

MMCI MACHINE LEARNING

Lab

MACHINE LEARNING LAB

- Founded by economist Anthony Goldbloom (2010)
- Platform to host predictive modeling competitions for companies (e.g. NASA)
- Sold to google
- kaggle
- Note for several projects:
 - Data (size)
 - Kernels
 - Evaluation

kaggle

MACHINE LEARNING LAB

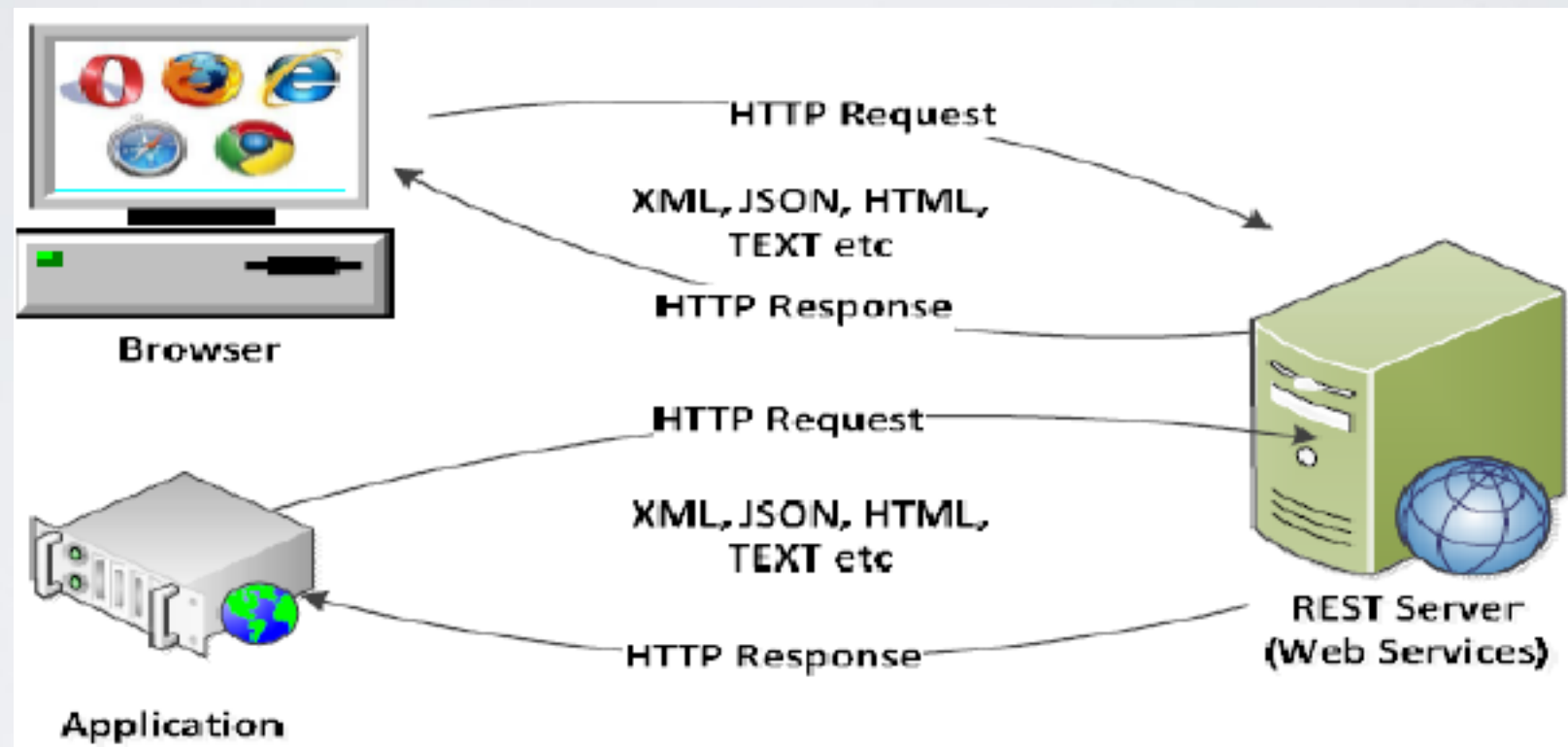
- 2nd Annual MMCI Predictive Modeling Competition
- Analytic:
 - Cohort(s)
 - Endpoint(s)
 - Predictor(s)
- Project:
 - Data Storage
 - Data Compute
 - Results



DATA DELIVERY (IDEAL)

- PROBLEM: Create algorithm that allows us to predict the number of repositories for an MMCI cohort.
- List the “points of the triangle” (cohort, endpoint, predictors).
- GitHub API:
 - Cohort: <https://api.github.com/repos/Duke-Translational-Bioinformatics/mmci-practical-datascience/forks>
 - Endpoint: <https://api.github.com/users/benneely>
 - Predictors: <https://api.github.com/users/benneely>

WEB SERVICE IN ACTION



DATA DELIVERY (LAB)

- Problem: Create an algorithm that allows Duke Endocrinology department to predict readmissions.
- Cohort: <fill in>
- Endpoint: <fill in>
- Predictors: <fill in>

STORAGE/COMPUTE/RESULTS

- Microsoft Azure Machine Learning Studio (google)
- Set a bookmark here.
- Interactively Explore Interface.



DATA STORAGE

- Data Sets
- + New
- From Local File
- Choose File from disk
- Add description
- OK



DATA EXPLORATION

- Open in Notebook
- Adult Census Income:
 - Problem: Create Algorithm that predicts adult income using the 2010 Census Data.
- Questions:
 - “Triangle”: Cohort / Endpoints / Predictors?
 - Missingness?
 - Data Representation?
 - Clinical Relevance?
 - What Kind of models can I use? What makes sense?

DATA EXPLORATION

- Open in Notebook
- Problem: Create an algorithm that allows Duke Endocrinology department to predict readmissions.
 - Questions:
 - “Triangle”: Cohort / Endpoints / Predictors?
 - Missingness?
 - Data Representation?
 - Clinical Relevance?
 - What is the sample size?
 - What is the event rate?



MODELING PIPELINE (DEMO)

`sklearn.pipeline.Pipeline`

`class sklearn.pipeline.Pipeline(steps, memory=None)`

[\[source\]](#)

Pipeline of transforms with a final estimator.

Sequentially apply a list of transforms and a final estimator. Intermediate steps of the pipeline must be 'transforms', that is, they must implement `fit` and `transform` methods. The final estimator only needs to implement `fit`. The transformers in the pipeline can be cached using `memory` argument.

The purpose of the pipeline is to assemble several steps that can be cross-validated together while setting different parameters. For this, it enables setting parameters of the various steps using their names and the parameter name separated by a `'_'`, as in the example below. A step's estimator may be replaced entirely by setting the parameter with its name to another estimator, or a transformer removed by setting to `None`.

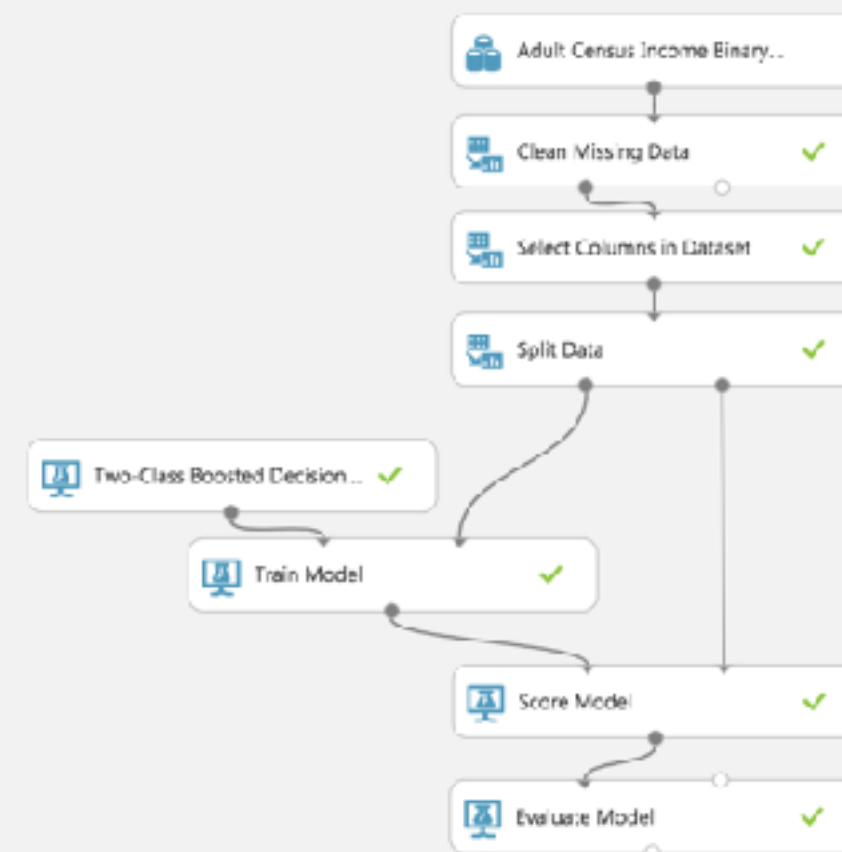
Read more in the [User Guide](#).

Parameters: `steps` : list

List of (name, transform) tuples (implementing `fit/transform`) that are chained, in the order in which they are chained, with the last object an estimator.

`memory` : None, str or object with the `joblib.Memory` interface, optional

Used to cache the fitted transformers of the pipeline. By default, no caching is performed. If a string is given, it is the path to the caching directory. Enabling caching triggers a



EXPERIMENTS

- + New
- Blank Experiment



EXPERIMENTS

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The top navigation bar includes the logo, the text "Microsoft Azure Machine Learning Studio", the workspace name "Ben Neely-Free-Workspace", and icons for help, collaboration, feedback, and user profile.

The left sidebar contains a search bar labeled "Search experiment items" and a list of categories with expandable arrows:

- Saved Datasets
- Transforms
- Data Format Conversions
- Data Input and Output
- Data Transformation
- Feature Selection
- Machine Learning
- OpenCV Library Modules
- Python Language Modules
- R Language Modules
- Statistical Functions
- Text Analytics
- Time Series

The main workspace area is titled "Experiment created on 6/22/2018" and "In draft". It contains a central canvas with the instruction "To create your experiment, drag and drop datasets and modules here". A dashed box on the canvas is labeled "Drag Items Here" with an arrow pointing to it. Below this, a "Mini Map" window is visible. The canvas shows a workflow diagram with three dashed boxes connected by arrows, representing a sequence of steps in the experiment.

The right sidebar contains the "Properties" and "Project" tabs. Under "Properties", there are sections for "Experiment Properties" (showing "STATUS CODE" as "InDraft") and "Summary" (with a text area for a description). Below these is a "Description" section with another text area. At the bottom of the sidebar is a "Quick Help" section.



EXPERIMENTS

Define Graph: (1) Load Data (2) Pare Down Columns **Node**

The screenshot displays the Orange3 data mining software interface. The main workspace is titled "2018 MMCI Diabetic" and shows a workflow graph with two nodes: "training.csv" (a data source node) and "Select Columns in Dataset" (a data transformation node). The nodes are connected by a vertical arrow. The interface includes a left sidebar with a search bar and a list of data transformation nodes under the "Manipulation" category. A "Mini Map" window at the bottom left shows a smaller version of the workflow graph. The right sidebar contains the "Properties" and "Project" panels, which include sections for "Experiment Properties", "Summary", and "Description". The "Experiment Properties" section shows fields for "START TIME", "END TIME", "STATUS CODE", and "STATUS DETAILS". The "Summary" section has a text area for describing the experiment. The "Description" section has a text area for a detailed description. The "Quick Help" section is also visible at the bottom of the right sidebar.



DATA EXPLORATION (TAKE 2)

- Open your experiment
- Remember the Problem: Create an algorithm that allows Duke Endocrinology department to predict readmissions.
 - Use Azure Machine Learning Studio to complete:
 - Write down variables you choose for your models
 - Make note of:
 - Missingness
 - Data Representation
 - Don't forget you need to think about the “triangle”!



EXPERIMENTS

Missing Data Treatment / Imputation

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The top bar shows the workspace name 'Ben Neely-Free-Workspace'. The left sidebar contains a search bar and a list of datasets under 'My Datasets' and 'Samples'. The main workspace area shows a workflow titled '2018 MMCI Diabetic' with the status 'Finished running'. The workflow consists of three steps: 'training.csv', 'Select Columns in Dataset', and 'Clean Missing Data'. The 'Clean Missing Data' step is highlighted with a blue border and numbered '1' and '2'. The right sidebar shows the 'Properties' panel for the 'Clean Missing Data' step, with settings for 'Columns to be cleaned', 'Minimum missing value ratio', 'Maximum missing value ratio', 'Cleaning mode', and 'Replacement value'. The 'Columns to be cleaned' section shows 'Selected columns: race,gender,weight' and a 'Launch column selector' button. The 'Minimum missing value ratio' is set to '0' and the 'Maximum missing value ratio' is set to '1'. The 'Cleaning mode' is set to 'Custom substitution value' and the 'Replacement value' is set to 'n'. A 'Quick Help' section at the bottom of the right sidebar provides additional information about handling missing values.

Microsoft Azure Machine Learning Studio

Ben Neely-Free-Workspace

2018 MMCI Diabetic

Finished running

training.csv

Select Columns in Dataset

Clean Missing Data

1 2

Properties Project

Clean Missing Data

Columns to be cleaned

Selected columns: race,gender,weight

Launch column selector

Minimum missing value ratio

0

Maximum missing value ratio

1

Cleaning mode

Custom substitution value

Replacement value

n

Quick Help

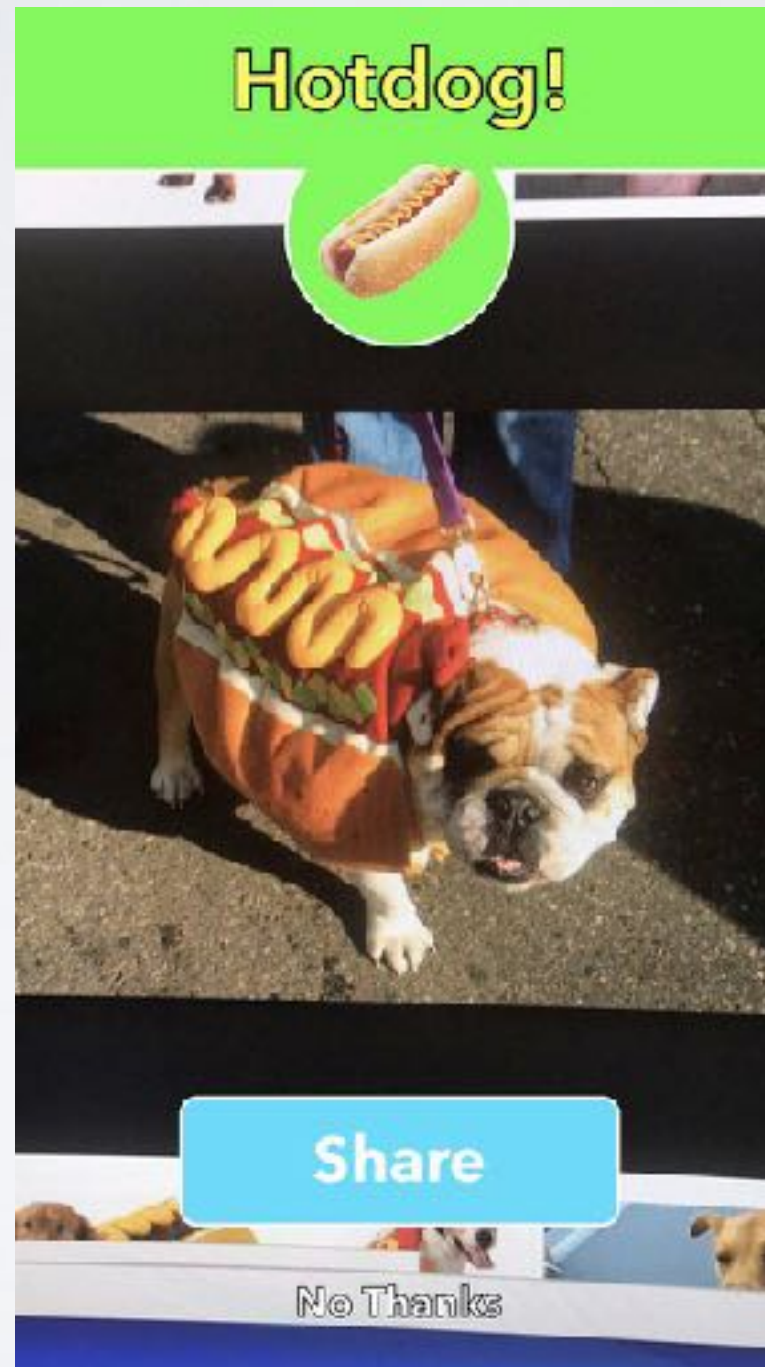
Specifies how to handle the values missing from a dataset

(more help...)

- different strategies for different columns?
- what is the right amount of missingness?
- are these default methods enough?



VALIDATION



VALIDATION

- Internal
 - Cross-Validation
 - Split-Sample
- External
 - Geographic / Temporal
 - Fully Independent

EXPERIMENTS

Split Sample Data

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The top navigation bar includes the title "Microsoft Azure Machine Learning Studio", the workspace name "Ben Neely-Free-Workspace", and icons for help, collaboration, and user profile. The left sidebar contains a search bar with the text "split" and a list of categories: "Saved Datasets" (with "Restaurant ratings" selected), "Data Transformation", and "Sample and Split" (with "Partition and Sample" and "Split Data" listed). The main workspace area is titled "2018 MMCI Diabetic" and shows a workflow diagram with four steps: "training.csv", "Select Columns in Dataset" (marked with a green check), "Clean Missing Data" (marked with a green check), and "Split Data" (marked with a blue check). The "Split Data" step is highlighted with a blue border and numbered 1 and 2. The right sidebar contains the "Properties" panel for the "Split Data" step, showing settings for "Splitting mode" (Set Rows), "Fraction of rows in the first output dataset" (0.8), "Randomized split" (checked), "Random seed" (0), and "Stratified split" (False). A "Quick Help" section at the bottom right provides a brief description of the step. The bottom toolbar includes icons for "NEW", "RUN HISTORY", "SAVE", "SAVE AS", "DISCARD CHANGES", "RUN", "SET UP WORKSPACE", and "PUBLISH TO GALLERY".

Microsoft Azure Machine Learning Studio

Ben Neely-Free-Workspace

2018 MMCI Diabetic

In draft

Draft saved at 1:56:08 PM

split

Search

Saved Datasets

- Samples
 - Restaurant ratings
- Data Transformation
- Sample and Split
 - Partition and Sample
 - Split Data

training.csv

Select Columns in Dataset

Clean Missing Data

Split Data

1 2

Properties Project

Split Data

Splitting mode

Set Rows

Fraction of rows in the first output dataset

0.8

☒ Randomized split

Random seed

0

Stratified split

False

Quick Help

Split the rows of a dataset into two distinct sets
([more help...](#))

+ NEW

RUN HISTORY

SAVE

SAVE AS

DISCARD CHANGES

RUN

SET UP WORKSPACE

PUBLISH TO GALLERY



TRAINING

Regression

Two-Class Logistic Regression

Two-Class Bayes Point Machine

Classification

Two-Class Decision Jungle

Ordinal Regression

Poisson Regression

Linear Regression

Two-Class Averaged Perceptron

Clustering

Two-Class Boosted Decision Tree

Decision Forest Regression

Anomaly Detection

Two-Class Support Vector Machine

Two-Class Locally-Deep Support Vec...

Two-Class Neural Network

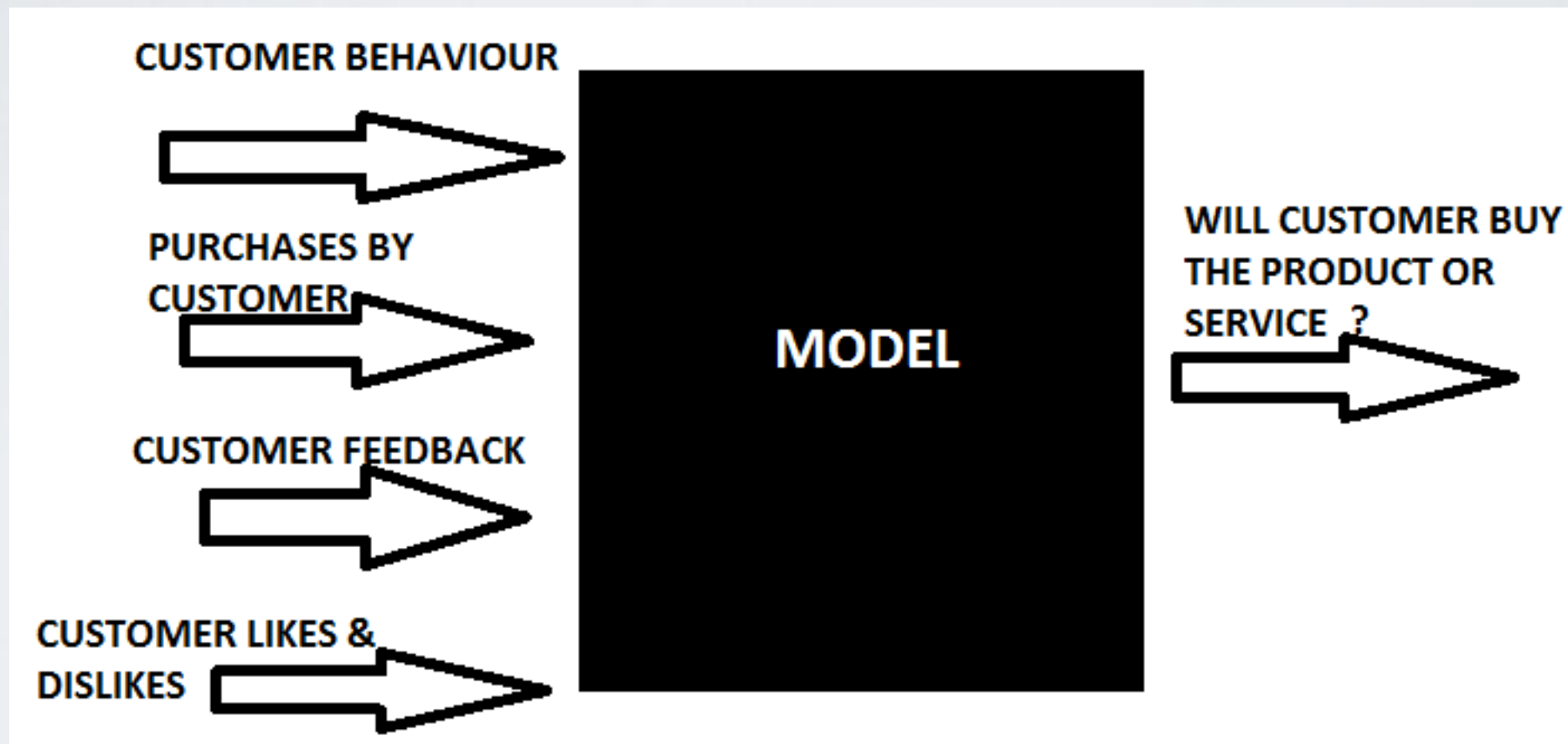
EXPERIMENTS

Training

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The top navigation bar includes the title 'Microsoft Azure Machine Learning Studio', the workspace name 'Ben Neely-Free-Workspace', and various utility icons. The left sidebar contains a search bar and a list of experiment categories: Saved Datasets, Transforms, Data Format Conversions, Data Input and Output, Data Transformation, Feature Selection, and Machine Learning. The Machine Learning section is expanded, showing options for Evaluate, Initialize Model, Anomaly Detection, and Classification. The Classification section is further expanded, listing Multiclass Decision Forest, Multiclass Decision Jungle, and Multiclass Logistic Regression. The main workspace area shows a training pipeline for the '2018 MMCI Diabetic' dataset, which is currently in draft status. The pipeline consists of the following steps: training.csv, Select Columns in Dataset (checked), Clean Missing Data (checked), Split Data, Two-class Logistic Regression, and Train Model. The bottom toolbar includes icons for NEW, RUN HISTORY, SAVE, SAVE AS, DISCARD CHANGES, RUN, SET UP WEB SERVICE, and PUBLISH TO GALLERY. The right sidebar contains tabs for Properties and Project, with the Properties tab active. It displays Experiment Properties (START TIME, END TIME, STATUS CODE: In Draft, STATUS DETAILS: None) and a Summary section with a text input field for describing the experiment (up to 140 characters). A Description section and a Quick Help icon are also visible.

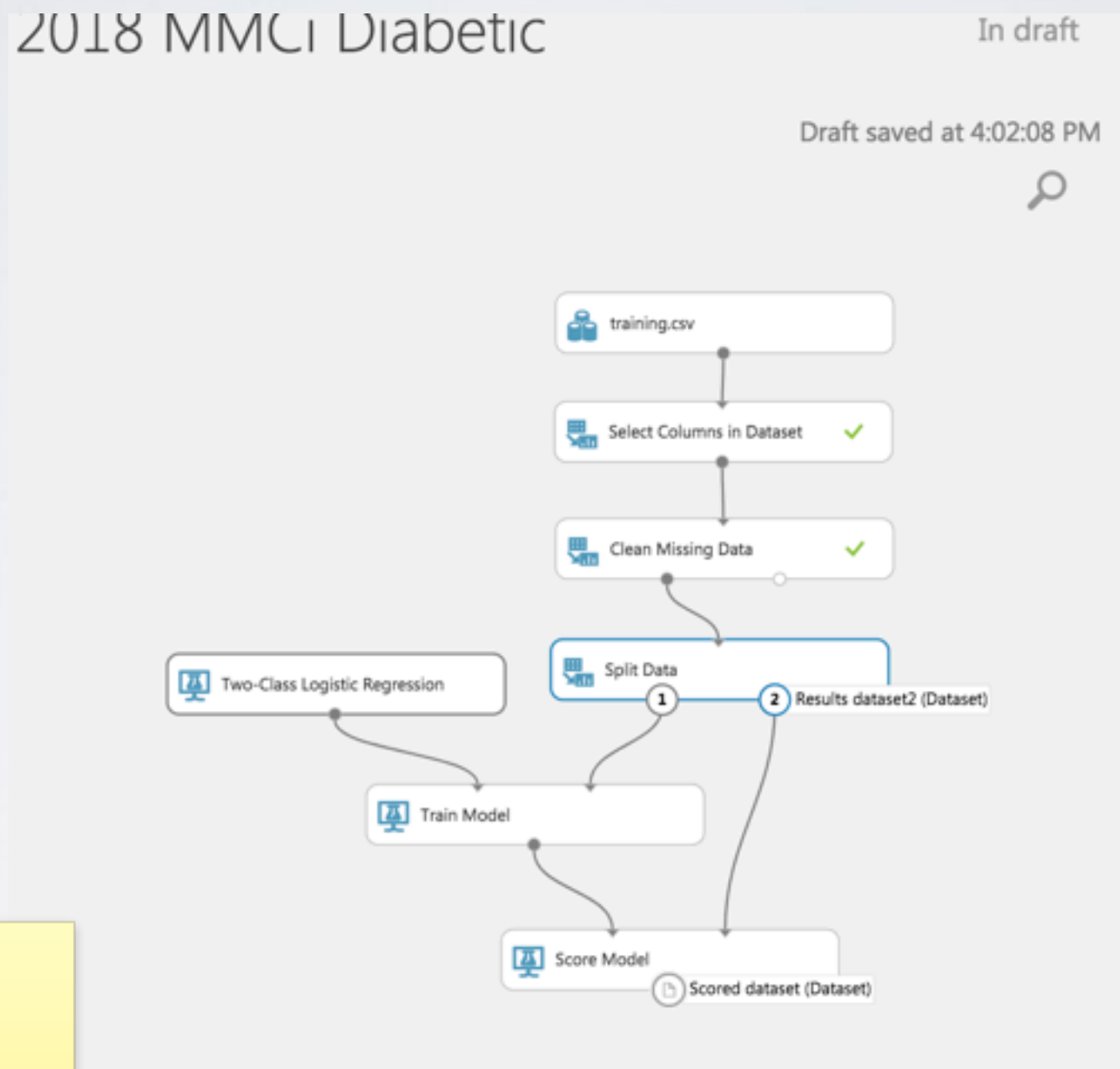


SCORE / PREDICT



EXPERIMENTS

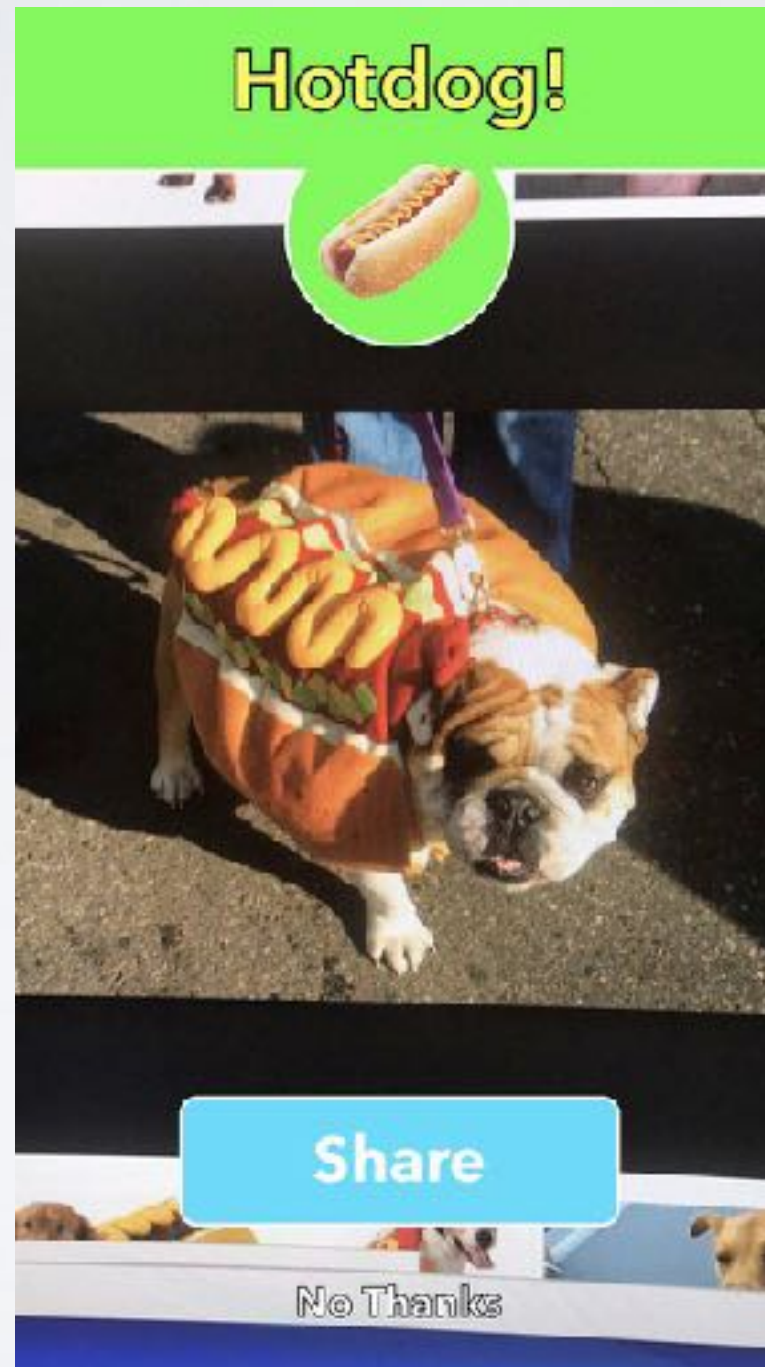
Scoring



- run experiment here and you can inspect the “black box”



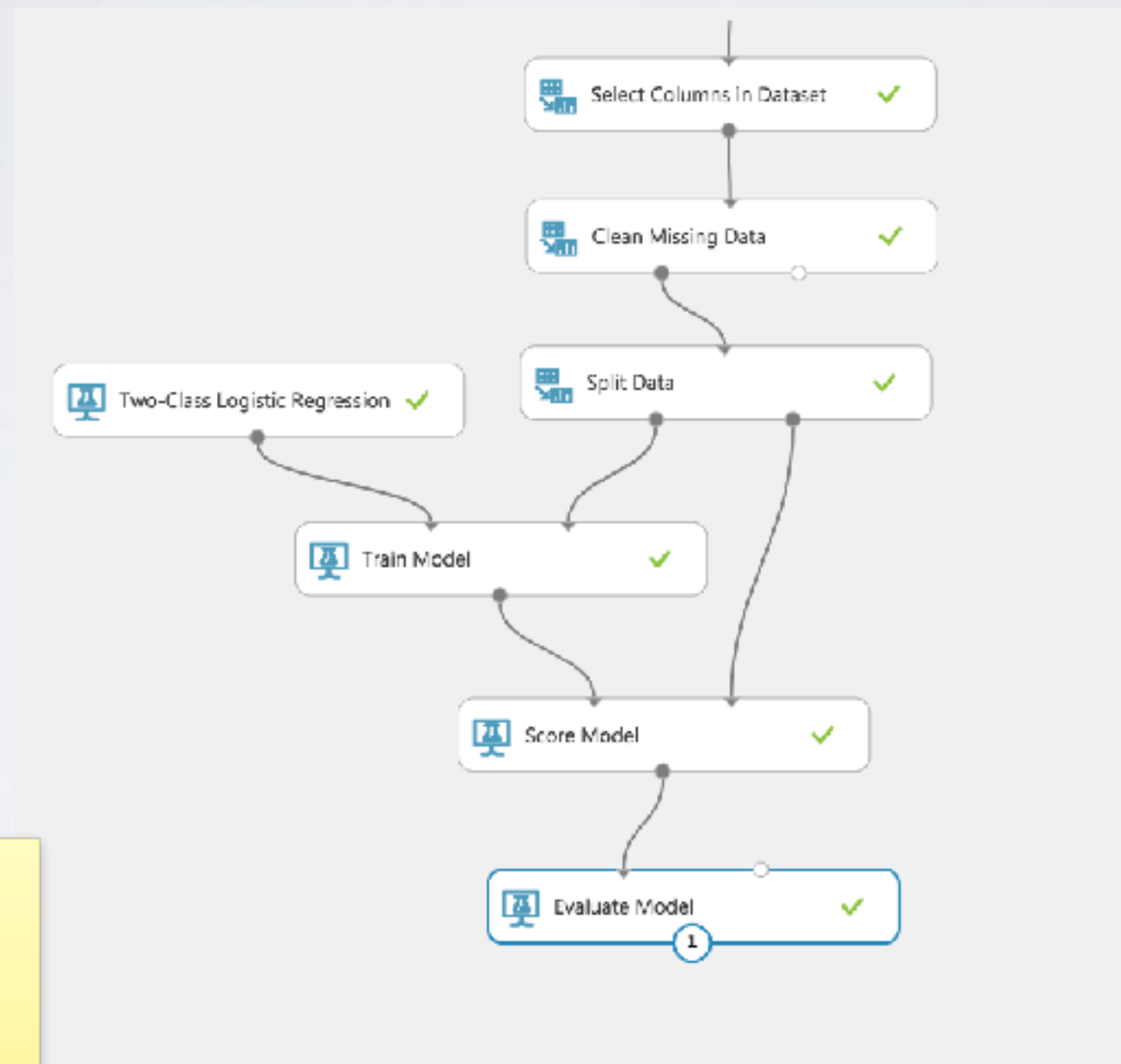
EVALUATION



- Favorite two models:
- perfect sensitivity
- perfect specificity

EXPERIMENTS

Evaluation



- self reporting AUC (shout outs)
- is this the end of the story?

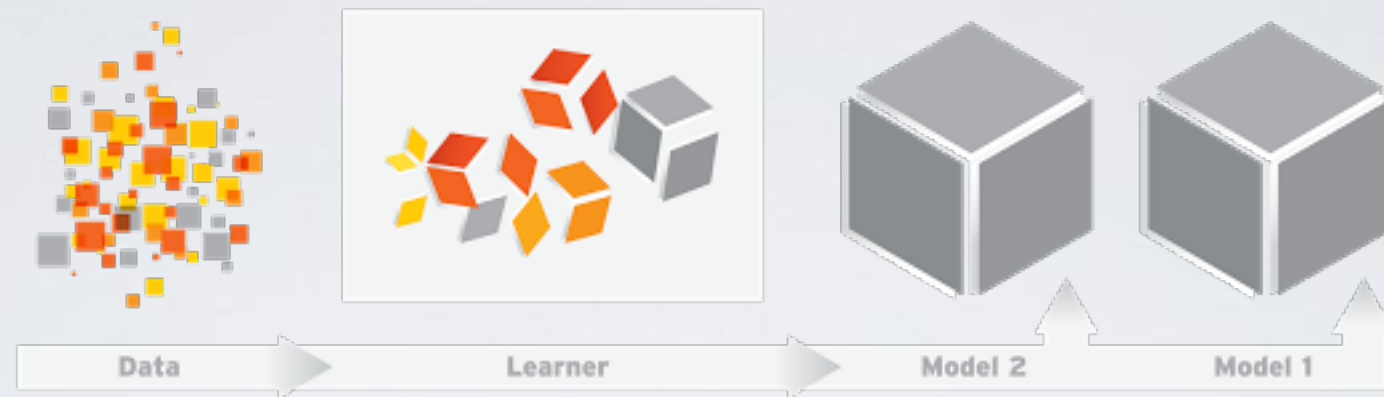


ISOLATED

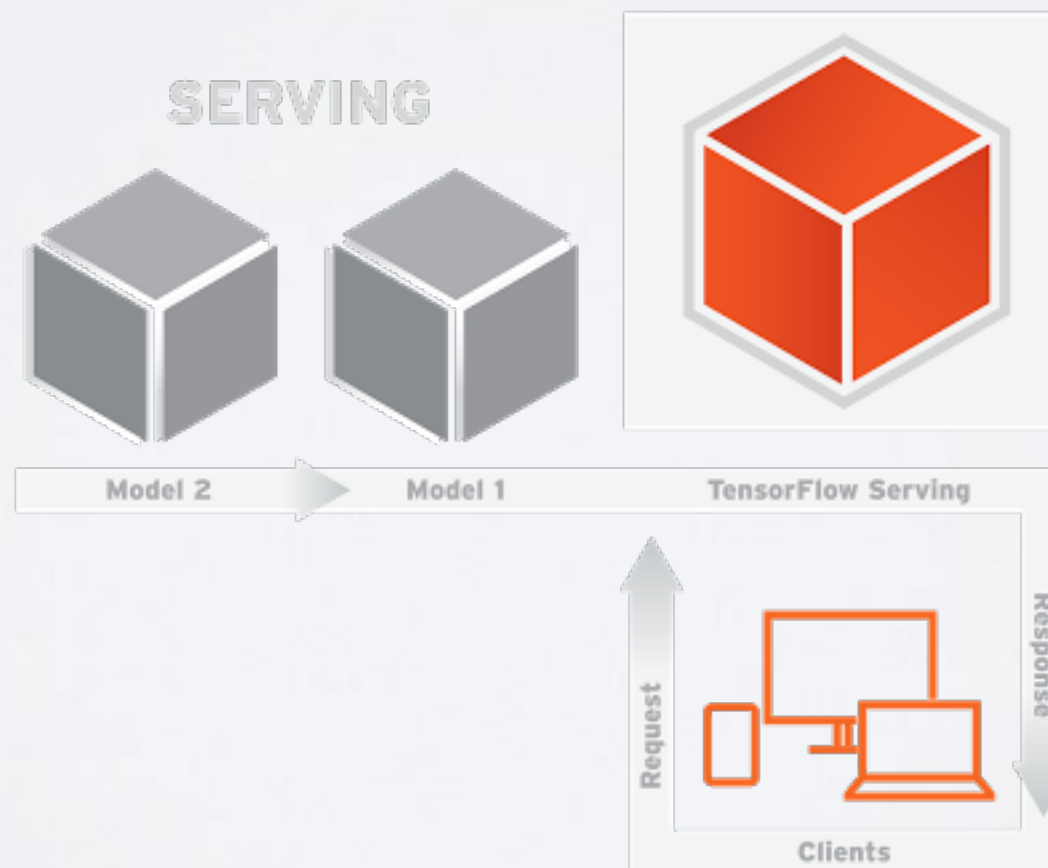


DEPLOYMENT

CONTINUOUS TRAINING PIPELINE



SERVING



DEPLOYMENT (ML STUDIO)

- Within Experiment:
 - click Save
 - click Run (play)
 - Hover Over **Set-Up Web Services:**
 - Predictive Web Service [Recommended]



WEB SERVICE IN ACTION

- Click on Experiments
- Choose Experiment that is the Web Service
- Click Run
- Publish to Gallery
 - Give Good Metadata (naming)
 - Don't need to upload image
 - See Screen shot to right for choosing visibility
 - Deploy Web Service

Visibility

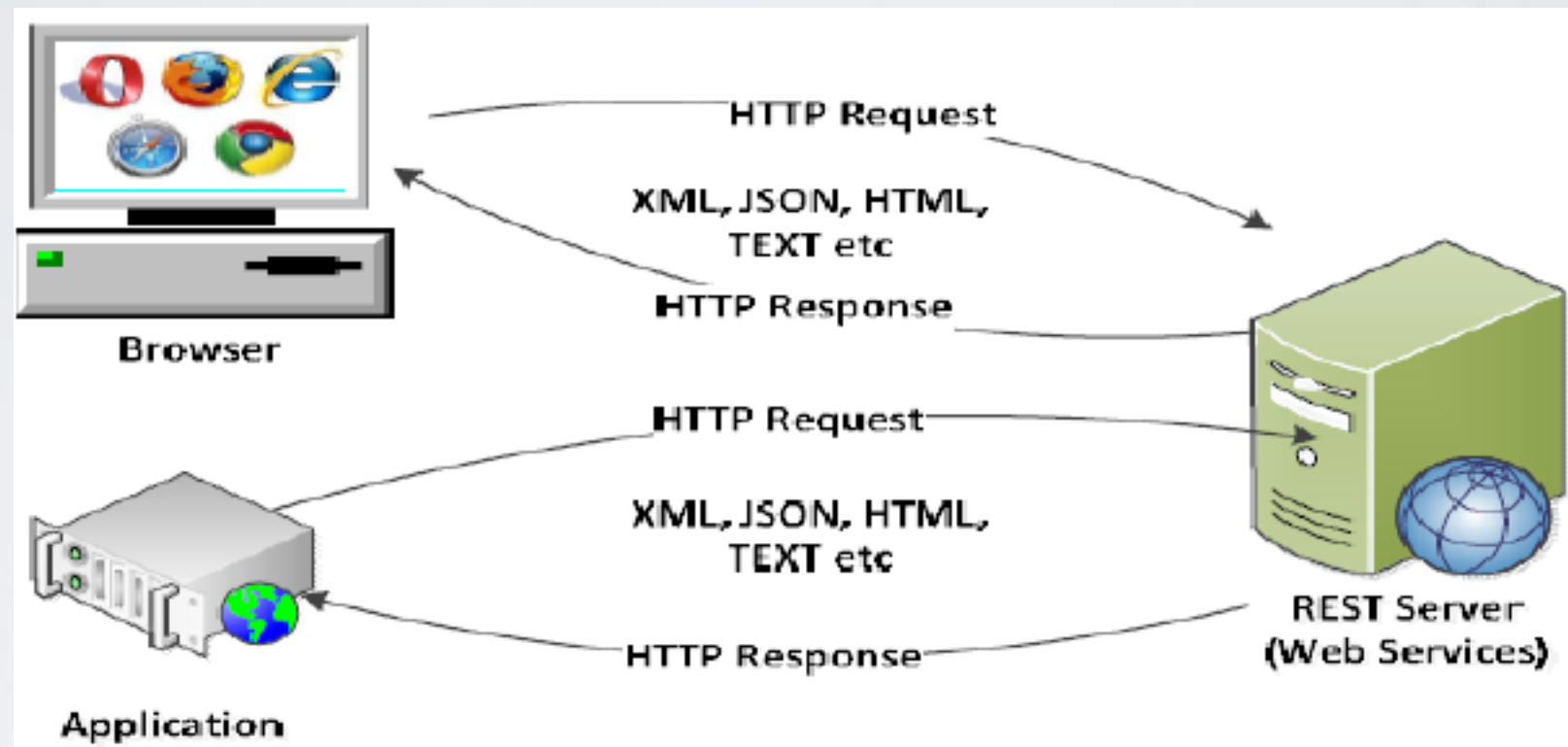
☒ PUBLIC
Anyone can view and copy the experiment. Once public, it cannot be unlisted.

☐ UNLISTED
Only people with the direct link can view and copy the experiment.

☒ I AGREE THAT PUBLISHING THE EXPERIMENT IS SUBJECT TO THE [MICROSOFT AZURE WEBSITE TERMS OF USE.](#)



WEB SERVICE IN ACTION



TEST YOUR WEB SERVICE!

- Click Web Services
- Open your web service (click hyperlink under NAME)
- Test (Blue Box) - Fill In Data - Click check to score.
- Click Test (Not Blue Box) - Do same as above.
- Discuss experience.
- Potential Use cases?



POST INPUTS

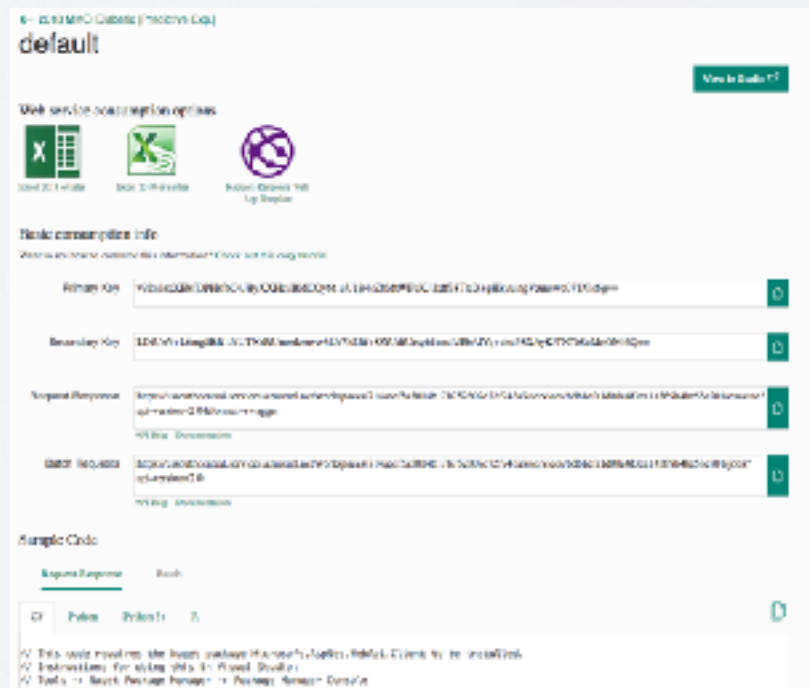
- RACE: Caucasian
- GENDER: Male
- AGE: [30, 40)
- WEIGHT: [150-175) (uh oh, what are the units???)
- NUM_LAB_PROCEDURES: 33

HEALTH SYSTEM DESIGN

- How should this tooling be implemented (bed-side tools?)
- Cadence of predictions?
- Are the algorithms proprietary?
- Do patients who have contributed to the algorithm deserve a cut?
- Do the algorithms contain PHI (can we legally make them publicly available if we want to?)

BAKE OFF

- Web Services
- Click hyperlink under “NAME”
- Next to General, click New Web Services Experience
- Click Consume
- You Should see:



BAKE OFF

- Go to Google Sheets Link and enter requested data.