



## Master the art of data science.

Enjoy the best of open source and collaborate in a social environment, built for data scientists by data scientists.



# Data Science Experience Local on IBM Integrated Analytics System Backup Hands-On Lab

*IBM NA Big Data Science Technical Team*

*October 2017*

---

## Contents

<b>INTRO</b>	<b>IBM DATA SCIENCE EXPERIENCE LOCAL ON IBM INTEGRATED ANALYTICS SYSTEM.....</b>	<b>3</b>
PART 1	JUPYTER NOTEBOOK THAT PREDICTS OUTDOOR EQUIPMENT PURCHASES .....	3
PART 2	USING THE MACHINE LEARNING BUILDER .....	6

---

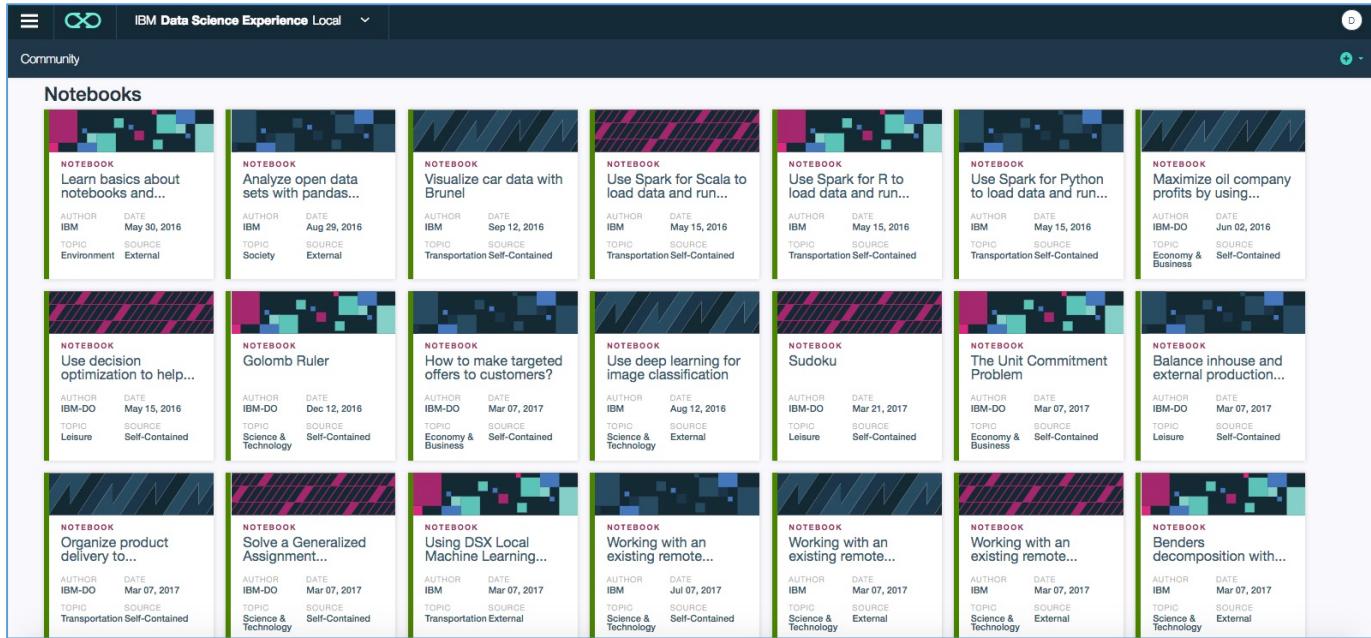
# Intro      IBM Data Science Experience Local on IBM Integrated Analytics System

IBM Data Science Experience (DSX) Local is an out-of-the box on premises enterprise solution for data scientists and data engineers. It offers a suite of data science tools that are integrated with proprietary IBM technologies. It also seamlessly integrates with RStudio, Spark, Jupyter and Zeppelin notebook technologies. The intuitive user interface provides a collaborative Projects space for teams and individuals to reduce the time to value.

## **Part 1      Jupyter Notebook that Predicts Outdoor Equipment Purchases**

This part of the lab will import a Jupyter notebook that will predict product line purchases based on a consumer's gender, age, marital status and job. This notebook will show you how to load and explore data, create a machine learning model, persist the model, predict locally and visualize, deploy and score the model.

- a) Open your browser and navigate to <http://ibm.biz/dsxlocal> to access the DSX Local sign in screen. Use your 'team##' with a password of 'sailfish##', where ## is your team number.
- b) After a successful sign in, the following Community screen will appear. Please take a few moments to view the various assets available to get started quickly.



The screenshot shows the 'Community' screen of the IBM Data Science Experience Local interface. At the top, there is a navigation bar with icons for Home, Sign In, and Help. Below the navigation bar, the title 'Community' is displayed. The main area is titled 'Notebooks' and contains a grid of 21 thumbnail cards, each representing a different Jupyter notebook. The notebooks are organized in three rows of seven. Each card includes a small preview image, the title, a brief description, and metadata such as author, date, topic, and source. The topics covered include various data science and machine learning applications, such as Scala, R, Python, and Scala on Brunei, as well as specific projects like 'Golomb Ruler' and 'Sudoku'.

Thumbnail	Title	Description	Author	Date	Topic	Source
	Learn basics about notebooks and...	Learn basics about notebooks and...	Author: IBM TOPIC: Environment	DATE: May 30, 2016	SOURCE: External	
	Analyze open data sets with pandas...	Analyze open data sets with pandas...	Author: IBM TOPIC: Society	DATE: Aug 29, 2016	SOURCE: External	
	Visualize car data with Brunei	Visualize car data with Brunei	Author: IBM TOPIC: Transportation	DATE: Sep 12, 2016	SOURCE: Self-Contained	
	Use Spark for Scala to load data and run...	Use Spark for Scala to load data and run...	Author: IBM TOPIC: Transportation	DATE: May 15, 2016	SOURCE: Self-Contained	
	Use Spark for R to load data and run...	Use Spark for R to load data and run...	Author: IBM TOPIC: Transportation	DATE: May 15, 2016	SOURCE: Self-Contained	
	Use Spark for Python to load data and run...	Use Spark for Python to load data and run...	Author: IBM TOPIC: Transportation	DATE: May 15, 2016	SOURCE: Self-Contained	
	Maximize oil company profits by using...	Maximize oil company profits by using...	Author: IBM-DO TOPIC: Economy & Business	DATE: Jun 02, 2016	SOURCE: Self-Contained	
	Use decision optimization to help...	Use decision optimization to help...	Author: IBM-DO TOPIC: Leisure	DATE: May 15, 2016	SOURCE: Self-Contained	
	Golomb Ruler	Golomb Ruler	Author: IBM-DO TOPIC: Science & Technology	DATE: Dec 12, 2016	SOURCE: Self-Contained	
	How to make targeted offers to customers?	How to make targeted offers to customers?	Author: IBM-DO TOPIC: Economy & Business	DATE: Mar 07, 2017	SOURCE: Self-Contained	
	Use deep learning for image classification	Use deep learning for image classification	Author: IBM-DO TOPIC: Science & Technology	DATE: Aug 12, 2016	SOURCE: External	
	Sudoku	Sudoku	Author: IBM-DO TOPIC: Leisure	DATE: Mar 21, 2017	SOURCE: Self-Contained	
	The Unit Commitment Problem	The Unit Commitment Problem	Author: IBM-DO TOPIC: Economy & Business	DATE: Mar 07, 2017	SOURCE: Self-Contained	
	Balance inhouse and external production...	Balance inhouse and external production...	Author: IBM-DO TOPIC: Leisure	DATE: Mar 07, 2017	SOURCE: Self-Contained	
	Organize product delivery to...	Organize product delivery to...	Author: IBM-DO TOPIC: Transportation	DATE: Mar 07, 2017	SOURCE: Self-Contained	
	Solve a Generalized Assignment...	Solve a Generalized Assignment...	Author: IBM-DO TOPIC: Science & Technology	DATE: Mar 07, 2017	SOURCE: Self-Contained	
	Using DSX Local Machine Learning...	Using DSX Local Machine Learning...	Author: IBM TOPIC: Transportation	DATE: Mar 07, 2017	SOURCE: External	
	Working with an existing remote...	Working with an existing remote...	Author: IBM TOPIC: Science & Technology	DATE: Jul 07, 2017	SOURCE: Self-Contained	
	Working with an existing remote...	Working with an existing remote...	Author: IBM TOPIC: Science & Technology	DATE: Mar 07, 2017	SOURCE: External	
	Benders decomposition with...	Benders decomposition with...	Author: IBM TOPIC: Science & Technology	DATE: Mar 07, 2017	SOURCE: External	

- c) Now click on the hamburger icon on the top left, then click on ‘Projects’ and ‘View all Projects’.

The screenshot shows the IBM Data Science Experience Local interface. In the top left, there's a hamburger menu icon. To its right is the title 'IBM Data Science Experience Local'. Below the title, there's a 'Community' section. Underneath that is a 'Projects' section with a dropdown arrow. A red circle highlights the 'View all Projects' button. Below this are sections for 'Recent Projects' and 'dsx-samples'. Further down are 'Model Management' and 'Tools' sections, each with a dropdown arrow. On the right side, there are two notebook cards: one titled 'NOTEBOOK' with the subtitle 'Analyze open data sets with pandas...', and another titled 'NOTEBOOK' with the subtitle 'Visualize car data with Brunel'.

- d) Click on ‘create project’.

The screenshot shows the 'Project List' page. At the top right, there's a red circle highlighting the '+ create project' button. The main area displays a table with columns for 'NAME' and 'LAST MODIFIED'. One row shows 'dsx-samples' last modified on '10-16-2017'. There are three dots at the bottom right of the table and a red circle on the far right edge of the page.

- e) Then enter your team name, ‘Team##’ for ‘Name’ and click on ‘Create’.

The screenshot shows the 'Create new project' page. At the top, there are three tabs: 'New', 'From File', and 'From Github'. The 'New' tab is selected. Below it, there's a 'Name \*' field containing 'Team00', which has a green checkmark next to it. A message 'This name is valid' appears below the field. There's also a 'Description' field with a placeholder 'Type your description here' and a character count of '94'.

- f) Click on ‘Create notebook’ from the pull down in the top right ‘+’ icon.

The screenshot shows the 'MyProject' overview page. At the top right, there's a red circle highlighting the 'Create notebook' option in the top right corner of the asset summary section. The page displays various metrics: 0 Assets, 0 Runtimes, 0 Data Sources, and 1 Collaborator. It also shows a 'Recent Assets' section with a table and a note 'You have no recently modified assets'.

- g) In the ‘Create Notebook’ page, click on ‘From URL’, then enter ‘DSXLocalSailfish’ for the ‘Name’, select ‘Jupyter’ for ‘Tool’, and enter the following for the ‘Notebook URL’.  
<https://raw.githubusercontent.com/dxkikuchi/DSX/master/DSXLocalSailfish/DSXLocalSFBU.ipynb>  
Then click on ‘Create’.

IBM Data Science Experience Local

Projects > My Project > Create Notebook

## Create Notebook

Blank    From File    From URL

Name \*  ✓  
This name is valid 50

Description  
 500

Notebook URL

Tool \*  ▾

- h) This will import the notebook from GitHub into your DSX Local environment.

IBM Data Science Experience Local

Projects > MyProject > DSXLocalSailfish.ipynb

Predicting Outdoor Equipment Purchases with Data Science Experience Local on the IBM Integrated Analytics System

This notebook provides steps and code to load data from the IBM Integrated Analytics System (IIAS), explore this data, create a predictive model, deploy and start scoring with new data. In addition, it introduces basic data cleansing and exploration, pipeline creation, model training, model persistence, model deployment, predictions and scoring.

Some familiarity with Python is helpful. This notebook uses Python 2.0 and Apache® Spark 2.0.

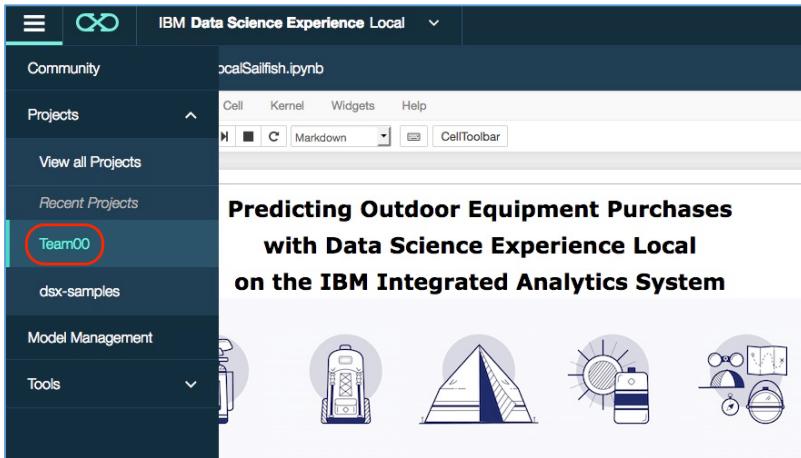
The data is publicly available and contains anonymous outdoor equipment purchases. It can be downloaded here [GoSales Transactions](#) but has already been loaded on IIAS. Based on this data, interest in product lines (golf accessories, camping equipment, etc.) will be predicted based on a consumer's gender, age, marital status and job.

- i) Please walk through the notebook by reading and executing the code cells to understand what is being accomplished.

## Part 2 Using the Machine Learning Builder

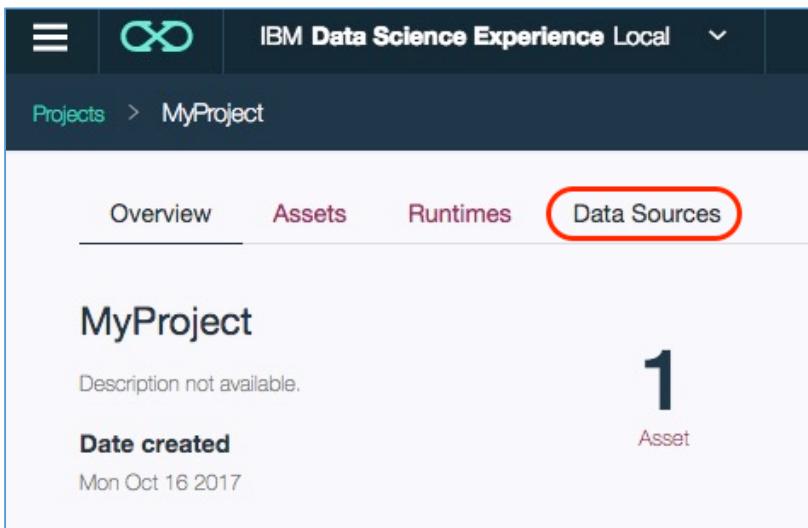
The following will walk through the steps to use the machine learning builder wizard.

- j) Click on the hamburger icon on the top left and select 'Projects' then your team project, 'TeamXX'.



A screenshot of the IBM Data Science Experience Local interface. The top navigation bar shows 'IBM Data Science Experience Local'. The left sidebar has sections for 'Community', 'Projects' (which is expanded), 'View all Projects', 'Recent Projects' (with 'TeamXX' highlighted and circled in red), 'dsx-samples', 'Model Management', and 'Tools'. The main content area displays a project titled 'Predicting Outdoor Equipment Purchases with Data Science Experience Local on the IBM Integrated Analytics System'. Below the title are four circular icons representing different outdoor equipment: a backpack, a tent, a camera, and a bicycle.

- k) First create a data source that will be recognized by the machine learning builder. Click on 'Data sources'.



A screenshot of the 'MyProject' overview page. The top navigation bar shows 'IBM Data Science Experience Local' and the path 'Projects > MyProject'. Below the path, there are tabs for 'Overview', 'Assets', 'Runtimes', and 'Data Sources' (which is circled in red). The main content area shows the project name 'MyProject', a description 'Description not available.', and a large number '1' indicating one asset. Below this, it shows 'Date created' as 'Mon Oct 16 2017'.

- l) Click on 'add data source'.



A screenshot of the 'Data Sources' creation page. The top navigation bar shows 'IBM Data Science Experience Local' and the path 'Projects > MyProject'. Below the path, there are tabs for 'Overview', 'Assets', 'Runtimes', and 'Data Sources'. The main content area has a header 'Data Sources (0)' with a 'add data source' button (circled in red). A table with columns 'NAME' and 'TYPE' is shown, with a note 'you have no data sources' at the bottom.

- m) On the first part of this screen enter:  
 'GOSales' for the 'Data Source Name'  
 'dashDB' for the 'Data Source Type'  
 'jdbc:db2://dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB' for the 'JDBC URL'  
 'dash7268' for the 'Username'  
 'mV@\$e7q4RPzL' for the 'Password' and  
 Keep 'Shared' UNCHECKED.

IBM Data Science Experience Local

Projects > Team00 > Create Data Source

## Create Data Source

Data Source Name \*

GOSales

Description

Type your description here

Data Source Type \*

dashDB

JDBC URL \*

'jdbc:db2://dashdb-entry-yp-dal09-09.services.dal.bluemix.net:50000/BLUDB'

Username \*

dash7268

- n) Now scroll down, click on 'Add Remote Data Set' and enter the following:  
 'GOSales' for the 'Remote Data Set Name'  
 'dash7268.GOSALES' for the 'Table'  
 Leave 'Schema' blank.  
 Now click 'Create'.

Add Remote Data Set

Remote Data Set Name \*

GOSales

Table \*

dash7268.GOSALES

Schema

Type table schema here

- o) The 'GOSales' data source should now appear.  
 Click on 'Add model' under '+' pull down on the top right.

IBM Data Science Experience Local

Projects > Team00

Overview Assets Runtimes Data Sources

Data Sources (1)

NAME	TYPE
GOSales	dashDB

+ add

- Create notebook
- Add data set
- Add model**

- p) On the 'Create Model' screen, enter your team name 'TeamXX' for the 'Name' and select 'Manual'. Please note that it states 'Need something more flexible? Create a notebook'. This was done in Part 1 of this lab. Now click on 'Create'.

- q) Select the 'GOSales' data source that we just created and click on 'Next'.

- r) Wait until the data is loaded. A small sample of the data will be displayed. Please keep 'Auto Data Preparation' and click on 'Next'.

- s) Now select 'PRODUCT\_LINE' from the 'Column value to predict (Label Col)'. Notice that a 'Multiclass Classification' is indicated for the 'Suggested technique'. Now click on 'Add Estimators'.

The screenshot shows the 'Select a technique' step in the IBM Data Science Experience Local interface. The 'PRODUCT\_LINE' column is selected as the label column. The 'Multiclass Classification' technique is highlighted as suggested. A red box highlights the 'Add Estimators' button.

- t) Select all three (Decision Tree Classifier, Random Forest Classifier, Naïve Bayes) for comparison. Now click 'Add'.

The screenshot shows the 'Select estimator(s)' dialog box. Three estimators are selected: 'Decision Tree Classifier', 'Random Forest Classifier', and 'Naïve Bayes'. A red box highlights the 'Add' button.

- u) Now click on 'Next'.

The screenshot shows the 'Select a technique' step. The three selected estimators are listed under 'Configured estimators': 'Decision Tree Classifier', 'Random Forest Classifier', and 'Naïve Bayes'. Each estimator has a red box highlighting it.

- v) The models will now be trained. This may take some time so please be patient.

## ) Training models

- w) The results of the 3 estimators will be listed. In this case, the ‘Decision Tree Classifier’ provided the best results. Please select the estimator with the best performance and click on ‘Save’. Then click on ‘Save’ again in the pop-up window.

Note: The accuracy of these models is ‘Poor’ or ‘Fair’. Greater accuracy may be achieved by using a notebook as in ‘Part 1’ and provide more data, regularization, normalization or other techniques. For purposes of this lab to show the process, we will continue with the best of the 3.

IBM Data Science Experience Local		Model Performance								
Select Data		Select model								
Prepare		ESTIMATOR TYPE	PERFORMANCE	WEIGHTED TRUE POSITIVE RATE	WEIGHTED FALSE POSITIVE RATE	WEIGHTED PRECISION	WEIGHTED F MEASURE	WEIGHTED RECALL	LAST VALIDATION	ACTIONS
Train		<input checked="" type="radio"/> Decision Tree Classifier	Poor	0.57554	0.20833	0.54721	0.55026	0.57554	22 Oct 2017, 10:17 PM	...
Evaluate		<input type="radio"/> Random Forest Classifier	Poor	0.56715	0.24405	0.58116	0.50894	0.56715	22 Oct 2017, 10:17 PM	...
		<input type="radio"/> Naive Bayes	Fail	0.45694	0.31059	0.34607	0.37458	0.45694	22 Oct 2017, 10:17 PM	...

- x) Under 'Deployments' click on 'Add Deployment'.

The screenshot shows the IBM Data Science Experience Local interface. At the top, there's a navigation bar with a menu icon, the title "IBM Data Science Experience Local", and a user profile icon. Below the title, the path "Projects > Team00 > Team00 Model" is displayed. The main content area is titled "Team00 Model" and includes a "Machine learning service" section with details: Label Column (PRODUCT\_LINE), Algorithm (org.apache.spark.ml.classification.NaiveBayesModel), and Training date (22 Oct 2017, 10:19 PM). A "Deployments" section follows, showing a table with columns "NAME", "DEPLOYMENT TYPE", and "ACTIONS". A red box highlights the "Add Deployment" button in the "ACTIONS" column.

IBM Data Science Experience Local

Projects > Team00 > Team00 Model

## Team00 Model

Machine learning service

IBM Machine Learning

Label Column

PRODUCT\_LINE

Algorithm

org.apache.spark.ml.classification.NaiveBayesModel

Training date

22 Oct 2017, 10:19 PM

## Deployments

NAME DEPLOYMENT TYPE ACTIONS

Your model is not deployed.

(+) Add Deployment

- y) In the ‘Deploy model’ screen, select ‘Online’ for ‘Deployment Type’ and your team name ‘TeamXX’ for the ‘Name’. Now click on ‘Deploy’.

Deploy model

Deployment Type  
Online

Name  
Team00

Close Deploy

- z) Under ‘Deployments’ for your team name, click on ‘View’.

Your deployment was successfully created.

Machine learning service	IBM Machine Learning
Label Column	PRODUCT_LINE
Algorithm	org.apache.spark.ml.classification.NaiveBayesModel
Training date	22 Oct 2017, 10:19 PM

**Deployments**

NAME	DEPLOYMENT TYPE	ACTIONS
Team00	Online	<span style="color: green;">View</span> <span style="color: blue;">...</span> <span style="color: red;">Delete</span>

aa) Notice that the scoring endpoints are displayed. Click on ‘TestAPI’.

DEPLOYMENT NAME: Team00

INTERNAL SCORING ENDPOINT: https://internal-nginx-svc.ibm-private-cloud.svc.cluster.local:12443/v2/scoring/online/6d87b70a-46d5-4d6d-a7a3-311ad31de1a1

EXTERNAL SCORING ENDPOINT: https://169.53.39.151/v2/scoring/online/6d87b70a-46d5-4d6d-a7a3-311ad31de1a1

PUBLISHER: team00

TYPE: Online

DATE DEPLOYED: 22 Oct 2017, 10:21 PM

NEXT EVALUATION DATE:

ASSOCIATED MODEL NAME: Team00 Model

REQUEST HEADER:

**Model Schema**

Input Schema				Output Schema			
NAME	TYPE	NULLABLE	METADATA	NAME	TYPE	NULLABLE	METADATA
age	integer	true	empty	age	integer	true	empty
gender	string	true	empty	gender	string	true	empty

bb) Use the provided input or specify something different, then click on ‘Predict’. The scoring results will be displayed.

**Input data**

- GENDER: M
- AGE: 43
- MARITAL\_STATUS: Married
- PROFESSION: Other

**Predicted value for PRODUCT\_LINE**

Probabilities

Camping Equipment

Outdoor Protection

Golf Equipment

Mountaineering Equipment

Personal Accessories

44.94%

23.05%

12.57%

16.42%

3.01%

A pie chart showing the predicted probabilities for different product lines. The largest segment is Camping Equipment at 44.94%, followed by Outdoor Protection at 23.05%, Personal Accessories at 16.42%, Mountaineering Equipment at 12.57%, and Golf Equipment at 3.01%.

cc) As an aside, the scoring endpoints could be used in a web application to provide real-time scoring.

The screenshot shows a web application interface for generating predictions. At the top, there are four customer profiles displayed in cards:

- Alice** (Professional): Gender F, Age 19 years, Marital Status Single
- Gregory** (Trades): Gender M, Age 57 years, Marital Status Married
- Joanna** (Executive): Gender F, Age 40 years, Marital Status Married
- Alexander** (Sales): Gender M, Age 36 years, Marital Status Single

Below the profiles is a red button labeled "Generate Predictions". In the center, there is a section titled "67% Personal Accessories" with an icon of a compass and map. The text states: "Based on your selection of **ProductLinePrediction** and your customer, it is predicted that **Alice** is 67% certain to buy **Personal Accessories**". To the right, under "Additional Recommendations", are listed: 17% Mountaineering Equipment, 13% Camping Equipment, 2% Outdoor Protection, and 1% Golf Equipment. A small "X" icon is next to the camping equipment entry.

Congratulations, you have completed this exercise.



Great work and congratulations, you have completed this exercise.



Screen captures in this document may vary slightly from yours.

## NOTES



---

© Copyright IBM Corporation 2016. Author: Daniel Kikuchi

The information contained in these materials is provided for informational purposes only, and is provided AS IS without warranty of any kind, express or implied. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, these materials. Nothing contained in these materials is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software. References in these materials to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. This information is based on current IBM product plans and strategy, which are subject to change by IBM without notice. Product release dates and/or capabilities referenced in these materials may change at any time at IBM's sole discretion based on market opportunities or other factors, and are not intended to be a commitment to future product or feature availability in any way.

IBM, the IBM logo and ibm.com are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at [www.ibm.com/legal/copytrade.shtml](http://www.ibm.com/legal/copytrade.shtml).



---