Microsoft Azure

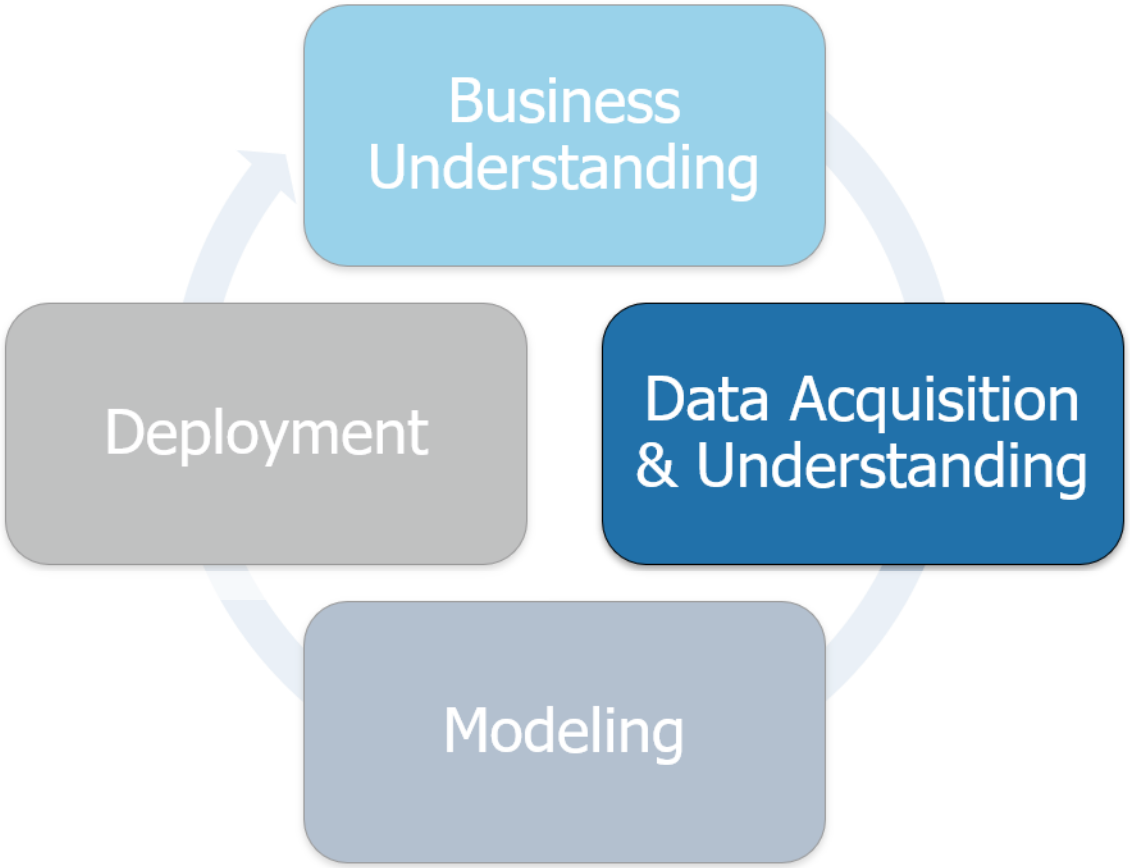
EDSP Mentoring Program

Data Preparation, EDA, and Feature Engineering

Jon Tupitza
Sr. Cloud Solution Architect – Data & Analytics – Azure Data Scientist
Professor of Data Science at University of Virginia
Jon.Tupitza@Microsoft.com

The Microsoft Team Data Science Process

Iterative & Exploratory: Largely Based on Conducting Experiments



Business Understanding
- Identify the Problem Domain
- Identify the Solution Scenario

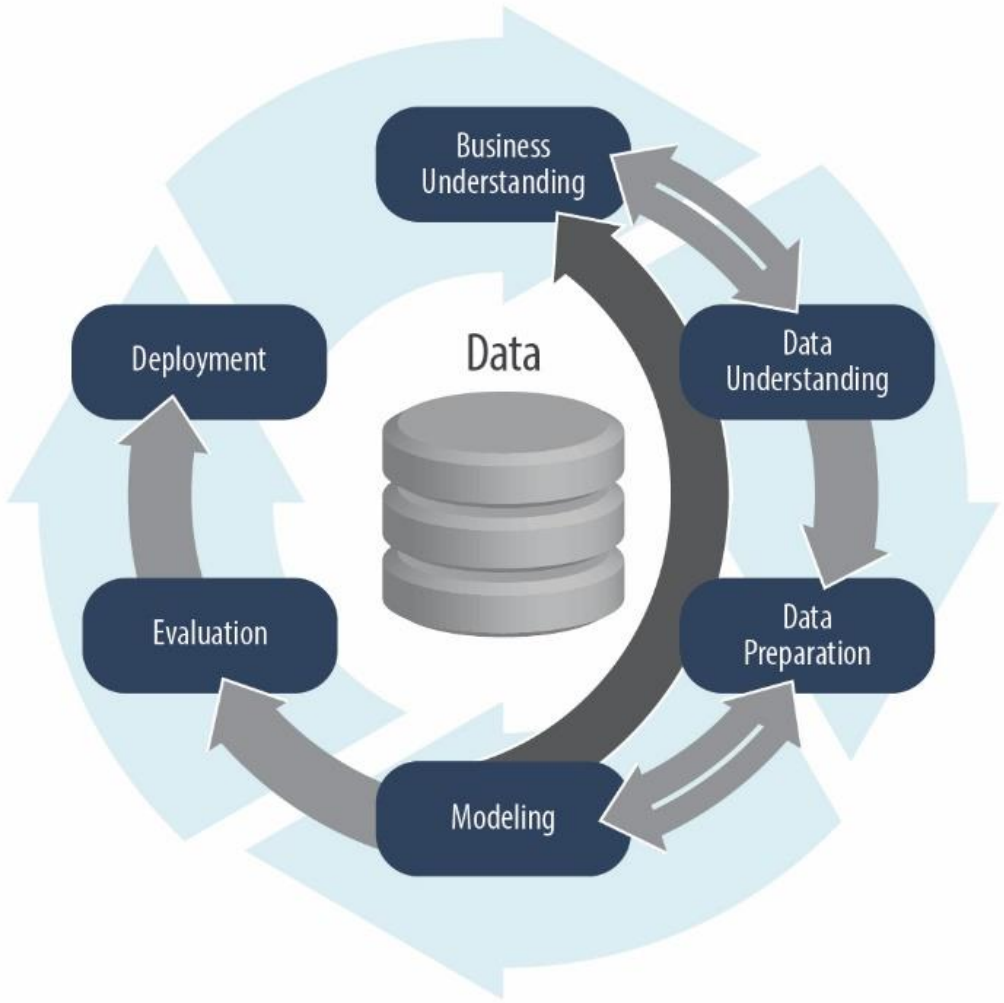
Acquire & Understand Data
- Load, Prepare & Explore Data
- Identify Influential Features

Develop Machine Learning Models
- Select & Engineer Features
- Train, Evaluate & Tune Models

Deployment
- Publish Models as Webservices
- Consume Models Visually and Programmatically

CRISP-DM: Cross-Industry Standard Process-Data Mining

First Introduced in 1996!



Business Understanding
- Identify the Problem Domain
- Identify the Solution Scenario

Data Understanding
- Load and Explore Data
- Identify Influential Features

Data Preparation
- Remove Duplicates & Nulls
- Impute Missing Values
- Select & Engineer Features

Modeling
- Train Models Using a Variety of Algorithms
- Tune Hyper-parameters

Evaluation
- Test Models' Performance & Predictive Power
- Cross-Validate to Appraise Goodness-of-Fit
- Select Most Effective Model for Deployment

Deployment
- Publish Models On-premises or in the Cloud
- Consume Models Visually & Programmatically

Data Profiling: Understanding the Data

Sourcing and Understanding Data that Pertains to the Business Case

Identify the *Shape* of the Dataset:

- Number of Observations (*Rows*) and their Meaning
- Number of Features (*Columns*) and their Meaning
  - DataType of Each Feature (*e.g., Numerical or Categorical*)
  - Cardinality (*Number of Unique Values*) and Rare Categories
  - Distribution of Values (*e.g., Normal, Bernoulli, Binomial, Poisson, Exponential)*
    - Statistics: *Mean, Standard Deviation, Min, Max, Interquartile Range (IQR)*

Identify any *Flaws* in the Data:

- Duplicate Observations *or* Redundant Features
- Missing Data – NULL or NaN (*Not a Number*)
- Inappropriate Data Type Definitions
- Outliers – *Are they erroneous or do they represent the general truth?*

Data Understanding: Define the Unit of Analysis

What Does Each Row (Observation) Represent?

Loan Risk Data

Education	Married	Purpose	Late Payments	Annual Income	Loan Is Bad
1	Yes	Car Purchase	0	$107,000	0
3	No	Small Business	3	$99,000	1
1	Yes	House Purchase	5	$85,000	1
2	No	Marriage	1	$72,000	0
2	No	Debt Consolidation	0	$120,000	0

Data Understanding: Define the Target

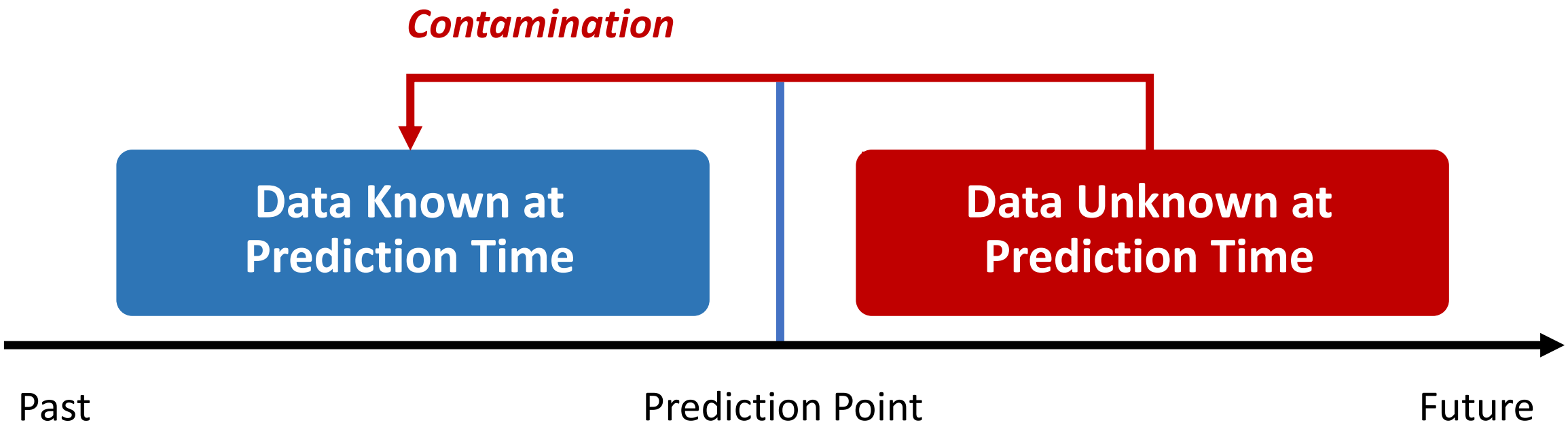
What Do You Want to Predict?

Loan Risk Data

Education	Married	Purpose	Late Payments	Annual Income	Loan Is Bad
1	Yes	Car Purchase	0	$107,000	0
3	No	Small Business	3	$99,000	1
1	Yes	House Purchase	5	$85,000	1
2	No	Marriage	1	$72,000	0
2	No	Debt Consolidation	0	$120,000	0

Data Understanding: Target Leakage

Data Not Known at the Time of Prediction



Data Understanding: Target Leakage

Data Not Known at the Time of Prediction

Loan Risk Data

Education	Married	Purpose	Late Payments	Annual Income	Loan Is Bad
1	Yes	Car Purchase	0	$107,000	0
3	No	Small Business	3	$99,000	1
1	Yes	House Purchase	5	$85,000	1
2	No	Marriage	1	$72,000	0
2	No	Debt Consolidation	0	$120,000	0

Prepare Data, Engineer & Select Features

Data Science & Machine Learning Development

Data Preparation: Feature Engineering

- Impute Missing Values
- Date/Time Parsing
- Scaling
- Encoding
- Generalization
- Discretization (Binning)
- Credibility Estimates

DateOfSale	DayOfWeek	MonthOfYear	IsWeekend
03/07/2020	Saturday	March	1
06/07/2020	Sunday	June	1
09/02/2020	Wednesday	September	0
10/08/2020	Thursday	October	0
12/07/2020	Monday	December	0

Age

Age Group

DateOfBirth	Age	Age Group
02/13/1969	51	Fifties
03/05/1972	48	Forties
04/11/1984	36	Thirties
05/21/1995	25	Twenties
06/24/2002	18	Teens

Feature Engineering: Numerical Features

- **Impute** missing values and create a flag to indicate imputed values:
  - Mean, Median, Mode
  - Custom or Calculated (e.g., MICE)

Price	Mean	Median	Imputed Flag
7	7	7	0
null	7	5	1
5	5	5	0
NaN	7	5	1
9	9	9	0

- **Scaling:**
  - **Standardize**: Rescale so mean (μ) = 0 and Standard Deviation (σ) = 1

$$z = (x_i - \mu)/\sigma$$

  - **Normalize**: Rescale so the range falls between 0 and 1

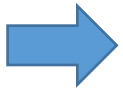
$$z = (x - \min(x) \,/\, \max(x) - \min(x)$$

Feature Engineering: Encoding Categorical Features

Because machine learning algorithms cannot interpret text data, categorical features must first be transformed (encoded) into numerical values.

One-Hot Encoding

Group
A
B
C
A
A

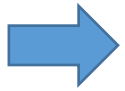


Group A	Group B	Group C
1	0	0
0	1	0
0	0	1
1	0	0
1	0	0

aka, Dummy Features

Count Encoding

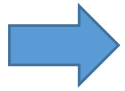
Group
A
B
C
A
A



Count
3
1
1
3
3

Ordinal Encoding

Group
A
B
C
A
A

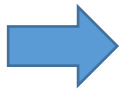


Ordinal
0
1
2
0
0

Feature Engineering: Credibility Estimates

Categorical features often have unequal member distributions.
Credibility estimates help compensate for this imbalanced representation.

Target	Group	Credibility Estimate
0	A	3 * (0.33 – 0.4)
0	B	1 * (0 – 0.4)
1	C	1 * (0 – 0.4)
1	A	3 * (0.33 – 0.4)
0	A	3 * (0.33 – 0.4)



The more of a value we observe in a group, the more we trust that group's deviation from the overall mean.

$$count_k \times (\bar{y}_k - \bar{y})$$

Feature Selection: Partial Dependence

A Univariate Method: The influence of each feature's influence is measured before modeling

Iteratively sets all observations in the column to each unique value contained in that column.

Then

Then observe the correlation each value of the column has to the response variable (target).

Original Values

Loan	Income	IsBad
$11,200	$108,000	False
$10,000	$65,000	False
$8,000	$20,000	True
$16,000	$110,000	True
$4,000	$155,000	False

All Values Set to $20k

Loan	Income	IsBad
$11,200	**$20,000**	False
$10,000	**$20,000**	False
$8,000	**$20,000**	True
$16,000	**$20,000**	True
$4,000	**$20,000**	False

All Values Set to $65k

Loan	Income	IsBad
$11,200	**$65,000**	False
$10,000	**$65,000**	False
$8,000	**$65,000**	True
$16,000	**$65,000**	True
$4,000	**$65,000**	False

Feature Selection: Permutation Importance

A Model-based Method: The performance of the model is measured before and after…

Randomly shuffling the values in each column, one-at-a-time, to break the correlation that each column has to the target variable

Then

Measuring the impact the change to each column's influence has upon the model's overall performance according to one of many applicable metrics:

Generates a stack-ranked list of features by their scores

Classification: Accuracy, Precision, or Recall
Regression: MAE, RMSE, RAE, RSE, R^2

Original Values

Loan	Income	IsBad
$11,200	$108,000	False
$10,000	$65,000	False
$8,000	$20,000	True
$16,000	$110,000	True
$4,000	$155,000	False

Shuffle Column 1

Loan	Income	IsBad
$8,000	$108,000	False
$4,000	$65,000	False
$16,000	$20,000	True
$10,000	$110,000	True
$11,200	$155,000	False

Shuffle Column 2

Loan	Income	IsBad
$11,200	**$65,000**	False
$10,000	**$155,000**	False
$8,000	**$108,000**	True
$16,000	**$20,000**	True
$4,000	**$110,000**	False