

AutoML on the Half Shell: How are our Oysters?

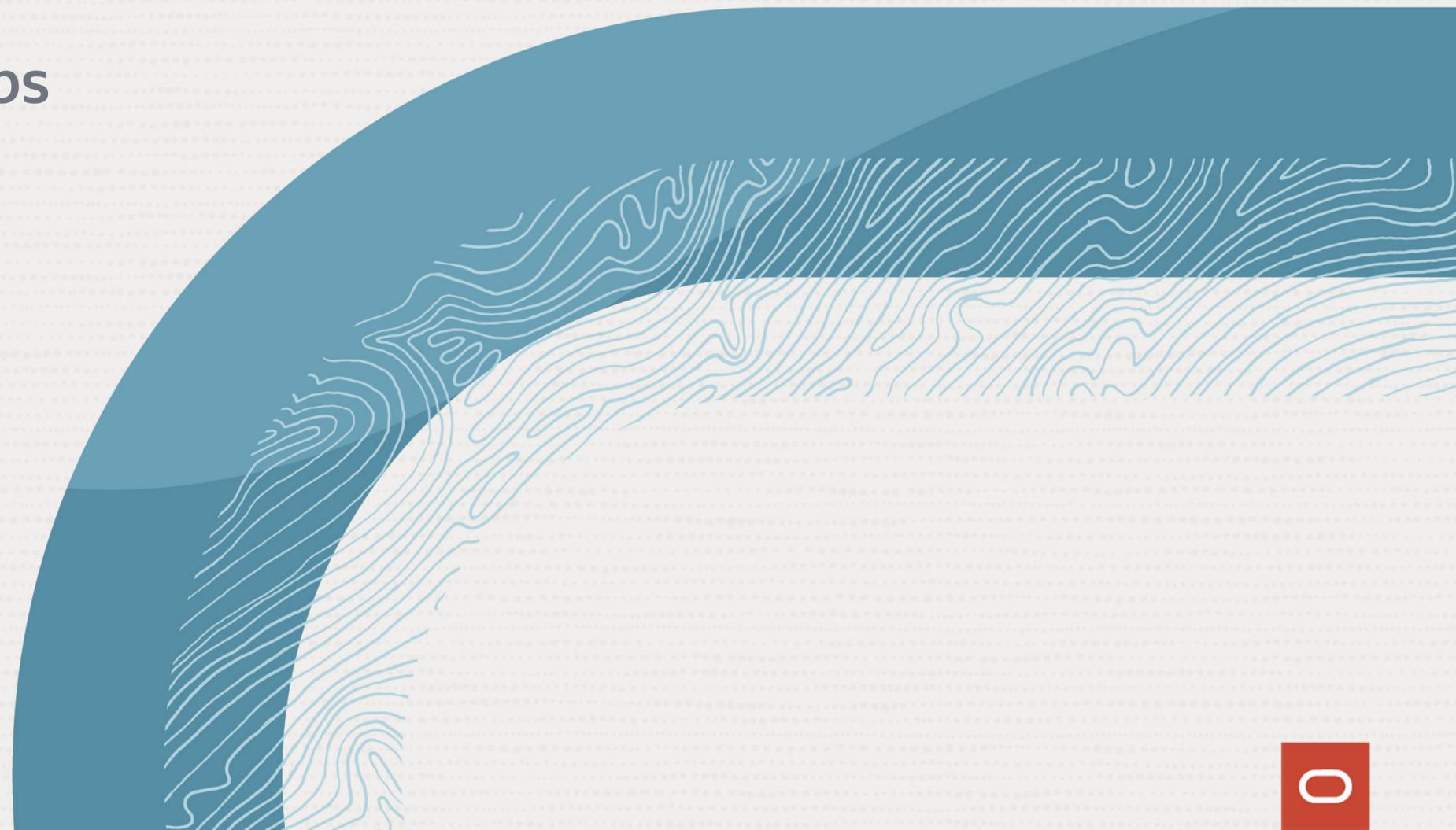
Hesam Fathi Moghadam

Senior Research Manager, Oracle Labs

March 14, 2023

Analytics & Data Summit

Redwood Shores, CA



Agenda

- Data Science Pipeline
 - How Automated Machine Learning with Explainability (AutoMLx) Fits In
- Application Example
 - Predicting Oyster Health
 - Demo

The data scientist pipeline

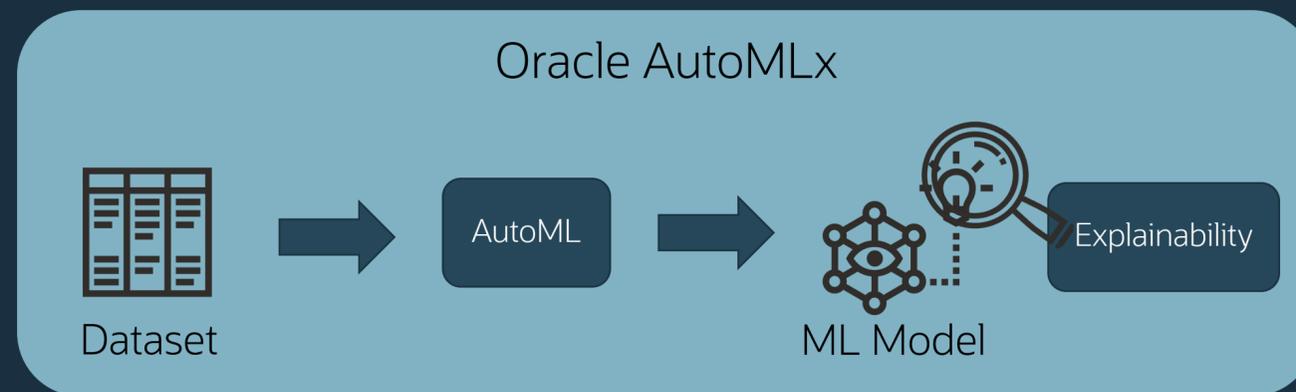


Creating a model:

- Which model?
- Are my features good (enough)?
- What hyper-parameter configuration?

Using a model:

- Can I trust my model?
- Is my model “fair”?
- Does it meet regulatory requirements?



AutoMLx

Easy-to-use
interface!

```
from automl import Pipeline
```

```
# 'regression' also supported;
```

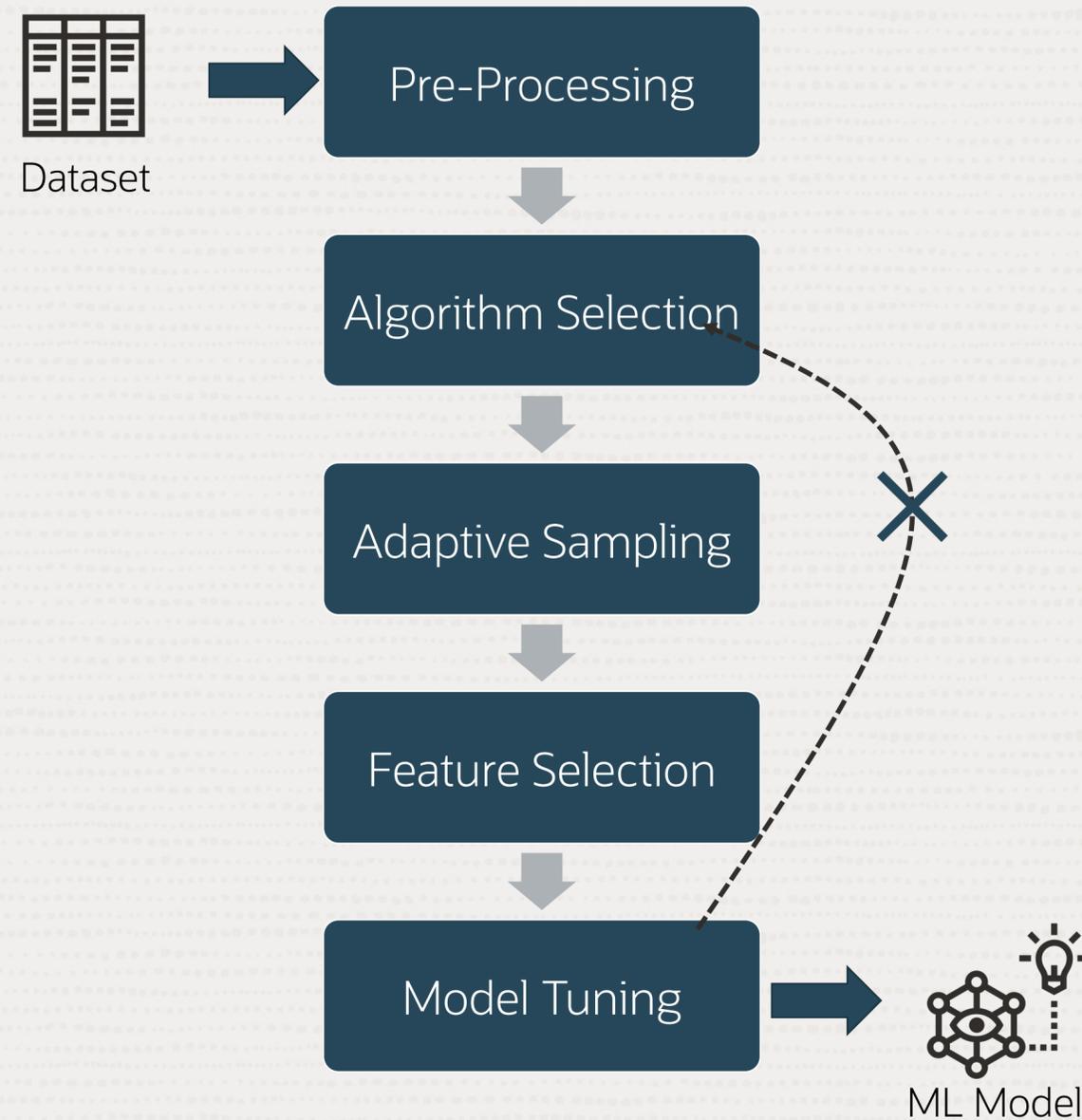
```
# 'forecasting' and 'anomaly detection' upcoming
```

```
pipeline = Pipeline(task='classification')
```

```
pipeline.fit(X, y)
```

```
y_pred = pipeline.predict(X_test)
```

Oracle's AutoML pipeline



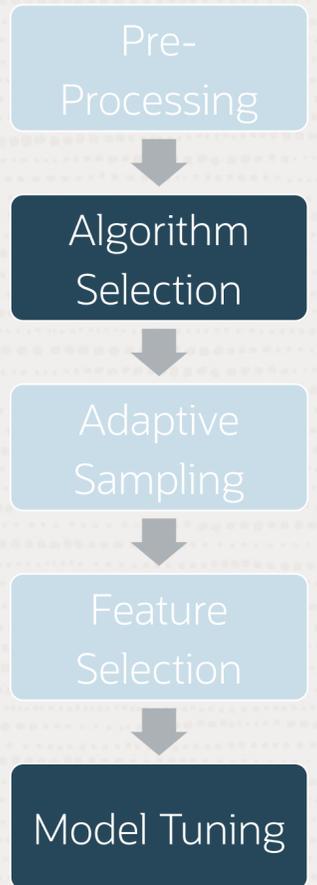
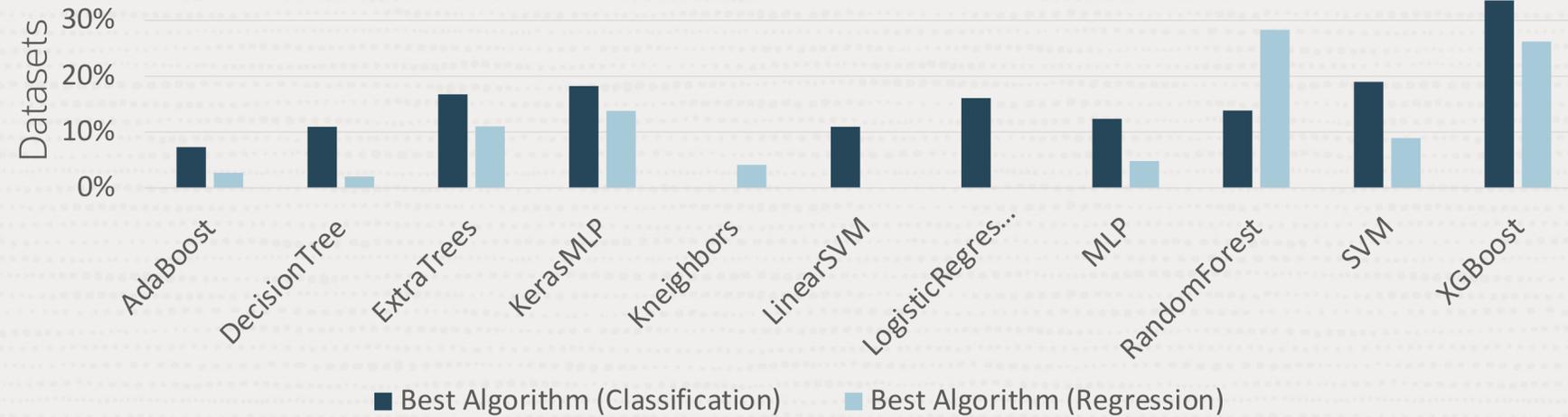
Traditional AutoML uses:

- Combined algorithm selection and hyper-parameter configuration

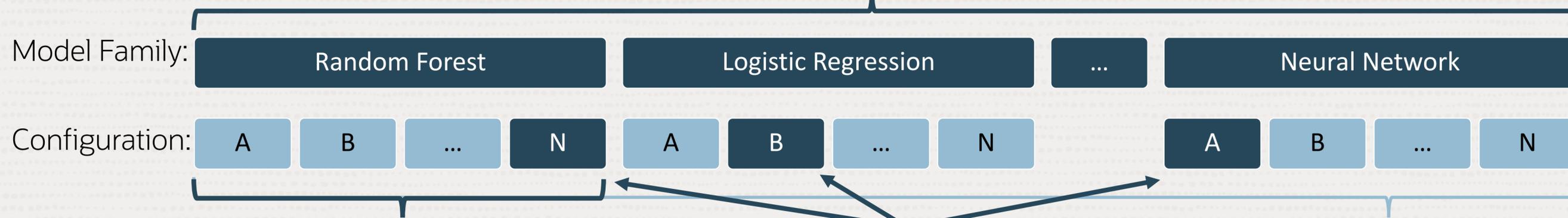
Our secret sauce?

- We never look back!

Algorithm selection & model tuning



Step 2 (Algorithm Selection):
We select between



Step 5 (Model Tuning):
Gradient-based search

Meta-learned proxy
model representatives

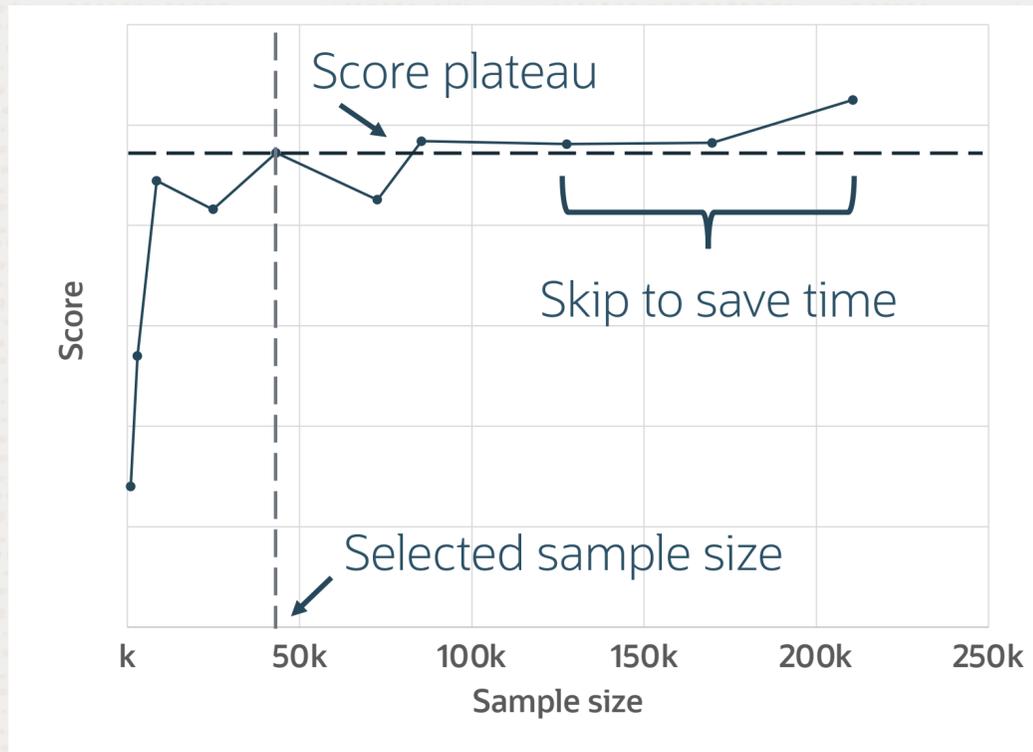
Traditional AutoML
selects between



Adaptive data reduction

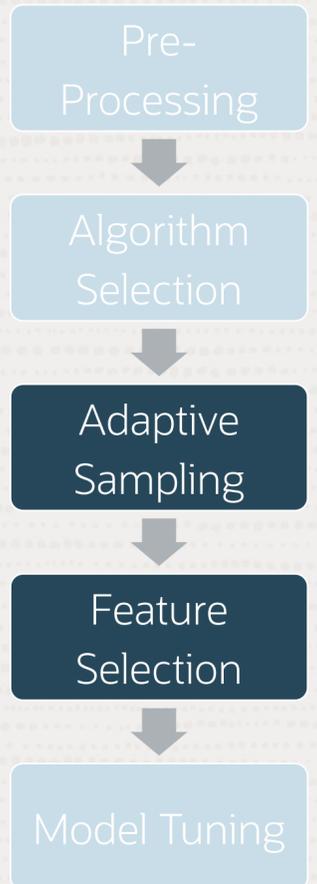
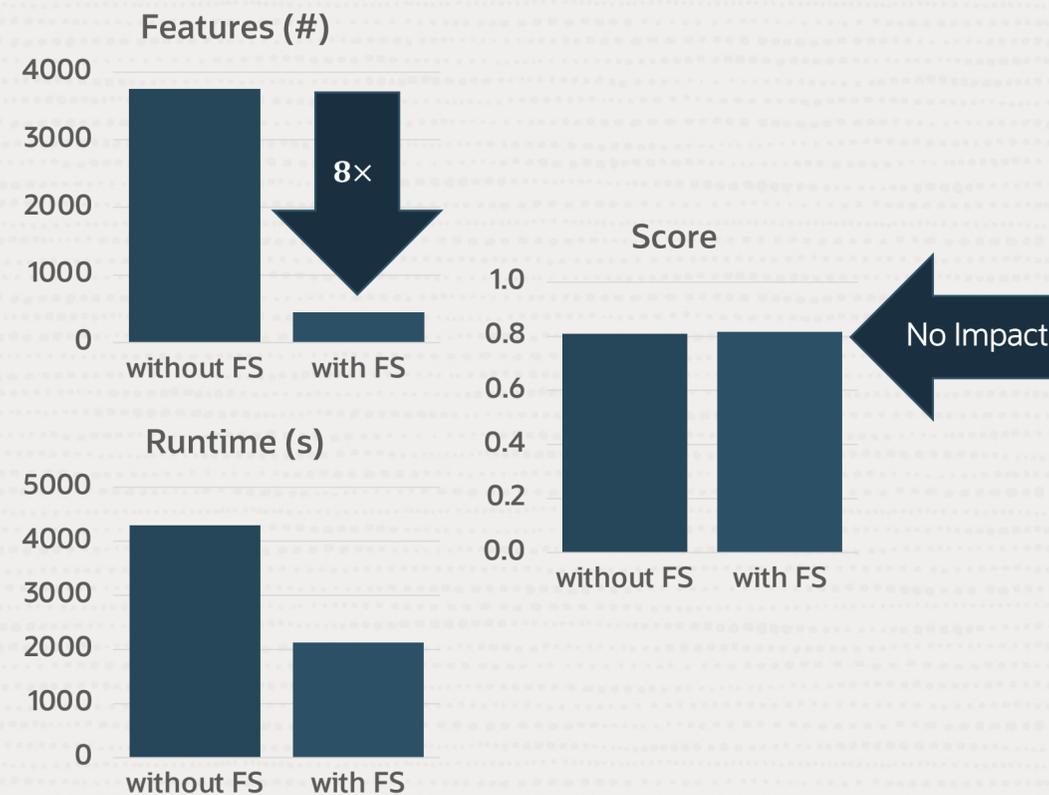
Adaptive Sampling

- Subsample rows for faster training
- Speeds up model search

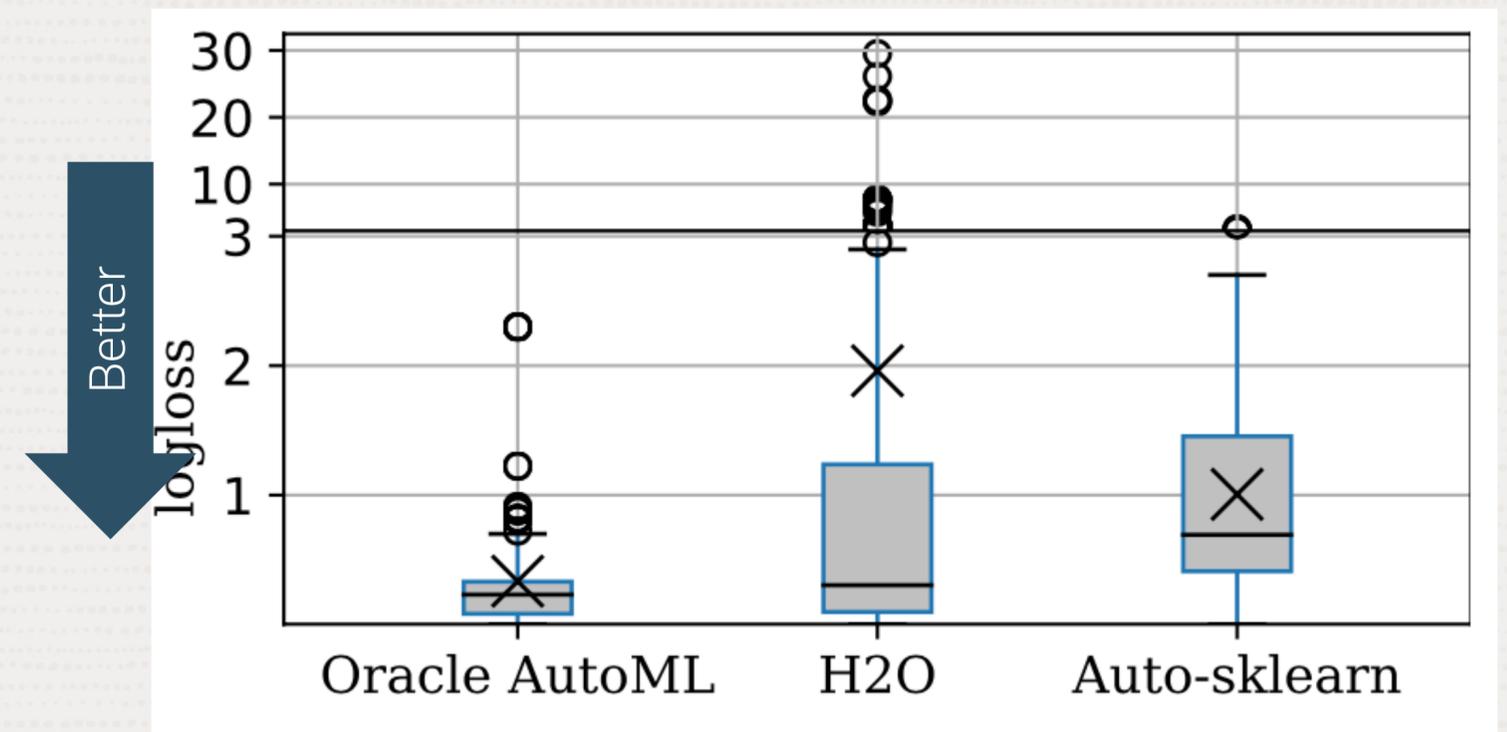
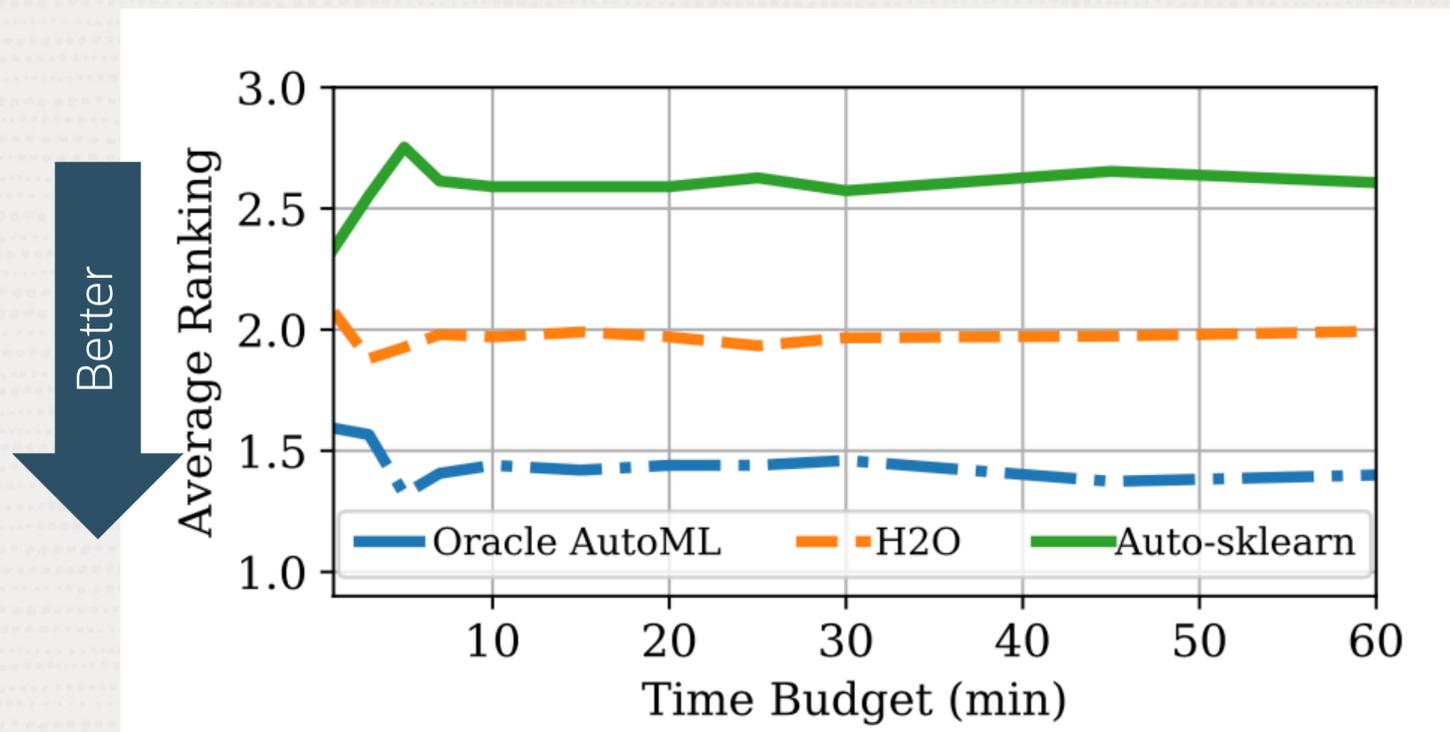


Feature Selection

- Subsample columns for faster training
- Can also reduce overfitting



Oracle AutoML Benchmarking



3.5 – 4× faster
and better scores

[Yakovlev, Anatoly, et al. "Oracle automl: a fast and predictive automl pipeline." *Proceedings of the VLDB Endowment* 13.12 \(2020\): 3166-3180.](#)



AutoMLx

Easy-to-use
interface!

```
from automl import MLEExplainer
```

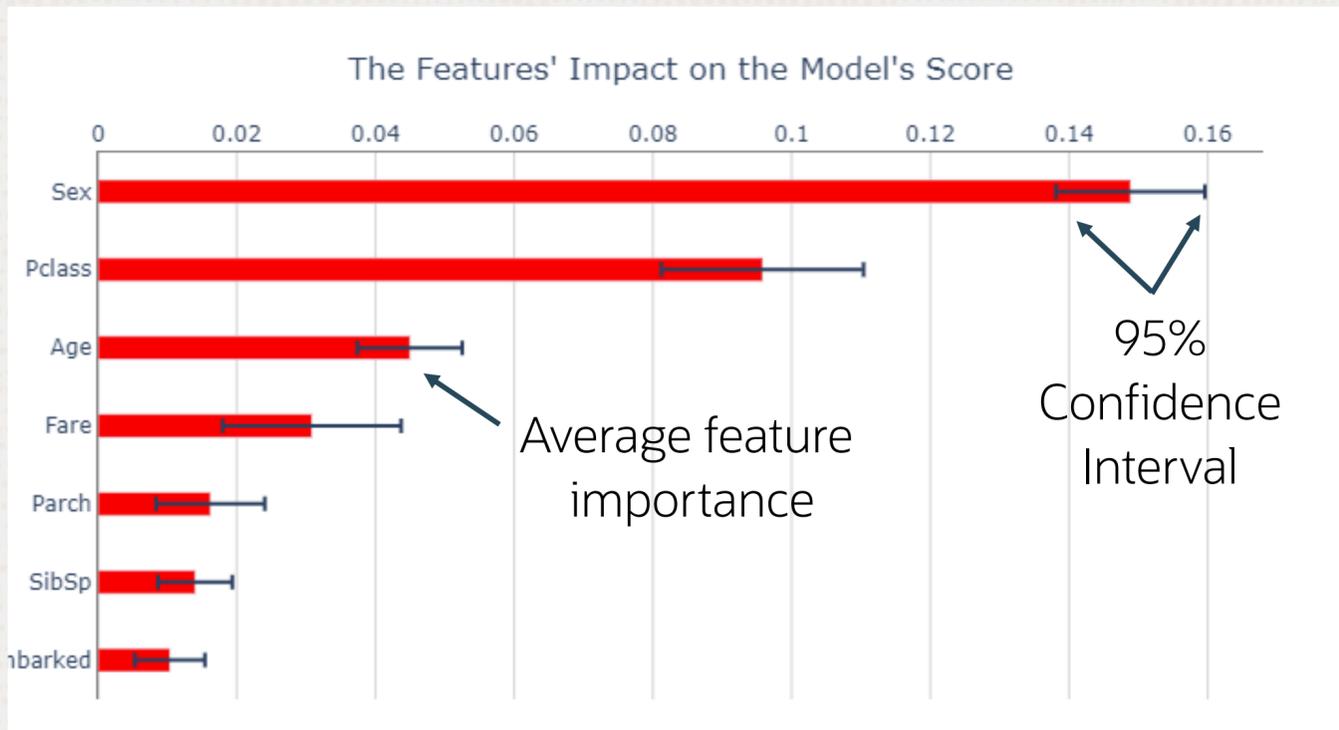
```
# Can be an AutoML pipeline or scikit-learn model  
explainer = MLEExplainer(model, X, y, task)
```

```
# Global feature importance  
explainer.explain_model()
```

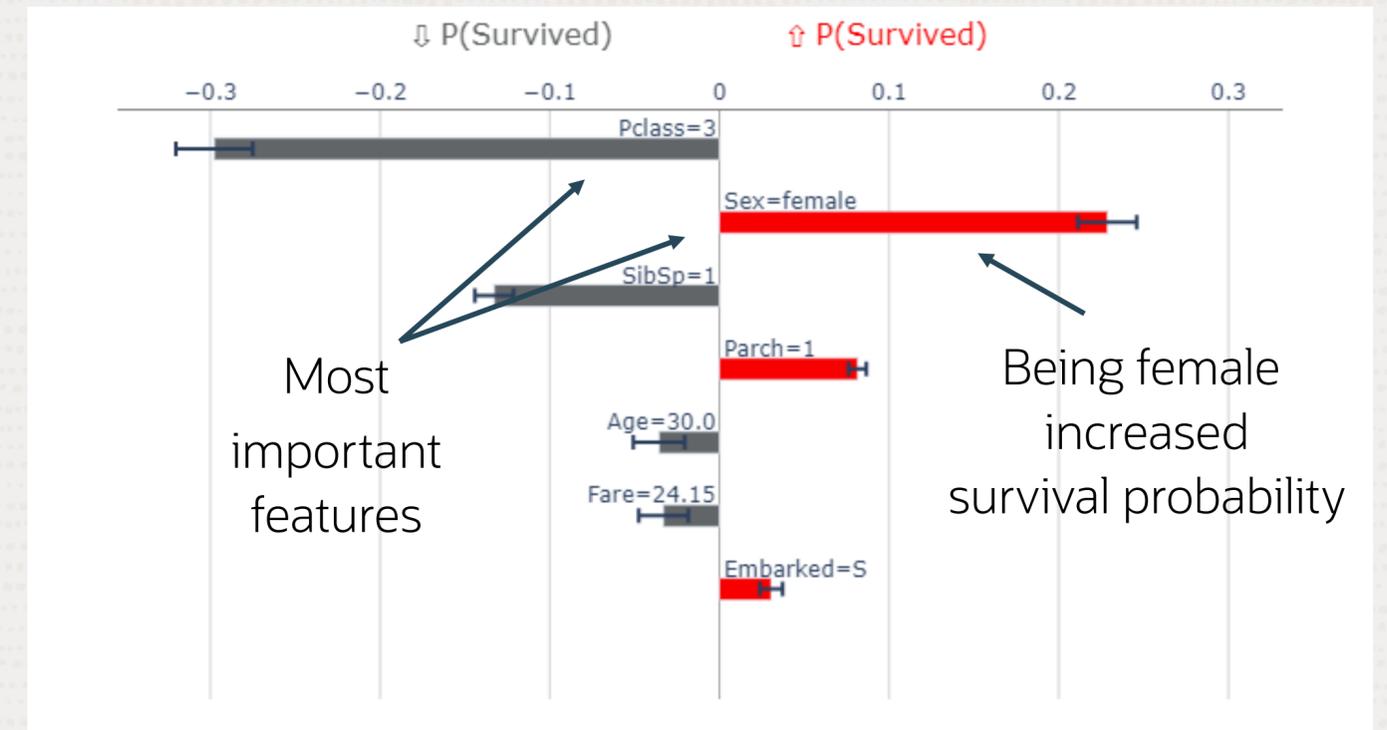
```
# Local feature importance  
explainer.explain_prediction(X_test)
```

```
# Partial dependence plot  
explainer.explain_feature_dependence(feature)
```

Feature importance examples – titanic dataset

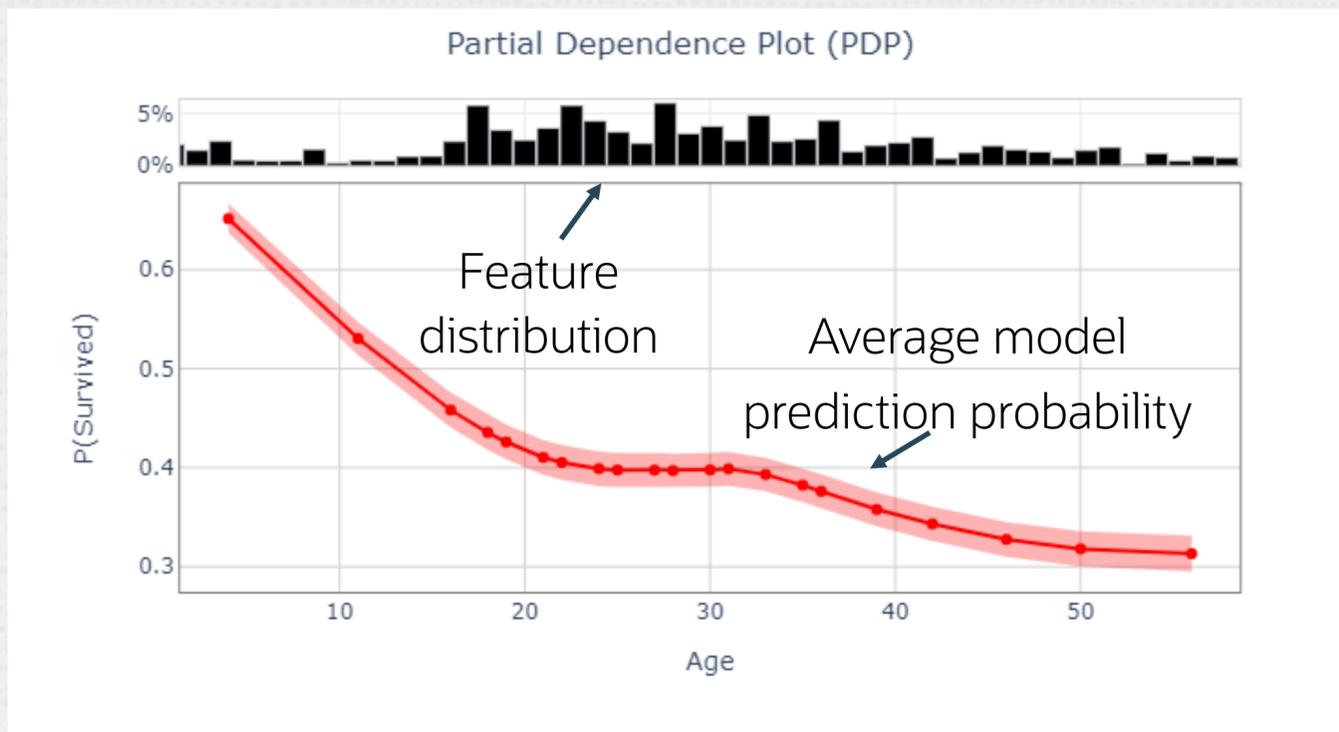


Global (model) feature importance
`explainer.explain_model()`



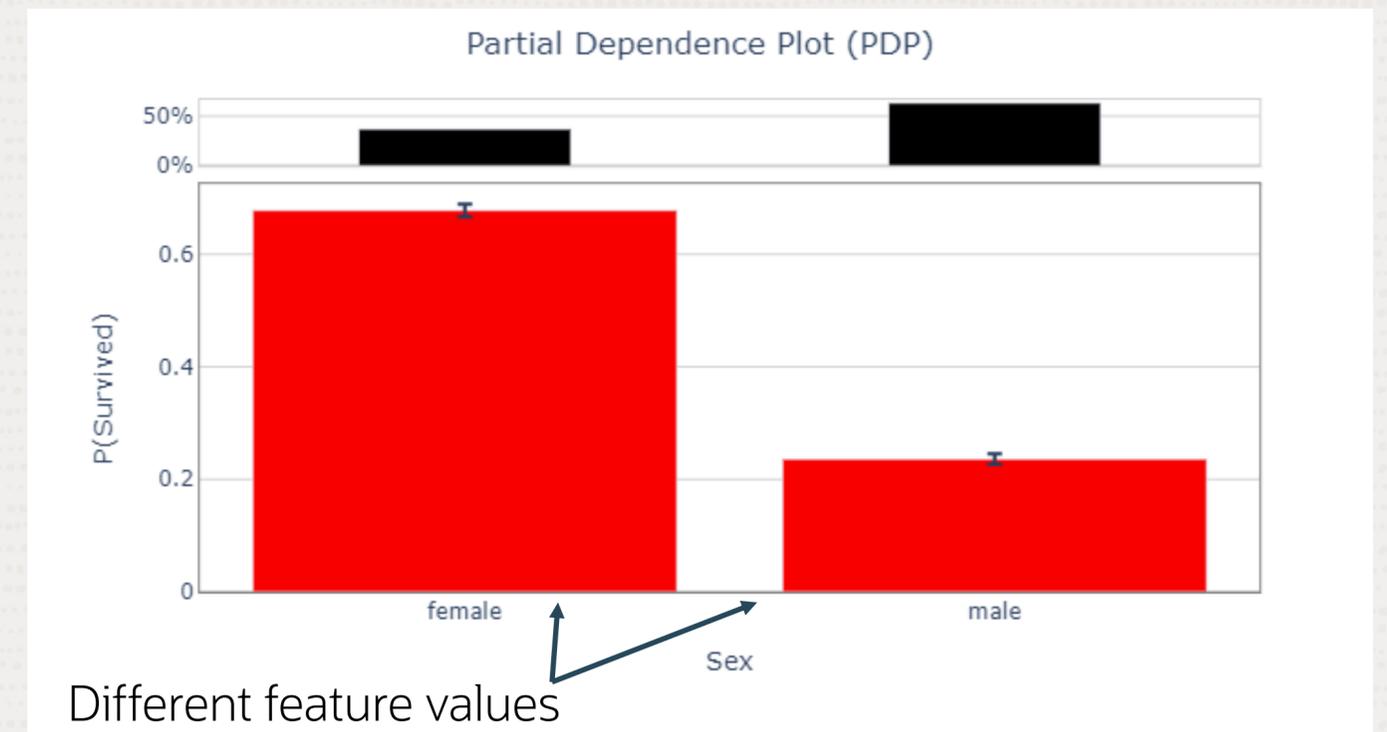
Local (prediction) feature importance
`explainer.explain_prediction(X_test)`

Feature dependence examples – titanic dataset



Continuous feature PDP

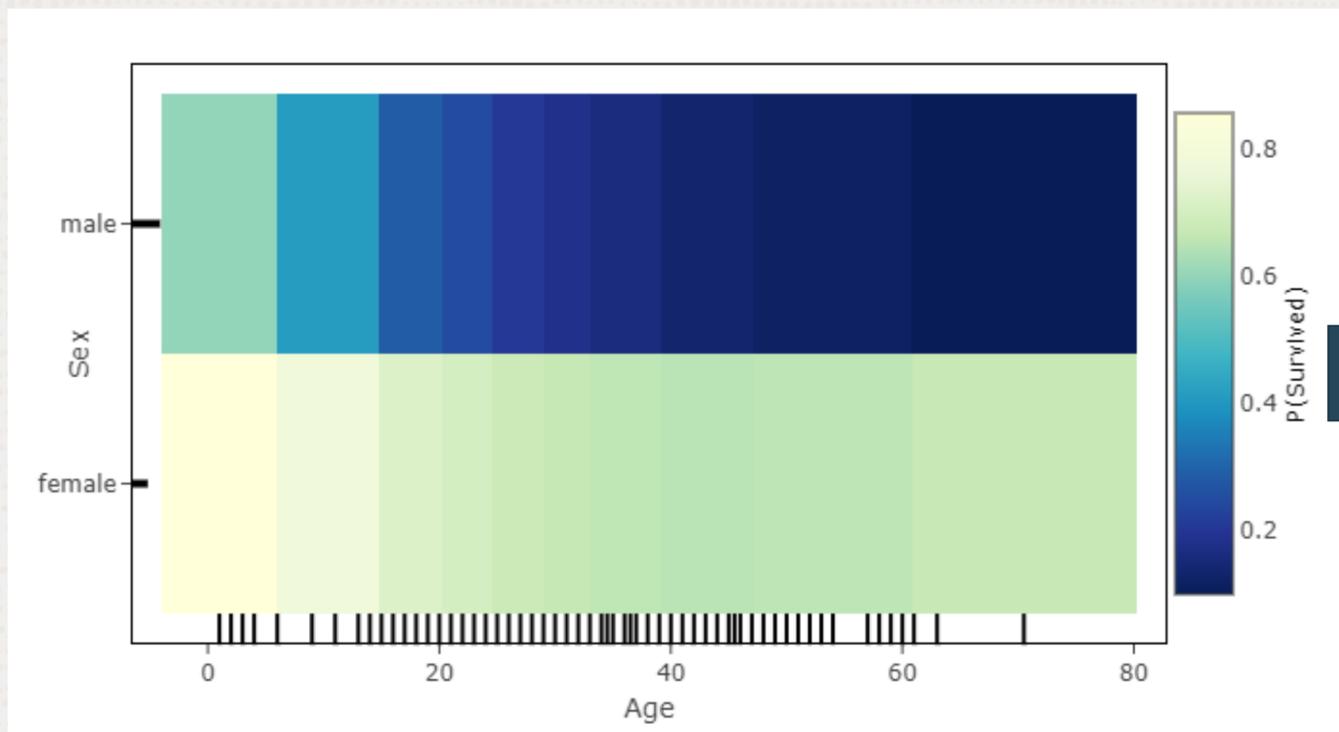
`explainer.explain_feature_dependence('age')`



Categorical feature PDP

`explainer.explain_feature_dependence('sex')`

Feature dependence examples – two features



Improved



Just specify the features!

Traditional two-feature PDP

Oracle AutoMLx two-feature PDP

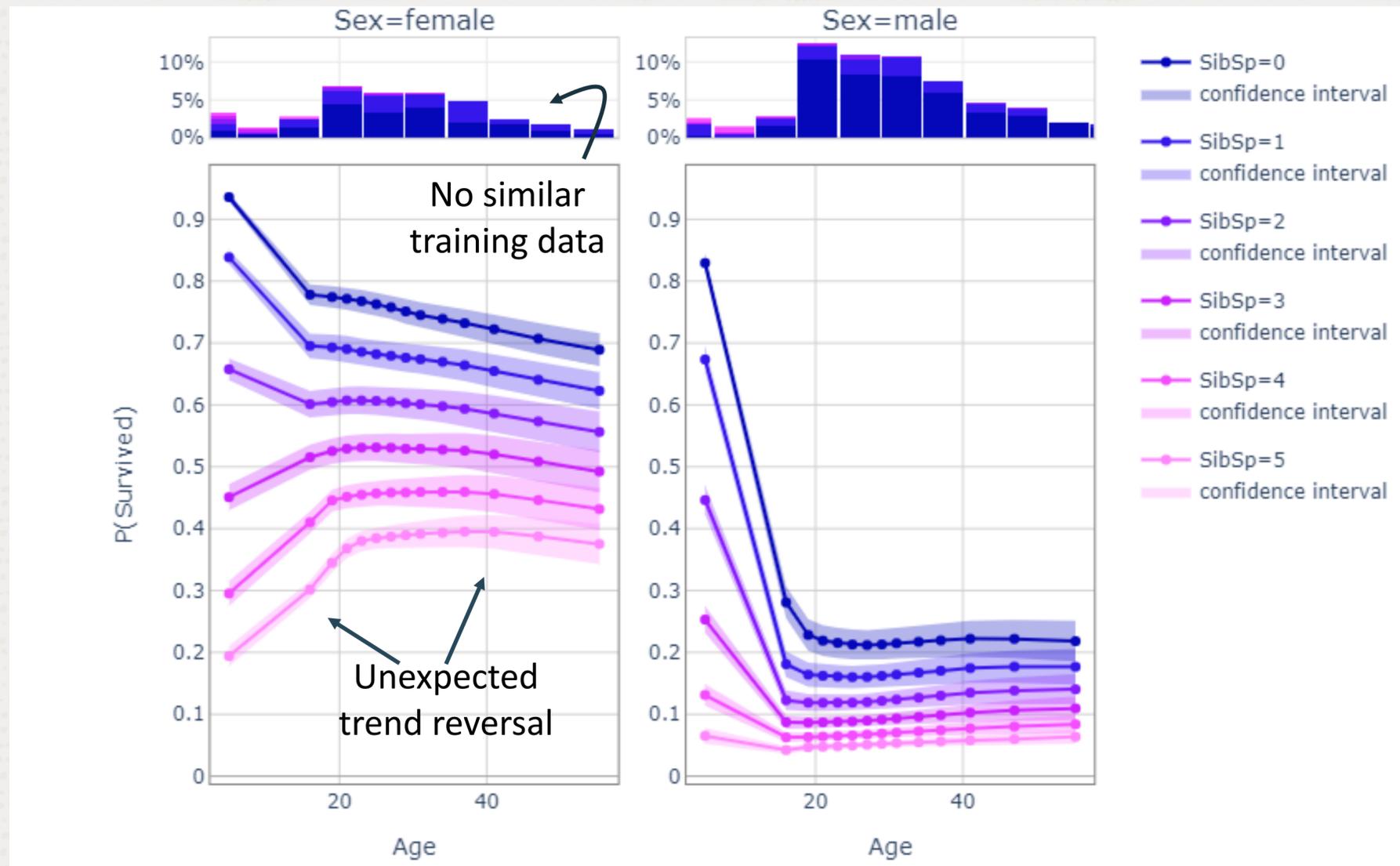
```
explainer.explain_feature_dependence(  
    ['age', 'sex'],  
)
```

Hard-to-understand heat map!
(Traditional)

Easy-to-read line chart!
(Oracle AutoMLx)



Feature dependence examples – three+ features



3-4 feature PDPs

Easy-to-read faceted line chart!
(Oracle AutoMLx)

AutoMLx available feature set

ML TASKS

Available

- Classification
- Regression
- Forecasting
- Anomaly Detection

SCORING METRICS

Optimize for any predefined scoring metric such as accuracy, F1, MSE, fairness, etc.

Optimize for any user-defined metric such as cost, throughput, etc.

ML ALGORITHMS

Classification/Regression

- Logistic/Linear Regression
- Extremely Randomized Trees
- Decision Trees
- Random Forest
- TabNet
- SVM
- LightGBM
- Naive Bayes
- Catboost
- KNN
- MLP

Anomaly Detection

- Isolation Forest
- SubspaceOD
- One Class SVM
- CLOF
- AutoEncoder
- MinCov OD
- HistogramOD
- KNN
- PCA

Forecasting

- Naive
- STLwES
- Prophet
- STLwARIMA
- Theta
- ETS
- Orbit
- ExpSmooth
- VARMAX
- DynFactor
- SARIMAX

DATA TYPES

Tabular

- Numerical, string, time (datetime, timedelta)

Text

Timeseries

- Univariate, multivariate, exogenous

EXECUTION PLATFORMS

Oracle DB

Dask

Python

- Multi-processing
- Multi-threading

ML EXPLAINABILITY

Prediction Explanations

- Permutation importance
- Shapley
- Surrogate-based (LIME+)
- Counterfactuals (FaCE, DiCE)

Model Explanations

- Permutation importance
- Shapley
- Partial dependence plots
- Individual conditional expectations
- Accumulated local effects
- Fairness feature importance

AutoMLx availability

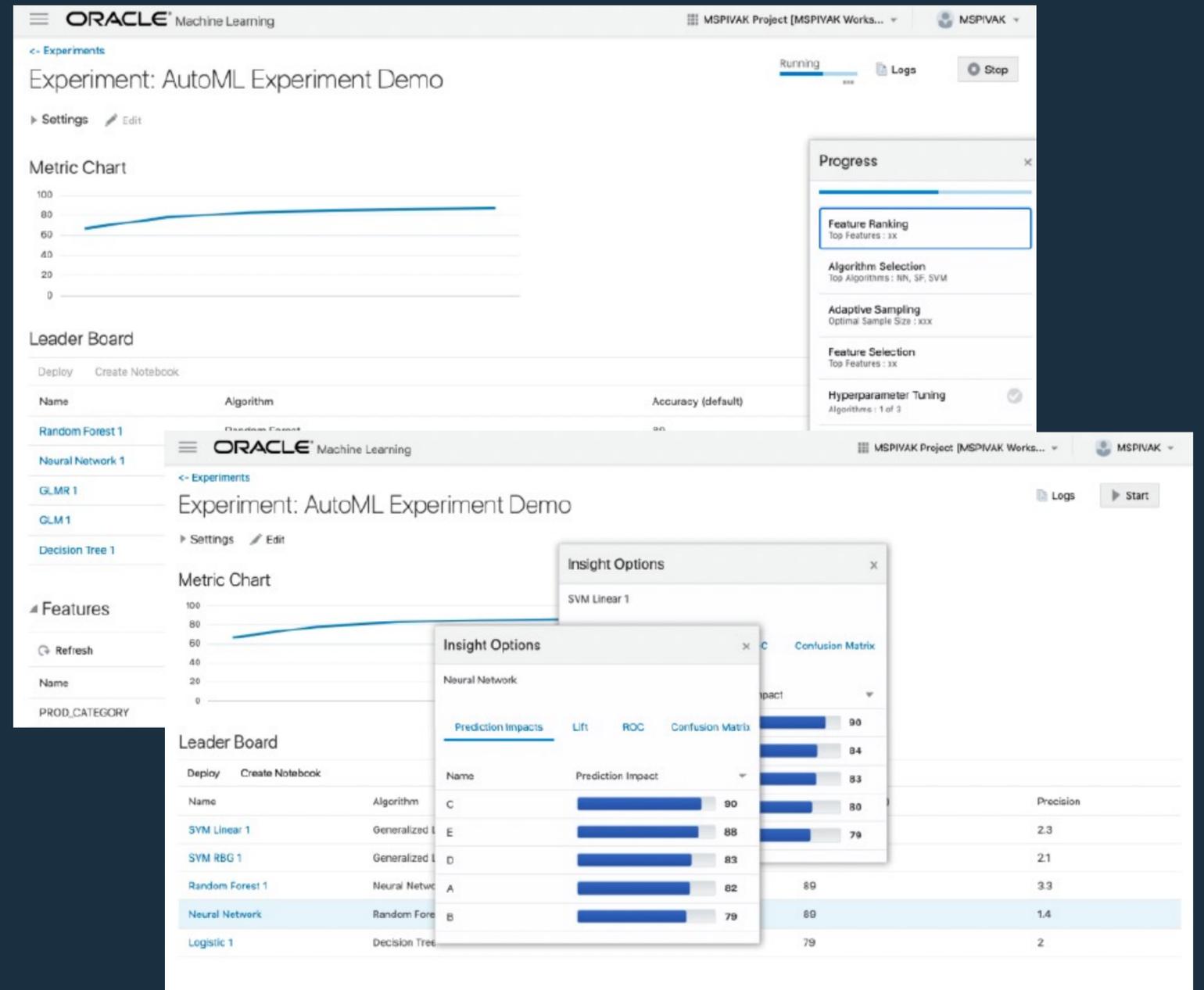
- Platforms

- Oracle Autonomous Database (OML)
 - Autonomous and on-prem
 - Graphical user interface (GUI) or notebook
- Oracle Cloud Infrastructure Data Science
 - Notebook
- MySQL DB (HeatWave ML)
 - MySQL console & notebook
- Available on Oracle's always-free cloud services

- Applications and verticals

- Oracle Transportation Management

- Others in progress



Application Example

Predicting Oyster Health



The Louisiana oyster industry

200

Years of history

40000

Jobs

\$300 million

Economic impact on Gulf States of the United States

Shapes the identity of entire communities

Foundation of the New Orleans and Gulf Coast food culture

Major impact on tourism

Image source: https://www.wlf.louisiana.gov/assets/Species_Guide/Fish_Shellfish/Images/1200x900pxOyster_1.jpg

What is dermo?



Perkinsus Marinus

Parasite causing the dermo disease in oysters.



Where

The eastern oyster is a species native to eastern North and South America.



Dermo Sentinel¹

Project aiming at assessing oyster infection along the US coast of the Gulf of Mexico.



Machine Learning

Can ML help oyster farmers in assessing the risk of dermo infection in their lots?

¹ <https://data.oystersentinel.cs.uno.edu/dermo>

A dark-themed map of North America, including the United States, Canada, and Mexico. The map is overlaid with a large white text block. In the southern United States and northern Mexico, there is a cluster of colored dots (red, orange, yellow, green) representing data points or locations. The text reads: "Attempts at eradicating the disease have proven ineffective, so prevention and timely intervention are crucial."

Attempts at eradicating the disease have proven ineffective, so prevention and timely intervention are crucial.

The dataset

Includes information about the environment of locations all around the US coast of the Gulf of Mexico, where oysters were collected and tested for the disease in the scope of the Dermo Sentinel project.¹

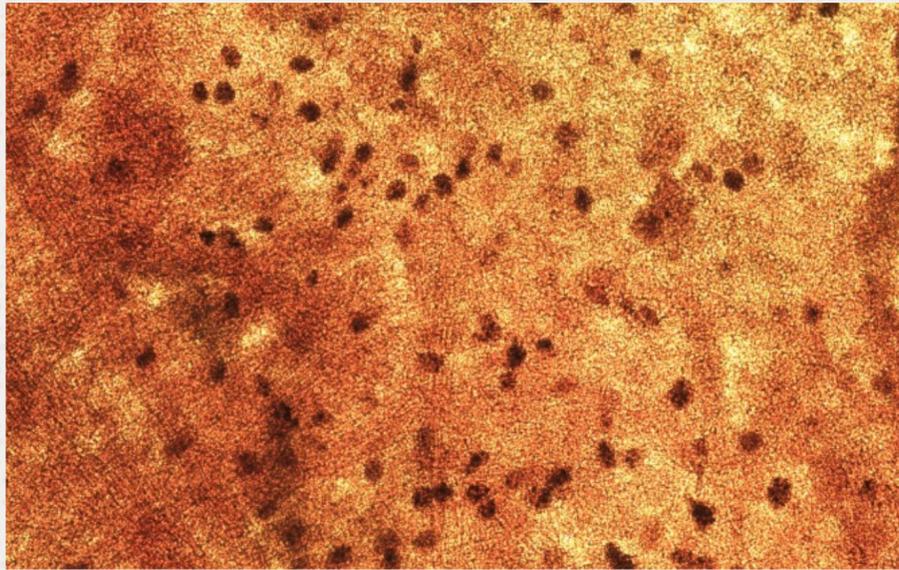
~5398 data samples, split into 90% training set and 10% test set

COLLECTION DATE	LATITUDE	LONGITUDE	TEMPERATURE (°C)	SALINITY (PPM)	JUVENILE OYSTERS	...	INFECTION INTENSITY (0.0 – 5.0)
2002-09-19	26.025936	-97.195015	28.8	40.0	False		2.26
2002-09-19	26.025936	-97.195015	28.8	40.0	True		1.899
2003-07-21	26.025936	-97.195015	31.5	36.0	False		2.266



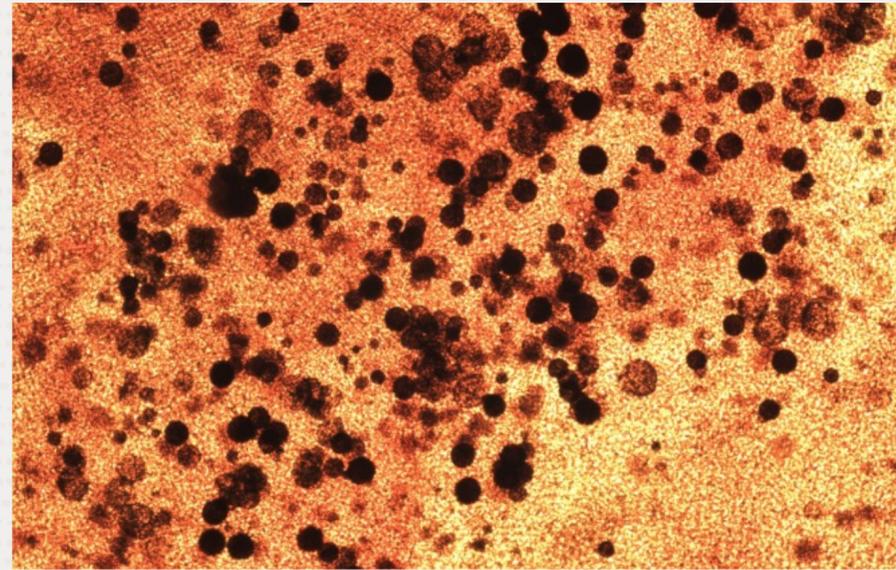
¹ <https://data.oystersentinel.cs.uno.edu/dermo>

Dermo risk assessment



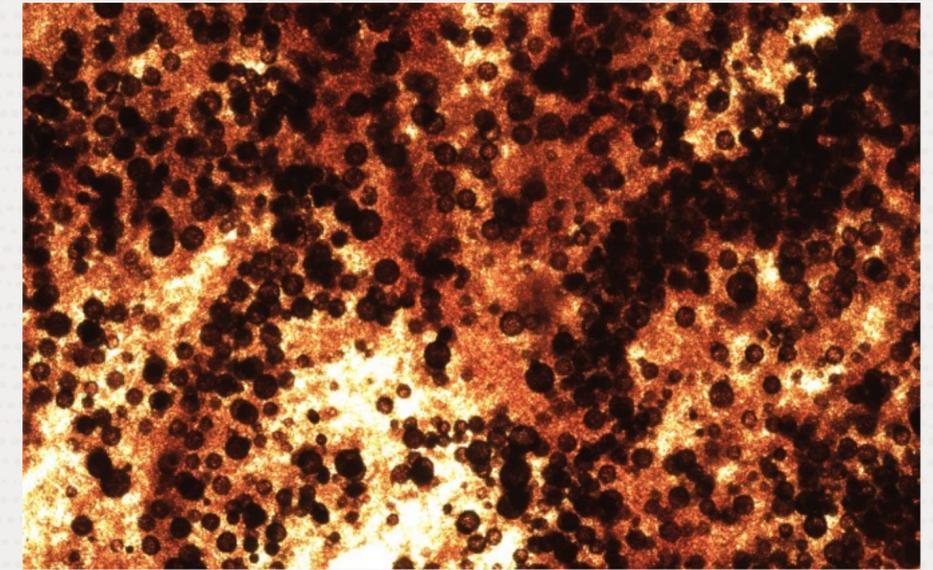
Low risk

If infection intensity is less than one.
While some infected oysters may be present, incidence of the disease is still under control.



Moderate risk

Infection intensity is between 1 and 2.
The area should be monitored closely and interventions should not be delayed.



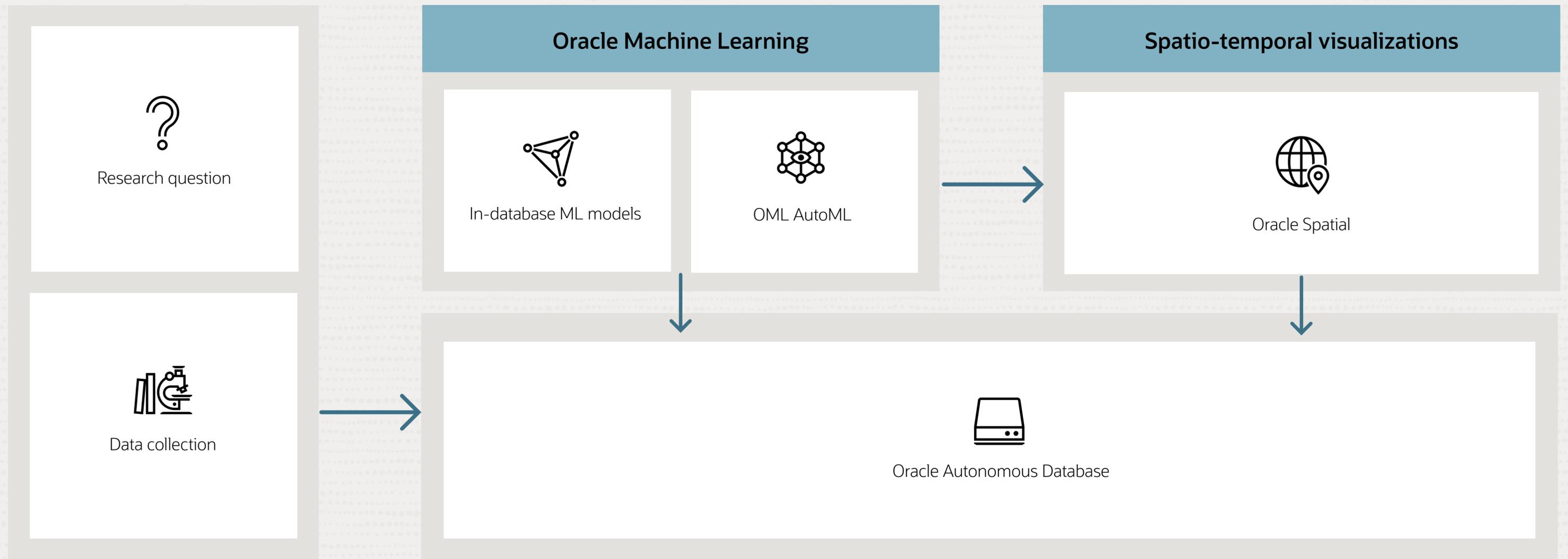
High risk

Infection intensity is above 2.
The disease has spread significantly; timely intervention is crucial to avoid further damage to the oyster population.

Image source: https://data.oystersentinel.cs.uno.edu/RFTM_SOP.pdf

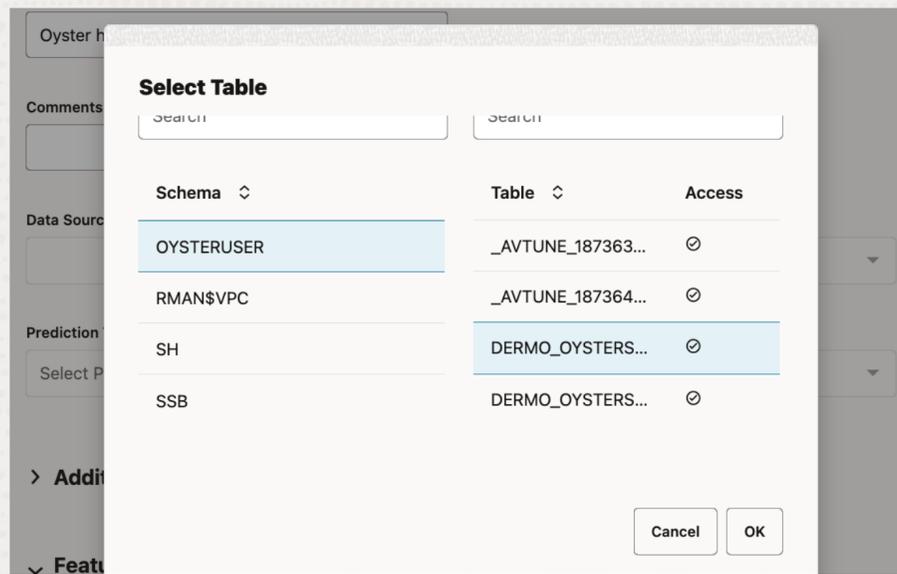
How Oracle can help

Oracle Autonomous database, spatial and graph, and machine learning technologies

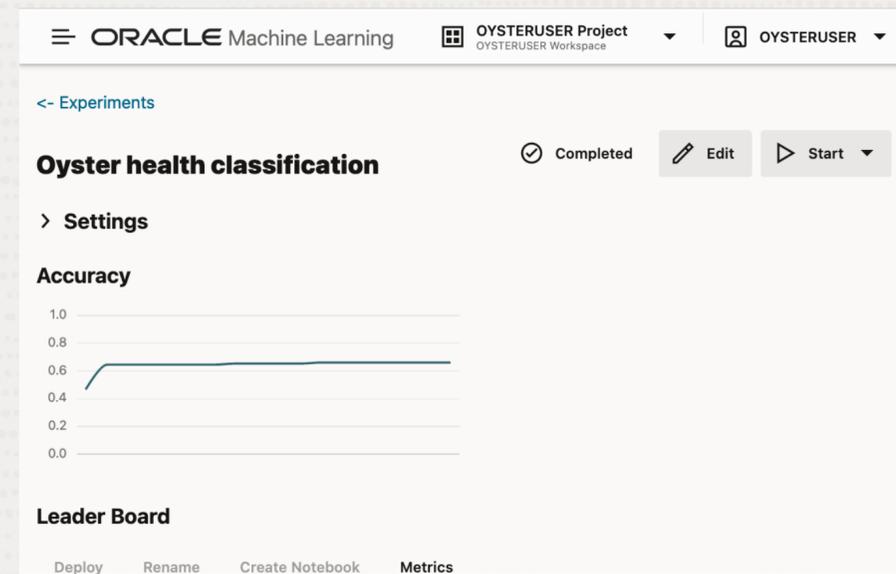


Zoom-in: OML

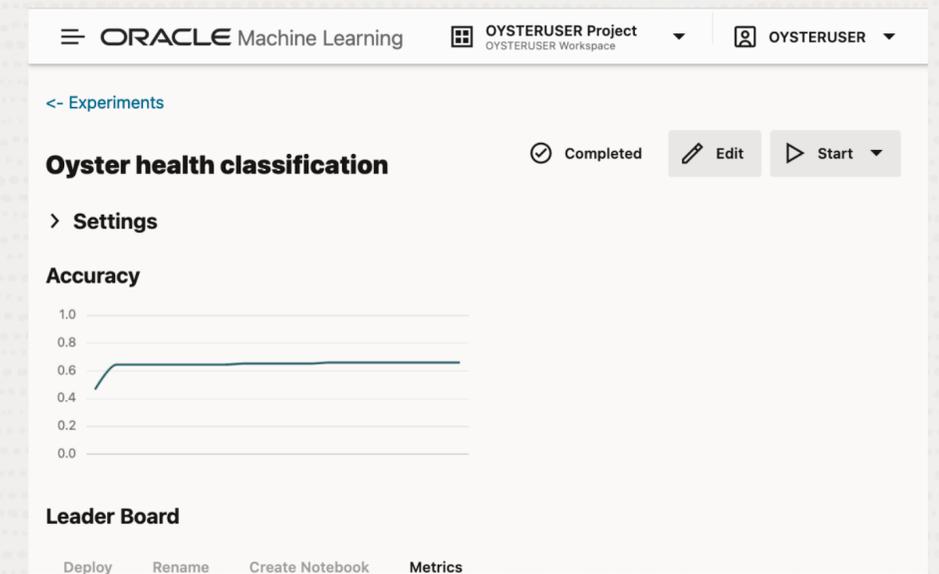
Eliminates the need to move data to dedicated machine learning systems



Load data directly from database tables



Create and manage projects with the OML AutoML UI

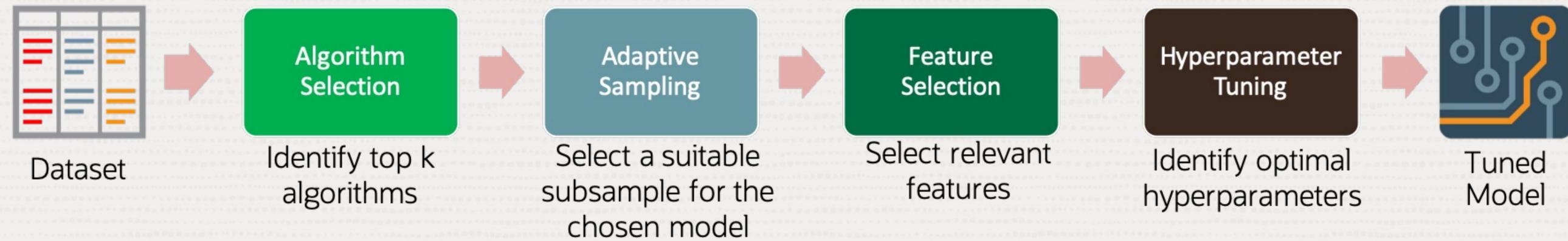


Further explore data and models with OML Notebooks

To learn more about OML: <https://docs.oracle.com/en/database/oracle/machine-learning/index.html>

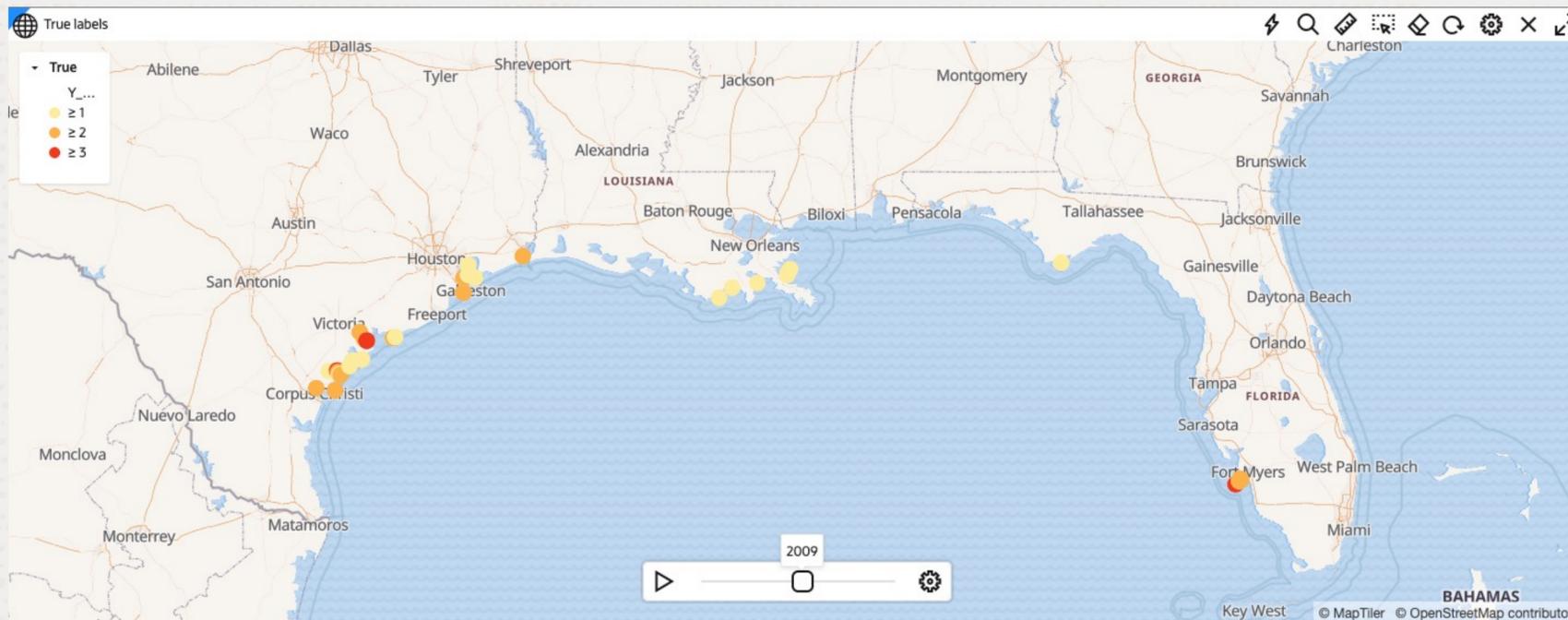
Zoom-in: AutoML

AutoML makes it easy to go from data to high performance machine learning models!



Zoom-in: Spatial Studio

Create spatio-temporal visualization of your data in the Autonomous Database



View and analyze the evolution of the Dermo disease in the area of interest from the model's predictions

Predicting Dermo Risk (Low, Medium, High)

72%

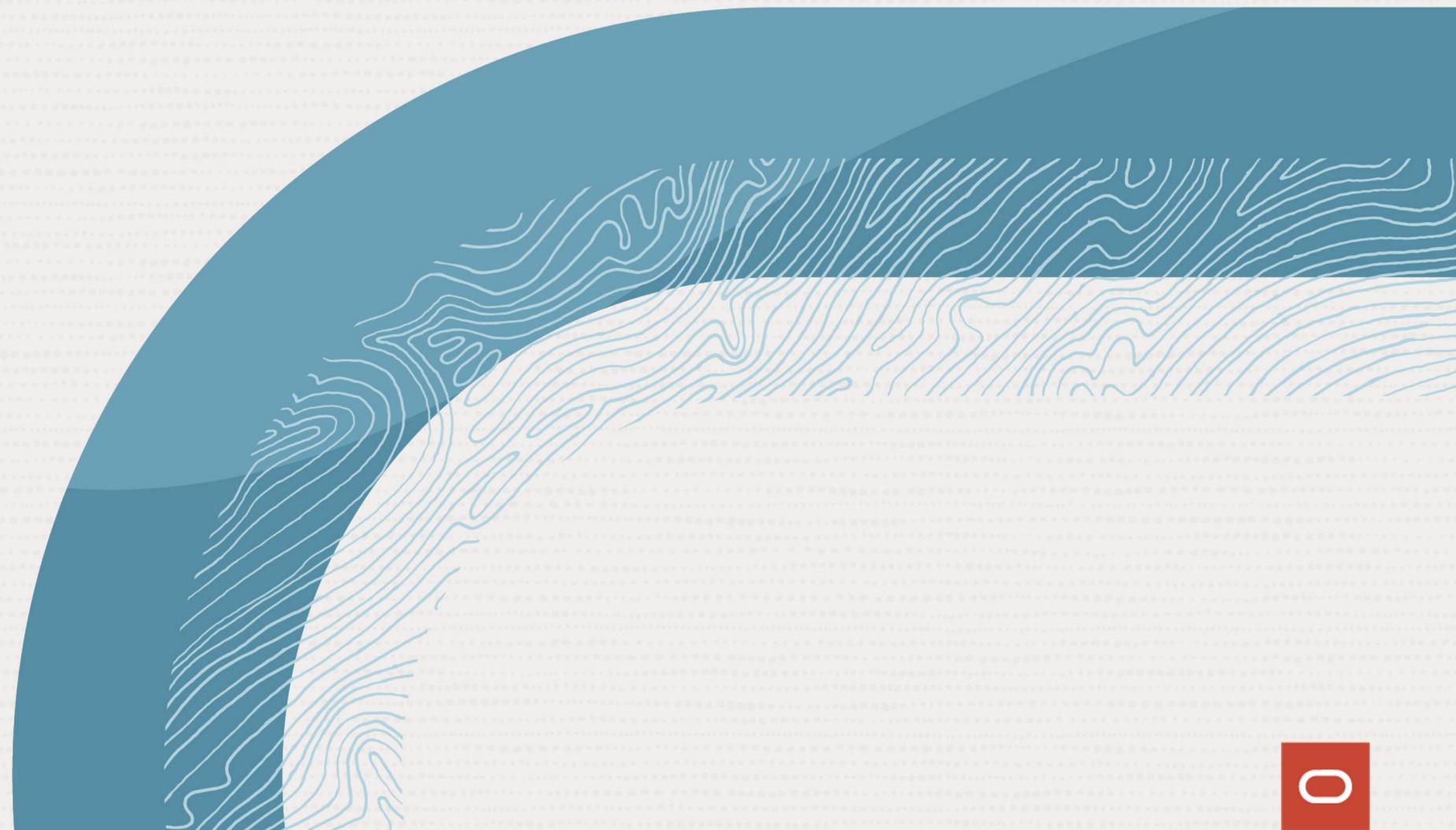
Accuracy

66%

Balanced Accuracy



Demo



Oracle Machine Learning UI – Configuration

ORACLE Machine Learning

OYSTERUSER Project
OYSTERUSER Workspace

OYSTERUSER

Experiment Settings: How are our oysters?

Start Cancel Save

Name
How are our oysters?

Comments

Data Source
OYSTERUSER.OYSTER_DERMO_TRAIN

Predict
INFECTION_INTENSITY_CAT

Prediction Type
Classification

Case ID
COLLECTION_ID

1. Select training data from database table
2. Select target column
3. ML task is inferred from target column
4. Select sample ID column

Oracle Machine Learning UI – Experiment overview

The screenshot displays the Oracle Machine Learning interface for an experiment titled "OYSTERUSER Project". The main heading is "How are our oysters?". The interface includes a "Settings" section, a "Balanced Accuracy" line graph, a "Leader Board" table, and a "Completed" modal window.

Balanced Accuracy Graph: The graph shows a line starting at approximately 0.35, rising to 0.6 by the 0.2 mark on the x-axis, and then remaining flat at 0.6 for the rest of the duration.

Leader Board Table:

Algorithm	Model Name	Balanced Accuracy
Random Forest	RF_C5554D6AFC	0.6356
Support Vector Machine (Gaus...	SVMG_E88D2FBD00	0.6157
Decision Tree	DT_B33A4A0A21	0.5921
Naive Bayes	NB_6E5254CCF7	0.5911

Completed Modal: Shows the experiment is completed in 0h 36m. The steps listed are: Initialization Completed, Algorithm Selection Completed, Adaptive Sampling Completed, Feature Selection Completed, Model Tuning Completed, and Random Forest Completed.

1. AutoML pipeline progress
2. Model score improvements
3. Model leaderboard, details and actions



Oracle Machine Learning – Notebooks

ORACLE Machine Learning

OYSTERUSER Project
OYSTERUSER Workspace

OYSTERUSER

Score model on data

FINISHED

```
%python
mod_predict = rf_mod.predict(test_data ,supplemental_cols = test_data[:, ['INFECTION_INTENSITY_CAT']]).pull()
y_true = mod_predict['INFECTION_INTENSITY_CAT']
y_pred = mod_predict['PREDICTION']
```

Took 1 sec. Last updated by OYSTERUSER at October 12 2022, 2:54:57 PM. (outdated)

Show model quality metric

FINISHED

```
%python
import sklearn as skl
balanced_acc_score = skl.metrics.balanced_accuracy_score(y_true, y_pred)
acc_score = skl.metrics.accuracy_score(y_true, y_pred)
print("Balanced accuracy:", balanced_acc_score.round(4))
print("Accuracy:", acc_score.round(4))
```

Balanced accuracy: 0.6752
Accuracy: 0.7352

Took 0 secs. Last updated by OYSTERUSER at October 12 2022, 2:54:57 PM. (outdated)

Compute attribute importances for each test sample

FINISHED

```
%python
descr = rf_mod.predict(test_data, supplemental_cols = test_data[:, ['COLLECTION_ID', 'INFECTION_INTENSITY_CAT']], topN_attrs=5).pull()

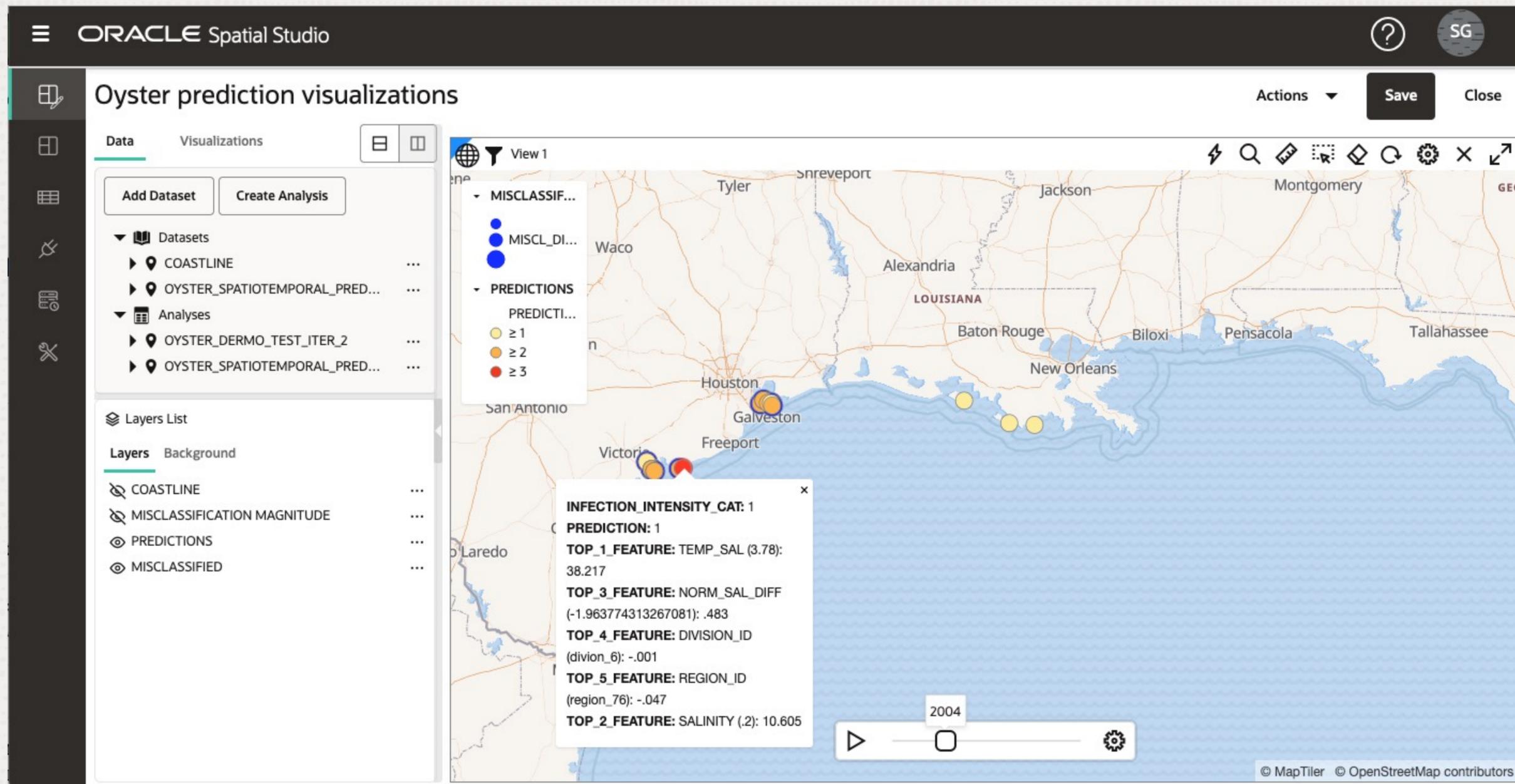
for i in range(1,6):
    feat_name = f'NAME_{i}'
    feat_value = f'VALUE_{i}'
    feat_weight = f'WEIGHT_{i}'
    descr[f"TOP_{i}_FEATURE"] = descr.apply(lambda x: f"{x[feat_name]} ({x[feat_value].round(3) if isinstance(x[feat_value], float) else x[feat_value]}): {x[feat_weight]}", axis=1)
    descr = descr.drop(columns=[feat_name, feat_value, feat_weight])

oml.create(descr, "OYSTER_DERMO_PREDICTIONS")

z.show(descr)
```

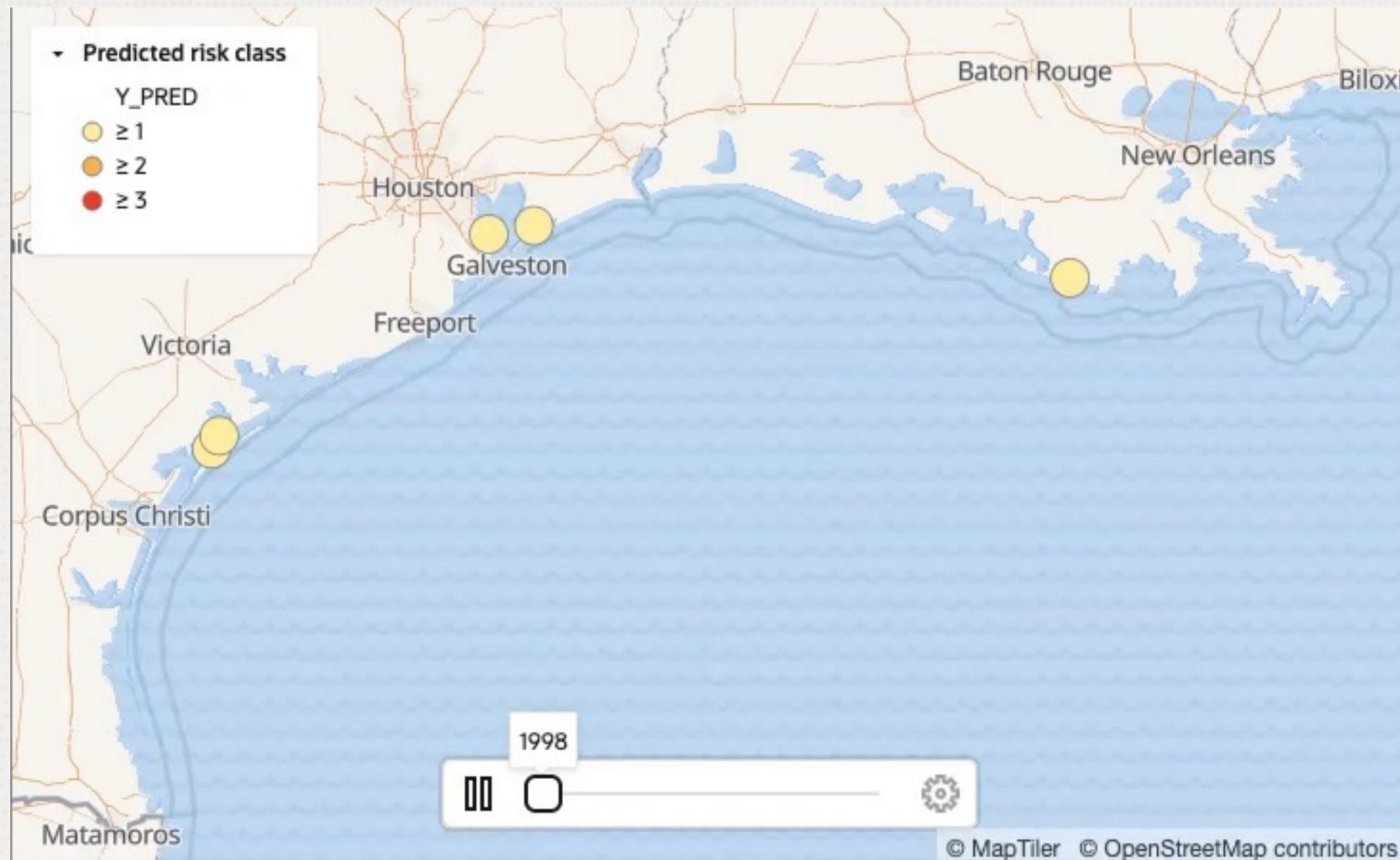
1. Generate code to re-create the model tuned by the AutoML pipeline
2. Score model on held-out test set
3. Compute feature importances for prediction explainability

Oracle Spatial Studio – Visualization



1. Visualize evolution of predictions over time
2. Review most important features used by the model for each prediction

Oracle Spatial Studio – Visualization



Oracle Spatial Studio – Spatial analysis

The screenshot displays the Oracle Spatial Studio interface for 'Oyster prediction visualizations'. The top navigation bar includes the Oracle logo, 'ORACLE Spatial Studio', a help icon, and a user profile icon 'SG'. The main content area is divided into a left sidebar and a main workspace. The sidebar contains 'Data' and 'Visualizations' tabs, with 'Data' selected. It includes buttons for 'Add Dataset' and 'Create Analysis', a 'Datasets' list with items like 'OYSTER_DERMO_TEST', 'OYSTER_DERMO_TRAIN', 'OYSTER_SPATIOTEMPORAL_PRED...', and 'US_COASTLINE', an 'Analyses' list with 'OYSTER_DERMO_TEST_V2', and a 'Layers List' with 'Background' and 'Layers' sections containing 'US_COASTLINE', 'OYSTER_DERMO_TEST', 'OYSTER_DERMO_TRAIN', 'MISCLASSIFICATION MAGNITUDE', 'PREDICTIONS', and 'MISCLASSIFIED'. The main workspace shows 'View 3' with a table of data and 'View 1' with a map. The table has columns: SAL, NORM_TEMP_DIFF, NORM_SAL_DIFF, COLLECTED_DAY, INFECTION_INTENSITY_CAT, GEOG_ID, and DISTANCE_FROM_SHORE. The map shows a coastal area with blue dots representing oyster samples and a red line representing the coastline. A red rounded rectangle highlights the 'DISTANCE_FROM_SHORE' column in the table.

SAL	NORM_TEMP_DIFF	NORM_SAL_DIFF	COLLECTED_DAY	INFECTION_INTENSITY_CAT	GEOG_ID	DISTANCE_FROM_SHORE
	-0.1639292069664988	-0.3326610444228901	4	2	20792	67.7943823925829
	-0.3626729501084615	-0.4717134182430388	29	3	21823	828.039903595077
	-0.343261158211728	-0.0900726163070113	21	3	21846	607.029718695044
	0.0995583631311756	0.2495481647851282	17	3	21823	599.895381940489
	-0.4576544267747705	0.267267096483009	31	2	21886	693.855494564001
	0.3155921933952339	-0.3284641310359706	7	1	21835	2229.52241372219
	1.45477140310347	1.4572040277017	10	1	21884	2187.05007845407

1. Compute minimum distance of dataset samples from shoreline to use as additional feature for future iterations of the model



Acknowledgements

This work was done in collaboration with and wouldn't be possible without significant contributions from:

- Dr. Thomas Soniat and his team from the University of New Orleans
- Krishna Shah, ML intern, and Giulia Carocari, Member of Technical Staff, Oracle Labs
- Hans Viehmann, Product Manager, and Ryota Yamanaka, Regional Product Manager, Spatial and Graph

ORACLE
CloudWorld
Thank you

Feel free to reach out to me with your questions!

Hesam Fathi Moghadam, Senior Manager
hesam.fathi.moghadam@oracle.com