Azure ML Studio
Overview for Data Engineers & Data Scientists
Rakesh Soni, Big Data Practice Director
Randi R. Ludwig, Ph.D., Data Scientist
Daniel Lai, Data Scientist
Intersys Company Summary

Overview

- Privately held IT services firm
- 150+ consultants spanning the full IT space
- Leverage a local, national and/or global model as appropriate for each customer and engagement

Key Industries

- Retail and Hospitality
- Financial Services
- Healthcare
- High Tech Manufacturing
- Media and Advertising

Core Values

- Be Accountable
- Bring Excellence
- Be Authentic
- Be in Service to Others
Core Practice Capabilities

Big Data & Analytics
- Big Data
- Analytics & Data Science
- Enterprise Search
- Business Intelligence
- Information Management

Application Services
- Application Development
- Modern Web
- Mobility
- Cloud
- DevOps / Agile

Technology Staffing
- Infrastructure
- Project Management
- Packaged Solutions
- Quality Assurance
- Cloud

Project & Program Management
Quality Assurance
Assessment
Strategy & Roadmap
Agenda

- Intersys Overview
- Machine Learning
- Azure Machine Learning Studio
- Main Features of Azure Machine Learning Studio
- Demo 1 – Predicting Income Category
- Demo 2 – Predicting Patient Readmission
- Q&A
Machine Learning – What Is It?
Machine Learning All Around Us

**Strategic Game Play**
Reinforced learning to take actions with the highest reward

Credit: Google Deep Mind, https://www.youtube.com/watch?v=vFr3K2DORc8

**Autonomous Technology**
Training computers to be make intelligent and intuitive decisions

Machine Learning in Real Life

Optimize Business Decisions

Real time insights into customer behavior

Credit: UiBS Microsoft Partner
http://www.uibs.net/uibscloud-cloud-circle-partner
Machine Learning – How Does it Work?

- **Data:** examples used to train and validate model
- **Model:** the system that makes predictions or classifications
- **Parameters:** the signals or factors used by the model to form its decisions
- **Learner:** the system that adjusts the parameters — and in turn the model — by looking at differences in predictions versus actual outcome.

Credit: Google Research, "Machine Learning 101"
Reported at: https://martechtoday.com/how-machine-learning-works-150366
Azure Machine Learning Studio
Azure ML Studio – What is it?

- Collaborative, drag-and-drop, fully managed cloud platform
- Build, test, and deploy predictive analytics
- Publish models as web services to be consumed by custom apps or BI tools.
Don’t want to build a machine learning model from scratch?

Avoid reinventing the wheel by browsing the public gallery for existing models that meet your needs.
To have full control, open your own Experiment.

For any part of the process you can use built in features. Find these via the search bar or browse the categories shown.

Once you’ve found the module you need, just drag and drop onto the workspace.
Data Import And Export Options

Bring in data in many forms. Export wherever you need it.
Easy To Create And Visualize Workflow

Easy to follow workflow for creating and sharing.

As you’re creating your model, you can see dependencies in your process.

When sharing your results, the workflow tells your story.
Dataset Statistics At Every Step

Includes quick column statistics on your whole dataset

Once your data is loaded into Azure ML, you can explore overall distributions of each feature.

You can also quickly identify columns with many missing values.
# Algorithm Selection

## Classification

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>Training time</th>
<th>Linearity</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>5</td>
</tr>
<tr>
<td>Support Vector Machine*</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>5</td>
</tr>
<tr>
<td>Neural Network</td>
<td>●</td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>Boosted decision tree†</td>
<td>●</td>
<td>○</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Decision forest</td>
<td>●</td>
<td>○</td>
<td></td>
<td>6</td>
</tr>
</tbody>
</table>

- ●: Good performance
- ○: Moderate performance

* Good for large feature sets
† Large memory footprint

## Regression

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>Training time</th>
<th>Linearity</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>4</td>
</tr>
<tr>
<td>Bayesian Linear Regression</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>2</td>
</tr>
<tr>
<td>Neural Network</td>
<td>●</td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>Boosted decision tree†</td>
<td>●</td>
<td>○</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Decision forest</td>
<td>●</td>
<td>○</td>
<td></td>
<td>6</td>
</tr>
</tbody>
</table>

## Anomaly Detection

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>Training time</th>
<th>Linearity</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal Component Analysis</td>
<td>○</td>
<td></td>
<td>●</td>
<td>3</td>
</tr>
<tr>
<td>K-means Clustering</td>
<td>○</td>
<td></td>
<td>●</td>
<td>4</td>
</tr>
</tbody>
</table>
Publish Output As A Web Service and BI Visualization

Azure Machine Learning: Basic workflow

Build models from data and operationalize a machine learning solution

Data collection & management
- Blobs & tables (Azure Storage)
- Hadoop (Azure HDInsight)
- Relational data (Azure SQL Database)
- Massive data stores (Azure Data Lake)

Machine Learning service
- ML Studio web app: Create machine learning models
- Output a web service that can be run on a scheduled basis and connected to a database.

Embedded ML model
- Add intelligence to apps or websites, or provide insights in BI tools
Azure Machine Learning Demo
Demo 1 - Predicting Income Category
Demo 2 – Predicting Patient Readmission
Azure Machine Learning Studio
Summary
Azure ML – Summary

- Interactive & visual workspace
- Various data sources supported: SQL Server, HIVE tables, CSV file etc.
- Dataset statistics and easy exploration
- Data cleaning & transformation
- Many modeling algorithms included out of the box
- SQL/R/Python code can be included in workflow
- Many built-in ways to evaluate and compare models using standard performance metrics
Microsoft's provided documentation is quite thorough and helpful:

Thank you. Any questions?
Q&A

Please submit your questions into the chat field.
Demo 1: This is a model that predicts whether a person earns > $50k.

Here we show how to build a model from scratch, including training the model, predicting values for the test set, and evaluating results.
Input your data

Adult Census Income Binary...

- **Clean Missing Data**
  - replace missing data with 0

- **Select Columns in Dataset**
  - select column that you want out of dataset
### Summary statistics

#### Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>34.647</td>
</tr>
<tr>
<td>Median</td>
<td>26.000</td>
</tr>
<tr>
<td>Min</td>
<td>1.000</td>
</tr>
<tr>
<td>Max</td>
<td>99.000</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>32.3474</td>
</tr>
<tr>
<td>Unique Values</td>
<td>94</td>
</tr>
</tbody>
</table>

#### Missing Values

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Numeric Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Values</td>
<td>0</td>
</tr>
</tbody>
</table>

#### Data Distribution

- **Distribution by Hours per Week**

![Histogram](histogram.png)

- **Distribution by Income**

![Income Distribution](income_dist.png)
How to Clean Missing Data

Clean Missing Data

Columns to be cleaned

Selected columns:
  All columns

Launch column selector

Minimum missing value ratio
  0

Maximum missing value ratio
  1

Cleaning mode
  Custom substitution value

Replacement value
  0

Generate missing value indicator column

START TIME 11/9/2016 4:13:53 PM
How to Limit the Number of Features in Your Model

Select Columns in Dataset

Select columns

Selected columns:
Column names:
age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, income

Launch column selector

START TIME 11/9/2016 4:13:57 PM
END TIME 11/9/2016 4:14:00 PM
ELAPSED TIME 0:00:03.087
STATUS CODE Finished
STATUS DETAILS None
View output log
Splitting into Training/Test Sets

### Split Data

- **Splitting mode:** Split Rows
- **Fraction of rows in the first output dataset:** 0.7
- **Randomized split:**
- **Random seed:** 0
- **Stratified split:** False

**Details:**
- **START TIME:** 11/9/2016 4:14:03 PM
- **END TIME:** 11/9/2016 4:14:06 PM
- **ELAPSED TIME:** 0:00:03
- **STATUS CODE:** Finished
- **STATUS DETAILS:** None

#### Quick Help

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

---

Two-Class Boosted Decision Tree Model with four key parameters

Train Model
- train model on training set

Permutation Feature Importance
- find out which features are important

Score Model
- predict on test set using training model

Evaluate Model
- evaluate accuracy of model
Choosing a Model

Two-Class Boosted Decision Tree

- Create trainer mode
- Maximum number of leaves per tree: 20
- Minimum number of samples per leaf node: 10
- Learning rate: 0.2
- Number of trees constructed: 100
- Random number seed:

Allow unknown categorical levels

START TIME
11/9/2016 4:13:52 PM

Quick Help
Creates a binary classifier using a boosted decision tree algorithm.
(more help...)
Training the Model
Which Features are Most Predictive?
### Which Features are Most Predictive?

#### Demo 1 › Permutation Feature Importance › Feature Importance

<table>
<thead>
<tr>
<th>Feature</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>marital-status</td>
<td>0.046683</td>
</tr>
<tr>
<td>capital-gain</td>
<td>0.044021</td>
</tr>
<tr>
<td>education-num</td>
<td>0.024877</td>
</tr>
<tr>
<td>age</td>
<td>0.017711</td>
</tr>
<tr>
<td>occupation</td>
<td>0.016687</td>
</tr>
<tr>
<td>capital-loss</td>
<td>0.013616</td>
</tr>
<tr>
<td>hours-per-week</td>
<td>0.006859</td>
</tr>
<tr>
<td>workclass</td>
<td>0.003583</td>
</tr>
<tr>
<td>fnlweat</td>
<td>0.002048</td>
</tr>
</tbody>
</table>
Predicting Values for Test Data
# Check Predictions (Scored Labels)

## Demo 1 > Score Model > Scored dataset

<table>
<thead>
<tr>
<th>rows</th>
<th>columns</th>
</tr>
</thead>
<tbody>
<tr>
<td>9768</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>education</th>
<th>education-num</th>
<th>marital-status</th>
<th>occupation</th>
<th>relationship</th>
<th>race</th>
<th>sex</th>
<th>capital-gain</th>
<th>capital-loss</th>
<th>hours-per-week</th>
<th>native-country</th>
<th>income</th>
<th>Scored Labels</th>
<th>Scored Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bachelors</td>
<td>13</td>
<td>Divorced</td>
<td>Craft-repair</td>
<td>Unmarried</td>
<td>White</td>
<td>Male</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>United-States</td>
<td>&lt;=50K</td>
<td>&lt;=50K</td>
<td>0.025473</td>
</tr>
<tr>
<td>10th</td>
<td>6</td>
<td>Never-married</td>
<td>Priv-house-serv</td>
<td>Own-child</td>
<td>White</td>
<td>Female</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>United-States</td>
<td>&lt;=50K</td>
<td>&lt;=50K</td>
<td>0.000855</td>
</tr>
<tr>
<td>HS-grad</td>
<td>9</td>
<td>Divorced</td>
<td>Adm-clerical</td>
<td>Unmarried</td>
<td>White</td>
<td>Female</td>
<td>0</td>
<td>0</td>
<td>48</td>
<td>United-States</td>
<td>&lt;=50K</td>
<td>&lt;=50K</td>
<td>0.024609</td>
</tr>
<tr>
<td>Some-college</td>
<td>10</td>
<td>Married-civ-spouse</td>
<td>Prof-specialty</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>United-States</td>
<td>&lt;=50K</td>
<td>&lt;=50K</td>
<td>0.377446</td>
</tr>
<tr>
<td>Masters</td>
<td>14</td>
<td>Married-civ-spouse</td>
<td>Prof-specialty</td>
<td>Wife</td>
<td>White</td>
<td>Female</td>
<td>0</td>
<td>0</td>
<td>45</td>
<td>United-States</td>
<td>&gt;50K</td>
<td>&gt;50K</td>
<td>0.903994</td>
</tr>
<tr>
<td>HS-grad</td>
<td>9</td>
<td>Married-civ-spouse</td>
<td>Exec-managerial</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>United-States</td>
<td>&lt;=50K</td>
<td>&lt;=50K</td>
<td>0.295065</td>
</tr>
</tbody>
</table>
Check Model Accuracy, Etc.
End of Demo 1

Select Columns in Dataset

Select columns

Selected columns:
Column names: age, workclass,fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, income

Launch column selector

START TIME: 11/9/2016 4:13:57 PM
END TIME: 11/9/2016 4:14:00 PM
ELAPSED TIME: 0:00:03.087
STATUS CODE: Finished
STATUS DETAILS: None

View output log
Demo 2: This model predicts whether a patient will be readmitted to a hospital for further treatment.

This model considers a variety of strong machine learning algorithms. It then tunes the strongest model to be more efficient, evaluation of prediction results, and implements custom code in from R, Python, and SQL for further visualization and examination of data.
Explore the Dataset
Impute Missing Values
Compare Models: Decision Jungle
Compare Models: Boosted Decision Tree
Compare Models: Logistic Regression
Compare Models: Neural Network
Cross Validation Results: Variation in Accuracy for All Algorithms
ROC Chart comparison for models
Move Forward with Most Accurate Model
How to Cross Validate a Tuned Model

1. Partition and Sample
   Prepare partitioning to sweep parameter's cross-validation mode.

2. Tune Model Hyperparameters
   Train and tune model 2 on training set.

3. Two-Class Boosted Decision...
   Model type 2 was best.
Set Model Parameter Ranges
Tune Model Parameters
Use Model to Predict Values for Test Set
Visualize Predicted Values

### Score Model
- Scored dataset

<table>
<thead>
<tr>
<th>Column</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
<th>Value 5</th>
<th>Value 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>pi/alpha</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>acarbose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mgptol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>troglitazone</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tosasme</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>insulin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>glukos/glukos</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rnetform/glukos</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rnetform/glukos</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rnetform/glukos</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>change</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>diabetes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>admission_type_description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>num_mets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>readm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scored Labels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scored Probabilities</th>
<th>Mean</th>
<th>Median</th>
<th>Max</th>
<th>Standard Deviation</th>
<th>Missing Values</th>
<th>Feature Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.6449</td>
<td>0.4487</td>
<td>0.9994</td>
<td>0.37</td>
<td>0</td>
<td>Numeric Score</td>
</tr>
</tbody>
</table>

### Visualizations
- Scored Probabilities
- Histogram

**Histogram**

- Scored Probabilities

**Scored Probabilities**

- Frequency

**Frequency**

- Scored Probabilities

**Scored Probabilities**

- Frequency
Permute Features to Find Most Important
Find Features Most Influential to the Model

<table>
<thead>
<tr>
<th>Feature</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>number_inpatient</td>
<td>0.050442</td>
</tr>
<tr>
<td>discharge_disposition_id</td>
<td>0.021291</td>
</tr>
<tr>
<td>admission_source_id</td>
<td>0.01332</td>
</tr>
<tr>
<td>payer_code</td>
<td>0.008735</td>
</tr>
<tr>
<td>number_outpatient</td>
<td>0.007752</td>
</tr>
<tr>
<td>num_lab_procedures</td>
<td>0.007206</td>
</tr>
<tr>
<td>number_diagnoses</td>
<td>0.007206</td>
</tr>
<tr>
<td>race</td>
<td>0.006223</td>
</tr>
<tr>
<td>number_emergency</td>
<td>0.005568</td>
</tr>
<tr>
<td>time_in_hospital</td>
<td>0.004913</td>
</tr>
<tr>
<td>diag_1</td>
<td>0.004695</td>
</tr>
<tr>
<td>age</td>
<td>0.004388</td>
</tr>
<tr>
<td>diag_2</td>
<td>0.004367</td>
</tr>
<tr>
<td>num_procedures</td>
<td>0.004258</td>
</tr>
<tr>
<td>weight</td>
<td>0.003712</td>
</tr>
</tbody>
</table>

**Statistics**
- Mean: 0.0039
- Median: 0.0011
- Min: -0.0021
- Max: 0.0504
- Standard Deviation: 0.0083
- Unique Values: 27
- Missing Values: 0

**Feature Type**: Numeric Feature

**Visualizations**
- Score
- Histogram
Can Export Results for Multiple Uses

- Convert to CSV: can export csv or open in a Jupyter notebook for further debugging or analysis.
- Execute Python Script: visualization script.
- Edit Metadata: edit column names for R import.
- Execute R Script: finding optimal cutoff point to minimize cost of bed predictions.
- Apply SQL Transformation: further feature pruning for more advanced models.
Export CSV for debugging in notebooks
Sample R Notebook

```r
In [1]: library("AzureML")
ws <- workspace()
dat <- download.intermediate.dataset(
ws,
  experiment = "f07d1cf0ba64b19ae0620f6a4d84d3-f-id.763f3d4d4fa7a48c9bcafc8fc30cc64ee",
  node_id = "d9a4d521-af4f-41e6-b4ff-7cf56ff4a8f2-139463",
  port_name = "Results dataset",
  data_type_id = "GenericCSV"
)

In [2]: head(dat)
```
SQL scripts

```sql
1 SELECT number_inpatient,
    discharge_disposition_id, 
    number_emergency, 
    readmi_class  
2 FROM t1
```
## Apply SQL Transformation

### Results dataset

<table>
<thead>
<tr>
<th>number_inpatient</th>
<th>discharge_disposition_id</th>
<th>number_emergency</th>
<th>readmi_class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>YES</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>YES</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>0</td>
<td>YES</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>0.190476</td>
<td>1</td>
<td>0.039474</td>
<td>YES</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>0.047610</td>
<td>3</td>
<td>0.013158</td>
<td>NO</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0</td>
<td>YES</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>YES</td>
</tr>
<tr>
<td>0.047619</td>
<td>6</td>
<td>0</td>
<td>YES</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>0.047619</td>
<td>3</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>0.142857</td>
<td>11</td>
<td>0.013158</td>
<td>NO</td>
</tr>
</tbody>
</table>

**Statistics**
- Unique Values: 2
- Missing Values: 0
- Feature Type: String Feature

**Visualizations**
- Histogram
  - compare to: None
Find optimal cutoff in R script
R result

```
Demo2 ➤ Execute R Script ➤ R Device

- **Standard Output**
  - RWorker pushed "port1" to R workspace.
  - Beginning R Execute Script

  [1] 56000
  Loading objects:
  - port1
  [1] "Loading variable port1..."
  [1] 0.15
  [1] "Saving variable dataset ...
  [1] "Saving the following item(s): .memi.oport1"

- **Standard Error**
  - R reported no errors.

- **Graphics**

```

![Image](image_url)

---

67
Python script for further visualizations
Python Visualization

Demo 2 ➔ Execute Python Script ➔ Python device

- Graphics
Evaluate Final Model
Results