数据科学好帮手

IBM数据科学及机器学习平台揭秘和案例分享

曾勇华
zengyh@cn.ibm.com
数据科学家及解决方案架构师

IBM 机器学习及数据科学部门

Agenda 大纲

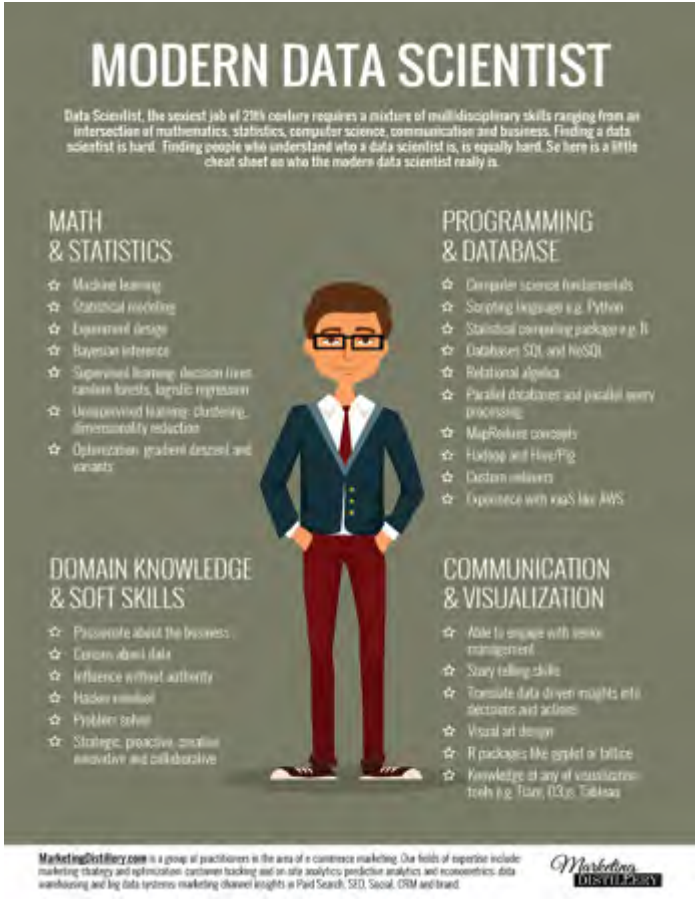
- 数据科学和机器学习概要
  - Data Science 101
  - Machine Learning 101
  - Data Science and ML Challenges
- IBM 数据科学平台介绍
  - IBM Data Science Experience
  - IBM Machine Learning
- 数据科学和机器学习案例演示

What is Data Science?

- **Data science**, also known as **data-driven science**, is an interdisciplinary field about scientific methods, processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured,[1][2] similar to Knowledge Discovery in Databases (KDD).
- Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data.[3] It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization.

From Wikipedia

Data Scientist: The Sexiest Job of the 21st Century



What abilities make a data scientist successful?

Think of him or her as a hybrid of

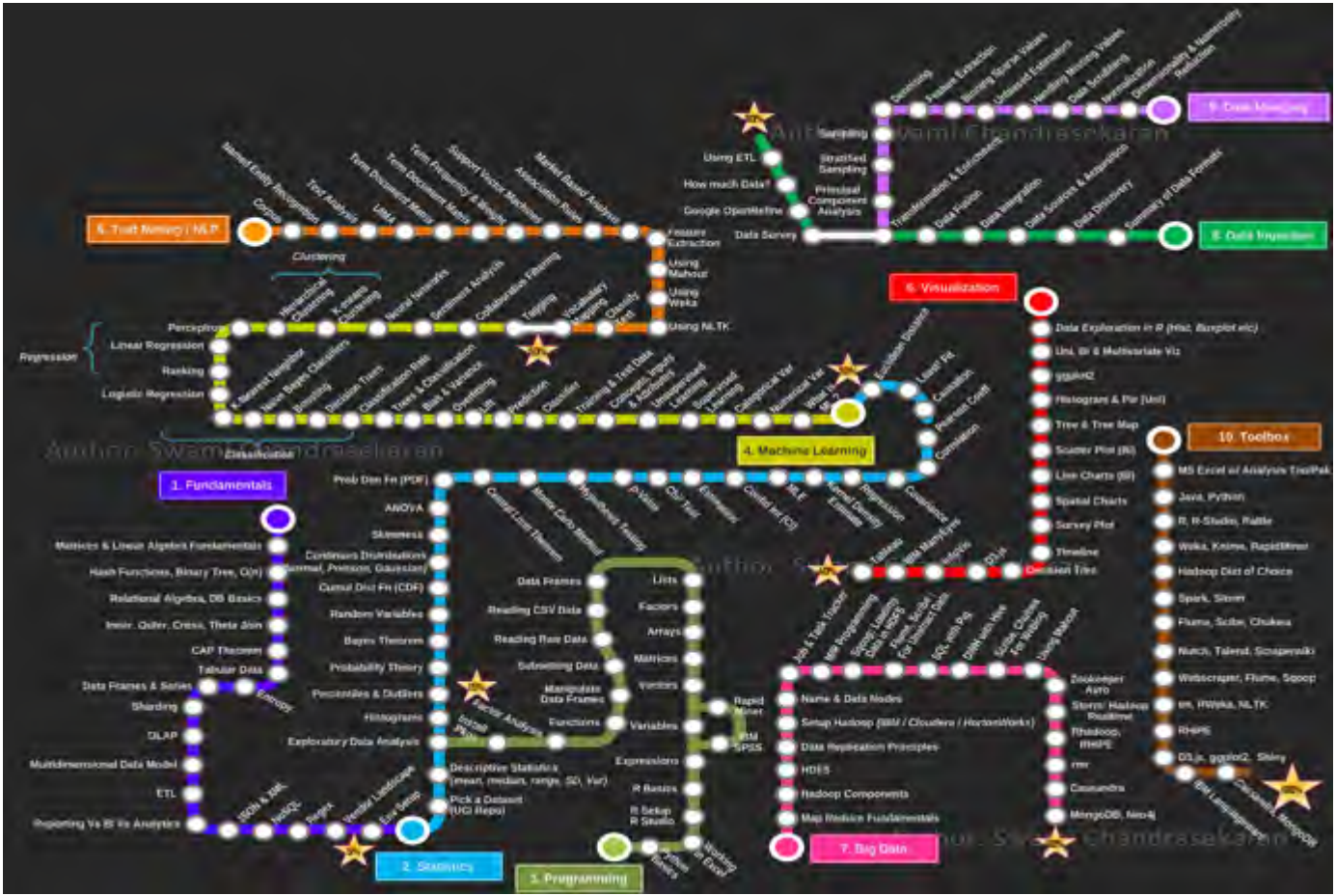
- data hacker
- analyst
- communicator
- trusted adviser

The combination is extremely powerful—and rare.

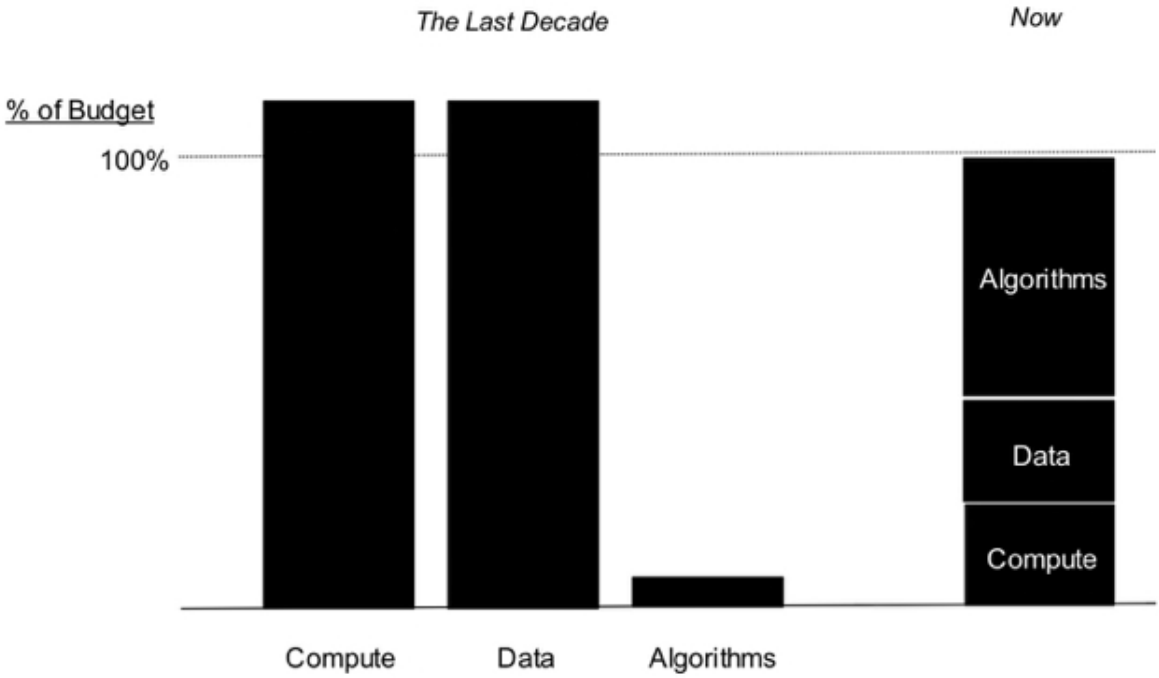
-----Harvard Business Review Oct 2012 Issue

数据科学家的硬技能

http://nirvacana.com/thoughts/becoming-a-data-scientist/



机器学习 第三次浪潮



What is Machine Learning?

Computers that …

Learn without being explicitly programmed

Grow and change when exposed to new data

Deliver personalized and optimized customer interactions

Identify Patterns
not readily foreseen by humans

Build Models
of behavior from those patterns

Achieving Business Value through Watson Machine Learning Capabilities

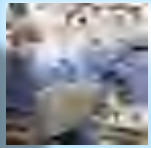
Machine Learning helps…

- *Constantly learns and adapts*
- *Avoids making the same mistakes*
- *Faster, deeper, improved insights*

Resulting in…

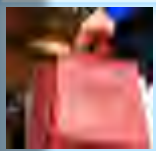
- ✓ **Smarter business outcomes**
- ✓ **Lower business risks and costs**
- ✓ **New business opportunities**

Churn analysis helps identify the cause of the churn and implement effective strategies for retention.

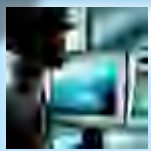


Detect and understand life-threatening medical conditions and design ever more effective treatment programs

Learn, predict weather patterns and energy production from renewable sources and integrate into grid more effectively



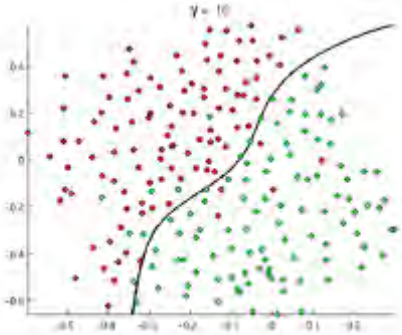
Product recommendation, next purchase prediction, targeted offers – individual tailored shopping experience.



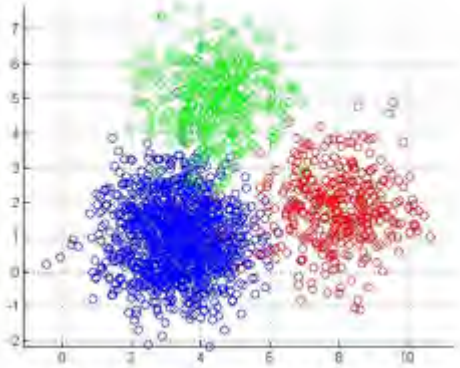
Identify suspicious behavior, predict and prevent threats / fraud – continually reduce business risks and costs

Machine Learning 101 : Types of machine learning

- **Classification**
  - Data points are labeled and are being used to predict a category
  - Two-class vs multi-class
  - Example:
    - Fraud detection (fraud vs non-fraud)
    - Spam email detection (spam vs non-spam)



- **Regression**
  - When a value is being predicted
  - Example:
    - Stock prices prediction



- **Clustering**
  - Data points are not labeled.
  - Goal is to group data into clusters to better organize the data

Machine Learning 101 : feature engineering

- A feature is a piece of information that might be useful for prediction
  - Example, predict the churn probability of a customer

- Labeled data is the desired output data
  - Example, CHURN_LABEL false representing a churn sample

CUST_ID	AVG_DAILY_TX	EDUCATION	EDUCATION_GROUP	INVESTMENT	AVG_TX_AMT	CHURN_LABEL	AGE
1009530866	0.917807996273040	[illegible]	Bachelors degree	114368	2090.32006835937	false	84
1009544000	0.950685024261474	2	Bachelors degree	90298	2095.0400390625	false	44
1009534260	0.920548021793365	2	Bachelors degree	94881	1723.4599609375	true	23
1009574010	0.994521021842955	2	Bachelors degree	112099	1297.419921875	true	24
[illegible]	0.917807996273040	5	Doctorate	84638	1333.17996[illegible]640625	false	67

NOT a feature

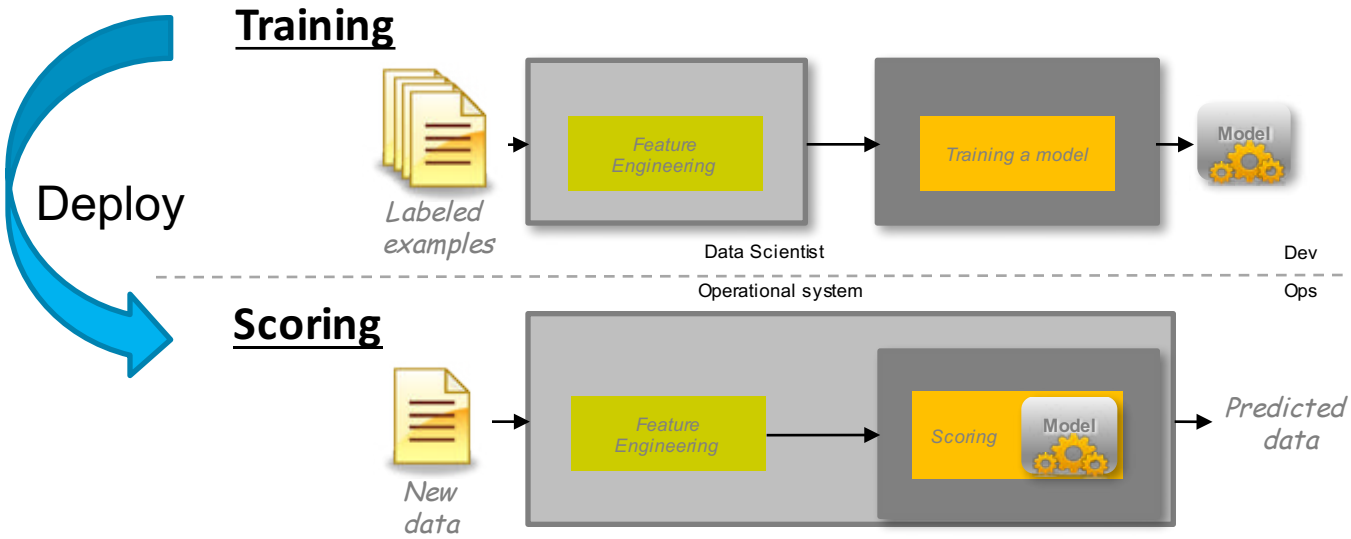
Feature

Feature

Feature

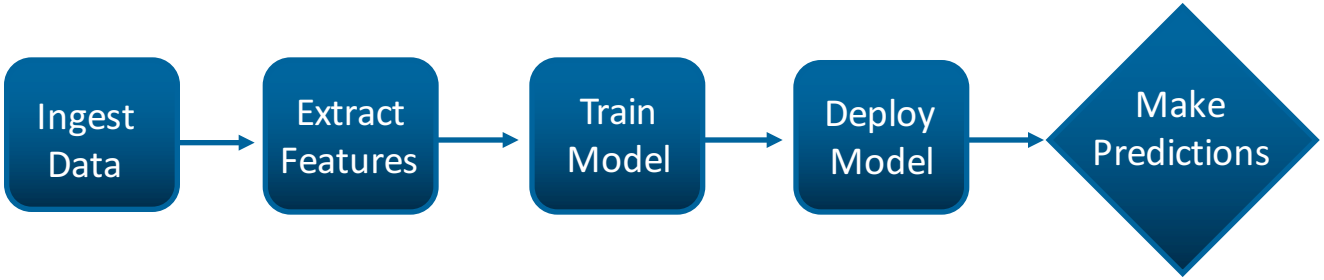
What is Machine Learning (机器学习概要)

a TrainOps (DevOps) story



The (incomplete) machine learning process

Takes significant development, deployment and management efforts



Human Intervention

Choose Best Model

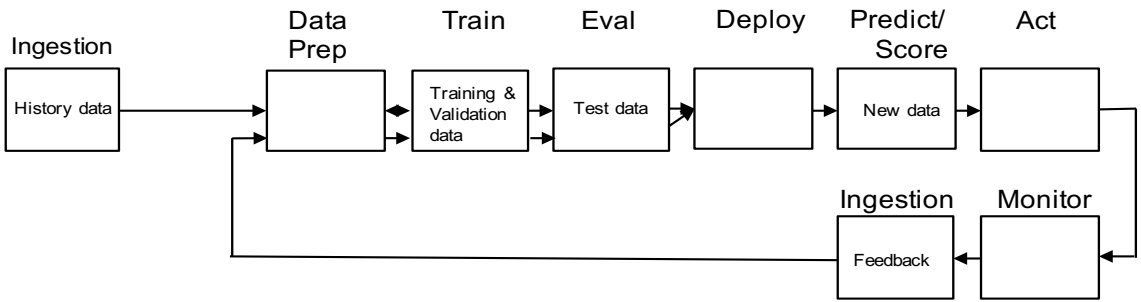
Identify Model Degradation

Prediction And Scoring

Manage Deployments

数据科学及机器学习新挑战

- 降低数据科学入门门槛 (Citizen Data Scientist)
- 管控机器学习全生命周期
- 提高持续交付能力
- 数据科学的可重复性



Agenda 大纲

- 数据科学和机器学习概要
  - Data Science 101
  - Machine Learning 101
  - Data Science and ML Challenges
- IBM 数据科学平台介绍
  - IBM Data Science Experience
  - IBM Machine Learning
- 数据科学和机器学习案例演示

IBM 数据科学工具箱

- IBM SPSS
- IBM Data Science Experience
- IBM Machine Learning

Figure 1. Magic Quadrant for Data Science Platforms

Source: Gartner (February 2017)

Data Science Experience (DSx) 主要特性

IBM Data Science Experience

社区

- 教材与数据集
- 连接数据科学家
- 提问
- 文章与论文
- 复制与分享项目

开源

- Scala/Python/R/SQL
- Jupyter and Zeppelin* Notebooks
- RStudio IDE and Shiny apps
- Apache Spark
- Your favorite libraries

IBM 提供的能力

- 数据预处理/Pipeline UI *
- 自动数据准备与建模*
- 高级可视化*
- 模型管理与部署
- 模型API文档*
- Spark云服务/Packaged Spark

❖ DSx Cloud Service
http://datascience.ibm.com

❖ DSx Local Edition

IBM Machine Learning for z/OS – 企业级机器学习平台

Notebook 和可视化建立模型

Cognitive Assistant for Data Scientists (CADS)

模型部署

模型管理

持续监控和反馈

IBM Machine Learning for z/OS 组件

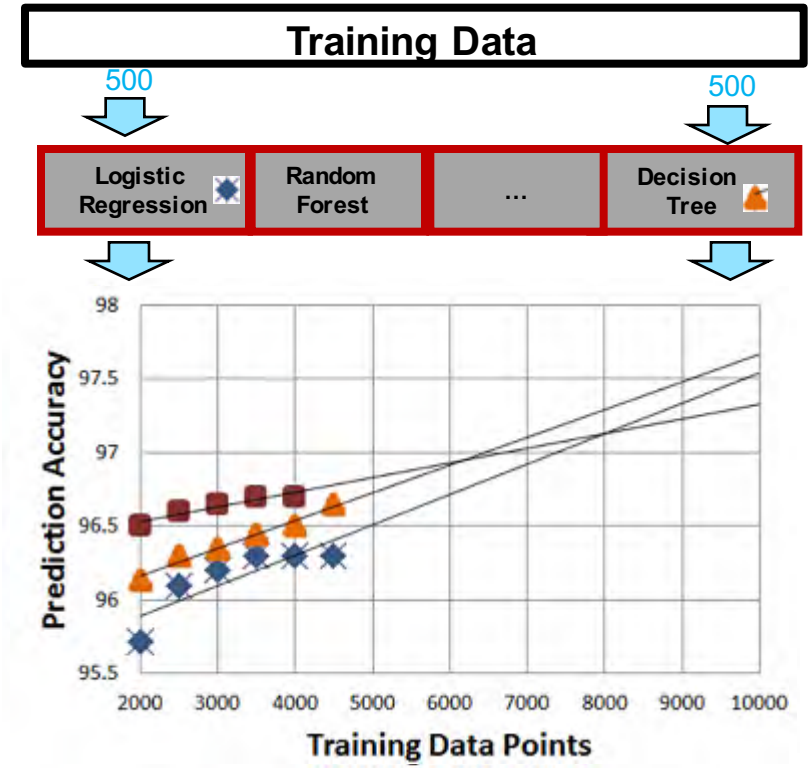
Feature Highlights – CADS 数据科学认知助手

- What is CADS?
  - Cognitive Assistant for Data Scientist which helps select the best fit algorithm for training

- Why Data Scientists need CADS?
  - Many algorithms for classification/regression tasks: SVM, Decision Trees/Forests, Naïve Bayes, Logistic Regression, etc.
  - Substantial cost in user and compute time to select the best algorithm
    - User spends time on trying various learners
    - Computational cost for training a single SVM can exceed 24h
    - Selection commonly based on data scientist bias and experience

Feature Highlights – CADS/HPO

- Minimize amount of data to be considered to make an informed selection of most suitable learner
- Given a data set try to select best approach by directly considering part of actual data



Feature Highlights

– Integrated Notebook Interface with flexible APIs

Load the data from DB2 for z/OS into dataframe and split for training, testing and evaluating

```
In [2]: import org.apache.spark.sql.SparkSession
        import org.apache.spark.sql.functions._

        //Load data from DB2 for z/OS using JDBC driver
        val churnDataRaw = spark.read.format("jdbc").
                                options(Map("driver" -> "com.ibm.db2.jcc.DB2Driver",
                                 "url" -> "jdbc:db2://9.125.72.72:430/LOCDB11",
```

Ingest data from DB2z table

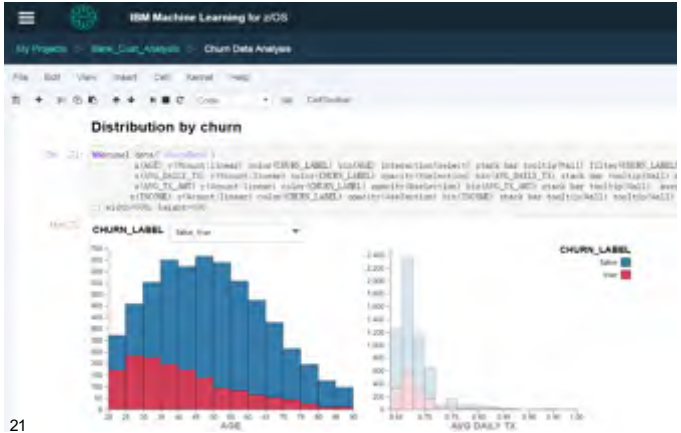
```
val toDouble

val churnData
churnData.sho

+---+--------
|AGE|ACTIVITY
+---+--------
| 84|       5
| 44|       1
| 23|       1
| 24|       4
| 67|       3
+---+--------
only showing
```

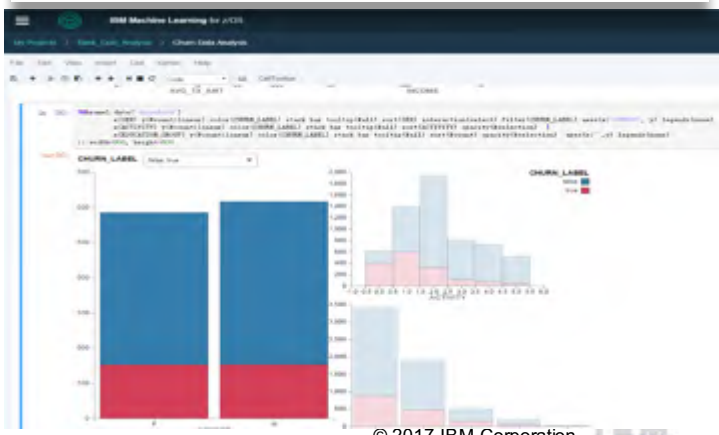
Use CADS(Cognitive Assistant of Data Science) to train & recommend the best model automatically from DT and LR

```
In [4]: //Feature definition

        val genderIndexer = new StringIndexer().setInputCol("SEX").setOutputCol("gender_code")
        val stateIndexer = new StringIndexer().setInputCol("STATE").setOutputCol("state_code")
        val labelIndexer = new StringIndexer().setInputCol("CHURN_LABEL").setOutputCol("label")

        val featuresAssembler = new VectorAssembler().setInputCols(Array("AGE",
                                                                         "ACTIVITY",
                                                                         "EDUCATION",
                                                                         "NEGTWEETS",
                                                                         "INCOME",
                                                                         "gender_code",
                                                                         "state_code")).setOutputCol("features")

In [5]: //Select model automatically in candidate algorithm - Logistic Regression, SVM or Decision Tree?
        val lr = new LogisticRegression().setRegParam(0.01).setLabelCol("label").setFeaturesCol("features")
        val decisionTree = new DecisionTreeClassifier().setMaxBins(50).setLabelCol("label").setFeaturesCol("features")
        val nb = new NaiveBayes().setLabelCol("label").setFeaturesCol("features")

In [6]: //Cognitive Assistant for Data Scientists - predict model performance based on sampled data
        val learners = List(Learner("DT", decisionTree), Learner("LR", lr))
        val cads = CADSEstimator().setEvaluator(new BinaryClassificationEvaluator().
                                setMetricName("areaUnderROC")).
                                setLearners(learners).
                                setKeepBestNLearnersParam(3).
                                setTarget(Target("rawPrediction", "label")).
                                setNumSampleFoldsParam(2)
        val pipeline = new IBMSparkPipeline().setStages(Array(labelIndexer, genderIndexer, stateIndexer, featuresAssembler, cads))
        val model = pipeline.fit(trainingDF)
```

Data transformation and training

Feature Highlights
– Data Visualization with Brunel (https://github.com/Brunel-Visualization/Brunel)





Feature Highlights

– Visual Model Builder, the guided Machine Learning Interface

Ingest data and transform

Prepare data set

Select Data
Prepare
Train
Select Model
Evaluate

Add a Transformer

TWITTERID	CUST_ID	AVG_DAILY_TX	EDUCATION	EDUCATION_GROUP		AVG_TX_AMT	CHURN_LA
0	1009530860	0.9178079962730408	2	Bachelors degree	114368	2090.320068359375	false
0	1009544000	0.9506850242614746	2	Bachelors degree	90298	2095.0400390625	false
0	1009534260	0.9205480217933655	2	Bachelors degree	94881	1723.4599609375	true
0	1009574010	0.9945210218429565	2	Bachelors degree	112099	1297.419921875	true
0	1009578620	0.9178079962730408	5	Doctorate	84638	1333.179931640625	false
0	1009575250	0.9452049732208252	5	Doctorate	80194	1175.570068359375	true

Configured transformers

StringIndexer
GenderCode

StringIndexer
ChurnLabel

StringIndexer
StateCode

VectorAssembler
AllFeatures

Training and evaluation

Select model

Select Data
Prepare
Train
Select Model
Evaluate

Binary Classification

	MODEL NAME	ESTIMATOR TYPE	PERFORMANCE	AREA UNDER ROC CURVE	AREA UNDER PR CURVE	LAST VALIDATION	ACTIONS
●	BankingChurnLRModel	logistic_regression	Excellent	0.96481	0.96867	19 Feb 2017, 7:48 PM	...
○	BankingChurnLRWith4FeaturesModel	logistic_regression	Good	0.84418	0.79224	8 Mar 2017, 9:32 AM	...
○	BankingChurnLRWith3FeaturesModel	logistic_regression	Poor	0.68375	0.64866	8 Mar 2017, 9:33 AM	...

Feature Highlights – Model Management

Dashboard | **Models** | Deployments

Churn

Manage model, create deployment

MODEL NAME	MODEL ID	OWNED BY	DATE CREATED	DATE UPDATED			ACTIONS
ChurnCADS	12	steve	2017-03-06 13:22	2017-03-06 13:22	ACTIVE		
ChurnModel	11	wmlz16	2017-03-06 06:53	2017-03-06 06:53	ACTIVE	1	
AdarshChurnCADSModel	7	adarsh	2017-02-28 19:16	2017-02-28 19:16	ACTIVE	1	
ChurnCADSModelLR	4	steve	2017-02-19 11:55	2017-02-19 11:55	ACTIVE	1	
BankingChurnLRModel	3	steve	2017-02-19 11:50	2017-02-19 11:50	ACTIVE	1	
ChurnCADSModel	2	steve	2017-02-19 11:49	2017-02-19 11:49	ACTIVE	1	
ChurnCADSWithNotebook	1	steve	2017-02-19 11:46	2017-02-19 11:46	ACTIVE	1	

- View details
- Create deployment
- Update
- Delete

Dashboard | Models | **Deployments**

Notebook

Manage deployment

DEPLOYMENT NAME	DEPLOYMENT ID	MODEL	EVALUATION SCHEDULED	ATED	ACTIONS
TentCADSNotebookModelD02	11	TentCADSNotebookModel	NO	2 -27 14:36	
TentCADSNotebookModel	4	TentCADSNotebookModel	NO	2017-02-19 20:03	
ChurnCADSWithNotebook	2	ChurnCADSWithNotebook	YES	2017-02-19 20:02	

- View details
- Test
- Schedule evaluation
- Delete

Feature Highlights

– Easily consumable RESTful API for online Scoring within Application Code

API Details

Scoring Endpoint	http://9.125.72.72:10082/wml/scoring/spark/deployments/143/predict
Number Of Invocation	0
Average Elapsed Time	0 ms

Request Header

Authorization: Bearer <token>	Required. Pass the token from IBM Wat
Content-Type: application/json	Required if the request body is sent in J

RESTful API for online scoring and prediction

```
function getMoksaProfileValue($gender, $age, $marital_status, $job_type) {
    if ($_COOKIE["usesystemz"] == "no") {
        echo "<!-- bypass z system -->\n";
        // bypass z Systems call
        return getMoksaProfileValue2($gender, $age, $marital_status, $job_type);
    }
    echo "<!-- use z system -->\n";

    $ch = curl_init("http://9.125.72.72:10082/wml/scoring/spark/deployments/143/predict");
    $data_string = "{\"Record\":[\"{$gender}\", $age, \"{$marital_status}\", \"{$job_type}\"]}";

    curl_setopt($ch, CURLOPT_CUSTOMREQUEST, "POST");
    curl_setopt($ch, CURLOPT_POSTFIELDS, $data_string);
    curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
    curl_setopt($ch, CURLOPT_HTTPHEADER, array(
            'Content-Type: application/json',
            'Authorization: eyJhbGciOiJSUzUxMiIsInR5cCI6IkpXVCJ9.eyJlc2VySWQiOiJ0dXNlcjAxIiwiaXNzIjoiaHR0cHM6Ly93
            'Content-Length: ' . strlen($data_string))
            );
    $result = curl_exec($ch);
    curl_close($ch);
    $array = json_decode($result, true);

    return $array["prediction"];
}
```

Feature Highlights

– Feedback and Continuous Monitoring

Watson Machine Learning for z/OS

Dashboard Models Deployments

Evaluation Highlights

MODEL	EVALUATION TIME	METRICS	THRESHOLD	RESULT
GCSalesTentLR4StevaD01	2017-01-23 02:01:23	areaUnderROC	0.7	▲ 0.64
GCSalesTentLR4StevaD01	2017-01-22 02:03:35	areaUnderROC	0.7	■ 0.34
TentFeedback	2017-01-20 07:33:02	areaUnderROC	0.95	▲ 0.64
TentFeedback	2017-01-20 06:57:03	areaUnderROC	0.95	▲ 0.64

Model Health

92%
Health Index

24 MODELS IN TOTAL
2 models have been highlighted

Spark Status

Alive Workers: 4
Cores in use: 50%
Memory in use: 75%
Applications: 3 Running, 120 Completed
URL: spark://WINITM.CN.IBM.COM:7077

Top models by number of invocations

Number of invocation

140 120 100 80 60 40 20 0

In-tent dt-customer in-reg

Top models by average elapsed time

Average elapsed time(ms)

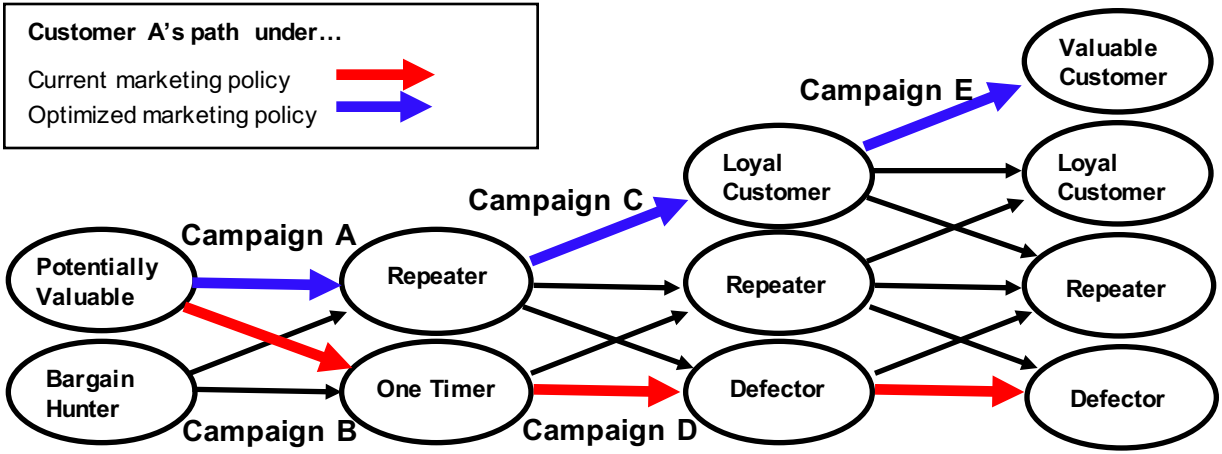
140 120 100 80 60 40 20 0

In-tent dt-customer in-reg

数据科学和机器学习的方法学

机器学习平台演示

客户流失预测



案例： “认知银行” – 了解你的客户

- **业务问题：**

我愿意提供优惠来防止客户流失，问题是我不知道谁会流失，过去都是流失之后我才知道的

- **解决方案：**

用机器学习找到现有已流失客户的特点，用来预测现有客户的流失可能性

- **知道之后应该怎么做：**

推荐系统之“认知银行”，以后介绍完整场景

- IBM Machine Learning Platform

谢谢