

HKSJU Cloud Pak for Data

- Existing Services

INFUSE

Operationalize AI throughout the business

ANALYZE

Build and scale AI models

Watson Studio - the environment and tools to collaborate and work on data

- AutoAI – graphical tool to automatically analyzes data and generates candidate model pipelines customized for predictive modelling
- Data Refinery – pre-defined operations to transform raw data to consumable, quality data ready for analysis
- Jupyter Notebook – web-based environment for interactive programming

Watson Machine Learning – full range of tools and services to build, train and deploy machine learning models

ORGANIZE

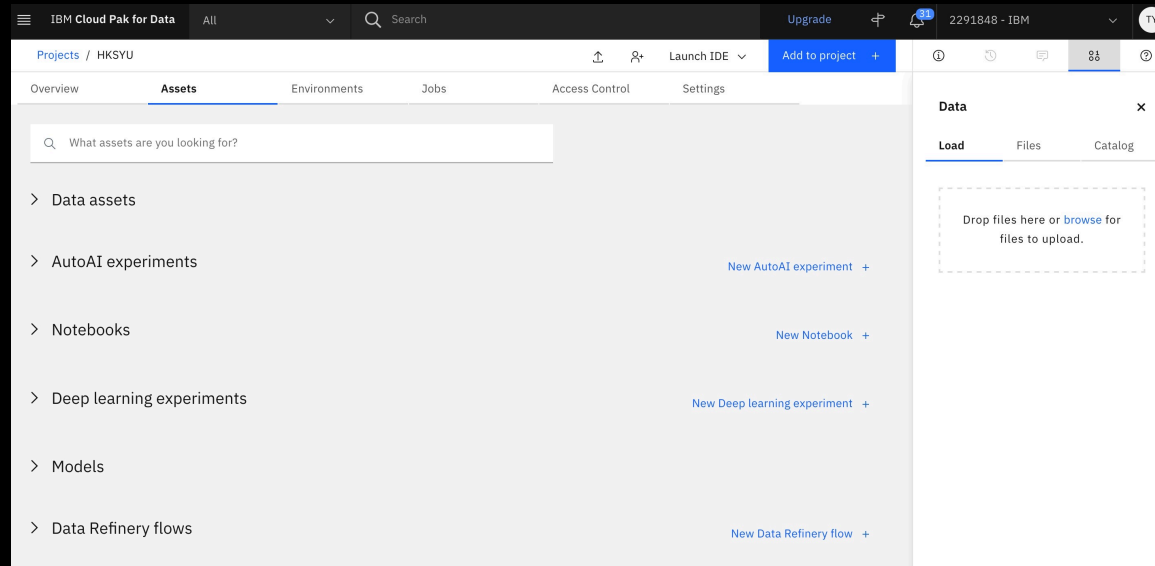
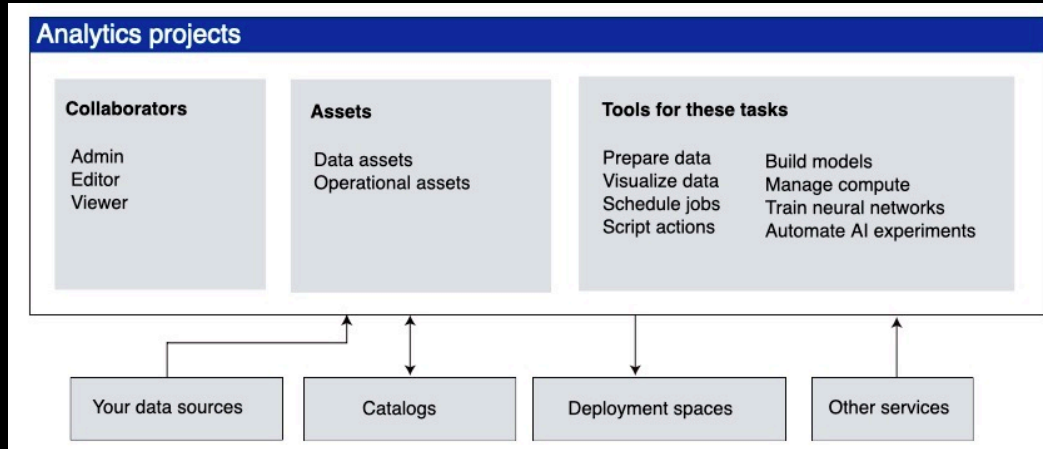
Create a business-ready analytics foundation

COLLECT

Make your data simple and accessible

Watson Studio

Custom environment and tools to collaborate and work on data



Multiple Tools for Data

Provide an integrated platform for both IBM and open source tools with best practices

Business Value: provides the extensibility and flexibility needed for migrating data science team's open source work to Cloud Pak for Data for increased security and productivity

Project Centered

You can have these types of resources in a project:

- Data assets
- Data Refinery: Prepare and visualize data.
- Jupyter notebook editor
- JupyterLab IDE: Code Jupyter notebooks and Python scripts with Git integration.

Collaboration

- Automatic **code generation** of AutoAI pipelines; Users can save pipeline code in Python notebook and run in their favorite IDE.
- AutoAI performance improvements and configuration options

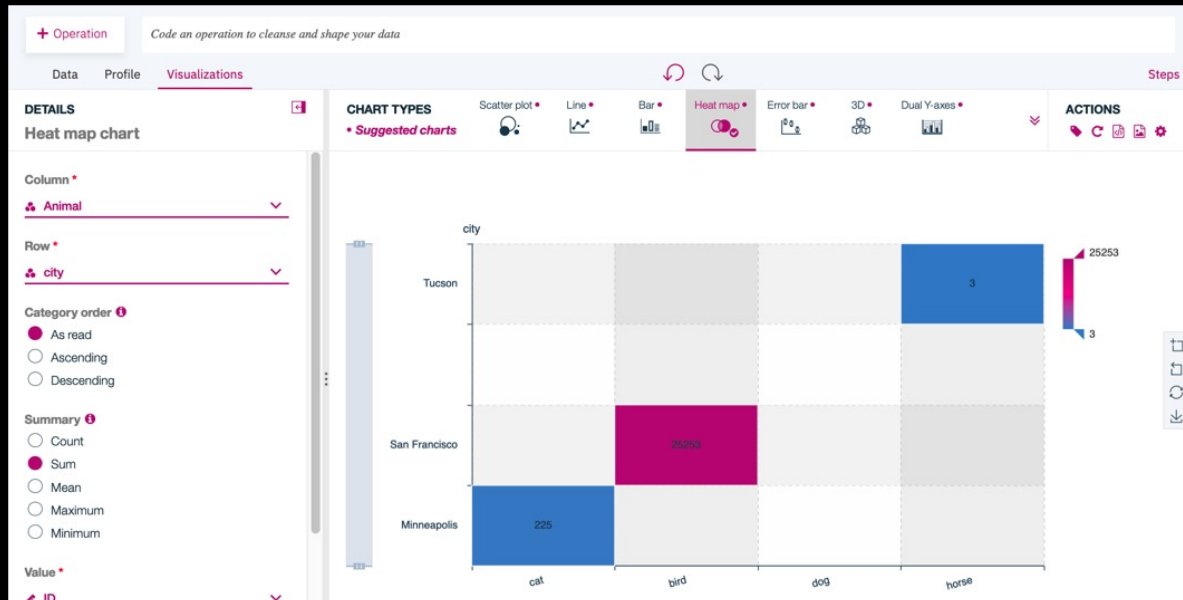
Business Value:

- AutoAI is already placing in the **top 20% in open data science competitions** – can reduce FTEs and skills requirements
- Build ML/AI models in **minutes instead of weeks**
- **Zero lock-in** with autogenerated Python code

Data Refinery

Custom environment and tools to collaborate and work on data

The screenshot shows the Data Refinery interface. On the left, there's a sidebar with a search bar and a list of operations: Calculate, Convert column type, Filter, Math, Remove, Rename, Sort ascending, and Sort descending. The main area displays a table with columns: Animal (String), city (String), ID (Integer), and Phone (Integer). The data rows are: cat (Minneapolis, 225, 444), bird (San Francisco, 25253, 664), dog (NA, 866), and horse (Tucson, 3, NA). On the right, there's a '1 STEPS' section showing the 'Data Source' as 'mydata.csv' and a 'Convert column type' step marked as 'AUTOMATIC' with a description: 'Automatically converted one or more columns to inferred data types.'



Analyze and transform data

Interactively discover, cleanse, and transform your data with over 100 built-in operations. No coding skills are required.

Profile and visualize data

Understand the quality and distribution of your data using dozens of built-in charts, graphs, and statistics. Automatically detect data types and business classifications.

Schedule job execution

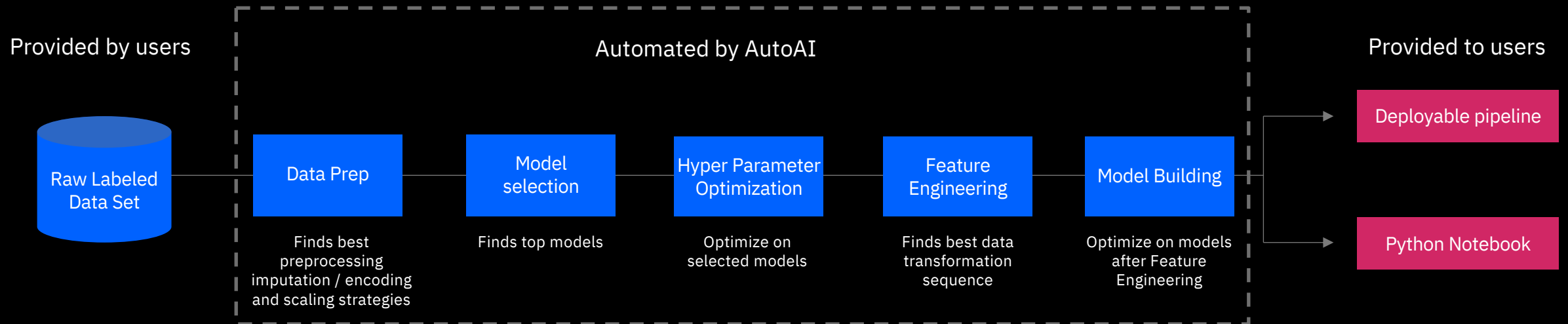
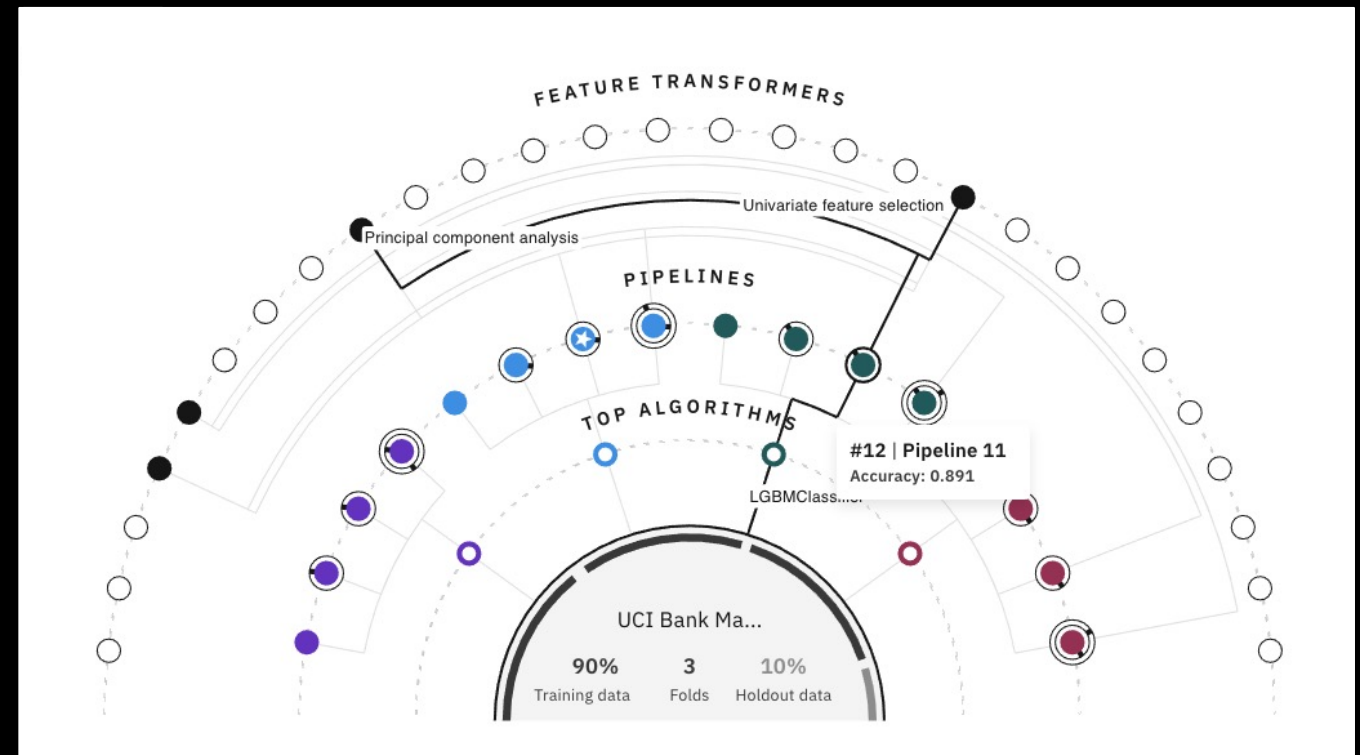
Schedule data flow executions for repeatable outcomes. Monitor results and receive notifications.

Serverless Execution

Easily scale out via Apache Spark to apply transformation recipes on full data sets. No management of Apache Spark clusters needed.

AutoAI

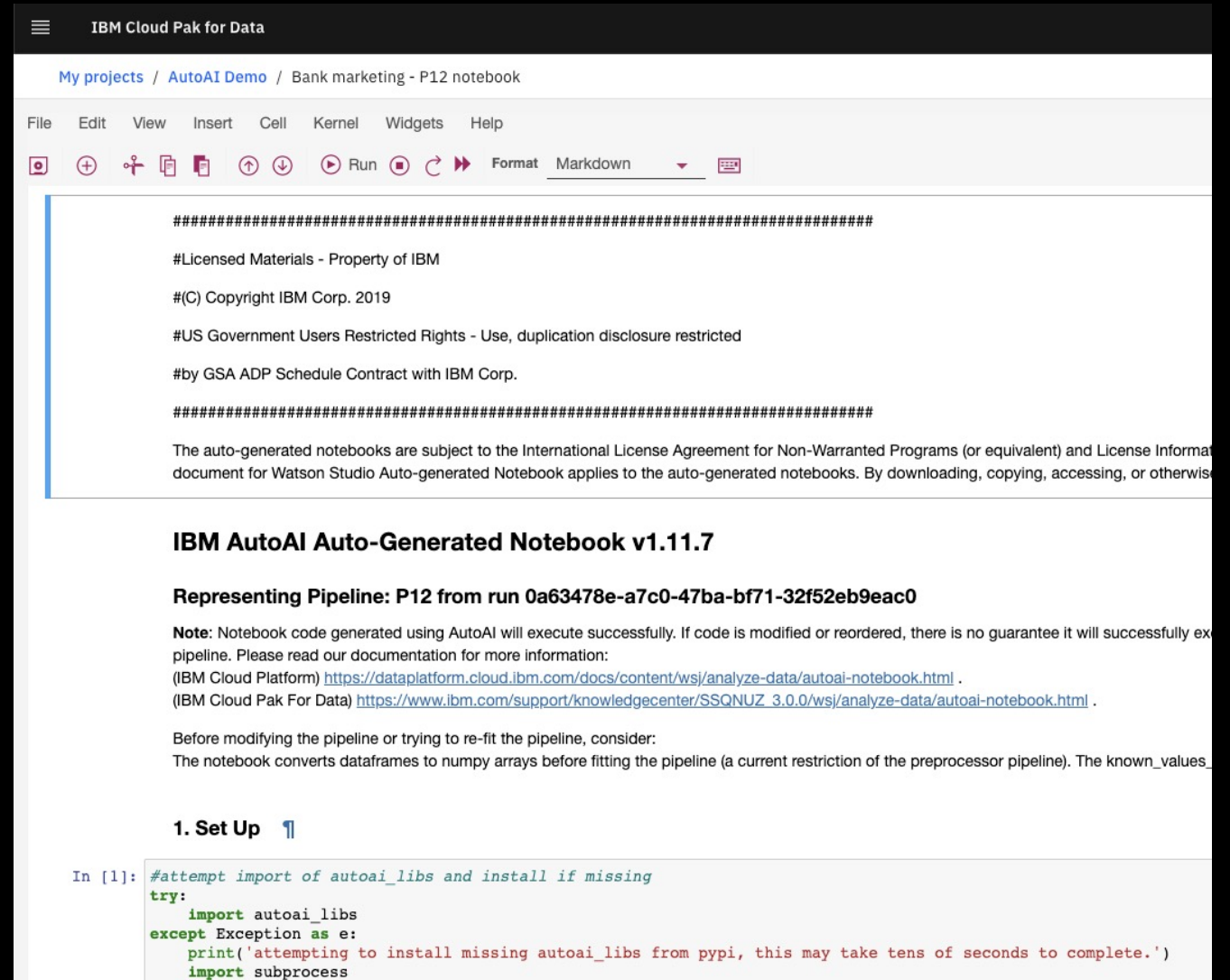
- Deliver AI proof-of-concept in **weeks** instead of months¹
- Increase model performance by **200%** or more²
- Reduce **53%** human error³



AutoAI

No black box with automatic code generation

- Python code **automatically written in seconds instead manually writing it for days**
- **Zero lock-in**, exportable Python code
- Seamless integration with model deployment service

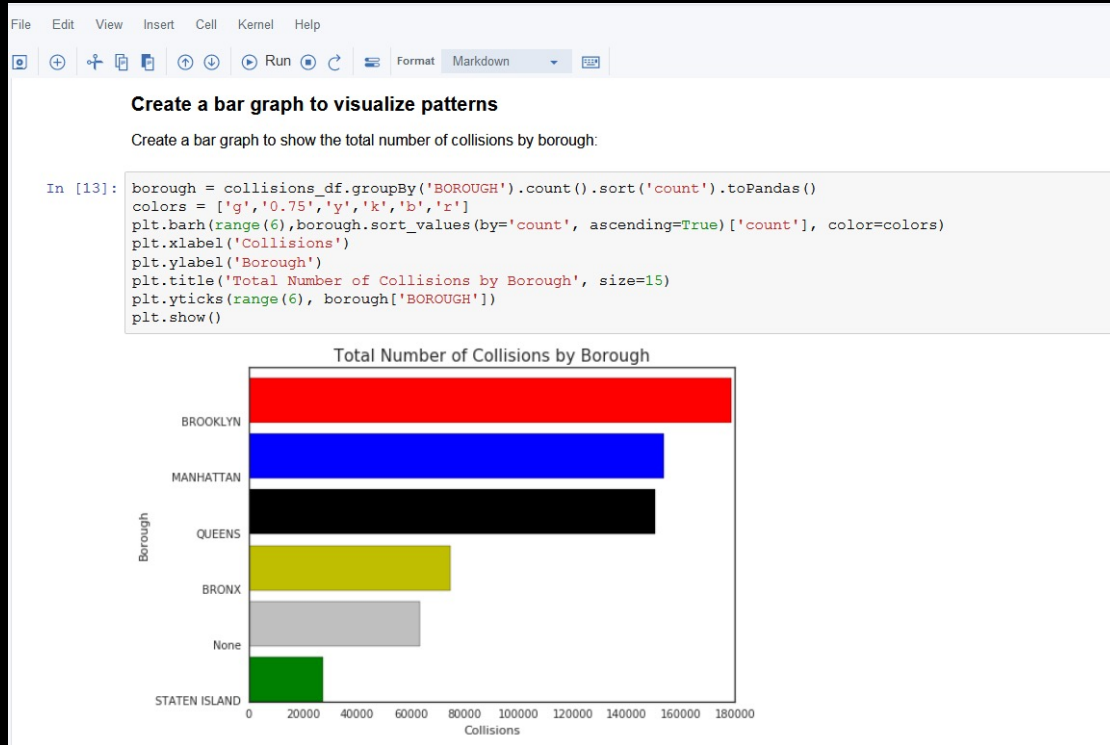


The screenshot displays the IBM Cloud Pak for Data interface. At the top, the title bar reads 'IBM Cloud Pak for Data'. Below it, the breadcrumb navigation shows 'My projects / AutoAI Demo / Bank marketing - P12 notebook'. A menu bar includes 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', 'Widgets', and 'Help'. A toolbar contains icons for file operations, a 'Run' button, and a 'Format' dropdown set to 'Markdown'. The main content area shows a notebook with a license notice and a section titled 'IBM AutoAI Auto-Generated Notebook v1.11.7'. This section includes a 'Representing Pipeline: P12 from run 0a63478e-a7c0-47ba-bf71-32f52eb9eac0' and a 'Note' explaining that the code is auto-generated and may not always execute successfully. It provides links to documentation for both the IBM Cloud Platform and IBM Cloud Pak for Data. A warning advises against modifying the pipeline before understanding its structure. The notebook concludes with a '1. Set Up' section containing a code cell with Python code to attempt importing 'autoai_libs' and install it if missing.

```
In [1]: #attempt import of autoai_libs and install if missing
try:
    import autoai_libs
except Exception as e:
    print('attempting to install missing autoai_libs from pypi, this may take tens of seconds to complete.')
    import subprocess
```

Notebook

Web-based environment for interactive computing



Include all building blocks needed for data

Including:

- Data
- The code computations that process the data
- Visualizations of the results

Includes preinstalled libraries for Python & R

Including:

- Spark libraries, such as SparkSQL, Spark Streaming, Spark Mllib
- [Python project-lib library](#)
- [R project-lib library](#)

Best Practice

- IBM provide industry accelerators that help you solve a specific business problems, whether it's preventing credit card fraud in the banking industry or optimizing the efficiency of your contact center.
- Full list of [IBM Industry Accelerator](#)
- Most accelerators include a Sample analytics project with everything you need to analyze data, build a model, and display results. The sample projects include detailed instructions, data sets, Jupyter notebooks, models, and R Shiny applications.

Watson Machine Learning

Deployment space to manage models

Deployments / ModelOps Space					
ModelOps Space					
Assets Deployments Jobs Manage					
What assets are you looking for?					
Models (2) Import model +					
Name	Type	Software specification	Tags	Last modified	↓
customer_churn_model_1	scikit-learn_0.23	default_py3.7		Mar 1, 2021 4:00 PM	
customer_churn_model	scikit-learn_0.23	default_py3.7		Mar 1, 2021 3:13 PM	
Data assets (3) Import Data assets +					
Name	Type	Tags	Last modified	↓	
Churn Results - Batch API.csv	Data asset		Mar 1, 2021 4:05 PM		
CutoomerChurnScoringResults.csv	Data asset		Mar 1, 2021 3:50 PM		
new_customers.csv	Data asset		Mar 1, 2021 3:48 PM		

Deployments (4)						
Name	Type	Status	Asset	Tags	Last modified	↓
Churn Deployment via API-Batch	Batch	Deployed	customer_churn_model_1		Mar 1, 2021 4:04 PM	
Churn Deployment via API-Online	Online	Deployed	customer_churn_model_1		Mar 1, 2021 4:01 PM	
ModelOps	Batch	Deployed	customer_churn_model		Mar 1, 2021 3:41 PM	
ModelOps	Online	Deployed	customer_churn_model		Mar 1, 2021 3:31 PM	

Options to deploy

Use the tools available from a deployment space to deploy and run models and scripts. The types of deployment available for a model and the type of input supported depends on the model framework.

- Online → real time scoring
- Batch → process batches of input from data files

Why deploy?

Deployment is the process of configuring an analytic asset for integration with other applications or access by business users. Several types of analytics assets can be deployed in CPD including Spark, PMML, AutoAI, SPSS, Scikit-Learn, XGBoost, TensorFlow, Keras, PyTorch, Python Function & Script, RScript

Full list [here](#)

Deployment of Notebook as Jobs Supported

This is the advanced topic. Few examples for jobs:

- A data scientist needs to run a script to prepare data for daily analysis. The script runs at night, and prepared data is available for the data scientist in the morning.
- A data scientist needs to prepare data for modeling. Since the data preparation task runs for a few hours, he would like to run it as a job.
- A data scientist wants to train or retrain a model in batch mode.