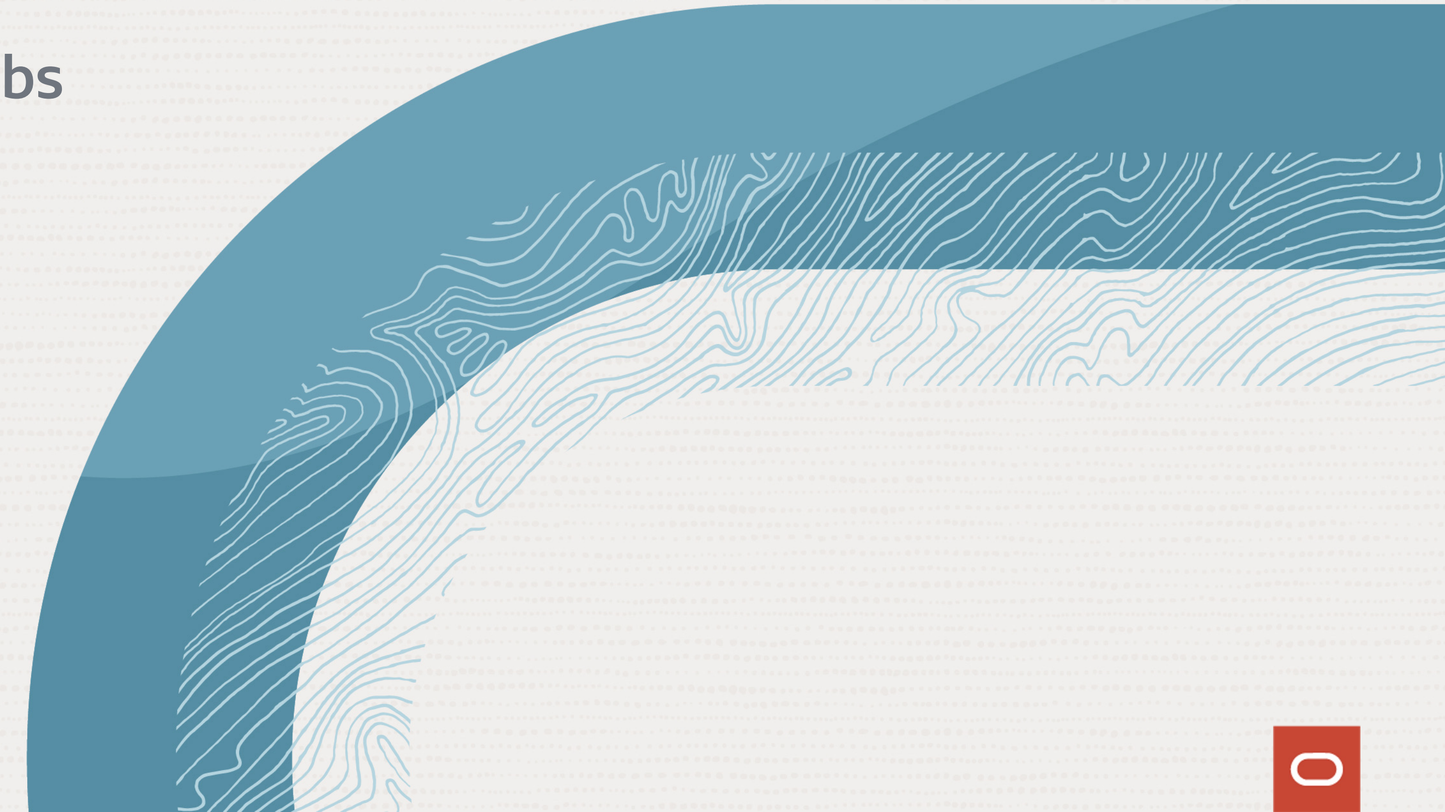
AutoML on the Half Shell: How are our Oysters?

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

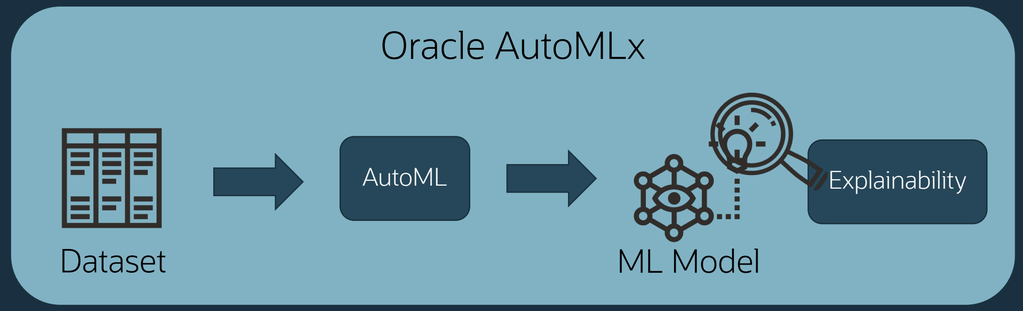
The data scientist pipeline

Creating a model:

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

Using a model:

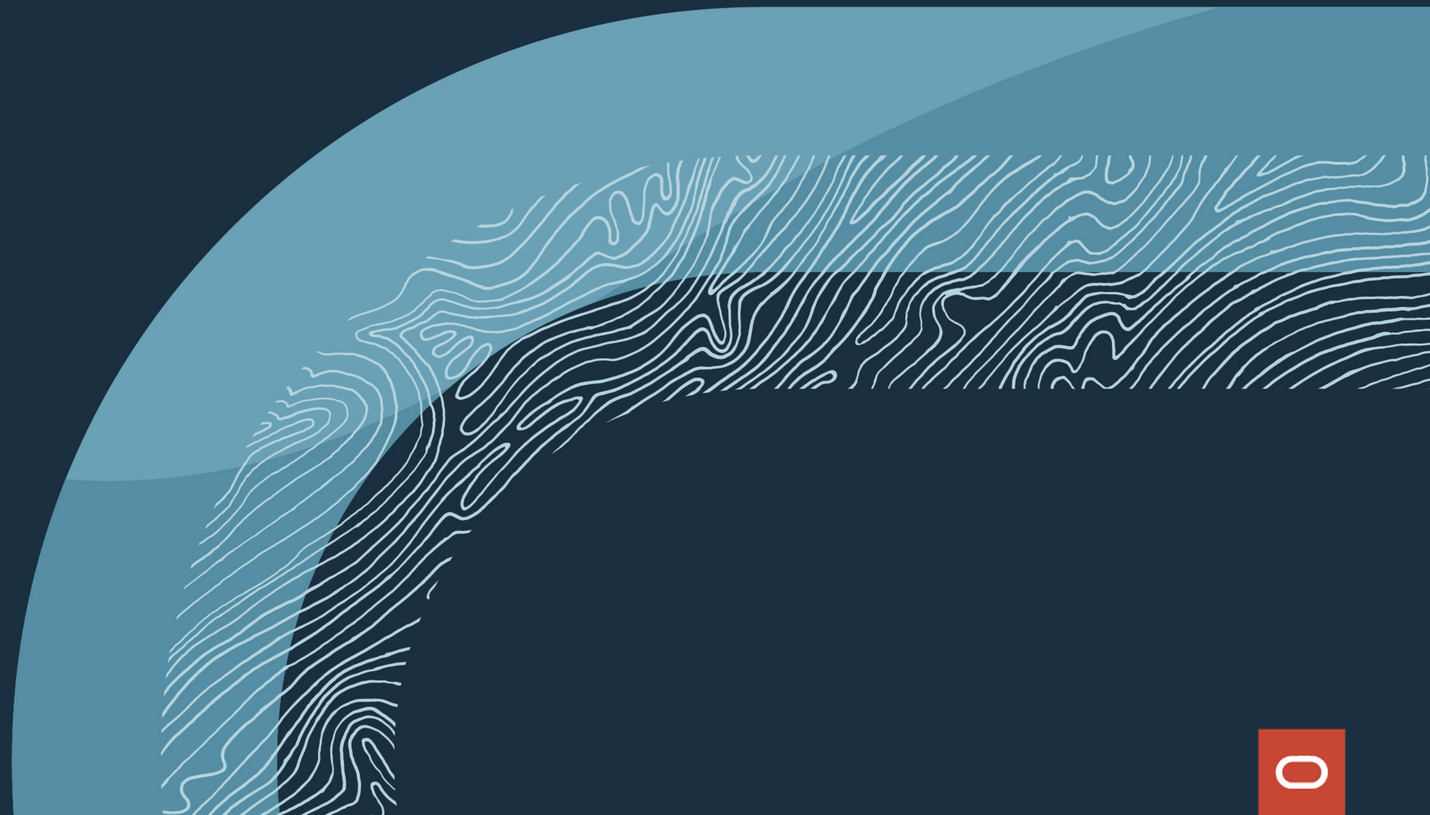
- Can I trust my model?
- Is my model “fair”?
- Does it meet regulatory requirements?



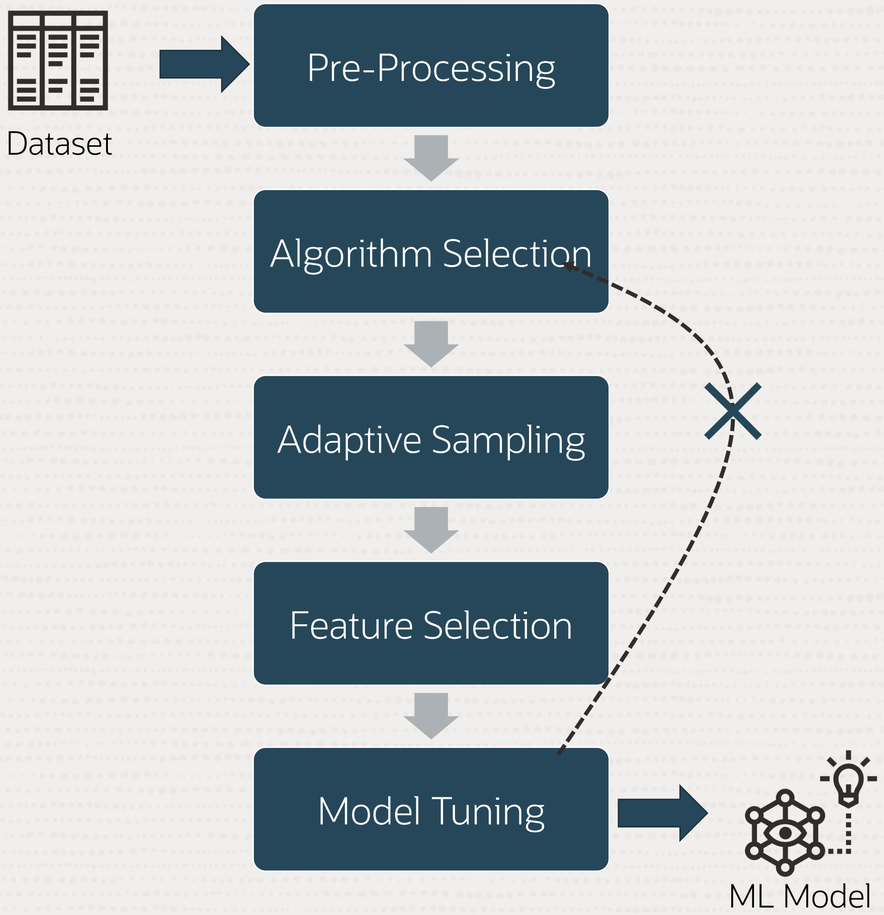
AutoMLx

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



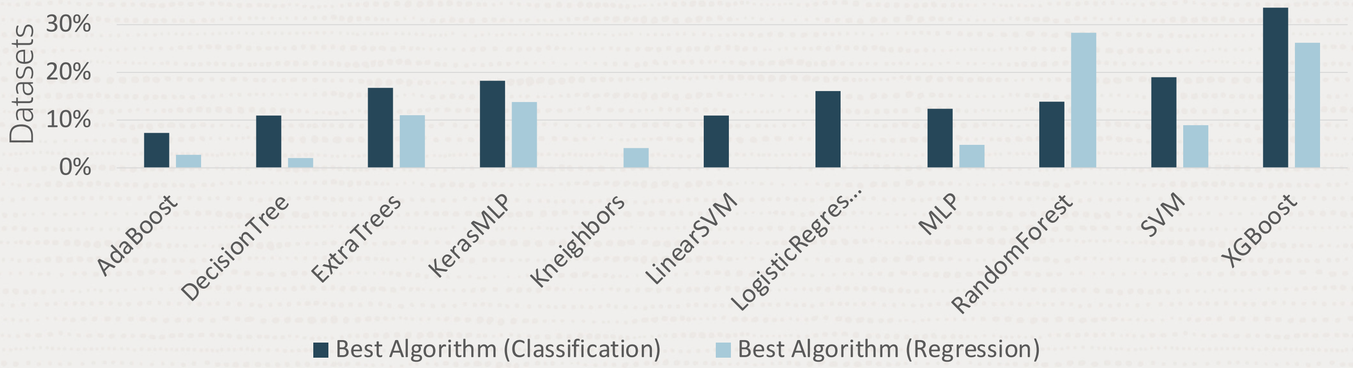
Traditional AutoML uses:

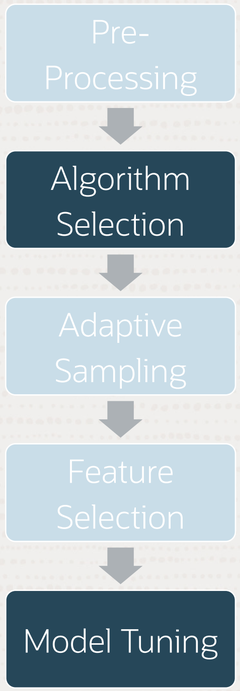
- Combined algorithm selection and hyper-parameter configuration

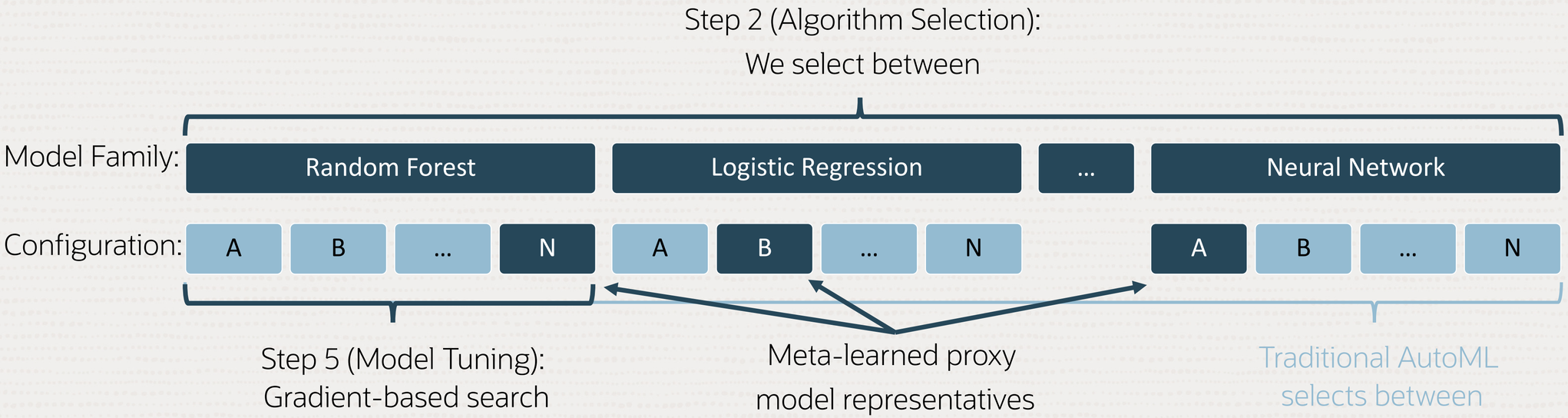
Our secret sauce?

- We never look back!

Algorithm selection & model tuning



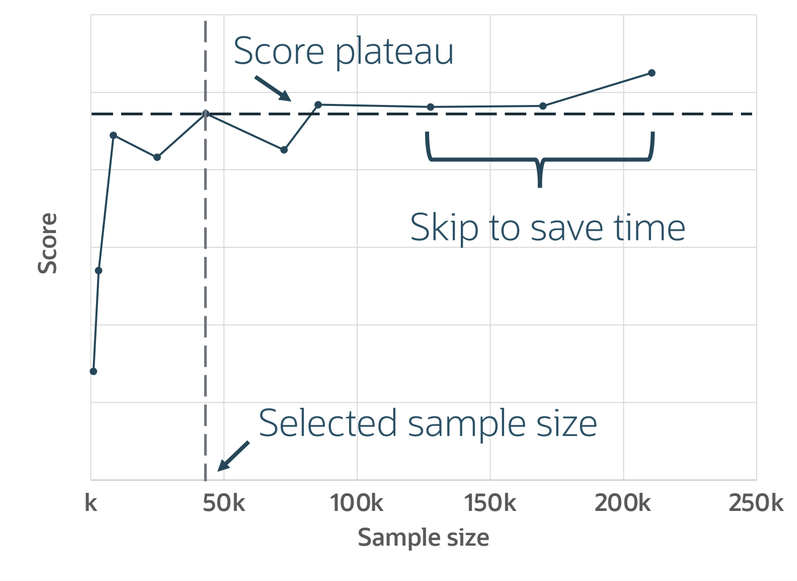




Adaptive data reduction

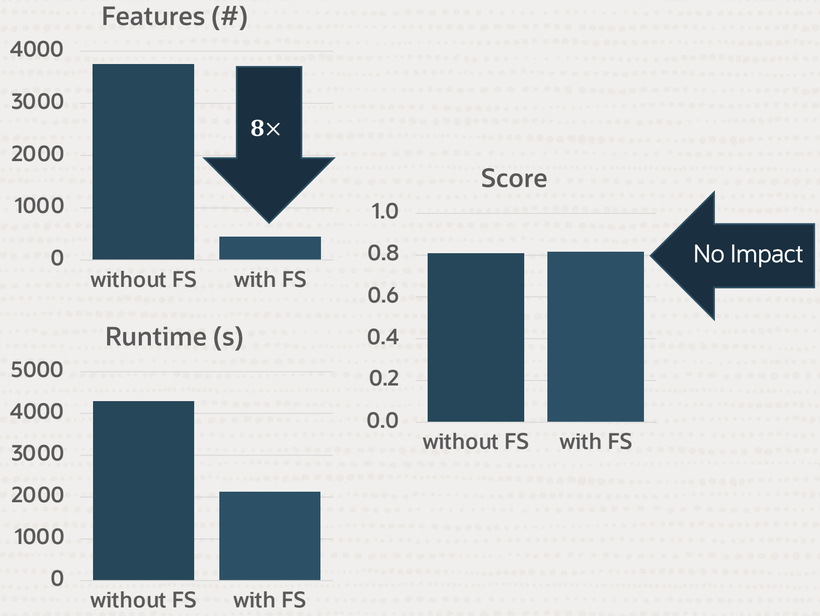
Adaptive Sampling

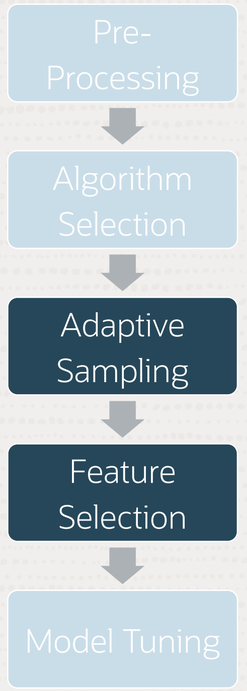
- Subsample rows for faster training
- Speeds up model search



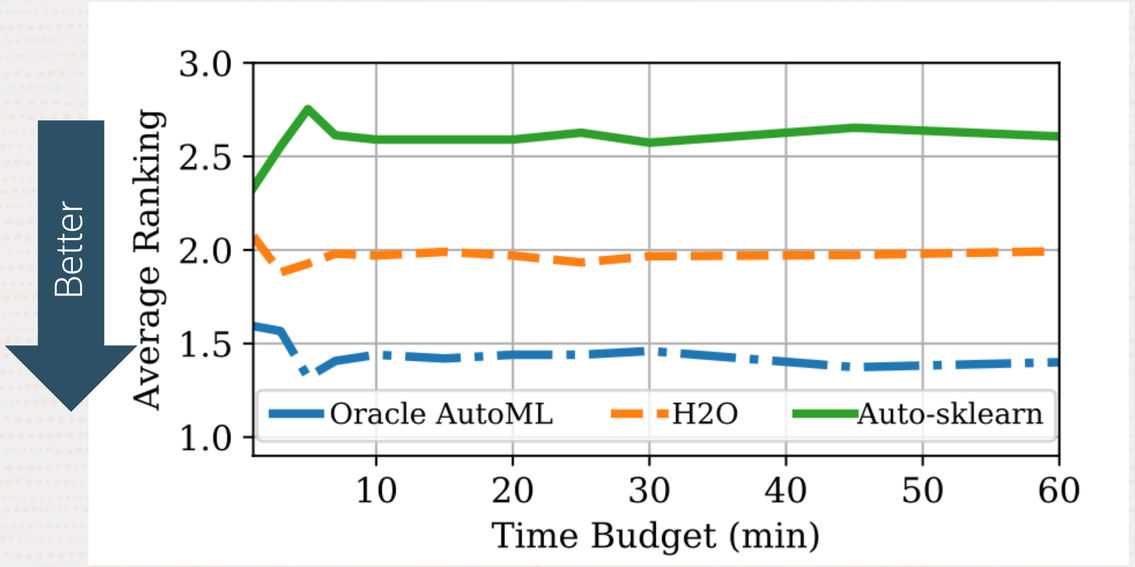
Feature Selection

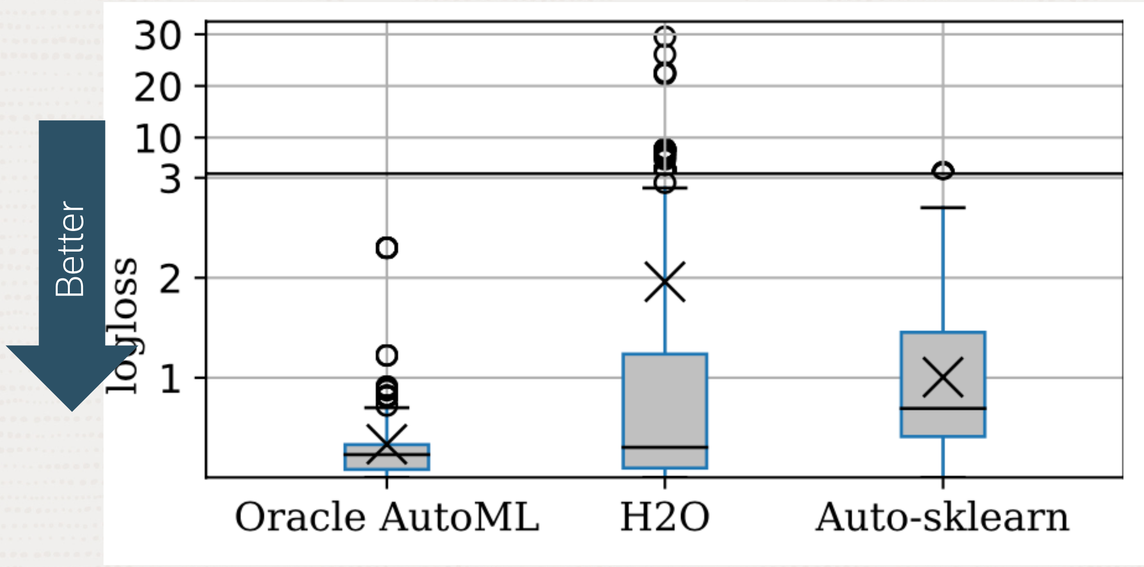
- Subsample columns for faster training
- Can also reduce overfitting





Oracle AutoML Benchmarking





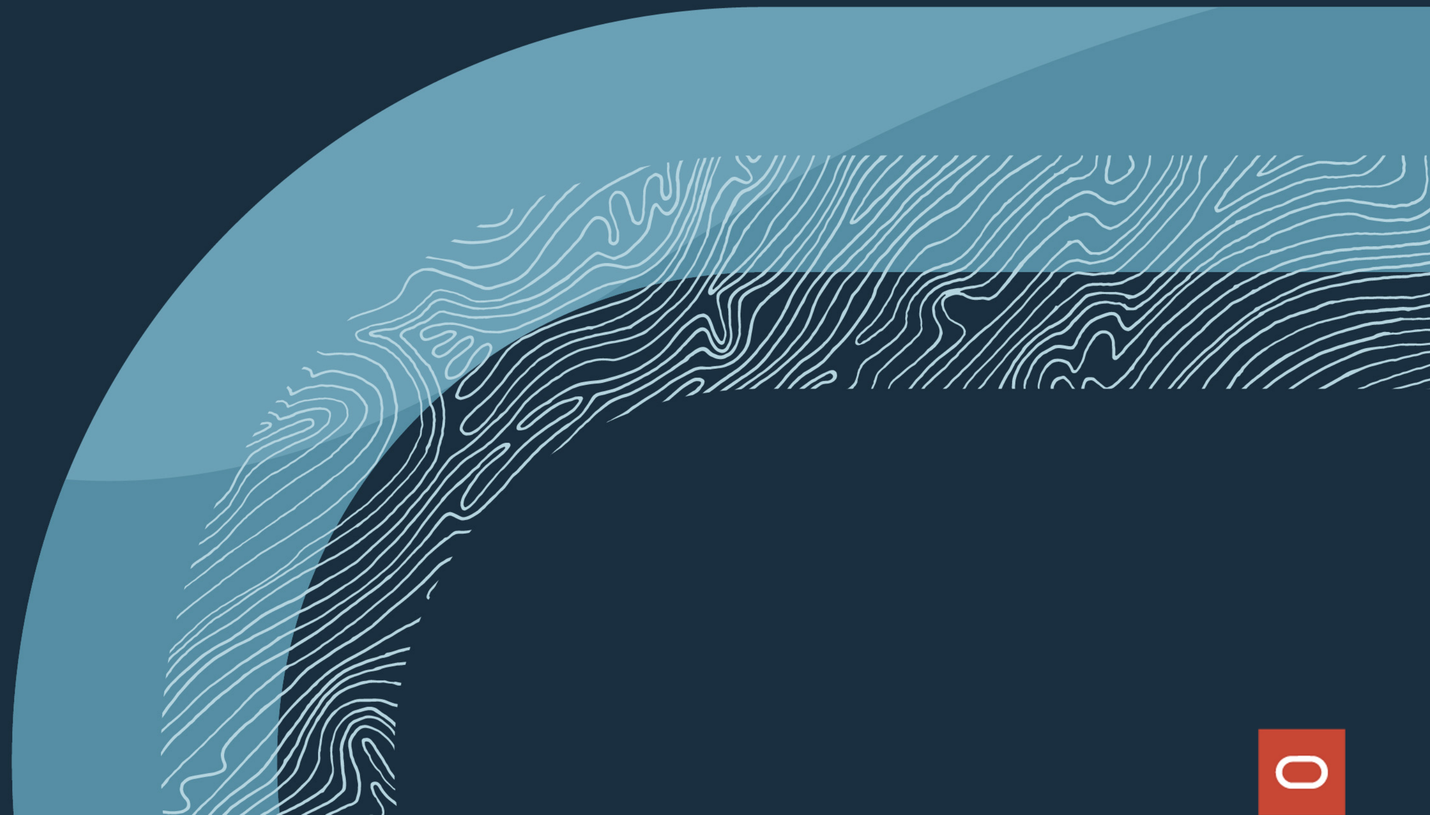
3.5 – 4× faster
and better scores

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

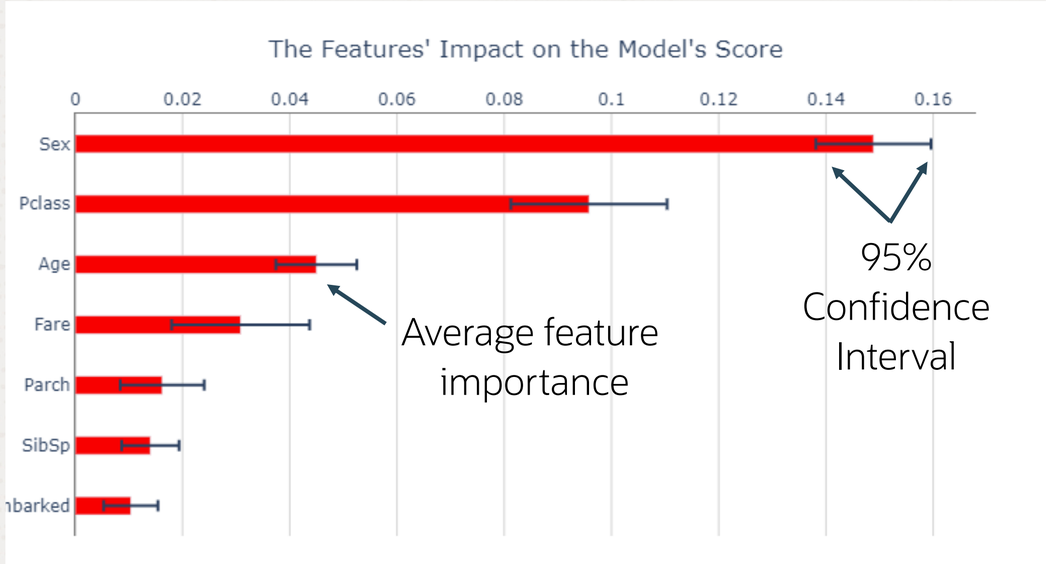
AutoMLx

Easy-to-use interface!

```
from automl import MLExplainer

# Can be an AutoML pipeline or scikit-learn model
explainer = MLExplainer(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)
```

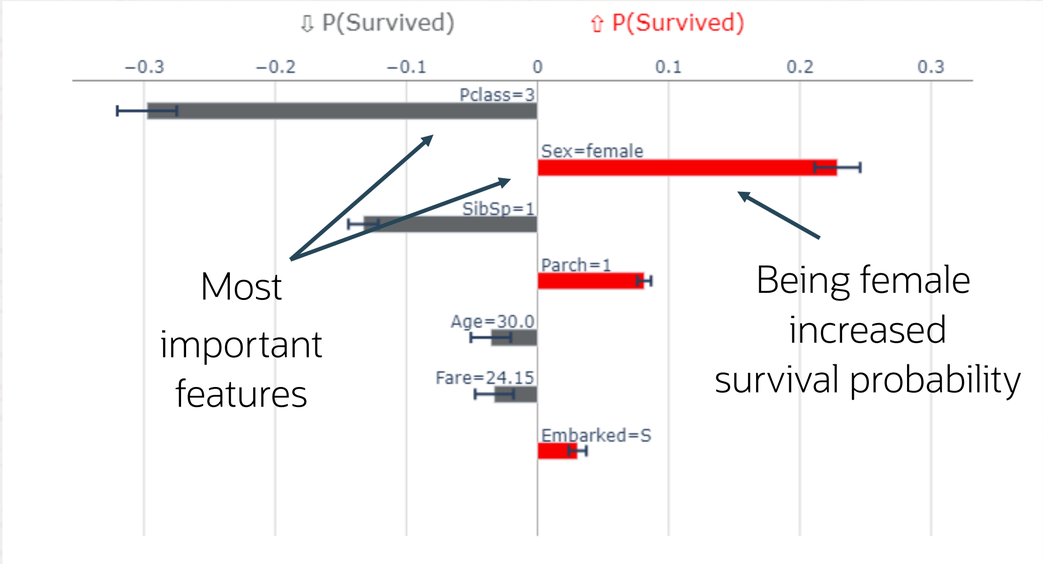


Feature importance examples – titanic dataset



Global (model) feature importance

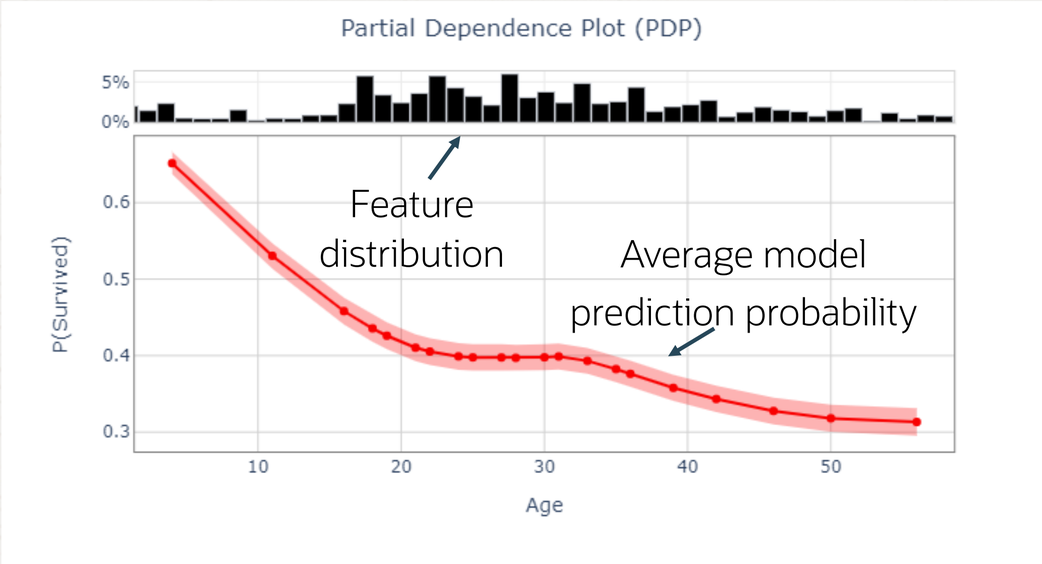
explainer.explain_model()



Local (prediction) feature importance

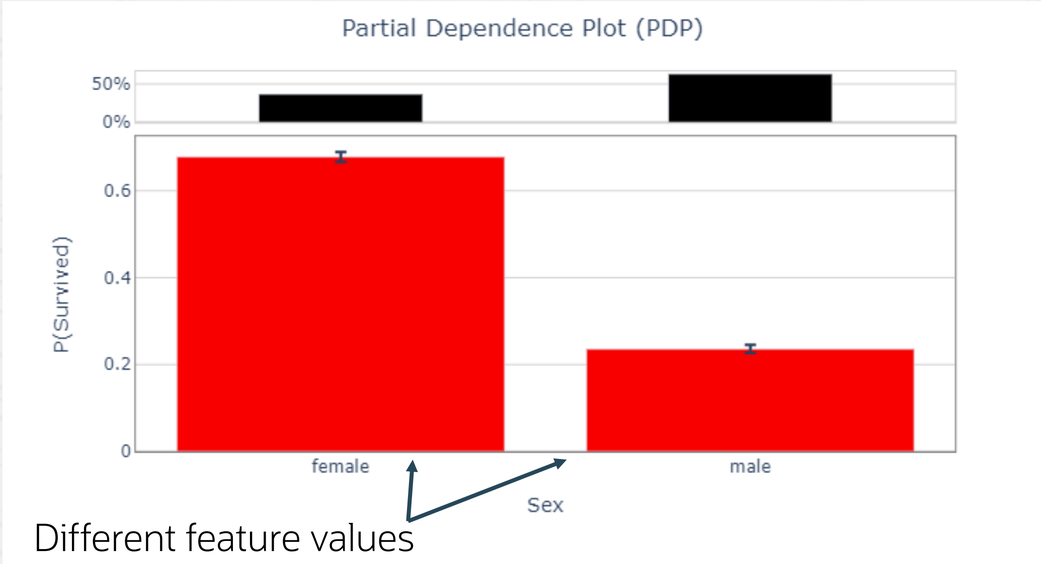
explainer.explain_prediction(X_test)

Feature dependence examples – titanic dataset



Continuous feature PDP

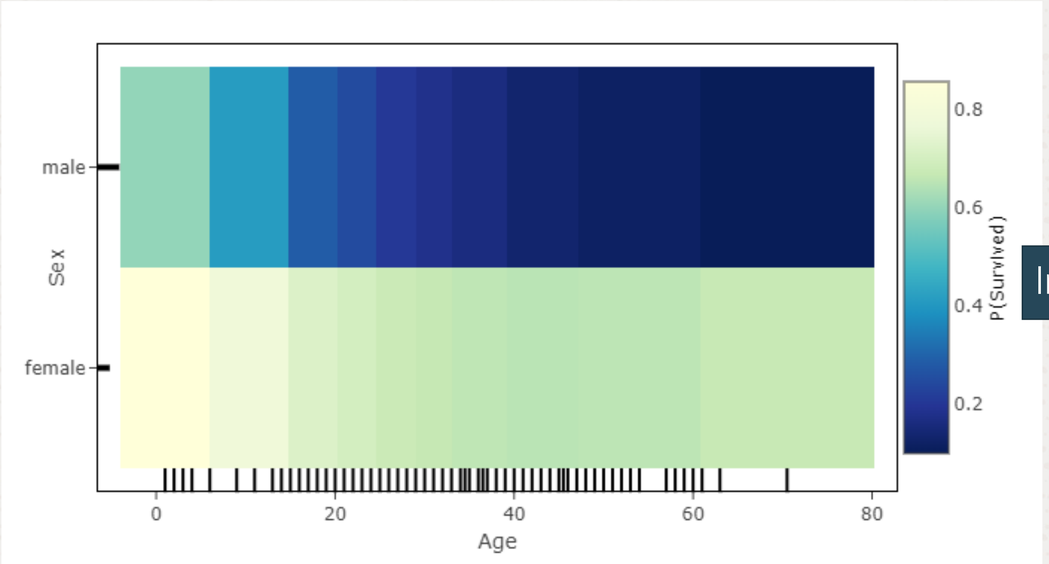
explainer.explain_feature_dependence('age')



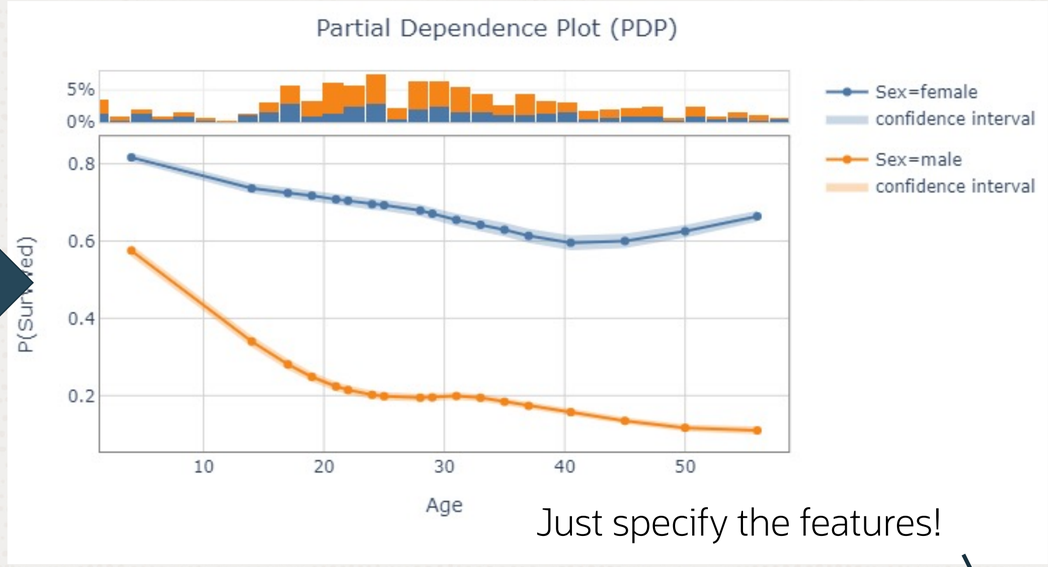
Categorical feature PDP

explainer.explain_feature_dependence('sex')

Feature dependence examples – two features



Improved



Just specify the features!

Traditional two-feature PDP

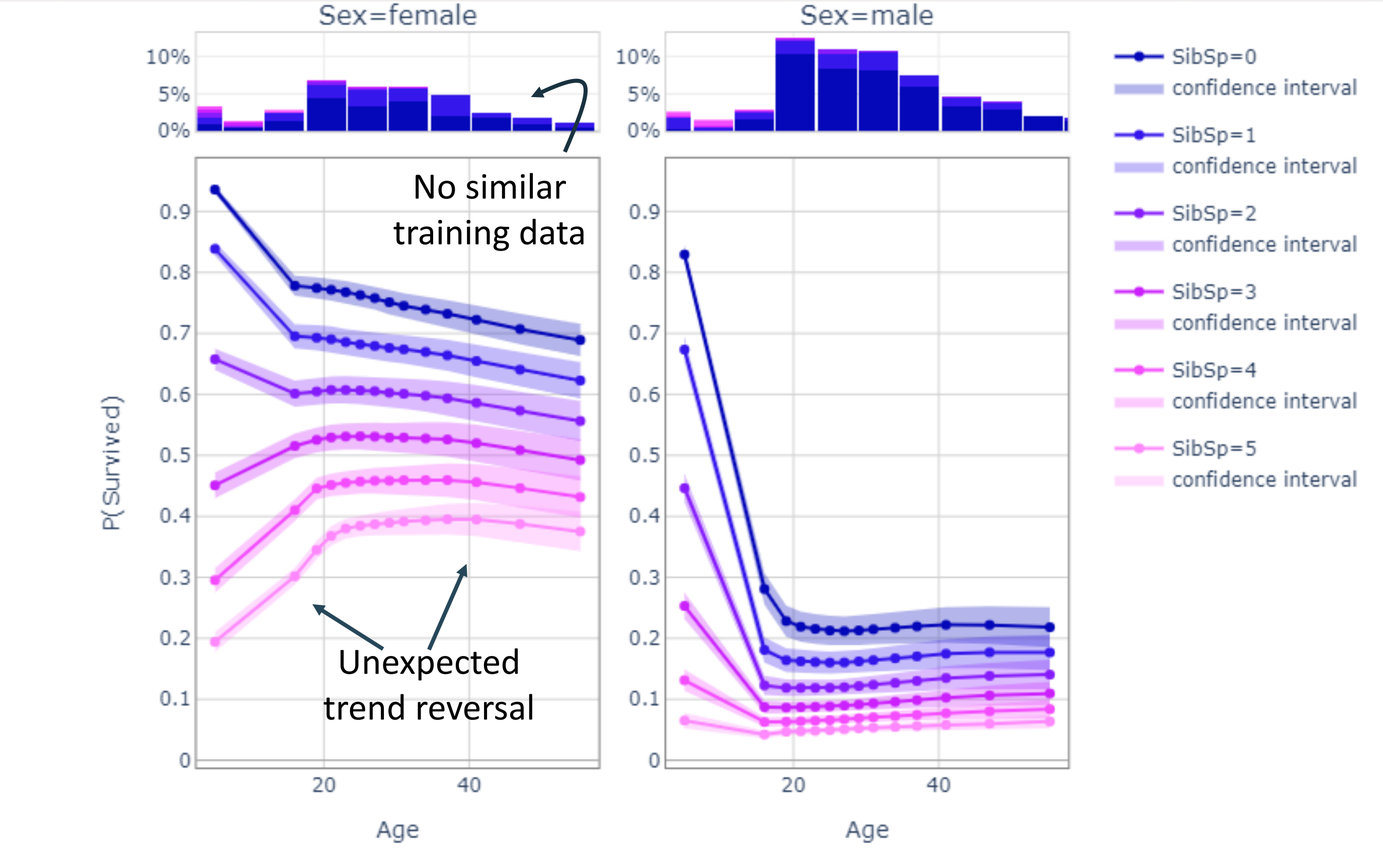
Hard-to-understand heat map!
(Traditional)

Oracle AutoMLx two-feature PDP

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

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

Feature dependence examples – three+ features



3-4 feature PDPs

Easy-to-read facetted 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
- Naïve Bayes
- XGB
- KNN
- Catboost
- 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

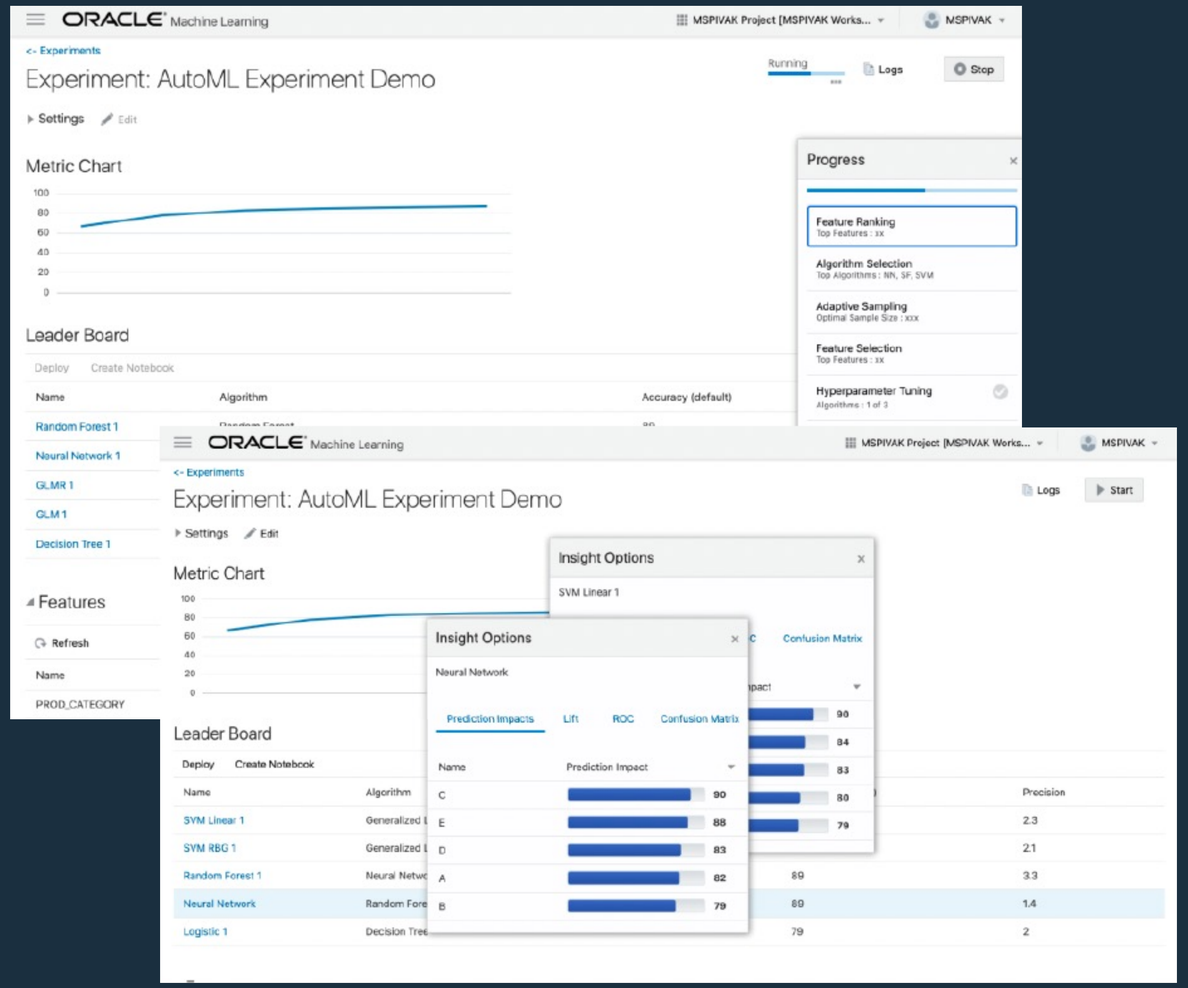
ML EXPLAINABILITY

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

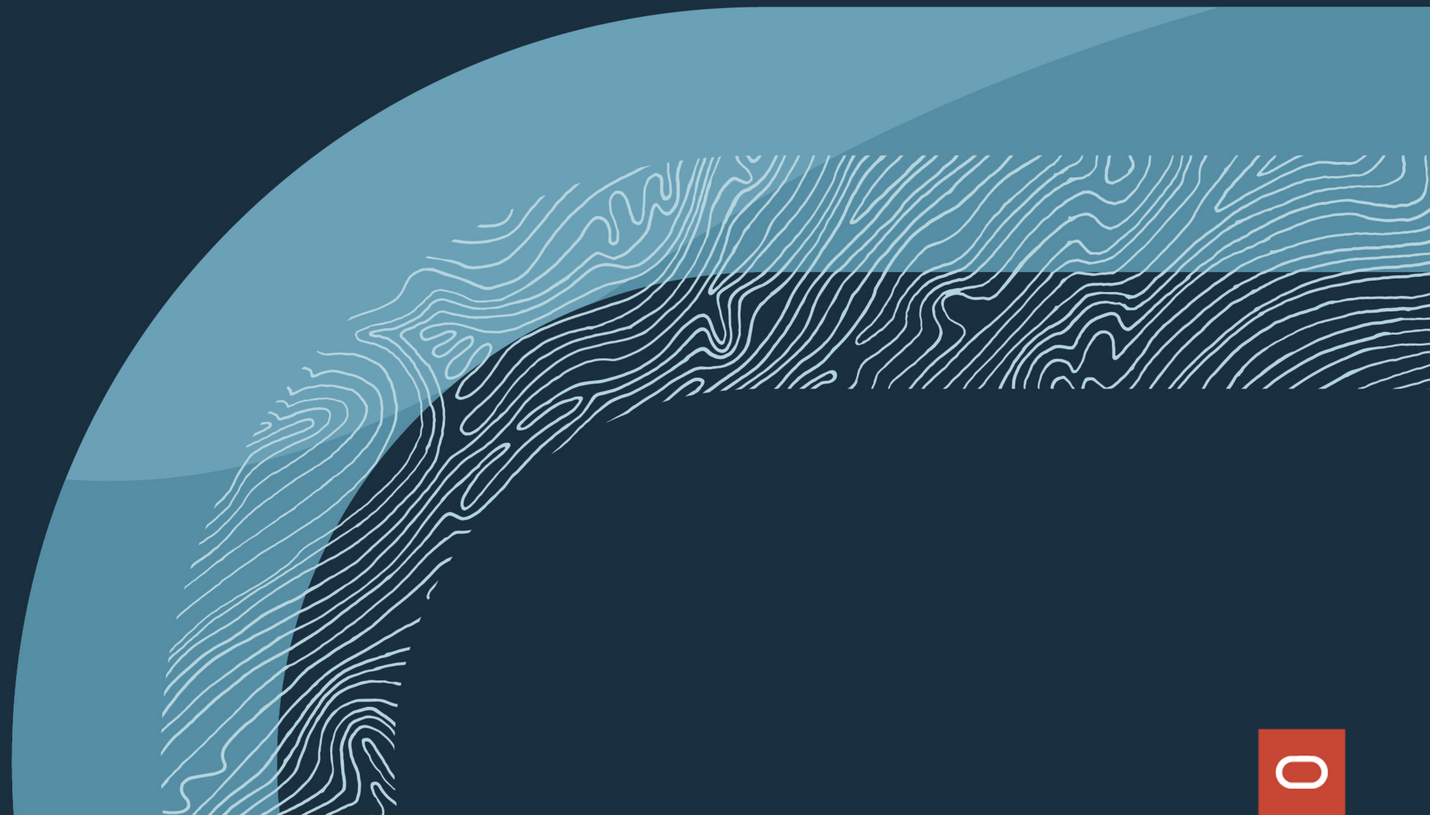
AutoMLx availability

- 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

4000
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 Sentinel[1]

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

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

The dataset

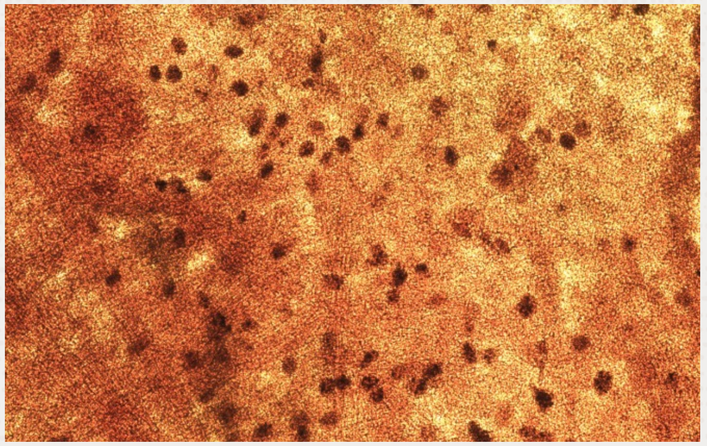
Includes information about the environment of locations all around the US coast of the Gulf of Mexico, where oysters were collected and tested for the disease in the scope of the Dermo Sentinel project.[1]

~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

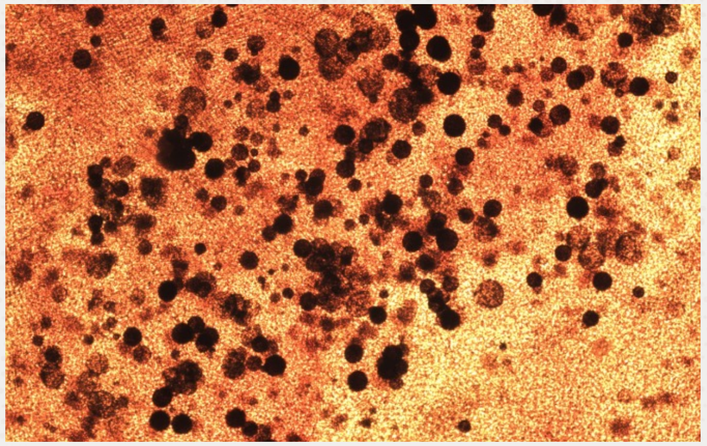
Dermo risk assessment



Low risk

If infection intensity is less than one.

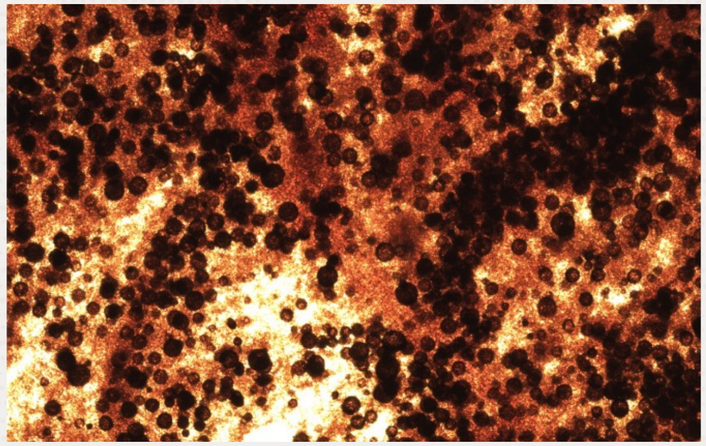
While some infected oysters may be present, incidence of the disease is still under control.



Moderate risk

Infection intensity is between 1 and 2.

The area should be monitored closely and interventions should not be delayed.



High risk

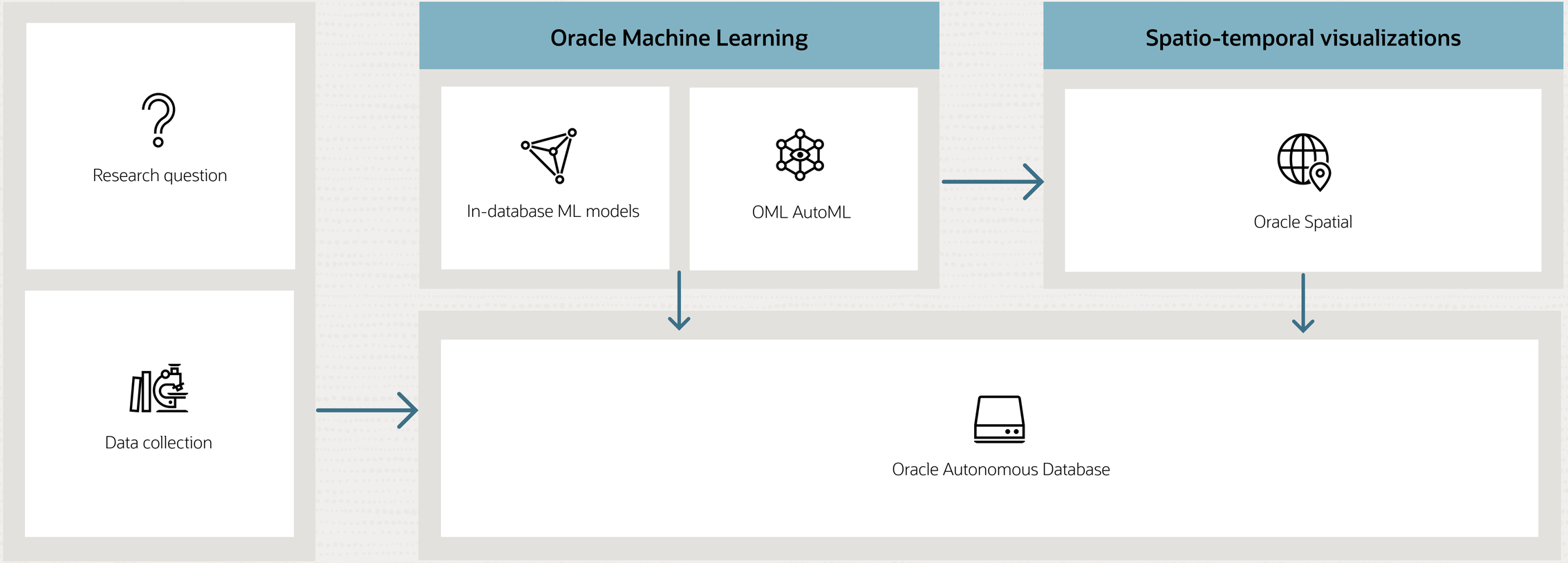
Infection intensity is above 2.

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

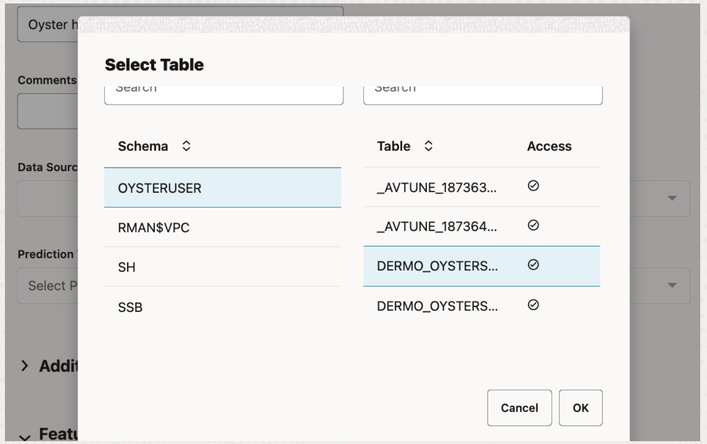
How Oracle can help

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

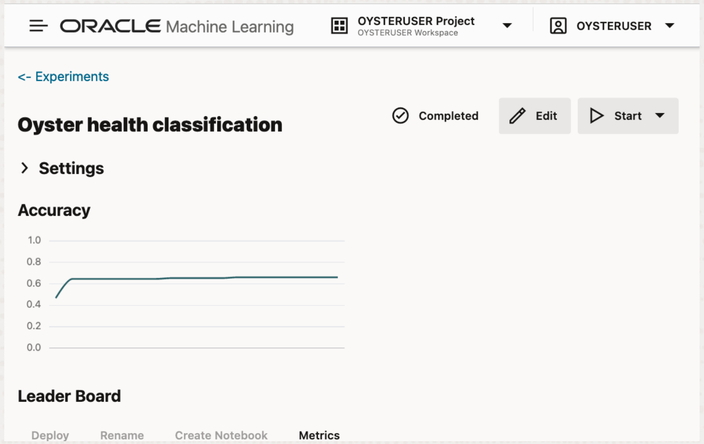


Zoom-in: OML

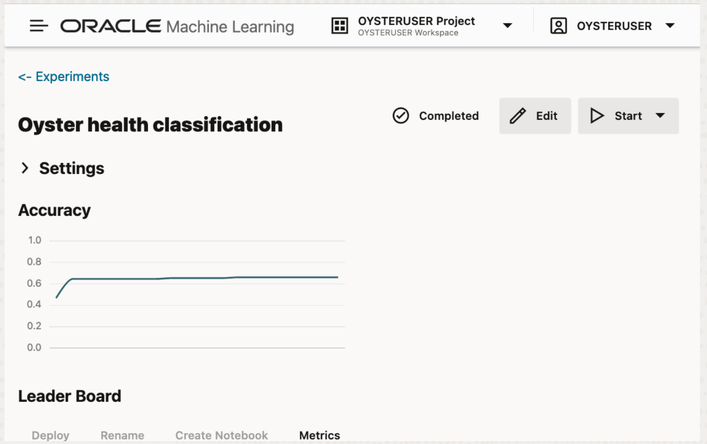
Eliminates the need to move data to dedicated machine learning systems



Load data directly from database tables



Create and manage projects with the OML AutoML UI

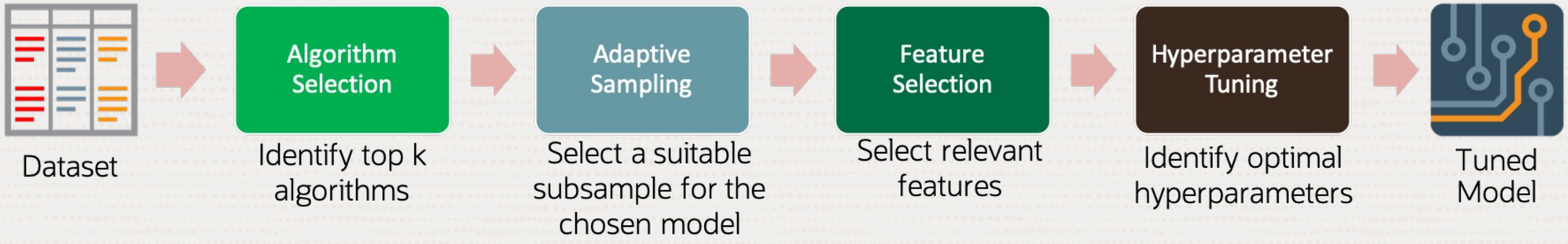


Further explore data and models with OML Notebooks

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

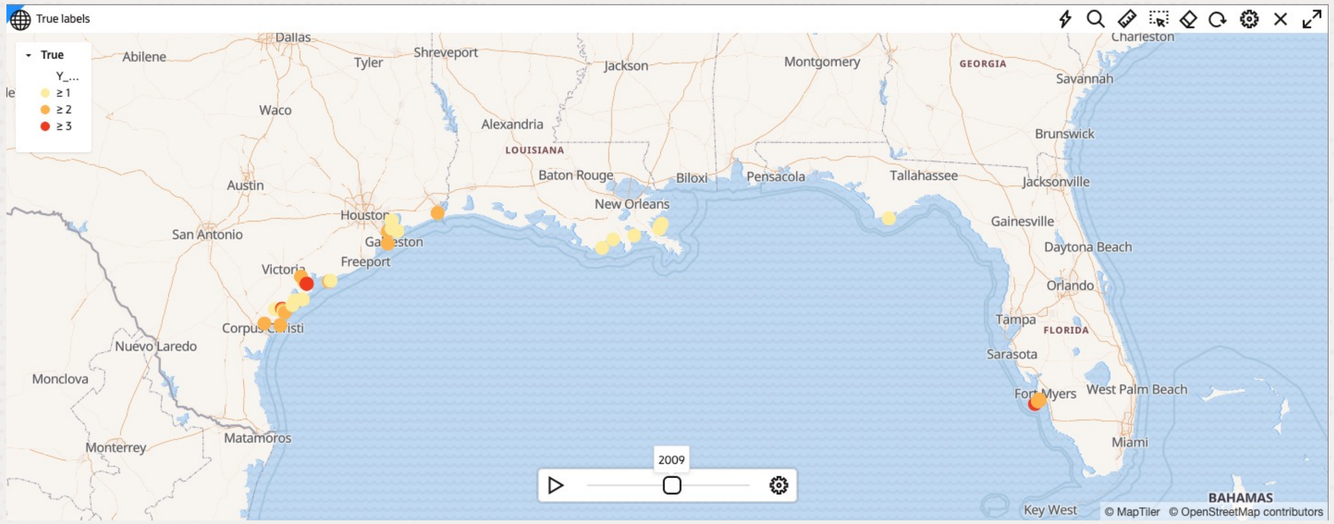
Zoom-in: AutoML

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)

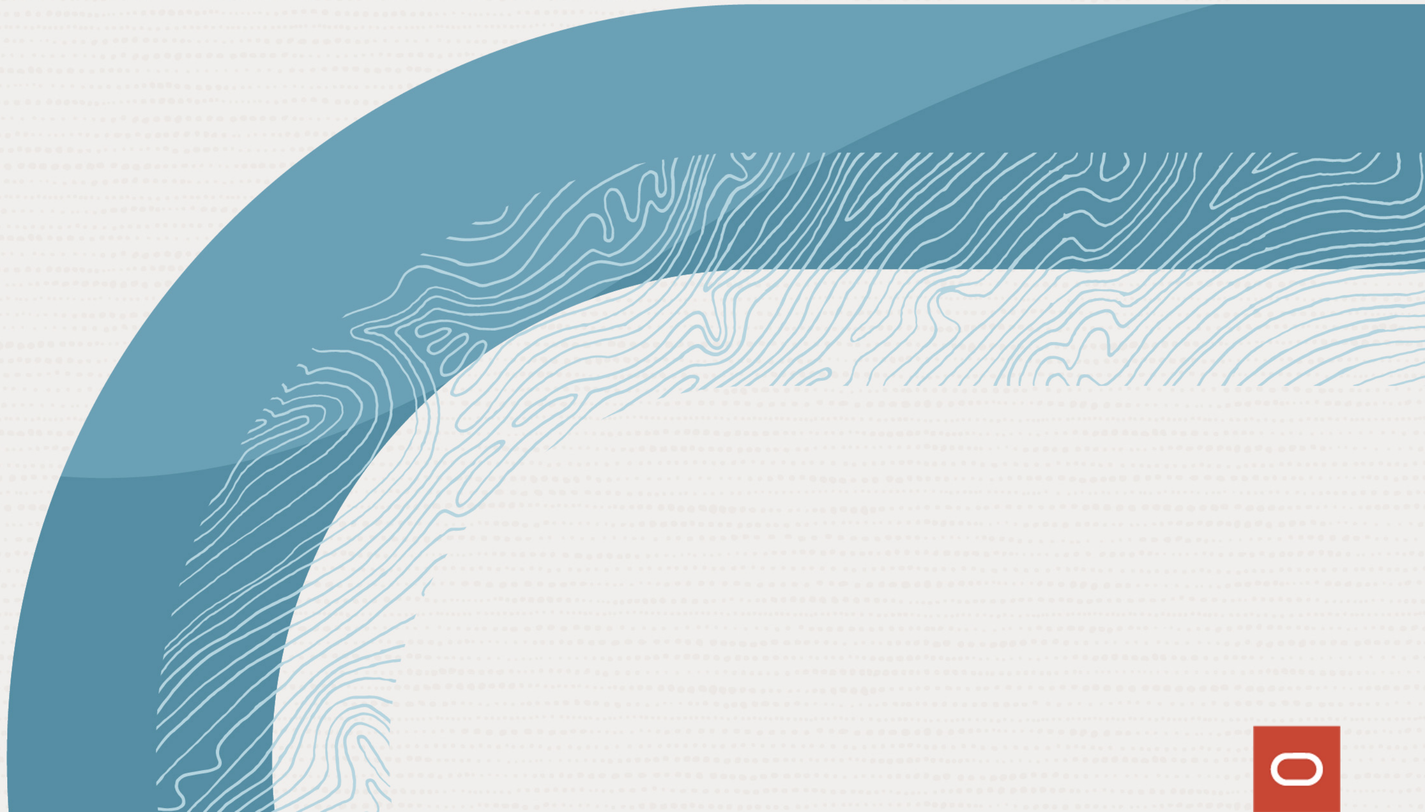
72%

Accuracy

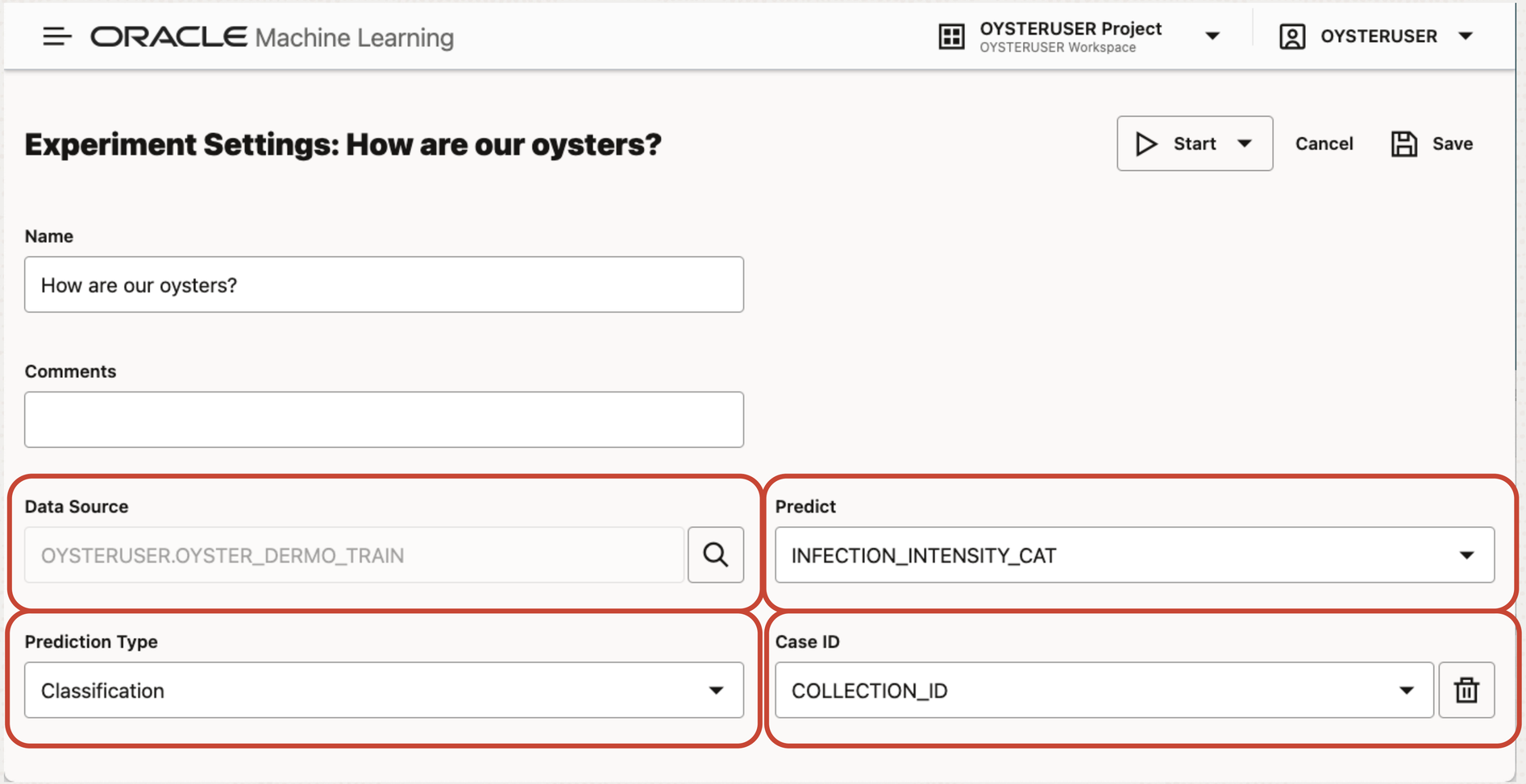
66%

Balanced Accuracy

Demo

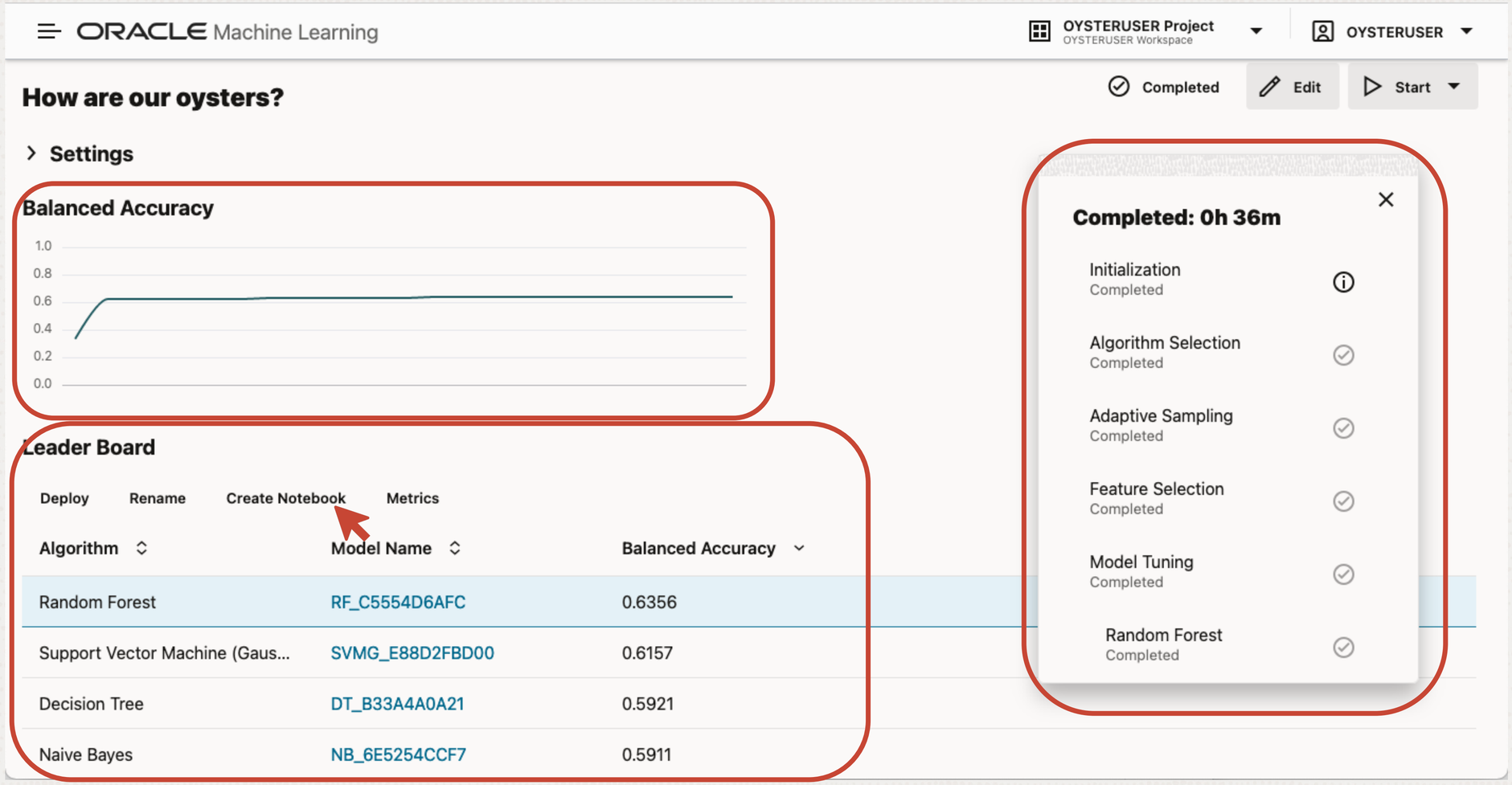


Oracle Machine Learning UI – Configuration



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

Oracle Machine Learning UI – Experiment overview

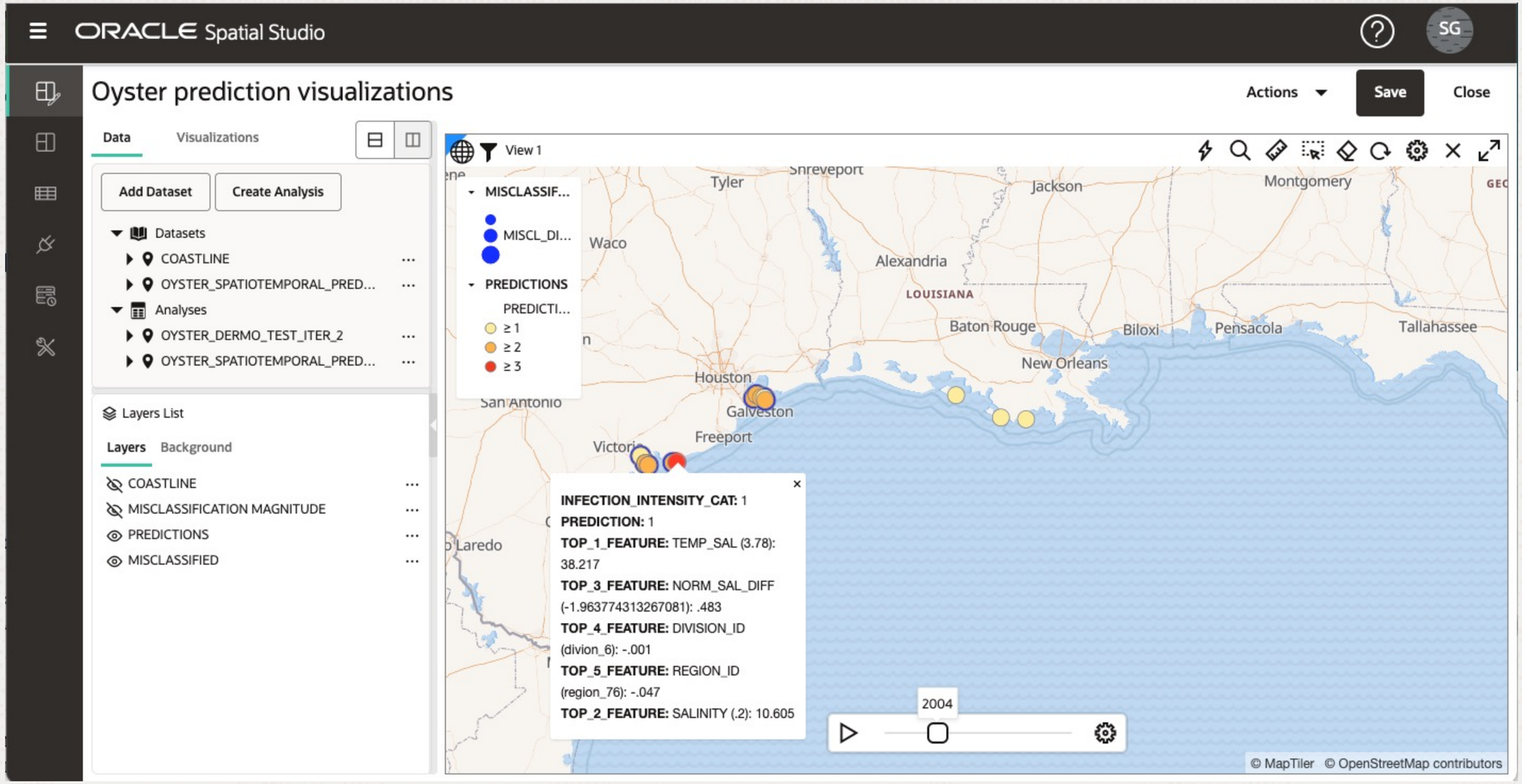


1. AutoML pipeline progress
2. Model score improvements
3. Model leaderboard, details and actions

Oracle Machine Learning – Notebooks

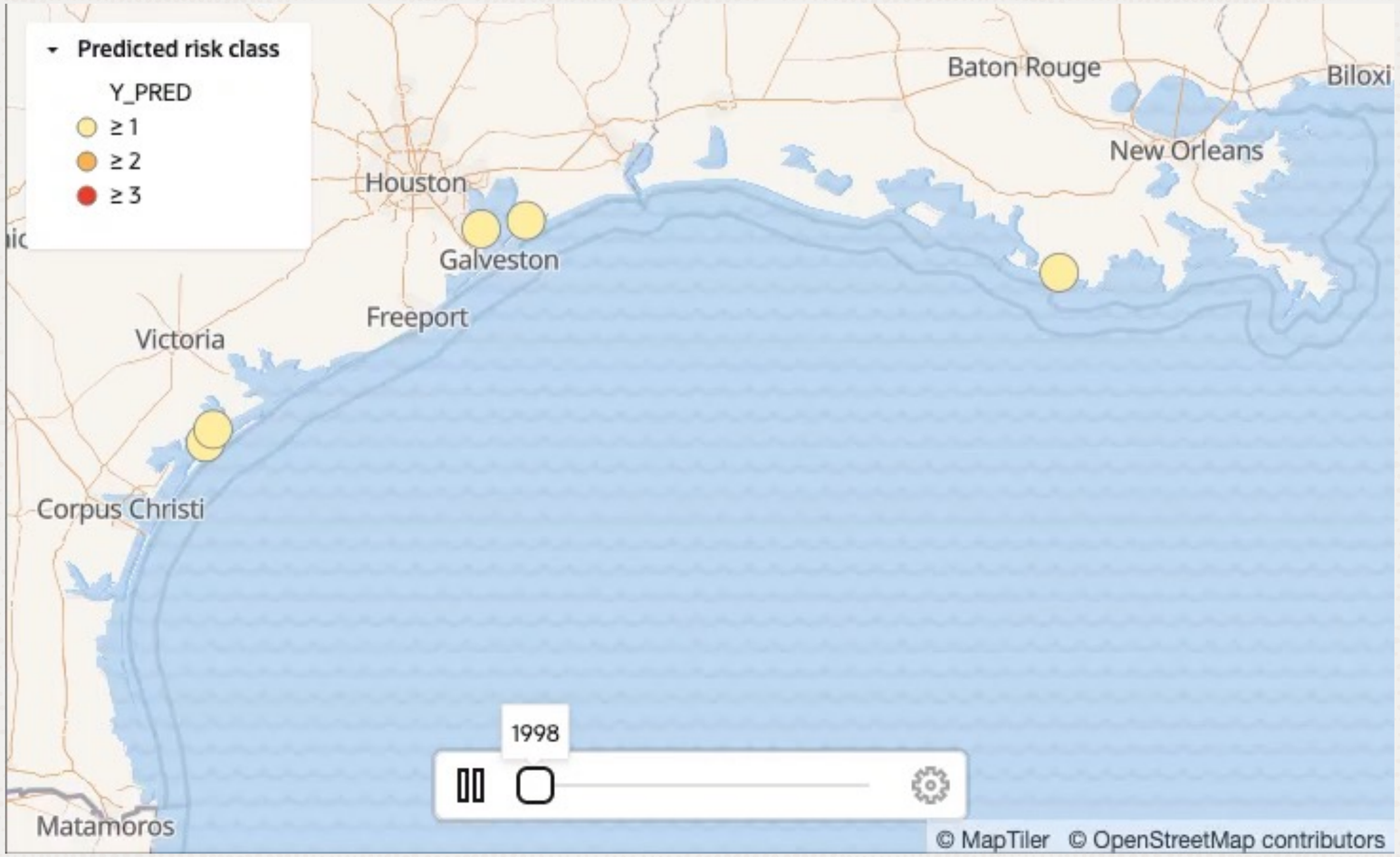
1. Generate code to re-create the model tuned by the AutoML pipeline
2. Score model on held-out test set
3. Compute feature importances for prediction explainability

Oracle Spatial Studio – Visualization

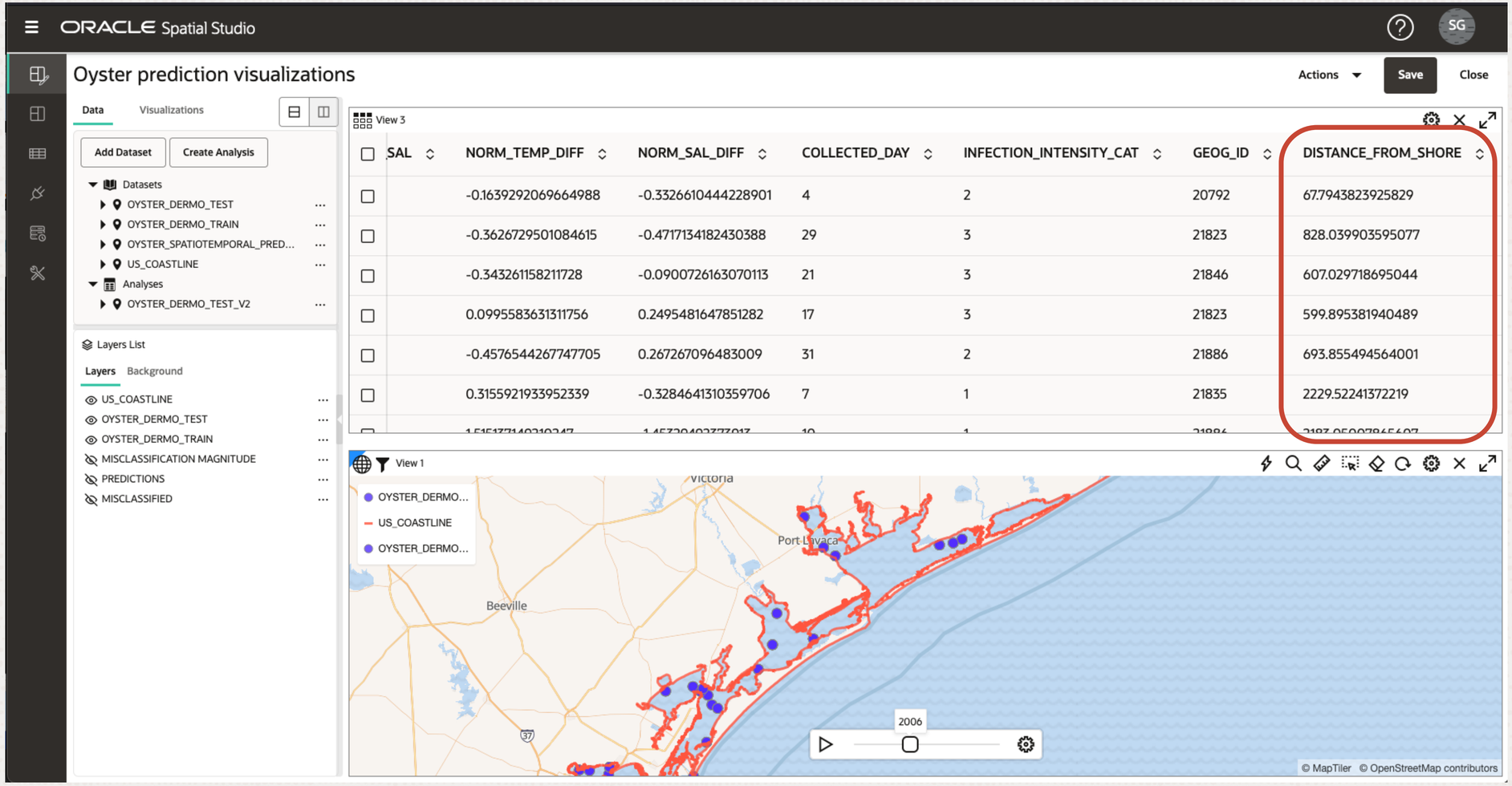


1. Visualize evolution of predictions over time
2. Review most important features used by the model for each prediction

Oracle Spatial Studio – Visualization



Oracle Spatial Studio – Spatial analysis



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

