuals in Step #1, $X - X^*$
3. perform linear, least-squares regression to predict Y from $Z_{1,\ldots,m-2}$.
4. calculate the residuals in Step #1,$Y - Y^*$
5. calculate the correlation coefficient, $\rho_{XY.Z} = \rho_{X-X^*,Y-Y^*}$

Feature Ranking Metrics

Ranking Method – Partial Correlation Coefficient

The partial correlation, provides a measure of the linear relationship between X and Y while controlling for the effect of Z other features on both, X and Y

To use this method we must assume:

- two variables to compare, X and Y
- other variables to control, $Z_{1,\ldots,m-2}$.
- linear relationships between all variables
- no significant outliers
- approximately bivariate normality between the variables

We are in pretty good shape, but we have some departures from bivariate normality.

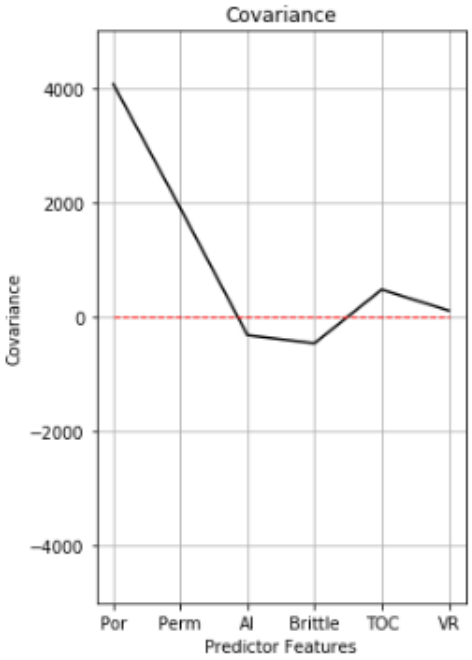
- We apply a Gaussian transform in the demonstration

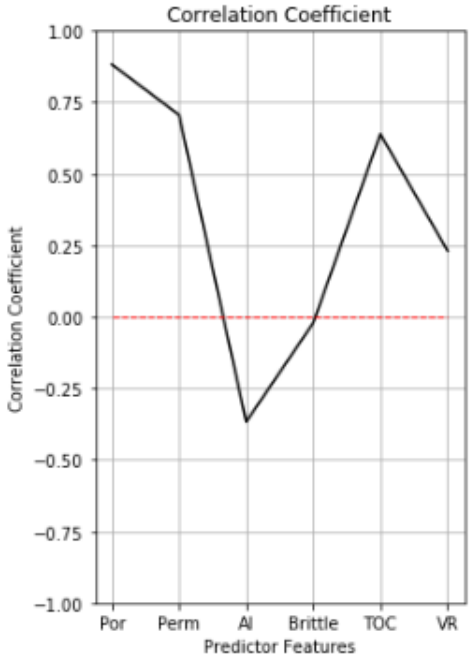
Feature Ranking Metrics

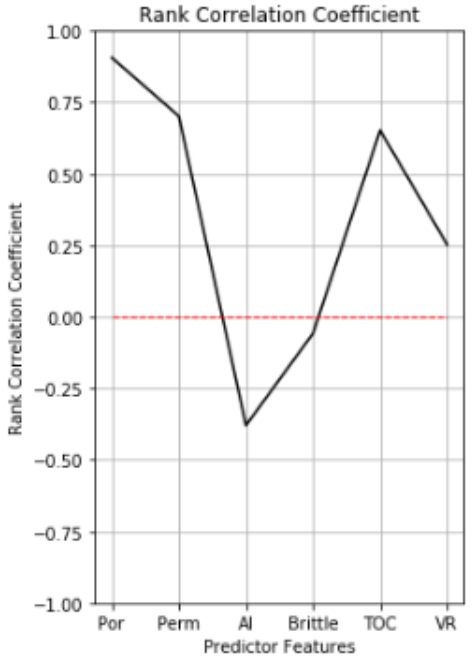
Ranking Methods - Summary

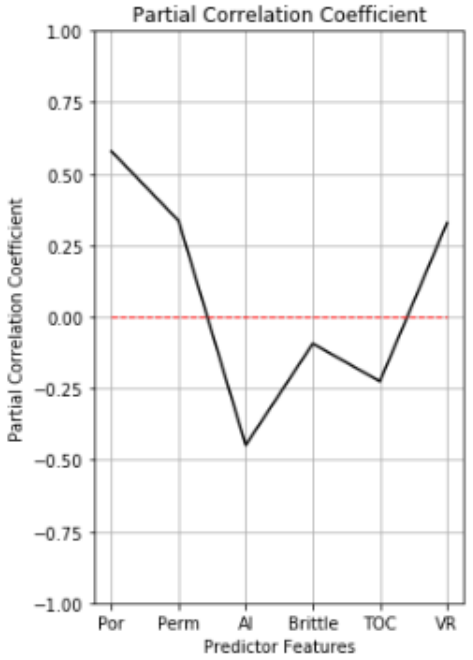
Are we converging on porosity, permeability and vitrinite reflectance as the most important variables with respect to linear relationships with the production?

- What about brittleness?









Feature Ranking Metrics

Ranking Method – Conditional Statistics

- Access the conditional distributions and probabilities of all predictor features given the response feature.

$$F_{X_\alpha \mid Y}(x_\alpha \mid y) = \frac{F_{X_\alpha, Y}(x_\alpha, y)}{F_Y(y)}$$

- We can access the difference between the conditional distributions given low and high case of the response feature.

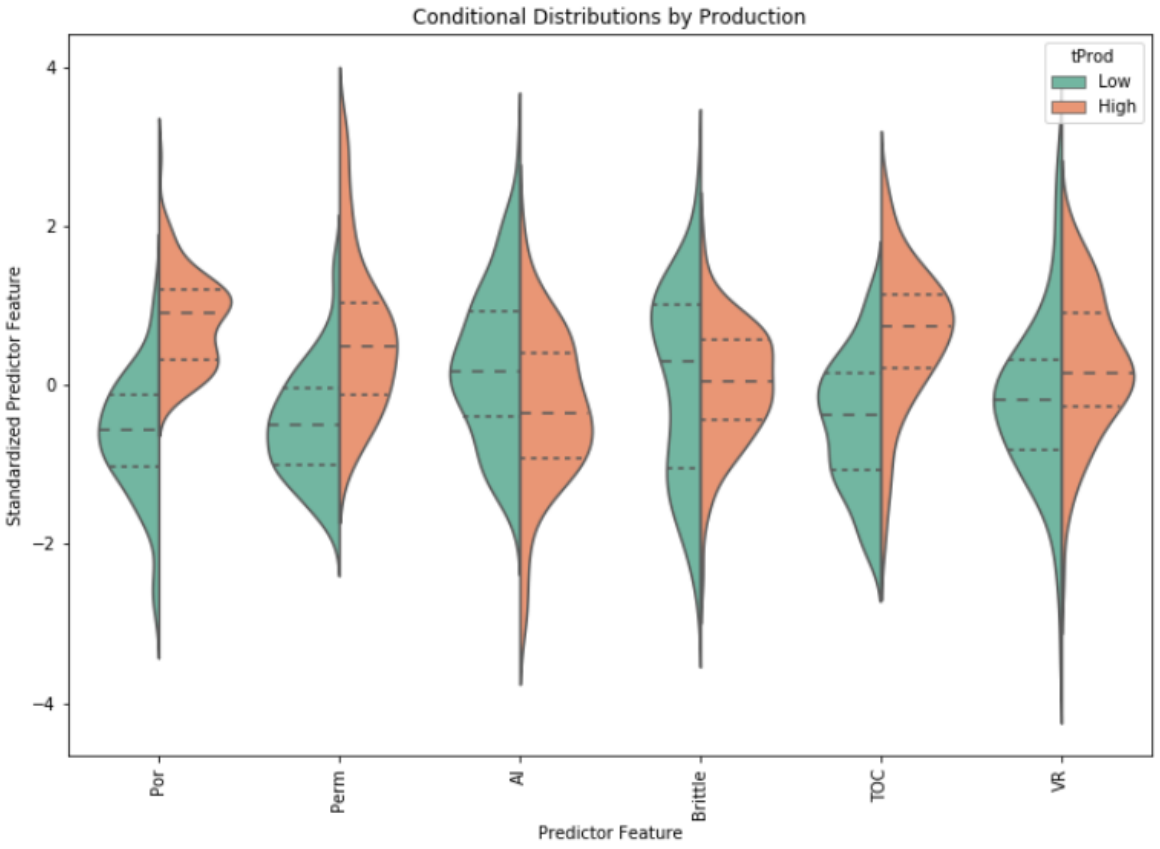
$$F_{X_\alpha \mid Y}(x_\alpha \mid y_{low}) \sim F_{X_\alpha \mid Y}(x_\alpha \mid y_{high})$$

then X_α does not provide information on Y.

Feature Ranking Metrics

Ranking Method – Conditional Statistics

- Standardized each feature, truncate the response feature (if continuous) the build a violin or box plot.



Violin plot for 6 predictor features vs. low and how production rate.

Feature Ranking Metrics

Ranking Method – Model-based Ranking – B coefficients

- We could also consider *B* coefficients from linear regression.

$$Y^* = \sum_{i=1}^{m} B_i X_i + c$$

- These are the linear regression coefficients without standardization of the variables.
- Sensitive to feature variance.
- We are capturing interactions between variables.

Caution: model-based feature ranking is dependent on having a good model. Always check the model prediction accuracy!

Feature Ranking Metrics

Ranking Method – Model-based Ranking – B (beta) coefficients

- We could also consider *B* coefficients from linear regression

$$Y^{s*} = \sum_{i=1}^{m} B_i X_i^s + c$$

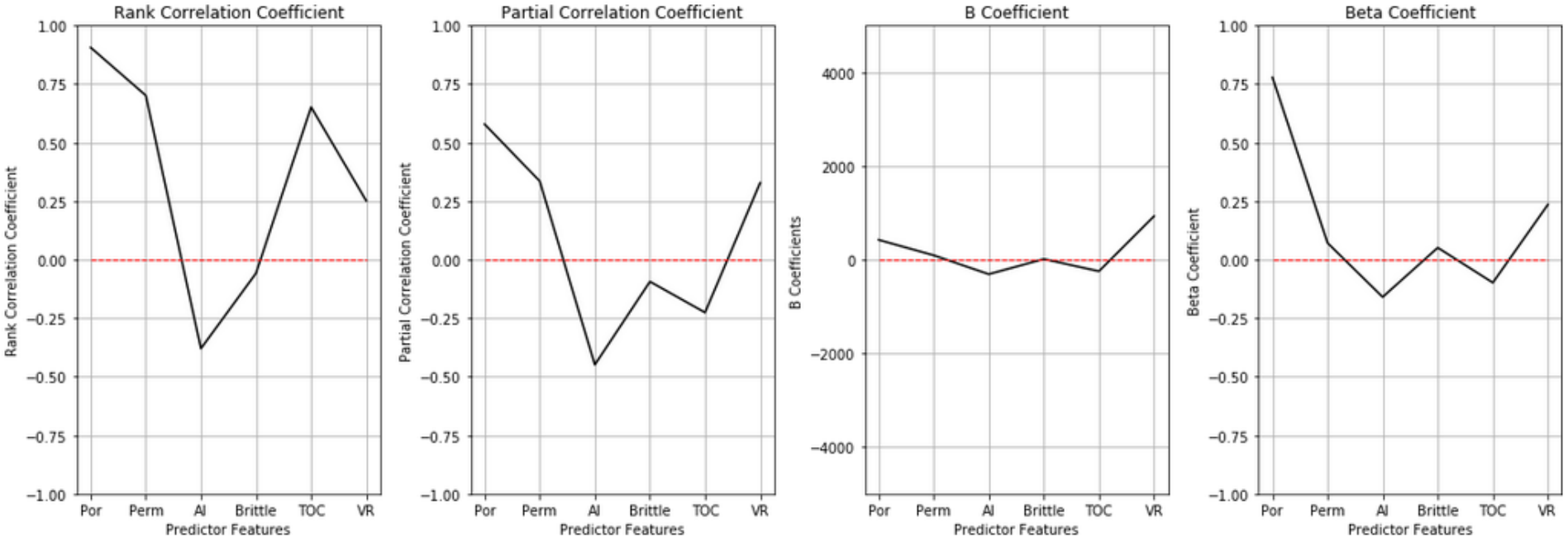
- These are the linear regression coefficients with standardization of the variables, X_i^s and Y^{s*} (variance = 1)
- Not sensitive to variance of the features
- We are capturing interactions between variables.

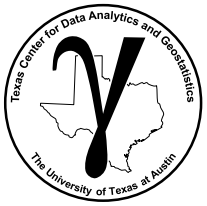
Feature Ranking Metrics

Ranking Methods – Summary of Results

Now what do we see?

- Beta demotes permeability!
- Porosity, acoustic impedance and vitrinite reflectance retain high metrics





PGE 383 Machine Learning

Feature Selection

Lecture outline . . .

- **Mutual Information**

Michael Pyrcz, The University of Texas at Austin

Feature Ranking Metrics

Ranking Method – Mutual Information

- From probability and information theory
  - A more general measure of amount of information from X_α about Y
  - Nonparametric measure without assumption of the form of the relationship
  - Units are ‘Shannons’ / ‘bits’
  - Measure of the difference between the joint $P(x, y)$ and the product of the marginals $P(x) \cdot P(y)$, integrated over all x $\in X$ and y $\in Y$.
  - Leveraging the definition of independence:

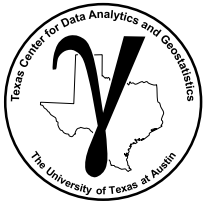
$$P(X, Y) = P(X) \cdot P(Y)$$

Joint and marginal definition of independence.

  - Some form of dependence exists when:

$$P(X, Y) \neq P(X) \cdot P(Y)$$

Joint and marginal definition of dependence.



Feature Ranking Metrics

Mutual Information

- Definition of independence for joint and marginal.

$$P(X,Y) = P(X) \cdot P(Y)$$

Joint and marginal definition of independence.

  - Recall:

$$P(X,Y) = P(Y|X)P(X)$$

Reordered the definition of conditional probability and substitute.

  - After substitution, the expression for conditional and marginal independence:

$$P(Y,X) = P(X) \cdot P(Y)$$

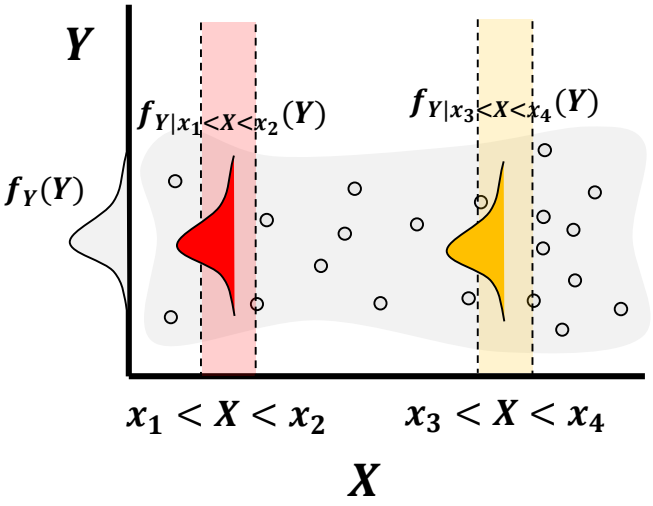
$$P(Y|X) = P(Y)$$

Interpretation:
Knowing something about X tells us nothing about Y!

Feature Ranking Metrics

Mutual Information

- Independence and conditional probabilities.



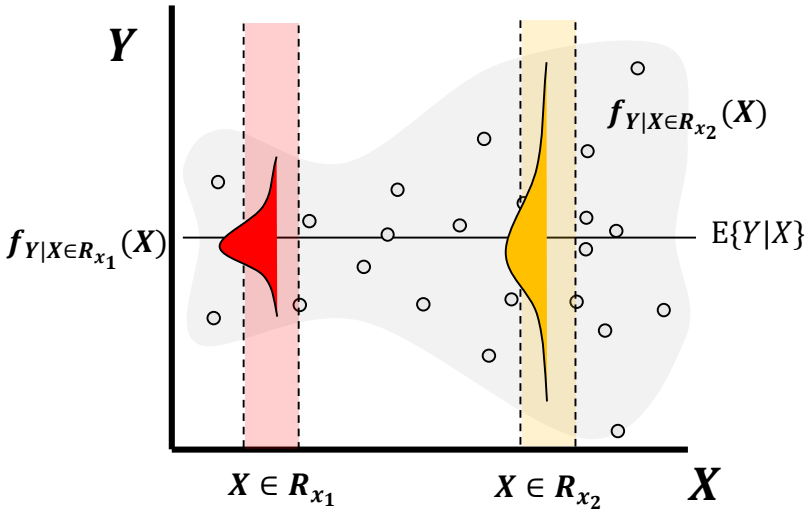
$$f_{Y|x_1<X<x_2}(Y) = f_{Y|x_3<X<x_4}(Y) = \cdots = f_Y(Y)$$

for any choice of bins x_1, x_2, x_3 and x_4.

Feature Ranking Metrics

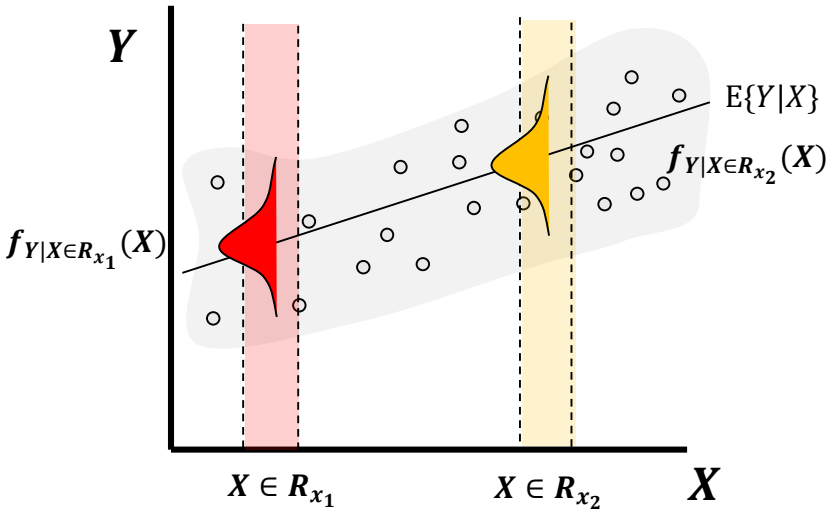
Mutual Information

- Examples of information sharin by visualization of conditional probabilities.



$$E\{Y|X \in R_{x_1}\} = E\{Y|X \in R_{x_2}\}$$

$$Var\{Y|X \in R_{x_1}\} \neq Var\{Y|X \in R_{x_2}\}$$



$$E\{Y|X \in R_{x_1}\} \neq E\{Y|X \in R_{x_2}\}$$

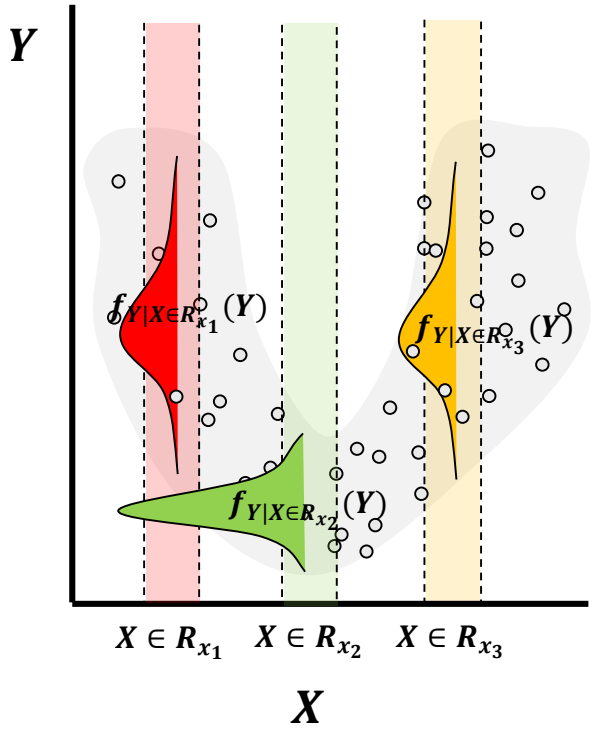
$$Var\{Y|X \in R_{x_1}\} = Var\{Y|X \in R_{x_2}\}$$

- If conditional distribution $f_{Y|X\in R_{x_1}}(y|X \in R_{x_1}) \neq f_{Y|X\in R_{x_2}}(y|X \in R_{x_2})$ for any choice of bins R_{x_1}, R_{x_2}. In more general terms, independence is $f_{Y|X\in R_{x_1}} = f_Y(Y)$

Feature Ranking Metrics

Mutual Information

- Limitation of correlation analysis, assumes monotonic (rank), and linear (Pearson)

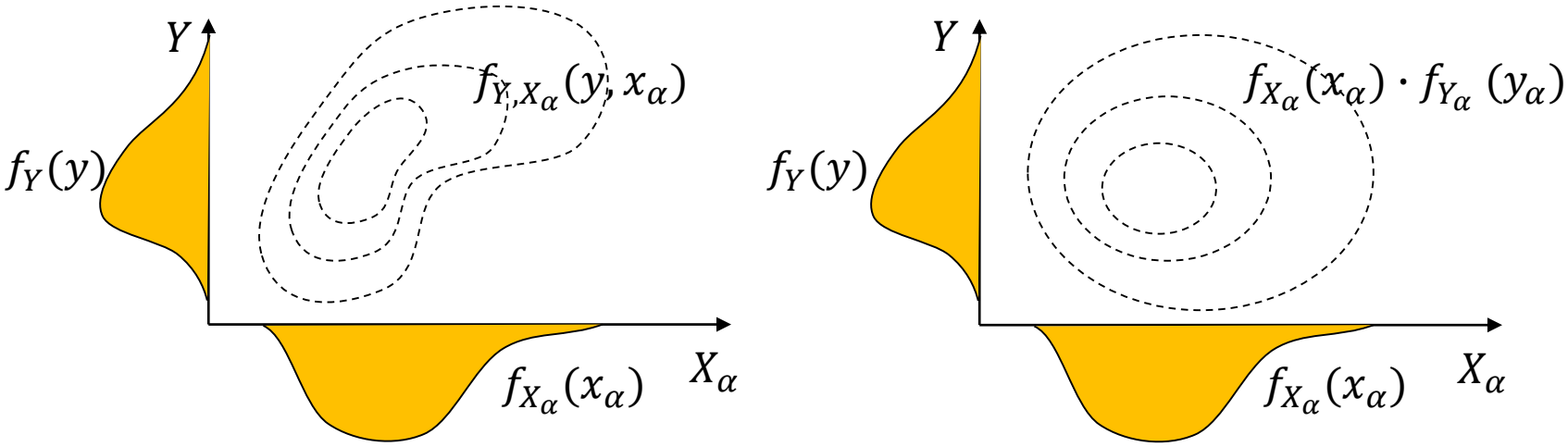


- For a quadratic relationship, $Y = X^2 + c$, highly nonlinear, nonmonotonic
  - $\rho_{X,Y} = 0$ and $\rho_{R_X, R_Y} = 0$! But knowing about X, helps know about Y!

Feature Ranking Metrics

Mutual Information – Proposed non-parametric metric

- Propose a non-parametric model (e.g. no reliance on linearity) given our previous observations
- Compare the actual joint density, $f_{Y,X_\alpha}(y, x_\alpha)$, to joint density assuming independence, $f_{X_\alpha}(x_\alpha) \cdot f_{Y_\alpha}(y_\alpha)$



Schematic of marginal and joint distributions for independence (left) and non-linear dependence (right).

Feature Ranking Metrics

Mutual Information – Proposed non-parametric metric

- Propose a non-parametric model (e.g. no reliance on linearity)
- Compare the actual joint density, $f_{Y,X_\alpha}(y, x_\alpha)$, to joint density assuming independence, $f_{X_\alpha}(x_\alpha) \cdot f_{Y_\alpha}(y_\alpha)$

$$I(X_\alpha; Y) = \int_Y \int_{X_\alpha} P_{X_\alpha,Y}(x_\alpha, y) \cdot log\left(\frac{P_{X_\alpha,Y}(x_\alpha, y)}{P_{X_\alpha}(x_\alpha) \cdot P_Y(y)}\right) dx\, dy$$

Weighting by local density

Measure of mismatch

$$\text{If } \frac{P_{X_\alpha,Y}(x_\alpha,y)}{P_{X_\alpha}(x_\alpha)\cdot P_Y(y)} = 1,\ \log(1) = 0$$

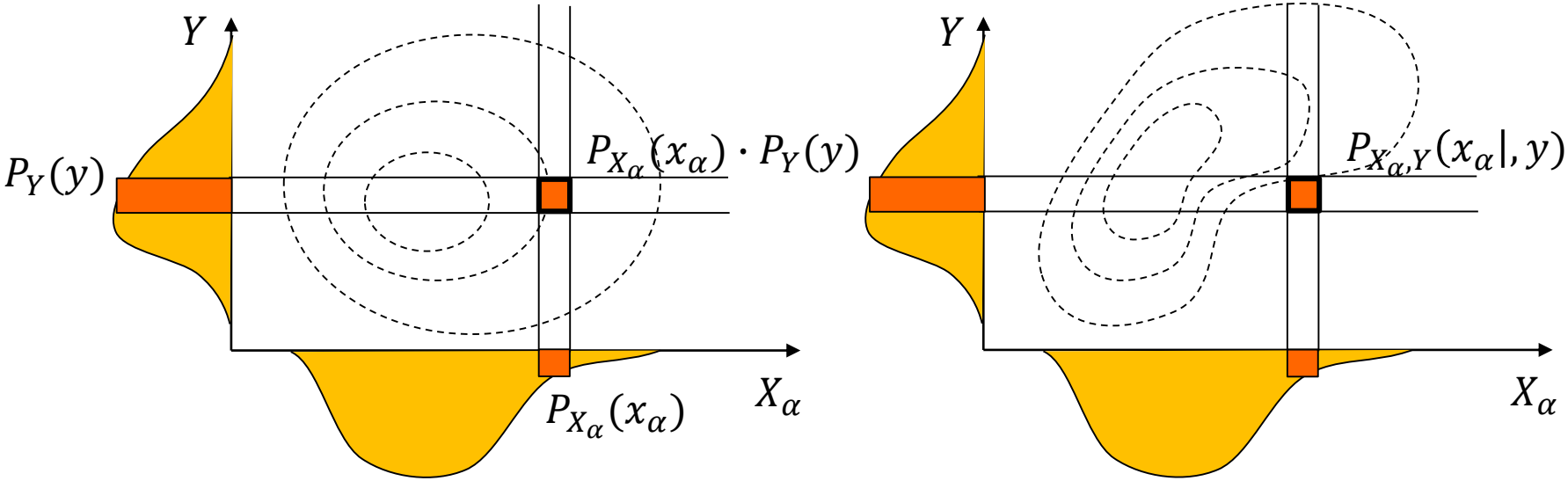
Measure of mismatch between actual joint density and that expected from marginals with independence assumption.

Feature Ranking Metrics

Mutual Information – Practical Calculation

Bin the joint PDF and compare to the expected joint given independence (recall $P(x,y) = P(x) \cdot P(y)$ if independent).

$$I(X_\alpha; Y) = \sum_{y \in Y} \sum_{x \in X_\alpha} P_{X_\alpha,Y}(x_\alpha, y)\, log\left(\frac{P_{X_\alpha,Y}(x_\alpha, y)}{P_{X_\alpha}(x_\alpha) \cdot P_Y(y)}\right)$$



Schematic of marginal and joint distributions for independence (left) and non-linear dependence (right).

Feature Ranking Metrics

Mutual Information

Calculation given continuous Marginal PDFs, $f_{X_\alpha}(x_\alpha)$, $f_Y(y)$ and Joint PDF, $f_{X_\alpha,Y}(x_\alpha, y)$

$$I(X_\alpha; Y) = \int_Y \int_{X_\alpha} f_{X_\alpha,Y}(x_\alpha, y) \cdot log\left(\frac{f_{X_\alpha,Y}(x_\alpha, y)}{f_{X_\alpha}(x_\alpha) \cdot f_Y(y)}\right) dx\, dy$$

Calculation given a sample dataset, with binning decision:

$$I(X_\alpha; Y) = \sum_{y \in Y} \sum_{x \in X_\alpha} P_{X_\alpha,Y}(x_\alpha, y)\, log\left(\frac{P_{X_\alpha,Y}(x_\alpha, y)}{P_{X_\alpha}(x_\alpha) \cdot P_Y(y)}\right)$$

Nonnegativity:

$$I(X_\alpha; Y) \geq 0$$

by Jensens's Inequality

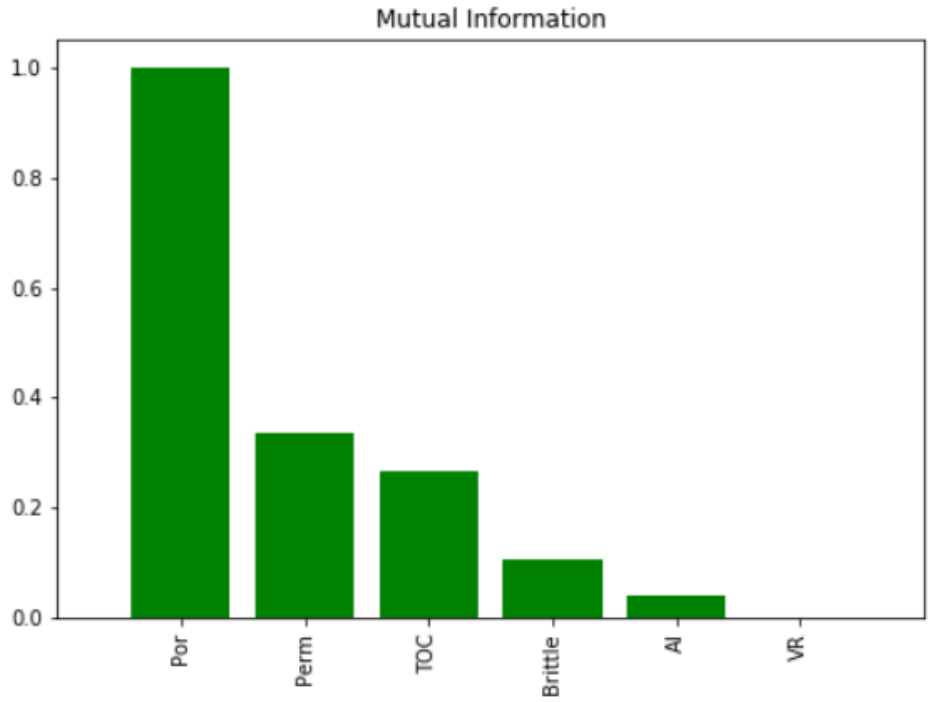
Symmetry:

$$I(X_\alpha; Y) = I(Y; X_\alpha)$$

Feature Ranking Metrics

Ranking Method – Mutual Information

- The simplest method is to select predictor features with the highest mutual information with the response feature.



Normalized mutual information between production and each response feature.
Maximum set to 1.0.

Feature Ranking Metrics

Mutual Information and Entropy

- Entropy, $H(\cdot)$, is a measure of uncertainty about a random variable
- Mutual information may be related to marginal and conditional entropy.

$$H(Y) = -\sum_{y \in Y} P_Y(y) \log\left(P_Y(y)\right) \text{ and } H(X) = -\sum_{x \in X} P_X(x) \log\left(P_X(x)\right)$$

$$H(Y|X) = \sum_{y \in Y} P_{Y|X}(y|x) \log\left(P_{Y|X}(y|x)\right)$$

- We can now define mutual information relative to entropy

$$I(X_\alpha; Y) = H(Y) - H(Y|X_\alpha)$$

Uncertainty reduction in Y given X is known

Uncertainty in Y

Uncertainty in Y given X is known

Feature Ranking Metrics

Mutual Information for Feature Ranking with Max-Dependency Criteria (Peng et al., 2005).

Workflow to find the subset of features, S, p predictor features that maximize mutual information with a response feature, Y.

$$\max I(S,Y), D = \left(\left\{x_{p_1}, i = 1, \dots, p\right\}, Y\right),$$

- For p = 1, must calculate the mutual information for all $I(X_i; Y)$ and select maximum.

$$I(X_{p_1}; Y) = \int_Y \int_{x_{p_1}} f_{X_\alpha, Y}(x_{p_1}, y) \cdot log\left(\frac{f_{X_{p_1}, Y}(x_{p_1}, y)}{f_{X_\alpha}(x_{p_1}) \cdot f_Y(y)}\right) dx_{p_1} dy$$

- For p > 1, use incremental/one at a time trial and error approach, add one feature at a time and calculate. For example given we found the first, X_{p_1}, we find the second:

$$I(X_{p_1}, X_{p_2}; Y) = \int_Y \int_{x_{p_1}} \int_{x_{p_2}} f_{X_{p_1}, X_{p_2}, Y}(x_{p_1}, x_{p_2}, y) \cdot log\left(\frac{f_{X_{p_1}, X_{p_2}, Y}(x_{p_1}, x_{p_2}, y)}{f_{X_{p_1}, X_{p_2}}(x_{p_1}, x_{p_2}) \cdot f_Y(y)}\right) dx_{p_2} dx_{p_1} dy$$

- Note this can be difficult for high dimensional cases as there won't be sufficient data to samples the high dimensional joint distribution.

Feature Ranking Metrics

Mutual Information for Feature Ranking with Max-Dependency Criteria (Peng et al., 2005).

1. Calculate all mutual information, $I(X_i; Y)$, for $i = 1, \ldots, m$ predictor features. **Select the predictor feature, $\boldsymbol{X_{p_1}}$**, with maximum mutual information with response, Y.

2. Calculate all joint mutual information, $I(X_{p_1}, X_i; Y)$, for $i = 1, \ldots, m, i \neq p_1$. **Select the next predictor feature**, $\boldsymbol{X_{p_2}}$,$p_2 \neq p_1$

3. **Continue until p predictor features are selected**, $X_1, \ldots, X_p$.

This is a heuristic, short cut method, that avoids investigation of the full combinatorial of possible feature subsets.

- more on heuristics later

Feature Ranking Metrics

Max Relevance Minimum Redudancy (MRMR) (Peng et al., 2005)

- Peng et al. (2005) suggested a practical approach that avoids the high dimensional $I(\cdot)$:

$$I(X_1, \ldots, X_p; Y)$$

terms while accounting for the relevance and redundancy between predictor features.

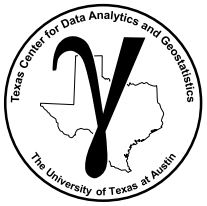
- For continuous predictor features and categorical response, first discretize the feature into a few states, $X_i \rightarrow X_{i,k}$, $Y \rightarrow Y_k$

$$mRMR = \max_{S} \left[\frac{1}{|S|} \sum_{X_i \in S} I(X_i, Y) - \frac{1}{|S|^2} \sum_{X_i \in S} \sum_{X_j \in S} I(X_i, X_j) \right]$$

Relevance of X_i with respect to Y

Redundancy of X_i with each other

where S is the predictor feature subset and $|S|$ is the number of features in the subset S.



PGE 383 Machine Learning

Feature Selection

Lecture outline . . .

- **Shapley Values**

Michael Pyrcz, The University of Texas at Austin

Shapley Value Introduction

Shapley Value Motivation

- **Model-based, Local and Global Feature Importance** by learning contribution of each feature to the prediction
- **Explainable Machine Learning:** complicated models are often required but have low interpretability.

Two choices to improve model interpretability:

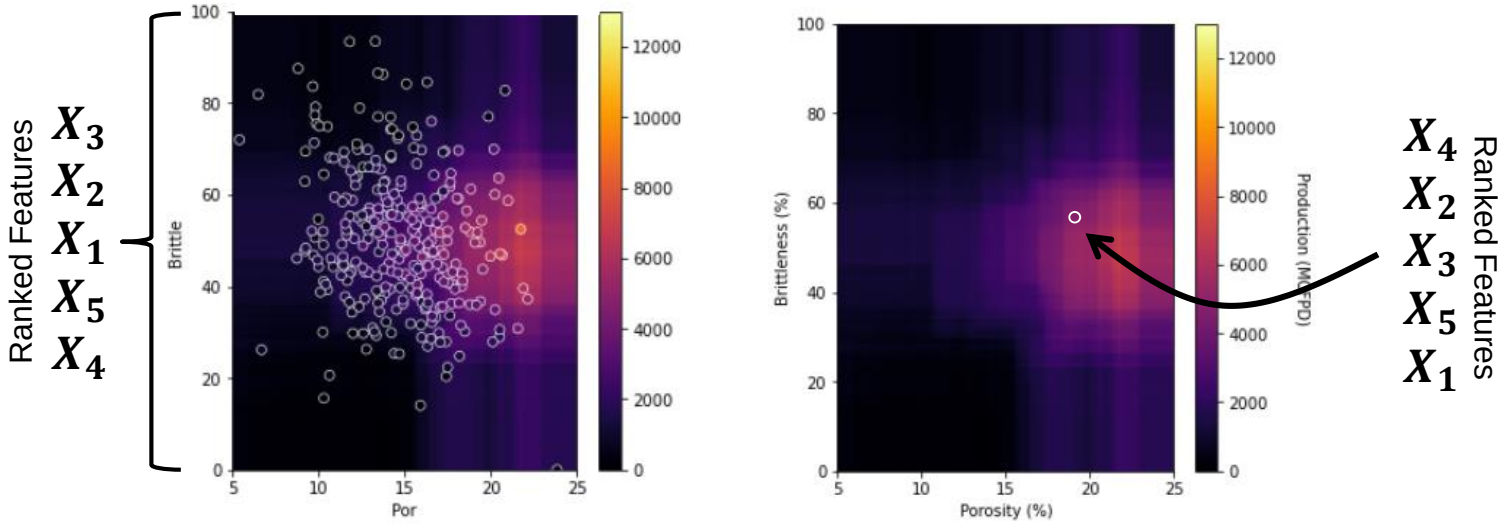
1. reduce the complexity of the models, but may also reduce model accuracy
2. develop improved, agnostic (for any model) model diagnostics

Shapley is a general method for feature importance, based on the model as a black box (model agnostic).

Shapley Value Introduction

Local vs. Global Feature Importance

- **Global Feature Importance** – impact of each feature summarized over multiple locations, testing / withheld data
- **Local Feature Importance** – impact of each feature for a single prediction, 1 testing / withheld datum



Given a model we can calculate global feature importance over various testing, withheld data (left) or local feature importance for one testing datum (right).

- Other model-based feature importance measures summarize over all testing predictions, Global Feature Importance. Shapley is local, but can be summarized for global feature importance.

Shapley Value Introduction

Shapley Values

- **Game theory approach**
  - Calculate the contribution of each predictor feature to push the response prediction away from the mean value of the response over training
  - Based on the Shapley value for allocating resources between 'players' based on a summarization of marginal contributions.

    Dividing up payment between players.
  - Marginal contributions and Shapley values are in units of the response feature.

Marginal Contribution – the impact of one player on the result in a single case. For example:

Marginal Contribution Player 1 = Outcome Player 1 – Outcome No Players

Marginal Contribution Player 1 = Outcome Player 1 & 2 – Outcome Player 2 Only

Shapley Value Introduction

What is the Contribution of Marcus Mumford to Mumford and Sons?

Mumford and sons photo by Alistair Taylor-Young from Universal Publishing Group.

Records Sold = Marcus Solo – ~~No One Playing Music~~ 0

Records Sold = Mumford and Sons with Marcus – Mumford and Sons without Marcus

Shapley Value = Summarization of Marginal Contribution in Record Sold

We could generalize this method to calculate contribution of all band members, but we don't want to encourage further band break ups!

Shapley Value Introduction

Shapley Values

Contribution as the summarization / weighted average of all possible marginal contributions.

2 people work together and make $125,000. How do we split the profits?

$$f(person_1) = \$50{,}000, f(person_2) = \$75{,}000, f(person_1, person_2) = \$125{,}000$$

Allocation person 1: $\frac{1}{2} f(person_1) + \frac{1}{2}\left(f(person_1, person_2) - f(person_2)\right)$

$$\frac{1}{2}(\$50k - 0) + \frac{1}{2}(\$125k - \$75k) = \$50k$$

Allocation person 2: $\frac{1}{2} f(person_2) + \frac{1}{2}\left(f(person_1, person_2) - f(person_1)\right)$

$$\frac{1}{2}(\$75k - 0) + \frac{1}{2}(\$125k - \$50k) = \$75k$$

Solution:

Pay person 1 $50k and person 2 $75k of the total $125k.

Shapley Value Introduction

Shapley Values

Contribution as the summarization / weighted average of all possible marginal contributions.

2 people work together and make $125,000. How do we split the profits?

Synergy

$$f(person_1) = \$50{,}000, f(person_2) = \$70{,}000, f(person_1, person_2) = \$150{,}000$$

Allocation person 1: $\frac{1}{2} f(person_1) + \frac{1}{2}\left(f(person_1, person_2) - f(person_2)\right)$

$$\frac{1}{2}(\$50k - 0) + \frac{1}{2}(\$150k - \$70k) = \$65k$$

Allocation person 2: $\frac{1}{2} f(person_2) + \frac{1}{2}\left(f(person_1, person_2) - f(person_1)\right)$

$$\frac{1}{2}(\$70k - 0) + \frac{1}{2}(\$150k - \$50k) = \$85k$$

Solution:

Pay person 1 $65k and person 2 $85k of the total $150k.

Shapley Value Introduction

Shapley Values

Contribution as the summarization / weighted average of all possible marginal contributions.

2 people work together and make $125,000. How do we split the profits?

Dysergy

$$f(person_1) = \$50{,}000, f(person_2) = \$70{,}000, f(person_1, person_2) = \$90{,}000$$

Allocation person 1: $\frac{1}{2} f(person_1) + \frac{1}{2}\left(f(person_1, person_2) - f(person_2)\right)$

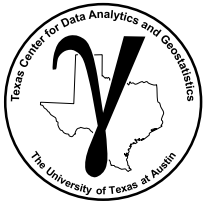
$$\frac{1}{2}(\$50k - 0) + \frac{1}{2}(\$90k - \$70k) = \$35k$$

Allocation person 2: $\frac{1}{2} f(person_2) + \frac{1}{2}\left(f(person_1, person_2) - f(person_1)\right)$

$$\frac{1}{2}(\$70k - 0) + \frac{1}{2}(\$90k - \$50k) = \$55k$$

Solution:

Pay person 1 $35k and person 2 $55k of the total $90k.



Shapley Value Introduction

Shapley Values

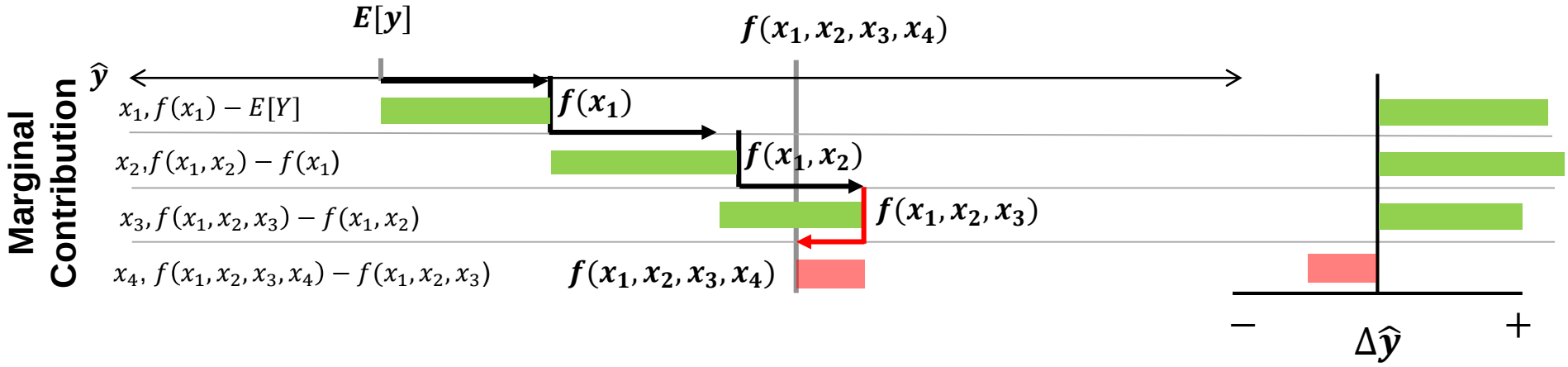
We work out the contribution of each player through summarization over marginal contributions.

- now change player → feature, x_i
- and earnings → model prediction, $f(x)$

Shapley Value Issues

Feature Contribution via Local Feature Importance

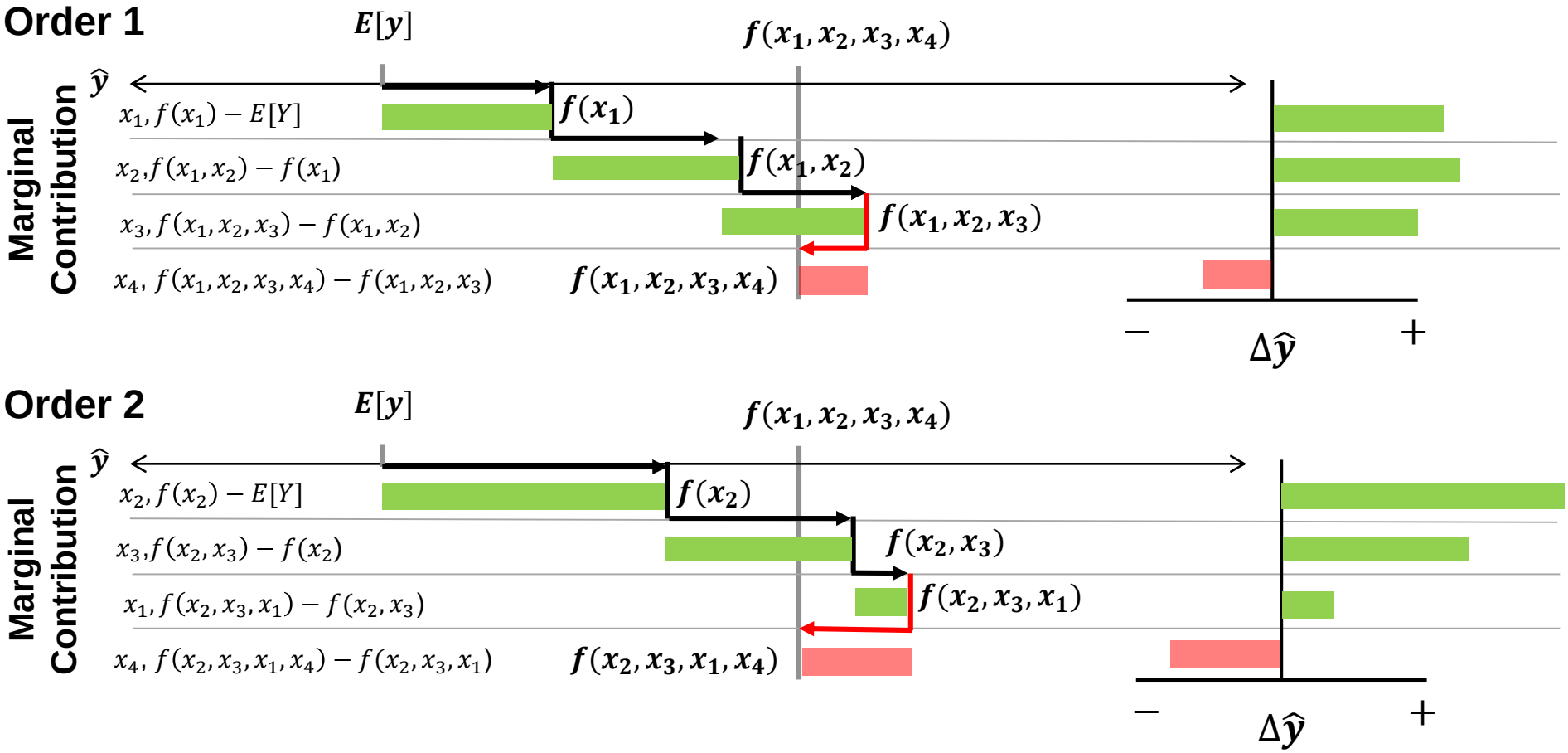
- Local Feature Importance – representing a specific prediction case $(x_1, \ldots, x_m)$.



- Recall $E[y]$ is the expectation of all response training values, i.e. no information from the predictor features.
- Issues:
  1. What if we change the order of adding features?
  2. We do not want to build multiple models, we want to assess feature importance for one specific model, $f(x_1, x_2, x_3, x_4)$.
  3. We may want a global importance for all possible predictions, not a specific case, $x_1, \ldots, x_m$.

Shapley Value Solutions

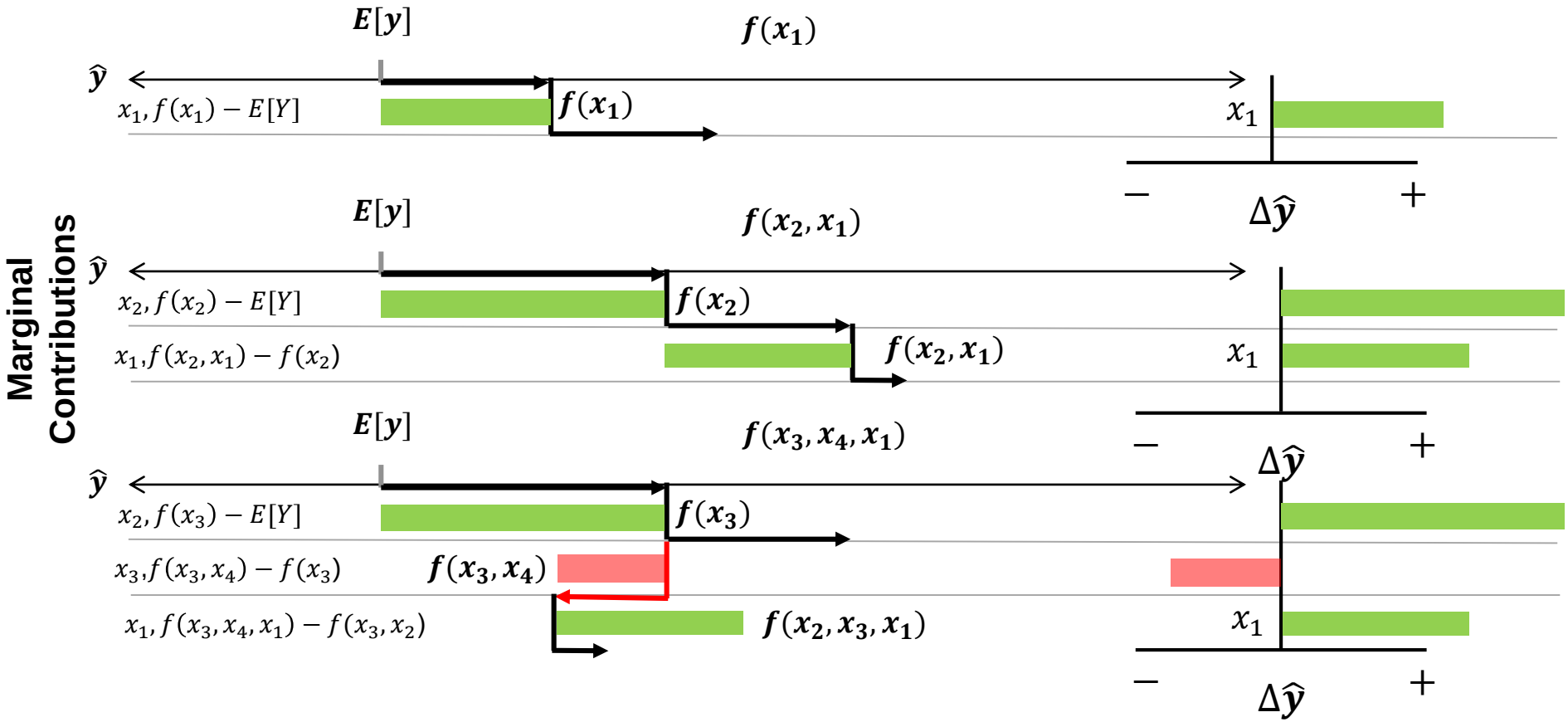
The Combinatorial of Feature Contributions



- Due to interactions between predictor features. The order matters!
- We will average the marginal contribution over the combinatorial of orders.

Shapley Value Solutions

The Combinatorial of Feature Contributions includes these models with 1, 2, 3 feature subsets:



- We will average the marginal contribution over the combinatorial of orders feature subsets.

Shapley Value Solutions

We need to take a single model, $\boldsymbol{f(x_1, x_2, x_3, x_4)}$, and make an estimate for all possible combinations feature subsets!

Note: the **naïve approach** is to train the full combinatorial of models. We don't want to do that if our goal is feature importance to diagnose our model, f. We want to support model explain-ability.

The variety of approaches are similar to imputation methods:

$$\boldsymbol{f(x_1, x_2, x_3\ ) = f(x_1, x_2, x_3, x_4 = E[x_4])}$$

$$\boldsymbol{f(x_1, x_2, x_3\ ) = f(x_1, x_2, x_3, x_4 = P50_{x_4})}$$

There is a unique method with tree-based models:

- Remove $\boldsymbol{x_4}$ by averaging response prediction over all branches with $\boldsymbol{x_4}$.

Shapley Value Solutions

Shapley Equation

Averaging over all possible subsets, orders of marginal contribution

Shapley value, ϕ_i, for the local importance of the i feature:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\ (|F| - |S| - 1)}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right]$$

Our model prediction with i

Our model prediction without i

where:

$|S|$ size of the subset before we add the ith feature

|F| number of features

$\left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right]$ the marginal contribution of i^{th} feature

$\frac{|S|!\ (|F| - |S| - 1)}{|F|!}$ is the weighting by the combinations for this occurrence

$S \subseteq F \setminus \{i\}$ is all possible subsets without i feature, so we can add i

$S \cup \{i\}$ is subset S with i added and S is a subset without i

Shapley Value Solutions

Shapley Equation

Let's explain the weighting applied to each case.

$$\frac{|S|!\ (|F| - |S| - 1)!}{|F|!}$$

Weight by the number same / reorder S cases divided by the total number of possible combinations.

$X_1, X_2 + X_3 = X_2, X_1 + X_3$

So we can just say 2x weight.

Example $F = X_1, X_2, X_3, X_4,\ \ S = X_1, X_2,\ \ i = X_3$

$|S|!\,(|F| - |S| - 1)! : 2\ -\ X_1, X_2, X_3, X_4$ and X_2, X_1, X_3, X_4

$2!\,1! = 2$

$|F|! : 24\ - [X_1, X_2, X_3, X_4], [X_1, X_3, X_2, X_4], \ldots, [X_4, X_3, X_2, X_1]$

$4! = 24$

Shapley Value Solutions

Shapley Equation

$$\sum_{S \subseteq F \setminus \{i\}}$$

We sum over all possible subsets without i, example subsets:

$$i = X_3, |F| = 3$$

Let's assume values, $x_1 = 10\%$, $x_2 = 150\ mD$, $x_3 = 13\%$, $x_4 = 0.54$

$f(X_3 = x_3) - E[y]$ ⟶ $f(X_1 = \overline{x_1}, X_2 = \overline{x_2}, X_3 = x_3) - E[y]$

$f(X_1 = x_1, X_3 = x_3) - f(X_1 = x_1)$ ⟶ $f(X_1 = x_1, X_2 = \overline{x_2}, X_3 = x_3) - f(X_1 = x_1, X_2 = \overline{x_2}, X_3 = \overline{x_3})$

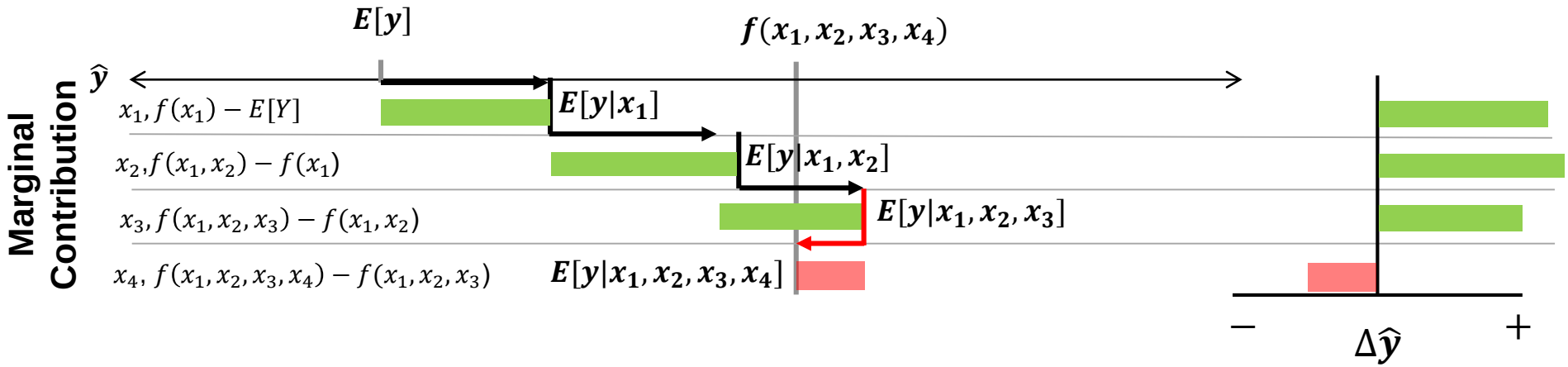
$f(X_1 = x_1, X_2 = x_2, X_3 = x_3) - f(X_1 = x_1, X_2 = x_2)$ ⟶ $f(X_1 = x_1, X_2 = x_2, X_3 = x_3) - f(X_1 = x_1, X_2 = x_2, X_3 = \overline{x_3})$

with feature X_3 without feature X_3

Shapley Value Output

Feature Contribution via Local Feature Importance

- Local Feature Importance – representing a specific prediction case $(x_1, \ldots, x_m)$.



- Recall $E[y]$ is the expectation of all response training values, i.e. no information from the predictor features.

Shapley Value Output

Local Feature Importance

Scatter plot the local feature importance over all feature values, X_i

- Check for local feature importance over a range of values.
- Observed the scatter due to interactions with other features.
- Label with additional features.

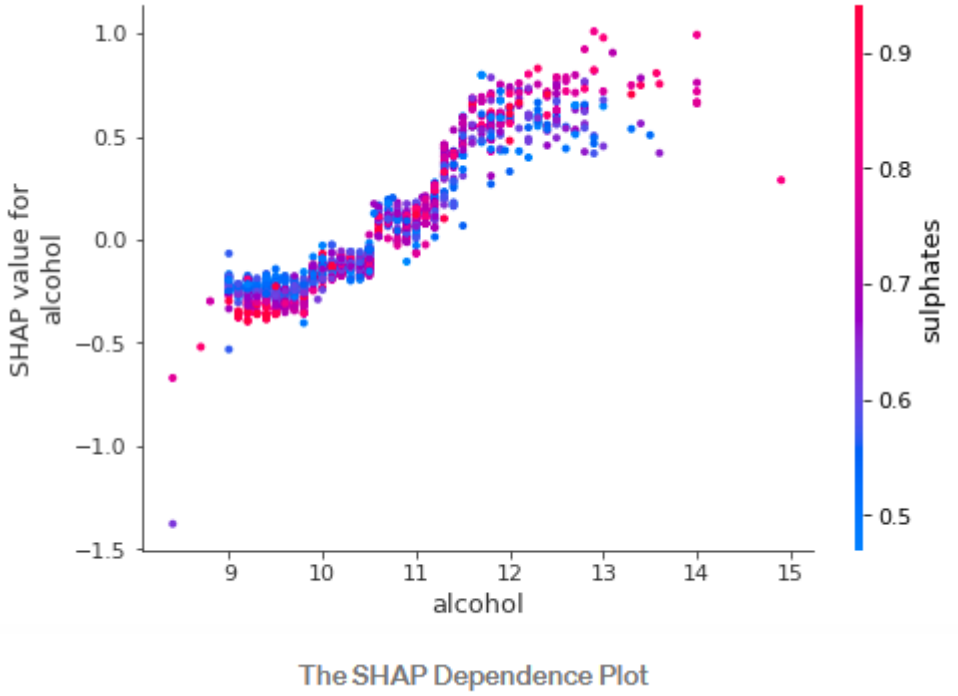


Image from https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d

Shapley Value Output

Global Feature Importance

Plot the local feature importance over all estimates, $\hat{y}$.

- Check for consistent SHAP and feature values.
- Summarize with the average SHAP value over all estimates.

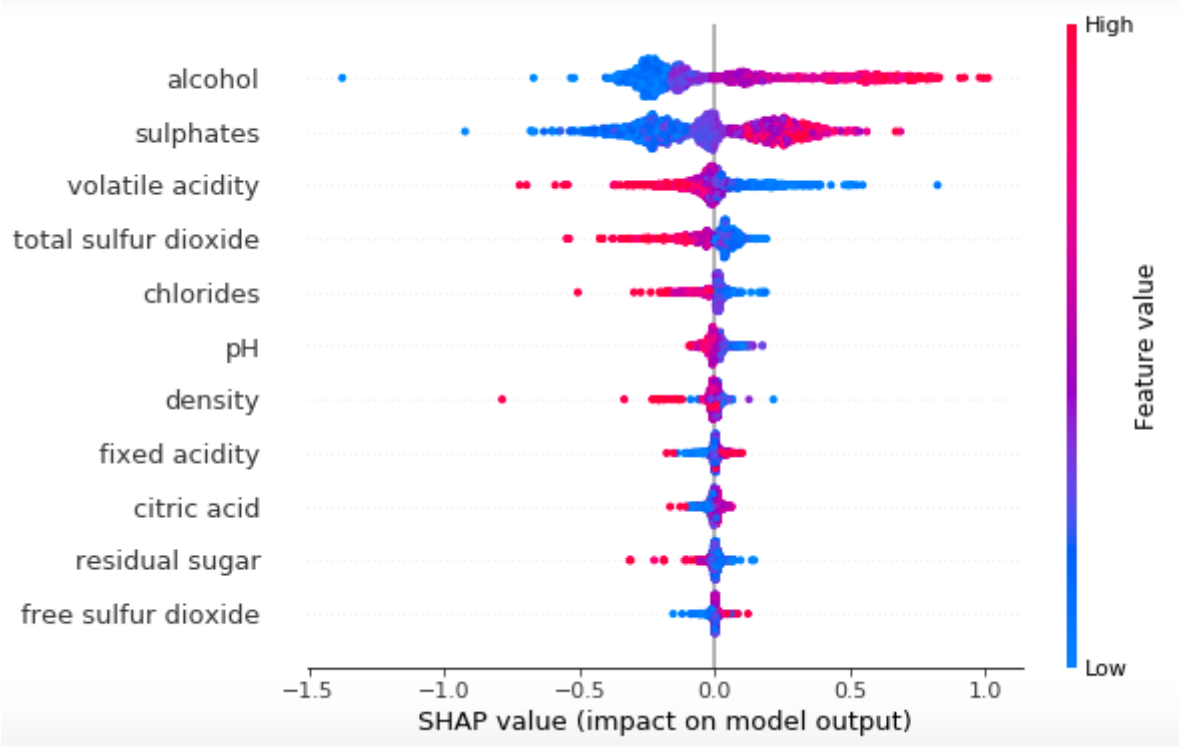


Image from https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d

PGE 383 Machine Learning

Feature Selection

Lecture outline . . .

- **More on Feature Selection**

Michael Pyrcz, The University of Texas at Austin

Feature Ranking Metrics

Ranking Methods – Recursive Feature Elimination

Recursive Feature Elimination (RFE) method works by recursively removing features and building a model with the remaining features.

- build a model with all features, calculate coefficient or feature importance (depending on which is available with the modeling method)
- remove the feature with the lowest coefficient or feature importance and rebuild the model
- repeat the process until only one feature remains
- any model could be used!
- the method assigns rank $1, \ldots, m$ for all features.

Feature Ranking Metrics

Assumptions and Issues

- Selection Bias – with many predictor features it is likely a few are randomly correlated with the response feature
  - they will rank highly and be retained
  - there is a limited sampling of the feature interactions
  - it would require a more complete sampling of predictor feature combinations to establish that the predictor feature is uninformative

- Also, there is a risk of overfit of the model

- More robust if resampling and validation are applied for each model
  - Cross validation and the boostrap

- An alternative is to check all possible combinations of features, this is not feasible if the combinatorial is very large

Feature Ranking Metrics

Ranking Methods – Recursive Feature Elimination

The recursive feature elimination method with a linear regression model provides these ranks:

1. Total Organic Carbon
2. Vitrinite Reflectance
3. Acoustic Impedance
4. Porosity
5. Permeability
6. Brittleness

There has been quite a bit of change from our previous metrics, let’s use a more flexbile model.

Feature Ranking Metrics

Ranking Methods – Recursive Feature Elimination

The recursive feature elimination method with a random forest model provides these ranks:

1. Porosity
2. Brittleness
3. Vitrinite Reflectance
4. Permeability
5. Total Organic Carbon
6. Acoustic Impedance

This is more consistent with our previous results. The advantages with the recursive elimination method:

- the actual model can be used in assessing feature ranks
- the ranking is based on the contribution of each feature to the model

PGE 383 Machine Learning

Feature Selection

Lecture outline . . .

- **Feature Selection Hands-on**

Michael Pyrcz, The University of Texas at Austin

Feature Ranking Demonstration in Python

Demonstration of the wide array approach for feature selection with a documented workflow.

Subsurface Data Analytics

Feature Selection for Subsurface Data Analytics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

Twitter | *GitHub* | *Website* | *GoogleScholar* | *Book* | *YouTube* | *LinkedIn*

Subsurface Machine Learning: Feature Ranking for Subsurface Data Analytics

Here's a demonstration of feature ranking for subsurface modeling in Python. This is part of my Subsuface Machine Learning Course at the Cockrell School of Engineering at the University of Texas at Austin.

Variable Ranking

There are often many predictor features, input variables, available for us to work with for subsurface prediction. There are good reasons to be selective, throwing in every possible feature is not a good idea! In general, for the best prediction model, careful selection of the fewest features that provide the most amount of information is the best practice.

Here's why:

- more variables result in more complicated workflows that require more professional time and have increased opportunity for blunders
- higher dimensional feature sets are more difficult to visualize
- more complicated models may be more difficult to interrogate, interpret and QC
- inclusion of highly redundant and colinear variables increases model instability and decreases prediction accuracy in testing
- more variables generally increase the computational time required to train the model and the model may be less compact and portable
- the risk of overfit increases with the more variables, more complexity

What is Feature Ranking?

Feature ranking is a set of metrics that assign relative importance or value to each feature with respect to information contained for inference and importance in predicting a response feature. There are a wide variety of possible methods to accomplish this. My recommendation is a **'wide-array'** approach with multiple metric, while understanding the assumptions and limitations of each metric.

Here's the general types of metrics that we will consider for feature ranking.

1. Visual Inspection of Data Distributions and Scatter Plots
2. Statistical Summaries
3. Model-based

File SubsurfaceDataAnalytics_Feature_Ranking.ipynb at https://git.io/fjm4p.

Multivariate New Tools

Topic	Application to Subsurface Modeling
Curse of Dimensionality	Reduce problem to lowest dimension possible. *Feature ranking determined that porosity may be predicted from acoustic impedance and rock type alone.*
Feature Selection	Apply wide array methods to explore the importance of each predictor feature with respect to the response feature. *Partial correlation reveals that rock type provides little additional information to acoustic impedance.*

Michael Pyrcz, The University of Texas at Austin

PGE 383 Machine Learning

Feature Selection

Lecture outline . . .

- **Curse of Dimensionality**
- **Feature Selection**
- **Feature Selection Hands-on**

Michael Pyrcz, The University of Texas at Austin