ORACLE

Move the Algorithms; Not the Data!

Oracle Machine Learning

What's New?

Charlie Berger

Sr. Director Product Management, Machine Learning, AI and Cognitive Analytics

charlie.berger@oracle.com @CharlieDataMine

September, 2020

Our mission is to help people
see data in new ways, discover insights,
unlock endless possibilities.

Oracle Database is a Converged Database

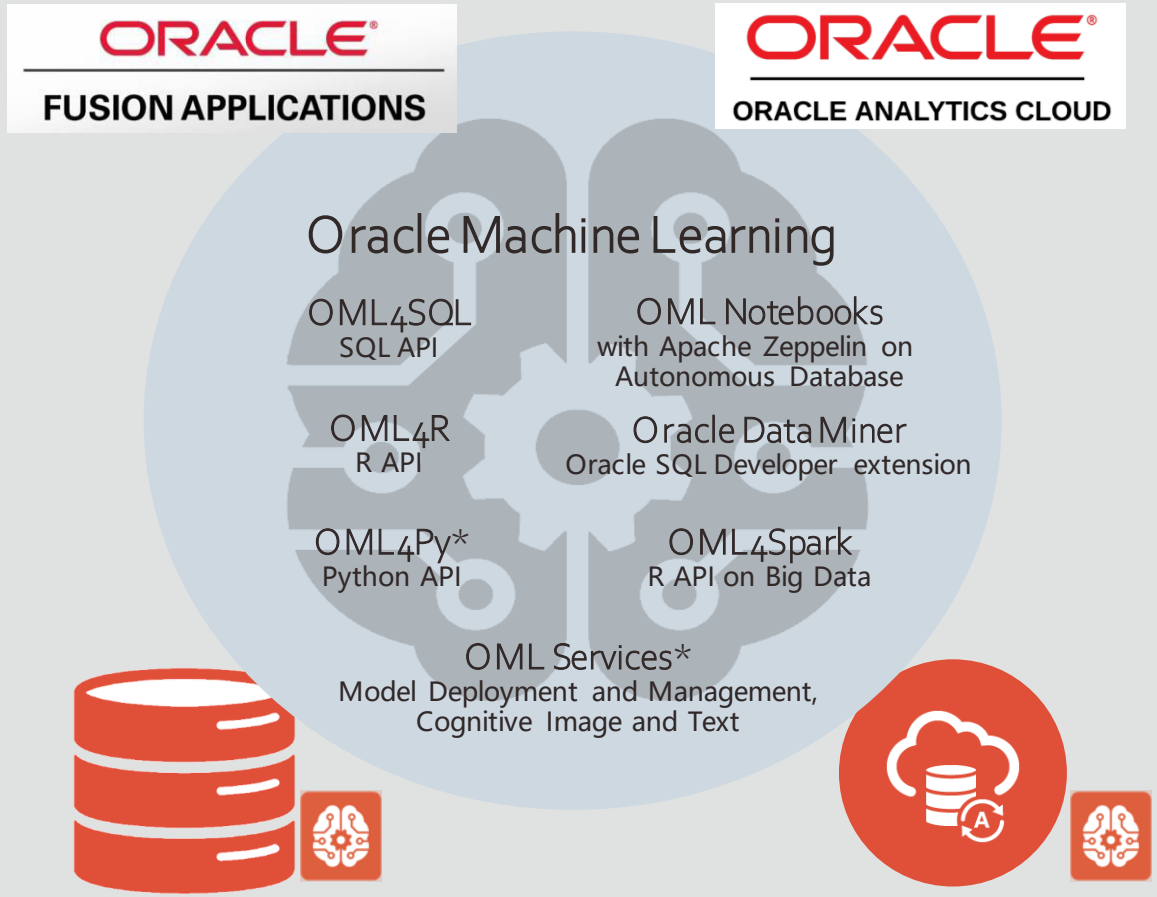
Key Benefits for Developers, Analysts, and Ops

- The richness of Data Driven Apps can make them ultra complex to implement
- By integrating the new data types and paradigms, Oracle makes it dramatically easier and faster to develop Data Driven Apps
  - Open Standard SQL across all data plus popular APIs: REST
  - Transactions across all data
  - Consistent, Queryable View of all data
  - Common Reliability, Scalability, Security
  - Common Management
  - Harvest more information, discover new insights and make predictions

Oracle Machine Learning

Oracle Machine Learning *extends* Oracle Database(s) and enables users to build "AI" applications and analytics dashboards

OML delivers powerful in-database machine learning algorithms, automated ML functionality via SQL APIs and integration with open source Python* and R

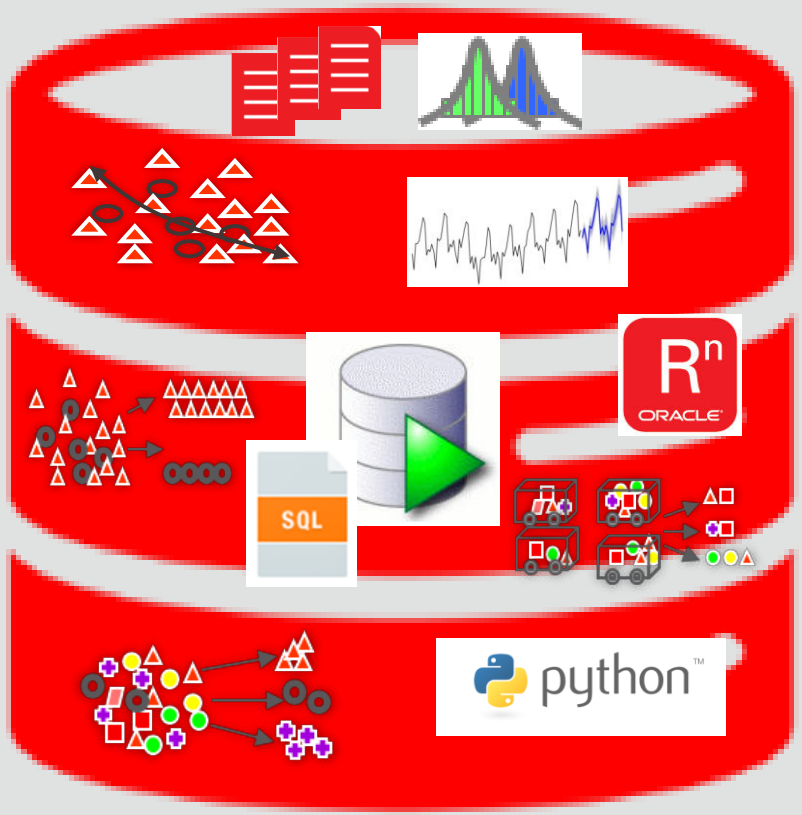


Oracle Makes Machine Learning Simple

** Coming this Fall*

Move the Algorithms, Not the Data!

In-Database Processing



Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

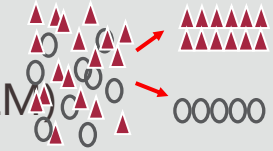
Posterior Probability

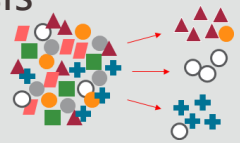
Predictor Prior Probability

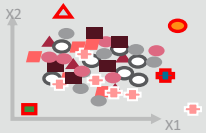
$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

An "AI Database"? A "Thinking Database"? It Changes *Everything*!

Oracle Machine Learning Algorithms

Classification

Naïve Bayes
Logistic Regression (GLM)
Decision Tree
Random Forest
Neural Network
Support Vector Machine
Explicit Semantic Analysis
XGBoost*

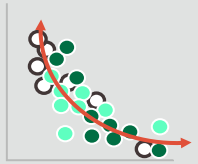
Clustering

Hierarchical K-Means
Hierarchical O-Cluster
Expectation Maximization (EM)

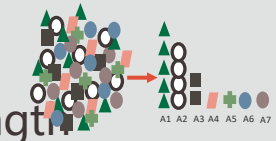
Anomaly Detection


One-Class SVM
MSET-SPRT*

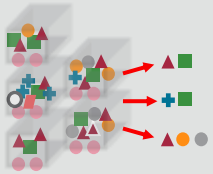
Time Series

Forecasting - Exponential Smoothing
Includes popular models e.g. Holt-Winters with trends, seasonality, irregularity, missing data

Regression

Linear Model
Generalized Linear Model
Support Vector Machine (SVM)
Stepwise Linear regression
Neural Network
XGBoost*

Attribute Importance


Minimum Description Length
Principal Comp Analysis (PCA)
Unsupervised Pair-wise KL Div
CUR decomposition for row & AI

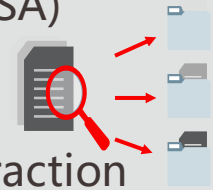
Association Rules

A priori/ market basket

Predictive Queries
Predict, cluster, detect, features

SQL Analytics


SQL Windows
SQL Patterns
SQL Aggregates

Feature Extraction
Principal Comp Analysis (PCA)
Non-negative Matrix Factorization
Singular Value Decomposition (SVD)
Explicit Semantic Analysis (ESA)

Text Mining Support

Algorithms support text
Tokenization and theme extraction
Explicit Semantic Analysis (ESA) for document similarity

Statistical Functions

Basic statistics: min, max, median, stdev, t-test, F-test, Pearson's, Chi-Sq, ANOVA, etc.

R & Python
** Coming this Fall*

Third-party R & Python Packag through Embedded Execution
Spark MLlib algorithm integrati

Model Deployment & Monitori
** Coming this Fall*

SQL—1st Class Objects
Oracle RESTful API (ORDS)
OML Services

Includes support for Partitioned Models, Transactional data and aggregations, Unstructured data, Geo-spatial data, Graph data. etc,

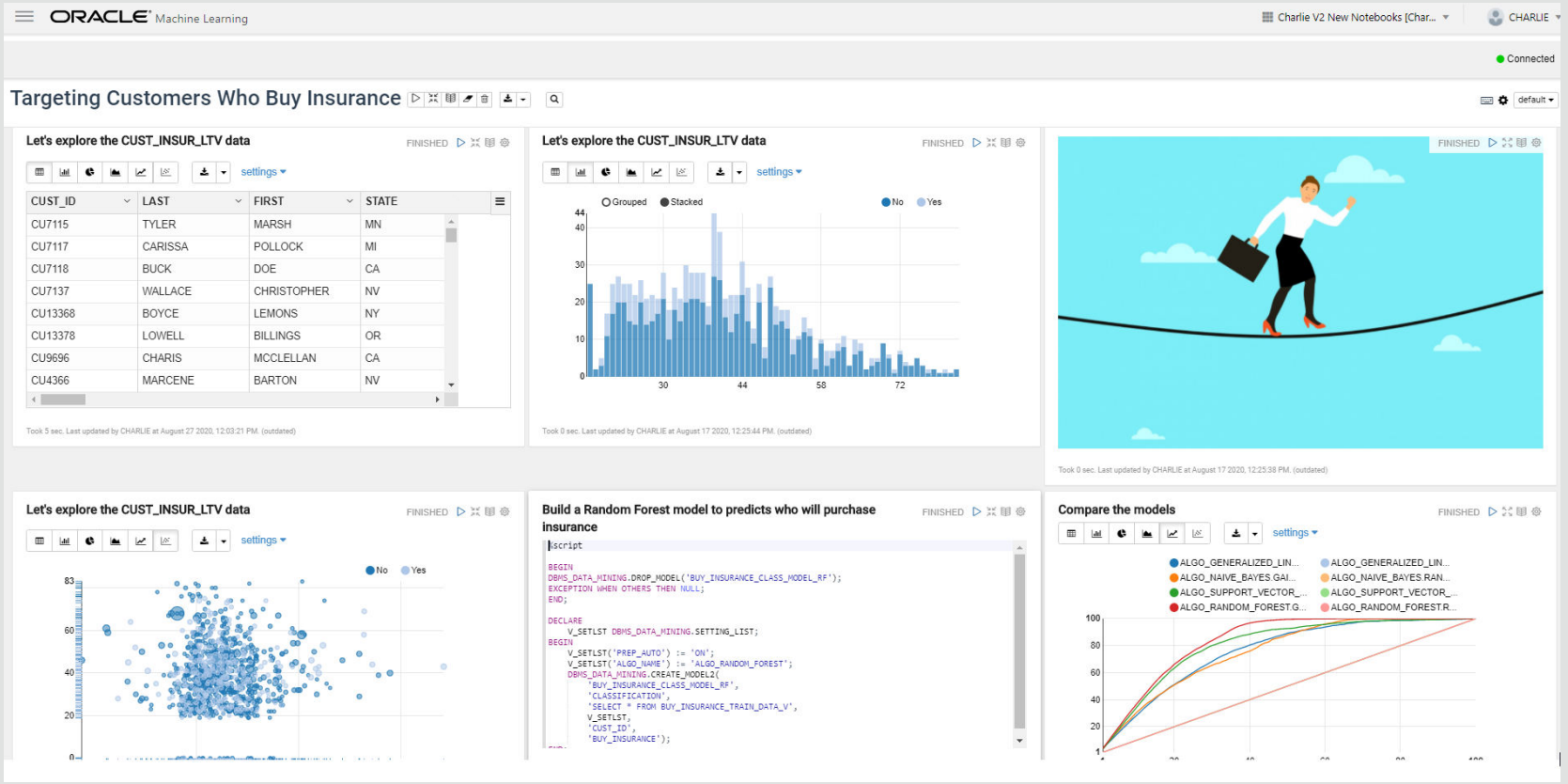
*New in 20c

Oracle Machine Learning

Oracle Machine Learning Notebooks included in Autonomous Databases

Key Features:

- Collaborative UI for data scientists & analysts
- Packaged with Autonomous Databases
- Easy access to shared notebooks, templates, permissions, scheduler, etc.
- Example notebooks
- OML4SQL
- OML4Py *coming this Fall*



OML to Predict Customer Behavior

Intuitive SQL API—OML4SQL

```
-- Build and train ML model to determine which customers will buy Travel Insurance

DECLARE
    v_setlst DBMS_DATA_MINING.SETTING_LIST;
BEGIN
    v_setlst('ALGO_NAME')  := 'ALGO_SUPPORT_VECTOR_MACHINES';
    V_setlst('PREP_AUTO')  := 'ON';

    DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME          => 'BUY_TRVL_INSUR',
        MINING_FUNCTION     => 'CLASSIFICATION',
        DATA_QUERY          => 'select * from CUSTOMERS',
        SET_LIST            => v_setlst,
        CASE_ID_COLUMN_NAME => 'CUST_ID',
        TARGET_COLUMN_NAME  => BUY_TRAVEL_INSURANCE');
END;
```

OML to Predict Customer Behavior

Intuitive SQL API—OML4SQL

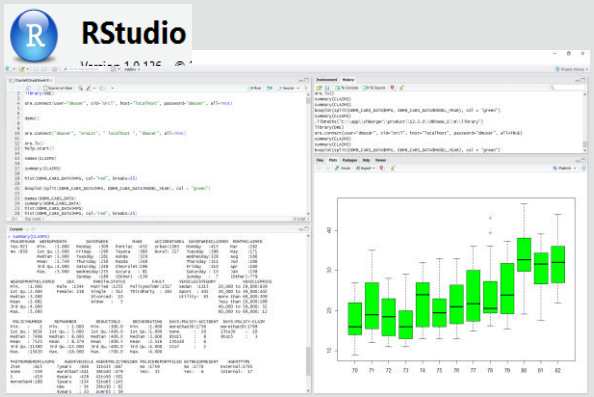
```
-- Build and train ML model to determine which customers will buy Travel Insurance

DECLARE
    v_setlst DBMS_DATA_MINING.SETTING_LIST;
BEGIN
    v_setlst('ALGO_NAME')  := 'ALGO_SUPPORT_VECTOR_MACHINES';
    V_setlst('PREP_AUTO')  := 'ON';
    DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME          => 'BUY_TRVL_INSUR',
        MINING_FUNCTION     => 'CLASSIFICATION',
        DATA_QUERY          => 'select * from CUSTOMERS',
        SET_LIST            => v_setlst,
        CASE_ID_COLUMN_NAME => 'CUST_ID',
        TARGET_COLUMN_NAME  => BUY_TRAVEL_INSURANCE');
END;
```

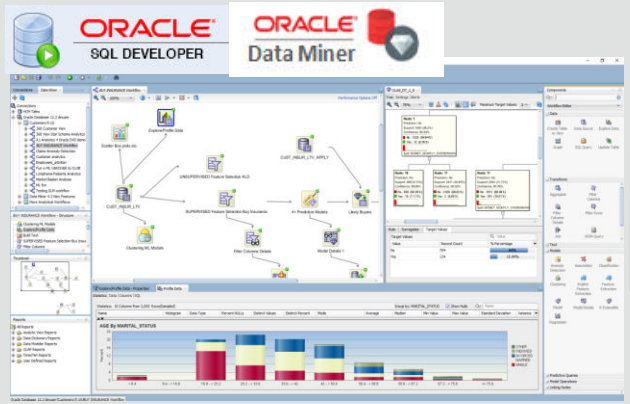
```
-- Apply ML model to sales to predict which customers are likely to buy

SELECT prediction_probability(BUY_TRVL_INSUR, 'Yes'
          USING 3500 as bank_funds, 37 as age, 'Married' as marital_status, 2 as num_previous_cruises)
FROM dual;
```

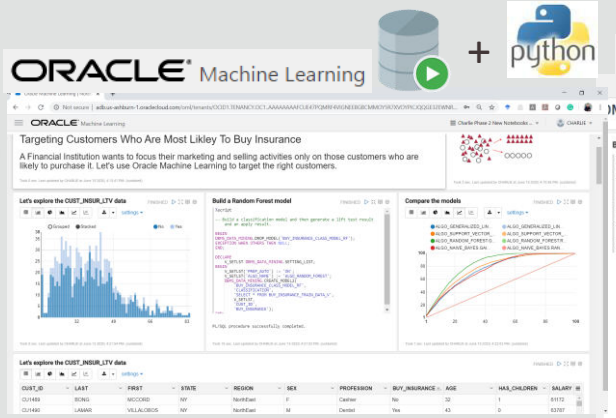
SQL | All Rows Fetched: 1 in 0.043 seconds

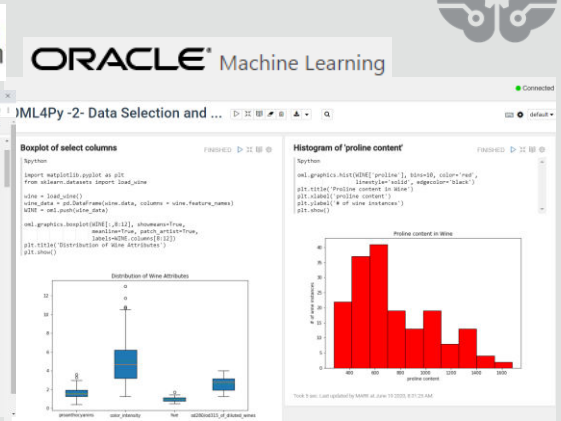
	PREDICTION_PROBABILITY(BUY_INSUR1,'YES'USING3500ASBANK_FUNDS,825ASCHECKING_AMOUNT,400ASCREDIT_BALANCE
1	0.9276956709910801

Goal: Manage and Analyze All Your Data









SQL / R / Python

Boil down the Data Lake

Big Data SQL

"Engineered Features"
– Derived attributes that reflect domain knowledge—key to best models e.g.:

- Counts
- Totals
- Changes over time

Coming Soon!

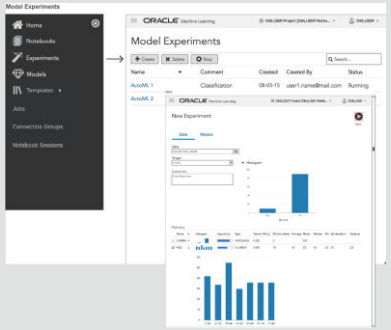
Oracle Machine Learning

What's Coming....

Autonomous Database

Oracle Machine Learning for Python (OML4Py) – *Target Fall 2020*

- OML4Py, Embedded Python execution, Python & REST APIs, SQL API afterwards
- AutoML functionality
- Support for 3rd party Python packages - *H2 FY21*



AutoML UI – *Target Fall 2020*

- Code-free interface with OML Notebooks
- Time series, anomaly detection – Target H2 FY21

OML Services – *Target Fall 2020*

- REST API for model management, deployment, ONNX and cognitive text

{ REST }

Oracle Machine Learning for R (OML4R) – *Target CY2020*

- OML4R, Embedded R execution will also support a REST API followed by a SQL API

Oracle Database 20c+ DBCS

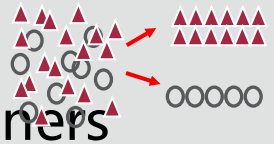
20c: New Algorithms: XGBoost, MSET-SPRT

OML4Py – *Target Fall 2020* (18c >)

Coming in 20c | New OML Algorithms

New Algorithms

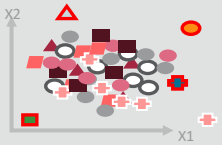
Gradient Boosted Trees (XGBoost)

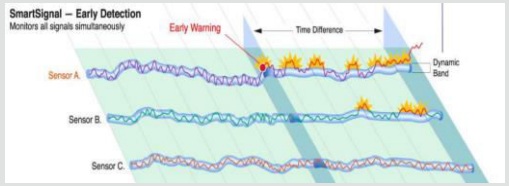


XGBoost

- 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
- Calibrates expected behavior from historical normal operational sequence of monitored signals
- Re-implemented and sped up in-DB and based on original Oracle Labs algorithm

Oracle Machine Learning for R / Python* *Coming soon

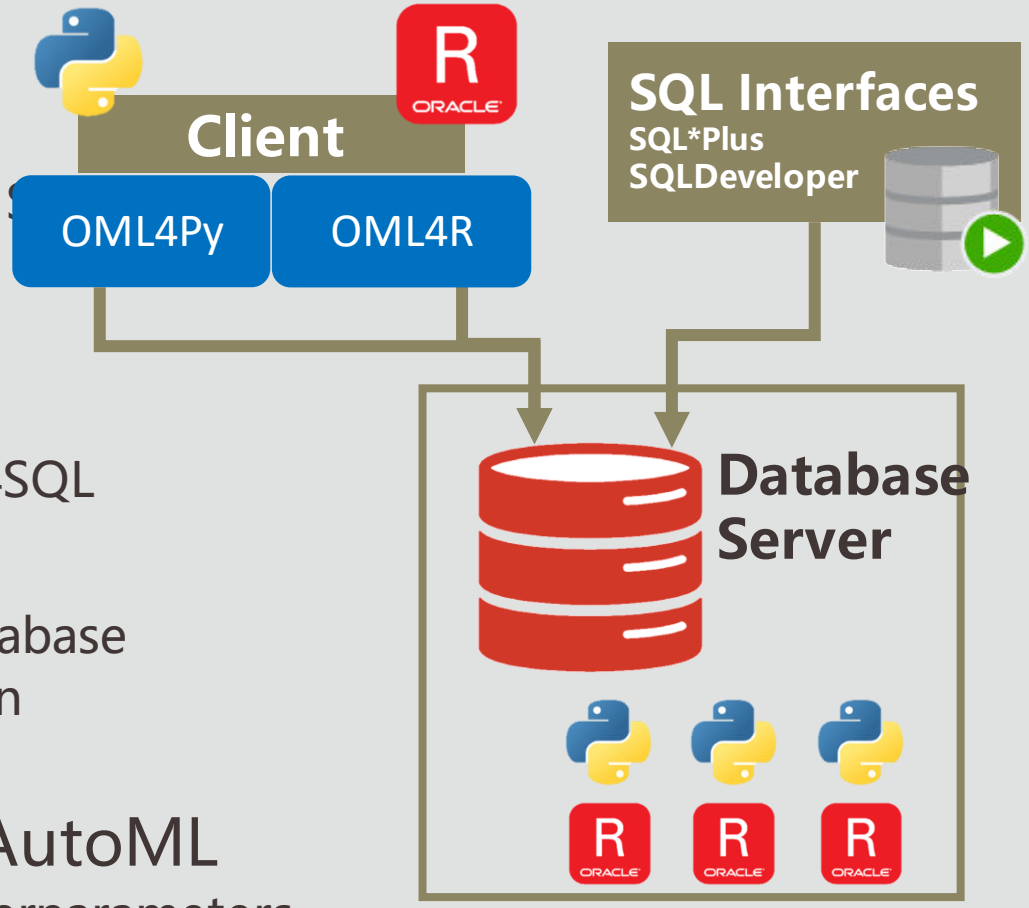
Integration with Open Source ML

Transparency layer
- Leverage proxy objects so data remain in database
- Overload native functions translating functionality to S
- Use familiar R/Python syntax to manipulate database

Parallel, distributed algorithms
- Scalability and performance
- Exposes in-database algorithms available from OML4SQL

Embedded execution
- Manage and invoke R or Python scripts in Oracle Database
- Data-parallel, task-parallel, and non-parallel execution
- Use open source packages to augment functionality

OML4Py, Automated Machine Learning - AutoML
- Algorithm selection, feature selection, auto tune hyperparameters

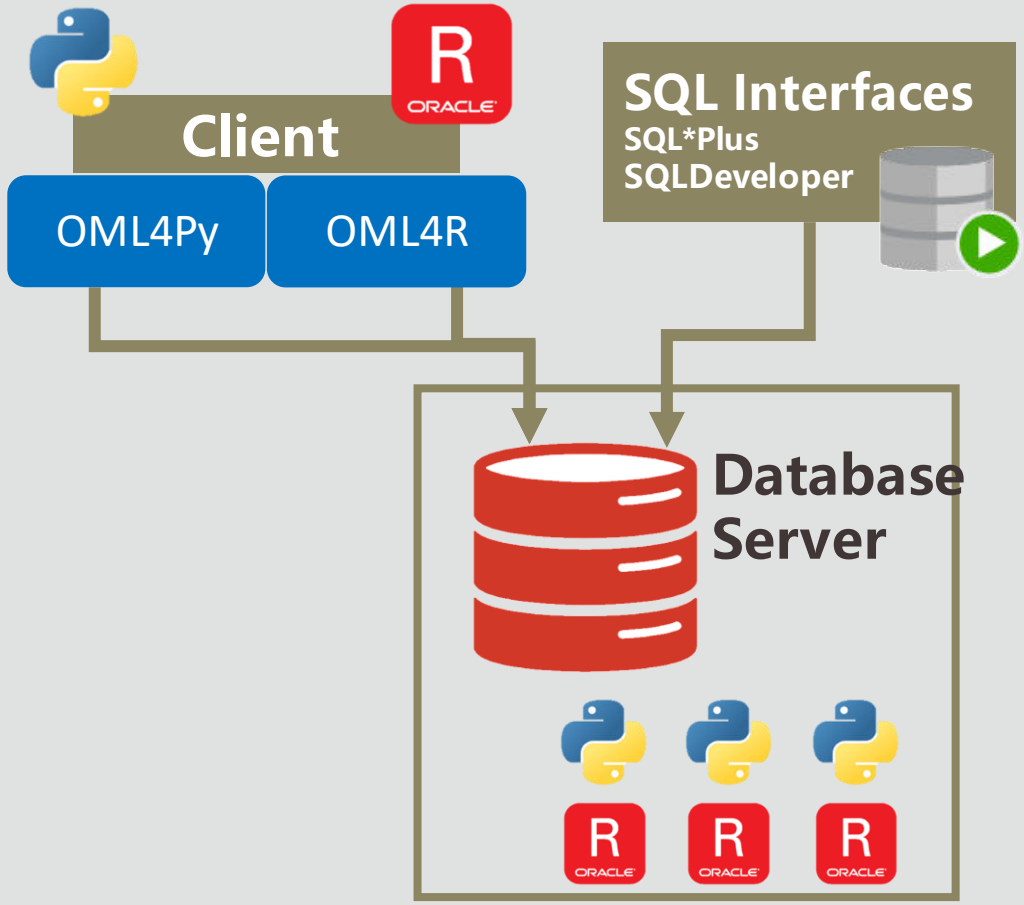


Oracle Machine Learning for R / Python* *Coming soon*

Preinstalled and supported Python Packages

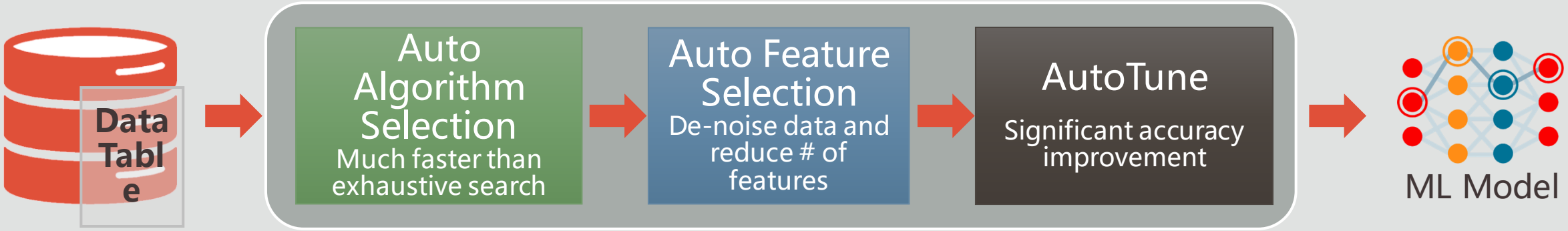
OML4Py details and supported packages

- Python 3.8.2
- cx_Oracle 7.3.0
- cycler 0.10.0
- joblib 0.14.0
- kiwisolver 1.1.0
- matplotlib 3.1.2
- numpy 1.18.1
- pandas 0.25.3
- pyparsing 2.4.0
- python_dateutil 2.8.1
- pytz 2019.3
- scikit-learn 0.22.1
- scipy 1.4.1
- six 1.13.0
- pyreadline 2.1* - windows only
- setuptools 41.2.0



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 Hyperparameters

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

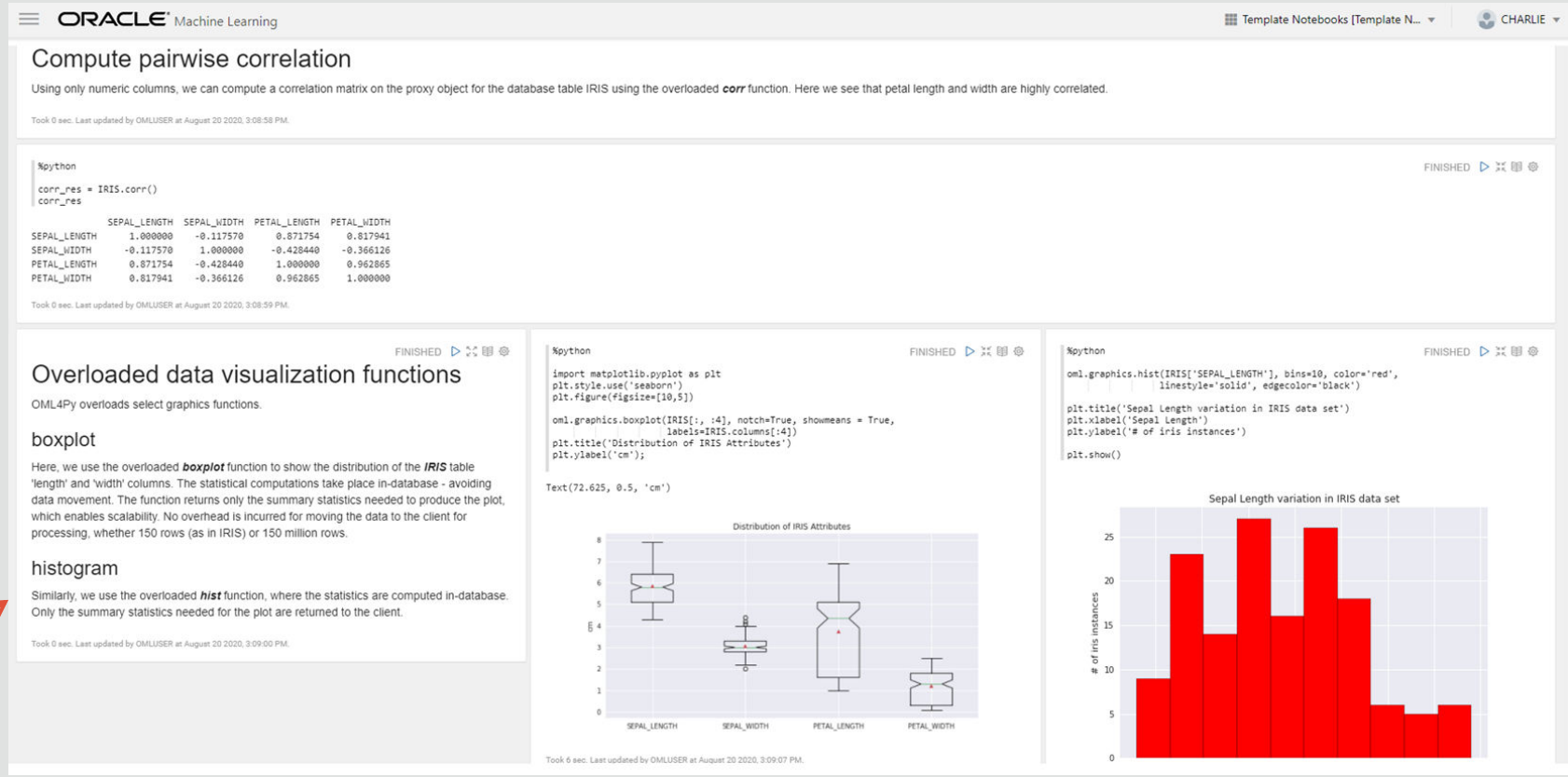
Enables non-expert users to leverage Machine Learning

Oracle Machine Learning

Oracle Machine Learning Notebooks included in Autonomous Databases

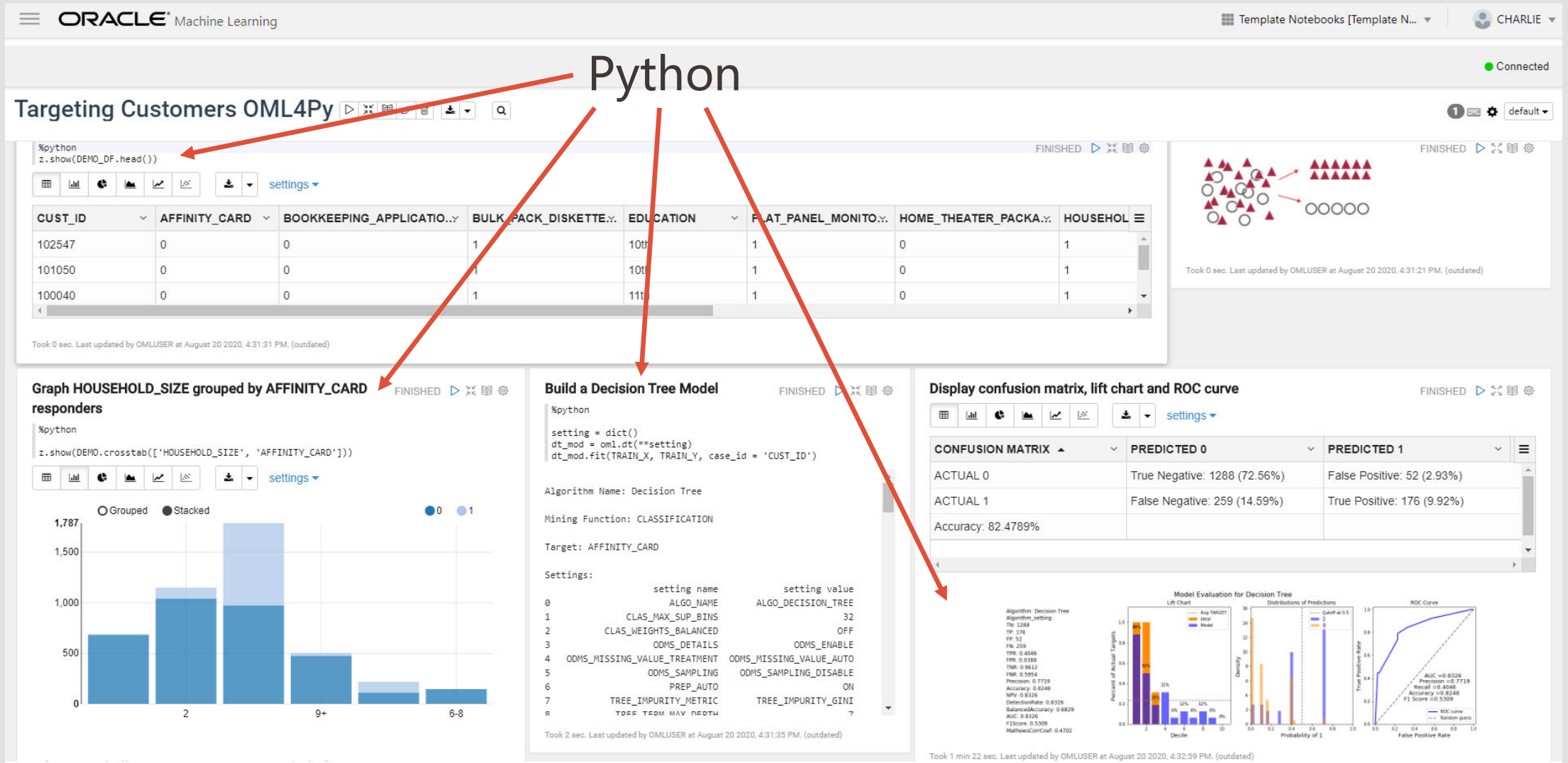
Key Features:

- Collaborative UI for data scientists & analysts
- Packaged with Autonomous Databases
- Easy access to shared notebooks, templates, permissions, scheduler, etc.
- Example notebooks
- OML4SQL
- OML4Py *coming this Fall*



Oracle Machine Learning

OML4Py Python Integration Added to OML on Autonomous Database



Coming Soon! | AutoML User Interface

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

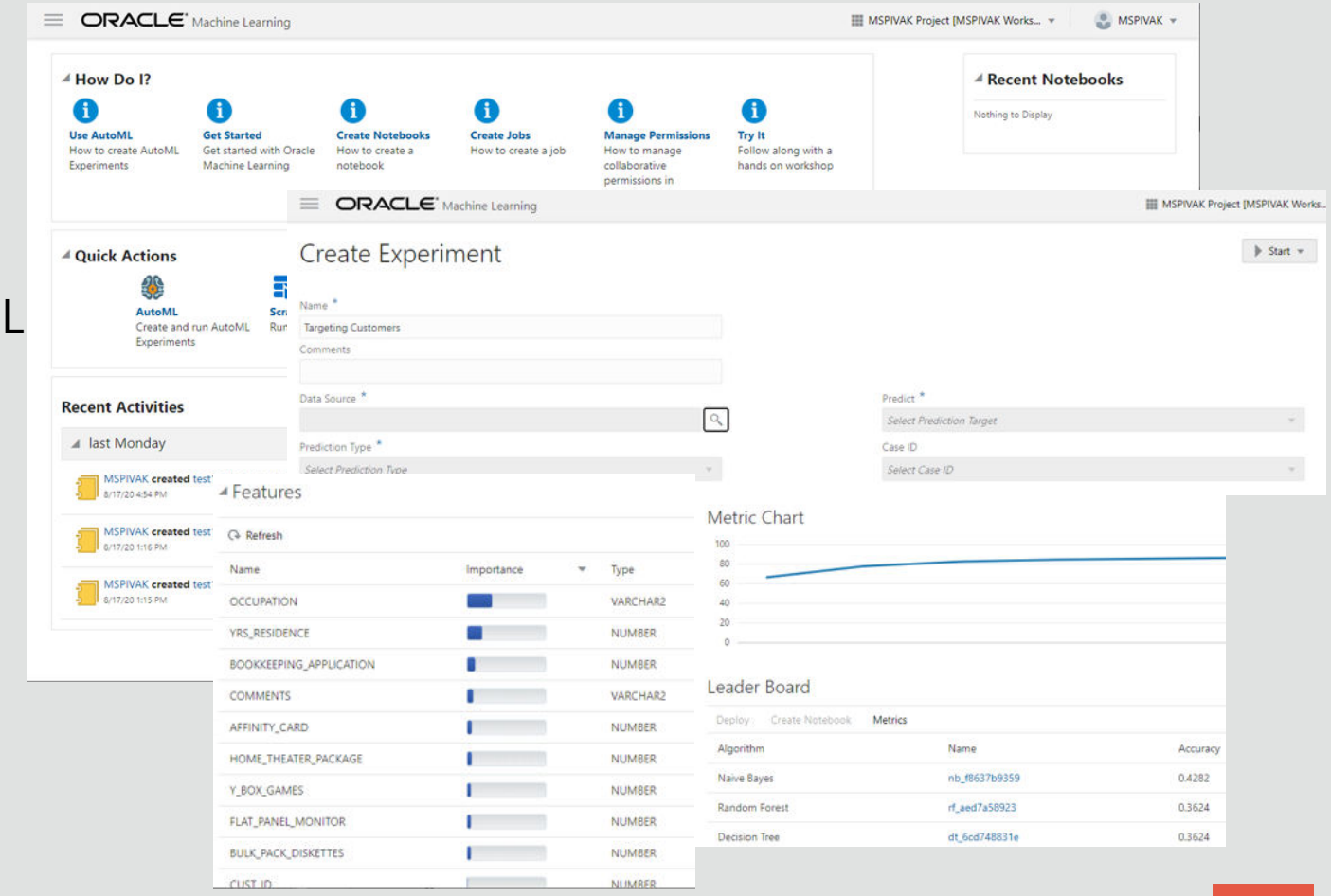
Enables Citizen Data Scientists

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

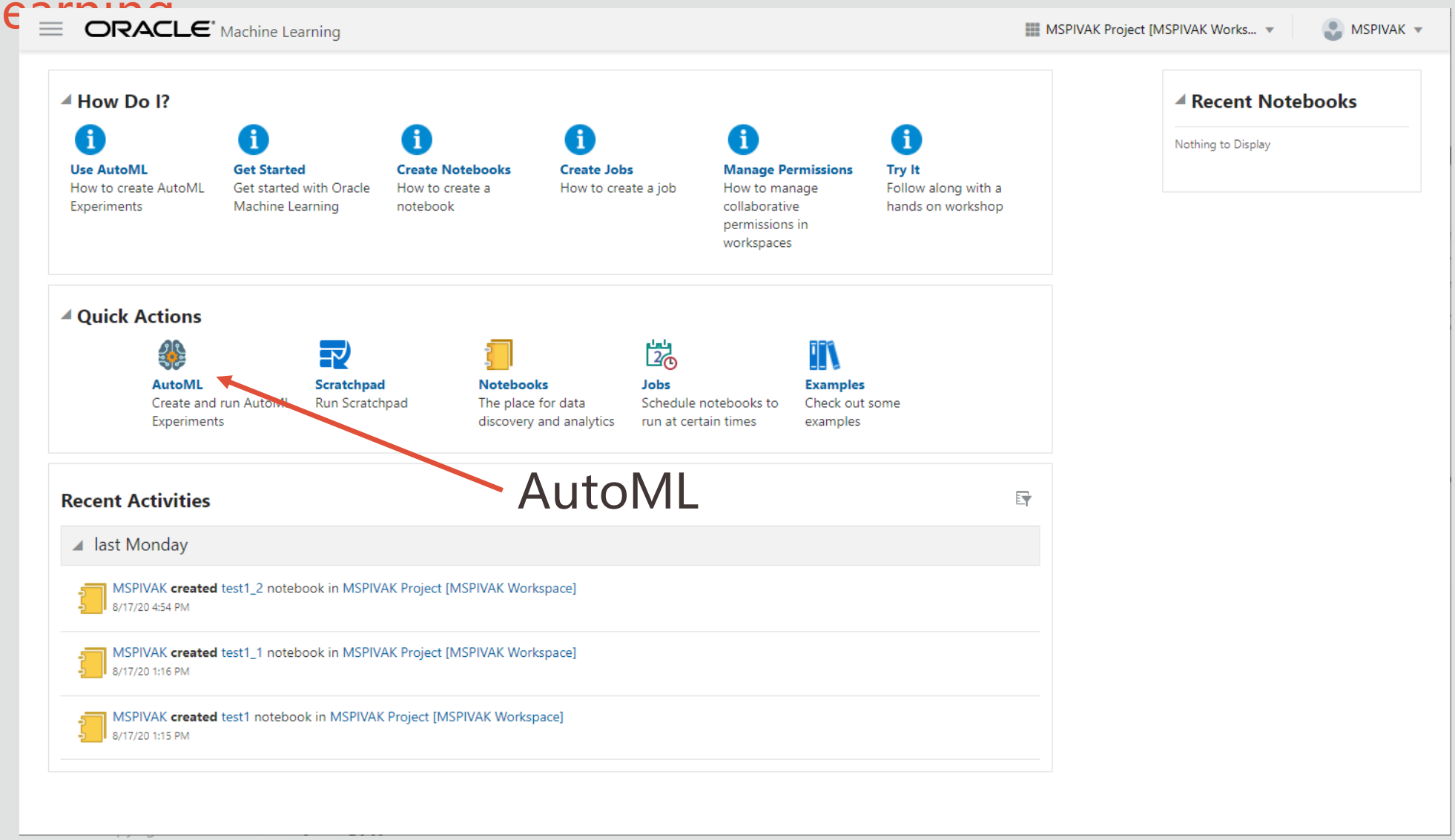
Features

- Minimal user input: data, target
- Model leaderboard
- Model deployment via REST
- Model monitoring
- Cognitive features for image and text



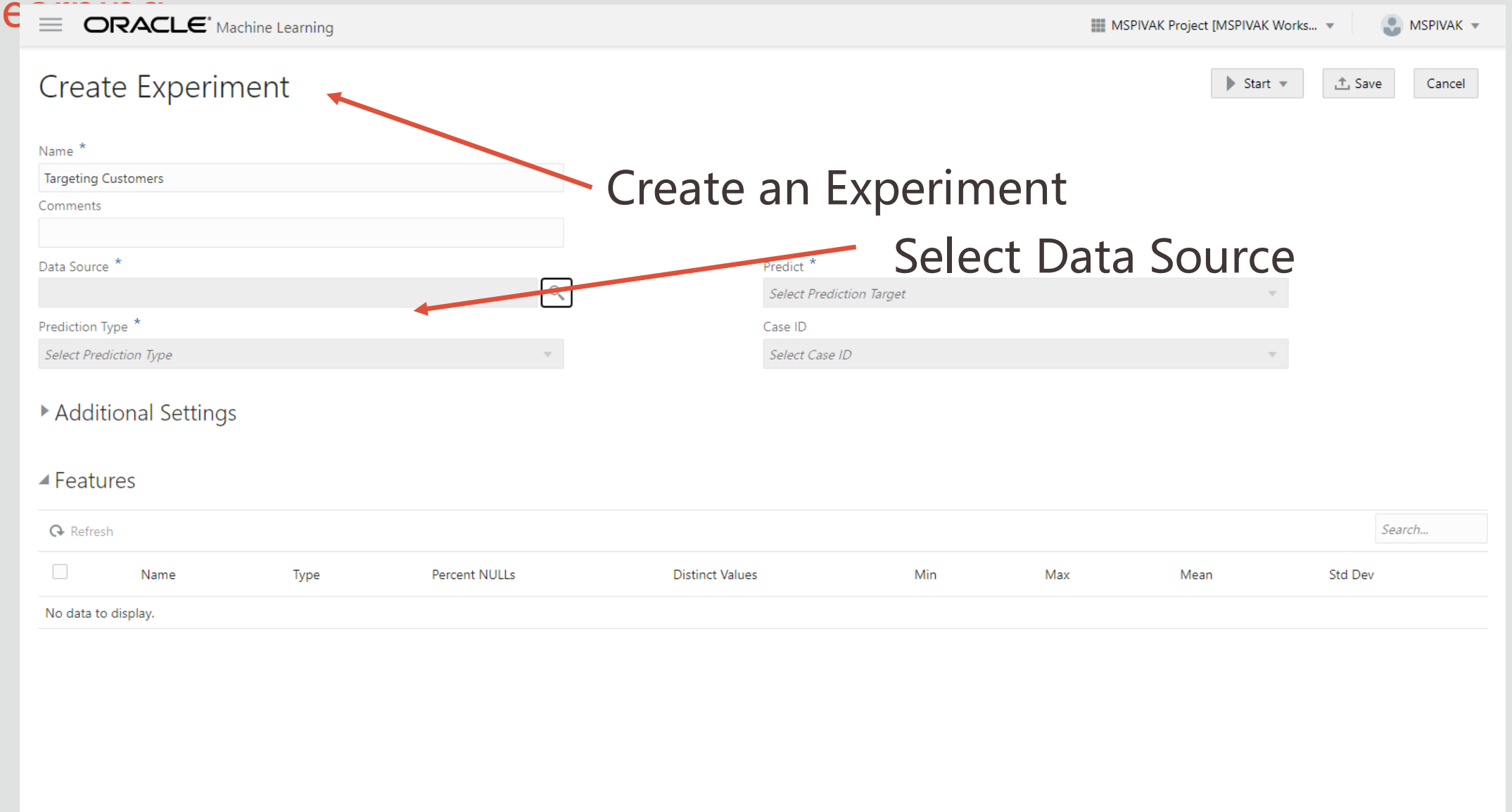
Coming Soon! | AutoML User Interface

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



Coming Soon! | AutoML User Interface

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

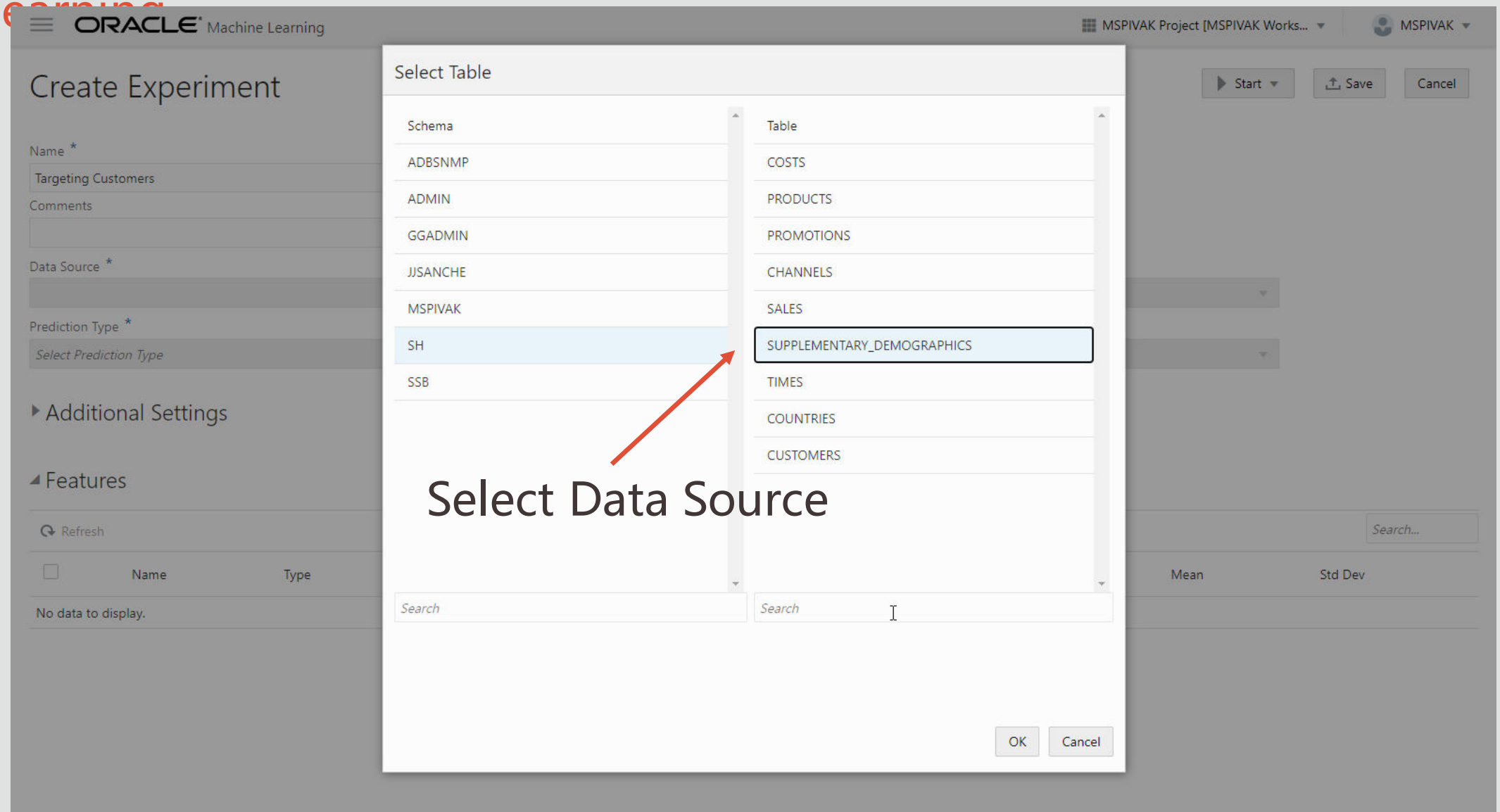


Create an Experiment

Select Data Source

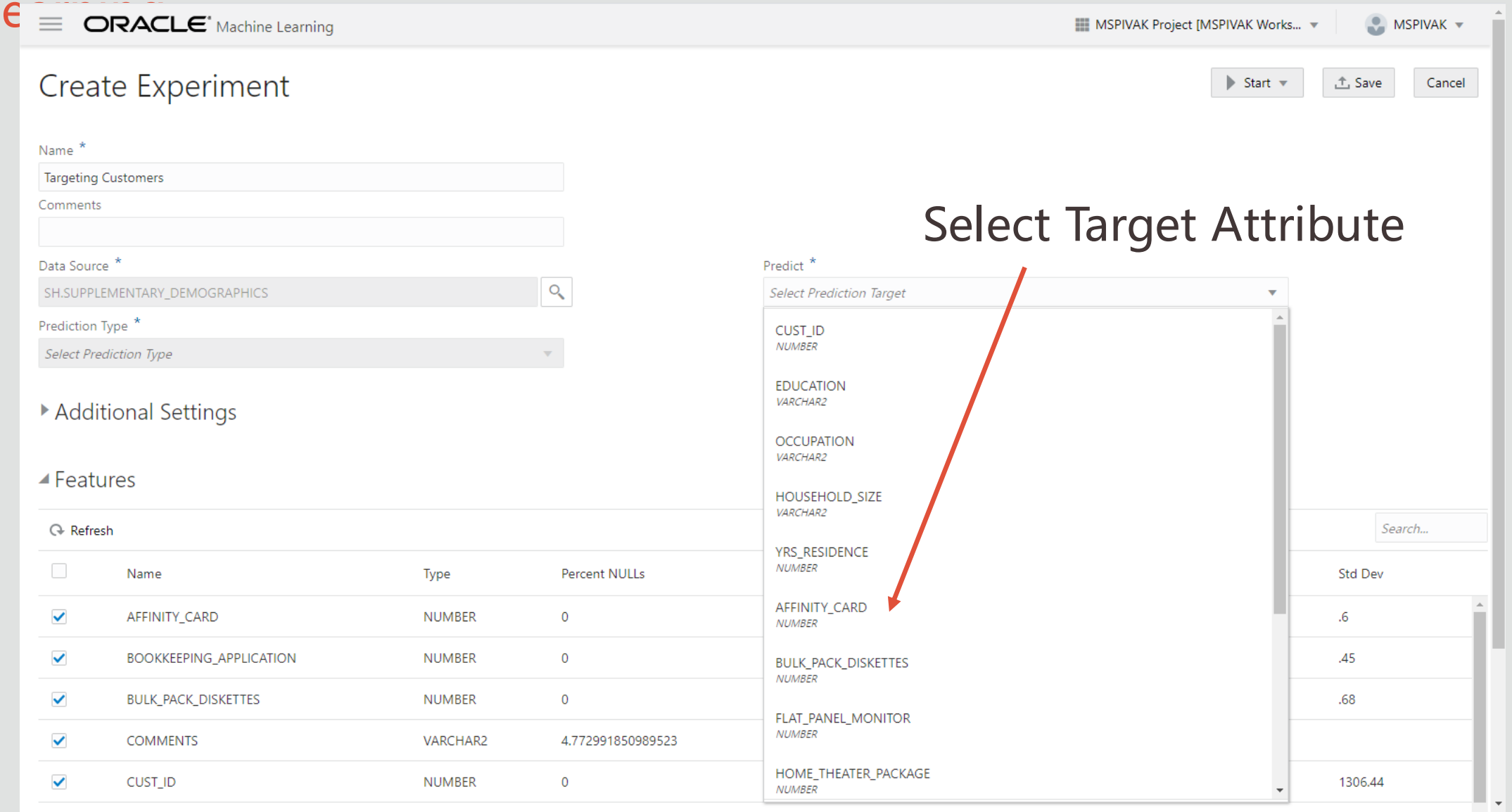
Coming Soon! | AutoML User Interface

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



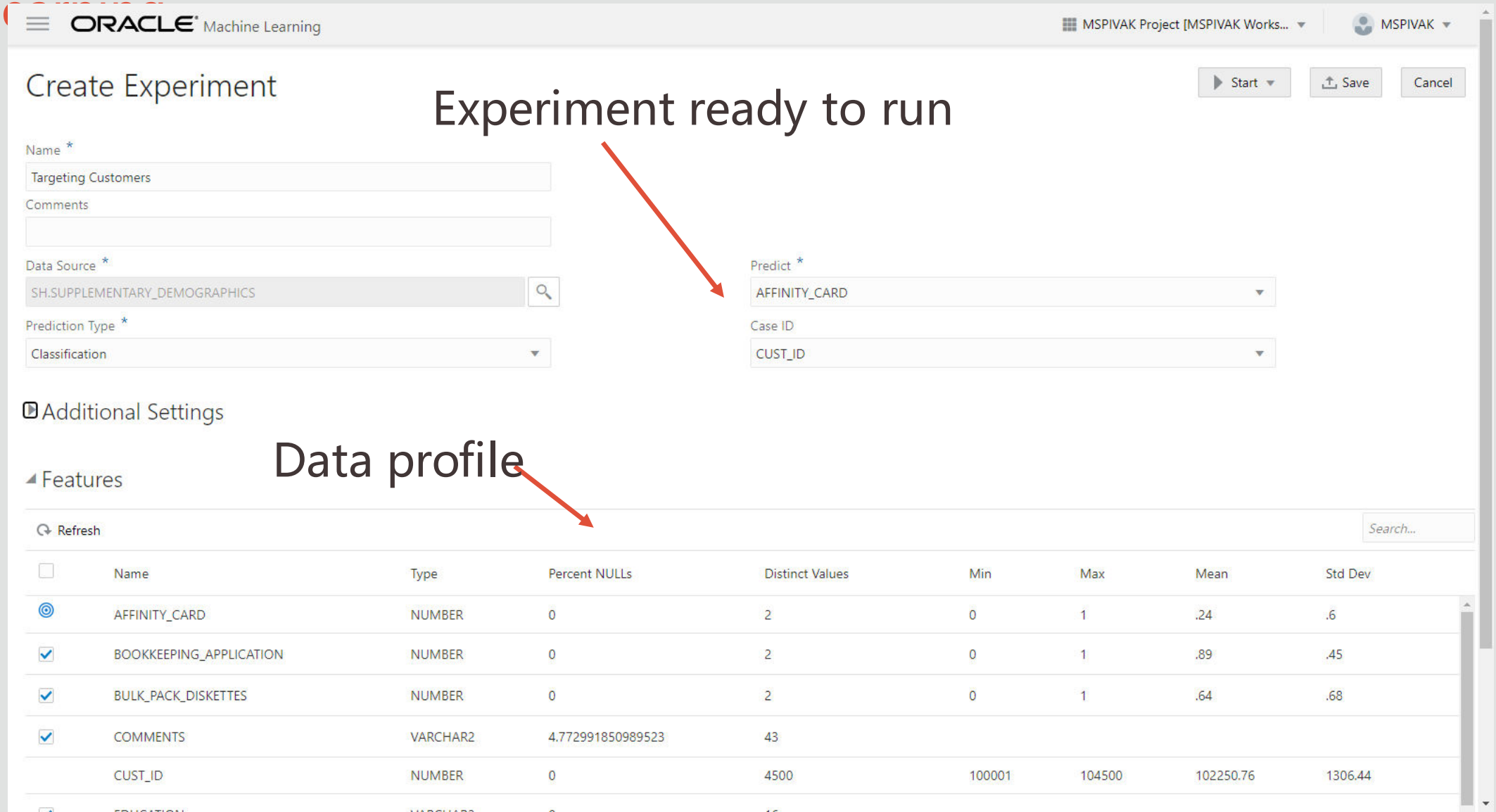
Coming Soon! | AutoML User Interface

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



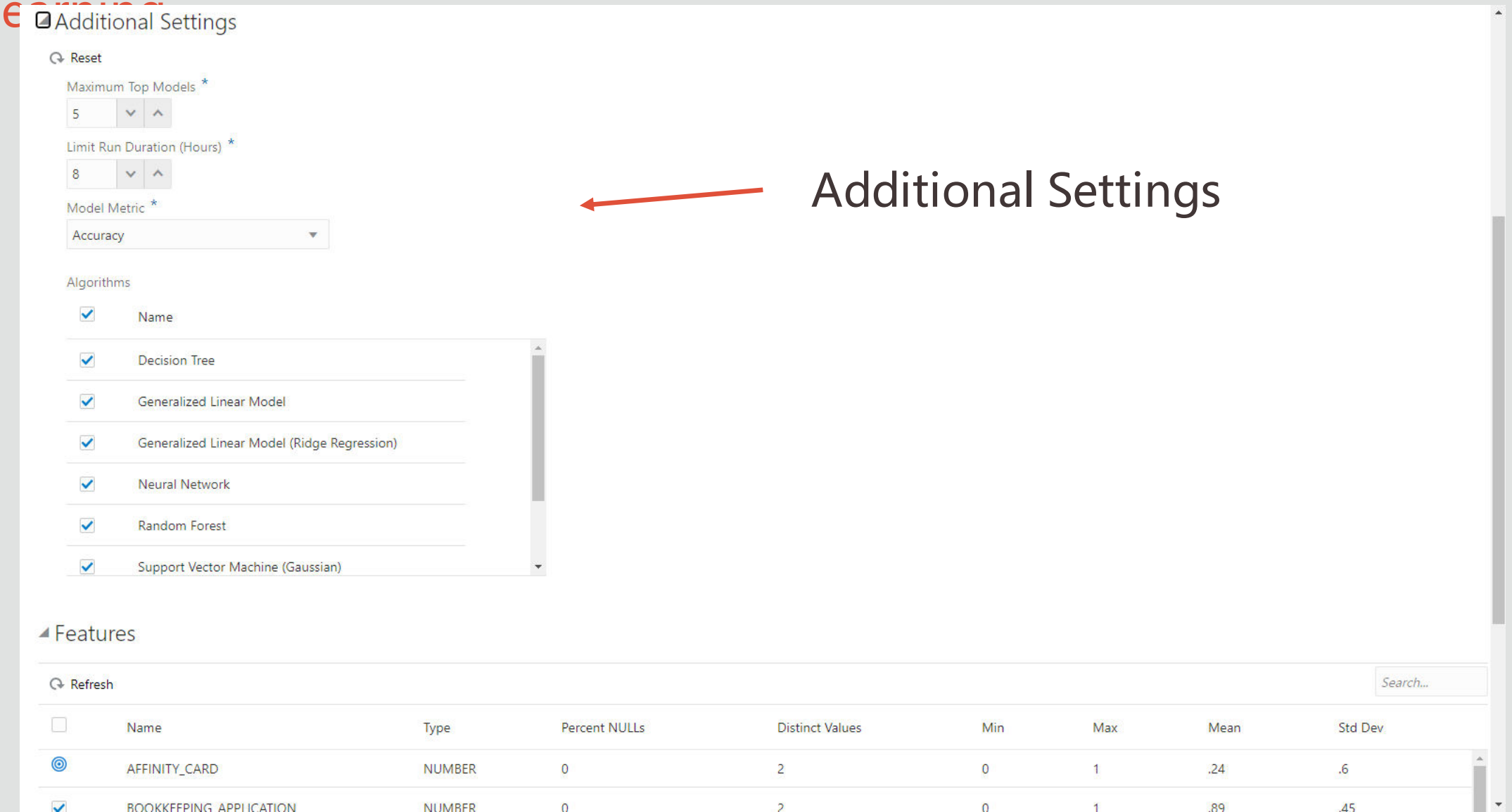
Coming Soon! | AutoML User Interface

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



Coming Soon! | AutoML User Interface

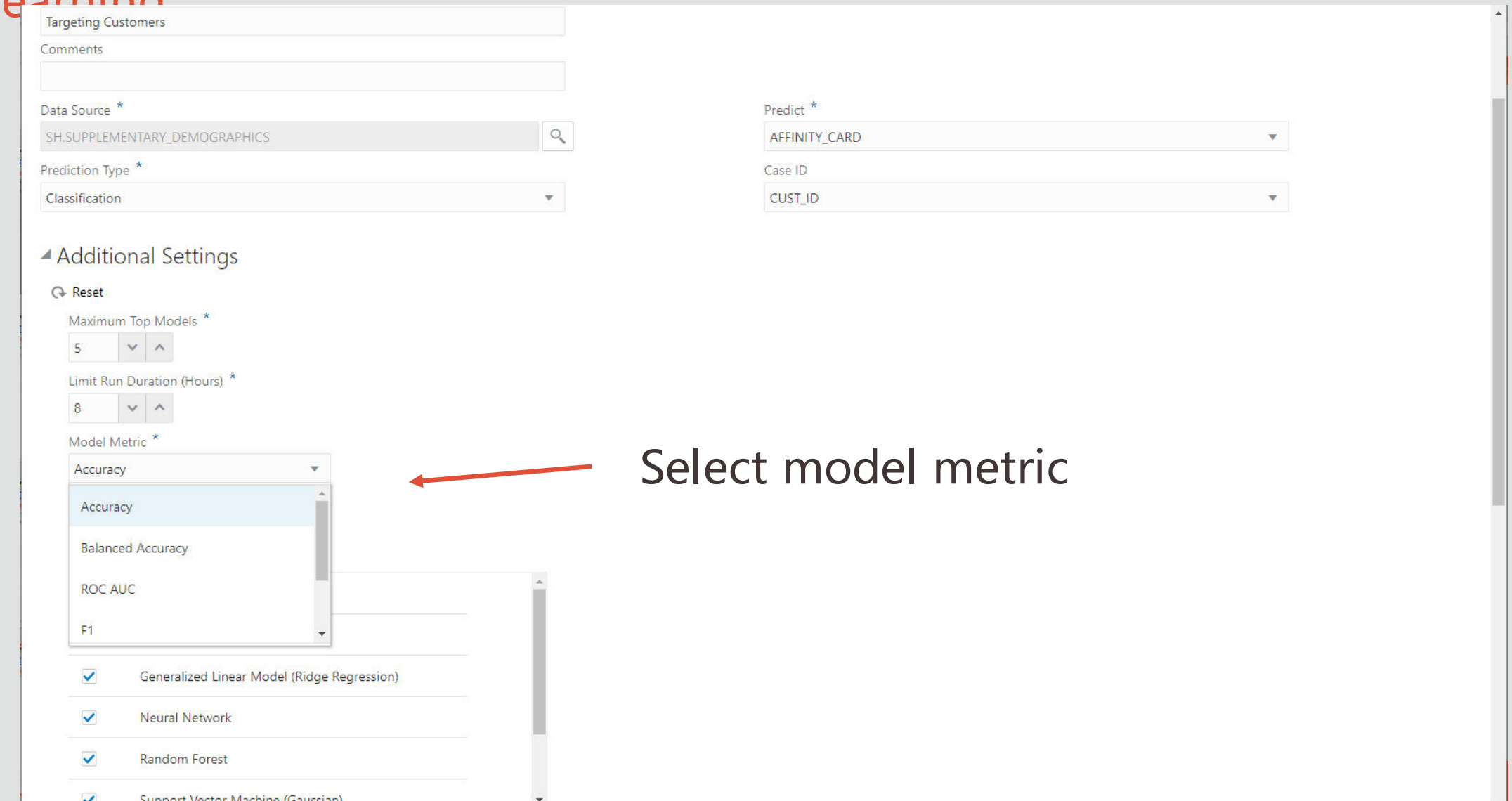
"Code-free" user interface supporting automated end-to-end machine learning



Additional Settings

Coming Soon! | AutoML User Interface

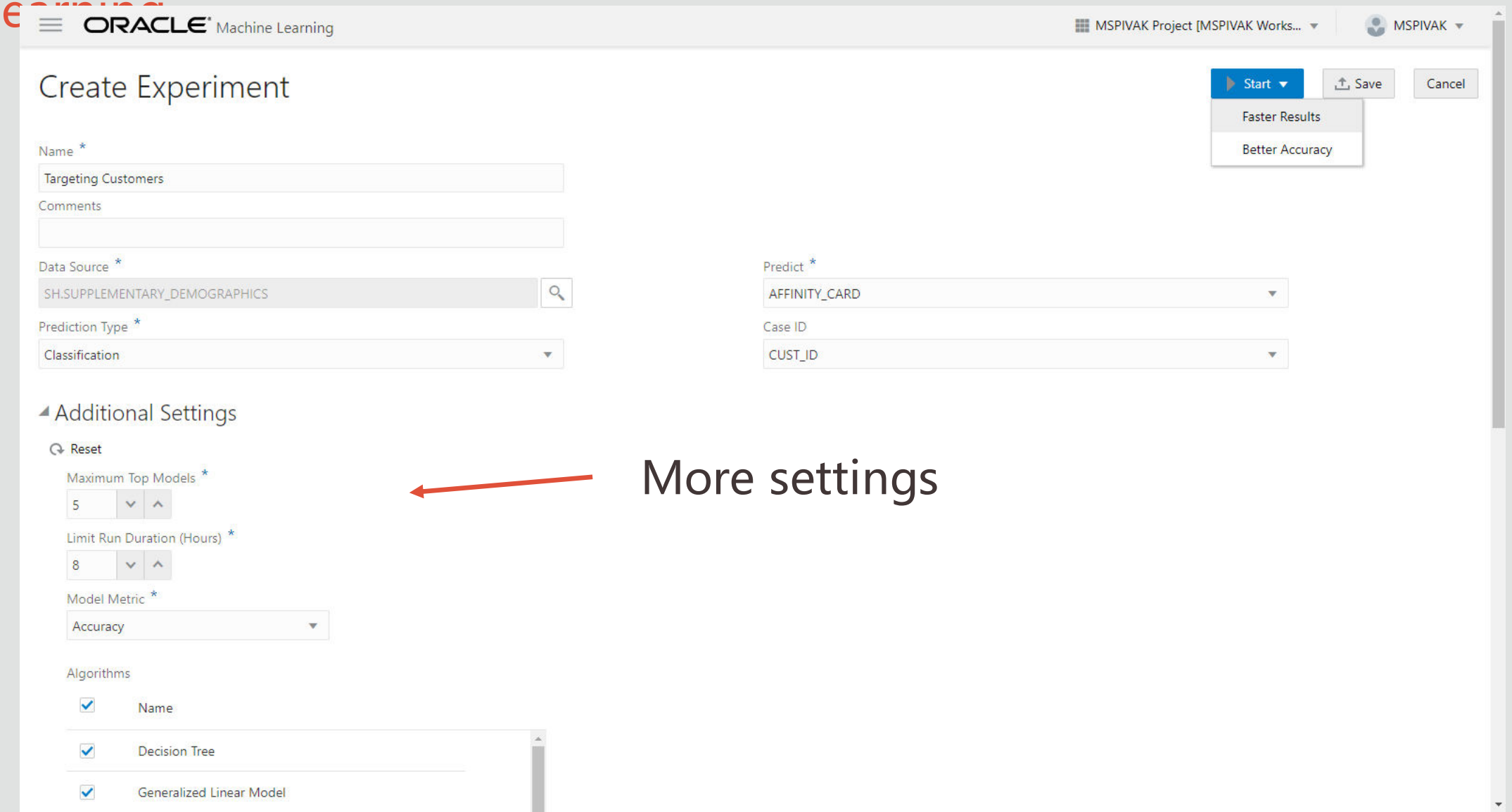
"Code-free" user interface supporting automated end-to-end machine learning



Select model metric

Coming Soon! | AutoML User Interface

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



More settings

Coming Soon! | AutoML User Interface

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

Leader board

Leader Board

Deploy | Create Notebook | Metrics

Algorithm	Name	Accuracy
Naive Bayes	nb_f8637b9359	0.4282
Random Forest	rf_aed7a58923	0.3624
Decision Tree	dt_6cd748831e	0.3624
Neural Network	nn_86b0b368e1	0.3529
Support Vector Machine (Linear)	svml_b3954f73b4	0.3224

Attribute importance

Features

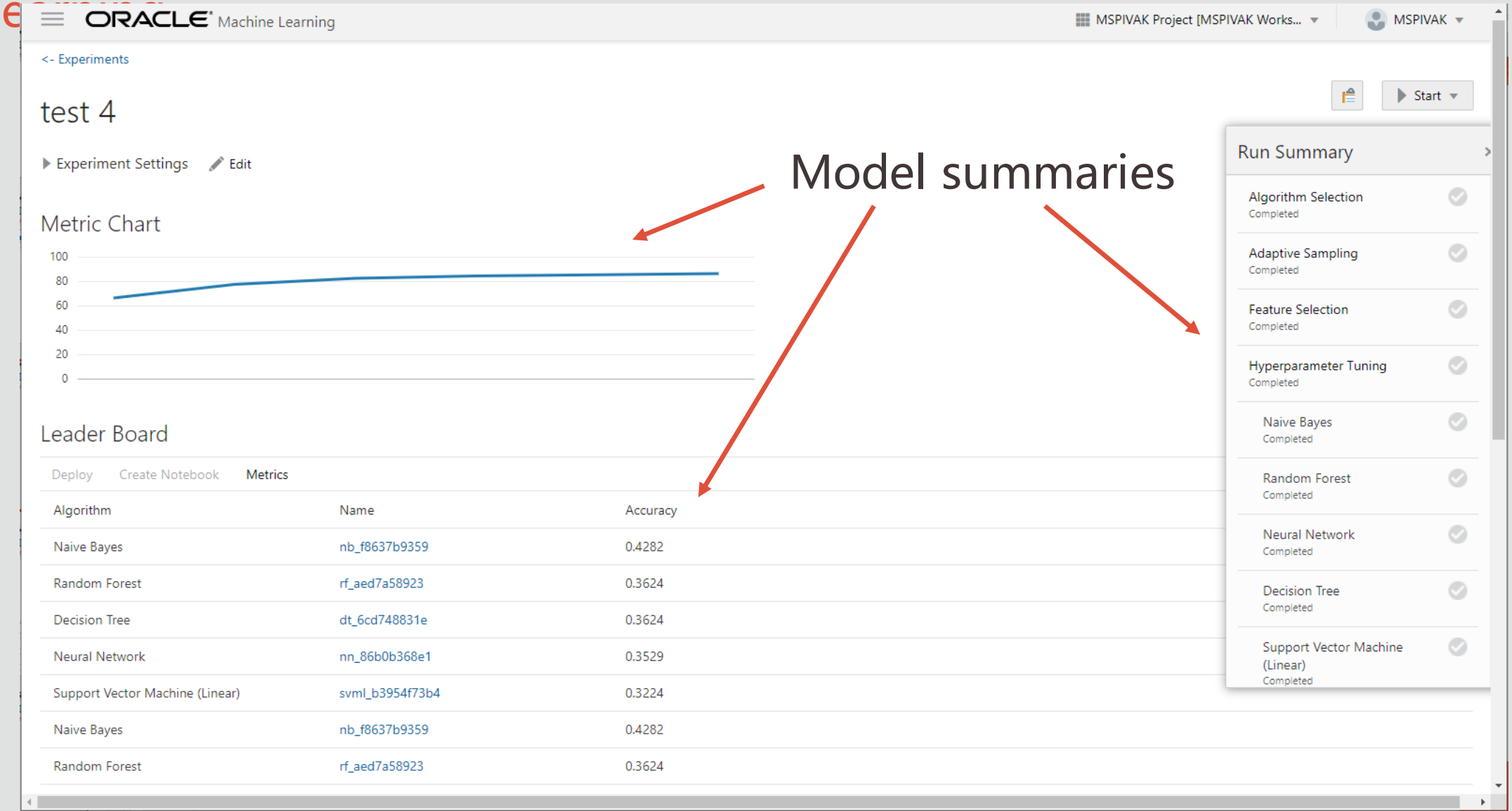
Refresh

Search...

Name	Importance	Type	Percent NULLs	Distinct Values	Min	Max	Mean	Std Dev
OCCUPATION		VARCHAR2	0	15				
YRS_RESIDENCE		NUMBER	0	15	0	14	4.02	1.97
BOOKKEEPING_APPLICATION		NUMBER	0	2	0	1	.89	.45
COMMENTS		VARCHAR2	4.772991850989523	43				
AFFINITY_CARD		NUMBER	0	2	0	1	.24	.6
HOME_THEATER_PACKAGE		NUMBER	0	2	0	1	.56	.7
Y_BOX_GAMES		NUMBER	0	2	0	1	.31	.66
FLAT_PANEL_MONITOR		NUMBER	0	2	0	1	.58	.7
BULK_PACK_DISKETTES		NUMBER	0	2	0	1	.64	.68

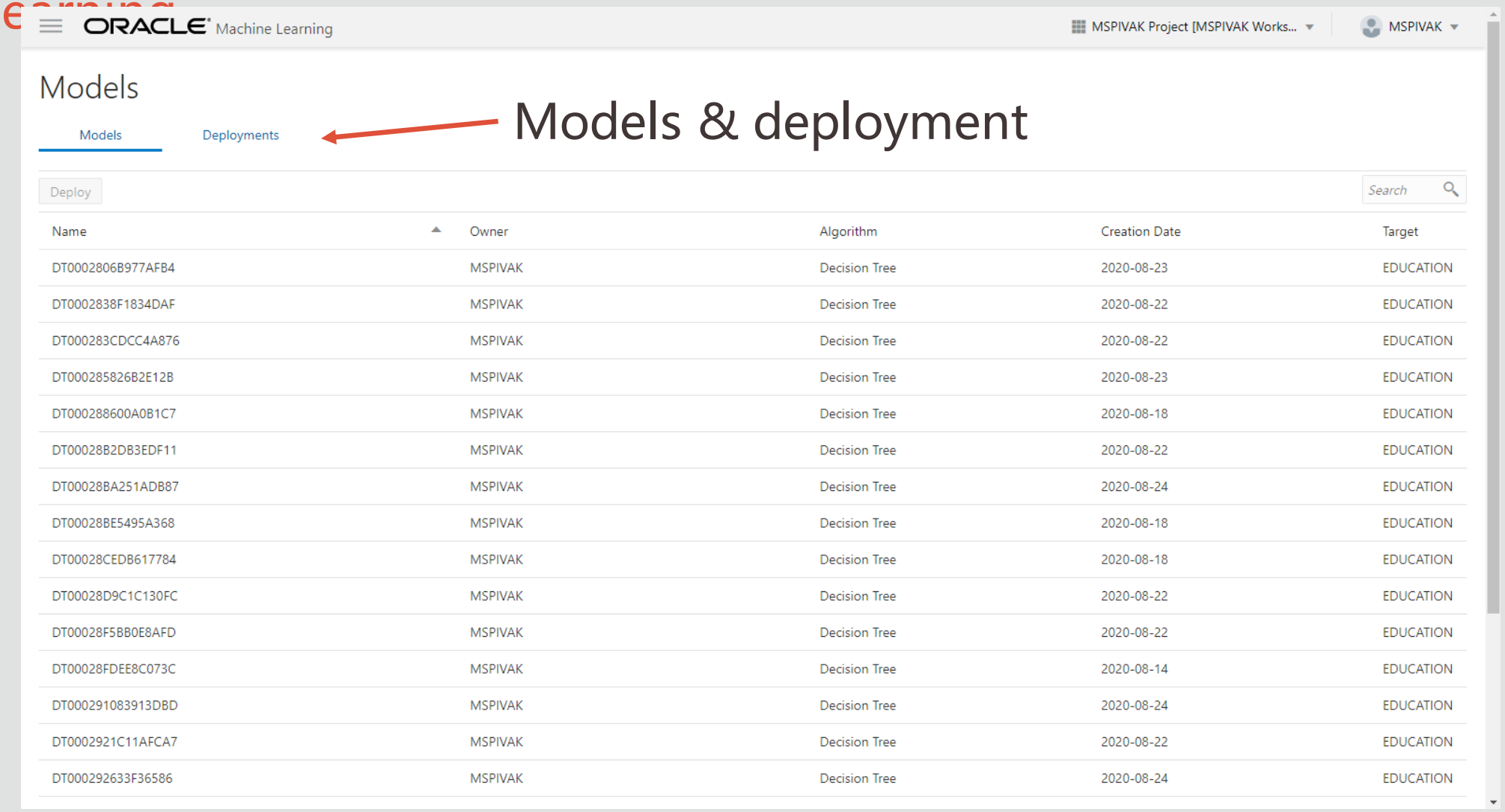
Coming Soon! | AutoML User Interface

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



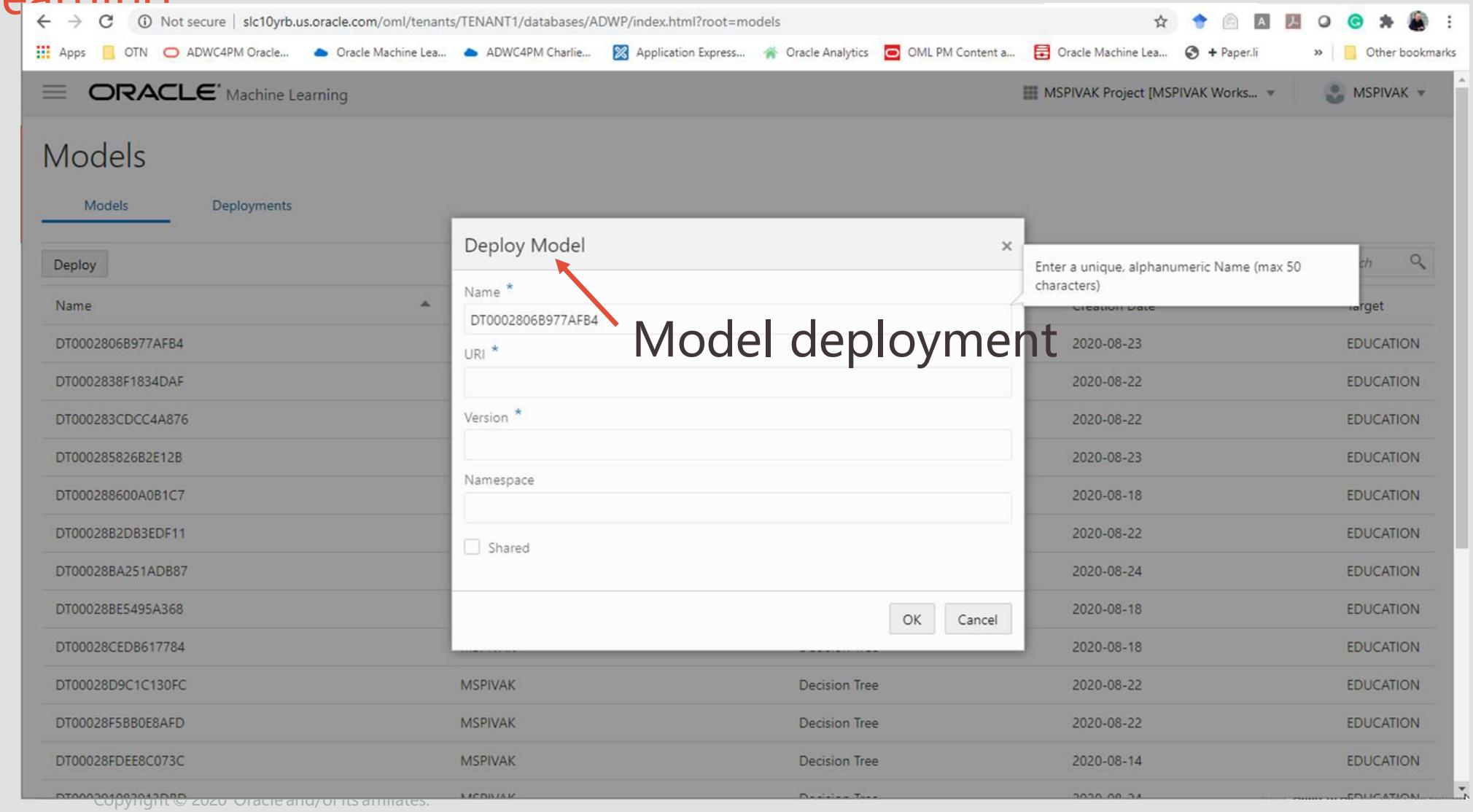
Coming Soon! | AutoML User Interface

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



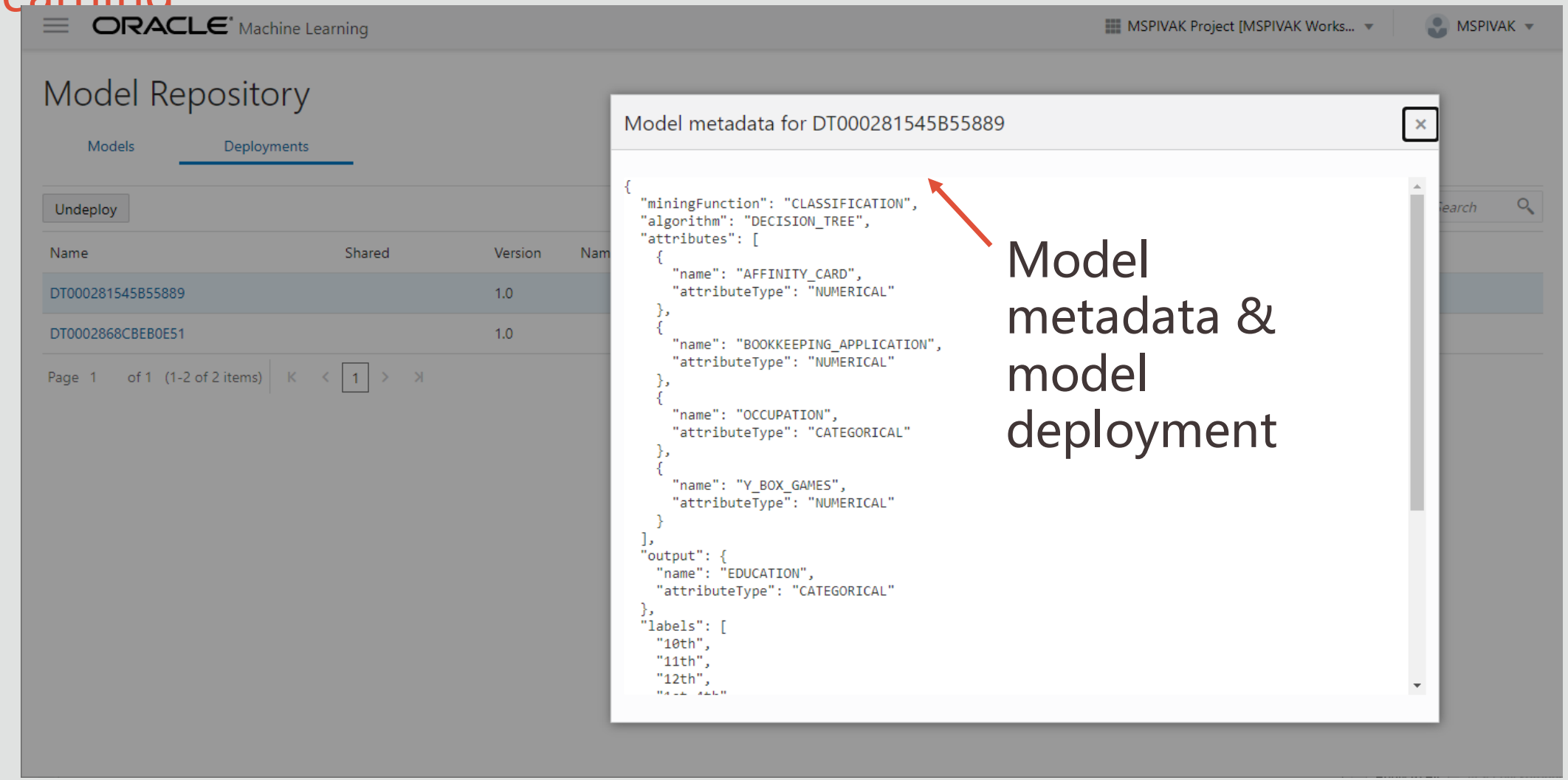
Coming Soon! | AutoML User Interface

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



Coming Soon! | AutoML User Interface

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



Coming Soon! | OML Services

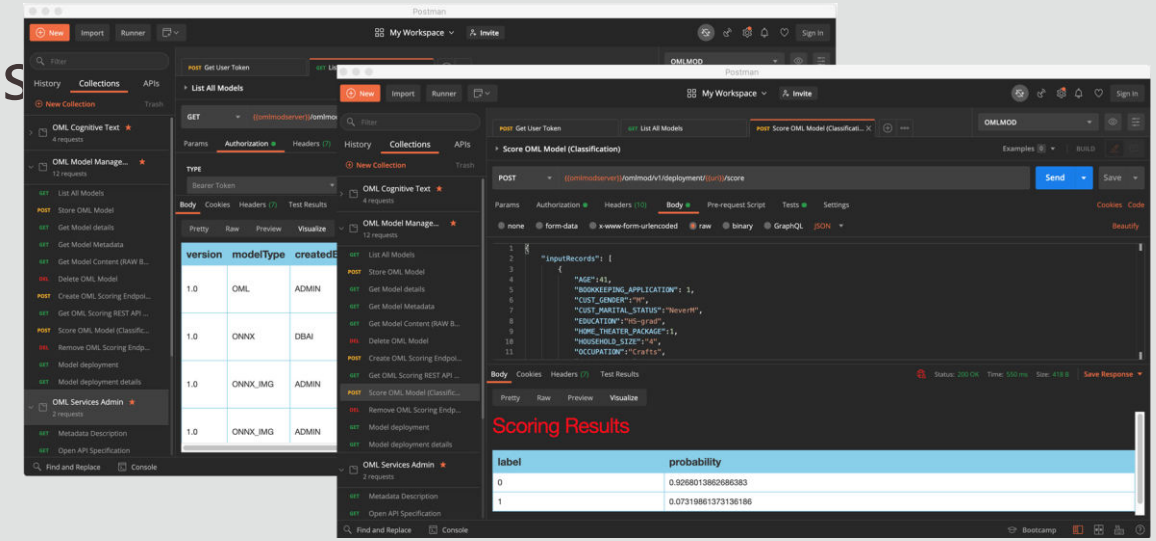
Supports model deployment and lifecycle management beyond Database

Support model deployment and model lifecycle management for both in-database OML and third-party classification for regression models via REST APIs

Supports ONNX (Open Neural Networks eXchange) format including Scikit-learn and TensorFlow and others

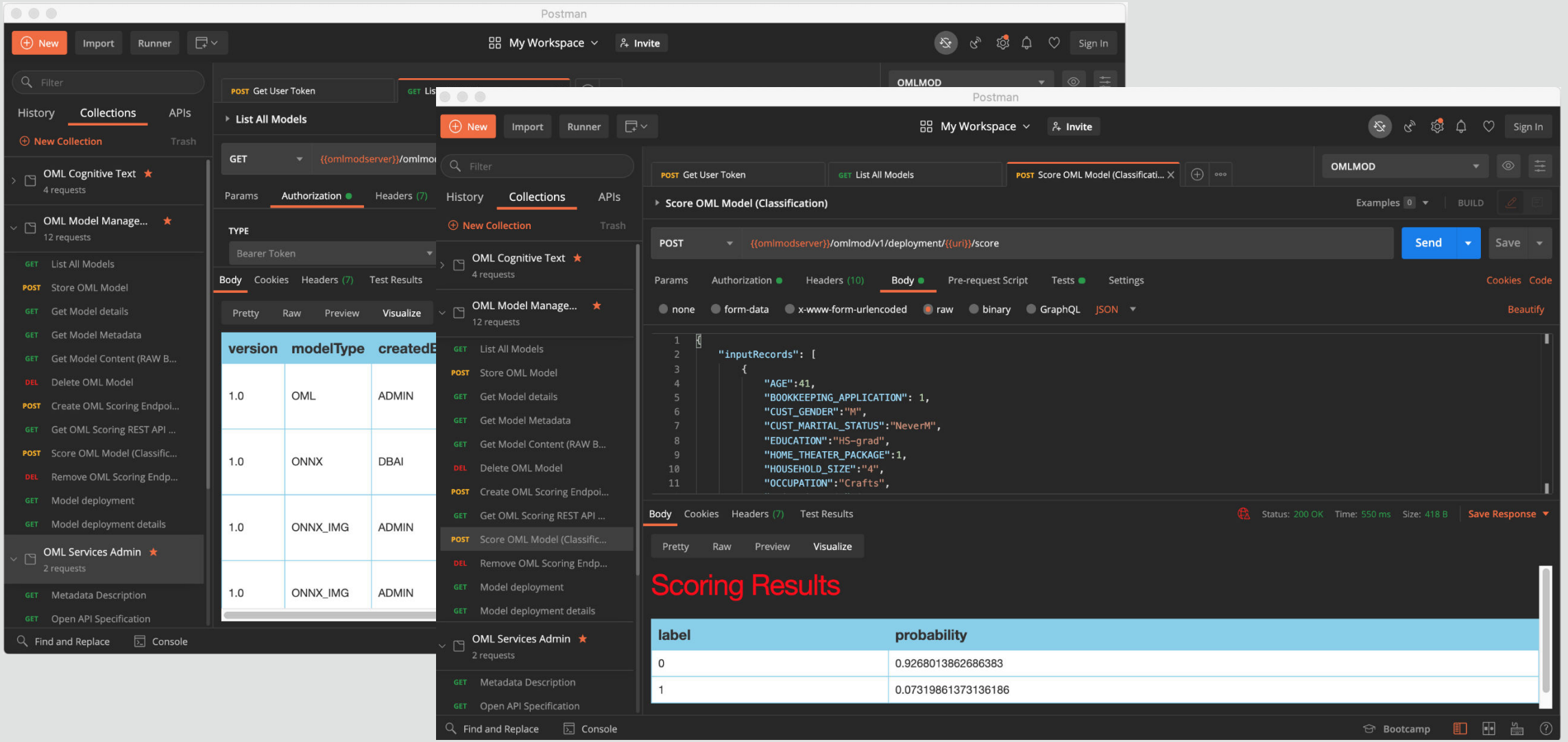
Supports OML cognitive text capabilities, with capabilities for topic discovery, keywords, summary, sentiment, and feature extraction (English, Spanish, and French initially).

Cognitive image supports ONNX format third-party model scoring for images or tensors



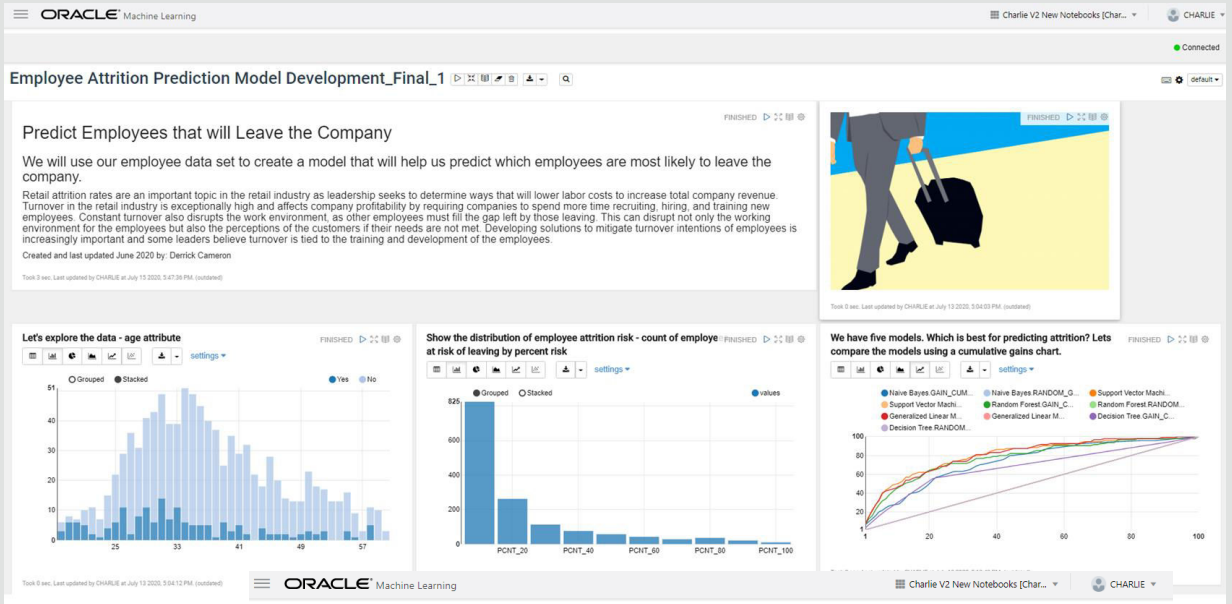
Coming Soon! | OML Services

Supports model deployment and lifecycle management beyond Database

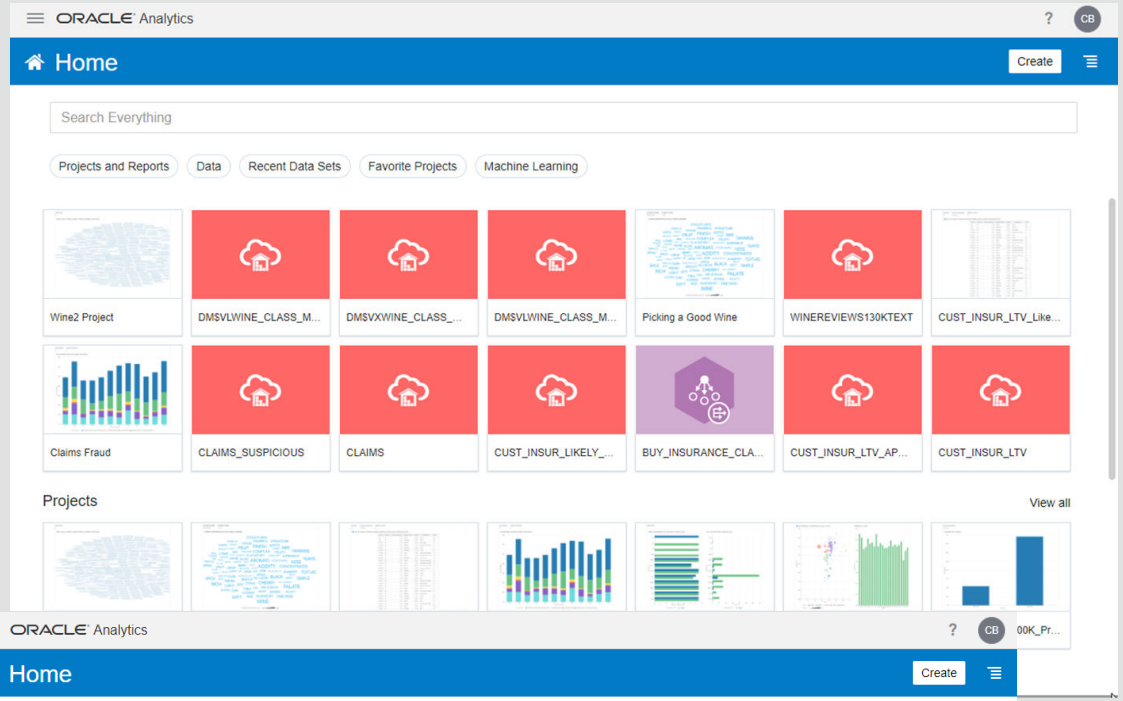


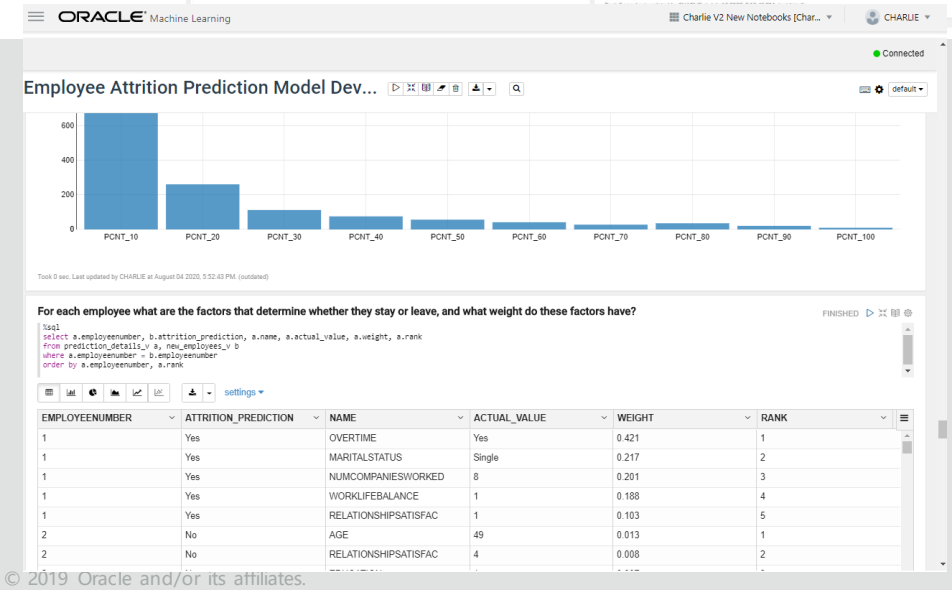
Oracle Analytics Cloud 5.7 + OML = ~“Solutions”

Oracle Machine Learning + Oracle Analytics Cloud

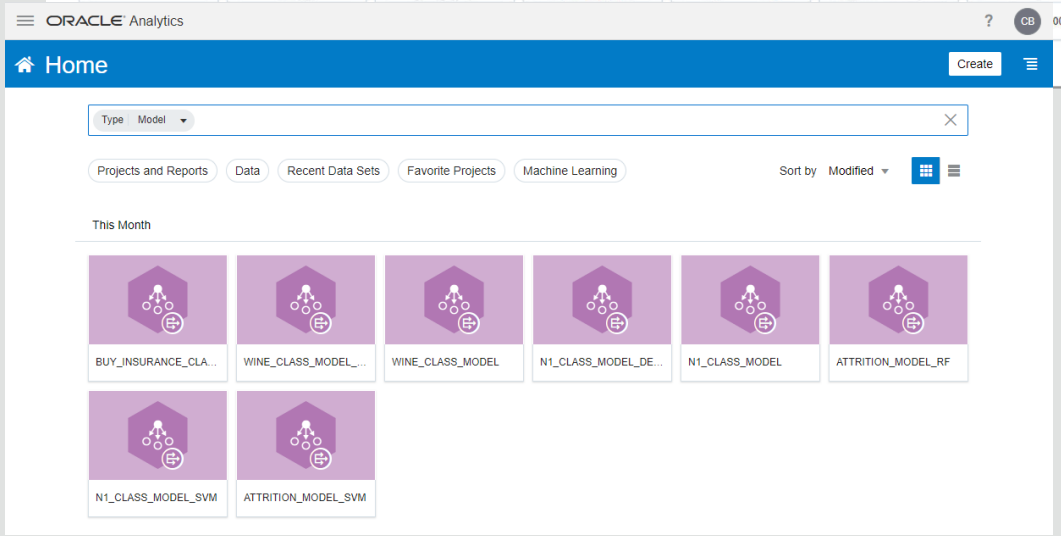


+

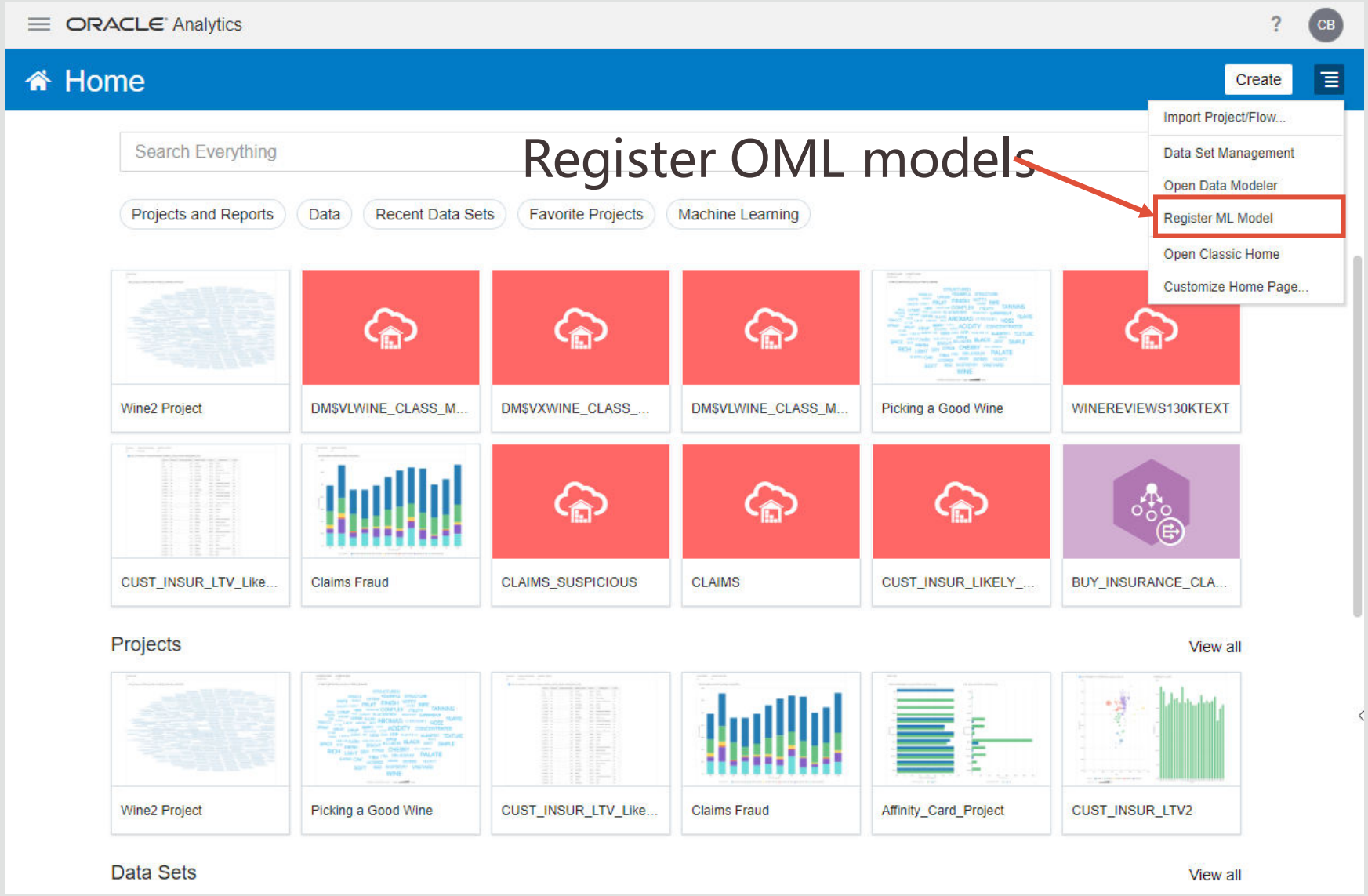




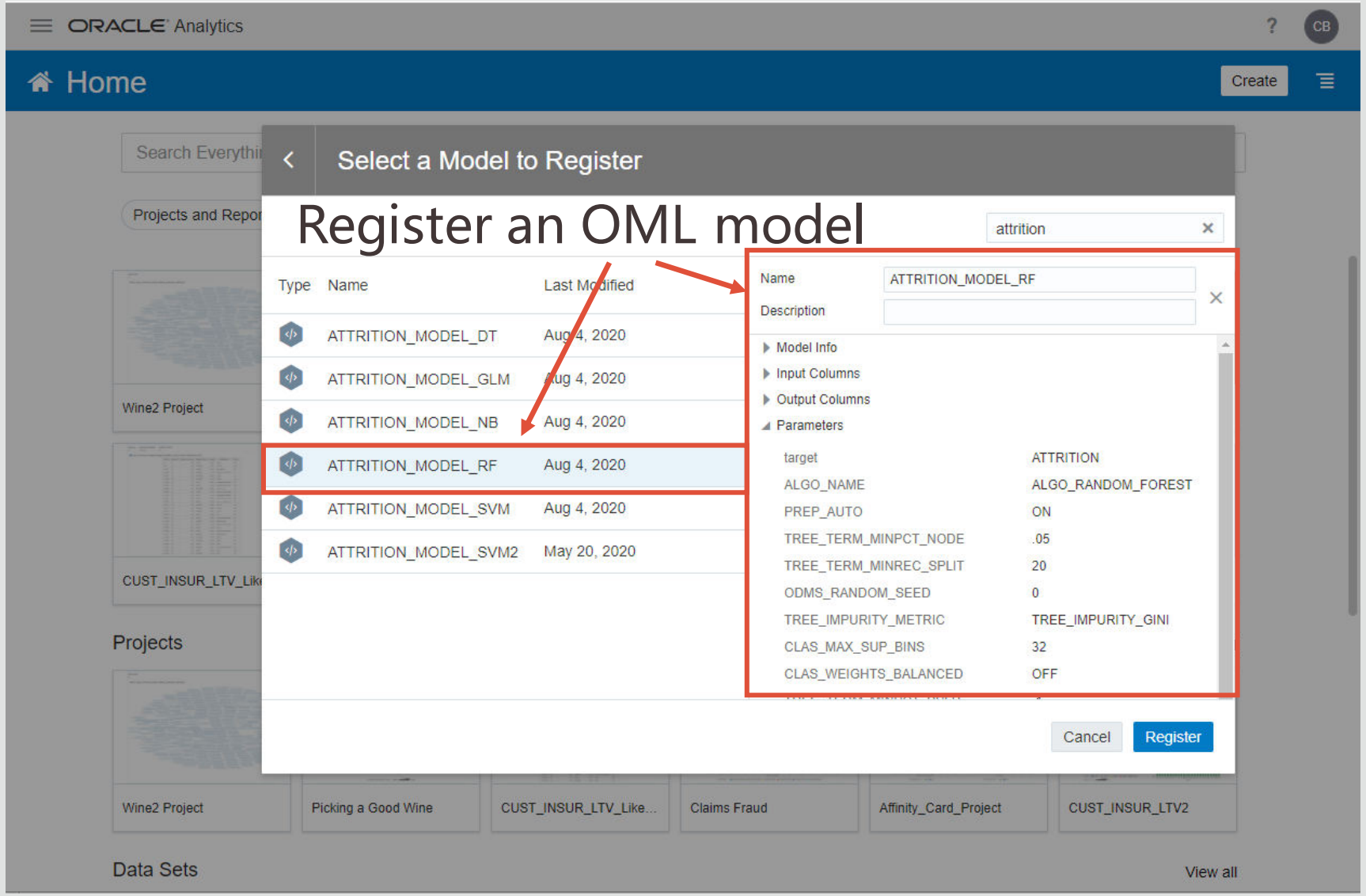
EMPLOYEENUMBER	ATTRITION_PREDICTION	NAME	ACTUAL_VALUE	WEIGHT	RANK
1	Yes	OVERTIME	Yes	0.421	1
1	Yes	MARITALSTATUS	Single	0.217	2
1	Yes	NUMCOMPANIESWORKED	8	0.201	3
1	Yes	WORKLIFEBALANCE	1	0.188	4
1	Yes	RELATIONSHIPSATISFAC	1	0.103	5
2	No	AGE	49	0.013	1
2	No	RELATIONSHIPSATISFAC	4	0.008	2



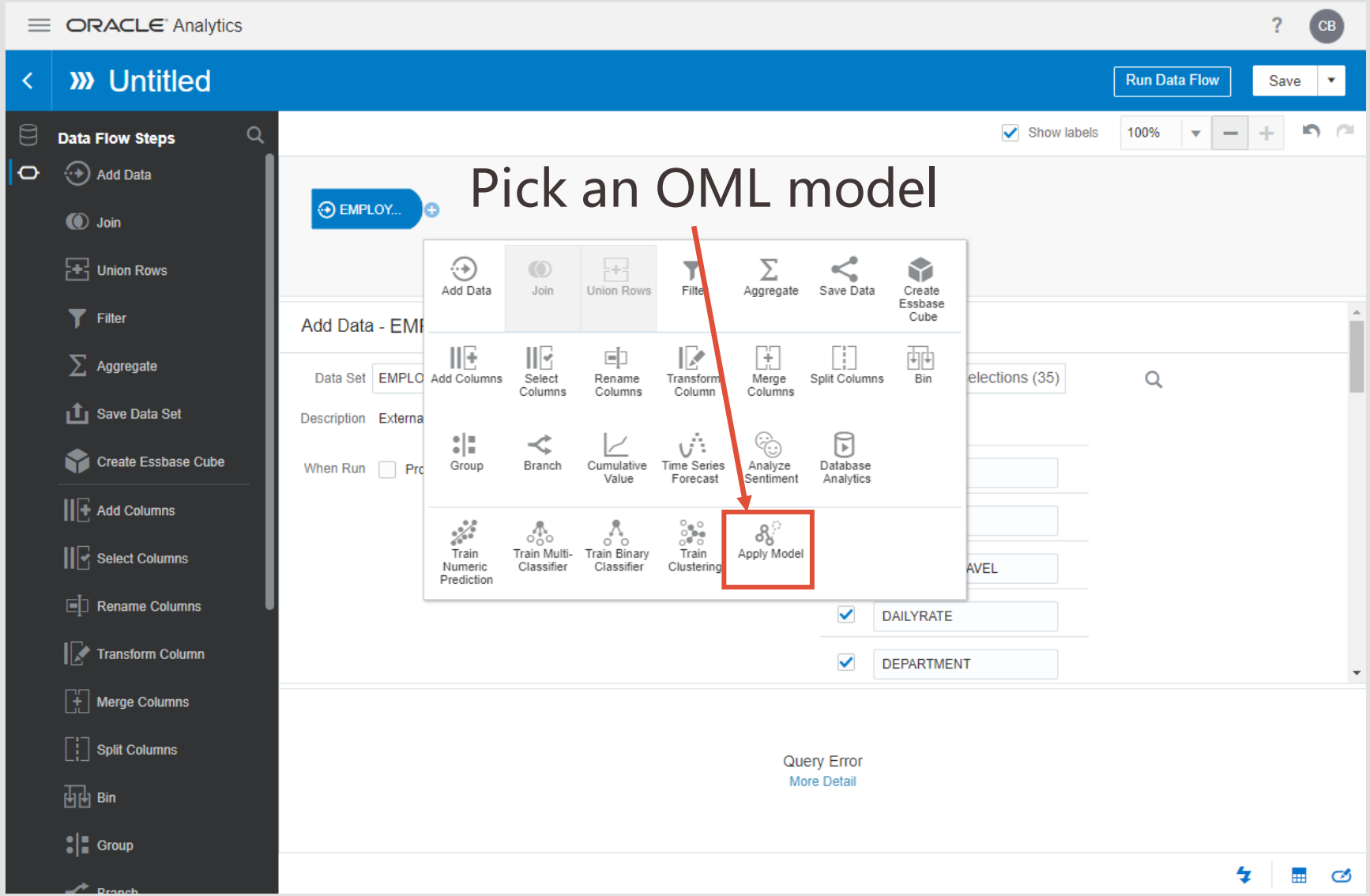
Register an OML Model in OAC



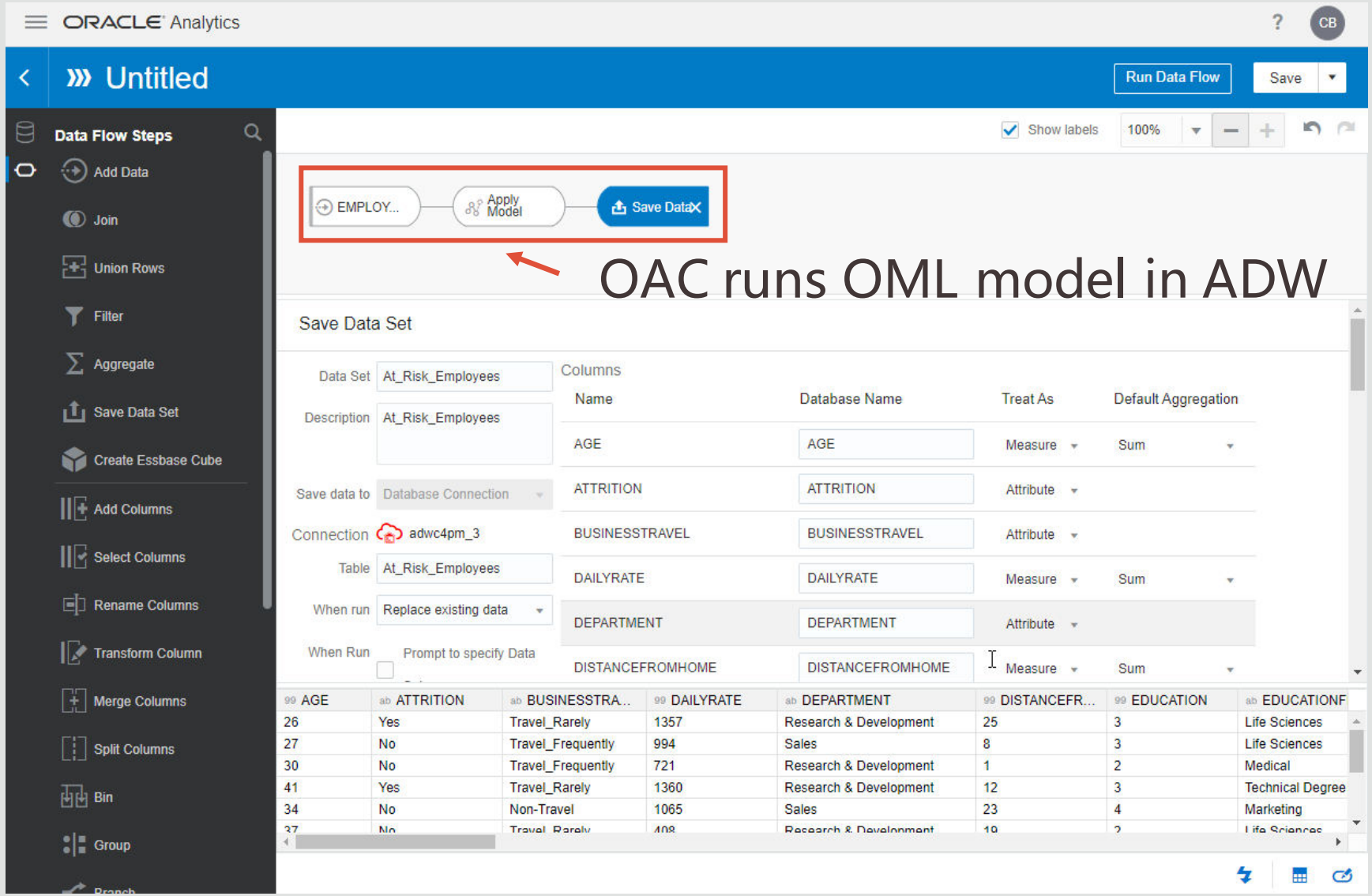
Register an OML Model in OAC



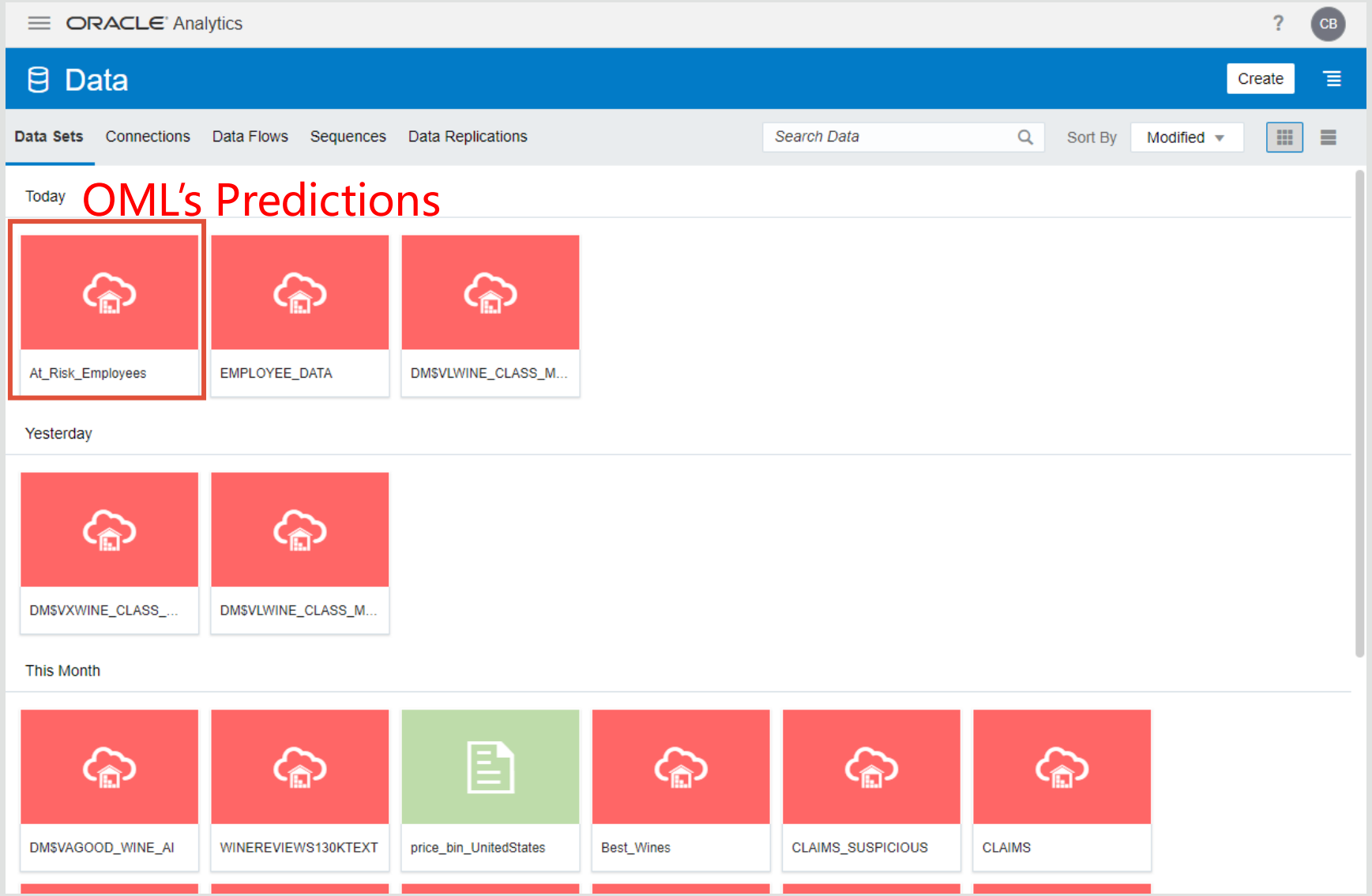
Define a Data Flow using OML in-DB Model



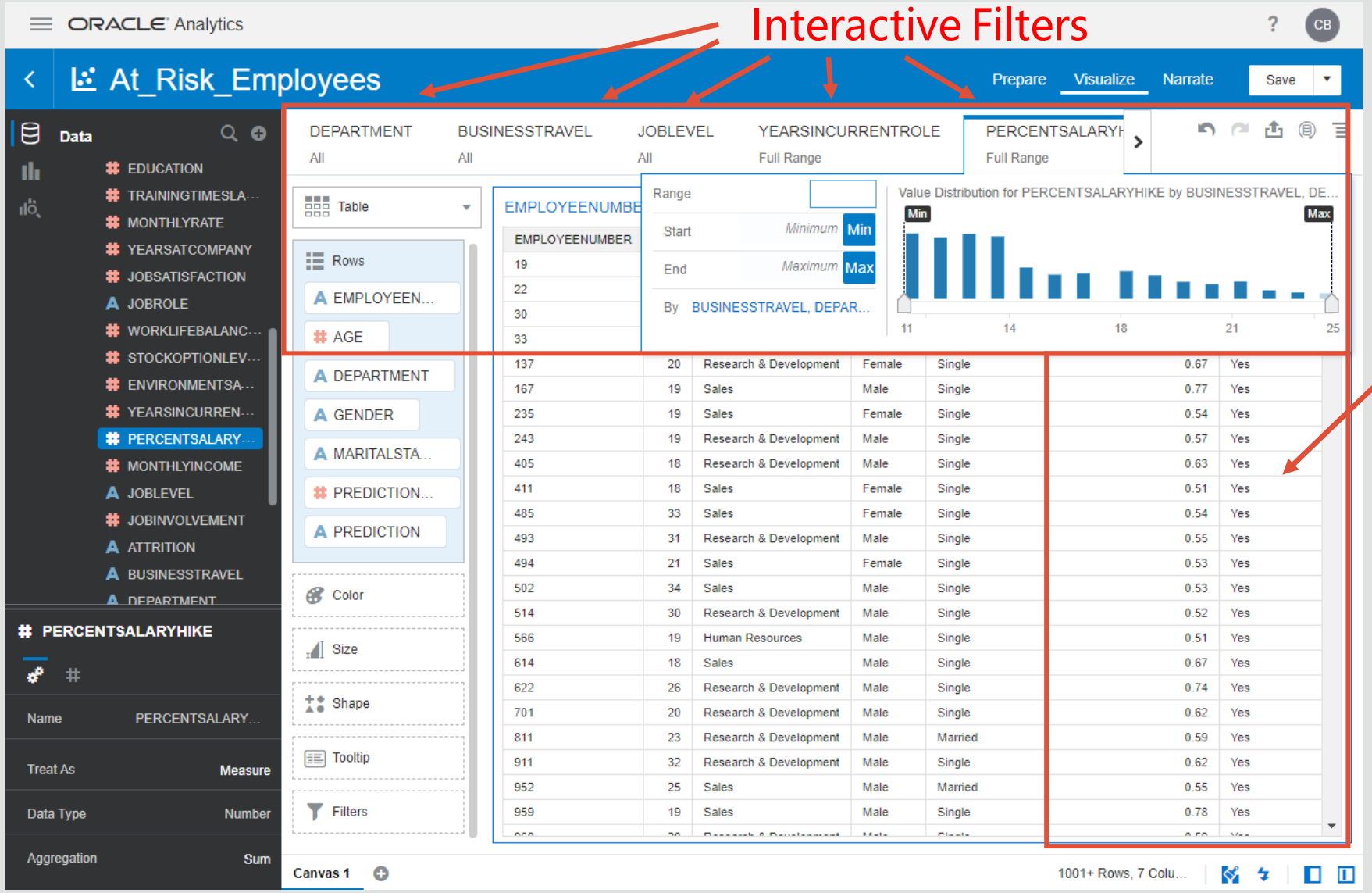
Define a Data Flow using OML in-DB Model



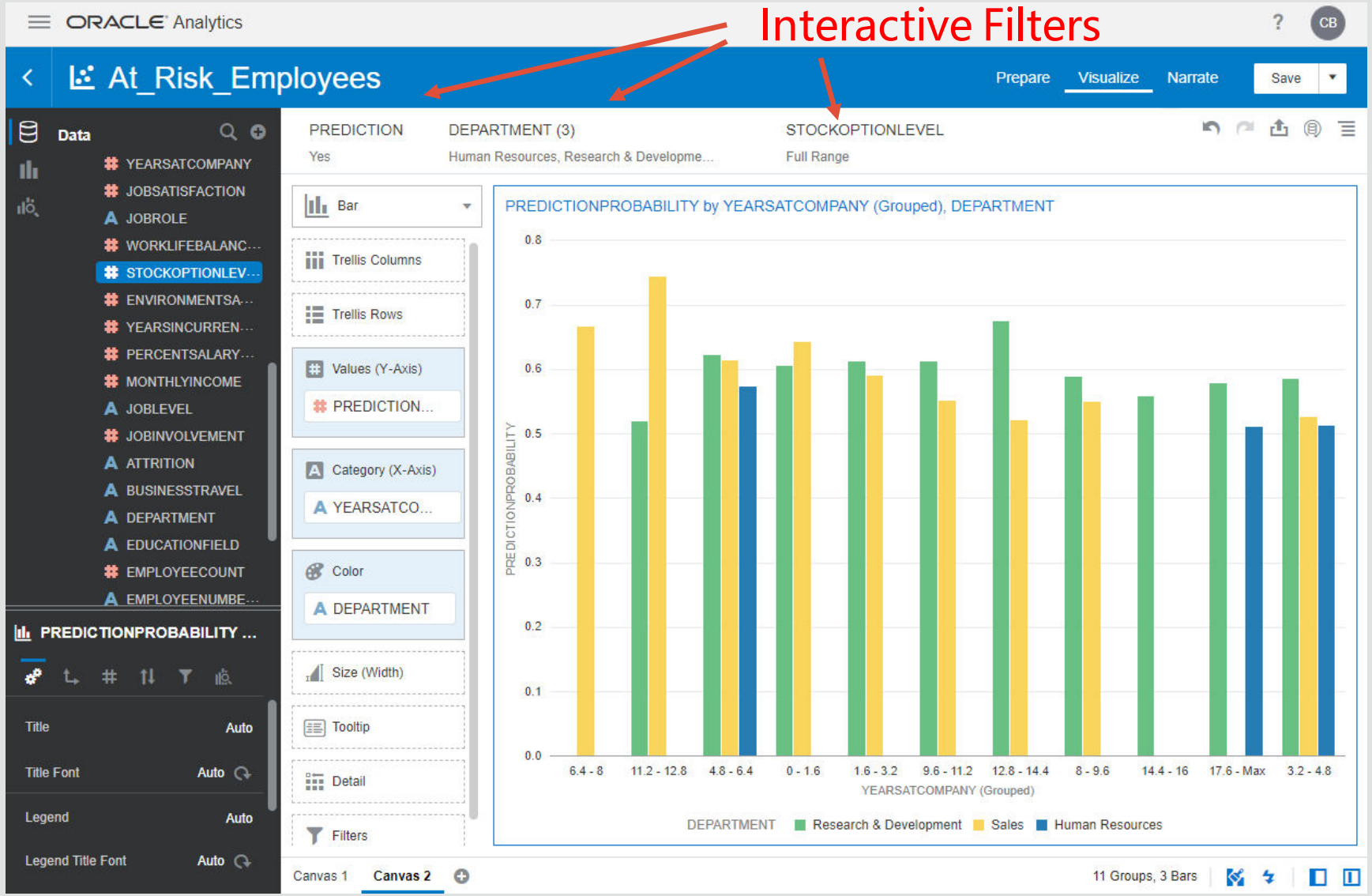
Investigate OML Insights and Predictions



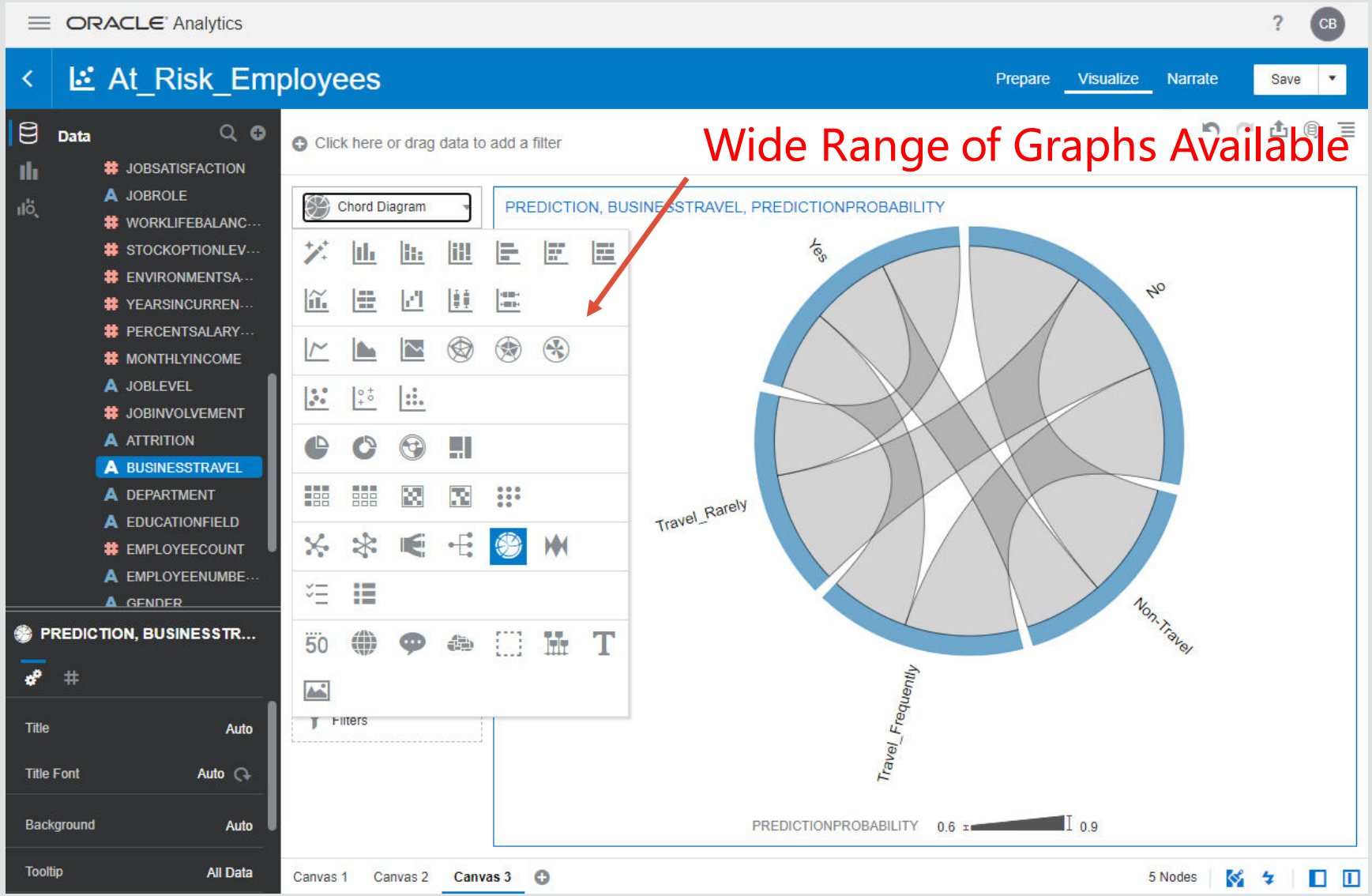
Investigate OML Insights and Predictions



Investigate OML Insights and Predictions



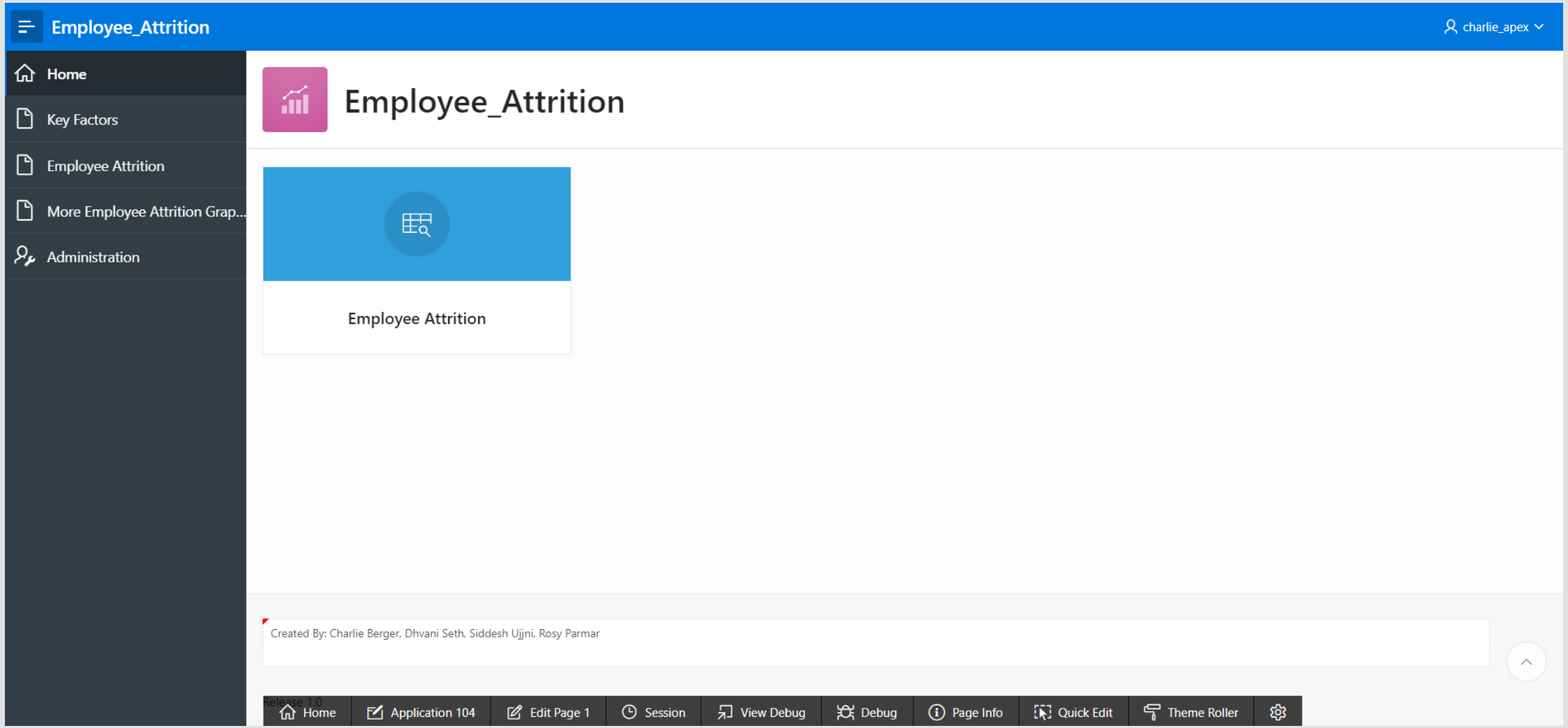
Investigate OML Insights and Predictions



Wait, there is more!

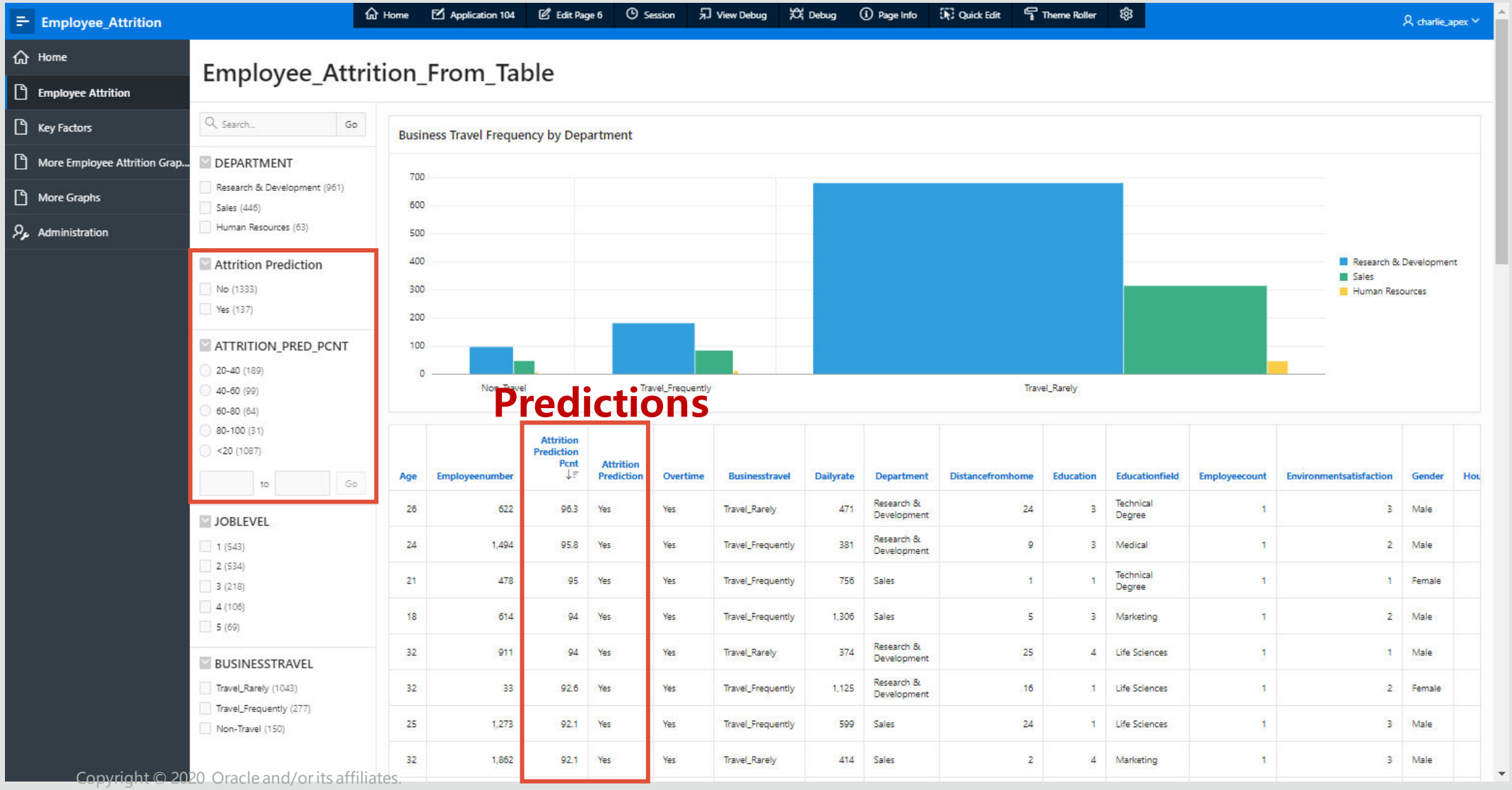
OML + APEX

Interactively Explore Data and OML Insights and Predictions



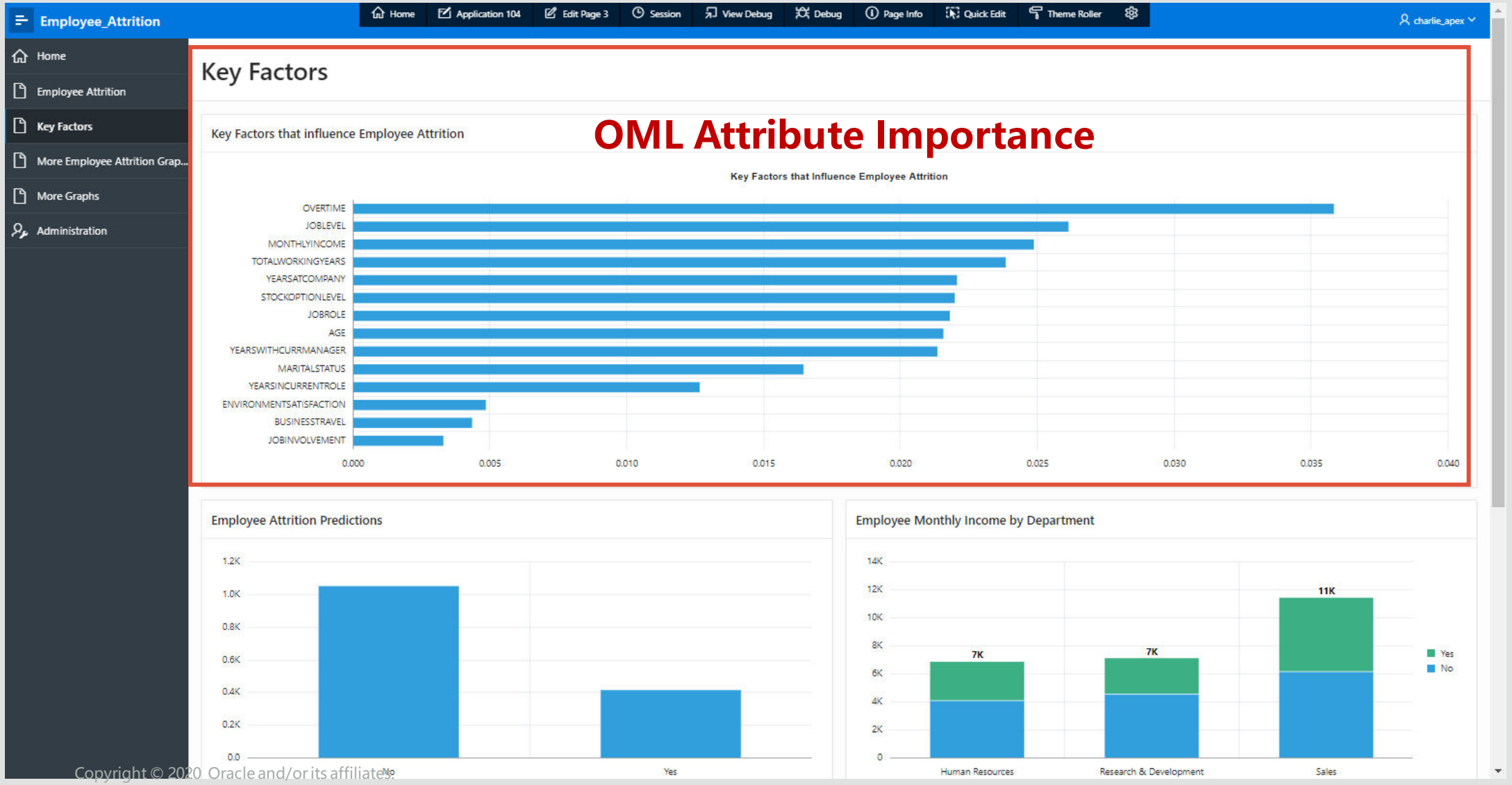
OML + APEX

Interactively Explore Data and OML Insights and Predictions



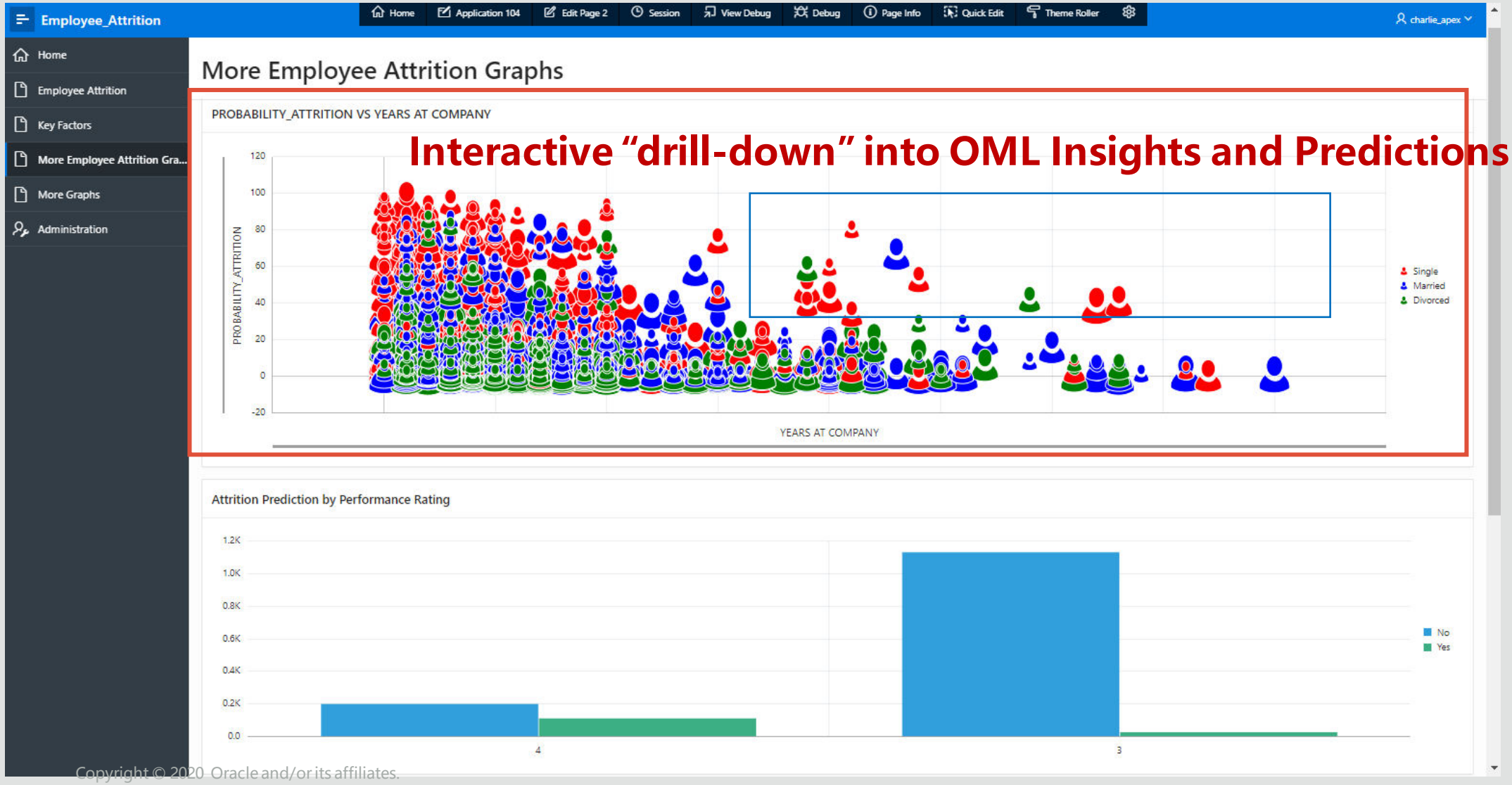
OML + APEX

Interactively Explore Data and OML Insights and Predictions



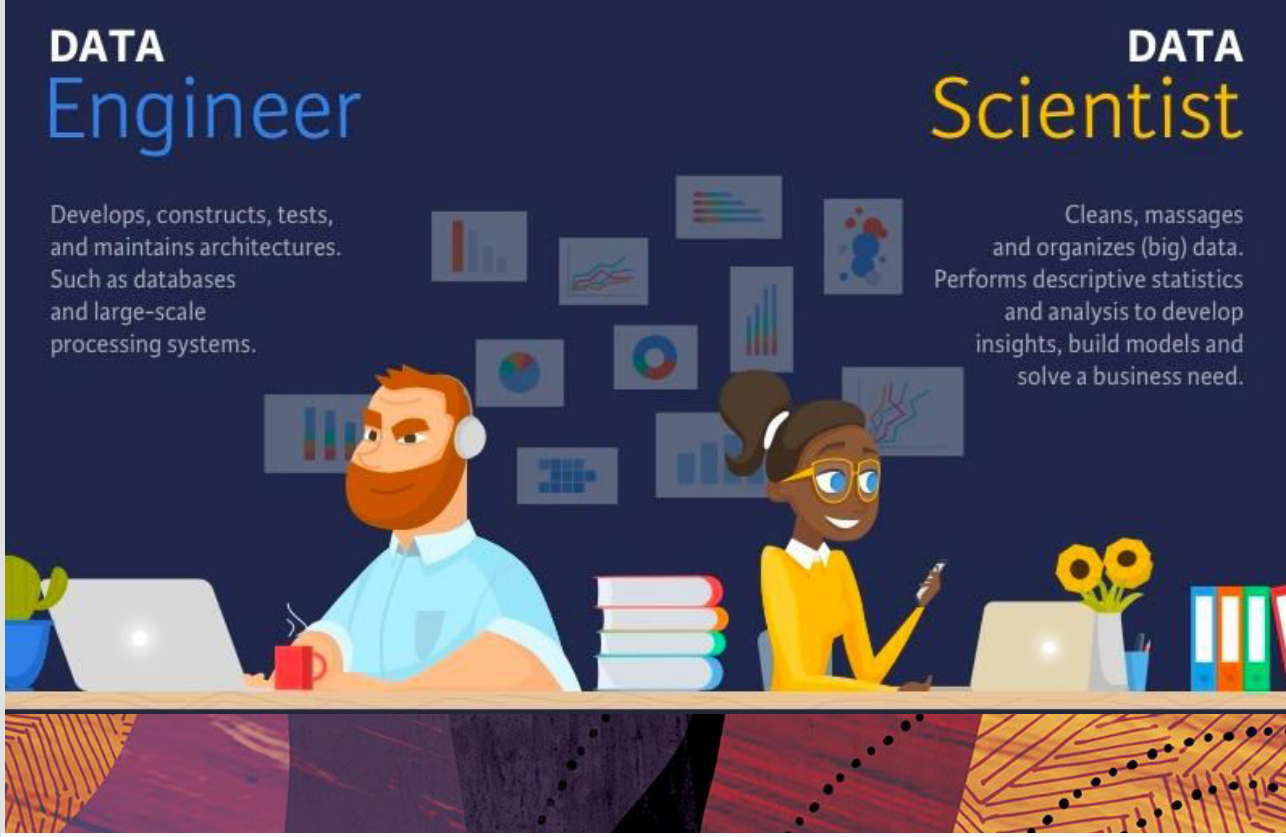
OML + APEX

Interactively Explore Data and OML Insights and Predictions



The Changing Role of the DBA:

From Oracle Data Professional to Oracle Data Scientist
in 6 Weeks!



https://www.datacamp.com/community/blog/data-scientist-vs-data-engineer
https://www.kdnuggets.com/2020/02/poll-automl-replace-data-scientists.html

Data Professional to Data Scientist Journey

You are Likely Already Doing Much of The Work!

Data extraction

Data wrangling

Deriving new attributes ("feature engineering")

...

...

Where the Machine Learning "Magic" Happens

Import predictions & insights

Translate and deploy ML models Oracle

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 w/

Data Management platform becomes combined/hybrid data management + machine learning platform

[1] - https://www.infoworld.com/article/3228245/data-science/the-80-20-data-science-dilemma.html

CRISP-DM Methodology

Six Major Steps

BUSINESS UNDERSTANDING

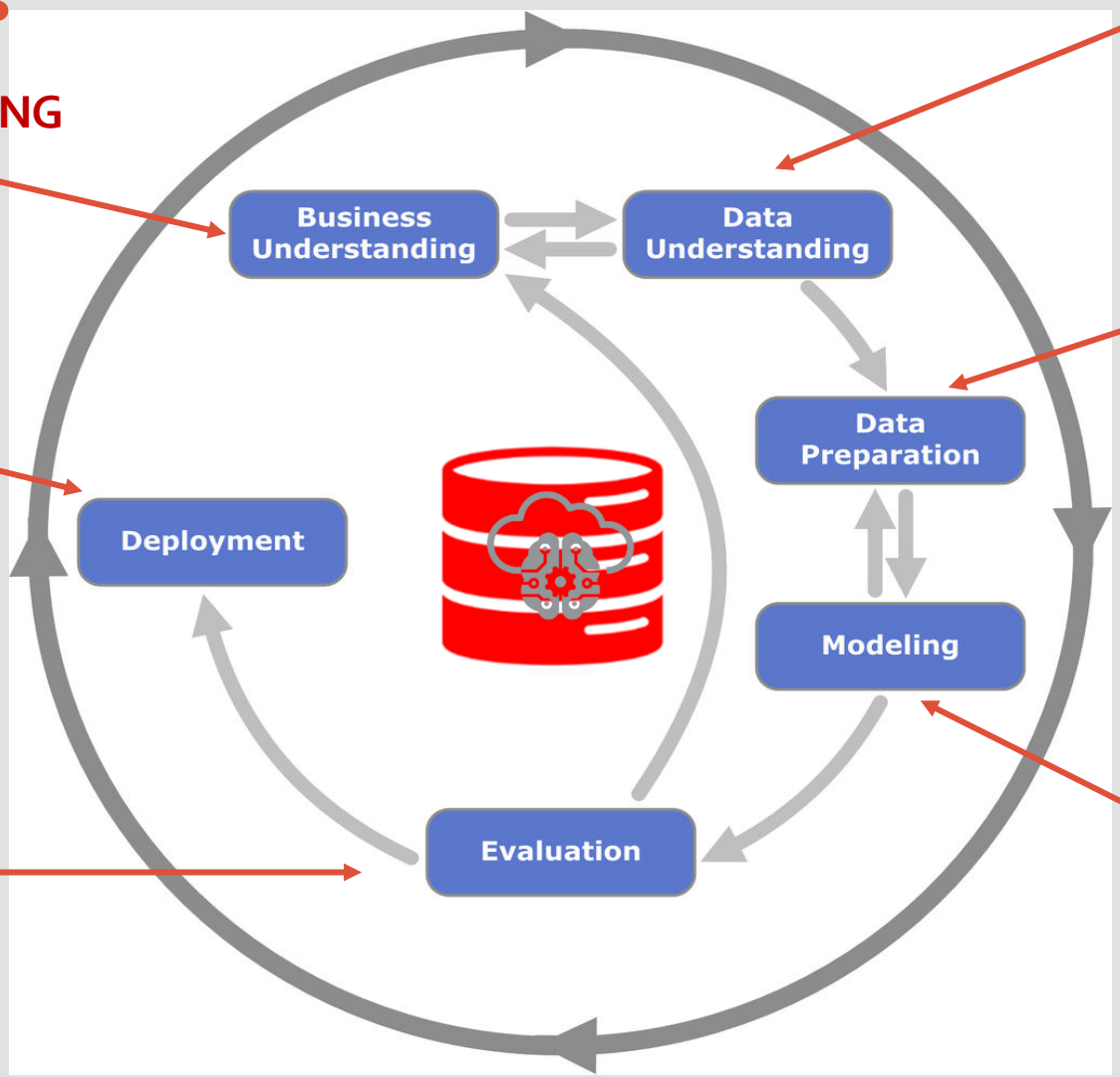
DATA UNDERSTANDING

DATA PREPARATION

DEPLOYMENT

MODELING

EVALUATION



https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

CRISP-DM Methodology

Six Major Steps

BUSINESS UNDERSTANDING

Well-defined
business problem

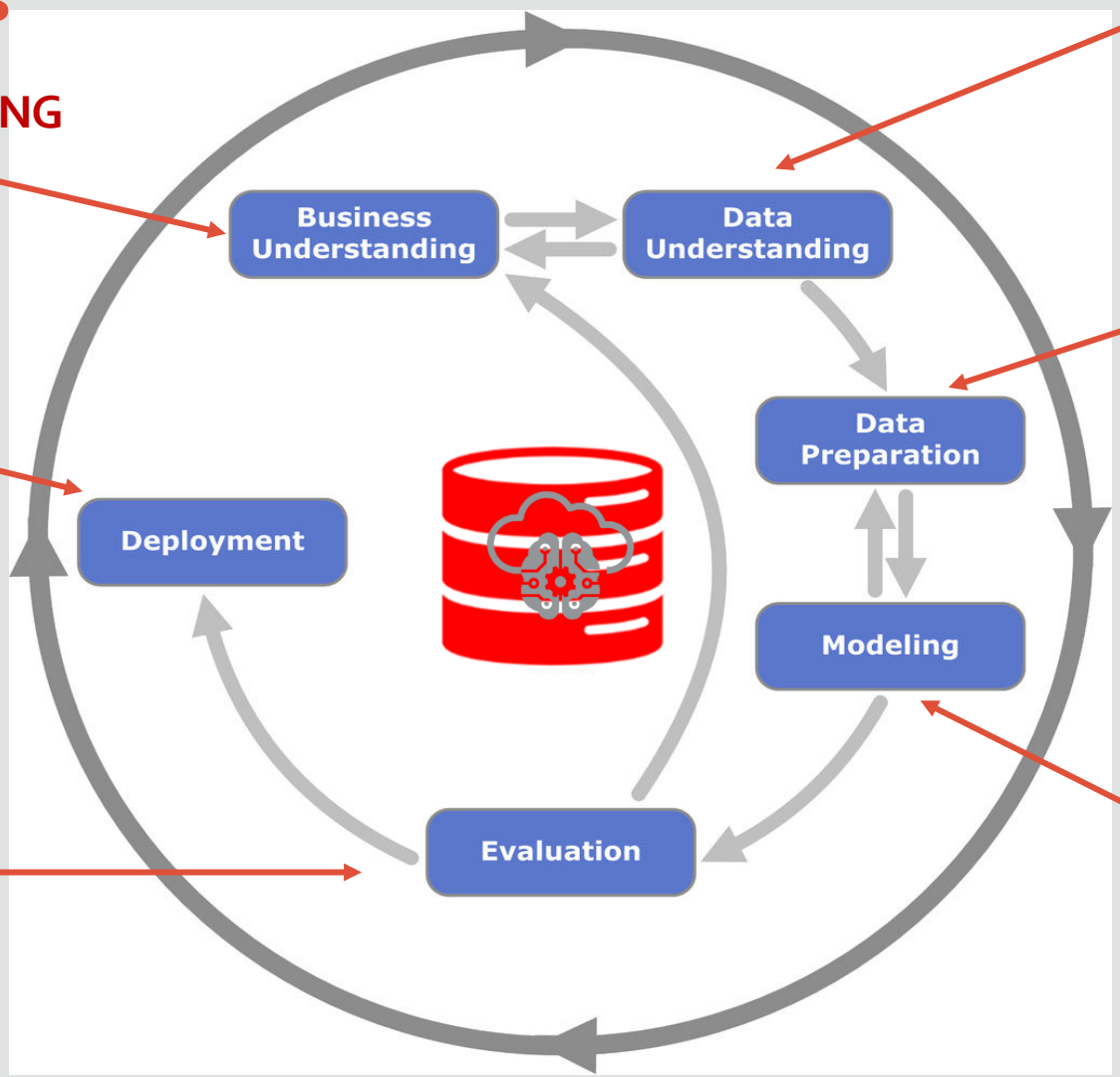
DEPLOYMENT

In-DB ML model apply
- Real-time ML apply
- In-database, REST

Embed methodology
- Applications
- Dashboards

EVALUATION

Model evaluation
Model comparison
Model selection



DATA UNDERSTANDING

Assemble the "right data"
Data profiling
- Data visualization
- Univariate statistics/group by
- Bi-variate statistics

DATA PREPARATION

Sampling/Stratified
Algorithm req'd transforms
- Auto Data Preparation
- Missing Values, Binning, Normalization, etc.
- Unstructured data
- Aggregations

Domain specific transforms
- "Engineered Features"

Features Selection

MODELING

Algorithm settings/defaults
- Stratified sampling
- Feature selection
- Build model(s)

https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

CRISP-DM Methodology

Six Major Steps

BUSINESS UNDERSTANDING

Well-defined
business problem

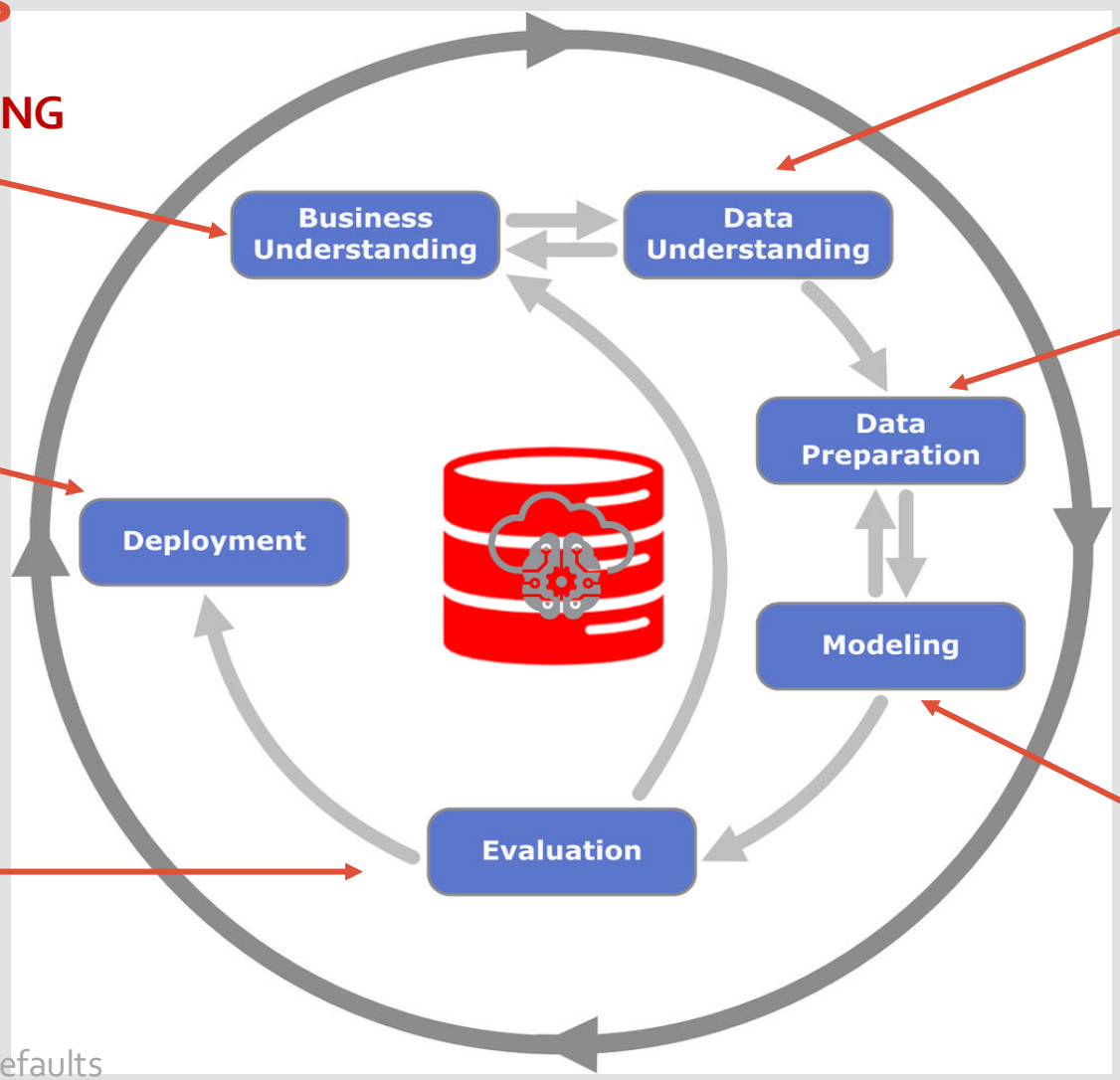
DEPLOYMENT

In-DB ML model apply
- Real-time ML apply
- In-database, REST

Embed methodology
- Applications
- Dashboards

EVALUATION

Model evaluation
Model comparison
Model selection



DATA UNDERSTANDING

Assemble the "right data"
Data profiling
- Data visualization
- Univariate statistics/group by
- Bi-variate statistics

DATA PREPARATION

Sampling/Stratified
Algorithm req'd transforms
- Auto Data Preparation
- Missing Values, Binning, Normalization, etc.
- Unstructured data
- Aggregations

Domain specific transforms
- "Engineered Features"

Features Selection

MODELING

Algorithm settings/defaults
- Stratified sampling
- Feature selection
- Build model(s)

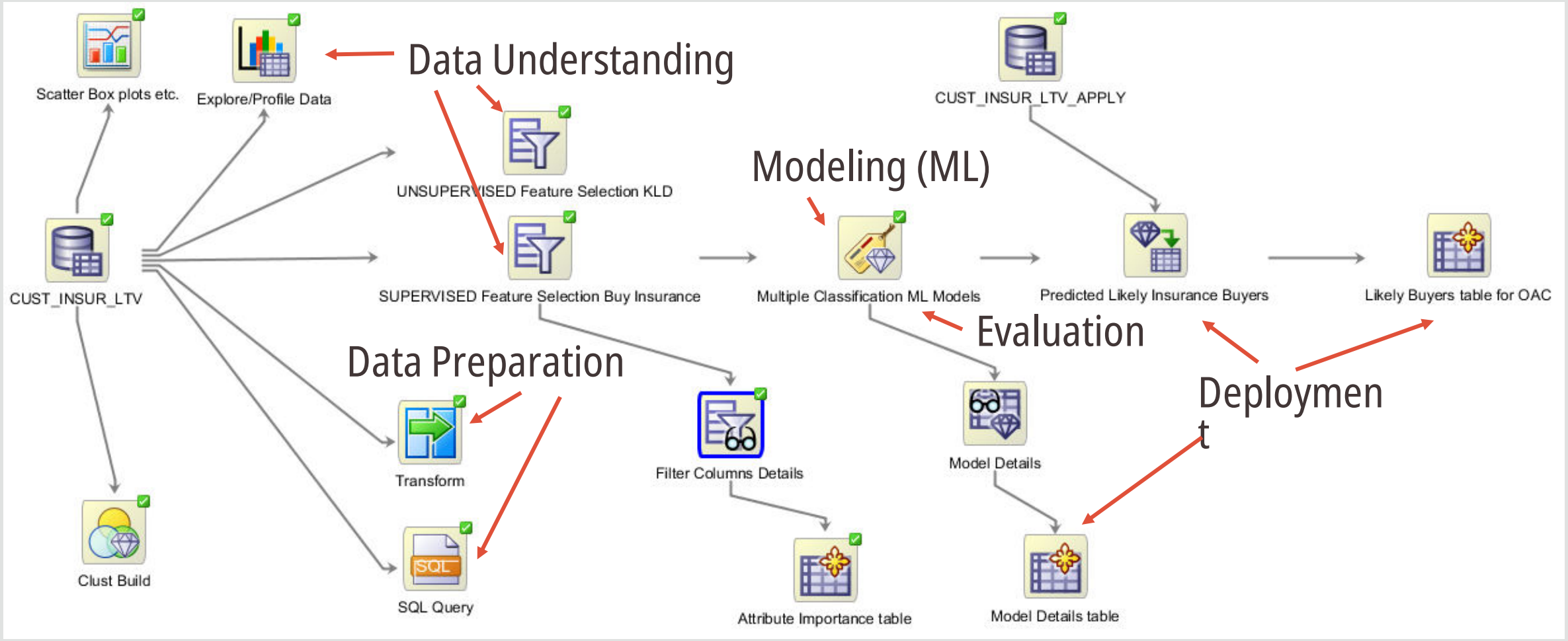
* Automated and/or system defaults

https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Oracle Machine Learning

SQL Developer Extension: Oracle Data Miner UI

Business Understanding: Target customers most likely to Buy Insurance



File Edit View History Bookmarks Tools Help

Oracle Machine Learning | Exan

https://adb.us-ashburn-1.oraclecloud.com/oml/tenants/OCID1.TENANCY.OC1 90% Search

Most Visited Oracle Data Mining 17... Oracle Cloud Infrastru... Autonomous Data Wa... Oracle Machine Learni...

ORACLE Machine Learning

USER07 Project [USER07 Workspa... USER07

Example Templates

Create Notebook

Search by...

Anomaly Detection
This notebook shows how to detect...
Author:
Date Added: 2/13/18 11:16 PM
Tags: 'Anomaly Detection' 'Machine...
★ 5 Likes 1601 76

Association Rules
Notebook to show the use of Assoc...
Author:
Date Added: 2/13/18 11:16 PM
Tags: 'SQL' 'Associations' 'Rules' 'M...
★ 2 Likes 772 106

Attribute Importance
Notebook to identify key attributes...
Author:
Date Added: 2/13/18 11:16 PM
Tags: 'SQL' 'Attribute Importance' 'K...
★ 2 Likes 530 29

Classification Prediction M...
Example notebook to predict custo...
Author:
Date Added: 2/13/18 11:16 PM
Tags: 'Classification' 'Prediction' 'De...
★ 3 Likes 1024 97

Clustering
This notebook shows how to identi...
Author:
Date Added: 2/13/18 11:16 PM
Tags: 'Clustering' 'K-Means' 'Expect...
★ 1 Likes 598 35

My First Notebook
Oracle Machine Learning example ...
Author:
Date Added: 2/13/18 11:16 PM
Tags: 'SQL' 'Data' 'Graph'
★ 4 Likes 914 27

Regression
This notebook shows how to predic...
Author:
Date Added: 2/13/18 11:16 PM
Tags: 'Regression' 'SVM' 'GLM' 'Logi...
★ 1 Likes 838 31

Statistical Function
Oracle Machine Learning example ...
Author:
Date Added: 2/13/18 11:16 PM
Tags: 'Statistics' 'ANOVA' 'T-test' 'F-...
★ 2 Likes 356 11

Time Series Forecasting
Oracle Machine Learning supports ...
Author:
Date Added: 9/5/19 4:14 AM
Tags: 'Prediction' 'Time Series' 'ESM'
★ 0 Likes 1 0

More Info, Quick Starts, HOLs, Docs and Tutorials

www.oracle.com/MachineLearning

https://twitter.com/CharlieDataMine

OML demos on YouTube Channel:
https://www.youtube.com/channel/UCtjWZqDAssJJ5sMi48Al85Q

Customer References:
https://www.oracle.com/database/technologies/datawarehouse-bigdata/oml-customers.html

Changing Role of the DBA: From Oracle Database Professional to Data Scientist in 6 Weeks!

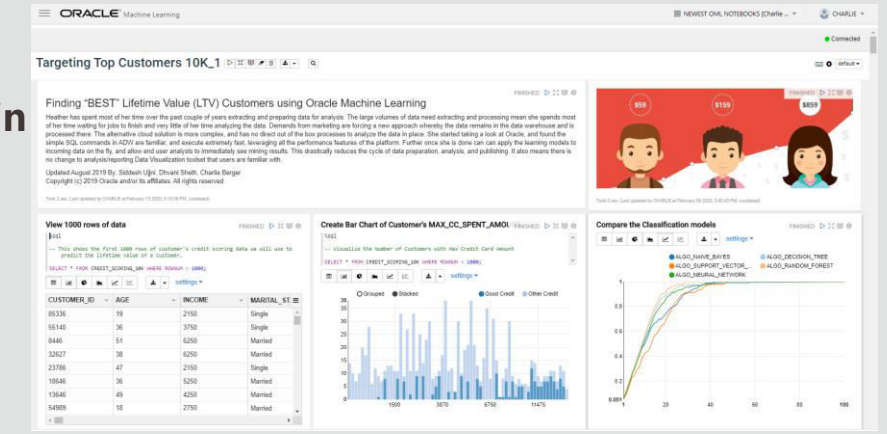
- **YouTube:** https://youtu.be/jFBMhOapGL8
- **Blog Series:** https://blogs.oracle.com/machinelearning/the-changing-role-of-the-dba%3a-from-database-developer-to-data-scientist-in-six-weeks

Hands-on Labs

OML + Wine + ADW + OAC + APEX + REST HOL Workshop
https://blogs.oracle.com/machinelearning/hands-0n-lab%3a-how-to-pick-a-good-wine-for-30%3c-using-oracle-autonomous-database%2c-oracle-machine-learning%2c-apex%2c-oracle-analytics-cloud-and-rest-services

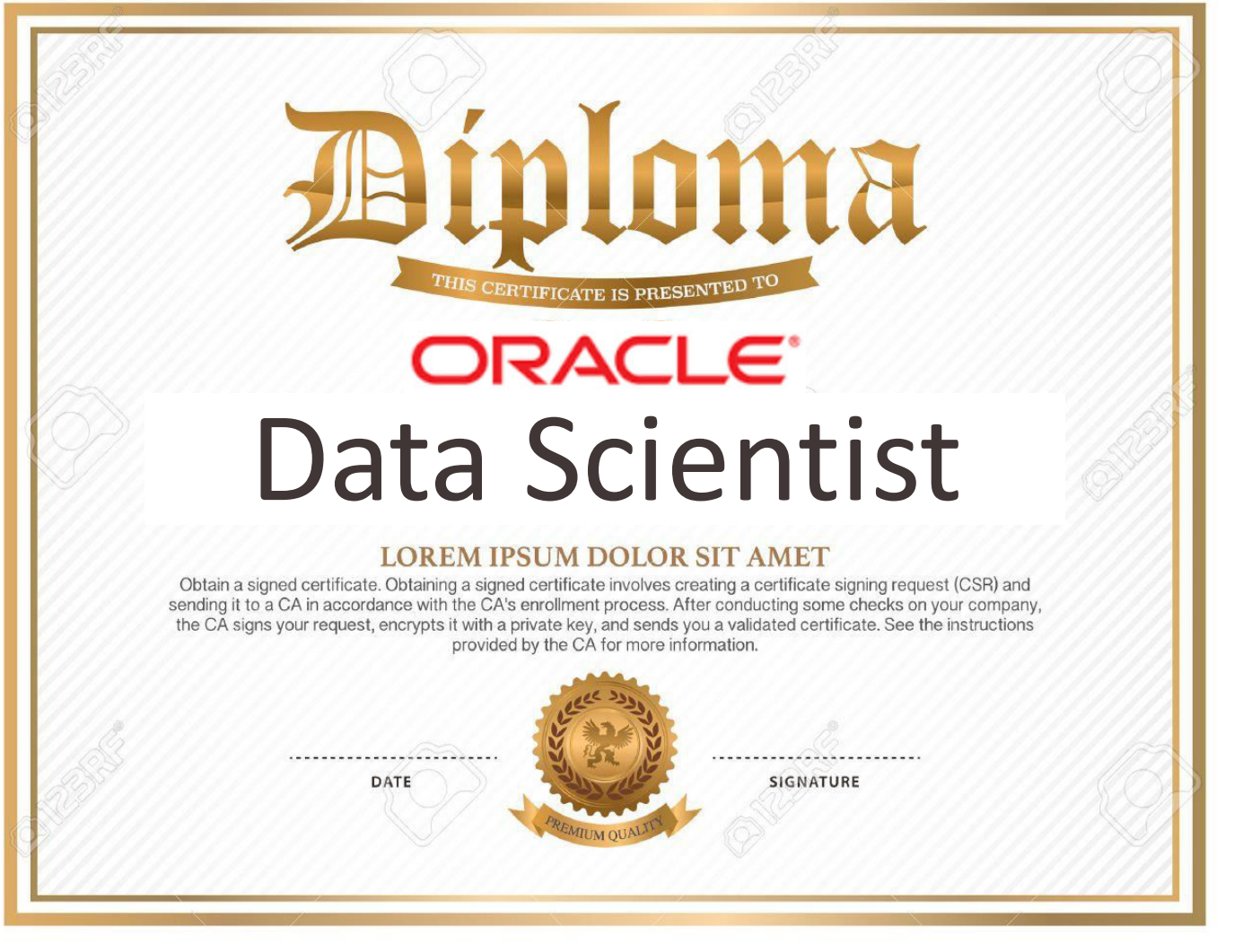
OML Oracle Data Miner HOL Workshop
https://blogs.oracle.com/machinelearning/learn-how-to-use-oracle-data-miner-ui-in-45-minutes

Oracle Machine Learning Notebooks + ADW HOL Workshop
https://blogs.oracle.com/machinelearning/learn-to-use-oracle-machine-learning-notebooks



Congratulations!

You are an Oracle Data Scientist!



Thank You

Charlie Berger

Senior Director, Product Management
Machine Learning, AI and Cognitive Analytics
Charlie.berger@oracle.com.com
CharlieDataMine