Is It Corked? Wine Machine Learning Predictions with OAC

Francesco Tisiot - @Ftisiot
Analytics Tech Lead - Rittman Mead

rittmanmead
A DATA AND ANALYTICS COMPANY

Agenda

- OAC
- Data Scientist
- Become a Data Scientist

Oracle Analytics Cloud

- Platform Services (PaaS)
- Delivered entirely in the cloud:
  - No infrastructure footprint
  - Flexibility
  - Simplified, metered licensing
- Several options to suit your needs:
  - BYOL
  - Functionality bundled into 2 editions
    - Professional
    - Enterprise

Functions

OAC supports **Every** type of analytics

Classic

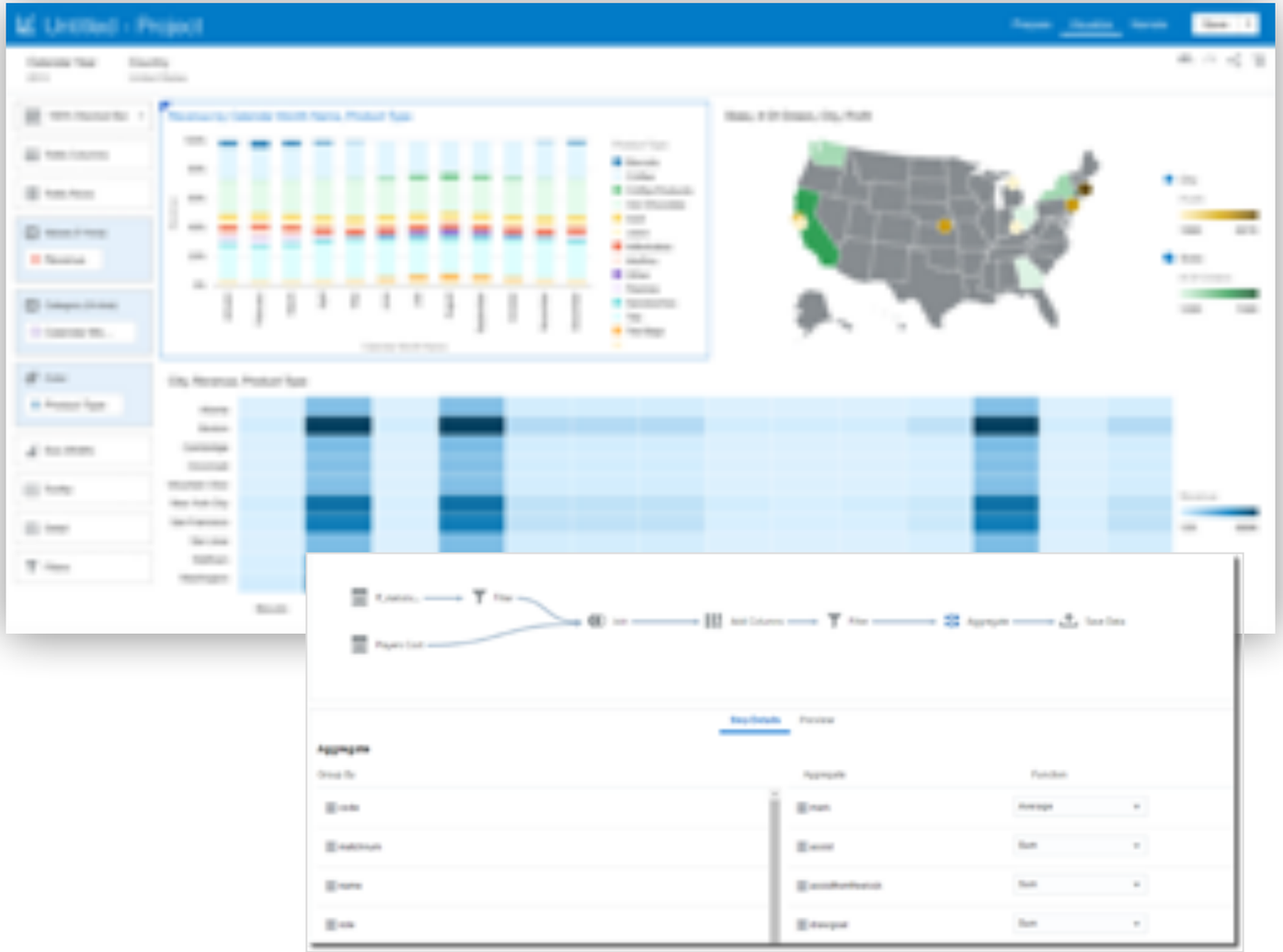
Modern

Classic Enterprise BI

- Similar to OBIEE 12c
  - Centrally maintained & governed
  - Semantic model
- Interactive Dashboards
  - KPI measurement & monitoring
  - Guided navigation paths
- BI Publisher
  - Highly formatted, burst outputs
- Action Framework
  - Navigation actions
  - Scheduled agents

Modern Data Discovery

- Data Preparation
  - Acquire data
  - Clean/Enrich
  - Transform
  - Repeatable Flows
- Data Visualisation
  - Create visual insights rapidly
  - Construct narrated storyboards
  - Share findings



Unified Analytics

Centralised Reporting

Free Discovery

Unique Source of Truth

Data Enrichment and Cleaning

Specific Access Control

Raw Data To Insights

https://speakerdeck.com/ftisiot/become-an-equilibrista-find-the-right-balance-in-the-analytics-tech-ecosystem

Augmented Analytics

Data Enrichment Suggestions

Natural Language Processing

Explain

One-Click Advanced Analytics

Advanced Machine Learning

OAC and Data Science

Basic Operations

What are the Drivers for My Sales?

Based on my Experience I can Guess….

Statistically Significant Drivers for Sales Are …

Augmented Analytics

Basic Operations

Is this Client going to accept the Offer?

YES/NO

50%

Basic ML Model

70%

Before Starting…. Define the Problem!

Problem Definition:
Predicting Wine Quality

Rule Based

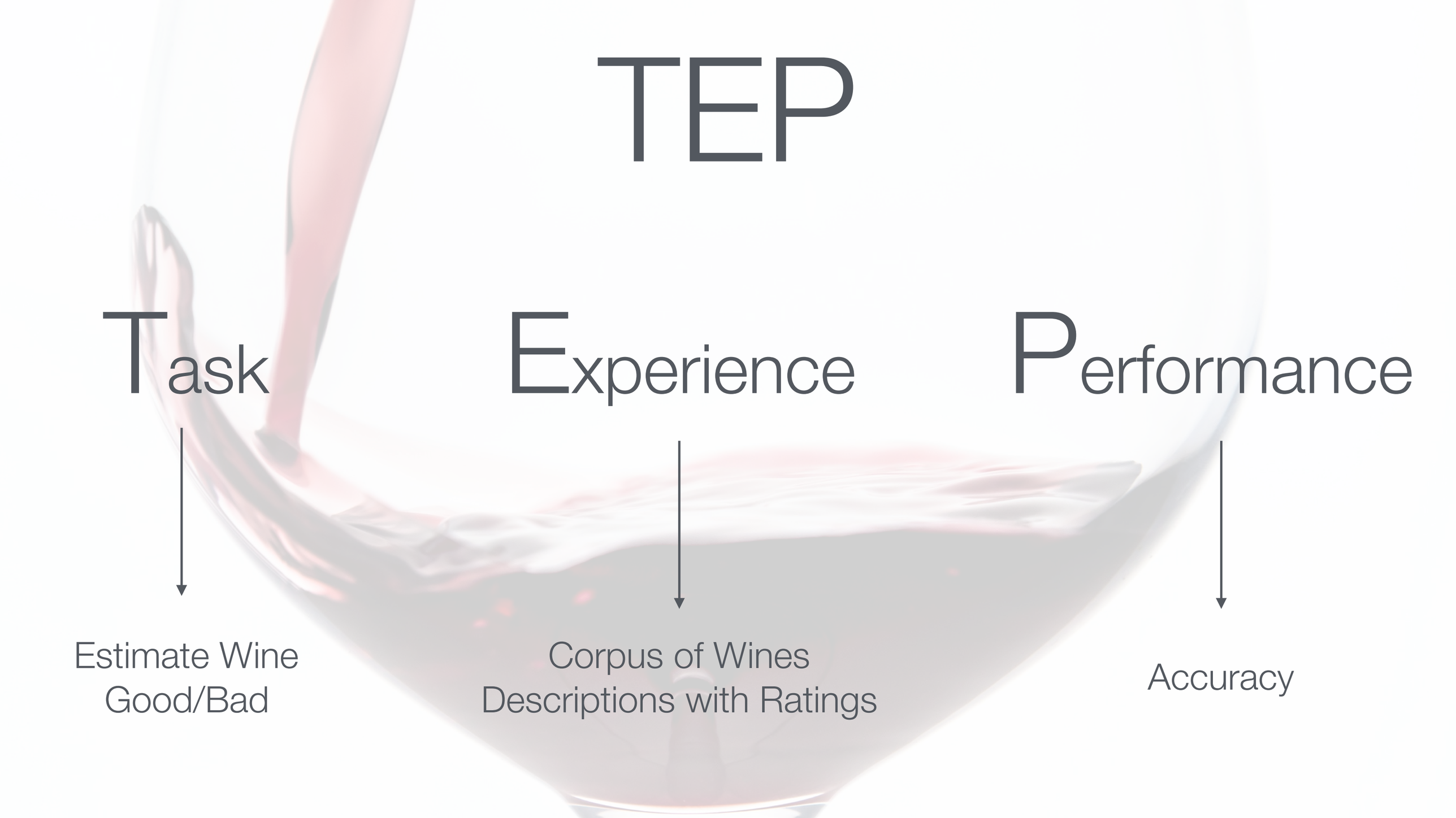
Italy or France -> Good

Rest of the World -> Bad

Price >= 10 Euros -> Good

Price < 10 Euros -> Bad

Price > 30 & Production Zone = Veneto & …. -> 6.5



TEP

Task

Estimate Wine
Good/Bad

Experience

Corpus of Wines
Descriptions with Ratings

Performance

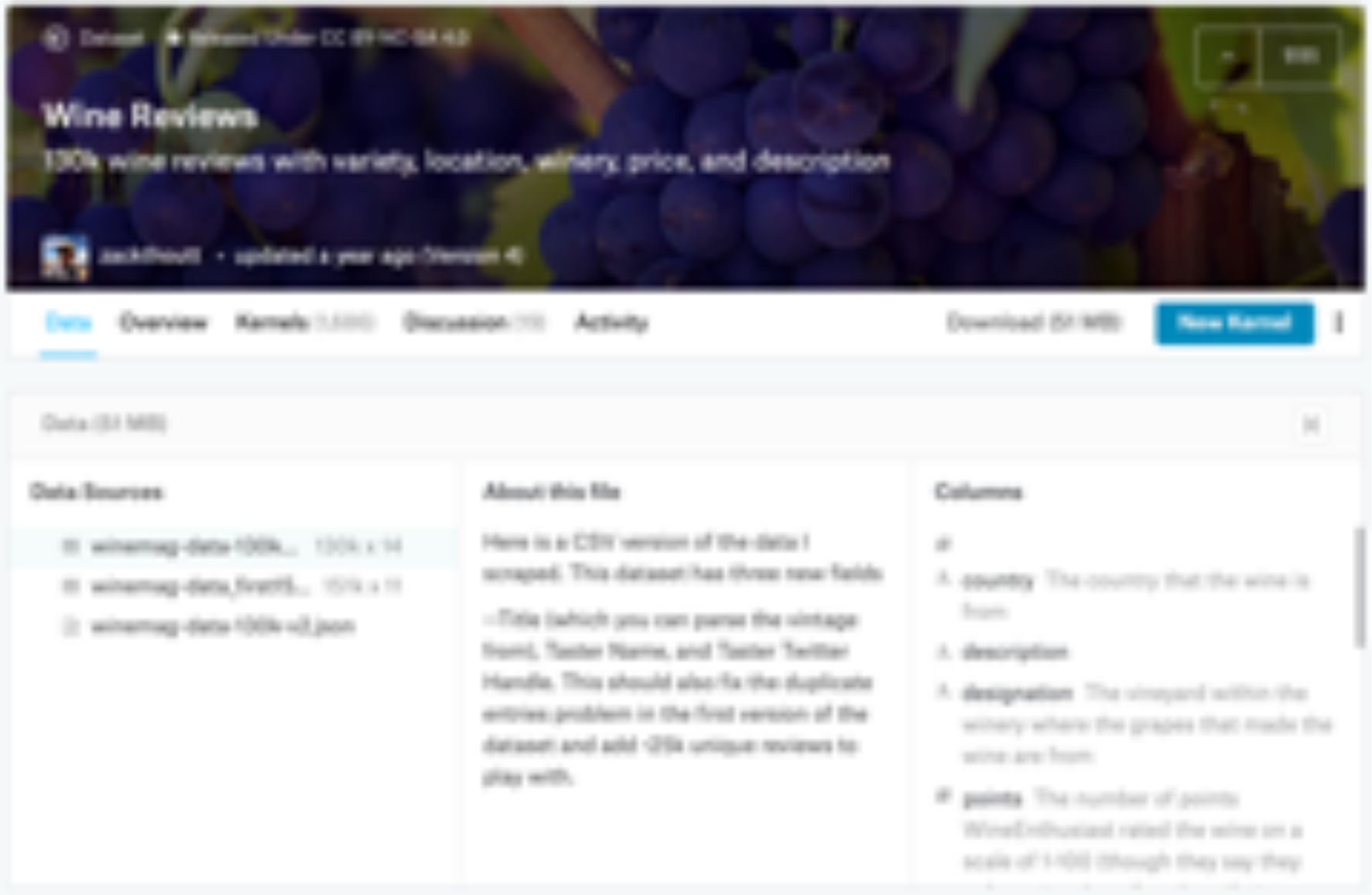
Accuracy

Accuracy

Real Value \ Predicted Value	Good	Bad
Good		
Bad		

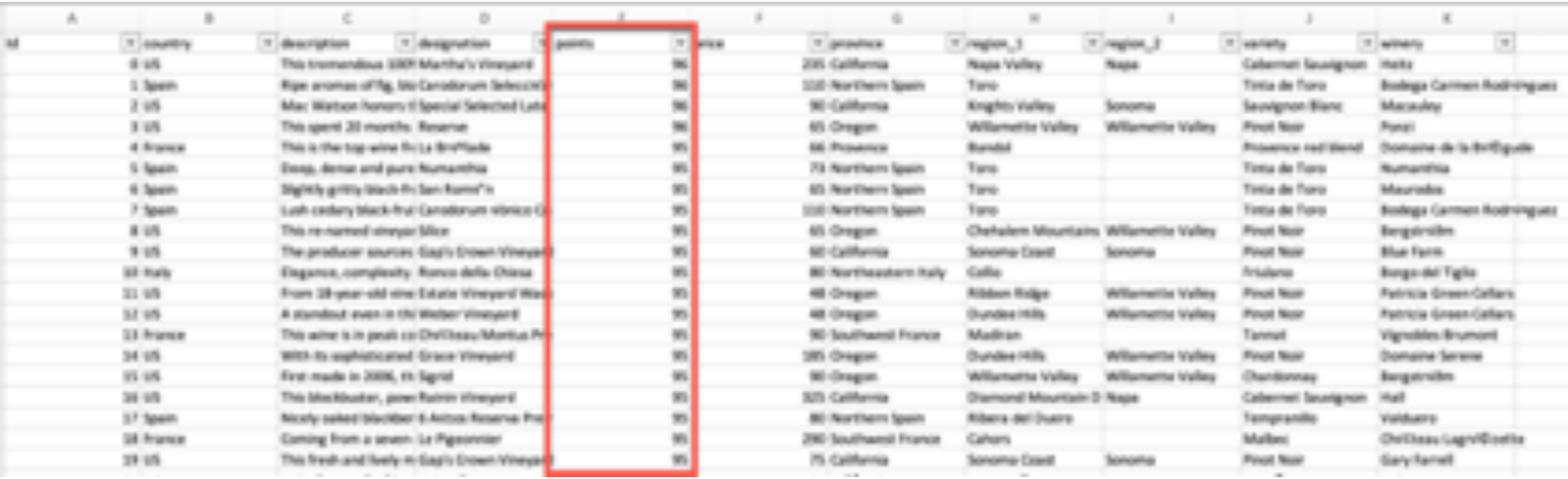
Accuracy =  / ( + )

Dataset



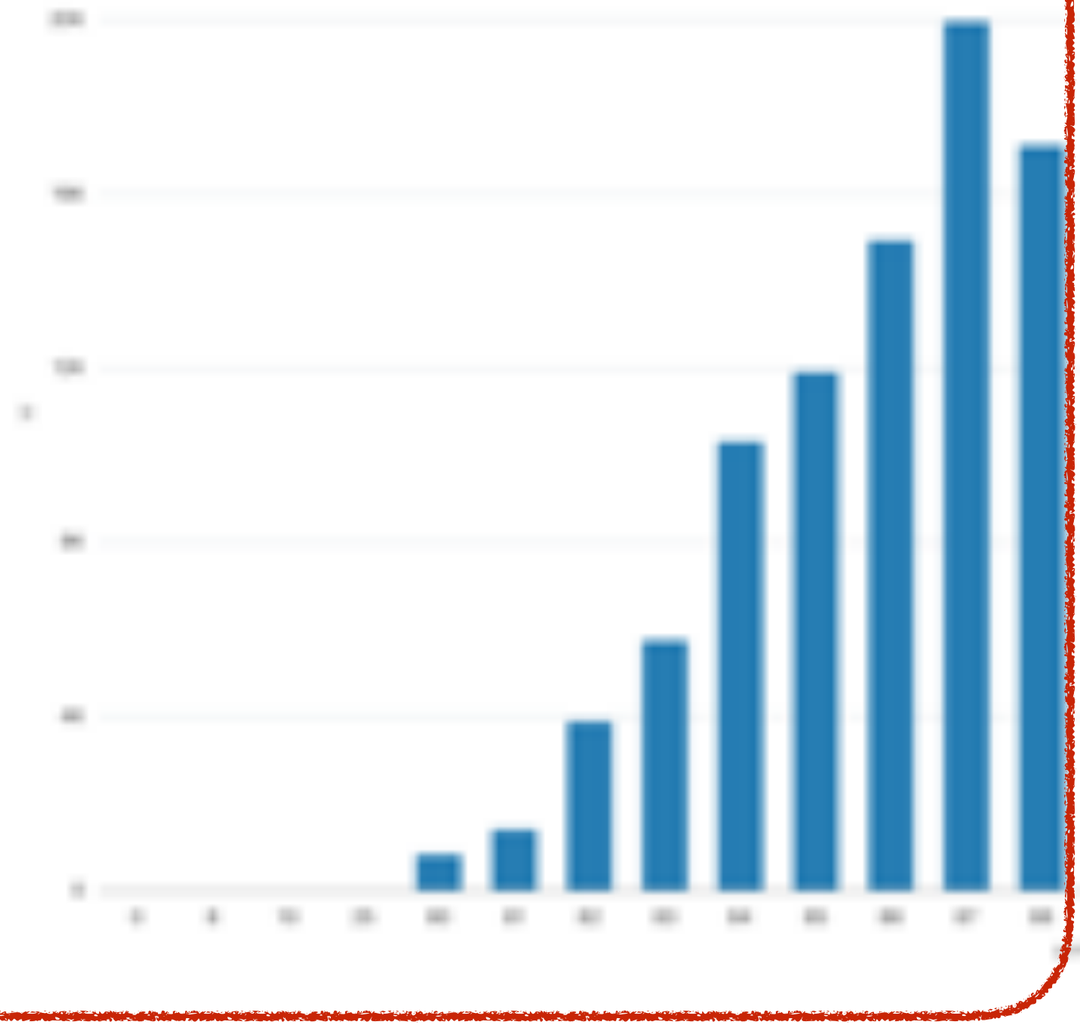
https://www.kaggle.com/zynicide/wine-reviews

The Data

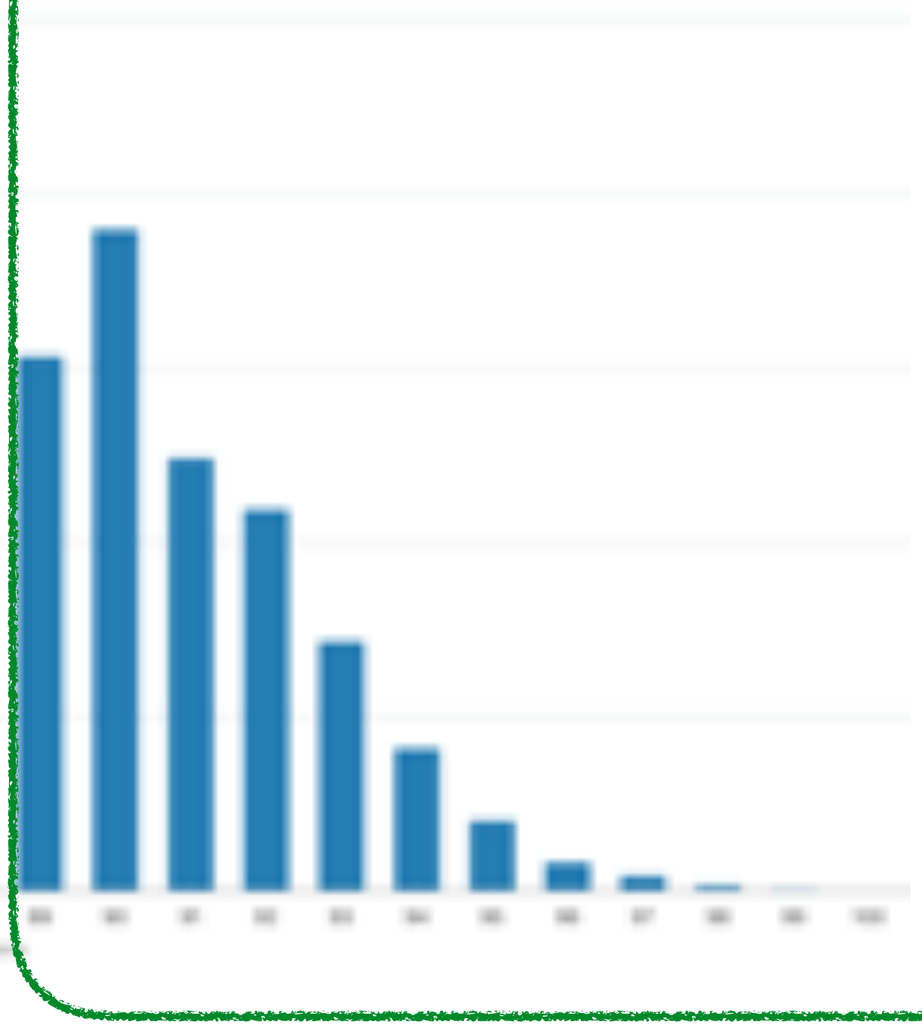


id	country	description	designation	points	price	province	region_1	region_2	variety	winery
0	US	This tremendous 100%	Martha's Vineyard	96	235	California	Napa Valley	Napa	Cabernet Sauvignon	Heitz
1	Spain	Ripe aromas of fig, bla	Carodorum Selección	96	110	Northern Spain	Toro		Tinta de Toro	Bodega Carmen Rodríguez
2	US	Mac Watson honors t	Special Selected Late	96	90	California	Knights Valley	Sonoma	Sauvignon Blanc	Macauley
3	US	This spent 20 months	Reserve	96	65	Oregon	Willamette Valley	Willamette Valley	Pinot Noir	Ponzi
4	France	This is the top wine fr	La Brûlade	95	66	Provence	Bandol		Provence red blend	Domaine de la Bégude
5	Spain	Deep, dense and pure	Numanthia	95	73	Northern Spain	Toro		Tinta de Toro	Numanthia
6	Spain	Slightly gritty black-fr	San Román	95	65	Northern Spain	Toro		Tinta de Toro	Maurodos
7	Spain	Lush cedary black-frui	Carodorum Único	95	110	Northern Spain	Toro		Tinta de Toro	Bodega Carmen Rodríguez
8	US	This re-named vineyar	Silice	95	65	Oregon	Chehalem Mountains	Willamette Valley	Pinot Noir	Bergström
9	US	The producer sources	Gap's Crown Vineyard	95	60	California	Sonoma Coast	Sonoma	Pinot Noir	Blue Farm
10	Italy	Elegance, complexity	Ronco della Chiesa	95	80	Northeastern Italy	Collio		Friulano	Borgo del Tiglio
11	US	From 18-year-old vine	Estate Vineyard Wa	95	48	Oregon	Ribbon Ridge	Willamette Valley	Pinot Noir	Patricia Green Cellars
12	US	A standout even in thi	Weber Vineyard	95	48	Oregon	Dundee Hills	Willamette Valley	Pinot Noir	Patricia Green Cellars
13	France	This wine is in peak co	Château Montus Pr	95	90	Southwest France	Madiran		Tannat	Vignobles Brumont
14	US	With its sophisticated	Grace Vineyard	95	185	Oregon	Dundee Hills	Willamette Valley	Pinot Noir	Domaine Serene
15	US	First made in 2006, th	Sigrid	95	90	Oregon	Willamette Valley	Willamette Valley	Chardonnay	Bergström
16	US	This blockbuster, pow	Rainin Vineyard	95	325	California	Diamond Mountain D	Napa	Cabernet Sauvignon	Hall
17	Spain	Nicely oaked blackber	6 Años Reserva Pr	95	80	Northern Spain	Ribera del Duero		Tempranillo	Valduero
18	France	Coming from a seven	Le Pigeonnier	95	290	Southwest France	Cahors		Malbec	Château Lagrézette
19	US	This fresh and lively m	Gap's Crown Vineyard	95	75	California	Sonoma Coast	Sonoma	Pinot Noir	Gary Farrell

Bad



Good



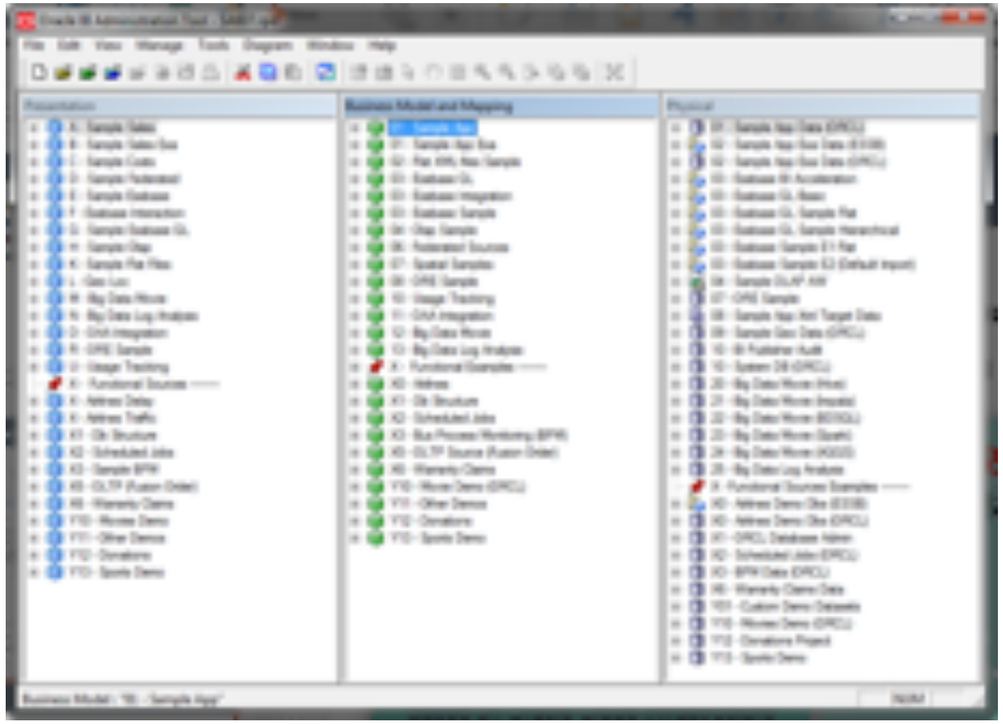
Become a Data Scientist with OAC

Connect	Clean	Transform & Enrich
Analyse	Train & Evaluate	Predict

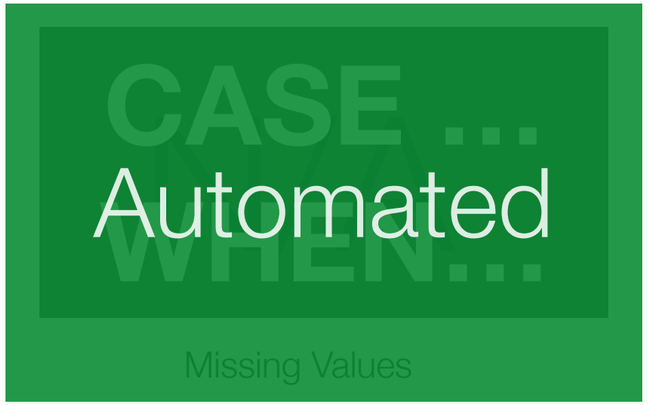
Connection Options in OAC

External
Data Sources

Pre-Defined
Data Models



Cleaning What?



Missing Values

Wrong Values

Irrelevant Observations

Labelling Columns

Handling Outliers

Feature Scaling

Aggregation

Train/Test Set Split

Feature Engineering

Location -> ZIP Code

Additional Data Sources?

Name -> Sex

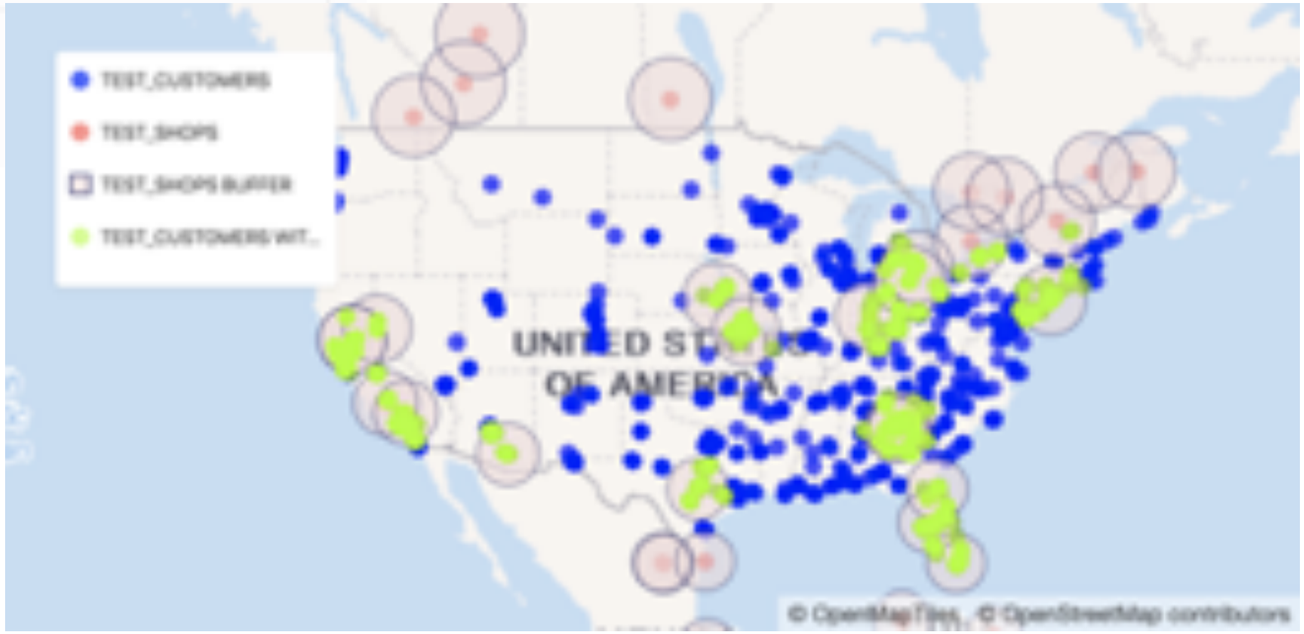
2 Locations -> Distance

Data Flow

Day/Month/Year -> Date

Data Preparation Recommendations

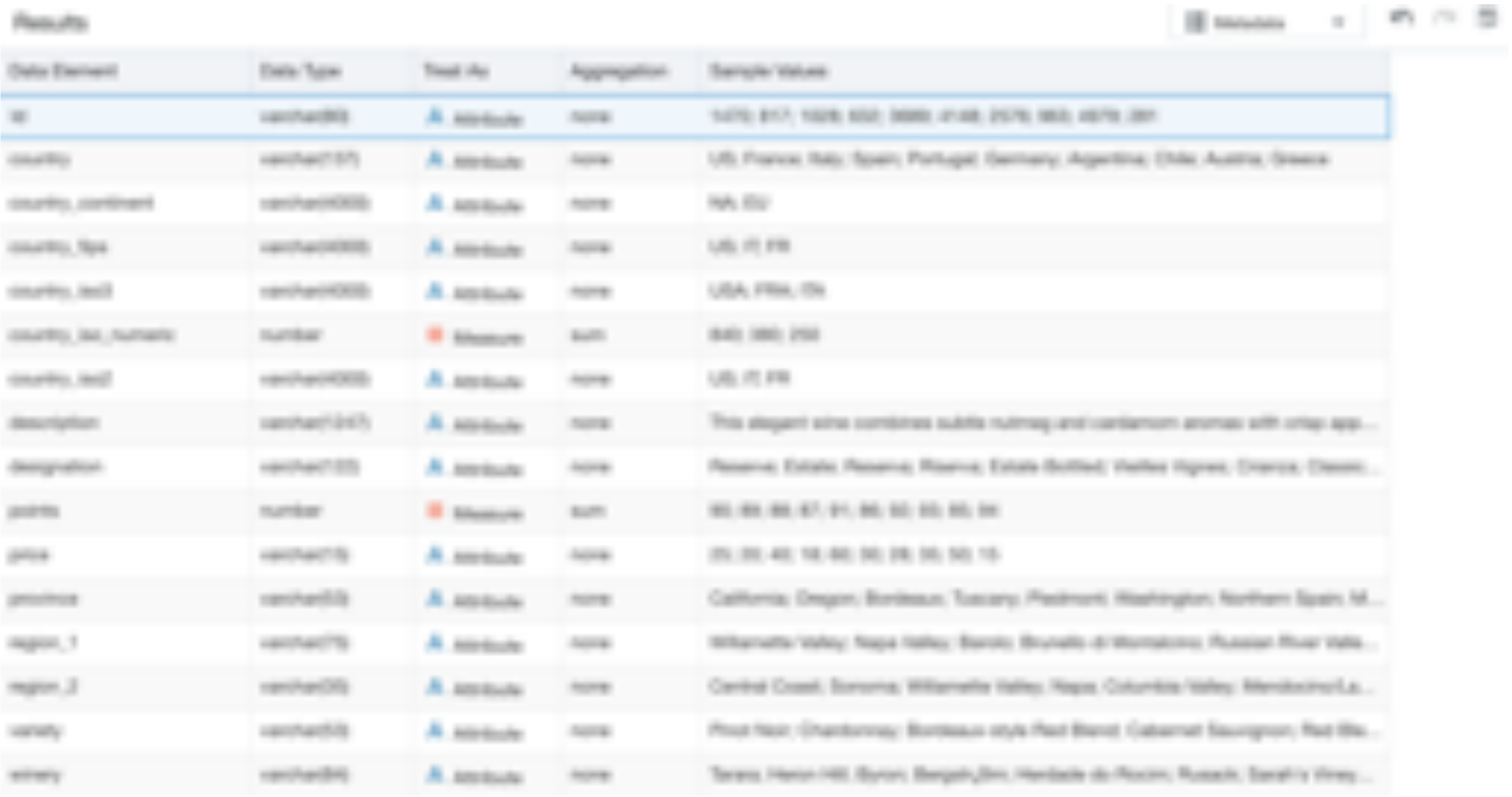
Spatial Enrichment



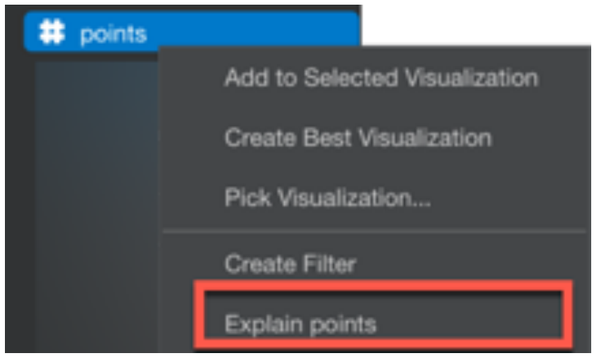
Oracle Spatial Studio

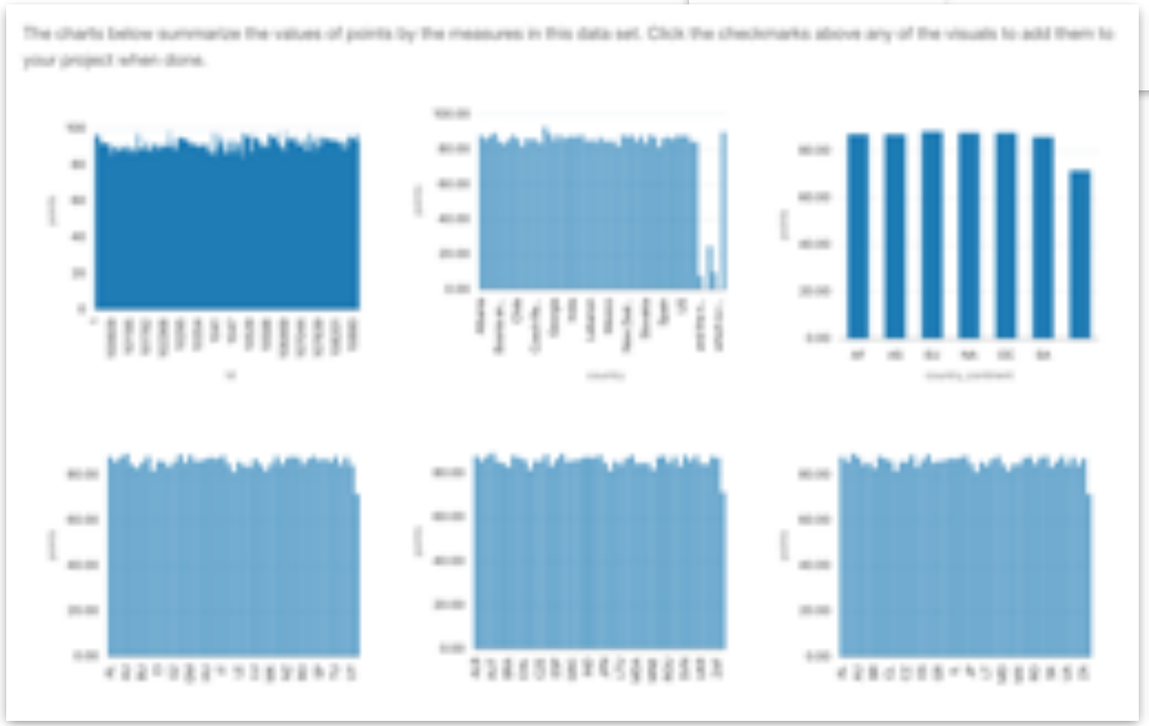
http://ritt.md/spatial-studio

Data Overview



Analyse - Explain



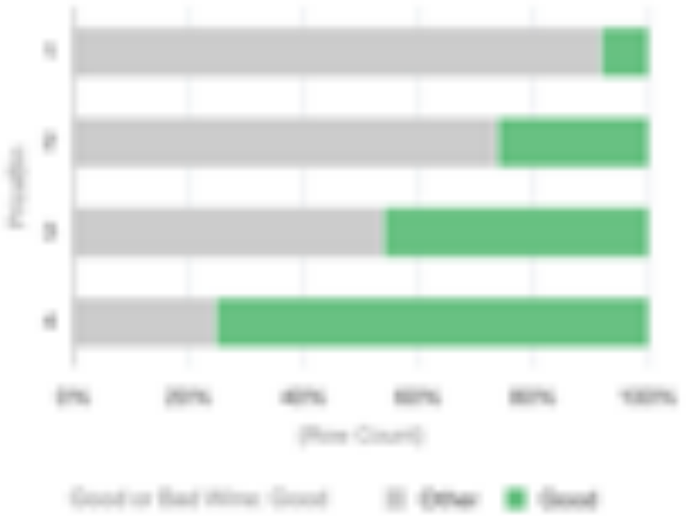


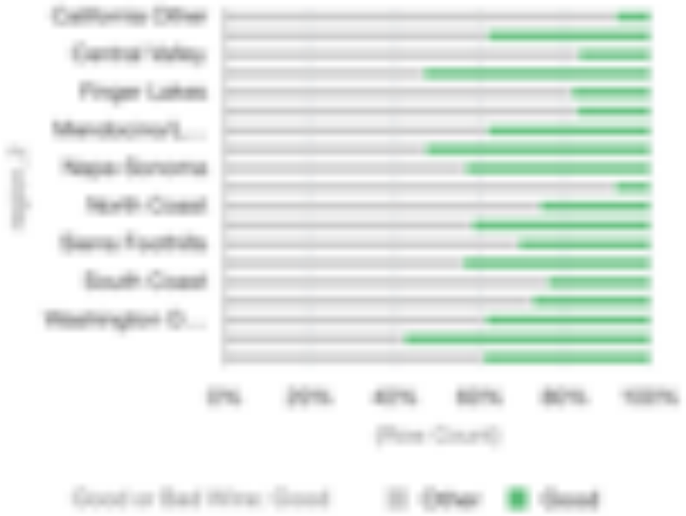
Explain - Key Drivers

Key Drivers of Good or Bad Wine

Based on Good or Bad Wine: Good the 2 attributes that are most strongly correlated are: PriceBin, region_2

The charts below show the distribution of Good or Bad Wine values across each of the key drivers. Click the checkmarks above any of the visuals to add them to your project when done.



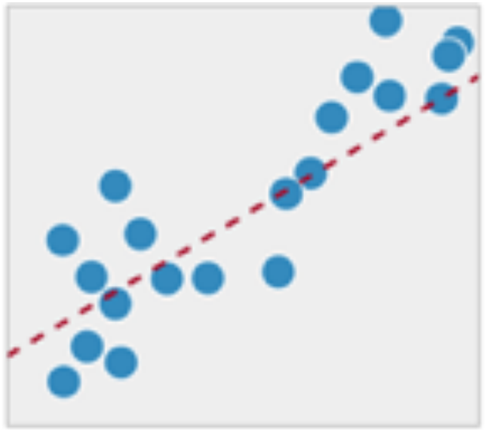


Train - What Problem are we Trying to Solve?

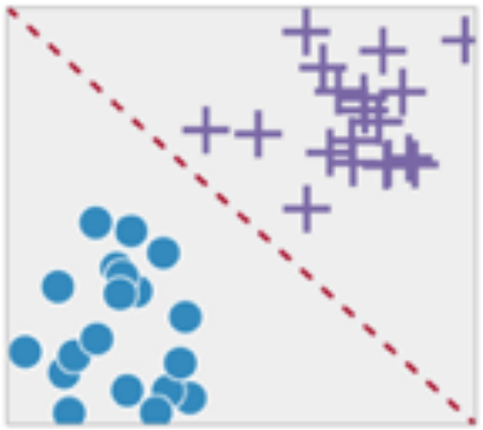
Supervised

"I want to predict the value of Y,
here are some examples"

Regression



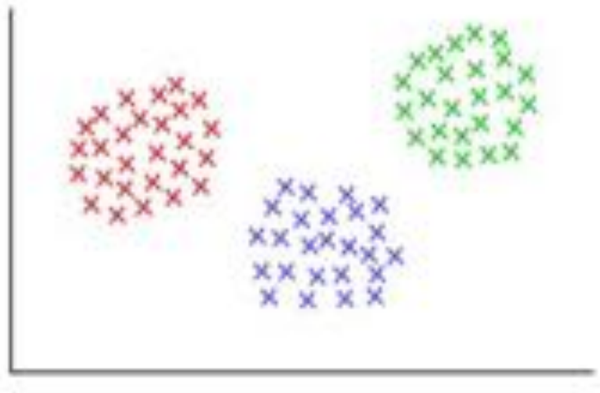
Classification



Unsupervised

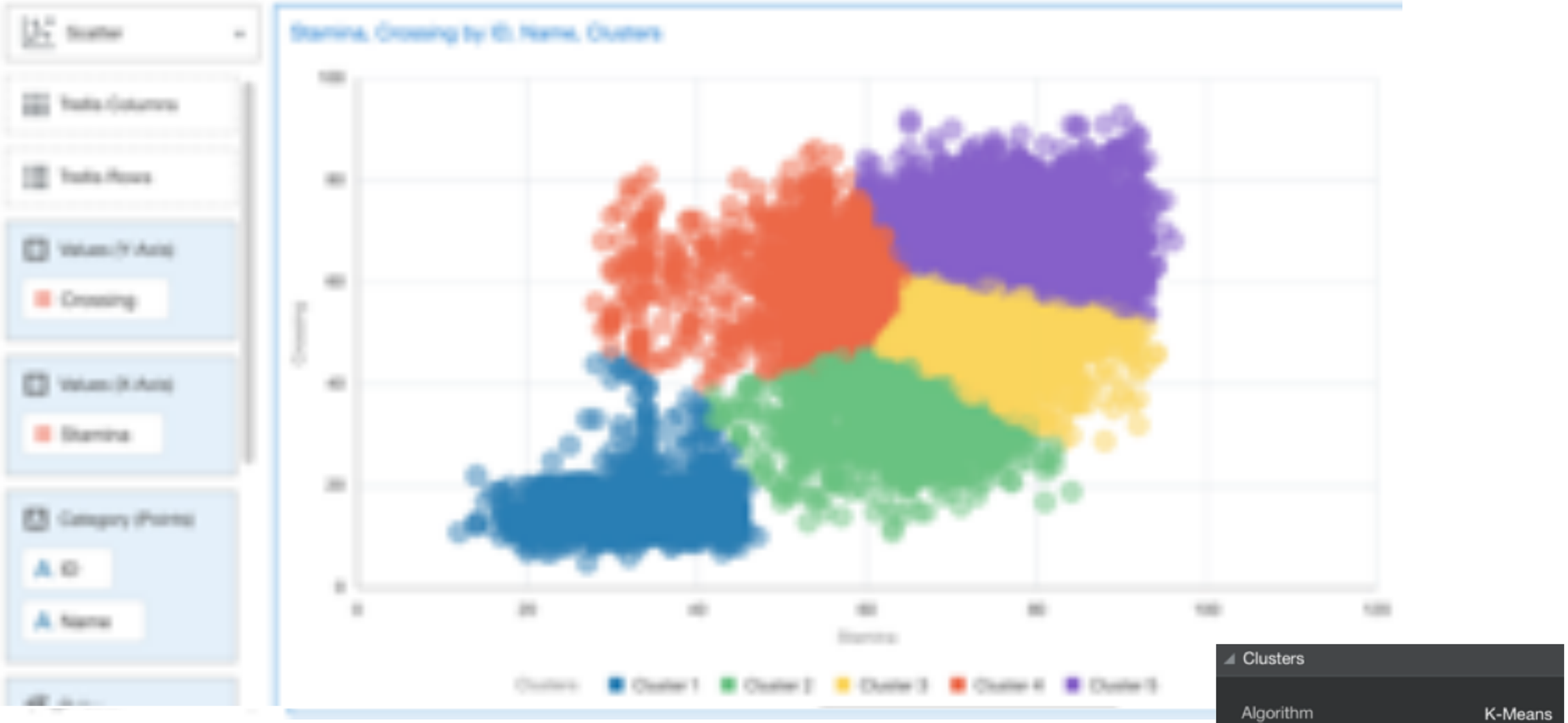
"Here is a dataset,
make sense out of it!"

Clustering

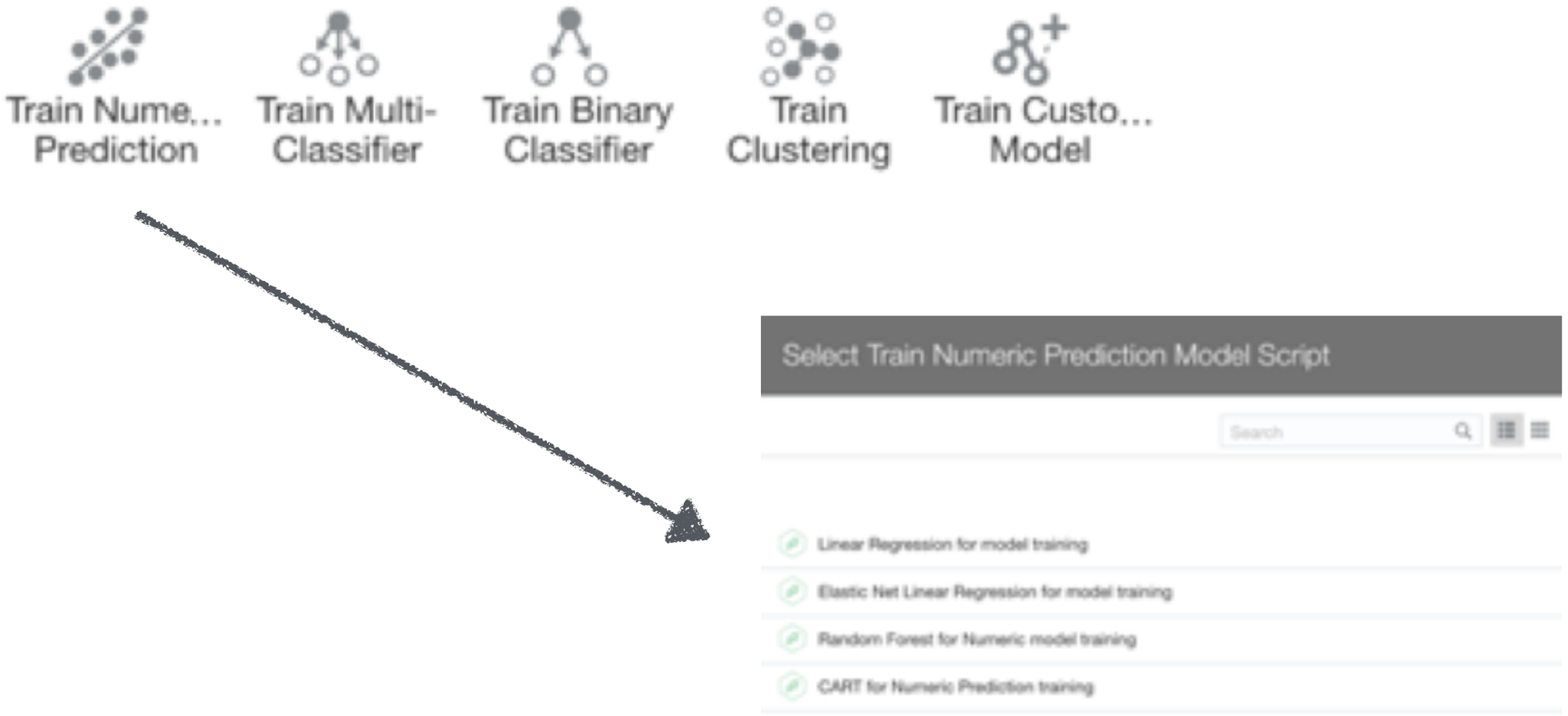


https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

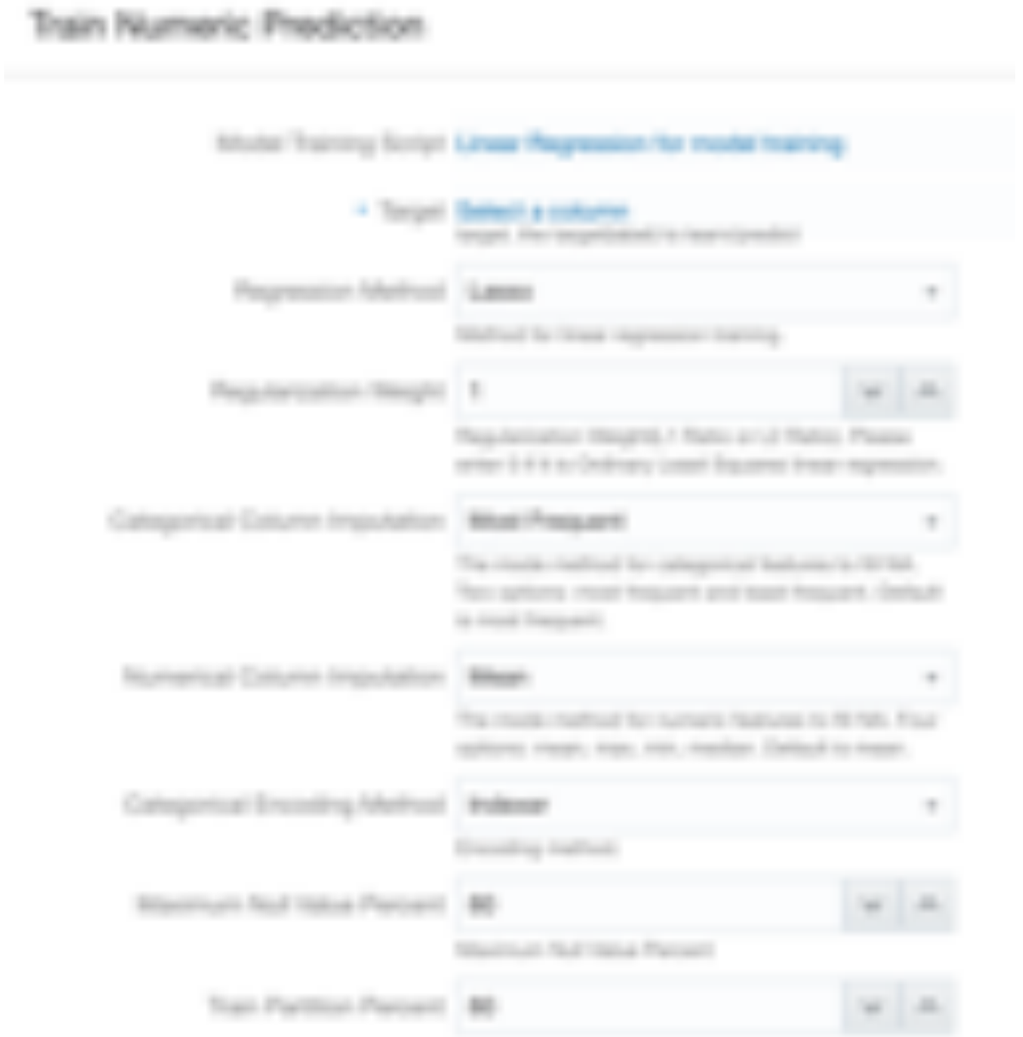
Model Training - Easy Models



DataFlow Train Model



Which Model - Parameters To Pick?



Select, Try, Save, Change, Try, Save …..

Train Numeric Prediction

Model Training Script: Elastic Net Linear Regression for model t

* Target: points

L1 Ratio: 0.5

L2 Ratio: 0.5

Categorical Column Imputation: Most Frequent

Select Train Numeric Prediction Model Script

Search

Linear Regression for model training

Elastic Net Linear Regression for model training

Random Forest for Numeric model training

CART for Numeric Prediction training

Data Sets | Connections | Data Flows | Sequences

Type | Name

ELN1

LR2

LR1

Machine Learning

Scripts | Models

Type | Name

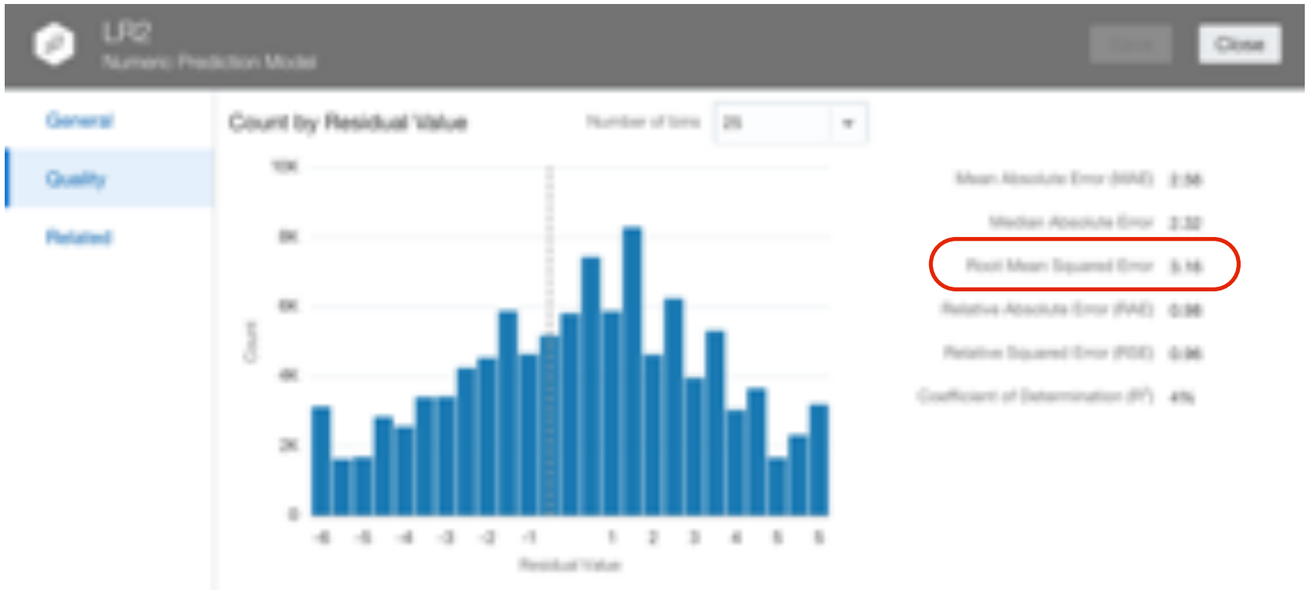
ELN1

LR2

LR1

Compare


LR1
Numeric Prediction Model
Close
General
Quality
Related
Count by Residual Value
Number of bins
25
Count
Residual Value
Mean Absolute Error (MAE) 2.60
Median Absolute Error 2.24
Root Mean Squared Error 3.23
Relative Absolute Error (RAE) 1.00
Relative Squared Error (RSE) 1.00
Coefficient of Determination (R²) 0%

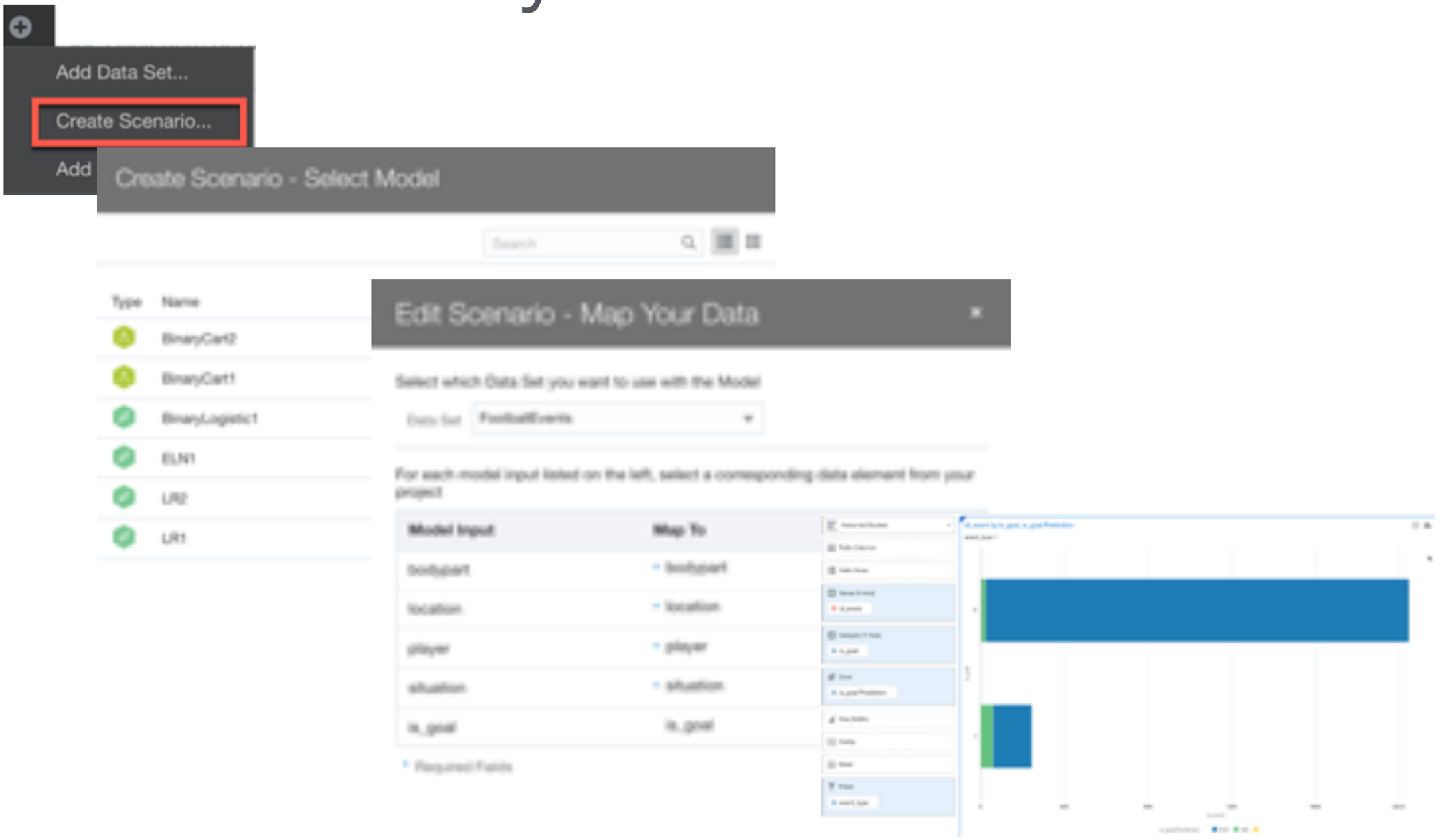

LR2
Numeric Prediction Model
Close
General
Quality
Related
Count by Residual Value
Number of bins
25
Count
Residual Value
Mean Absolute Error (MAE) 2.56
Median Absolute Error 2.32
Root Mean Squared Error 3.16
Relative Absolute Error (RAE) 0.98
Relative Squared Error (RSE) 0.96
Coefficient of Determination (R²) 4%

Compare - Classification

Predicted Values

Actual Values	0.0	1.0	Total
0.0	40439	471	40910 (90%)
1.0	3761	866	4627 (10%)
Total	44200 (97%)	1337 (3%)	45537 (100%)

Use On the Fly or with a Dataflow




Demo

Congratulations!

…You are now a Data Scientist!

ML Production Deployment

Data Scientist

ML -> Data

Oracle Machine Learning

ORACLE APPLICATIONS

ORACLE FUSION APPLICATIONS

Oracle Machine Learning

OML4SQL
Oracle Advanced Analytics
SQL API

OML Notebooks
with Apache Zeppelin on
Autonomous Database

OML4R
Oracle R Enterprise
R API

Oracle Data Miner
Oracle SQL Developer extension

OML4Py*
Python API

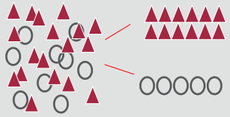
OML4Spark
Oracle R Advanced Analytics
for Hadoop

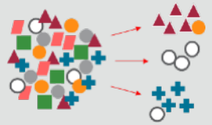
OML Microservices*
Supporting Oracle Applications
Image, Text, Scoring, Deployment,
Model Management

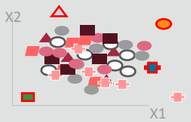
APACHE Spark

* Coming soon

Oracle Machine Learning Algorithms

Classification

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

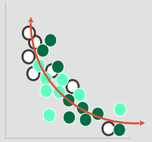
Clustering

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

Anomaly Detection


One-Class SVM

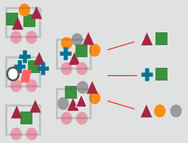
Time Series

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
LASSO

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

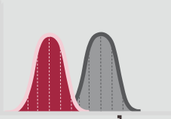
SQL Analytics


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.

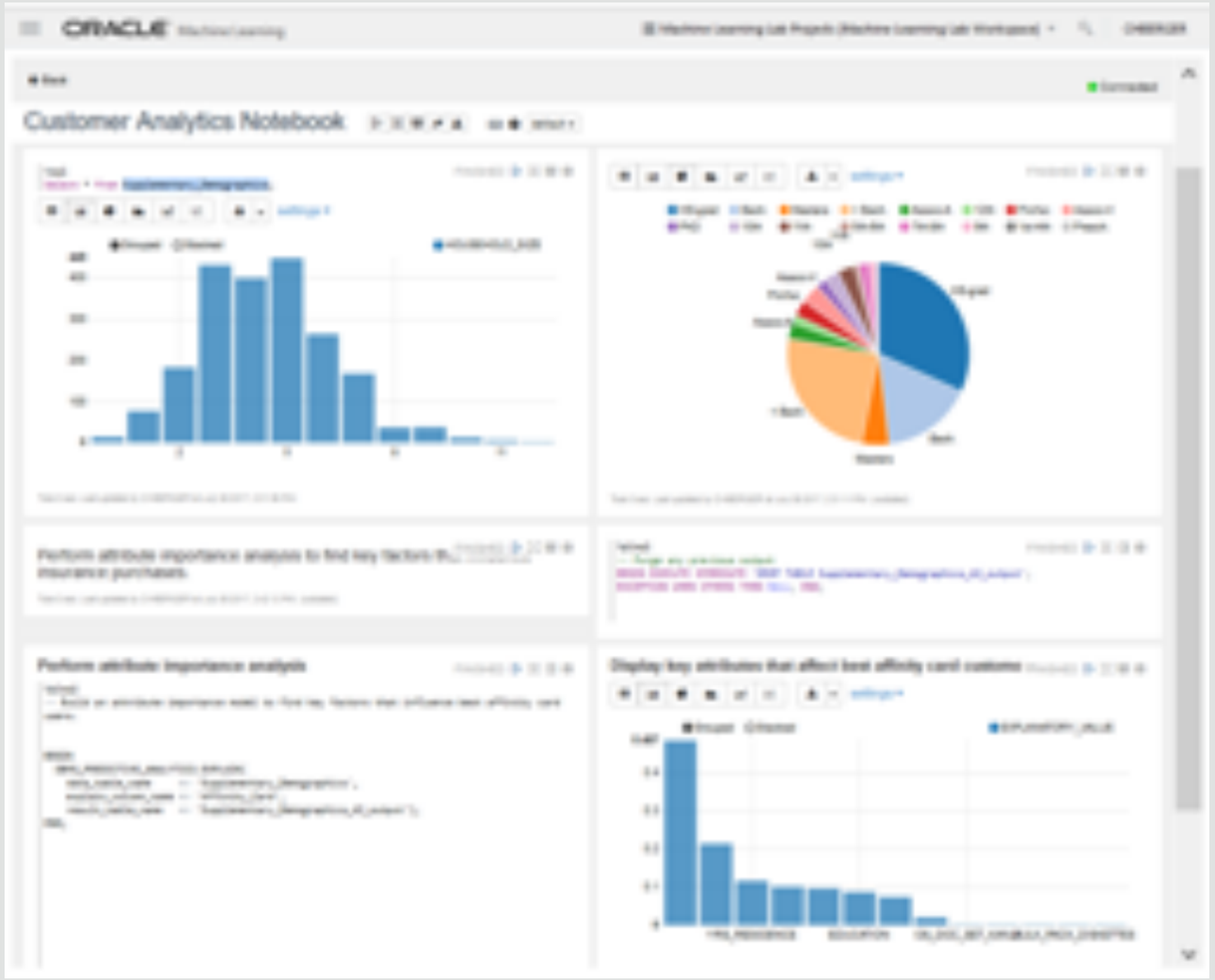
R Packages
Third-party R Packages through Embedded Execution
Spark MLlib algorithm integration

Model Deployment
SQL—1st Class Objects
Oracle RESTful API (ORDS)
OML Microservices (for Apps)

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

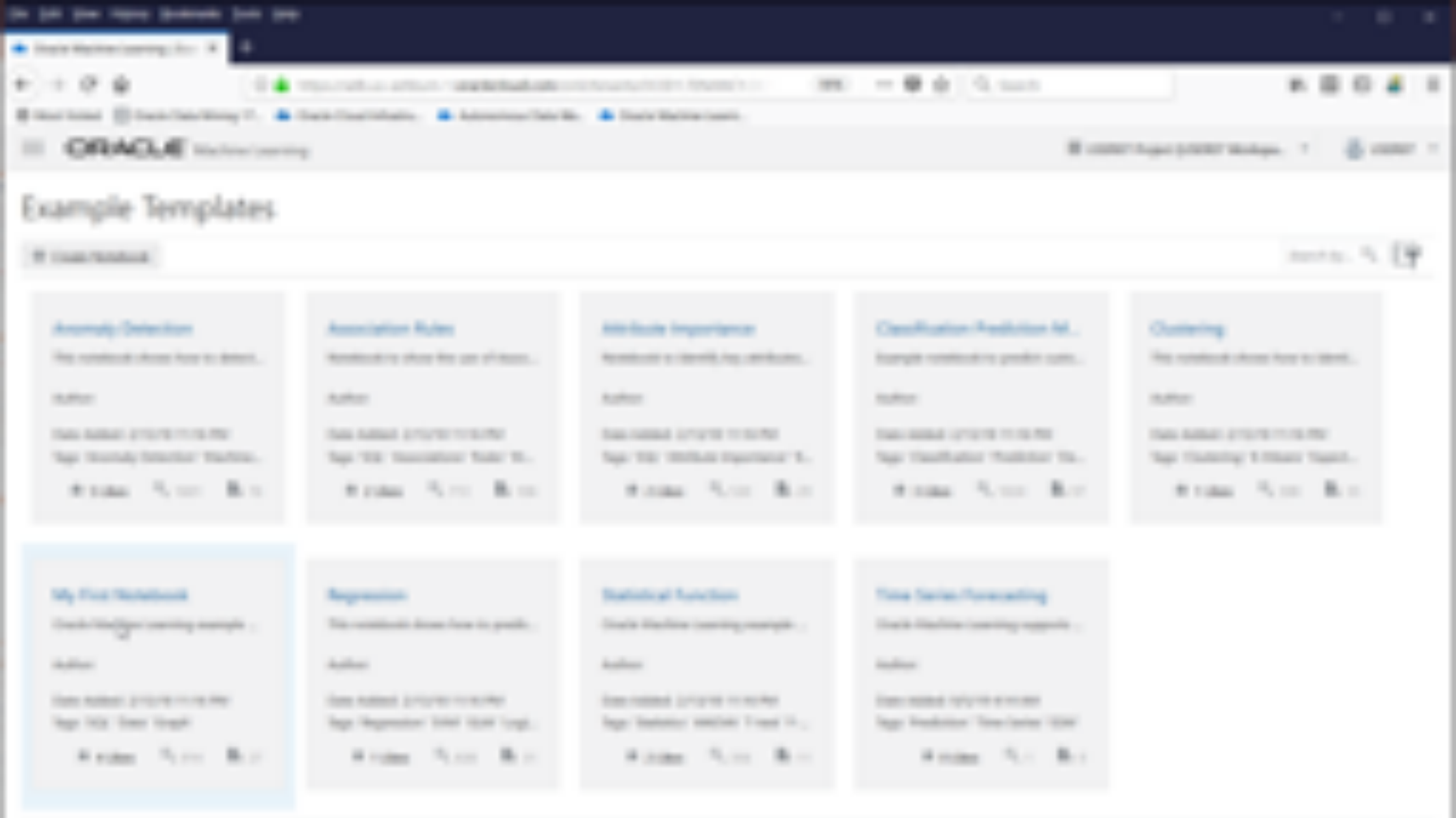
Oracle Machine Learning

Machine Learning Notebook for Autonomous Data Warehouse Cloud



Key Features:

- Collaborative UI for data scientists
  - Packaged with Autonomous Data Warehouse Cloud Easy access to shared notebooks, templates, permissions, scheduler, etc.
  - SQL ML algorithms API
- Supports deployment of ML analytics


Example Templates
Anomaly Detection
Association Rules
Attribute Importance
Classification/Prediction Al...
Clustering
My First Notebook
Regression
Statistical Function
Time Series Forecasting

Become a Data Scientist with OAC

http://ritt.md/OAC-datascience


ML in Action with OAC
Hello
Wine
Goodbye
Problems
http://ritt.md/OAC-ML-Video

Insights Lab

https://www.rittmanmead.com/insight-lab/