Microsoft Advanced Analytics

Juan Carlos Rodriguez García
jurodr@microsoft.com
Data Platform Solution Architect

Microsoft

Introducción

Modelo de Madurez Analítica

Fuente: Gartner

Hay Proyectos Muy Avanzados…

Banca
Omnicanal

MAX
Maximizar el Tiempo,
Todo el Tiempo

Visión
Artificial

Medicina
Preventiva

Análisis
Conductual
en Tienda

… Pero el Futuro es más Complejo

Los desplegables “tontos” van a desaparecer

Las metodologías Agile permiten aplicar conocimiento a los productos muy rápidamente

Herramientas Microsoft

Microsoft R Server

R Open

R+CRAN

RSR Connector

Microsoft R Server

DevelopR

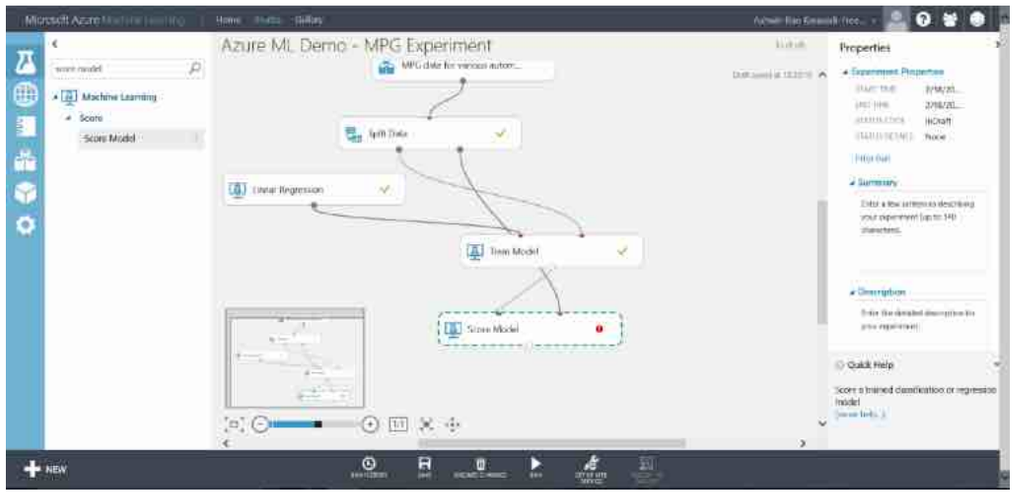
DeployR

ConnectR

ScaleR

DistributedR

Azure Machine Learning



Cognitive Services

Speech API

Vision API

Bot Framework

Anomaly Detection

Customer Churn

Text Analytics

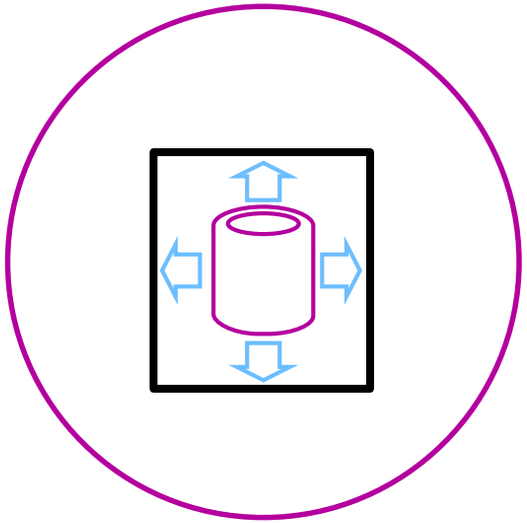
Natural Language

Speech recognition

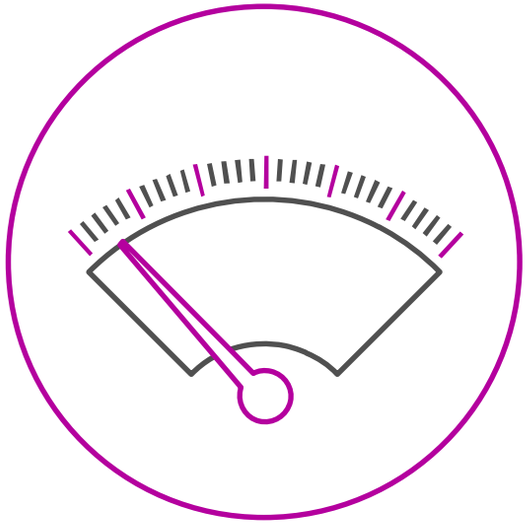
Emotion API

Microsoft R Server

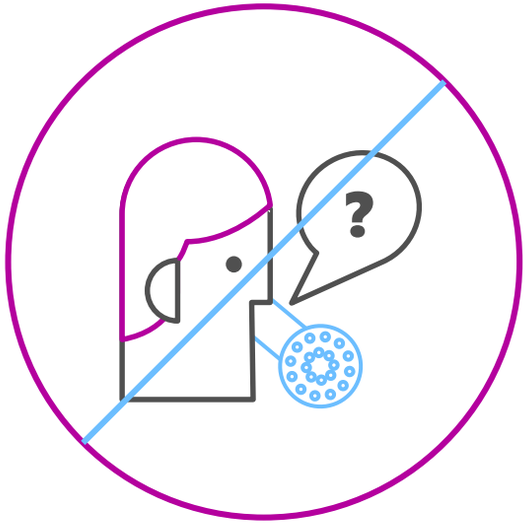
Retos de R



Volumen de Datos



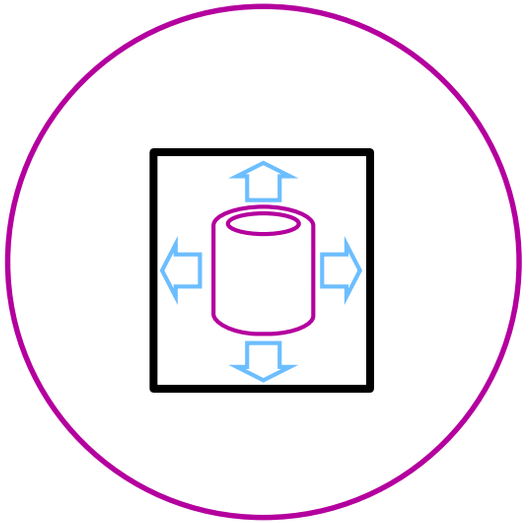
Paralelización



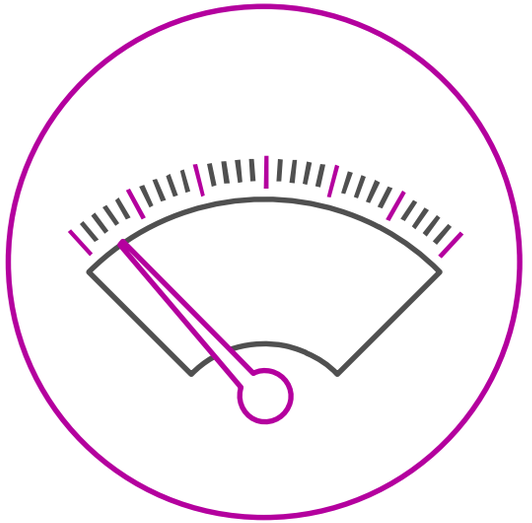
Soporte

Despliegue

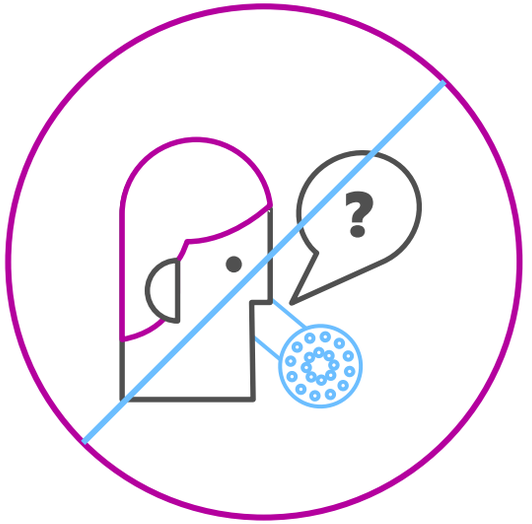
Soluciones de Microsoft R Server



Streaming de Datos



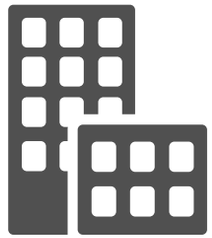
Single Source Multi Thread Multi Nodo



Soporte MSFT

Despliegue Sobre Cluster y Cloud

Microsoft R Server



Microsoft R

Comunidad

Microsoft
R Open

Comercial

SQL Server
R Services

Windows

Microsoft
R Server

Linux

Hadoop | Teradata

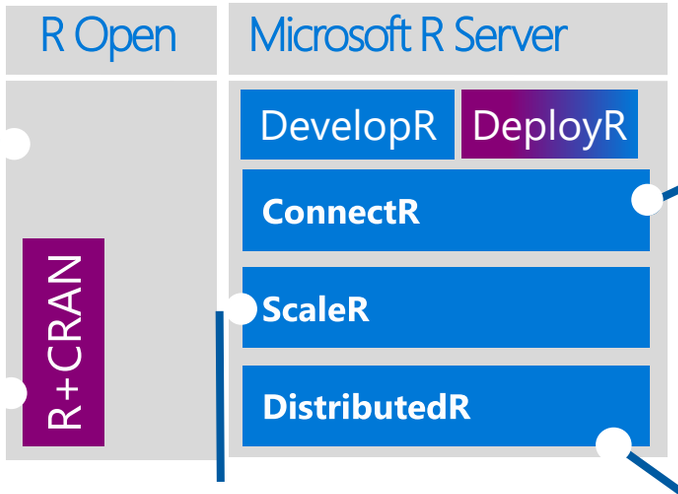
Plataforma Microsoft R Server

R+CRAN

- Open source R interpreter
  - R 3.1.2
- Freely-available huge range of R algorithms
- Algorithms callable by RevoR
- Embeddable in R scripts
- 100% Compatible with existing R scripts, functions and packages

RevoR

- Performance enhanced R interpreter
- Based on open source R
- Adds high-performance math library to speed up linear algebra functions



ScaleR

- Ready-to-Use high-performance big data big analytics
- Fully-parallelized analytics
- Data prep & data distillation
- Descriptive statistics & statistical tests
- Range of predictive functions
- User tools for distributing customized R algorithms across nodes
- Wide data sets supported – thousands of variables

ConnectR

- High-speed & direct connectors

Available for:

- High-performance XDF
- SAS, SPSS, delimited & fixed format text data files
- Hadoop HDFS (text & XDF)
- Teradata Database & Aster
- EDWs and ADWs
- ODBC

DistributedR

- Distributed computing framework
- Delivers cross-platform portability

Modelos Paralelizados

Data Step

- Data import – Delimited, Fixed, SAS, SPSS, OBDC
- Variable creation & transformation
- Recode variables
- Factor variables
- Missing value handling
- Sort, Merge, Split
- Aggregate by category (means, sums)

Descriptive Statistics

- Min / Max, Mean, Median (approx.)
- Quantiles (approx.)
- Standard Deviation
- Variance
- Correlation
- Covariance
- Sum of Squares (cross product matrix for set variables)
- Pairwise Cross tabs
- Risk Ratio & Odds Ratio
- Cross-Tabulation of Data (standard tables & long form)
- Marginal Summaries of Cross Tabulations

Statistical Tests

- Chi Square Test
- Kendall Rank Correlation
- Fisher's Exact Test
- Student's t-Test

Sampling

- Subsample (observations & variables)
- Random Sampling

Predictive Models

- Sum of Squares (cross product matrix for set variables)
- Quantiles (approx.)
- Generalized Linear Models (GLM) exponential family distributions: binomial, Gaussian, inverse Gaussian, Poisson, Tweedie. Standard link functions: cauchit, identity, log, logit, probit. User defined distributions & link functions.
- Covariance & Correlation Matrices
- Logistic Regression
- Classification & Regression Trees
- Predictions/scoring for models
- Residuals for all models

Variable Selection

- Stepwise Regression

Simulation

- Simulation (e.g. Monte Carlo)
- Parallel Random Number Generation

Cluster Analysis

- K-Means

Classification

- Decision Trees
- Decision Forests
- Gradient Boosted Decision Trees
- Naïve Bayes

Combination

- rxDataStep
- rxExec
- PEMA-R API Custom Algorithms

New

Single Source… Multi Thread, Multi Node

Procesamiento Paralelo Local

Se establece dónde se ejecutará el modelo

```
### SETUP LOCAL ENVIRONMENT VARIABLES ###
    myLocalCC <- "localpar"

### LOCAL COMPUTE CONTEXT ###
    rxSetComputeContext(myLocalCC)


### CREATE LINUX, DIRECTORY AND FILE OBJECTS ###
    localFS <- RxNativeFileSystem()
    AirlineDataSet <- RxXdfData("AirlineDemoSmall.xdf",
    fileSystem = localFS)
```

Modelo funcional, no afectado por los contextos

```
### ANALYTICAL PROCESSING ###
### Statistical Summary of the data
    rxSummary(~ArrDelay+DayOfWeek, data= AirlineDataSet, reportProgress=1)

### CrossTab the data
    rxCrossTabs(ArrDelay ~ DayOfWeek, data= AirlineDataSet, means=T)

### Linear Model and plot
    hdfsXdfArrLateLinMod   <- rxLinMod(ArrDelay ~ DayOfWeek + 0 , data = AirlineDataSet)
    plot(hdfsXdfArrLateLinMod$coefficients)
```

Single Source… Multi Thread, Multi Node

Se establece dónde se ejecutará el modelo

Procesamiento Paralelo Local

```
### SETUP LOCAL ENVIRONMENT VARIABLES ###
    myLocalCC <- "localpar"

### LOCAL COMPUTE CONTEXT ###
    rxSetComputeContext(myLocalCC)


### CREATE LINUX, DIRECTORY AND FILE OBJECTS ###
    localFS <- RxNativeFileSystem()
    AirlineDataSet <- RxXdfData("AirlineDemoSmall.xdf",
    fileSystem = localFS)
```

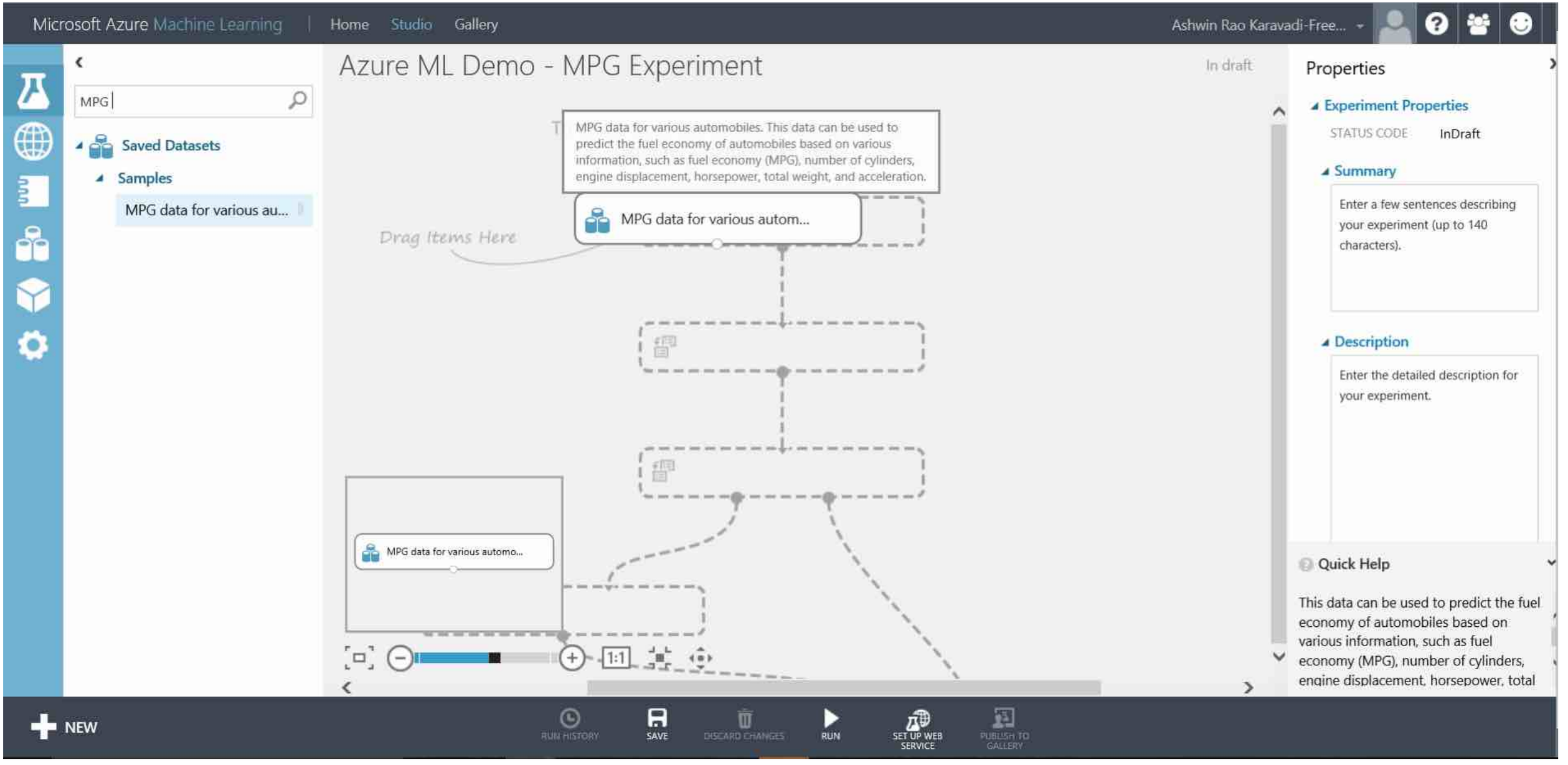
Procesamiento Distribuido

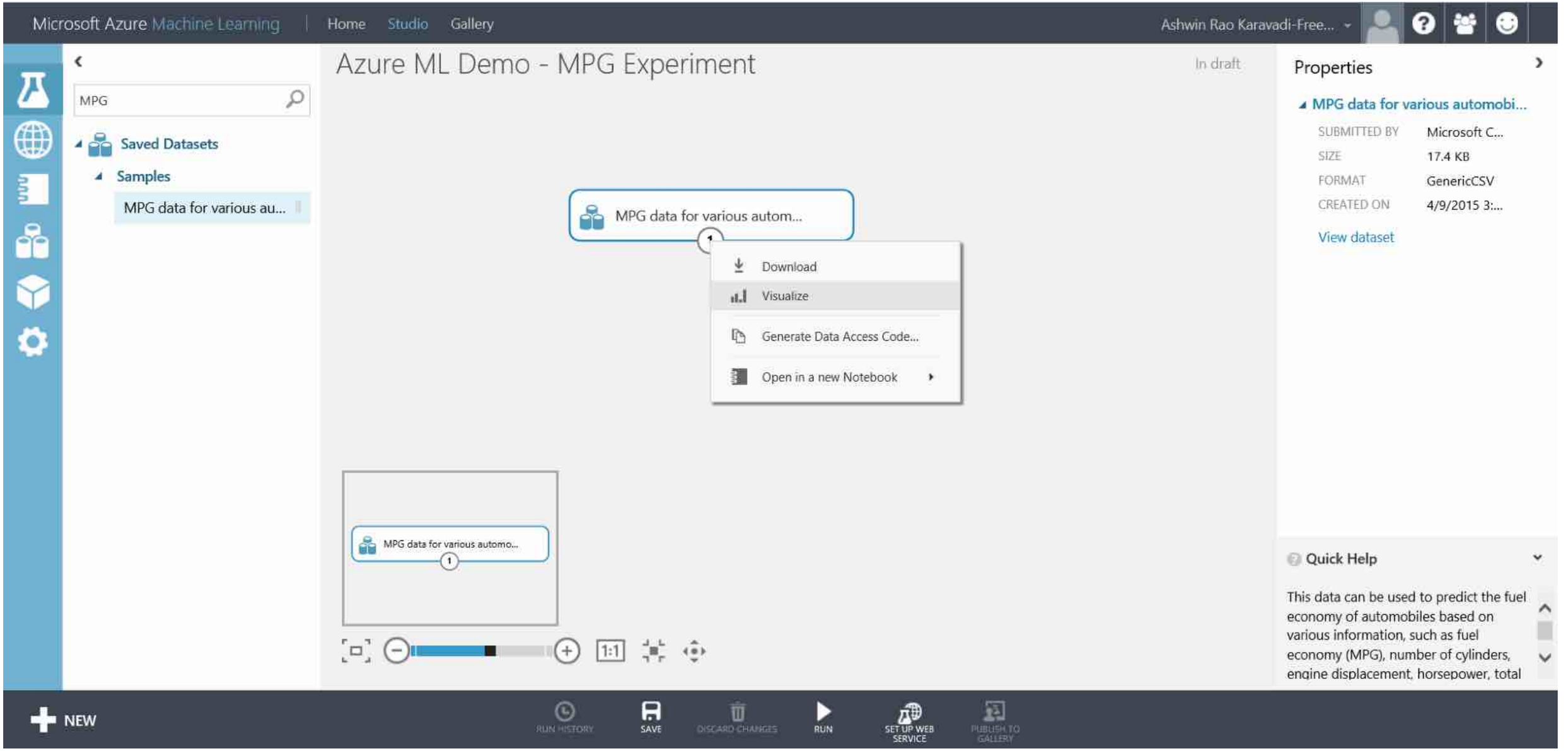
```
### SETUP HADOOP ENVIRONMENT VARIABLES ###
    myHadoopCC <- RxHadoopMR()

### HADOOP COMPUTE CONTEXT ###
    rxSetComputeContext(myHadoopCC)

### CREATE HDFS, DIRECTORY AND FILE OBJECTS ###
    hdfsFS <- RxHdfsFileSystem()
    AirlineDataSet <- RxXdfData("AirlineDemoSmall.xdf"),
    fileSystem = hdfsFS)
```

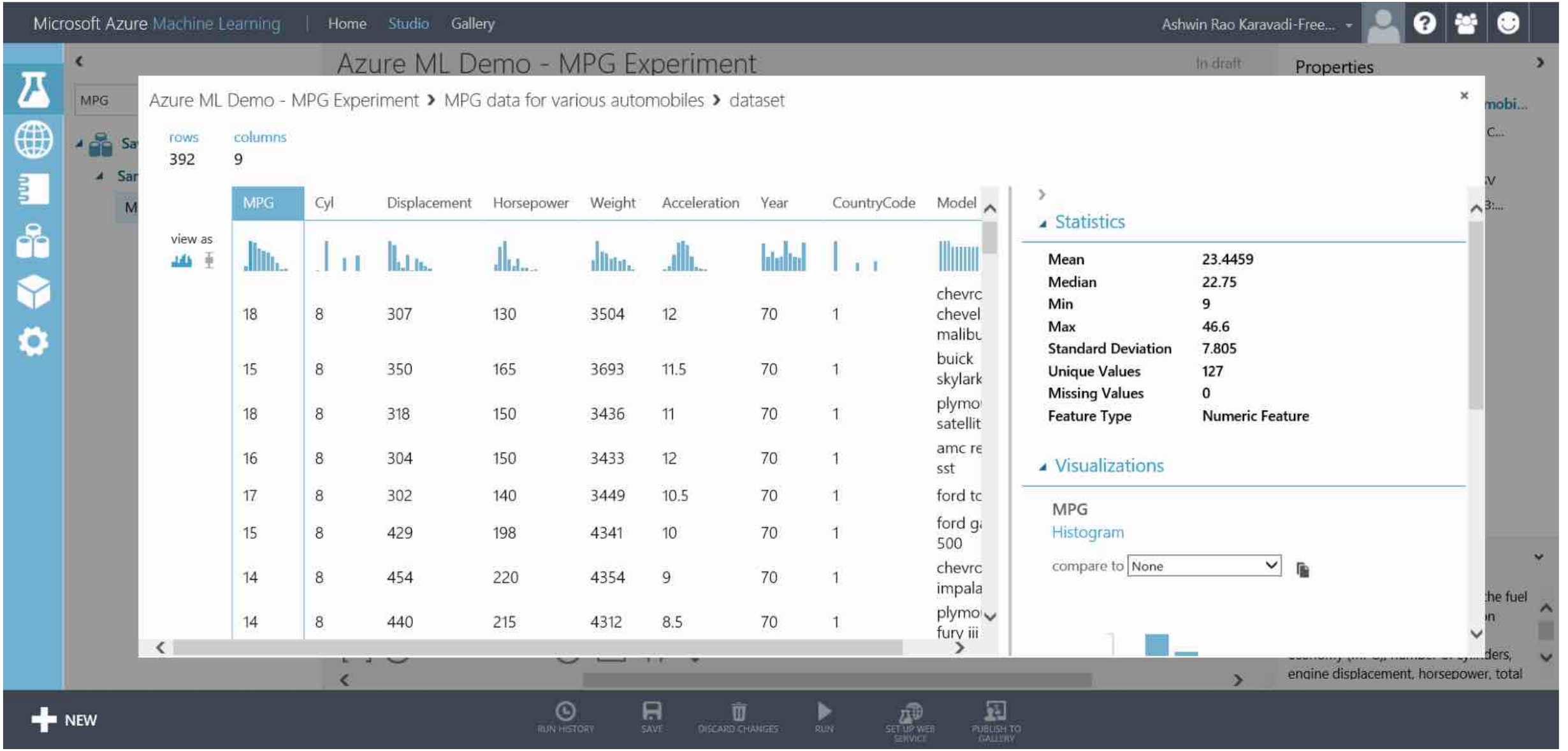
Modelo funcional, no afectado por los contextos

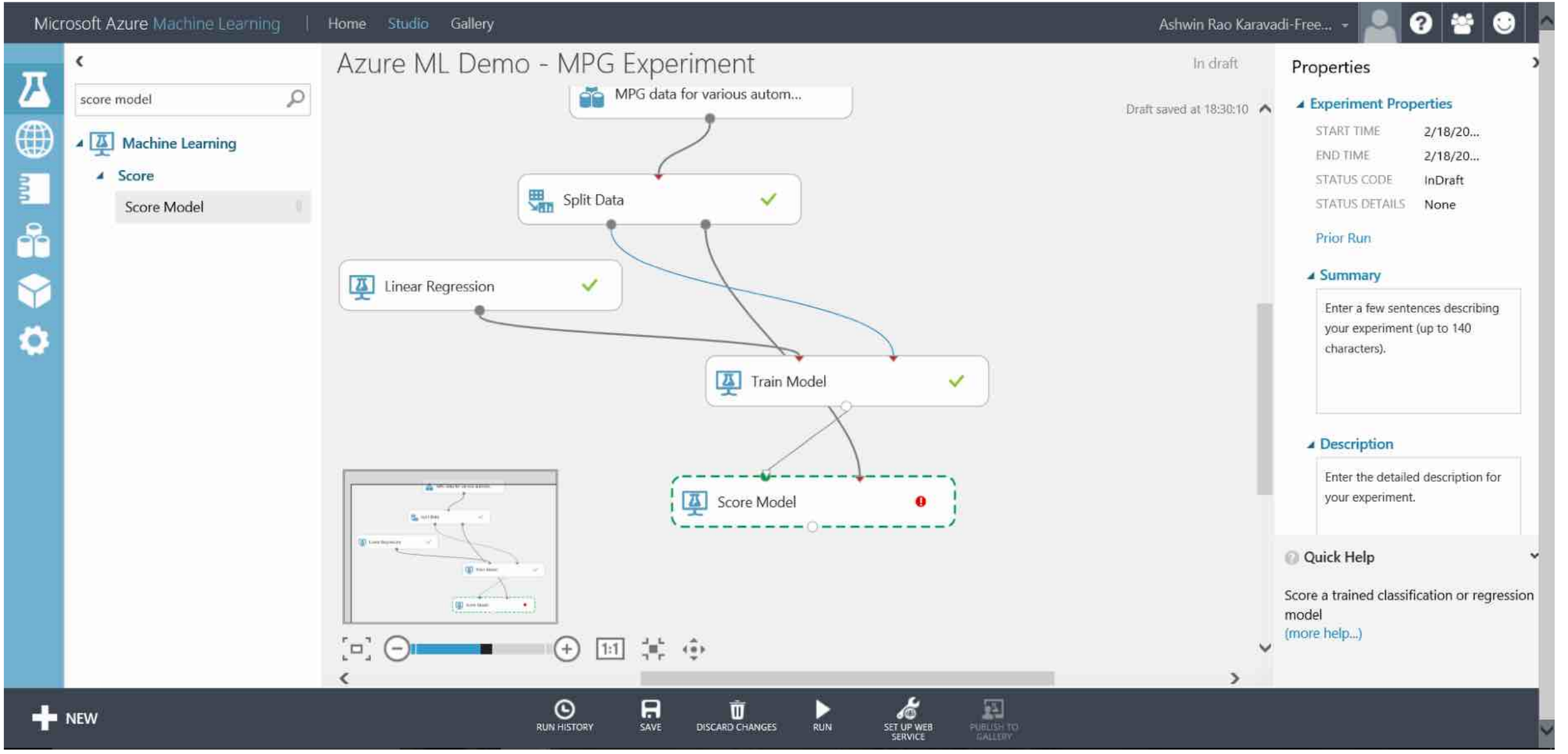
```
### ANALYTICAL PROCESSING ###
### Statistical Summary of the data
    rxSummary(~ArrDelay+DayOfWeek, data= AirlineDataSet, reportProgress=1)

### CrossTab the data
    rxCrossTabs(ArrDelay ~ DayOfWeek, data= AirlineDataSet, means=T)

### Linear Model and plot
    hdfsXdfArrLateLinMod   <- rxLinMod(ArrDelay ~ DayOfWeek + 0 , data = AirlineDataSet)
    plot(hdfsXdfArrLateLinMod$coefficients)
```

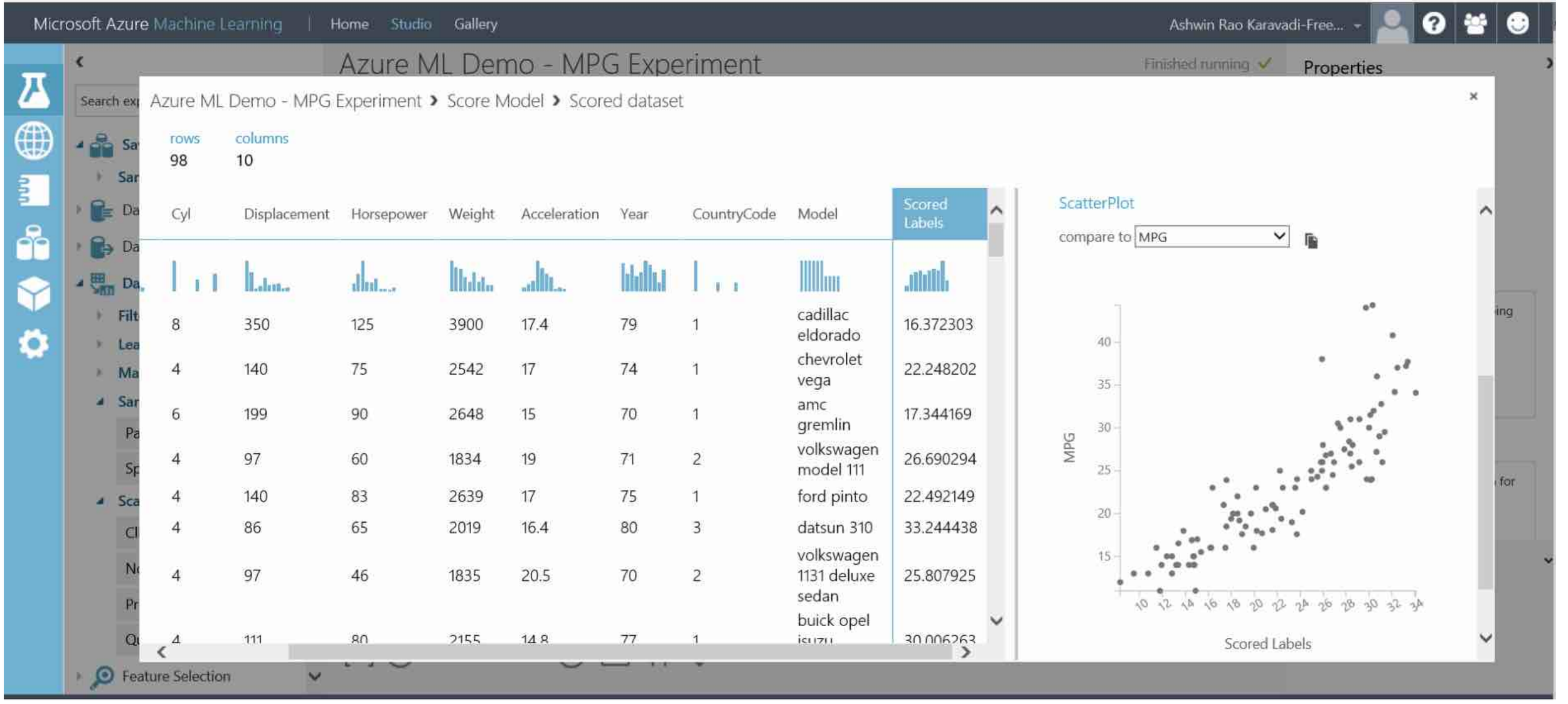
Azure Machine Learning


Microsoft Azure Machine Learning
Home
Studio
Gallery
Ashwin Rao Karavadi-Free...
MPG
Saved Datasets
Samples
MPG data for various au...
Azure ML Demo - MPG Experiment
In draft
MPG data for various automobiles. This data can be used to predict the fuel economy of automobiles based on various information, such as fuel economy (MPG), number of cylinders, engine displacement, horsepower, total weight, and acceleration.
MPG data for various autom...
Drag Items Here
MPG data for various automo...
1:1
Properties
Experiment Properties
STATUS CODE
InDraft
Summary
Enter a few sentences describing your experiment (up to 140 characters).
Description
Enter the detailed description for your experiment.
Quick Help
This data can be used to predict the fuel economy of automobiles based on various information, such as fuel economy (MPG), number of cylinders, engine displacement, horsepower, total
NEW
RUN HISTORY
SAVE
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY


Microsoft Azure Machine Learning
Home
Studio
Gallery
Ashwin Rao Karavadi-Free...
MPG
Saved Datasets
Samples
MPG data for various au...
Azure ML Demo - MPG Experiment
In draft
MPG data for various autom...
Download
Visualize
Generate Data Access Code...
Open in a new Notebook
MPG data for various automo...
1:1
Properties
MPG data for various automobi...
SUBMITTED BY
Microsoft C...
SIZE
17.4 KB
FORMAT
GenericCSV
CREATED ON
4/9/2015 3:...
View dataset
Quick Help
This data can be used to predict the fuel economy of automobiles based on various information, such as fuel economy (MPG), number of cylinders, engine displacement, horsepower, total
NEW
RUN HISTORY
SAVE
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY


Microsoft Azure Machine Learning
Home
Studio
Gallery
Ashwin Rao Karavadi-Free...
Azure ML Demo - MPG Experiment
In draft
Properties
Azure ML Demo - MPG Experiment › MPG data for various automobiles › dataset
rows
392
columns
9
view as
MPG Cyl Displacement Horsepower Weight Acceleration Year CountryCode Model
18 8 307 130 3504 12 70 1 chevrc chevel malibu
15 8 350 165 3693 11.5 70 1 buick skylark
18 8 318 150 3436 11 70 1 plymo satellit
16 8 304 150 3433 12 70 1 amc re sst
17 8 302 140 3449 10.5 70 1 ford tc
15 8 429 198 4341 10 70 1 ford g 500
14 8 454 220 4354 9 70 1 chevrc impala
14 8 440 215 4312 8.5 70 1 plymo fury iii
Statistics
Mean 23.4459
Median 22.75
Min 9
Max 46.6
Standard Deviation 7.805
Unique Values 127
Missing Values 0
Feature Type Numeric Feature
Visualizations
MPG
Histogram
compare to None
NEW
RUN HISTORY
SAVE
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY


Microsoft Azure Machine Learning
Home
Studio
Gallery
Ashwin Rao Karavadi-Free...
score model
Machine Learning
Score
Score Model
Azure ML Demo - MPG Experiment
In draft
Draft saved at 18:30:10
MPG data for various autom...
Split Data
Linear Regression
Train Model
Score Model
Properties
Experiment Properties
START TIME 2/18/20...
END TIME 2/18/20...
STATUS CODE InDraft
STATUS DETAILS None
Prior Run
Summary
Enter a few sentences describing your experiment (up to 140 characters).
Description
Enter the detailed description for your experiment.
Quick Help
Score a trained classification or regression model
(more help...)
NEW
RUN HISTORY
SAVE
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY


Microsoft Azure Machine Learning
Home
Studio
Gallery
Ashwin Rao Karavadi-Free...
Azure ML Demo - MPG Experiment
Finished running
Properties
Azure ML Demo - MPG Experiment › Score Model › Scored dataset
rows
98
columns
10
Cyl
Displacement
Horsepower
Weight
Acceleration
Year
CountryCode
Model
Scored Labels
8 350 125 3900 17.4 79 1 cadillac eldorado 16.372303
4 140 75 2542 17 74 1 chevrolet vega 22.248202
6 199 90 2648 15 70 1 amc gremlin 17.344169
4 97 60 1834 19 71 2 volkswagen model 111 26.690294
4 140 83 2639 17 75 1 ford pinto 22.492149
4 86 65 2019 16.4 80 3 datsun 310 33.244438
4 97 46 1835 20.5 70 2 volkswagen 1131 deluxe sedan 25.807925
buick opel
ScatterPlot
compare to MPG
MPG
Scored Labels
Feature Selection

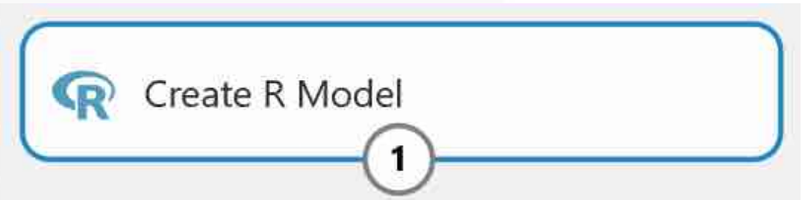
Modelos Disponibles

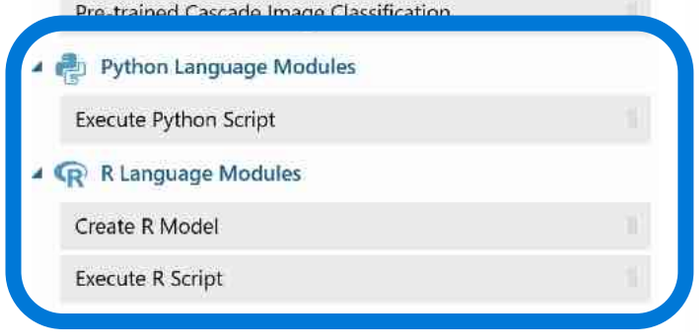
- **Anomaly Detection**
  - One-Class Support Vector Machine
  - PCA-Based Anomaly Detection
- **Classification**
  - Multiclass Decision Forest
  - Multiclass Decision Jungle
  - Multiclass Logistic Regression
  - Multiclass Neural Network
  - One-vs-All Multiclass
  - Two-Class Averaged Perceptron
  - Two-Class Bayes Point Machine
  - Two-Class Boosted Decision Tree
  - Two-Class Decision Forest
  - Two-Class Decision Jungle
  - Two-Class Locally-Deep Support Vector Machine
  - Two-Class Logistic Regression
  - Two-Class Neural Network
  - Two-Class Support Vector Machine
- **Clustering**
  - K-Means Clustering
- **Regression**
  - Bayesian Linear Regression
  - Boosted Decision Tree Regression
  - Decision Forest Regression
  - Fast Forest Quantile Regression
  - Linear Regression
  - Neural Network Regression
  - Ordinal Regression
  - Poisson Regression
- **OpenCV Library Modules**
  - Import Images
  - Pre-trained Cascade Image Classification
- **Python Language Modules**
  - Execute Python Script
- **R Language Modules**
  - Create R Model
  - Execute R Script
- **Text Analytics**
  - Detect Languages
  - Extract N-Gram Features from Text
  - Feature Hashing
  - Latent Dirichlet Allocation
  - Named Entity Recognition
  - Preprocess Text
  - Score Vowpal Wabbit Version 7-10 Model
  - Score Vowpal Wabbit Version 7-4 Model
  - Score Vowpal Wabbit Version 8 Model
  - Train Vowpal Wabbit Version 7-10 Model
  - Train Vowpal Wabbit Version 7-4 Model
  - Train Vowpal Wabbit Version 8 Model

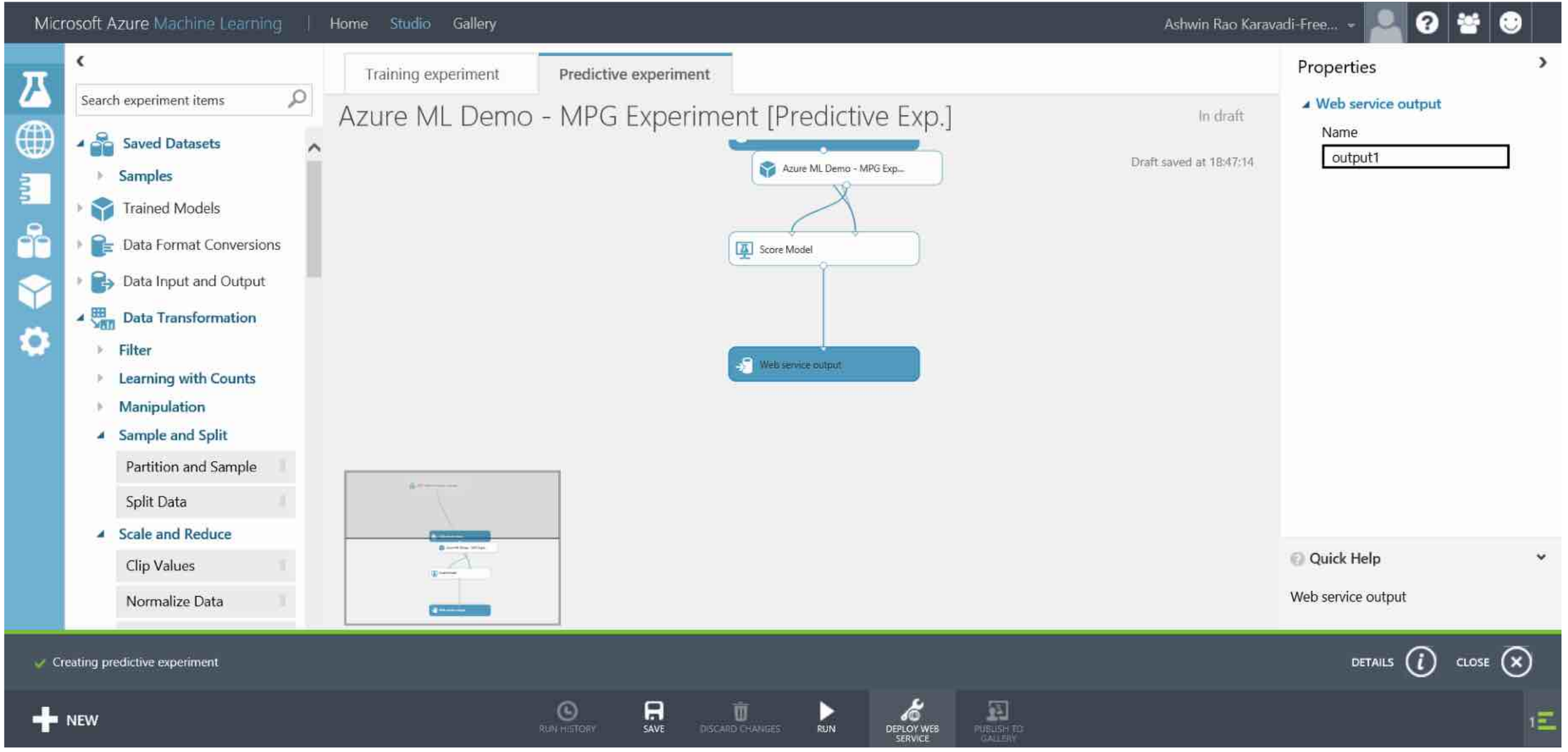
Python + R

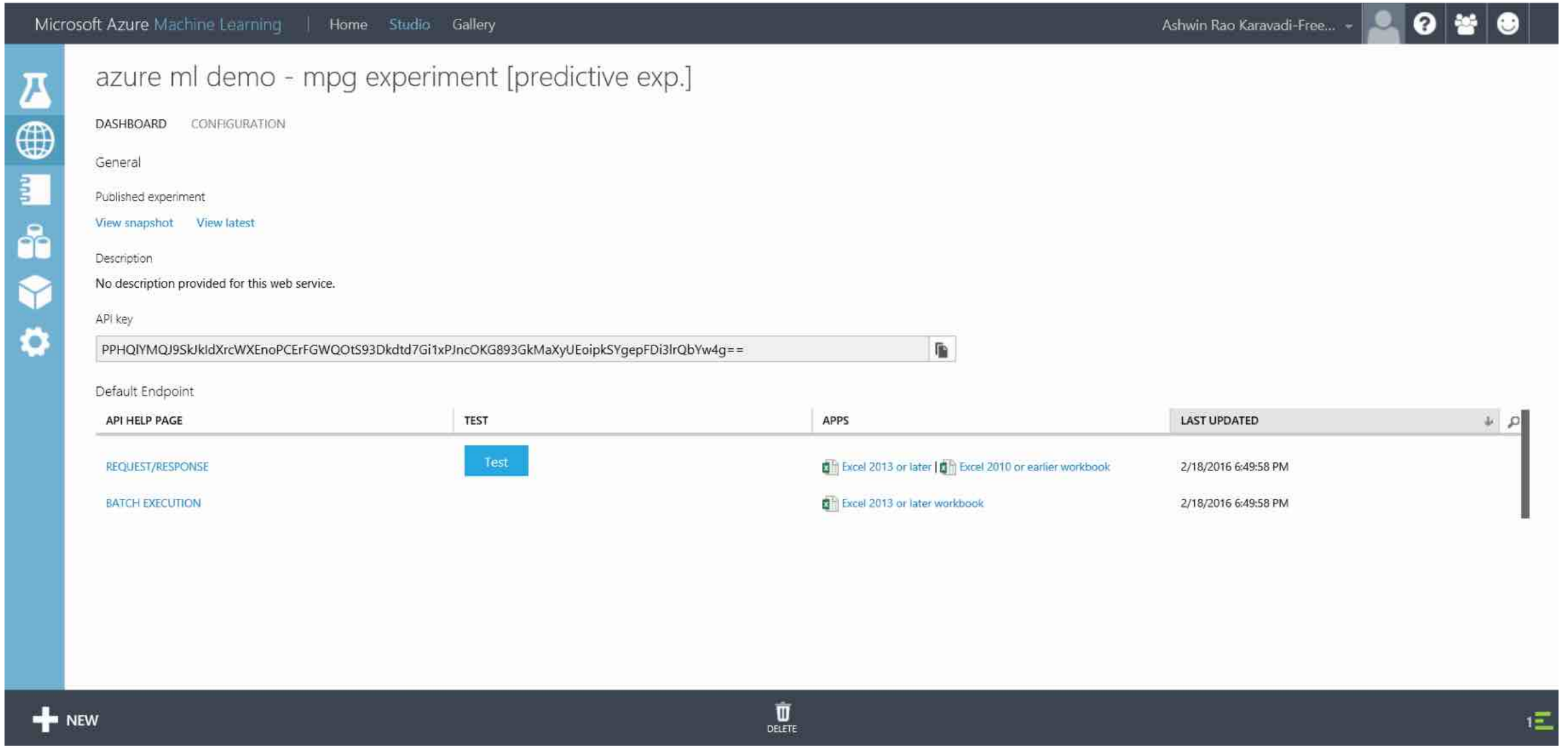

Execute Python Script
1
2


Execute R Script
1
2

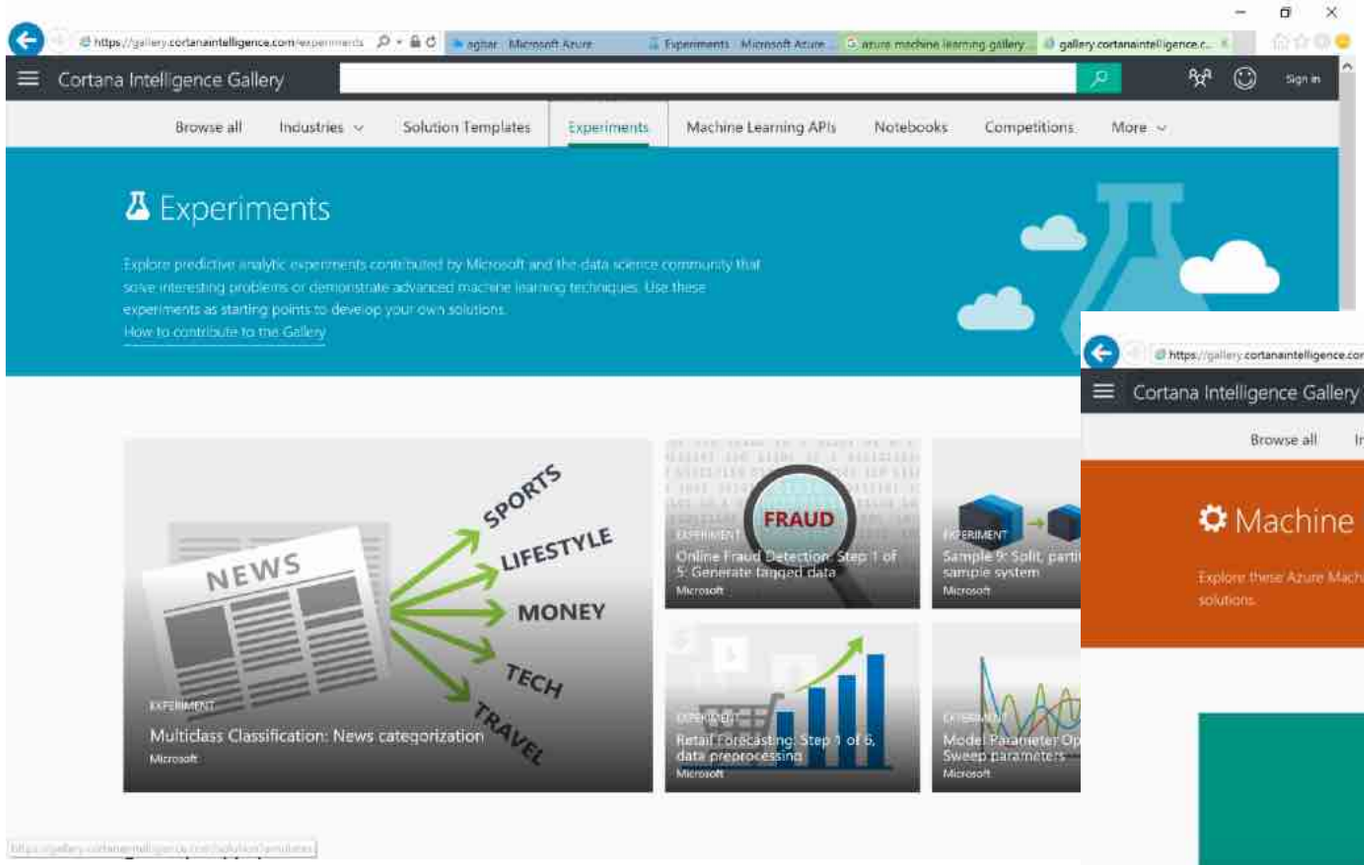

Create R Model
1

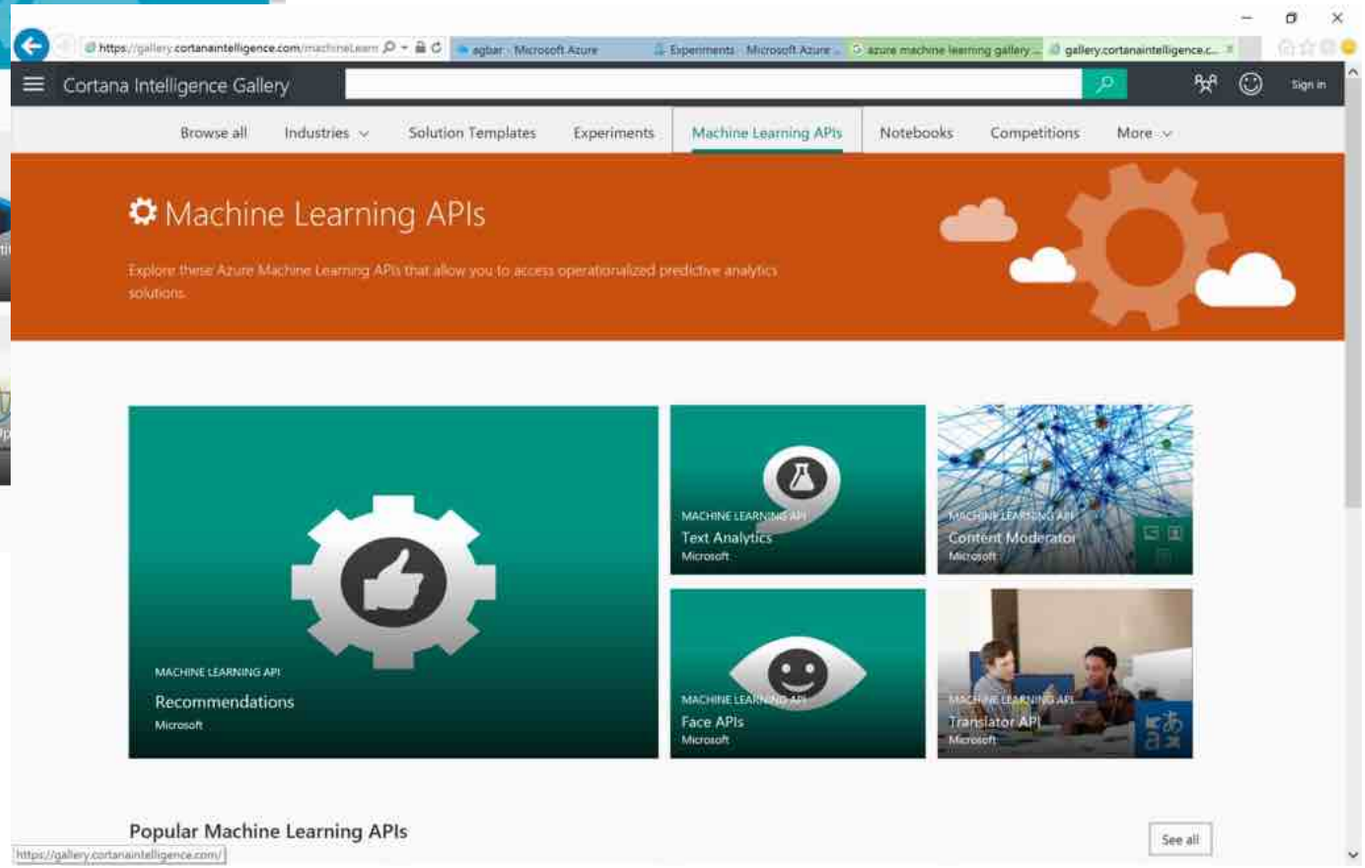

OpenCV Library Modules
Import Images
Python Language Modules
Execute Python Script
R Language Modules
Create R Model
Execute R Script


Microsoft Azure Machine Learning
Home
Studio
Gallery
Ashwin Rao Karavadi-Free...
Search experiment items
Saved Datasets
Samples
Trained Models
Data Format Conversions
Data Input and Output
Data Transformation
Filter
Learning with Counts
Manipulation
Sample and Split
Partition and Sample
Split Data
Scale and Reduce
Clip Values
Normalize Data
Training experiment
Predictive experiment
Azure ML Demo - MPG Experiment [Predictive Exp.]
In draft
Draft saved at 18:47:14
Azure ML Demo - MPG Exp...
Score Model
Web service output
Properties
Web service output
Name
output1
Quick Help
Web service output
Creating predictive experiment
DETAILS
CLOSE
NEW
RUN HISTORY
SAVE
DISCARD CHANGES
RUN
DEPLOY WEB SERVICE
PUBLISH TO GALLERY


Microsoft Azure Machine Learning
Home
Studio
Gallery
Ashwin Rao Karavadi-Free...
azure ml demo - mpg experiment [predictive exp.]
DASHBOARD
CONFIGURATION
General
Published experiment
View snapshot
View latest
Description
No description provided for this web service.
API key
PPHQIYMQJ9SkJkIdXrcWXEnoPCErFGWQOtS93Dkdtd7Gi1xPJncOKG893GkMaXyUEoipkSYgepFDi3IrQbYw4g==
Default Endpoint
API HELP PAGE
TEST
APPS
LAST UPDATED
REQUEST/RESPONSE
Test
Excel 2013 or later
Excel 2010 or earlier workbook
2/18/2016 6:49:58 PM
BATCH EXECUTION
Excel 2013 or later workbook
2/18/2016 6:49:58 PM
NEW
DELETE

Galería


Cortana Intelligence Gallery
Sign in
Browse all
Industries
Solution Templates
Experiments
Machine Learning APIs
Notebooks
Competitions
More
Experiments
Explore predictive analytic experiments contributed by Microsoft and the data science community that solve interesting problems or demonstrate advanced machine learning techniques. Use these experiments as starting points to develop your own solutions.
How to contribute to the Gallery
NEWS
SPORTS
LIFESTYLE
MONEY
TECH
TRAVEL
EXPERIMENT
Multiclass Classification: News categorization
Microsoft
FRAUD
EXPERIMENT
Online Fraud Detection: Step 1 of 5: Generate tagged data
Microsoft
EXPERIMENT
Sample 9: Split, parti
sample system
Microsoft
EXPERIMENT
Retail Forecasting: Step 1 of 6, data preprocessing
Microsoft
EXPERIMENT
Model Parameter Op
Sweep parameters
Microsoft


Cortana Intelligence Gallery
Sign in
Browse all
Industries
Solution Templates
Experiments
Machine Learning APIs
Notebooks
Competitions
More
Machine Learning APIs
Explore these Azure Machine Learning APIs that allow you to access operationalized predictive analytics solutions.
MACHINE LEARNING API
Recommendations
Microsoft
MACHINE LEARNING API
Text Analytics
Microsoft
MACHINE LEARNING API
Content Moderator
Microsoft
MACHINE LEARNING API
Face APIs
Microsoft
MACHINE LEARNING API
Translator API
Microsoft
Popular Machine Learning APIs
See all

Enlaces

Enlaces

- Microsoft R Server
  - https://www.microsoft.com/en/server-cloud/products/r-server/
- Azure Machine Learning
  - https://azure.microsoft.com/en-us/services/machine-learning/
- Cortana Intelligence Gallery
  - https://gallery.cortanaintelligence.com/
- Microsoft Cognitive Services
  - https://www.microsoft.com/cognitive-services/en-us/apis

Microsoft

Juan Carlos Rodriguez García
jurodr@microsoft.com
Data Platform Solution Architect