ORACLE

Introduction to Machine Learning

From DBA's to Data Scientists

Sandesh Rao

VP AIOps for the Autonomous Database

@sandeshr

https://www.linkedin.com/in/raosandesh/

https://www.slideshare.net/SandeshRao4

Traditionally DBAs are Responsible for:

Tasks Specific to Business and Innovation

- Architecture, planning, data modeling
- Data security and lifecycle management
- Application related tuning
- End-to-End service level management

Maintenance Tasks

- Configuration and tuning of systems, network, storage
- Database provisioning, patching
- Database backups, H/A, disaster recovery
- Database optimization

Autonomous Database Removes Generic Tasks

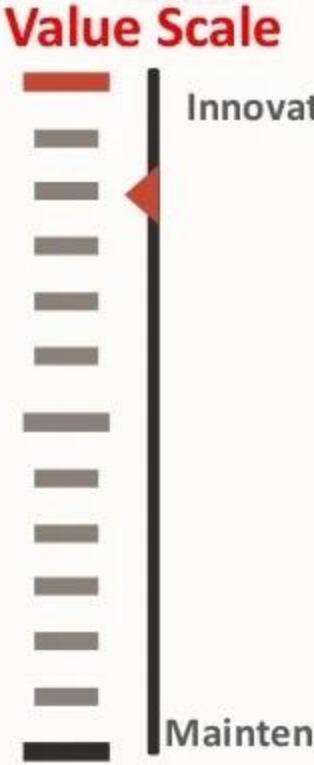
Freedom from Drudgery for DBA: More Time to Innovate and Improve the Business

Tasks Specific to Business and Innovation

- Architecture, planning, data modeling
- Data security and lifecycle management
- Application related tuning
- End-to-End service level management

~~Maintenance Tasks~~

- ~~Configuration and tuning of systems, network, storage~~
- ~~Database provisioning, patching~~
- ~~Database backups, H/A, disaster recovery~~
- ~~Database optimization~~



The Evolution of the DBA/Database Developer Role

Data Engineer
Architecture,
"data wrangler"

Data Security
Data classification,
Data life-cycle mgmt

Machine Learning
Solving data-driven problems
Discovering insights
Making predictions

Application Tuning
SQL tuning,
connection mgmt

Database Developer to Data Scientist Journey

You Are Probably Already Doing Most of This Work!

Data extraction

Data wrangling

Deriving new attributes ("feature engineering")

…

…

…

Import predictions & insights

Translate and deploy ML models

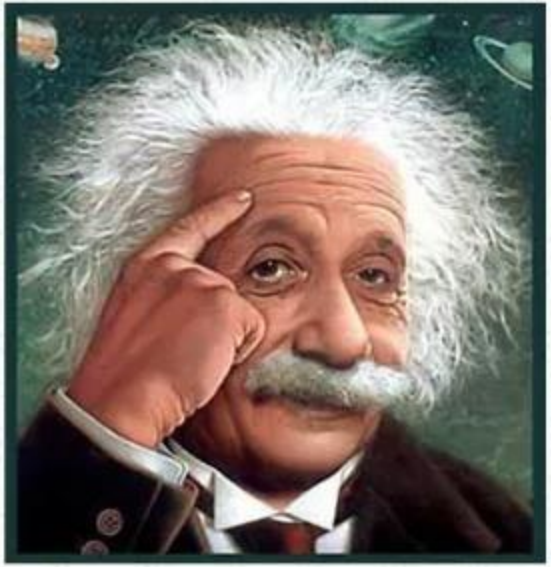
Automate

Typically 80% of the work

Most data scientists spend only 20 percent of their time on actual data analysis and 80 percent of their time finding, cleaning, and reorganizing huge amounts of data, which is an inefficient data strategy[1]

Eliminated or minimized with Oracle

Data Management platform becomes combine/hybrid DM + machine learning platform



"If I had an hour to solve a problem I'd spend 55 minutes thinking about the problem and 5 minutes thinking about solutions."

Albert Einstein

Why Machine Learning is important

Lots of Data needs to be crunched

- No time to manually sift through the data

Machine Learning has become accessible

- Software and algorithms are available
- Frameworks allow for massive training with no coding
- CI/CD available for MLOps
- It's not the algorithm you need to know about !!

Business use cases

- Find the use cases for maximum impact

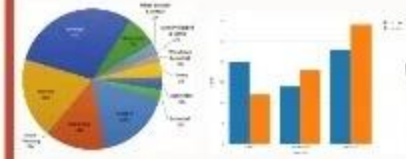
Analytics Value vs. Maturity

Automated ML Appls

"ML Enabled" Applications

What WILL happen?

Predictive / Machine Learning



Why it Happened?

Diagnostic Analysis & Reports

What Happened?

Reports & Dashboards

Value of Analytics

Data

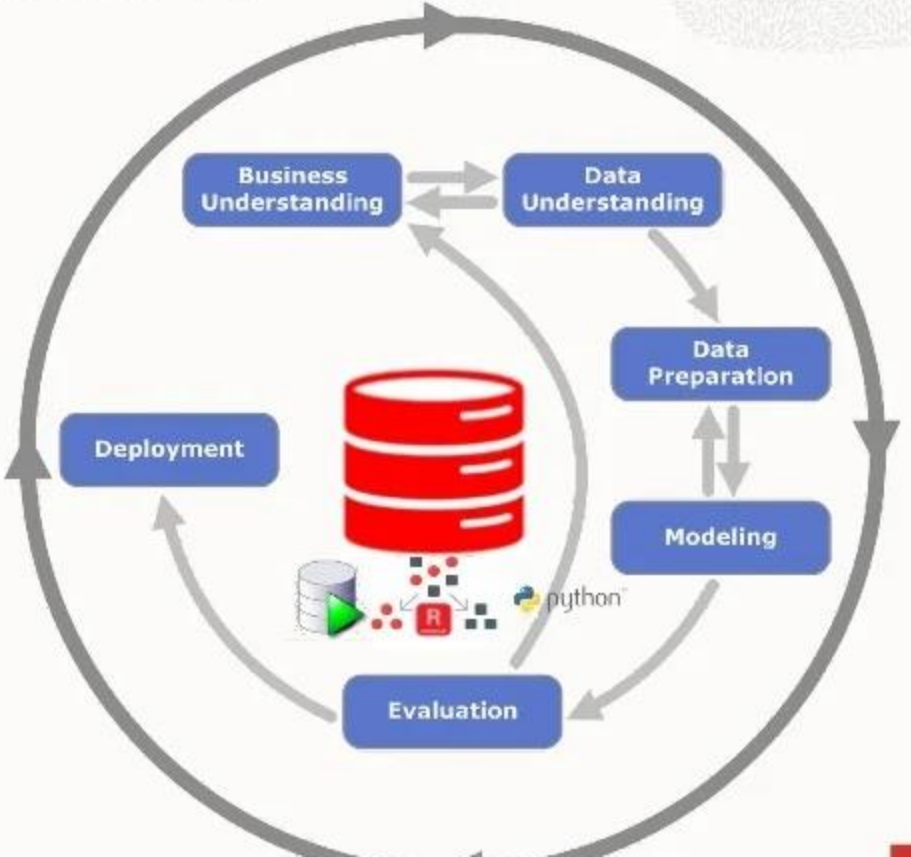
Information

Predictions & Insights

Appls with ML

Analytical Maturity

Database Developer to Data Scientist Journey



ML Project Workflow

1. Set the business objectives

2. Gather compare and clean data

3. Identify and extract features (important columns) from imported data

   This helps us identify the efficiency of the algorithm

4. Take the input data which is also called the training data and apply the algorithm to it

   For the algorithm to function efficiently, it is important to pick the right value for hyper parame (algorithm input parameters to the algorithm)

5. Once the training da the algorithm are combined we get a m

Types of Machine Learning

Supervised Learning

Predict future outcomes with the help of training data provided by human experts

Semi-Supervised Learning

Discover patterns within raw data and make predictions, which are then reviewed by human experts, who provide feedback which is used to improve the model accuracy

Unsupervised Learning

Find patterns without any external input other than the raw data

Reinforcement Learning

Take decisions based on past rewards for this type of action

TIME SERIES

Temporal Aspect

Hitting a threshold

Forecasting energy us

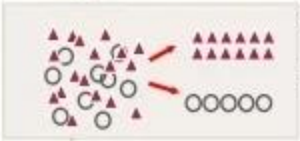
Seasonality of data

What is Machine Learning?

Algorithms ***automatically*** sift through large amounts of data to discover hidden patterns, new insights and make predictions

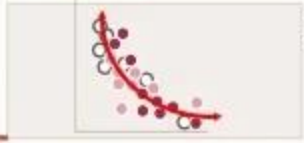
Supervised Learning

Identify most important factor (Attribute Importance)



Predict customer behavior (Classification)

Find profiles of targeted people or items (Classification



Predict or estimate a value (Regression)

Unsupervised Learning

Segment a population (Clustering)

Find fraudulent or “rare events” (Anomaly Detection)

Determine co-occurring items in a “basket” (Associations)

Machine Learning Algorithms

Clustering

- Hierarchical k-means, Orthogonal Partitioning Clustering, Expectation-Maximization

Feature Extraction/Attribute Importance / Component Analysis

Classification

- Decision Tree, Naive Bayes, Random Forest, Logistic Regression, Support Vector Machine

Regression

- Multiple Regression, Support Vector Machine, Linear Model, LASSO, Random Forest, Ridge Regression, Generalized Linear Model, Stepwise Linear Regression

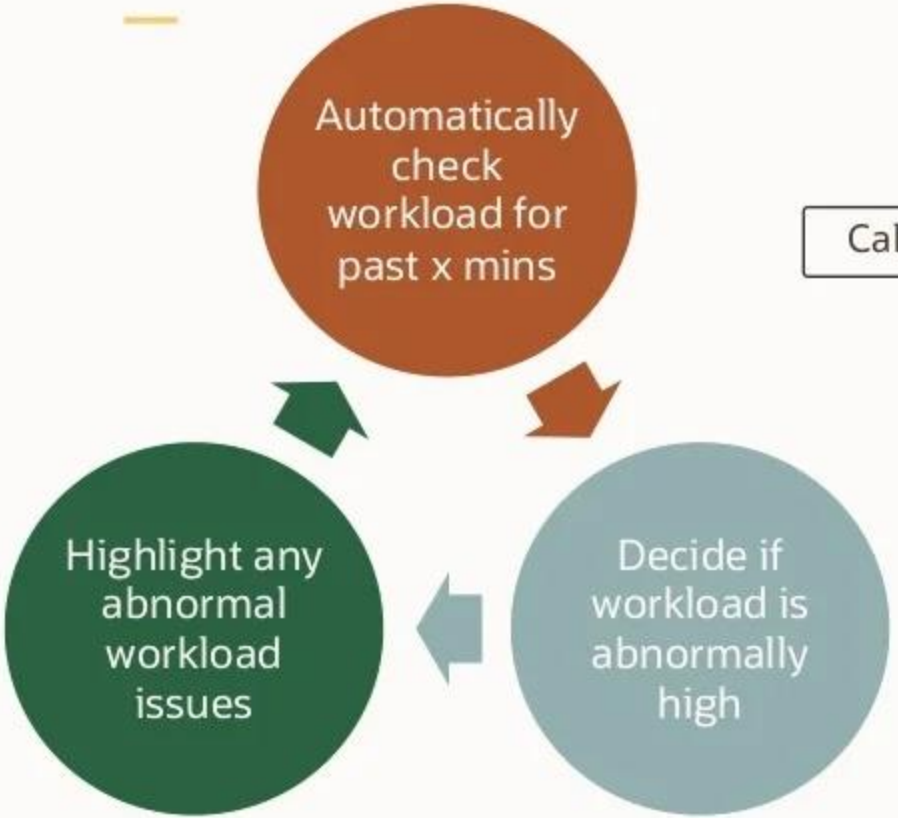
Association & Collaborative Filtering

Reinforcement Learning - brute force, Monte Carlo, temporal difference....

Neural network & deep Learning with Deep Neural Network

- Many different use cases

What is Workload



Calculated via machine learning

Optionally run on demand

Optionally snooze checking of a component

Prediction (Every 5 minutes)

5 X 1 min metrics captured for each dimension & ASH report captured for later analysis

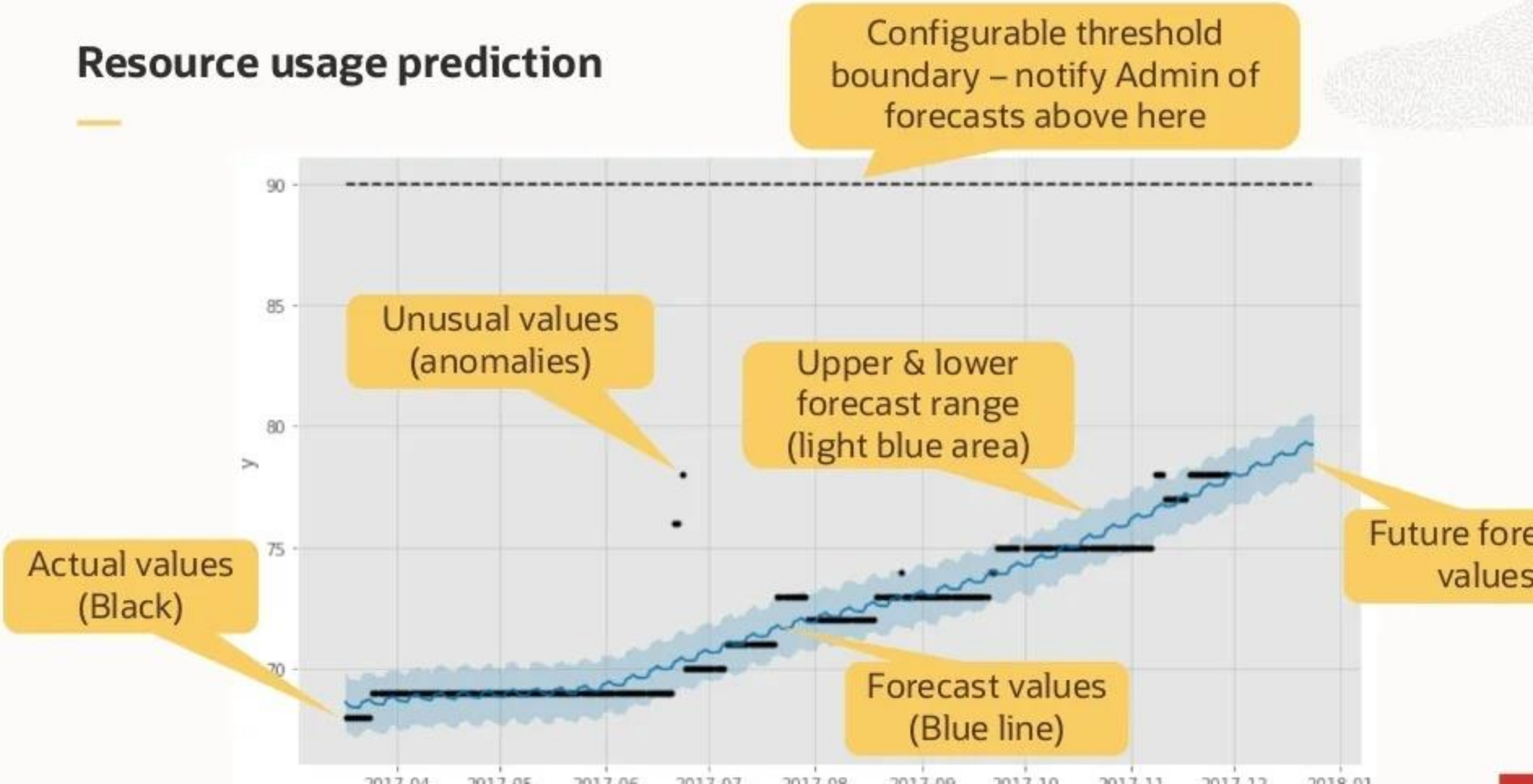
Each anomaly is compared against the SME rules to determine which dimension it applies to

Metrics evaluated by the primary model to determine if there are anomalies

If there is no primary model (i.e. <7 days of data or <=95% model confidence) then SME rules are used for anomaly detection

Any anomalies are rai
along with recentl
captured ASH repo

Resource usage prediction

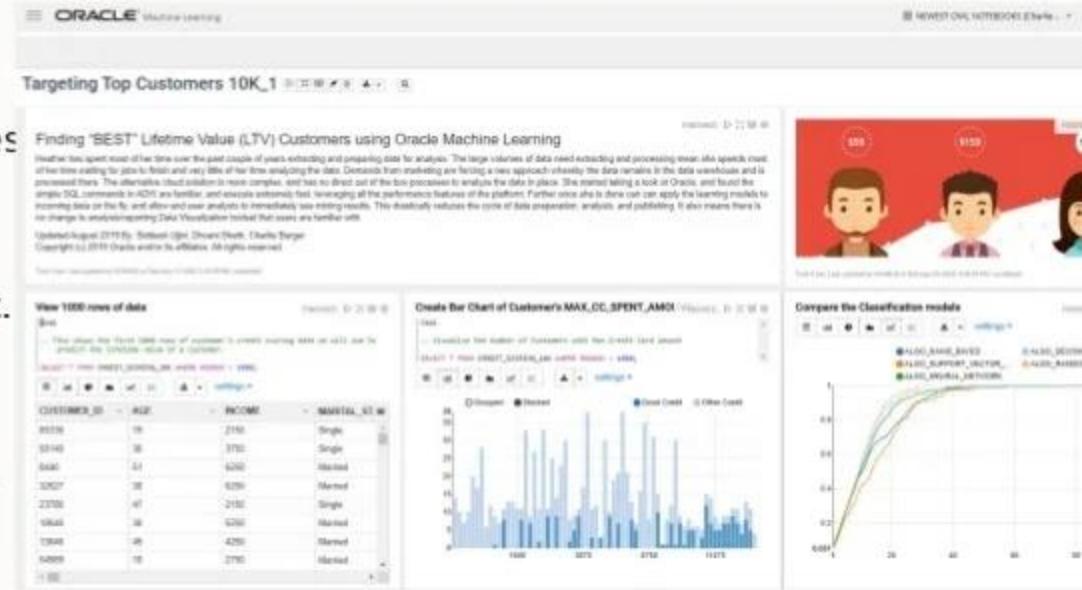


Oracle Machine Learning

Machine Learning Notebooks included in Autonomous Databases

Key Features:

- Collaborative UI for data scientist and analysts
- Packaged with Autonomous Databases
- Quick start Example notebooks
- Easy access to shared notebooks, templates, permissions, scheduler, etc.
- OML4SQL
- OML4Py *coming soon*
- Supports deployment of OML models



Statistical Functions

Simple SQL Syntax—Statistical Comparisons (t-tests)

Compare AVE Purchase Amounts Men vs. Women Grouped_By INCOME_LEVEL

```
SELECT SUBSTR(cust_income_level, 1, 22) income_level,
       AVG(DECODE(cust_gender, 'M', amount_sold, null)) sold_to_men,
       AVG(DECODE(cust_gender, 'F', amount_sold, null)) sold_to_women,
       STATS_T_TEST_INDEPU(cust_gender, amount_sold, 'STATISTIC', 'F') t_observed,
       STATS_T_TEST_INDEPU(cust_gender, amount_sold) two_sided_p_value
  FROM customers c, sales s
  WHERE c.cust_id = s.cust_id
  GROUP BY ROLLUP(cust_income_level)
  ORDER BY income_level, sold_to_men, sold_to_women, t_observed;
```

Query Result

All Rows Fetched: 14 in 1.523 seconds

	INCOME_LEVEL	SOLD_TO_MEN	SOLD_TO_WOMEN	T_OBSERVED	TWO_SIDED_P_VALUE
1	A: Below 30,000	105.2834897729...	99.42814466653473...	-2.05425922984...	0.0399647043795526...
2	B: 30,000 - 49,999	102.5965095067...	109.8296418272003...	2.9692233218892...	0.0029077419365879512
3	C: 50,000 - 69,999	105.6275880730...	110.1279310121247...	2.3496854009262...	0.0187922767711299937
4	D: 70,000 - 89,999	106.6302994897...	110.4726699326023...	2.2603928063380...	0.0233078312872170897
5	E: 90,000 - 109,999	103.3967414937...	101.6104162583700...	-1.2603509195 4...	0.207545662363282097
6	F: 110,000 - 129,999	106.7647596205...	105.9013119482142...	-0.6058001077 0...	0.5446485528703752
7	G: 130,000 - 149,999	108.8775321810...	107.3137690570293...	-0.8521978096 9...	0.394107754843487597
8	H: 150,000 - 169,999	110.9872579252...	107.1521911799573...	-1.9451485087 9...	0.0517626238993762487
9	I: 170,000 - 189,999	102.8082379709...	107.4395601412162...	2.1496692058 99...	0.0315878750783994557
10	J: 190,000 - 249,999	108.0405638372...	115.3433560297627...	2.5474906690 40...	0.0100549660212309457
11	K: 250,000 - 299,999	112.3779929260...	108.1960973300511...	-1.4115513680 6...	0.158091675654154387

STATS_T_TEST_INDEPU (SQL) Example;

P_Values < 05 show statistically significantly differences in the amounts purchased by men vs. women

Model Build and Real-time SQL Apply Prediction

Simple SQL Syntax—Attribute Importance - ML Model Build (PL/SQL)

```
BEGIN
   DBMS_DATA_MINING.CREATE_MODEL(
     model_name          => 'BUY_INSURANCE_AI',
     mining_function             => DBMS_DATA_MINING.ATTRIBUTE_IMPORTANCE,
     data_table_name             => 'CUST_INSUR_LTV',
     case_id_column_name         => 'cust_id',
     target_column_name          => 'BUY_INSURANCE',
     settings_table_name         => 'Att_Import_Mode_Settings');
END;
/
```

Model Results (SQL query)

```
SELECT attribute_name, rank , attribute_value
FROM BUY_INSURANCE_AI
ORDER BY rank, attribute_name;
```

ATTRIBUTE_NAME	RANK	ATTRIBUTE_VALUE
BANK_FUNDS	1	0.2161
MONEY_MONTLY_OVERDRAWN	2	0.1489
N_TRANS_ATM	3	0.1463
N_TRANS_TELLER	4	0.1156

Oracle Machine Leaning

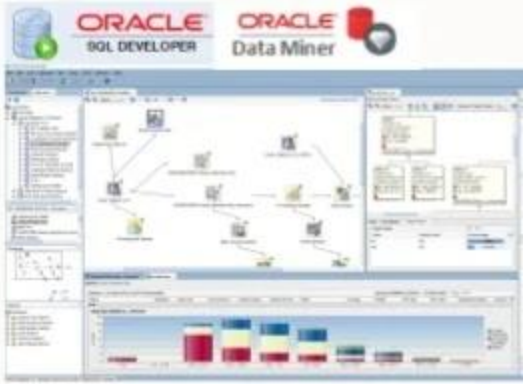
Multiple Languages UIs Supported for End Users & Apps Development

New!

R & Python Data Scientists

New!

Notebook Users & DS Teams



"Citizen" Data Scientists

Feature Engineering - examples

Create New Derived Attributes or "Engineered Features"

Source Attribute	New Attribute/"Engineered Feature"
Date of Birth	AGE
Address	DISTANCE_TO_DESTINATION
	COMMUTE_TIME
Call detail records (CDRs)	#_DROPPED_CALLS
	PERCENT_iNTERNATIONAL
Salary	PERCENT_VS_PEERS
Purchases	TOTALS_PER_CATEGORY *(e.g. Food, Clothing)*

Modeling and Machine Learning

First, Identify the Key Attributes That Most Influence the Target Attribute

Credit Score Predictions_1

default

Create Attribute Importance Machine Learning Model for Good Credi Customers

FINISHED

```
%script

/* Find the importance of attributes that independently impact the target attribute:
    CREDIT_SCORE_BIN */

DECLARE
v_sql varchar2(100);

BEGIN
BEGIN EXECUTE IMMEDIATE 'DROP TABLE ai_explain_output_credit_score_bin';
EXCEPTION WHEN OTHERS THEN NULL;
END;

BEGIN
  DBMS_PREDICTIVE_ANALYTICS.EXPLAIN(
    data_table_name     => 'CREDIT_SCORING_100K_V',
    explain_column_name => 'CREDIT_SCORE_BIN',
    result_table_name   => 'AI_EXPLAIN_OUTPUT_CREDIT_SCORE_BIN');
END;
```

Attribute Importance Model

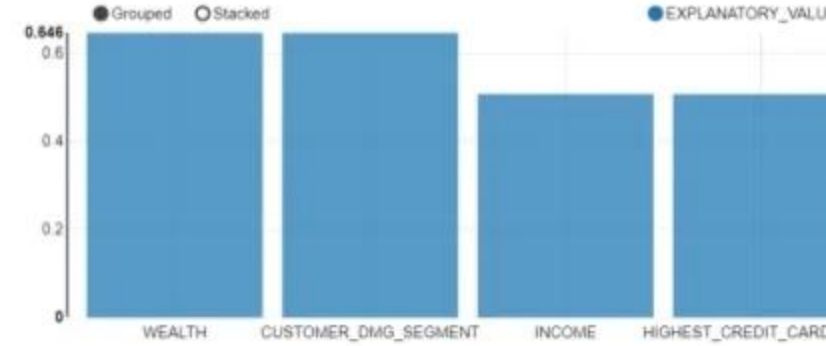
```
PL/SQL procedure successfully completed.
```

Took 1 min 49 sec. Last updated by CHARLIE at February 07 2019, 3:36:32 PM. (outdated)

Display the Top N Attributes for Good Credit Customers

FINISHED

settings



Took 4 sec. Last updated by CHARLIE at February 11 2019, 3:53:41 PM. (outdated)

Modeling and Machine Learning

Next, Build Predictive Models to Predict Customers who are Likely to Have Good_Credit

Preparatory Steps, Automation of Model Build and Test and Clean up using PL/SQL script

FINISHED

```
%script

/* Build a classification model and then generate a lift test result and an apply result.  Click on
    the arrow in the the upper right. */

DECLARE
v_sql varchar2(100);

BEGIN
/*
/* Split the Data into N1_TRAIN_DATA and N1_TEST_DATA */
EXECUTE IMMEDIATE 'CREATE OR REPLACE VIEW N1_TRAIN_DATA AS SELECT * FROM CREDIT_SCORING_100K_v SAMPLE
    (60) SEED (1)';
DBMS_OUTPUT.PUT_LINE ('Created N1_TRAIN_DATA');
EXECUTE IMMEDIATE 'CREATE OR REPLACE VIEW N1_TEST_DATA AS SELECT * FROM CREDIT_SCORING_100K_v MINUS
    SELECT * FROM N1_TRAIN_DATA';
DBMS_OUTPUT.PUT_LINE ('Created N1_TEST_DATA');

/* Create a Build Setting (DT) for Model Build */

EXECUTE IMMEDIATE 'CREATE TABLE n1_build_settings (setting_name VARCHAR2(30),setting_value VARCHAR2
    (4000))';
EXECUTE IMMEDIATE 'INSERT INTO n1_build_settings (setting_name, setting_value) VALUES (''ALGO_NAME'',
    ''ALGO_DECISION_TREE'')';
EXECUTE IMMEDIATE 'INSERT INTO n1_build_settings (setting_name, setting_value) VALUES (''PREP_AUTO'',
    ''ON'')';


DBMS_OUTPUT.PUT_LINE ('Created model build settings table: n1_build_settings ');

/*
-- Populate and Adjust Model Setting (DT) for Model Build
EXECUTE IMMEDIATE 'INSERT INTO n1_build_settings VALUES (''TREE_TERM_MAX_DEPTH'', 7)';
EXECUTE IMMEDIATE 'INSERT INTO n1_build_settings VALUES (''TREE_TERM_MINREC_SPLIT'', 20)';
EXECUTE IMMEDIATE 'INSERT INTO n1_build_settings VALUES (''TREE_TERM_MINPCT_SPLIT'', .1)';
EXECUTE IMMEDIATE 'INSERT INTO n1_build_settings VALUES (''TREE_TERM_MINREC_NODE'', 10)';
EXECUTE IMMEDIATE 'INSERT INTO n1_build_settings VALUES (''TREE_TERM_MINPCT_NODE'', 0.05)';
*/
```

Split Data into Train and Test

Build and Test Classification Mo

```
/* Build a Classification Model */

EXECUTE IMMEDIATE 'CALL DBMS_DATA_MINING.CREATE_MODEL(''N1_CLASS_MODEL'', ''CLASSIFICATION'',
    ''N1_TRAIN_DATA'', ''CUSTOMER_ID'','' CREDIT_SCORE_BIN'', ''n1_build_settings'')';
DBMS_OUTPUT.PUT_LINE ('Created model: N1_CLASS_MODEL ');

/* Test the Model by generating a apply result and then create a lift result */

EXECUTE IMMEDIATE 'CALL DBMS_DATA_MINING.APPLY(''N1_CLASS_MODEL'',''N1_TEST_DATA'',''CUSTOMER_ID''
    ,''N1_APPLY_RESULT'')';
DBMS_OUTPUT.PUT_LINE ('Created apply result: N1_APPLY_RESULT ');
EXECUTE IMMEDIATE 'CALL DBMS_DATA_MINING.COMPUTE_LIFT(''N1_APPLY_RESULT'',''N1_TEST_DATA''
    ,''CUSTOMER_ID'',''CREDIT_SCORE_BIN'',''N1_LIFT_TABLE'',''Good Credit'',''PREDICTION''
    ,''PROBABILITY'',100)';
DBMS_OUTPUT.PUT_LINE ('Created lift result: N1_LIFT_TABLE ');


END;
```

```
DROP TABLE n1_build_settings PURGE: drop unneccessary - no table exists
CALL DBMS_DATA_MINING.DROP_MODEL('N1_CLASS_MODEL'): drop unneccessary - no model
exists
DROP TABLE N1_APPLY_RESULT PURGE: drop unneccessary - no table exists
DROP TABLE N1_LIFT_TABLE PURGE: drop unneccessary - no table exists
Created N1_TRAIN_DATA
Created N1_TEST_DATA
Created model build settings table: n1_build_settings
Created model: N1_CLASS_MODEL
Created apply result: N1_APPLY_RESULT
Created lift result: N1_LIFT_TABLE
PL/SQL procedure successfully completed.
```

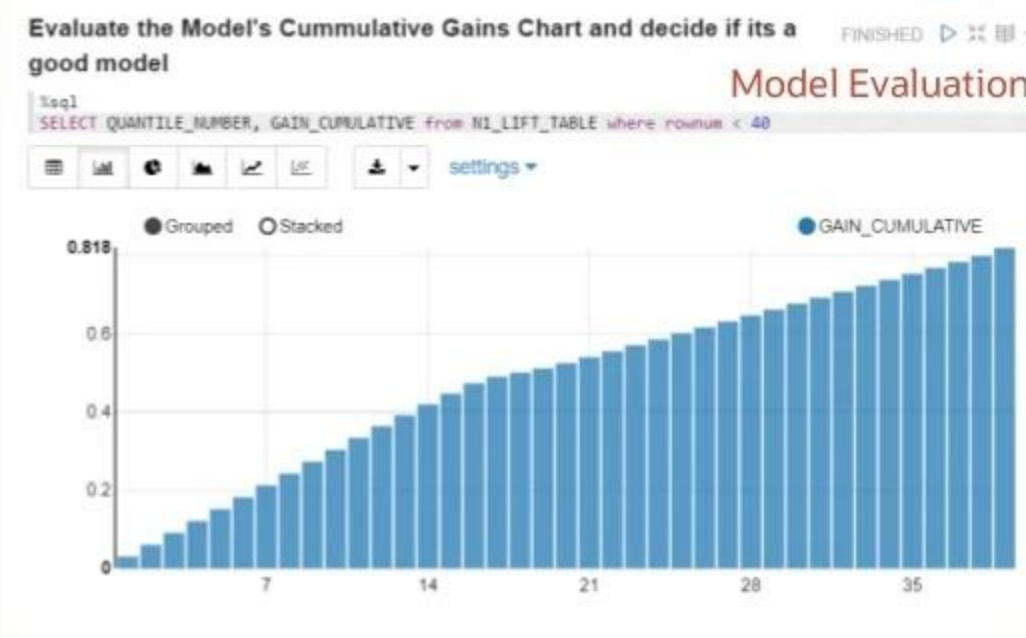
Took 25 sec. Last updated by ADMIN_WS2 at October 17 2018, 2:58:54 PM. (outdated)

Model Evaluation (Machine Learning)

Next, Build Predictive Models to Predict Customers who are Likely to Have Good_Credit

Test the ML model's accuracy

- Randomly selected "hold out" sample of data that was used to train the ML model
- Compute Cumulative Gains, Lift, Accuracy, etc.
- Review the attributes used in the model and model coefficients
- Make sure the model makes sense



Deployment

Apply the Models to Predict “Best Customers”

Simple SQL Apply scripts run 100% inside the Database for immediate ML model deployment

Apply the Oracle Machine Learning Model to New Customers to Show Customers Most Likely to Have Good Credit

FINISHED

Model Apply/”Scoring

```
%sql
select a.customer_id
    , a.prob_Credit_Score_Bin
    , b.age, b.income, b.tenure, b.loan_type, b.loan_amount, b.occupation, b.education_level, b
        .marital_status
from (select * from (select Customer_id, round(prob_Credit_Score_Bin *100,2) prob_Credit_Score_Bin from
    (select Customer_ID, prediction_probability(N1_CLASS_MODEL, NULL using *) prob_Credit_Score_Bin from
    credit_scoring_new_cust_v))) a
, credit_scoring_100k_v b
where a.customer_id = b.customer_id
order by a.prob_Credit_Score_Bin desc
```

CUSTOMER_ID	PROB_CREDIT_SCORE_BIN	AGE	INCOME	TENURE	LOAN_TYPE
34673	100	31	5250	8	Education
77936	100	37	6250	6	Need
56154	100	45	4250	10	Need
11610	100	28	6250	3	Housing
56733	100	63	4250	4	Housing
57999	100	54	4250	33	Housing

Took 0 sec. Last updated by ADWC_WS2 at October 17 2018, 2:51:03 PM. (published)

Coming Soon! | AutoML – *new* with OML4Py

Increase data scientist productivity – reduce overall compute time

Auto Algorithm Selection

- Identify in-database algorithm that achieves highest model quality
- Find best algorithm faster than with exhaustive search

Auto Feature Selection

- Reduce # of features by identifying most predictive
- Improve performance and accuracy

Auto Tune Hyperparamete

- Significantly improve model accuracy
- Avoid manual or exhaustive search techniques

Enables non-expert users to leverage Machine Learning

Coming Soon! | OML AutoML User Interface

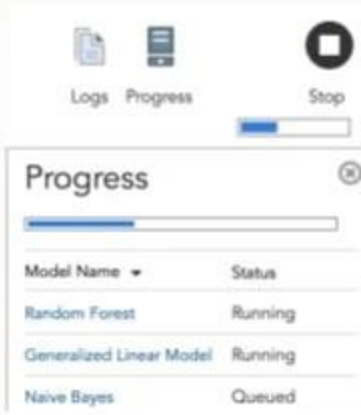
"Code-free" user interface supporting automated end-to-end machine learning

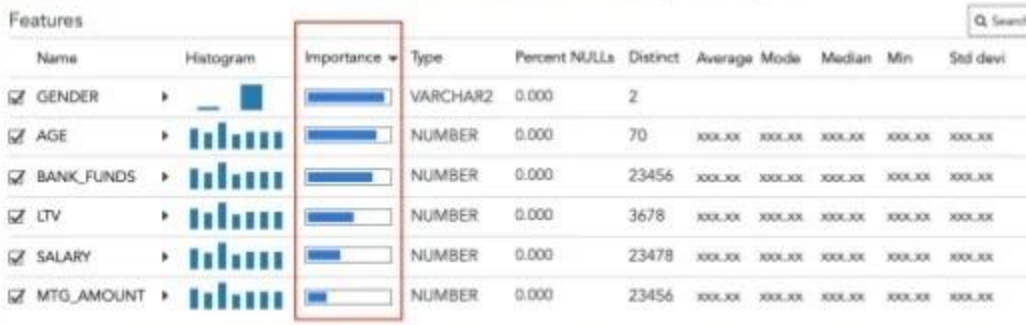
Automate production and deployment of ML models

- Enhance Data Scientist productivity and user-experience
- Enable non-expert users to leverage ML
- Unify model deployment and monitoring
- Support model management

Features

- Minimal user input: data, target
- Model leaderboard
- Model deployment via REST



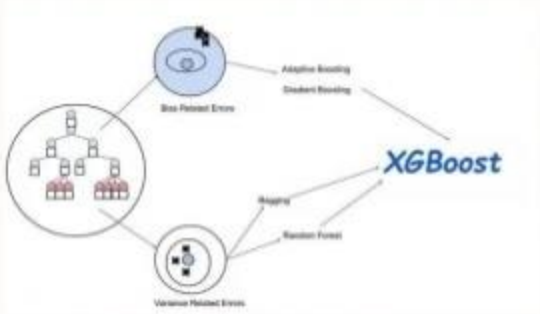


Coming Soon! | Algorithms for Database 20c

Two major new ML algorithms

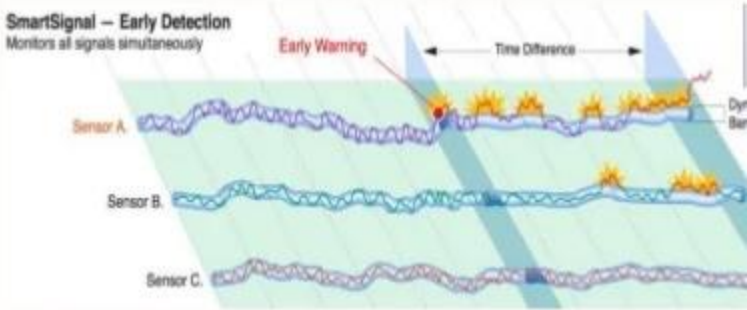
Gradient Boosted Trees (XGBoost)

- Highly popular and powerful algorithm – Kaggle winners
- Classification, regression, ranking, survival analysis



MSET-SPRT

- Multivariate State Estimation Technique - Sequential Probability Ratio Test (MSET-SPRT)
- Nonlinear, nonparametric anomaly detection algorithm designed to monitor critical processes.
- Detects subtle anomalies while also producing minimal false alarms.



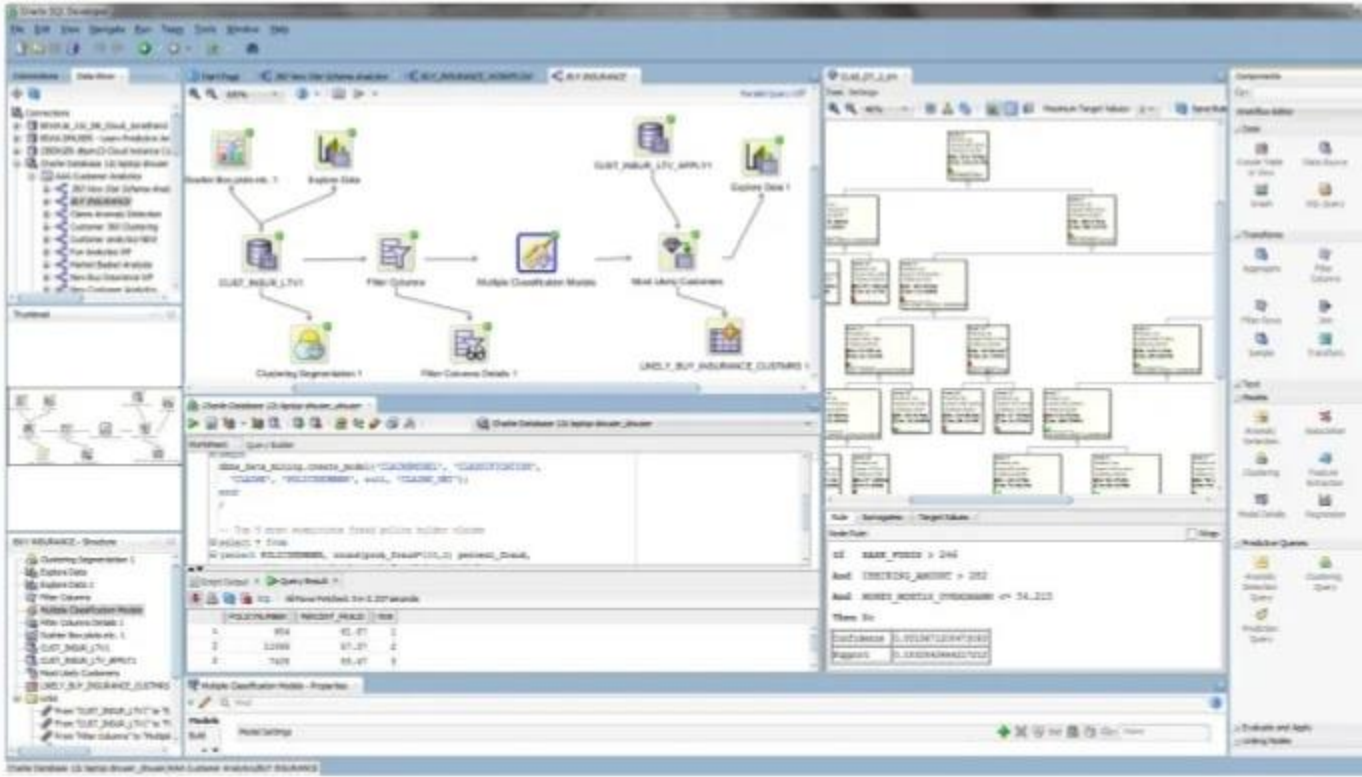
Oracle Data Miner UI

Drag and Drop, Workflows, Easy to Use UI for “Citizen Data Scientist”

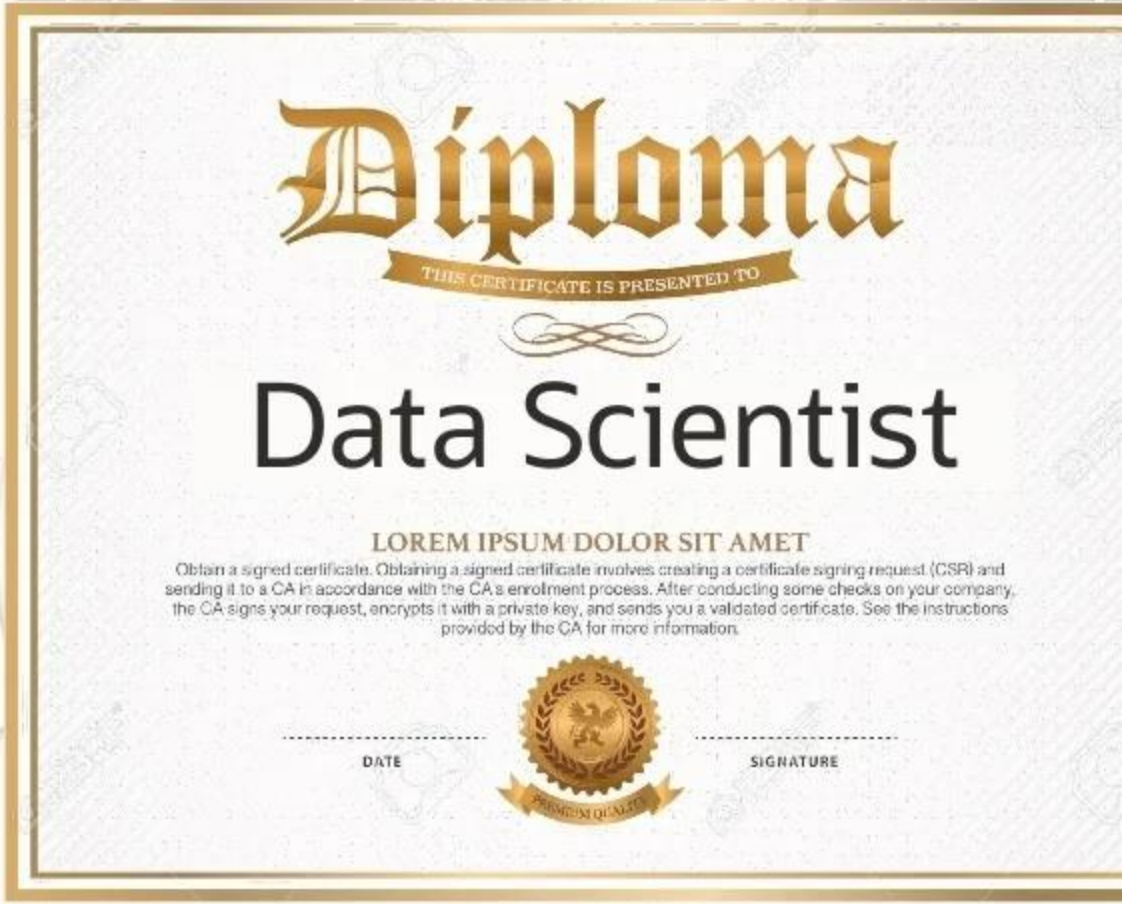
Easy to use to define analytical methodologies that can be shared

SQL Developer Extension

Workflow API and generates SQL code for immediate deployment



Congratulations!
Almost there ☺



Oracle Cloud Data Science Platform

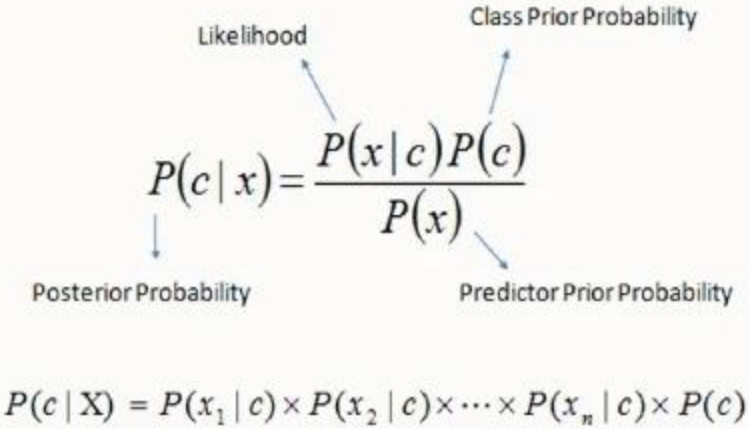
- Oracle Cloud Infrastructure Data Science
  - **AutoML**
    - Automated algorithm selection and tuning
    - Automates the process of running tests against multiple algorithms and hyperparameter configurations
    - Checks results for accuracy and confirms that the optimal model and configuration are selected for use.
    - This saves significant time for data scientists
  - **Feature Selection**
    - Automated predictive feature selection simplifies feature engineering by automatically identifying key predictive features from larger datasets.
  - **Model Evaluation**
    - Measure model performance against new data,
    - Rank models over time to enable optimal behavior in production
  - **Model Explanation**
    - Explanation of the relative weighting and importance of the factors that go into generating a prediction

Oracle Cloud Data Science Platform

- Oracle Cloud Infrastructure Data Science
  - **Notebook Sessions**
    - Built-in cloud-hosted JupyterLab notebook sessions enable teams to build and train models using Python.
  - **Visualization Tools**
    - Use popular open source visualization tools like plotly, matplotlib, and bokeh to visualize and explore data.
  - **Open Source Machine Learning Frameworks**
    - Launch notebook sessions with popular machine learning frameworks like TensorFlow, Jupyter, Dask, Keras, XGboost, and scikit-learn, or bring your own packages.

ML and AI are just “Algorithms”

Algorithms Operate on Data



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Move the Algorithms; Not the Data!;

It Changes *Everything*!

Thank You

Any Questions ?

Sandesh Rao
VP AIOps for the Autonomous Database


@sandeshr

https://www.linkedin.com/in/raosandesh/

https://www.slideshare.net/SandeshRao4