Data Science with IBM Watson Studio

Yann Gouedo
Data Scientist Leader – Machine Learning / Artificial Intelligence
Marketing / Risk / Fraud / Maintenance / Pricing
Distinguished Data Scientist, Open Group Certification

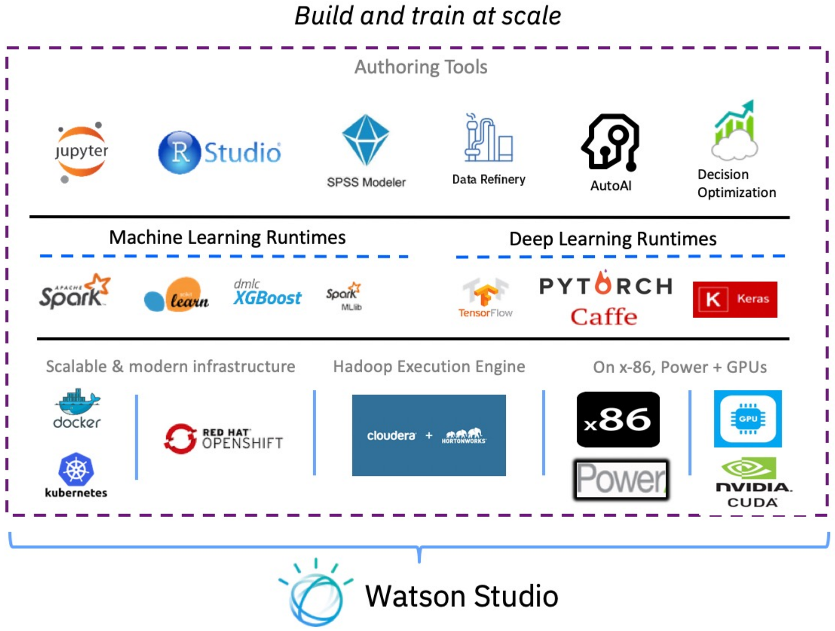
Watson Studio and Watson Machine Learning

Enterprise Data Science platform that helps your team work together to build models to make better data driven decisions for your business

- **Analyze any data, no matter where it lives**
Connect to and analyze your data without moving a single byte through dozens of connectors and multiple deployment options

- **Empower your entire organization with notebooks, visual productivity, and automation tools**
Leverage your entire organization with a variety of tools in a single integrated platform

- **One platform to rule them all from discovery to production**
Analyze data, build predictive models, and seamlessly integrate Watson Machine Learning to deploy at scale



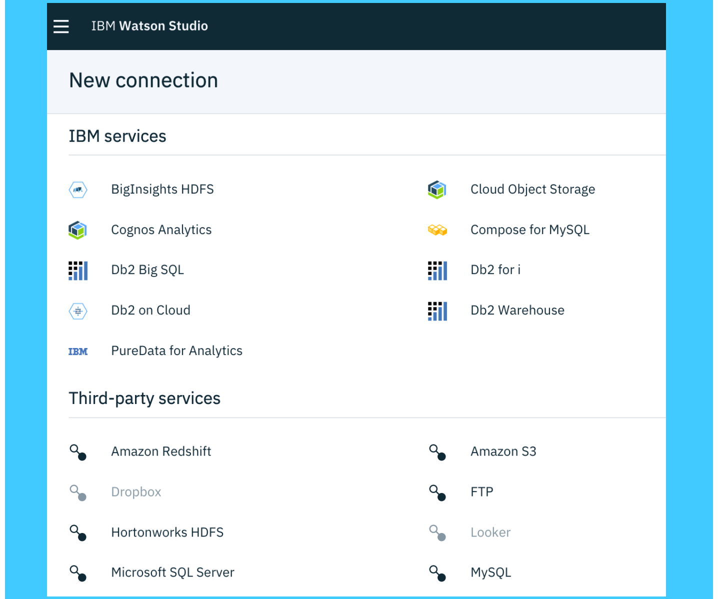
IBM Watson Studio

Enterprise Data Science platform that helps your team work together to build models to make better data driven decisions for your business

Analyze any data, no matter where it lives
Connect to and analyze your data without moving a single byte through dozens of connectors and multiple deployment options

Empower your entire organization with notebooks, visual productivity, and automation tools
Leverage your entire organization with a variety of tools in a single integrated platform

One platform to rule them all from discovery to production
Analyze data, build predictive models, and seamlessly integrate Watson Machine Learning to deploy at scale



IBM Watson Studio

New connection

IBM services

- BigInsights HDFS
- Cloud Object Storage
- Cognos Analytics
- Compose for MySQL
- Db2 Big SQL
- Db2 for i
- Db2 on Cloud
- Db2 Warehouse
- PureData for Analytics

Third-party services

- Amazon Redshift
- Amazon S3
- Dropbox
- FTP
- Hortonworks HDFS
- Looker
- Microsoft SQL Server
- MySQL

- IBM Services like **Cognos** & **DB2**
- 3rd Party Services like **Amazon S3**, **Hadoop**, & **Microsoft SQL Server**
- We have **Public Cloud**, **Private Cloud**, & **Desktop/Server** deployment options

IBM Watson Studio

Enterprise Data Science platform that helps your team work together to build models to make better data driven decisions for your business

Analyze any data, no matter where it lives

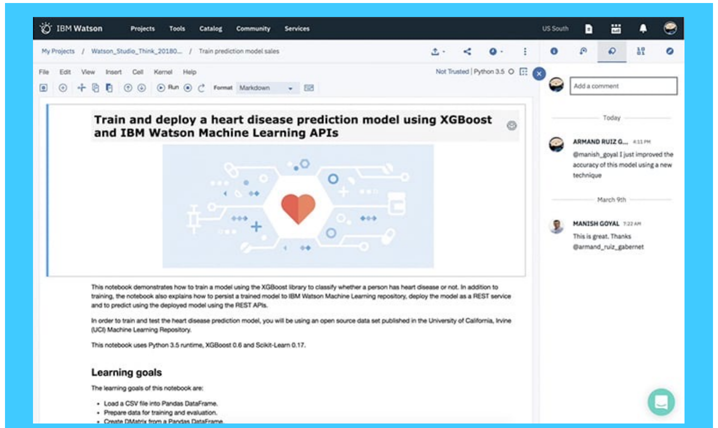
Connect to and analyze your data without moving a single byte through dozens of connectors and multiple deployment options

Empower your entire organization with notebooks, visual productivity, and automation tools

Leverage your entire organization with a variety of tools in a single integrated platform

One platform to rule them all from discovery to production

Analyze data, build predictive models, and seamlessly integrate Watson Machine Learning to deploy at scale



ARMAND RUIZ G... 4:11 PM
@manish_goyal I just improved the accuracy of this model using a new technique

MANISH GOYAL 7:22 AM
This is great. Thanks @armand_ruiz_gabernet

Train and deploy a heart disease prediction model using XGBoost and IBM Watson Machine Learning APIs

This notebook demonstrates how to train a model using the XGBoost library to classify whether a person has heart disease or not. In addition to training, the notebook also explains how to persist a trained model to IBM Watson Machine Learning repository, deploy the model as a REST service and to predict using the deployed model using the REST APIs.

In order to train and test the heart disease prediction model, you will be using an open source data set published in the University of California, Irvine (UCI) Machine Learning Repository.

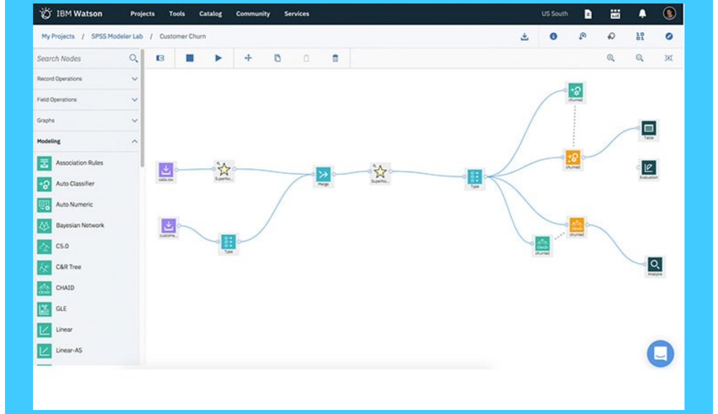
This notebook uses Python 3.5 runtime, XGBoost 0.6 and Scikit-Learn 0.17.

Learning goals

The learning goals of this notebook are:

- Load a CSV file into Pandas DataFrame.
- Prepare data for training and evaluation.
- Create DMatrix from a Pandas DataFrame.

Super charged Jupyter Notebooks & R Studio as most popular IDEs for data scientists well integrated with data connectors and rich set of default environments



Visual tools such as SPSS Modeler, Data Refinery, & AutoAI for non coders to analyze data and build models

IBM Watson Studio

Enterprise Data Science platform that helps your team work together to build models to make better data driven decisions for your business

Analyze any data, no matter where it lives

Connect to and analyze your data without moving a single byte through dozens of connectors and multiple deployment options

Empower your entire organization with notebooks, visual productivity, and automation tools

Leverage your entire organization with a variety of tools in a single integrated platform

One platform to rule them all from discovery to production

Analyze data, build predictive models, and seamlessly integrate Watson Machine Learning to deploy at scale

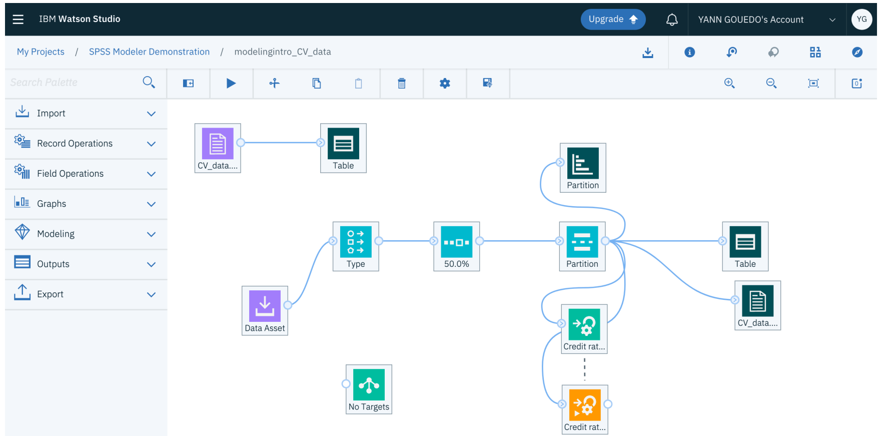
Seamless integration with:

- Watson Knowledge Catalog brings governance & catalog management
- Watson Machine Learning deploys, manages, & automatically retrains your models
- Watson OpenScale offers unmatched levels of trust in automation

VISUAL PROGRAMMING WITH MODELER FLOWS

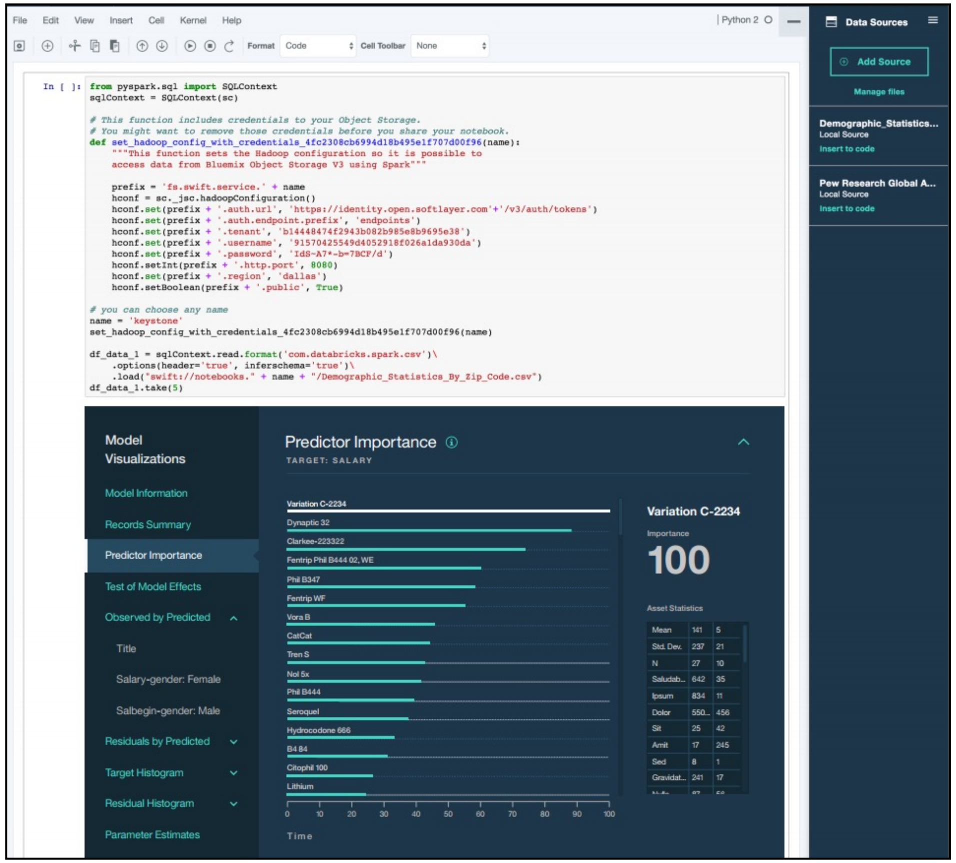
Data Science within IBM Watson Studio

Powerful workbench for code-optional predictive analytics



Delivered in the **cloud** through IBM Watson Studio (SPSS)

SPSS Algorithms in Python, R and Scala and Interactive Model Visualization through IBM Watson Studio

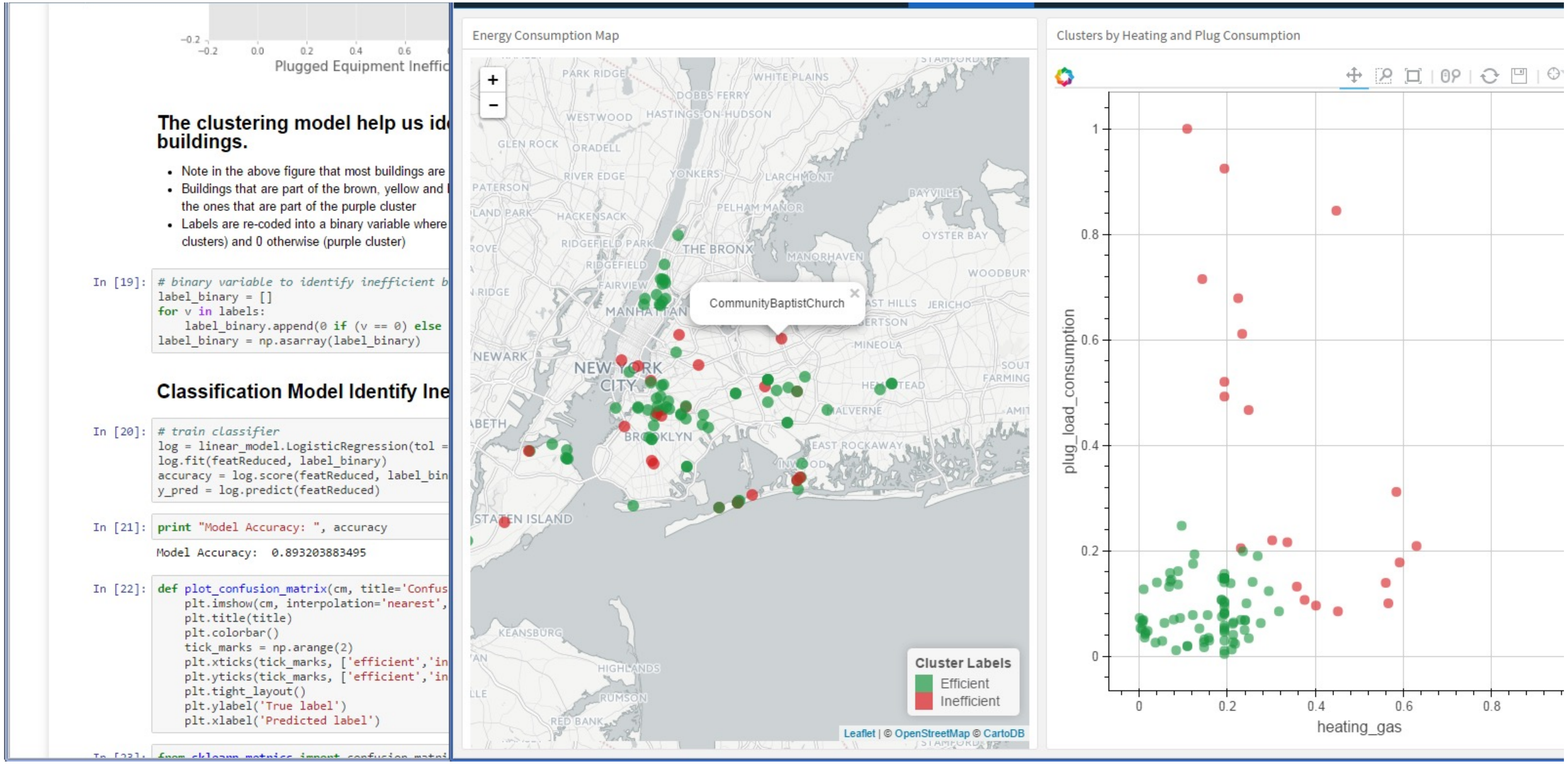


```
In [ ]: from pyspark.sql import SQLContext
        sqlContext = SQLContext(sc)

        # This function includes credentials to your Object Storage.
        # You might want to remove those credentials before you share your notebook.
        def set_hadoop_config_with_credentials_4fc2308cb6994d18b495e1f707d00f96(name):
            """This function sets the Hadoop configuration so it is possible to
            access data from Bluemix Object Storage V3 using Spark"""

            prefix = 'fs.swift.service.' + name
            hconf = sc._jsc.hadoopConfiguration()
            hconf.set(prefix + '.auth.url', 'https://identity.open.softlayer.com'+'/v3/auth/tokens')
            hconf.set(prefix + '.auth.endpoint.prefix', 'endpoints')
            hconf.set(prefix + '.tenant', 'b14448474f2943b082b985e8b9695e38')
            hconf.set(prefix + '.username', '91570425549d4052918f026a1da930da')
            hconf.set(prefix + '.password', 'IdS~A7*-b=7BCF/d')
            hconf.setInt(prefix + '.http.port', 8080)
            hconf.set(prefix + '.region', 'dallas')
            hconf.setBoolean(prefix + '.public', True)

        # you can choose any name
        name = 'keystone'
        set_hadoop_config_with_credentials_4fc2308cb6994d18b495e1f707d00f96(name)

        df_data_1 = sqlContext.read.format('com.databricks.spark.csv')\
            .options(header='true', inferschema='true')\
            .load("swift://notebooks." + name + "/Demographic_Statistics_By_Zip_Code.csv")
        df_data_1.take(5)
```

Mean	141	5
Std. Dev.	237	21
N	27	10
Saludab...	642	35
Ipsum	834	11
Dolor	550...	456
Sit	25	42
Amit	17	245
Sed	8	1
Gravidat...	241	17

Yann Gouedo – Any reproduction forbidden

Interactively explore the analysis of your data science team



The clustering model help us id buildings.

- Note in the above figure that most buildings are
- Buildings that are part of the brown, yellow and the ones that are part of the purple cluster
- Labels are re-coded into a binary variable where clusters) and 0 otherwise (purple cluster)

```
In [19]: # binary variable to identify inefficient b
         label_binary = []
         for v in labels:
             label_binary.append(0 if (v == 0) else
         label_binary = np.asarray(label_binary)
```

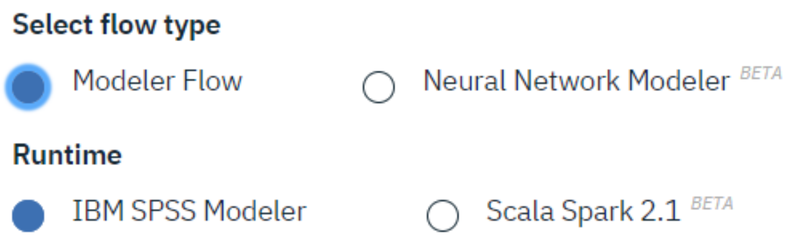
Classification Model Identify Ine

```
In [20]: # train classifier
         log = linear_model.LogisticRegression(tol =
         log.fit(featReduced, label_binary)
         accuracy = log.score(featReduced, label_bin
         y_pred = log.predict(featReduced)

In [21]: print "Model Accuracy: ", accuracy

         Model Accuracy:  0.893203883495

In [22]: def plot_confusion_matrix(cm, title='Confus
             plt.imshow(cm, interpolation='nearest',
             plt.title(title)
             plt.colorbar()
             tick_marks = np.arange(2)
             plt.xticks(tick_marks, ['efficient','in
             plt.yticks(tick_marks, ['efficient','in
             plt.tight_layout()
             plt.ylabel('True label')
             plt.xlabel('Predicted label')

In [23]: from sklearn.metrics import confusion_matri
```

Yann Gouedo – Any reproduction forbidden

Watson Studio Flow Modeler

- Watson Studio recently introduced a Flow Modeler capability:
  - Interactively build UI-driven Machine Learning flows
  - Support for 3 programming models:
    - SPSS flows
      - The flows build on the cloud are interoperable with the on-premises SPSS Modeler flows
    - Watson ML flows
      - Watson ML specific operators
    - Deep Learning flows
      - Produces TensorFlow, Keras, … code for Neural Networks



Select flow type

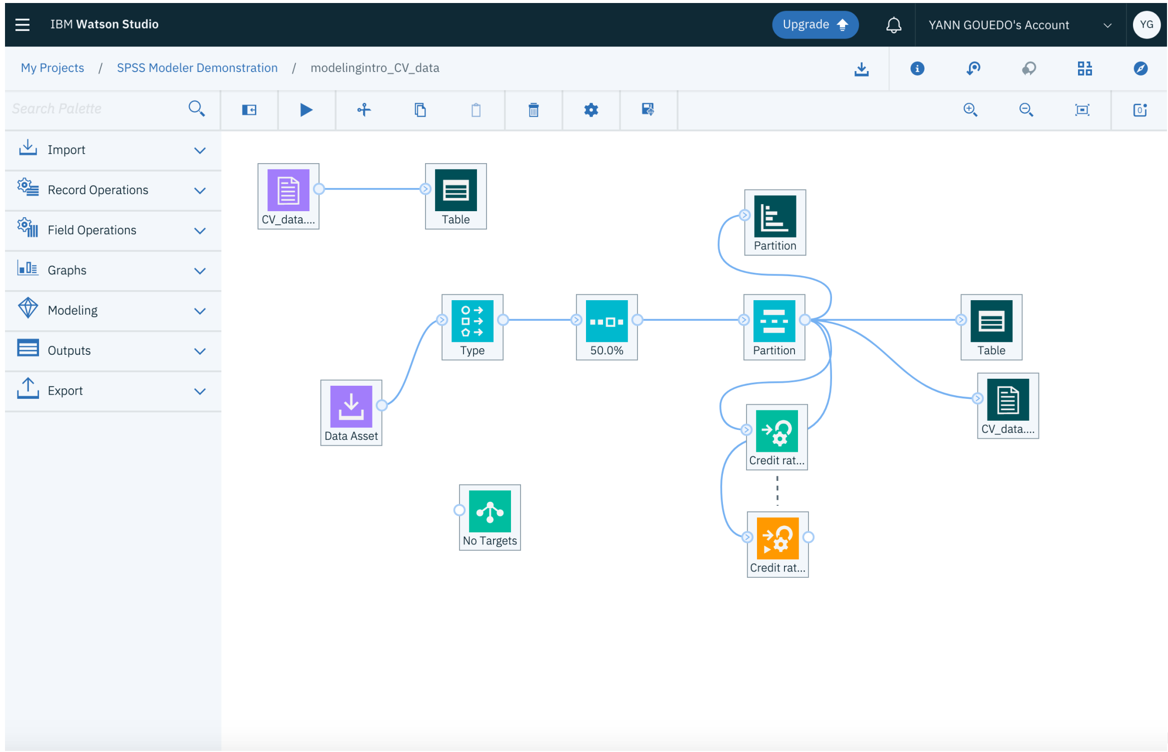
● Modeler Flow ○ Neural Network Modeler BETA

Runtime

● IBM SPSS Modeler ○ Scala Spark 2.1 BETA

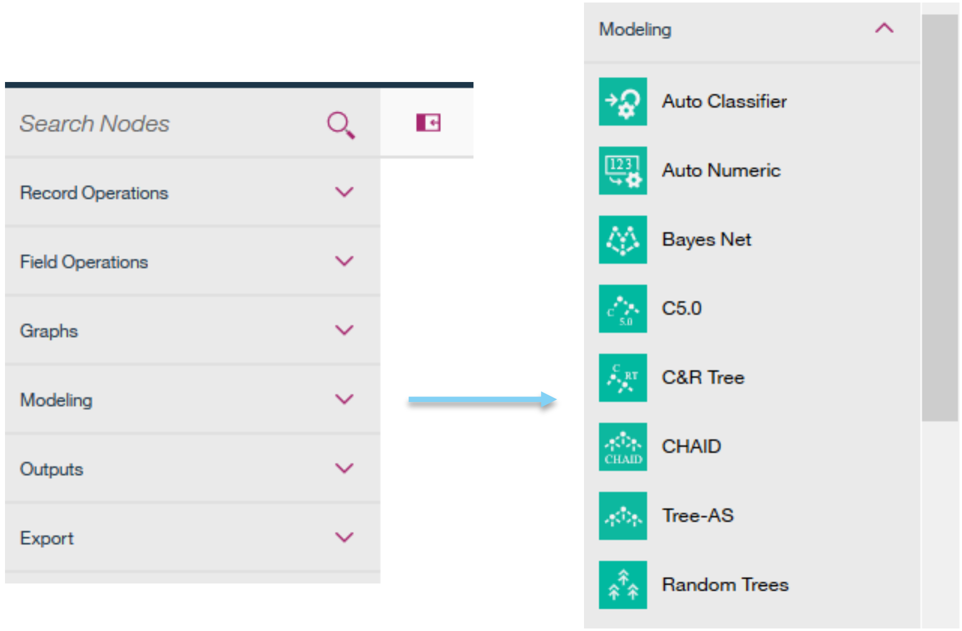
Watson Studio Flow Modeler: introduction

- Watson Studio provides a visual interface for implementing analytics
  - *Cloud version of SPSS Modeler*
  - *Supports data processing pipeline and Machine Learning*
- With this interface we create *flows*:
  - connect to a data source and apply various operations to data



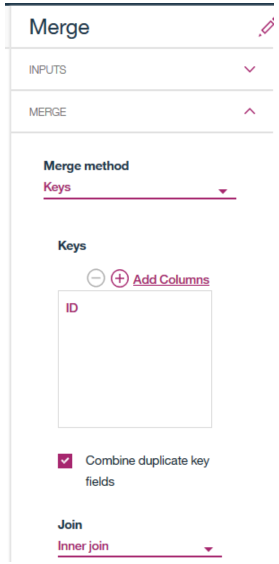
Watson Studio Flow Modeler: operations

- Data “flows” from one visual node to another and is transformed in the process, forming a data processing pipeline
- Operations are organized by type



Watson Studio Flow Modeler: nodes

- Each node (data preparation, graph, algorithm, etc.) provides a dialog box for modifying settings



Merge

INPUTS

MERGE

Merge method
Keys

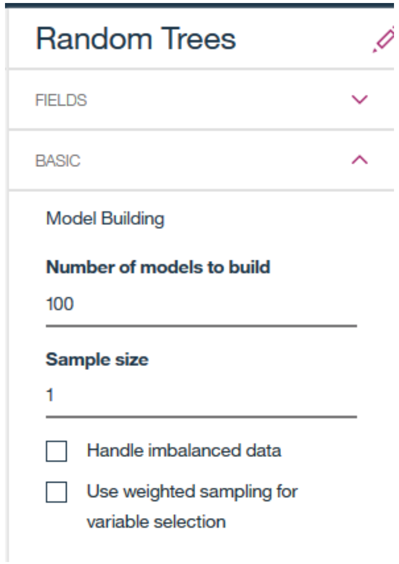
Keys

⊖ ⊕ Add Columns

ID

☑ Combine duplicate key fields

Join
Inner join



Random Trees

FIELDS

BASIC

Model Building

Number of models to build
100

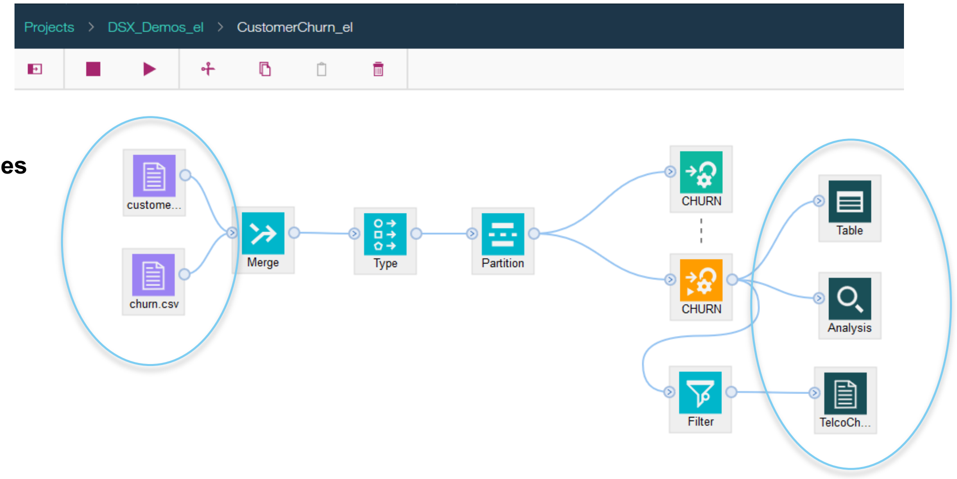
Sample size
1

□ Handle imbalanced data

□ Use weighted sampling for variable selection

Watson Studio Flow Modeler: structure

- All flows must start with the data source
- Most flows end with a data source export: write transformed data to a data source
  - Other “terminal nodes” may be data preview or graph nodes

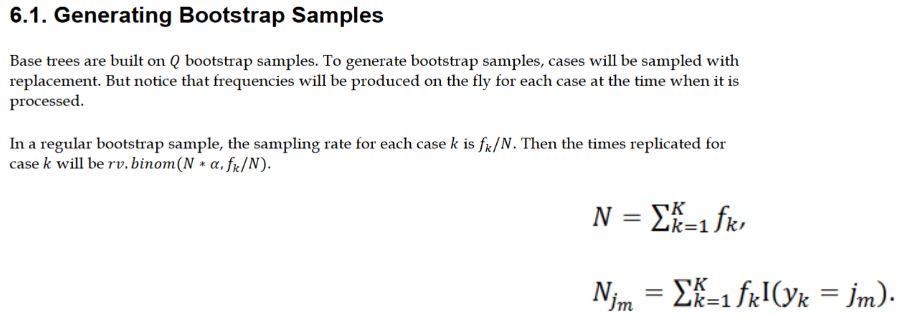


Data sources

Terminal nodes

Watson Studio Flow Modeler: Machine Learning

- Algorithms: IBM publishes the Algorithms Guide that explains how the algorithms are implemented
  - ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/AlgorithmsGuide.pdf
- End users don’t need to understand how algorithms are implemented in order to use it, but this information may be useful for statisticians/data scientists



6.1. Generating Bootstrap Samples

Base trees are built on Q bootstrap samples. To generate bootstrap samples, cases will be sampled with replacement. But notice that frequencies will be produced on the fly for each case at the time when it is processed.

In a regular bootstrap sample, the sampling rate for each case k is f_k/N. Then the times replicated for case k will be $rv.binom(N * \alpha, f_k/N)$.

$$N = \sum_{k=1}^{K} f_k,$$

$$N_{j_m} = \sum_{k=1}^{K} f_k \mathrm{I}(y_k = j_m).$$

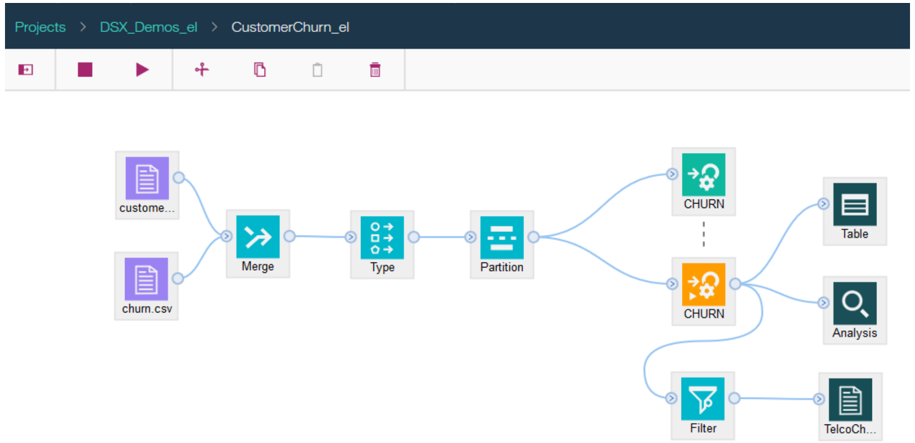
Analytical Techniques

Technique	Algorithms	Usage
Classification (Or Prediction)	Autoclassifiers, Decision Trees, Logistic, Support Vector Machines, Time Series	Predict Group Membership (e.g., Will This Employee Leave?) Or a Number (e.g., How Many Widgets Will I Sell?)
Segmentation	Autoclusters, K-Means, Anomaly Detection	Classify Data Points Into Groups That Are Internally Homogenous and Externally Heterogeneous, Identify Cases That Are Unusual
Association	Apriori, CARMA, Sequence	Find Events That Occur Together Or In a Sequence Market Basket
Geospatial	Space-Time Boxes	Ability to Improve Model Accuracy (for Any Model Type) By Including Inputs Derived From Geospatial Data Sources
Automated	Autoclassified, Autonumeric, Time Series, Clustering	Automatically Find the Right Algorithms Based On Data and Outcome to Create An Ensemble Model
Simulation	Monte Carlo	Run Different Scenarios to Identify Which Is Best From Historical Data Or Generated Data
Specialized	Text Analytics, Entity Analytics, Social Network Analysis	Improve Overall Model Accuracy
In-database	Netezza, DB2, Oracle, Microsoft	Provide User Friendly Interface On Top of Vendor Algorithms
Open Source	R/Spark/Python	Utilize Open Source Algorithms Within Modeler UI. Enhance By Easily Building Custom Dialogs (or downloading from community)

✓ A single analysis project may include multiple techniques

Watson Studio Flow Modeler: deployment

- The flow can be deployed as a web service into Watson Machine Learning
  - Exposed as a REST API
  - Flow Data sources define the input schema
  - Export nodes define the output schema



SPSS Modeler in Watson Studio Flow Modeler

- New flows created in Watson Studio Flow Modeler
  - Does not yet support all nodes that are available in SPSS Modeler Desktop
  - Deployment: supports batch scoring of SPSS flows
  - Differences between Modeler Desktop and Modeler in Watson Studio

- Existing SPSS streams can be imported into Flow Modeler
  - Even if a “visual node” is not yet supported in Watson Studio, the stream will still run

- SPSS streams in Watson Studio run in SPSS runtime (not Spark)
  - One of the main benefits of using SPSS in Watson Studio is a single platform for all analytics assets.

Thank You

Yann Gouedo
Data Scientist Leader – Machine Learning / Artificial Intelligence
Marketing / Risk / Fraud / Maintenance / Pricing
Distinguished Data Scientist, Open Group Certification

What is Watson Studio?

Watson Studio provides a IBM Cloud-based environment for Big Data & Analytics

- Federates Data Science Analysis resources under a single workbench

- **Dev:** Collaborative Tooling for Data Exploration, Analysis and Machine Learning
  - Programming environments: Jupyter and R Studio
  - Interactive UI-driven tools
    - Data Refinery (ETL)
    - Flow Modeler (SPSS, WML, DL)

- **Ops:** Runtime and Deployment
  - **Storage resources**
    - Cloud Object Storage
  - **Compute resources**
    - Managed operational runtime environments for notebooks
    - Spark Engines integration
    - Watson Machine Learning

Watson Studio Environment – Components

BUILD

jupyter

ML Wizard

python™

Scala

DATA

Object STORAGE

DATABASES

HADOOP

DEPLOY

APACHE Spark

XGBoost

PMML Predictive Model Markup Language

scikit learn

K

TensorFlow™

RUNTIME

- **Monitoring & Alerting**
- **Model retraining**
- **KPI Dashboards**
- **Model Refactoring**
- **Security**

Components of Watson Studio

Tools provided by the IBM Watson Studio. Some of the tools are based on or leverage open source and can interoperate with other platforms.

- **Jupyter Notebooks:** Notebooks are used by data scientists to clean, visualize, and understand data. DSX uses Jupyter Notebooks, but notebooks come in different flavors. In DSX you code your notebooks mainly in Python or Scala.
- **Object Storage:** DSX a part of the IBM Cloud and Watson Data Platform provides integrated Object Storage used to store large amounts of data
- **Apache Spark ML:** Spark ML is a library for building ML pipelines on top of Apache Spark. Spark ML includes algorithms and APIs for supervised and unsupervised machine learning problems.
- **IBM Watson ML:** Watson ML is a service for deploying ML models and making predictions at runtime. Watson ML provides a REST API to your ML models which can be called directly from your application or your middleware.
- **IBM Data Refinery Flow:** UI-driven data cleansing ETL and feature engineering pipeline which are managed and deployable.
- **IBM Flow Modeler:** Interactive UI-driven tool to build managed deployable machine learning pipelines, addressing SPSS, SparkML or Deep Learning frameworks.
- **IBM Dashboards:** UI-driven interactive dashboard building capability, which can be published
- **IBM Streaming Flows:** UI-driven streaming analytics that can act on real-time streaming data.