From DBA to DE: Becoming a Data Engineer

2022 Webinar Series
May 17, 2022

NYOUG

Jim Czuprynski
@JimTheWhyGuy
Zero Defect Computing, Inc.

Who Am I, and What Am I Doing Here?

ORACLE ACE Director

ORACLE Certified Professional

- E-mail me at jim@jimthewhyguy.com
- Follow me on Twitter (**@JimTheWhyGuy**)
- Connect with me on LinkedIn (**Jim Czuprynski**)

BEYOND TECH SKILLS

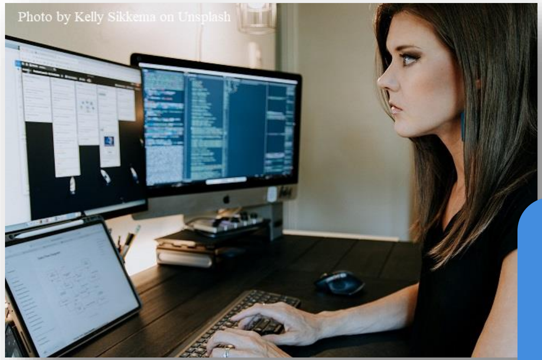
Liron Amitzi

Jim Czuprynski

The podcast that talks about everything tech – except tech.™

https://www.beyondtechskills.com

What Does a Modern Oracle DBA Spend Her Time On?



Protecting database **health, recoverability** and **security**

Tuning queries for **optimal performance and efficiency**

Building **flexible yet resilient** data models, thus ensuring data is **accurate and trustworthy**

Keeping data sources **as pristine as possible** to refresh data domains efficiently

Not Everyone Can Be A Data Scientist. Thank Goodness.

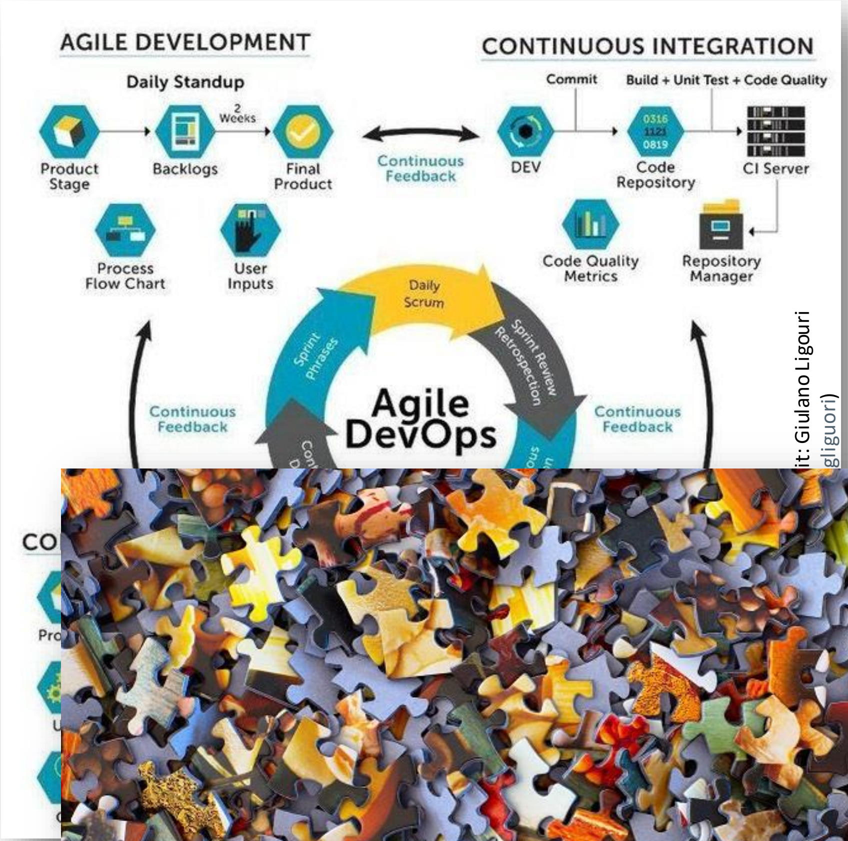
Data scientists report that they typically spend as much as **90%** of their time **cleansing data** …

… and that's when they're not **searching for relevant data**, in numerous places, in different formats …

… while ensuring their selected data is **sufficiently anonymized** to **protect subjects' privacy**

What they'd __rather__ be doing: **Training models** and **interpreting results** for **useful insights**

Data Science Is Just Like Application Development. (Not!)



DevOps: CI/CD Process Flow

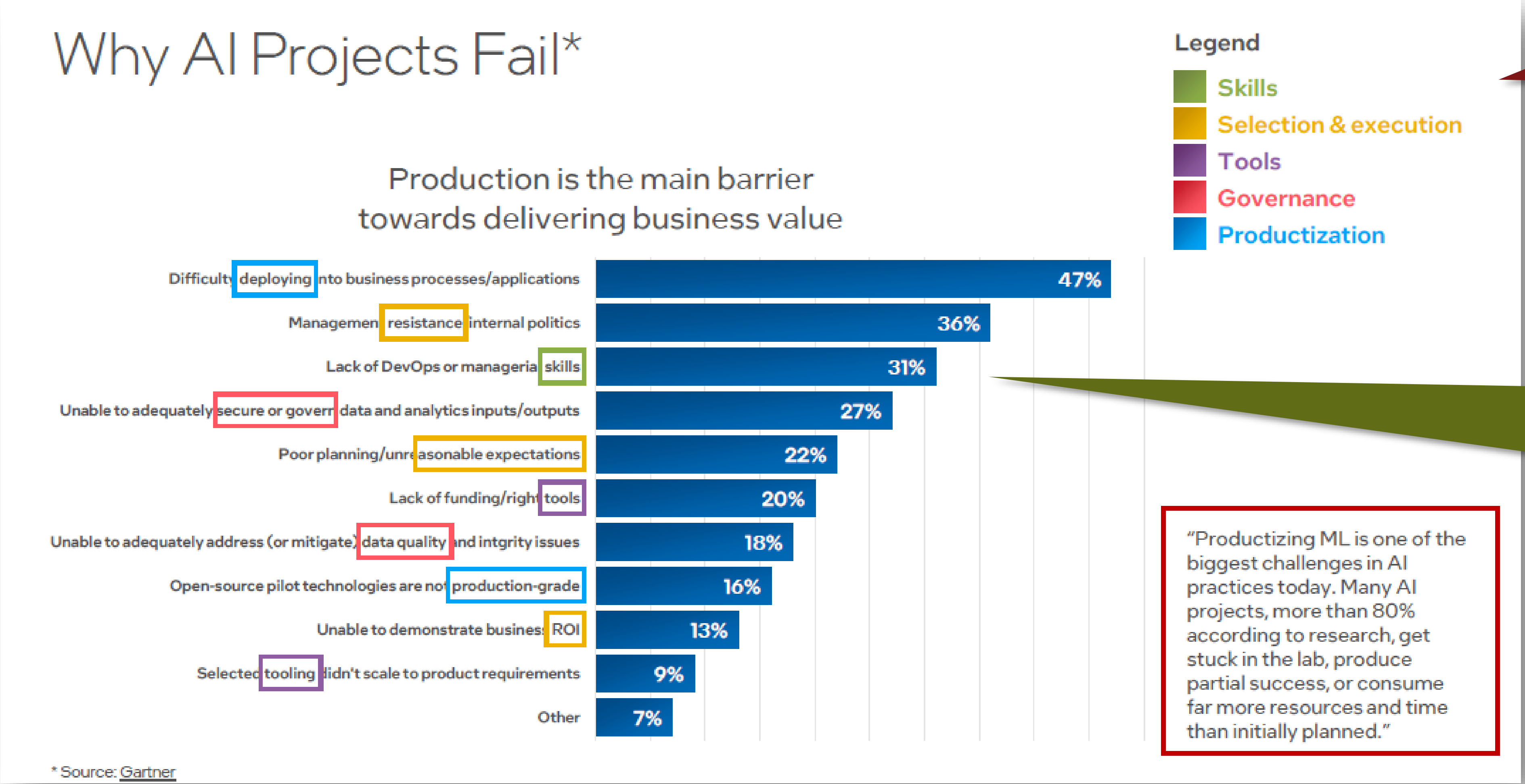
- **Focus:** Capturing, retaining, and reporting on data
- Errors are relatively, if not immediately, apparent
- Worst case: Roll back to a prior version of the application and its objects within the database*

* Assuming you've planned for that eventuality!

Data Science: Data > Useful Model(s)

- **Focus:** Accurate (and thus useful) models
- Machine Learning / AI involves *extremely complex* mathematics that devour computing cycles
- Worst case: A perfect model is now **utterly inaccurate**!
  - **Underfit:** Poor *initial* training data results in bad model precisely when *it's most needed*
  - **Overfit:** Good *initial* training data yields a good model initially … and then *new, never-before-seen* data screws up everything

Who Said AI/ML Was Easy?



From a recent seminar with Intel's AI/ML team

It turns out that most of the time, technology isn't the cause of project failure; rather, the human dimension is often the root cause

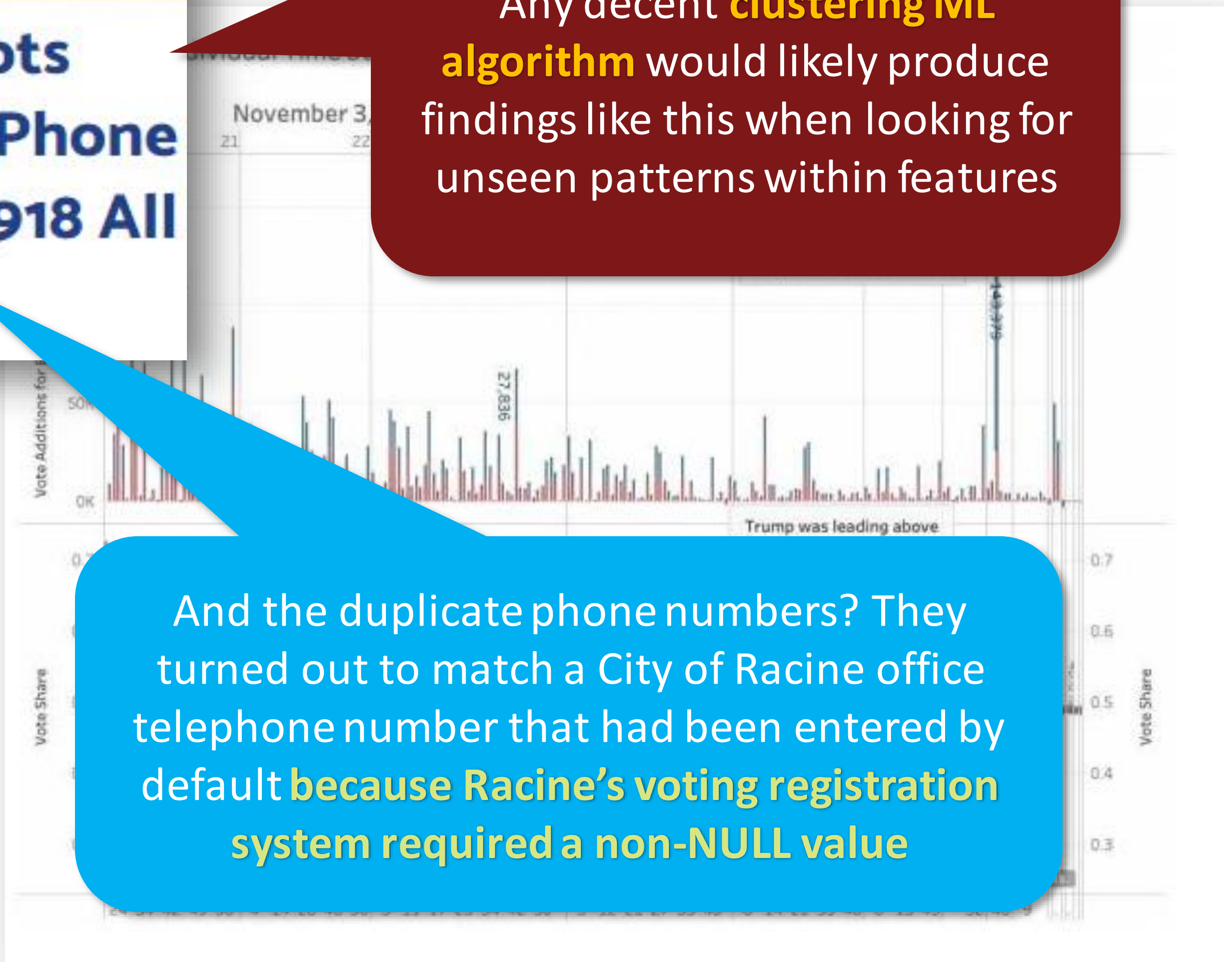
Of course, it's more complicated than this. Check out my recent blog post for deeper insights

The Scourge of Bad Data (1)


GATEWAY PUNDIT

We report the truth – and leave the Russia-Collusion fairy tale to the Conspiracy media

BREAKING – 'WISCONSIN HOT' – Grassroots Group Uncovers 23,000 Votes with Same Phone Number and 8,000 Voters Registered in 1918 All In One County!



Any decent **clustering ML algorithm** would likely produce findings like this when looking for unseen patterns within features

The reason for same dates? Values entered for birth date (**1/1/00**) and registration date (**1/1/18**) from some municipalities' voting records **during conversion to a centralized voter registration system** in 2002

And the duplicate phone numbers? They turned out to match a City of Racine office telephone number that had been entered by default **because Racine's voting registration system required a non-NULL value**

The Scourge of Bad Data (2)

An IT professional wanted to mess with California's Automatic License Plate Reader system … so he registered his vanity plate as the word **NULL**

BRIAN BARRETT SECURITY 08.13.2019 08:51 PM

How a 'NULL' License Plate Landed One Hacker in Ticket Hell

Security researcher Joseph Tartaro thought NULL would make a fun license plate. He's never been more wrong.

The next year, he got a **$35** ticket when he tried to renew his registration … because **NULL was no longer acceptable**

After he paid the ticket, the 3rd party administrator of the ticket fines collection system apparently connected his personal details to **all plates which LEOs had registered as missing or invalid**

$12,000 in fines later, he realized the joke was on him

The Scourge of Bad Data (3)

```
CREATE TABLE t_patients (
    pa_id                 NUMBER         NOT NULL
   ,pa_first_name         VARCHAR2(40)   NOT NULL
   ,pa_last_name          VARCHAR2(40)   NOT NULL
   ,pa_middle_initial     CHAR(01)       NOT NULL
   ,pa_sex                CHAR(01)       NOT NULL
   . . .
);
```

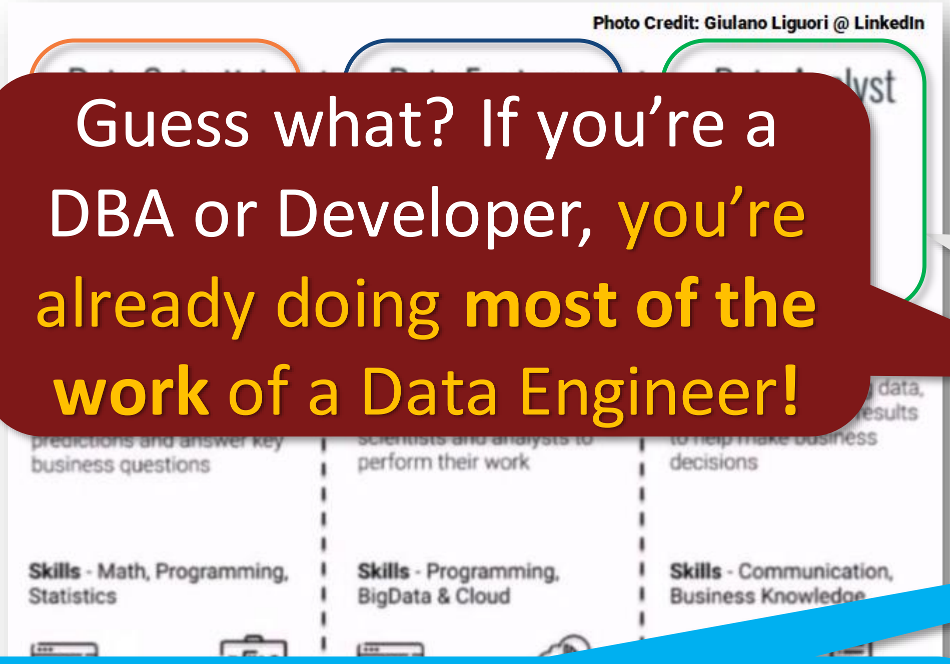
(M)ale and **(F)**emale are obvious choices …

What should be the **CHECK** constraint for this column?

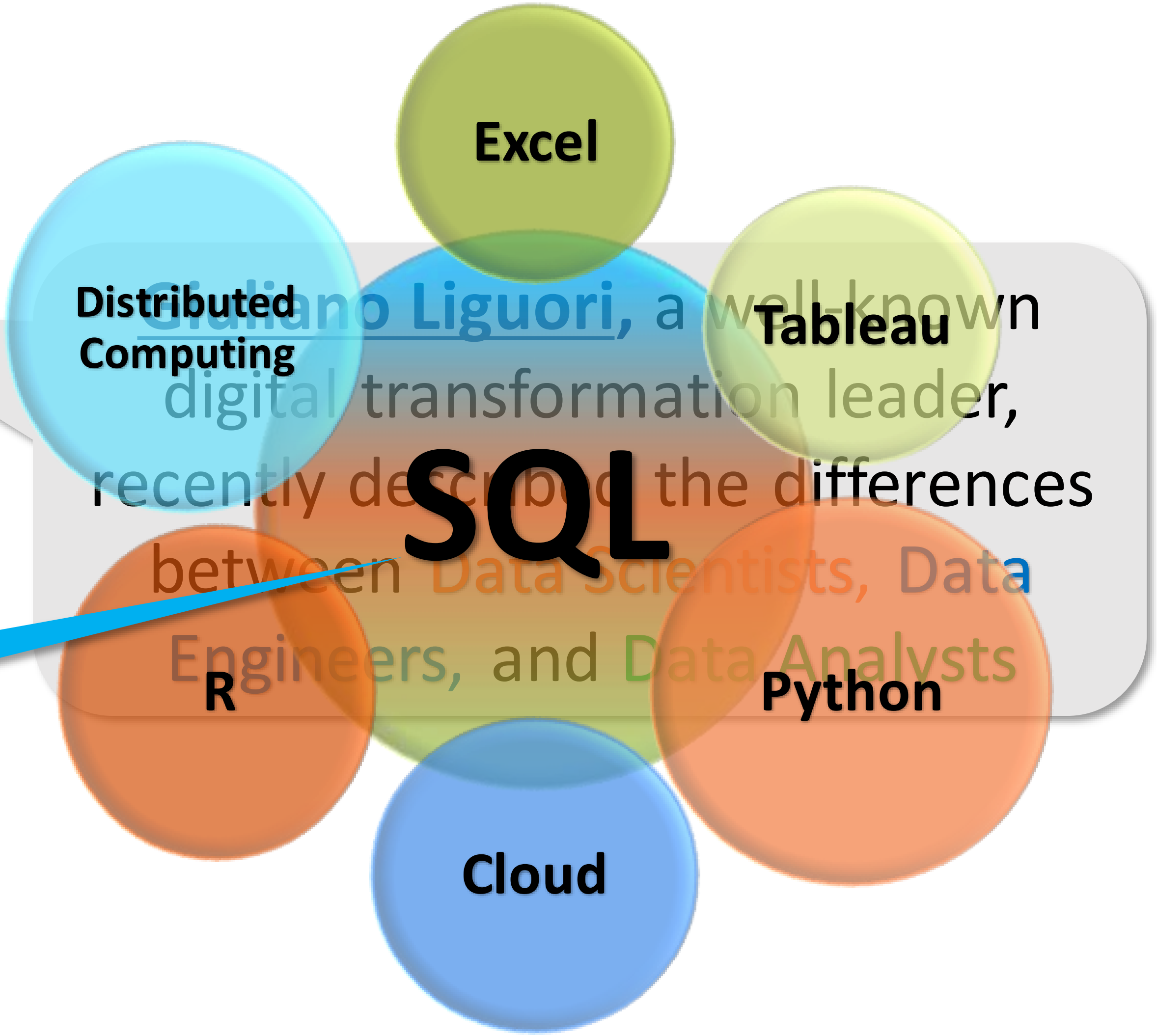
… but how do we classify trans-sexual people, or those who don't want to reveal their sex at all?

Note: We haven't even talked about the concept of *gender* yet.

So What *Does* a DE Do, Exactly?

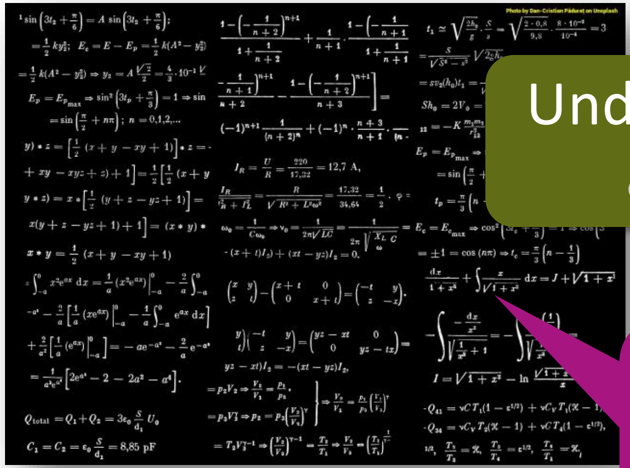


Guess what? If you're a DBA or Developer, you're already doing **most of the work** of a Data Engineer!



Notice what's at dead center of these skillsets? That's right. The sharpest tool **you already know.**

What Current DE Skills Do I Need?



Understand **statistics & probability**

Remember all that high school math you asked your teacher if you'd ever really use in real life? Yeah. It's this stuff.

Know how your DS team **extracts & processes data** (PANDA, etc.)

Yep, this means learning at least one other new language: Python

Learn to **clean & transform data** *before* your DS team needs to

Grasp **key metrics** of **model success**

Note: These are only *my* impressions of what skills are typically needed across a wide spectrum. So what skills do *your* Data Science team **really** need? **Ask them.**

How Do You Get To Carnegie Hall? Practice, Practice, Practice.

If you're still a "core" DBA, don't fret! You can start practicing all the skills you'll need to become a Data Engineer

It's easy to **leverage** the extremely powerful **Machine Learning** (ML) algorithms and **Analytic functions** already within the Oracle database …

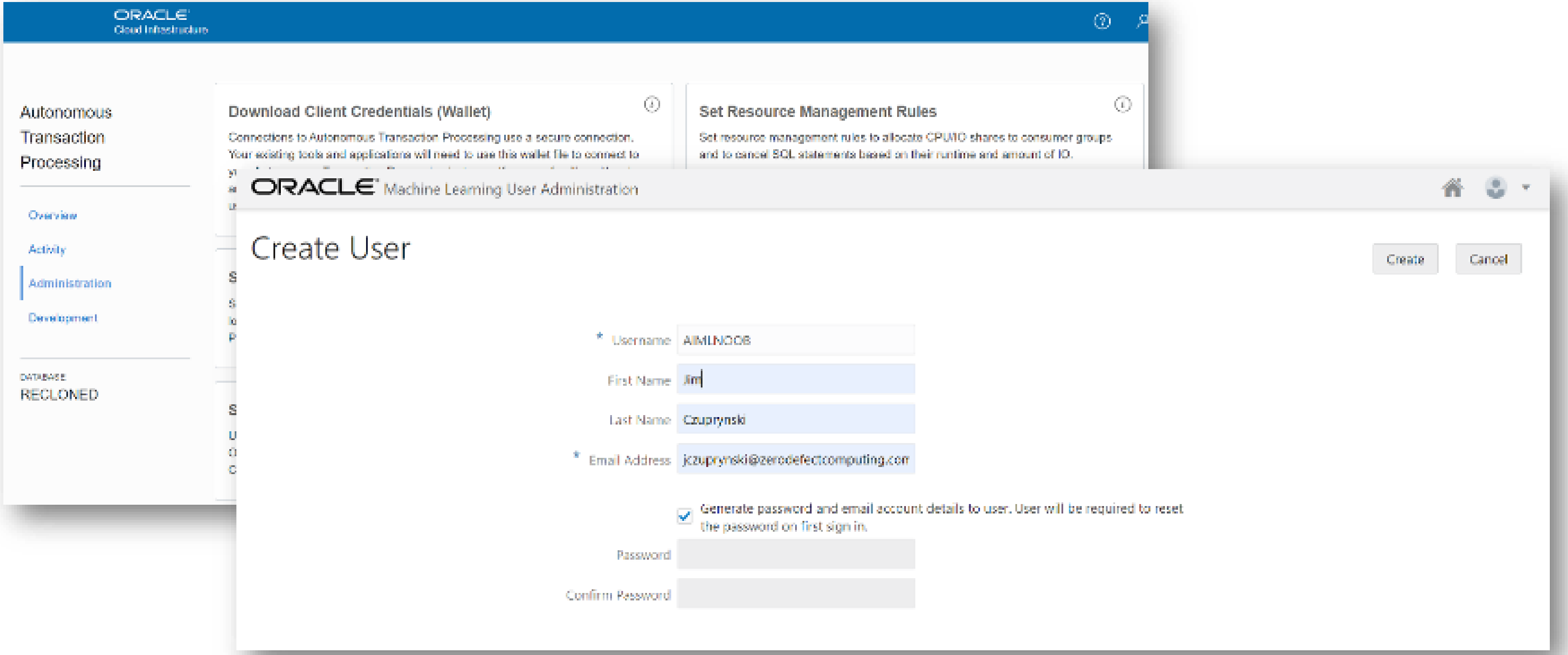
… because sometimes the <u>only</u> way to acquire the skills for a new career vector is to read > learn > do > teach

Check out the <u>newest and latest features</u> of Autonomous Database, including AutoML, OML4Py, OML4SQL, Property Graph support, and Graph Studio UI

Configuring Your OML Environment (1)

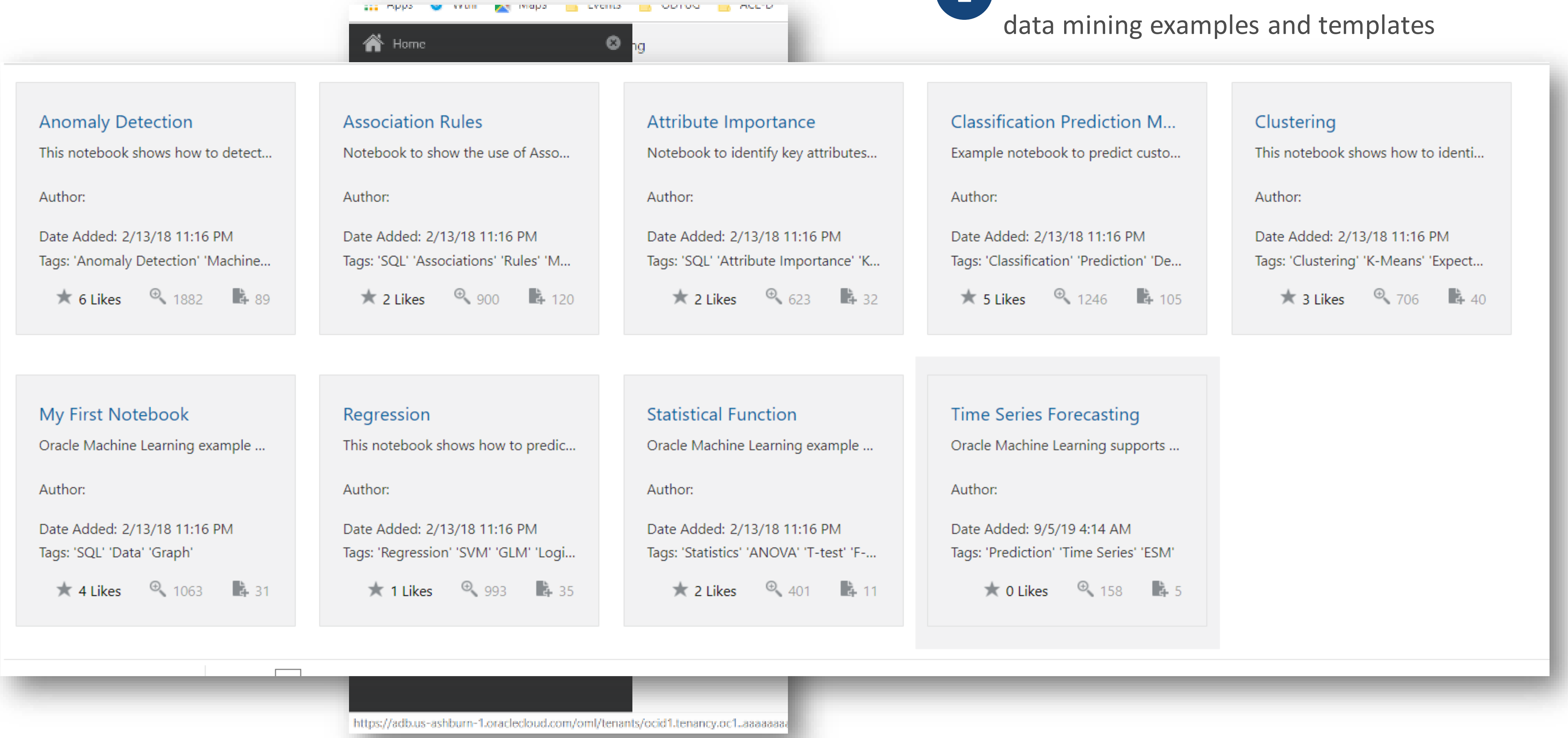
1 Request new ML User creation

2 Specify username, password, and details



Leveraging DBMS_DATA_MINING (1)

2 … and choose from a number of available data mining examples and templates



AutoML: Let the Database Decide!

This makes it easier for "citizen data scientists" to apply the power of **ML & Analytics** …

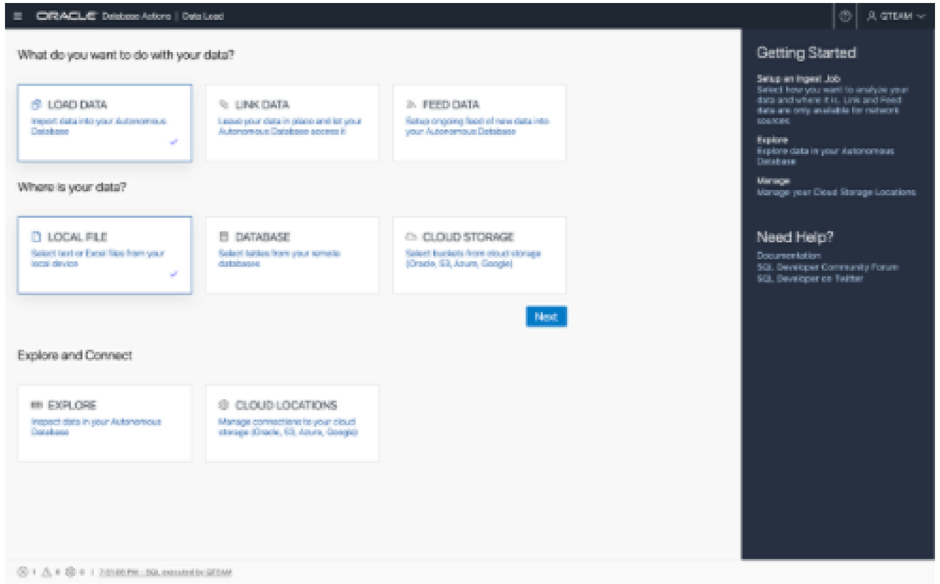
… the new **AutoML interface** makes selection of the proper algorithms a snap …

… and many more new features, including **Graph Studio**

New Innovations in Oracle Autonomous Data Warehouse

The latest release includes many new innovations, not only a broad set of capabilities that make it easier for analysts, citizen data scientists, and line-of-business developers to take advantage of the industry's first and only self-driving cloud data warehouse, but also features that deliver deeper analytics and tighter data lake integration. Key capabilities include:

- **Built-in Data Tools:** Business analysts now have a simple, self-service environment for loading data and making it available to their extended team for collaboration. They can load and transform data from their laptop or the cloud by simply dragging and dropping. They can then automatically generate business models; quickly discover anomalies, outliers and hidden patterns in their data; and understand data dependencies and the impact of changes.
- **Oracle Machine Learning AutoML UI:** By automating time-intensive steps in the creation of machine learning models, the AutoML UI provides a no-code user interface for automated machine learning to increase data scientist productivity, improve model quality and enable even non-experts to leverage machine learning.



Oracle Data Load

Check out the summary of all the latest AutoML enhancements!

Building a Data Source for AutoML to Devour

```
CREATE TABLE t_smartmeter_business_profiles AS
SELECT
 sm_id
    ,CD.cd_minority_owned
    ,CD cd family generations
    ,CD . . .
    ,CD
    ,CF
    ,CF
    ,SM
    ,SM
    ,SM
    ,SM
  FROM
    t_
. . .
```

```
 ,t_customer_demographics CD
 ,(SELECT
       sm_id
      ,ROUND(AVG(smr_kwh_used),2) AS avg_kwh_used
      ,ROUND(AVG(smr_solar_kwh),2) AS avg_solar_kwh
      ,ROUND(AVG(smr_solar_kwh) / AVG(smr_kwh_used) ,2) AS pct_solar
      ,CASE
         WHEN ROUND(AVG(smr_solar_kwh) / AVG(smr_kwh_used) ,2) >= 0.15
         THEN 1 ELSE 0
       END AS solar_superuser
    FROM
      t_smartmeters
     ,t_meter_readings
   WHERE smr_id = sm_id
   GROUP BY sm_id
   ORDER BY sm_id) SM
WHERE SM.sm_id = CF.cf_id
  AND SM.sm_id = CD.cd_id
ORDER BY sm_id;
```

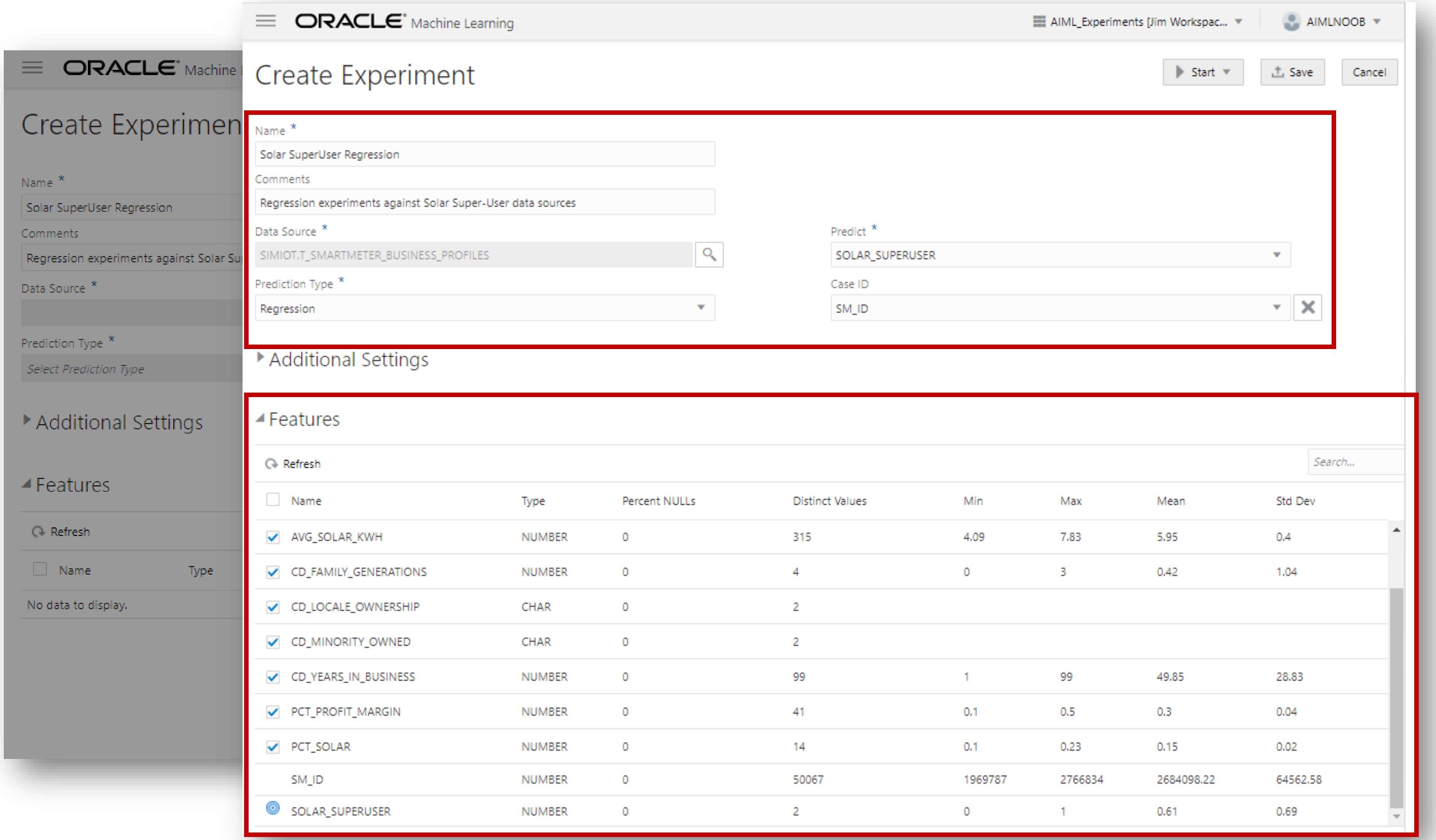
We're drawing on data summarized from a **Hybrid Partitioned table** containing **financial statistics** …

… as well as **customer demographics** and **solar energy usage data**

Regression Experiments with AutoML (1)

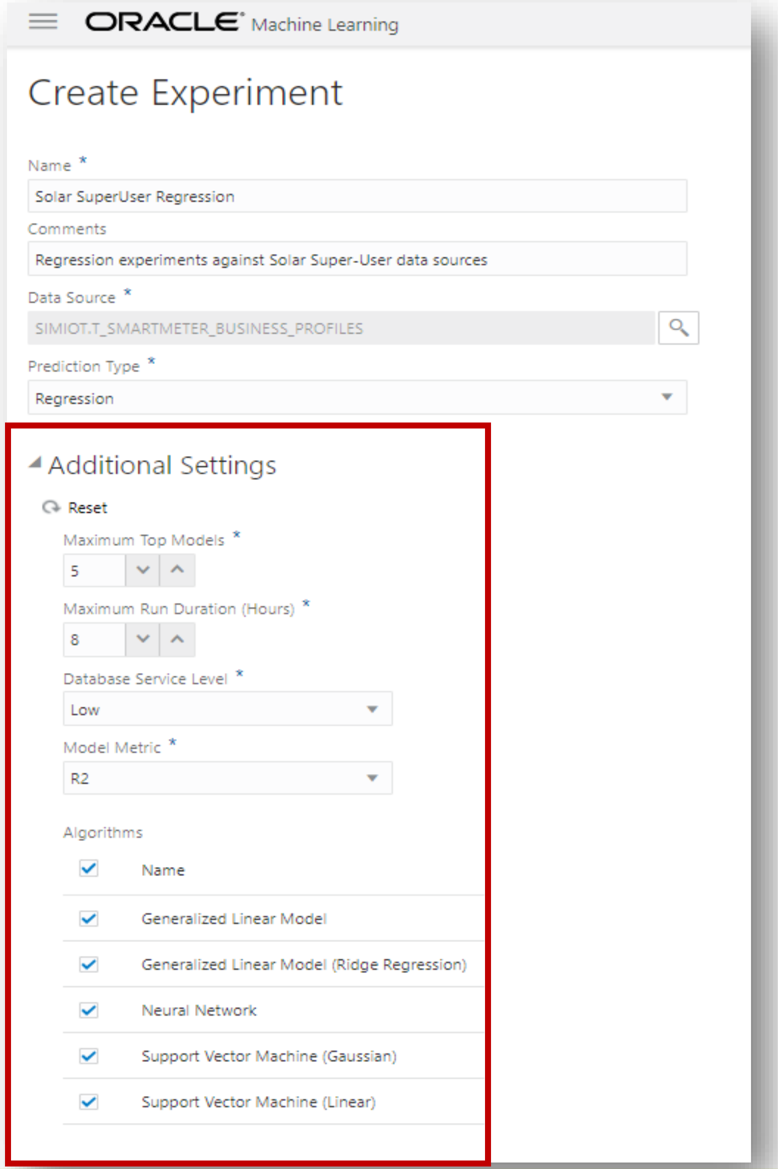
1. First, select an appropriate **data source**

2. AutoML automatically builds a list of potential **features** and their key **metrics**



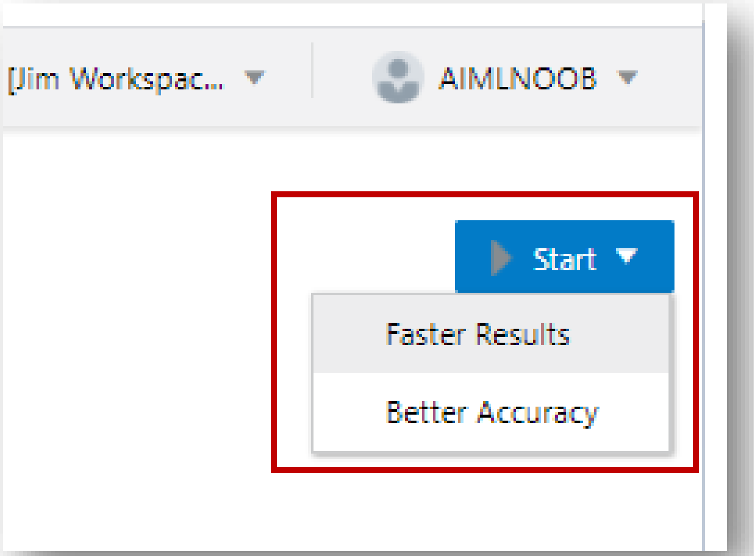
Name	Type	Percent NULLs	Distinct Values	Min	Max	Mean	Std Dev
AVG_SOLAR_KWH	NUMBER	0	315	4.09	7.83	5.95	0.4
CD_FAMILY_GENERATIONS	NUMBER	0	4	0	3	0.42	1.04
CD_LOCALE_OWNERSHIP	CHAR	0	2				
CD_MINORITY_OWNED	CHAR	0	2				
CD_YEARS_IN_BUSINESS	NUMBER	0	99	1	99	49.85	28.83
PCT_PROFIT_MARGIN	NUMBER	0	41	0.1	0.5	0.3	0.04
PCT_SOLAR	NUMBER	0	14	0.1	0.23	0.15	0.02
SM_ID	NUMBER	0	50067	1969787	2766834	2684098.22	64562.58
SOLAR_SUPERUSER	NUMBER	0	2	0	1	0.61	0.69

Regression Experiments with AutoML (2)



3 Review settings **for prediction type, run time, model metric,** and **ML algorithms to apply**

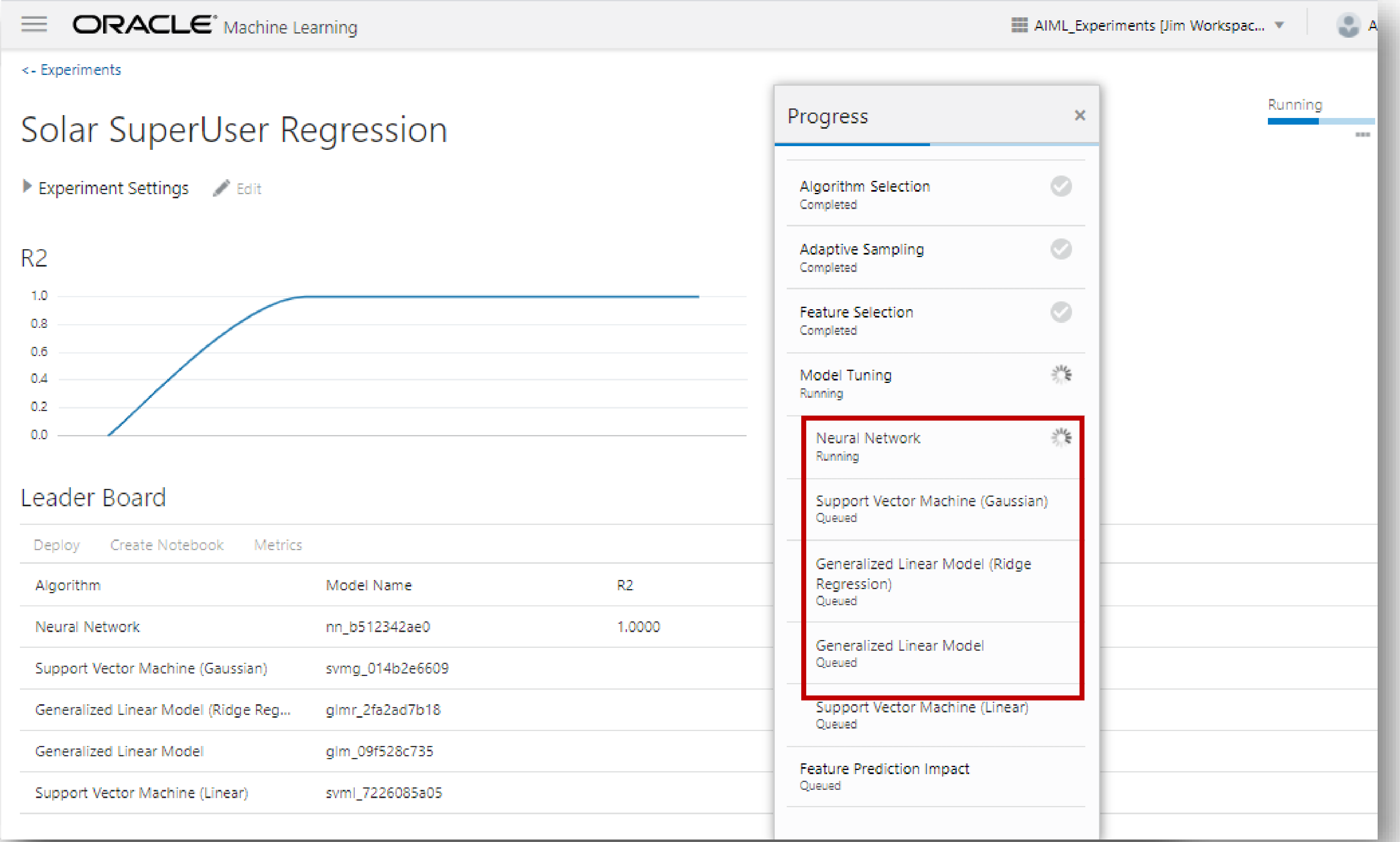
4 Start the experiment, choosing either **speed** or **accuracy**



Regression Experiments with AutoML (3)

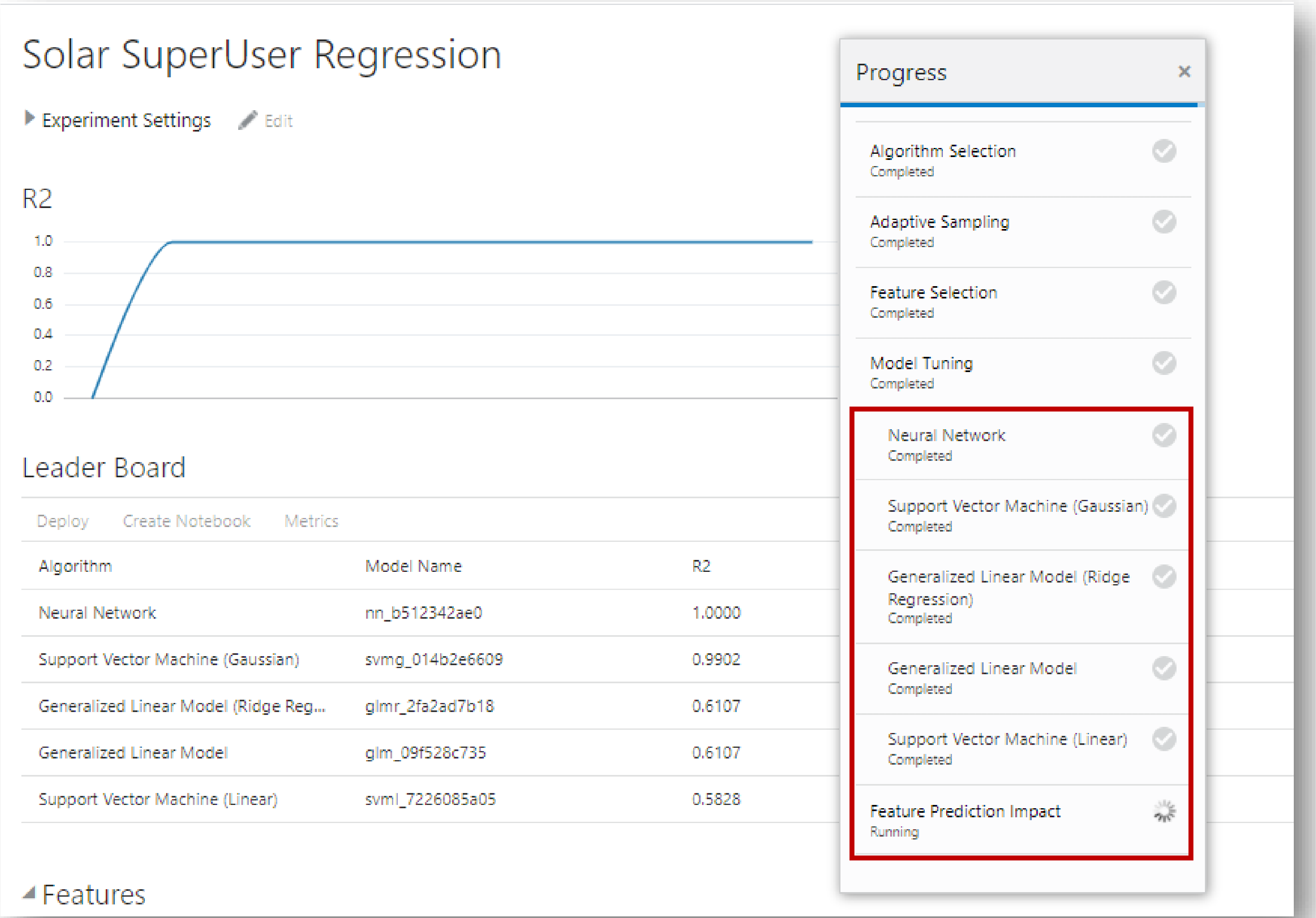
6 Next, AutoML begins building the selected **models**

5 AutoML now finishes any **sampling** needed and moves on to **feature selection**



Regression Experiments with AutoML (4)

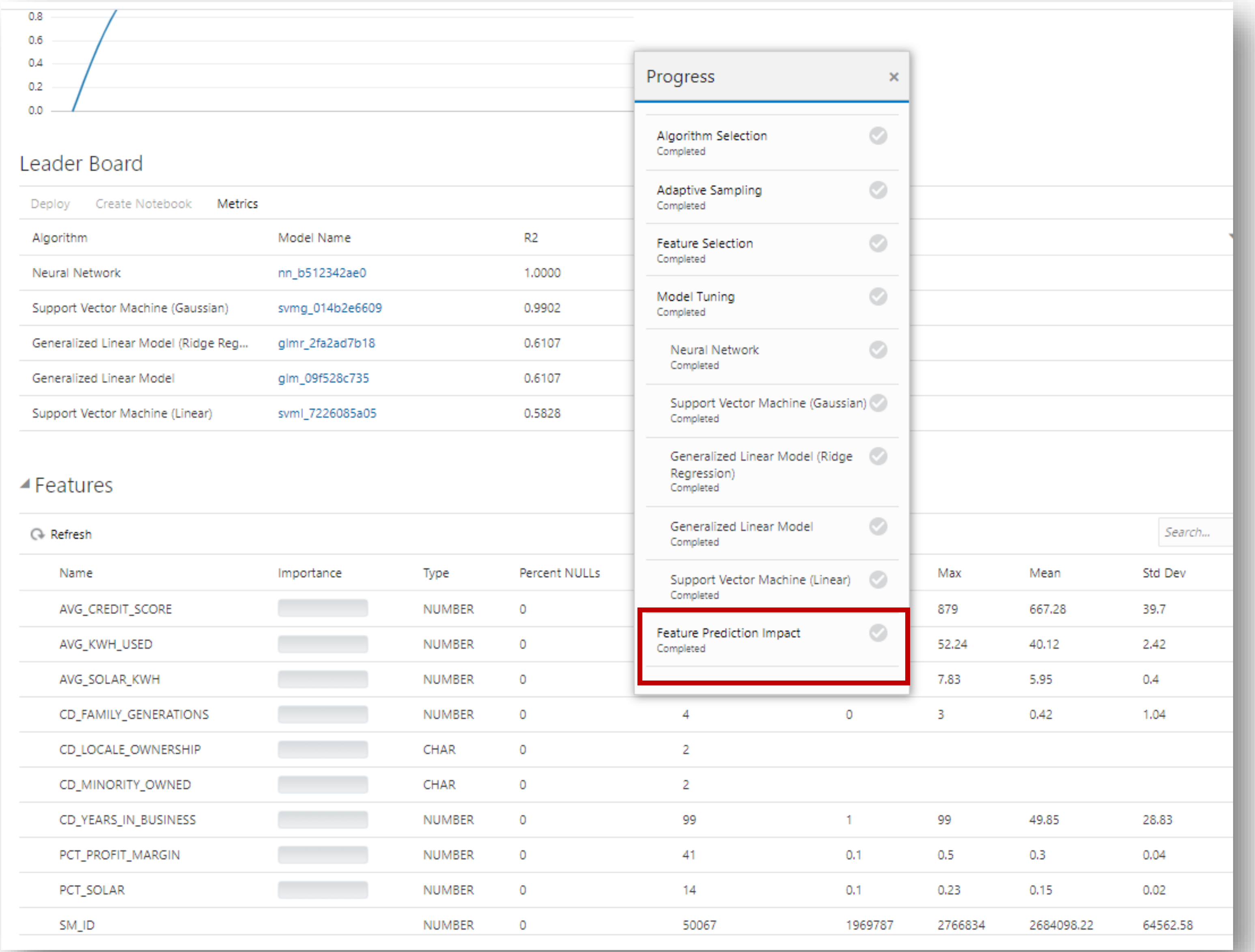
7 Model generation is complete! On to **Feature Prediction Impact** assessment ...



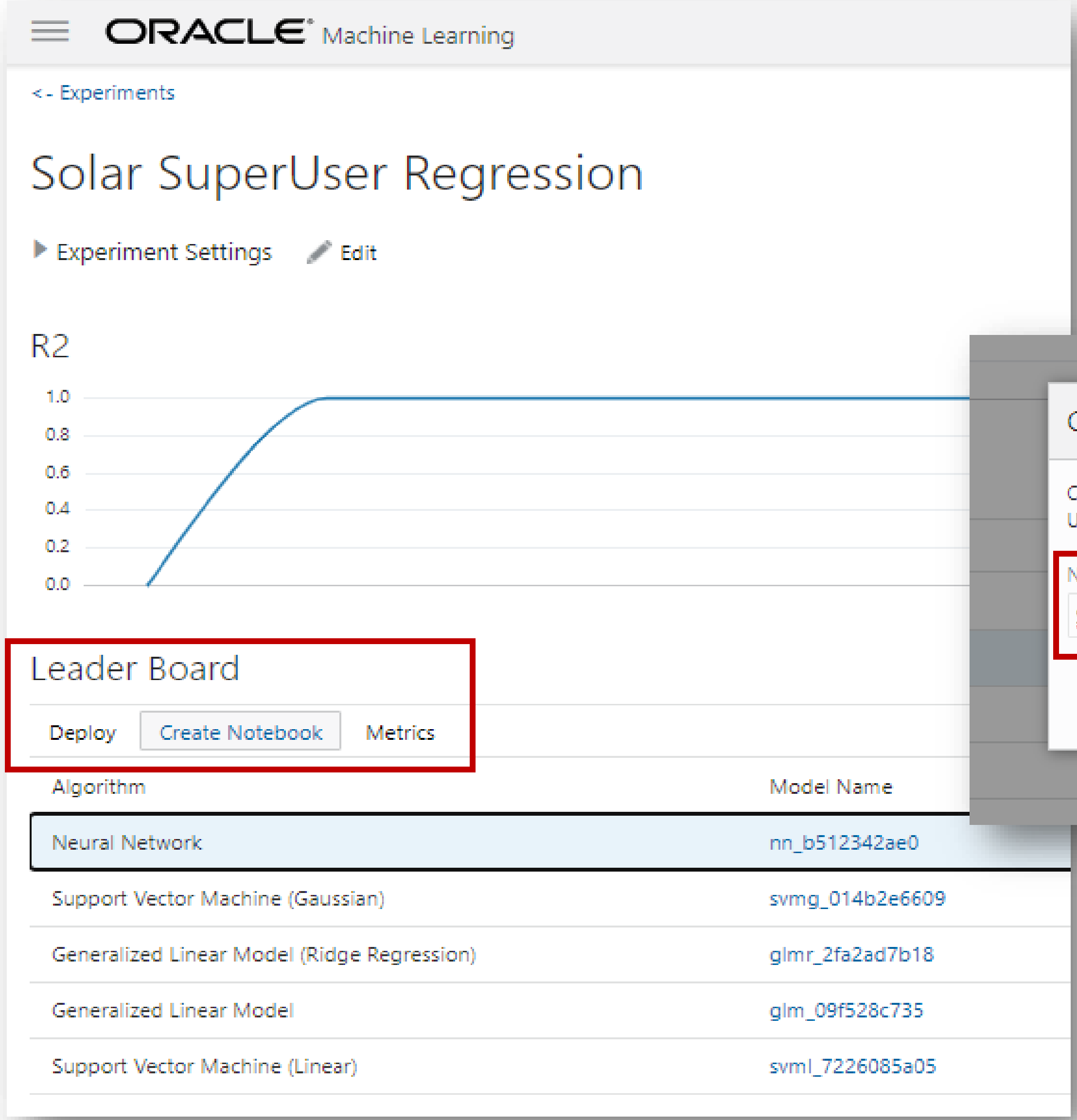
Regression Experiments with AutoML (5)

8

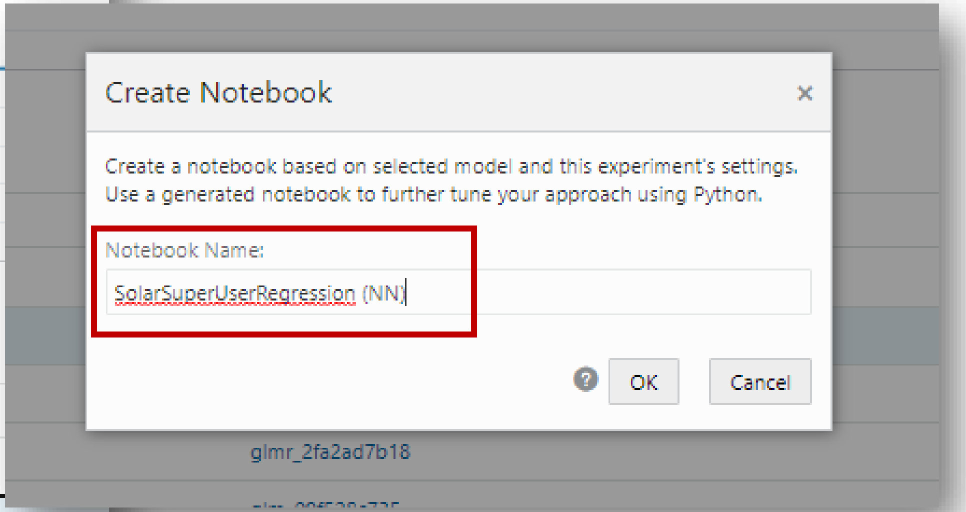
Regression(s) **complete**!
Now let's transform the **Neural Network** model into a **Zeppelin notebook**, with *just a few mouse clicks*



Transform an AutoML Experiment into a NoteBook (1)



2 **Name** the new notebook



Transform an AutoML Experiment into a NoteBook (2)

4 Don't know Python? No worries! The new notebook uses **OML4Py** to construct paragraphs for **data retrieval** and **modeling**

ORACLE Machine Learning AIML_Experiments [Jim Workspac... AIMLNOOB

Build Data READY

```
%python

import oml

columns = 'SM_ID' , 'AVG_CREDIT_SCORE' , 'AVG_KWH_USED' , 'AVG_SOLAR_KWH' , 'CD_FAMILY_GENERATIONS' , 'CD_LOCALE_OWNERSHIP' ,
    'CD_MINORITY_OWNED' , 'CD_YEARS_IN_BUSINESS' , 'PCT_PROFIT_MARGIN' , 'PCT_SOLAR' , 'SOLAR_SUPERUSER'
schema='SIMIOT'
table='T_SMARTMETER_BUSINESS_PROFILES'

column = ','.join(columns)
query = 'SELECT ' + column + ' FROM ' + schema + '.' + table

data_build = oml.sync(query=query)
z.show(data_build)
```

Create Train Data READY

```
%python

import oml

X_train = data_build[:,['SM_ID', 'AVG_CREDIT_SCORE', 'AVG_KWH_USED', 'AVG_SOLAR_KWH', 'CD_FAMILY_GENERATIONS',
    'CD_LOCALE_OWNERSHIP', 'CD_MINORITY_OWNED', 'CD_YEARS_IN_BUSINESS', 'PCT_PROFIT_MARGIN', 'PCT_SOLAR']]
y_train = data_build[:,'SOLAR_SUPERUSER']
```

Build 'NEURAL_NETWORK' Model READY

```
%python

import oml

nn_settings = {
   'ALGO_NAME' : 'ALGO_NEURAL_NETWORK' , 'NNET_HIDDEN_LAYERS' : '1' , 'NNET_ITERATIONS' : '100' , 'ODMS_SAMPLE_SIZE' : '1904' ,
       'NNET_REGULARIZER' : 'NNET_REGULARIZER_HELDASIDE' , 'NNET_ACTIVATIONS' : " 'NNET_ACTIVATIONS_LOG_SIG'" ,
       'NNET_NODES_PER_LAYER' : '18' , 'ODMS_DETAILS' : 'ODMS_DISABLE' , 'ODMS_SAMPLING' : 'ODMS_SAMPLING_ENABLE'
  }

nn_mod = oml.nn('regression',**nn_settings)

nn_mod = nn_mod.fit(X_train, y_train)
```

Transform an AutoML Experiment into a NoteBook (3)

5 ***Et voila!*** Here's your first results from a notebook completely generated via **AutoML**!

ORACLE Machine Learning

AIML_Experiments [Jim Workspac... AIMLNOOB

Connected

SolarSuperUserRegression (NN)

default

Build Data

FINISHED

```
%python

import oml

columns = 'SM_ID' , 'AVG_CREDIT_SCORE' , 'AVG_KWH_USED' , 'AVG_SOLAR_KWH' , 'CD_FAMILY_GENERATIONS' , 'CD_LOCALE_OWNERSHIP' , 'CD_MINORITY_OWNED' , 'CD_YEARS_IN_BUSINESS' ,
    'PCT_PROFIT_MARGIN' , 'PCT_SOLAR' , 'SOLAR_SUPERUSER'
schema='SIMIOT'
table='T_SMARTMETER_BUSINESS_PROFILES'

column = ','.join(columns)
query = 'SELECT ' + column + ' FROM ' + schema + '.' + table

data_build = oml.sync(query=query)
z.show(data_build)
```

settings

SM_ID	AVG_CREDIT_SCO...	AVG_KWH_USED ...	AVG_SOLAR_KW...	CD_FAMILY_GENERATION...	CD_LOCALE_OWNERS...	CD_MINORITY_OWN...
1969787	702	41.9	5.64	2	N	Y
2230604	621	40.26	6.32	2	Y	Y
2314443	689	41.01	5.83	0	N	Y
2320514	652	37.76	5.94	0	N	Y
2333622	618	38.01	5.94	0	N	Y
2390930	670	41.34	5.57	0	N	Y
2398356	785	43.08	6.14	1	N	Y
2398357	666	37.75	6.28	0	N	N

How Do I Keep My DE Career Relevant?

How did you keep your Developer / DBA career relevant?
How is this <u>any different</u>?

✓ Associate with **other DEs,** and help **uplift others** to DE status

✓ Attend **conferences and training sessions** on **latest industry trends**

✓ Consider **certifying** your **hard-won, newly-acquired skills**

They call it *life-long learning* for a reason - it **never, <u>ever</u> stops**!

Are There Any DE Professional Organizations? Maybe.

American Statistical Association (ASA)

Offers a wide array of meetings, publications, and training as well as the vaunted **PStat** and **GStat** accreditations

Data Science Council of America (DASCA)

Offers several different certifications in Big Data, Analytics, and Data Science

Institute for Operations Research and the Management Sciences (INFORMS)

Offers various trainings, events, publications, and certifications

The Association of Data Scientists (ADaSci)

Based in India, they offer a Chartered Data Scientist (CDS) certification exam and training

Further Reading In the Real World of Data Science

- **AI Projects Fail All Too Often. Successful Ones Share a Common Secret**
  https://gestaltit.com/tech-talks/intel/intel-2021/jimthewhyguy/ai-projects-fail-all-too-often-successful-ones-share-a-common-secret/

- **Machine Learning in Production: Why Is It So Hard and So Many Fail?**
  https://towardsdatascience.com/machine-learning-in-production-why-is-it-so-difficult-28ce74bfc732

- **Fact Check-Claims about 23,000 Wisconsin voters with the same phone number and 4,000 voters registered on 1/1/1918**
  https://www.reuters.com/article/factcheck-wisconsin-numbers/fact-check-claims-about-23000-wisconsin-voters-with-the-same-phone-number-and-4000-voters-registered-on-1-1-1918-missing-context-idUSL1N2RU1WC

- **How a 'NULL' License Plate Landed One Hacker in Ticket Hell**
  https://www.wired.com/story/null-license-plate-landed-one-hacker-ticket-hell/

Useful Oracle Documentation

- **What is Data Science?**

https://www.oracle.com/data-science/what-is-data-science/

- **Machine Learning Solutions with Oracle's Services and Tools**

https://www.oracle.com/a/ocom/docs/build-machine-learning-solutions-cloud-essentials.pdf

- **Oracle Cloud Infrastructure Data Catalog**

https://www.oracle.com/a/ocom/docs/ebook-cloud-infrastructure-data-catalog.pdf

- **OML Algorithms "Cheat Sheet"**

https://www.oracle.com/a/tech/docs/oml4sql-algorithm-cheat-sheet.pdf

- **Oracle 21*c* Machine Learning Basics (including AutoML)**

https://docs.oracle.com/en/database/oracle/machine-learning/oml4sql/21/dmcon/machine-learning-basics.html