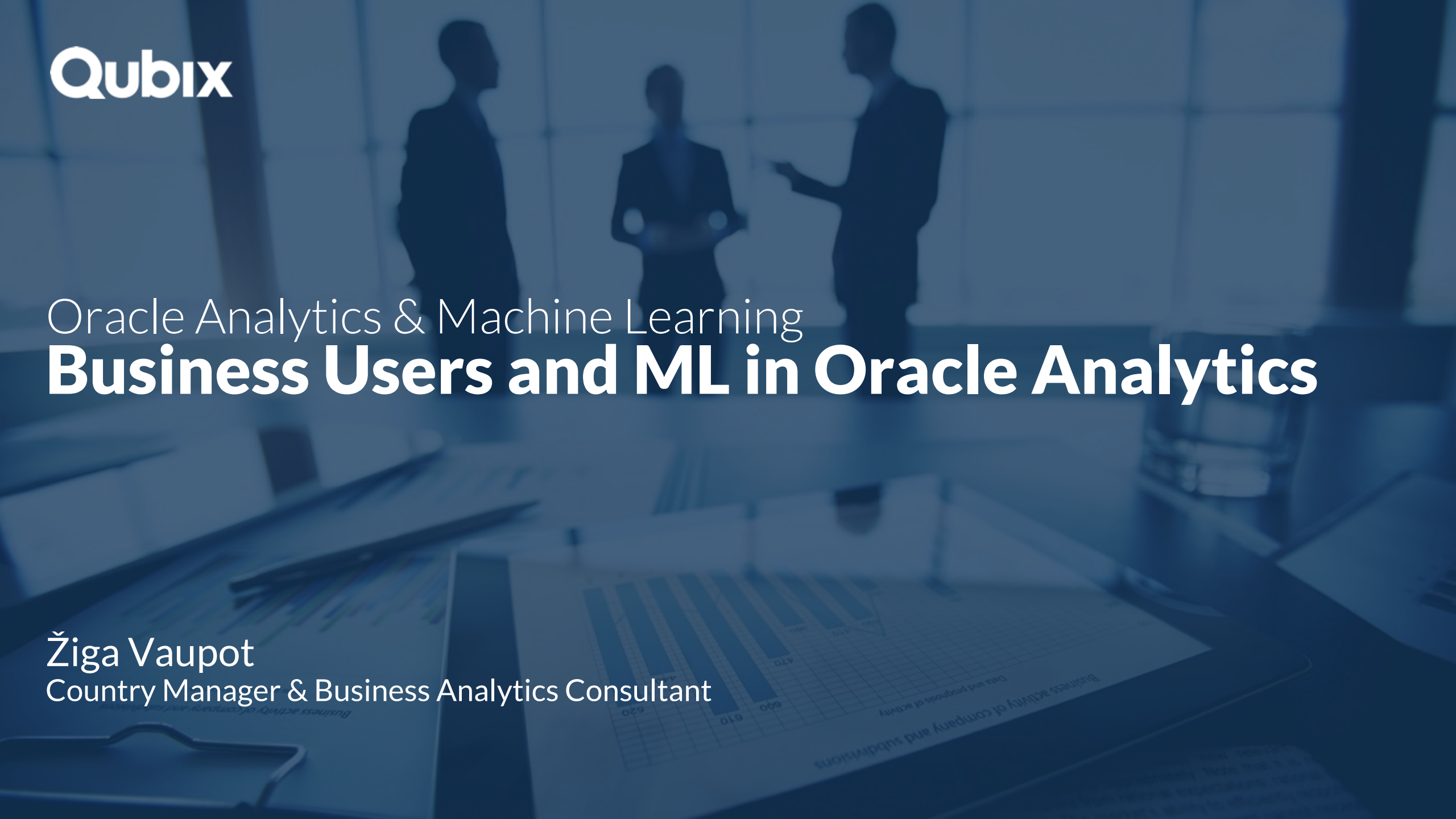


Qubix

Oracle Analytics & Machine Learning

Business Users and ML in Oracle Analytics

Žiga Vaupot
Country Manager & Business Analytics Consultant

Introductions

Žiga Vaupot

Qubix
Country Manager, Slovenia/CEE
Business Analytics Consultant

SIOUG
Board Member

ziga.vaupot@qubix.com
zigavaupot.blogspot.com
linkedin.com/in/zigavaupot

ORACLE ACE

Qubix

Introducing Qubix

Capabilities focused on ERP, EPM, Analytics and Data Engineering

1,000+ engagements

FOUNDED IN 1987

Offices in 5 countries

Global Oracle multi award-winning partner

Business Users and ML in Oracle Analytics

Typical users of Oracle Analytics

Business users

Enterprise type of reporting
Pixel-perfect
Predefined dashboarding and reporting

Business analysts

Effective data visualization and presentation
Ad-hoc
Self-service data preparation

Qubix

Machine Learning in Oracle Analytics

- (More or less) complete machine learning process cycle
- Effective data visualization
- Self-Service data preparation
- Self-Service model training and deployment
- Supports most popular supervised and unsupervised functions and algorithms (Classification, Regression, Clustering)
- Possibility to register, evaluate and use external data models (eg. from Autonomous DB)

Qubix

How typical user would most likely use ML in Oracle Analytics?

Business users

No idea how ML works, possibly quite limited understanding.

In most cases they are able to define business problems and ask questions.

Will wait for results to be presented and interpreted by others.

Business analysts

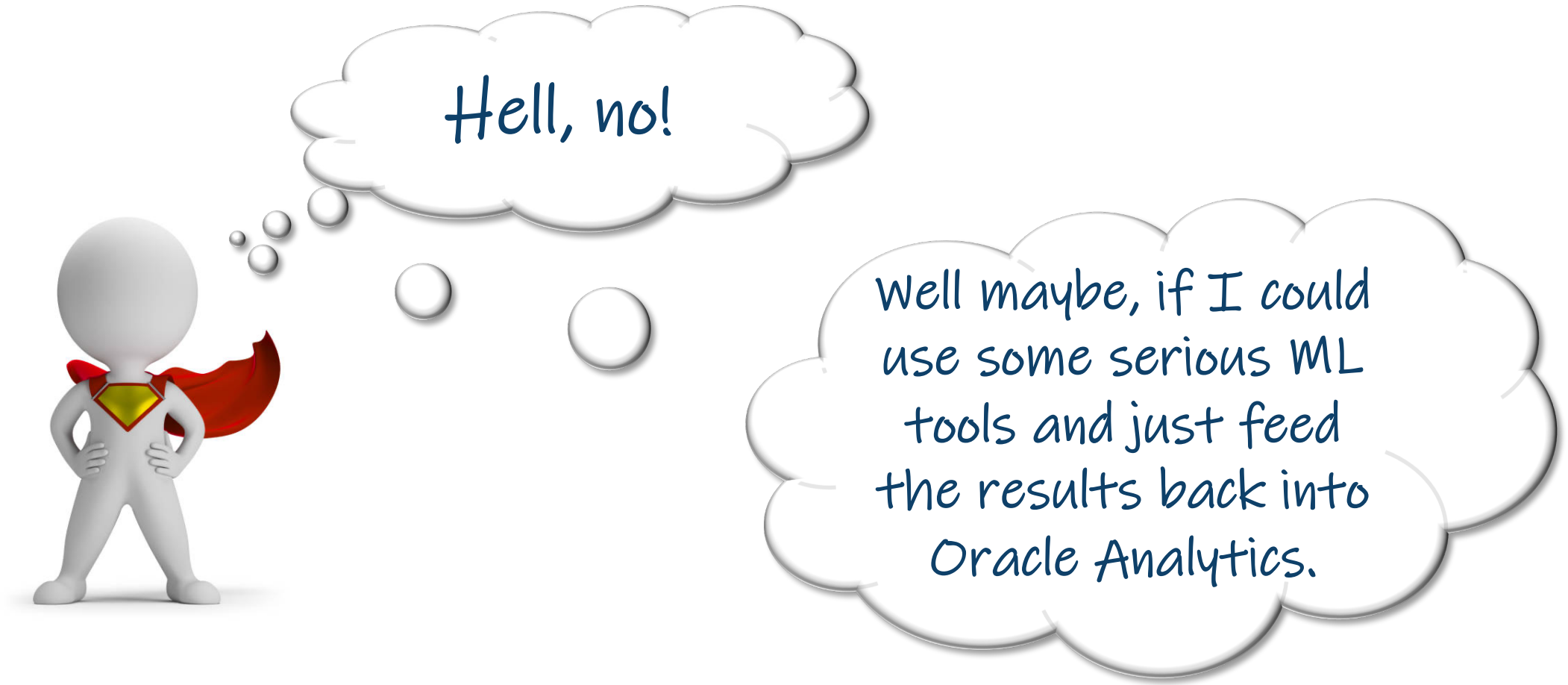
Possibly, have pretty good idea what ML is about and how it works.

Good candidates for self-service ML.

Will be able to understand the results and will use results in thier projects.

Qubix

A tool of choice for Data Scientists?



Qubix

So, what options have we got then?

a) Everything is done within Oracle Analytics

Very likely scenario:

Business analyst has enough business knowledge to define the the business problem.

Business analysis might posses enough (data flows/coding) skills to collect and prepare data. This one could be a tricky one (!).

Business analyst will be able to use Oracle Analytics toolset to train, evaluate and deploy ML models.

Business analyst might engage business user to obtain more business insights and/or data scientist for model improvements. With enough ML knowledge, business analyst might take option B.

b) Part of the process happens outside of Oracle Analytics

Very likely scenario:

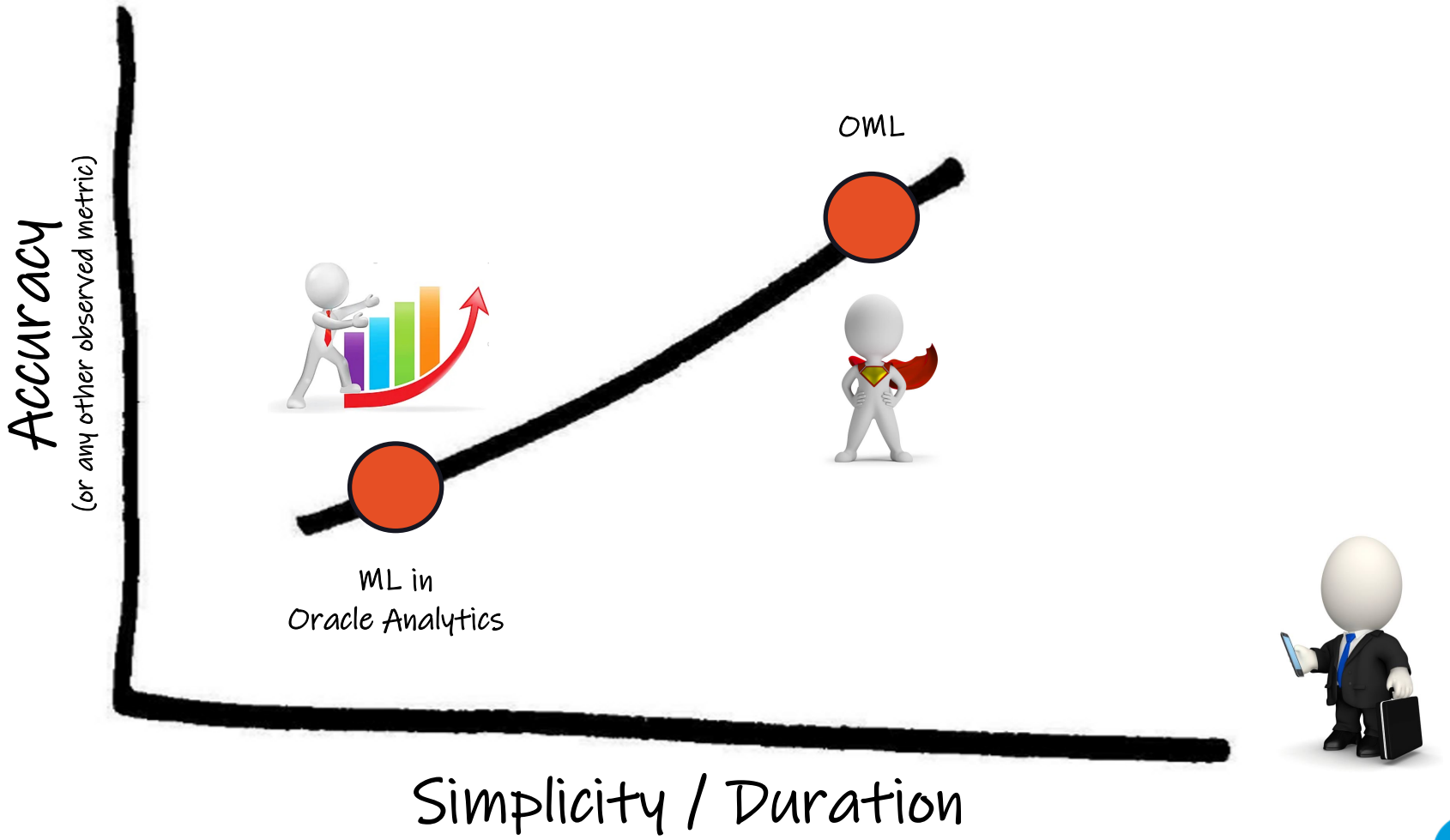
Business user will search for a data scientist (or will turn to Business analyst, but this is scenario A)

Data scientist will follow standard machine learning process using OML to come back with the best possible machine learning model. Model can already be deployed, in this case final results are presented to business user.

Model will be then registered or deployed results dataset can be created with Oracle Analytic. This can be done by (if skilled) Business user or more likely Business Analyst.

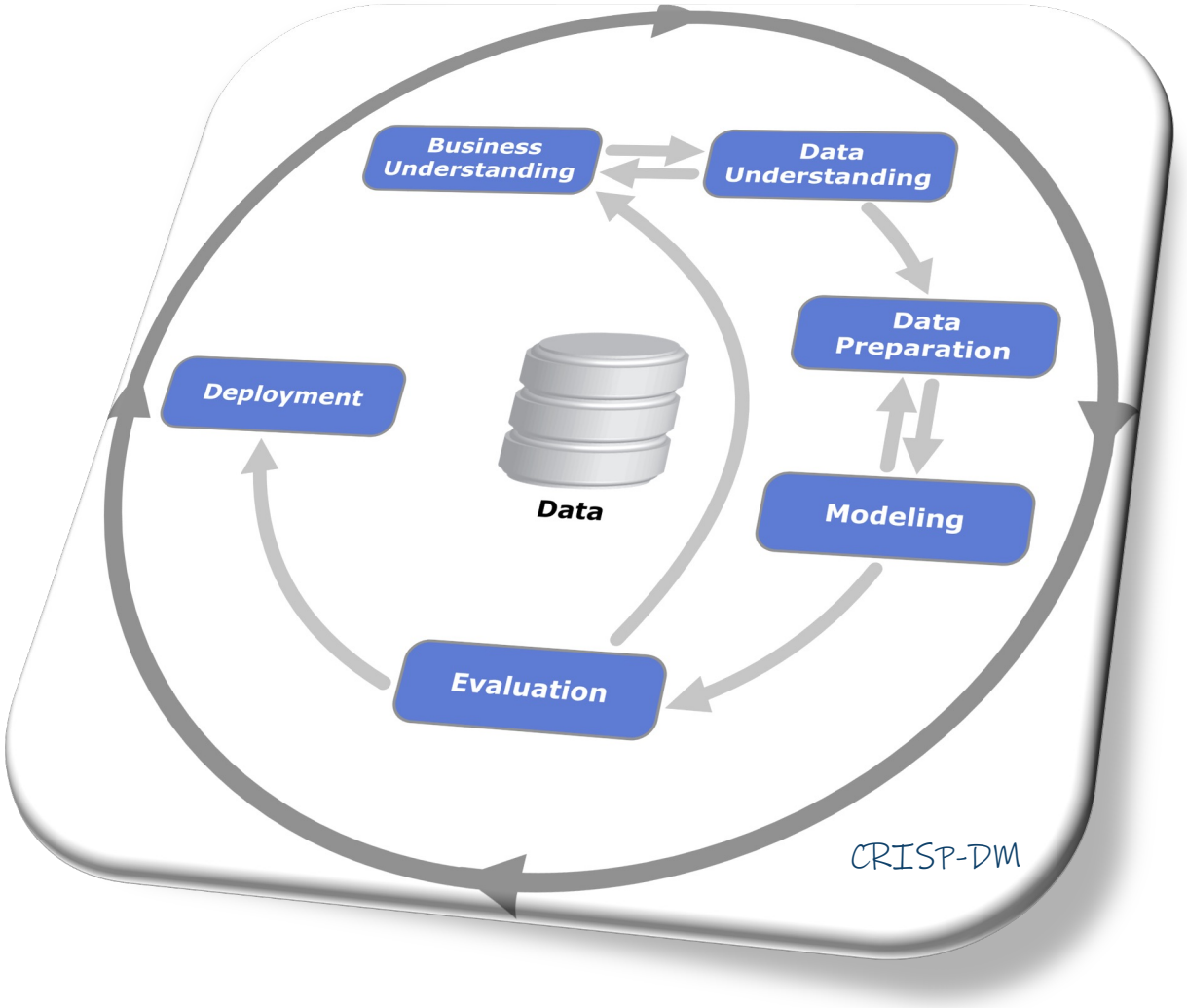
Qubix

Trade-off



Qubix

Regardless of scenario chosen, ML process applies!



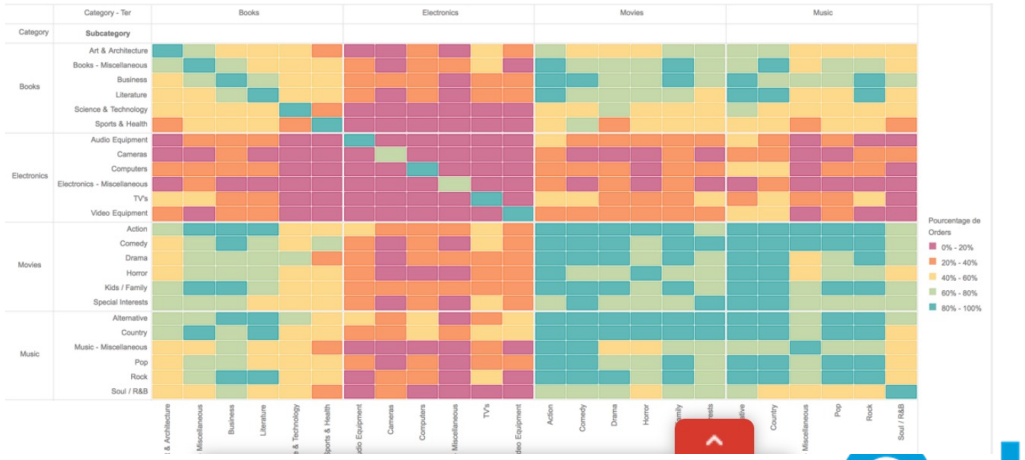
Qubix

Asking the Right Question

Which products/product subcategories are most often sold together?

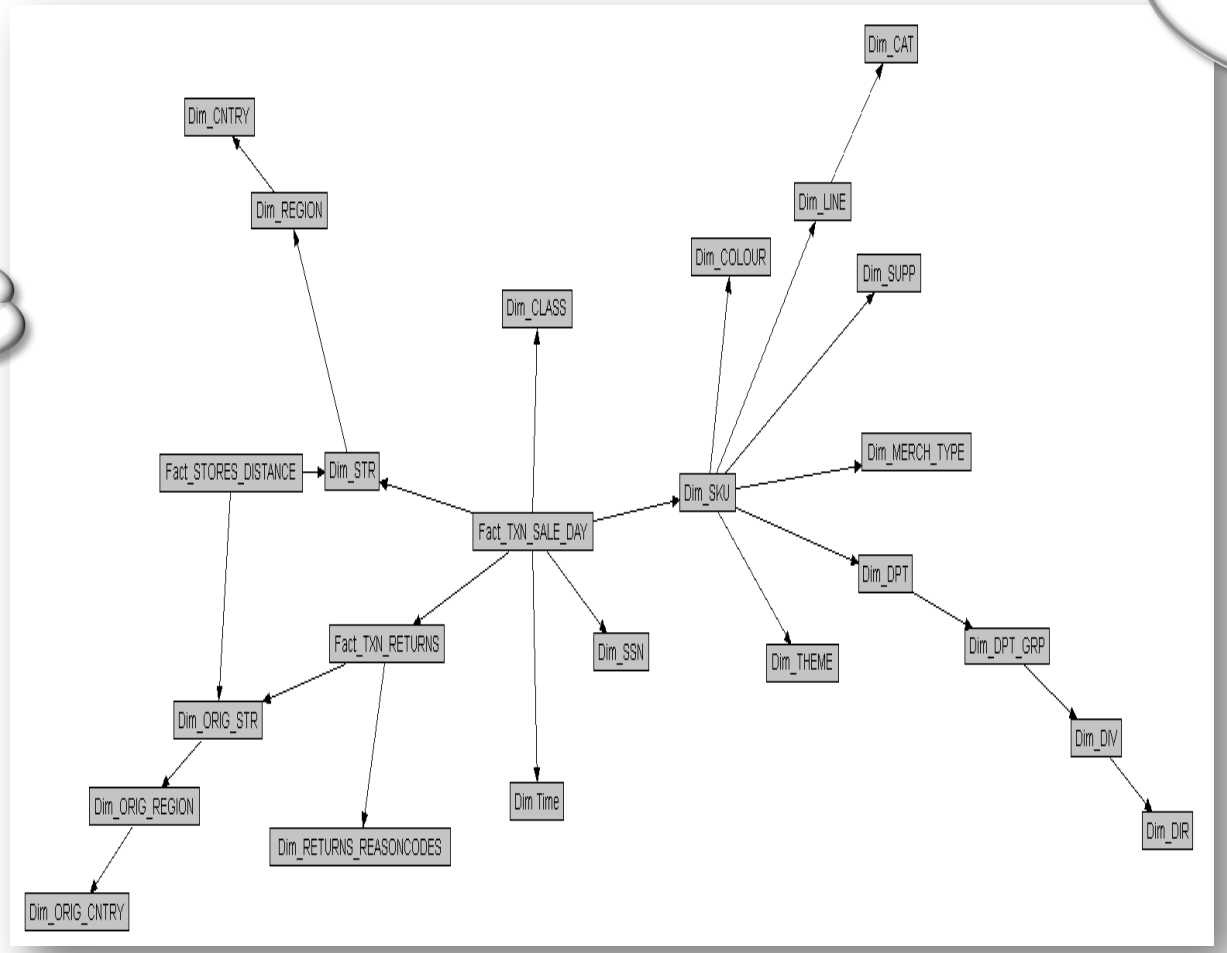
Which are the biggest opportunities for up/cross-sell?

Can we analyse this by each and every store by periods of the day?



Qubix

The Data



We are ready for some great predictions. We've got all of our data in a data warehouse!

Qubix

Data Collection & Exporatory Data Analysis

Oracle Analytics

- Data Visualizations: Correlation Matrix, Trendline (regression), Boxplot ...
- Explain, Autoinsights

Zeppelin Notebooks

- SQL, PL/SQL and Python for data wrangling
- SQL/Python Statistical Functions
- Visualizations

Qubix

Facts

Data preparation can take

60- 90 % of total ML project time.

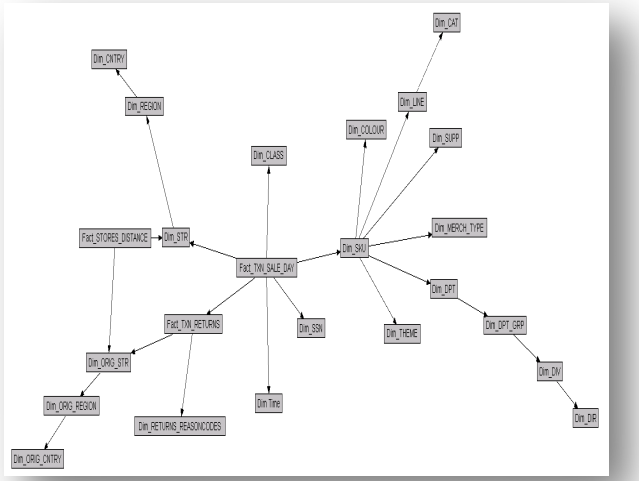
Data quality contributes **75% - 90%**

to overall ML project success.

Pyle, D, 2003: The handbook of data mining (Data Collection, Preparation, Quality, and Visualization)

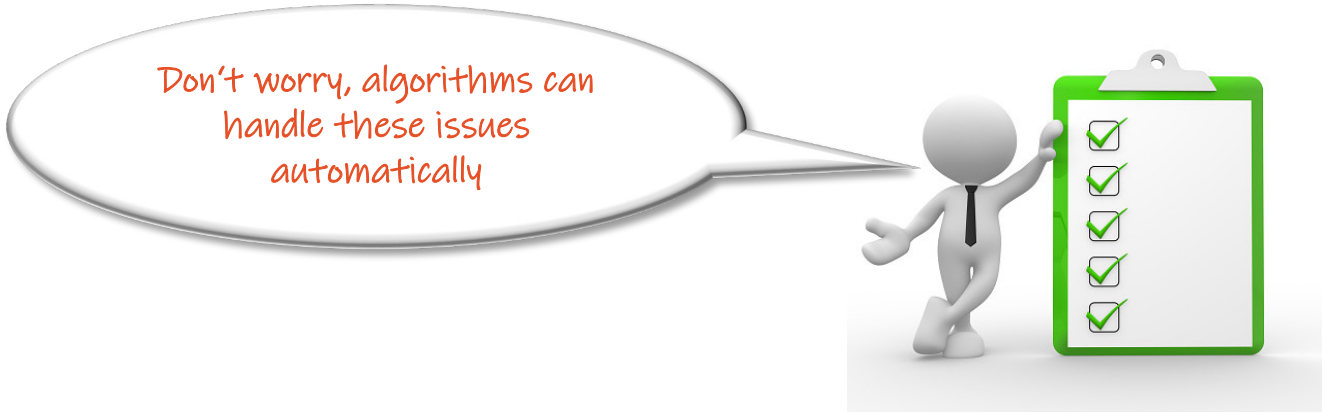
Qubix

As business analyst, these are things to consider

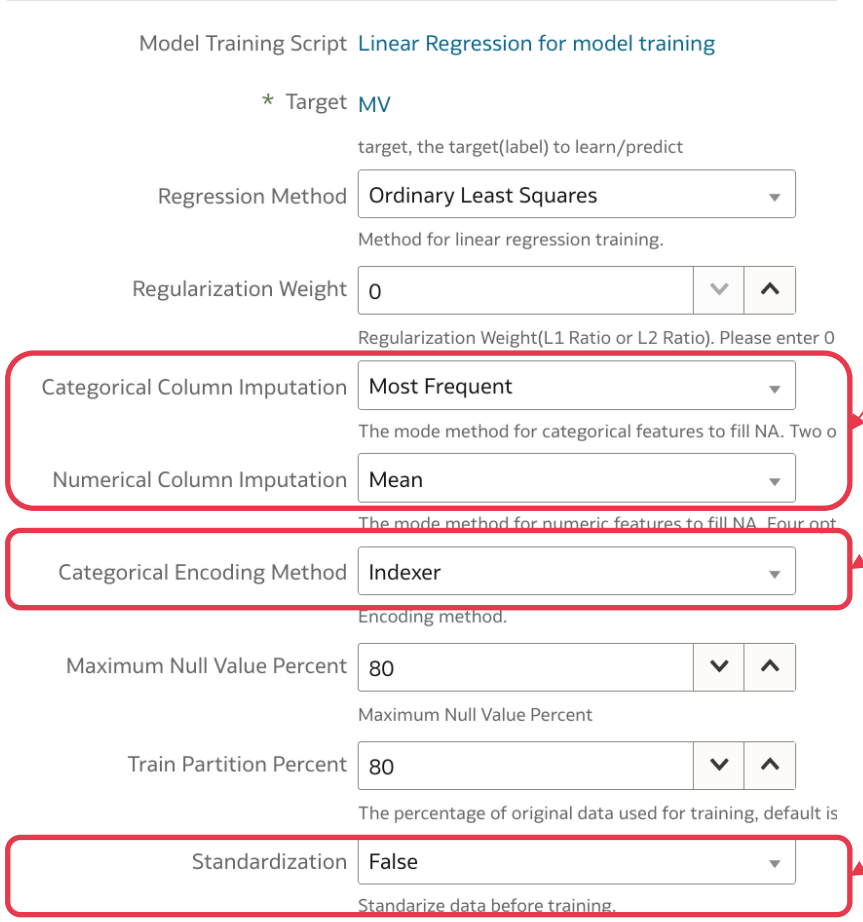


Feature 1	Feature 2	…	Feature *N*	Target

- Missing Values and Outliers
- Normalization
- Handling categorical values
- Dealing with correlations in data
- Transformations



Yes, this is true ...



Resolving missing values issue

Handling Categorical Values

Standardization/ Normalisation

... but pay attention ...

Which formula is used for standardization and what are the consequences?

$$v_i' = \frac{v_i}{10^k}$$

$$v_i' = \frac{v_i - min_v}{max_v - min_v} * (max_{v'} - min_{v'}) + min_{v'}$$

$$v_i' = \frac{v_i - \bar{V}}{\sigma_v}$$

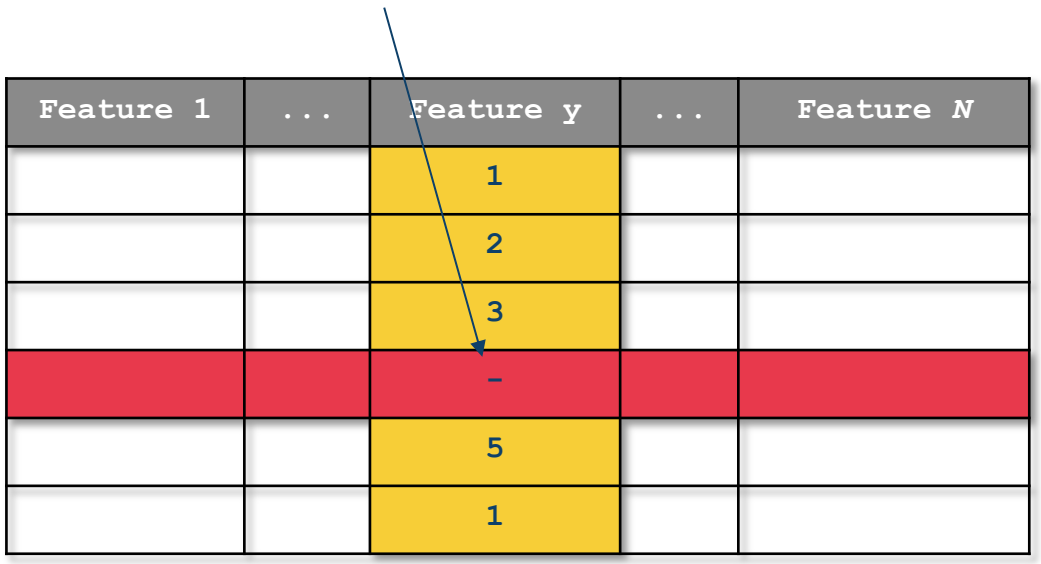
$$v_n = \frac{1}{1 + e^{-v_i}}$$

(replace v_i in formula above with $v_t = \frac{(v_i - V)}{\lambda(\sigma_V/2\pi)}$)

Qubix

… or consider the following:

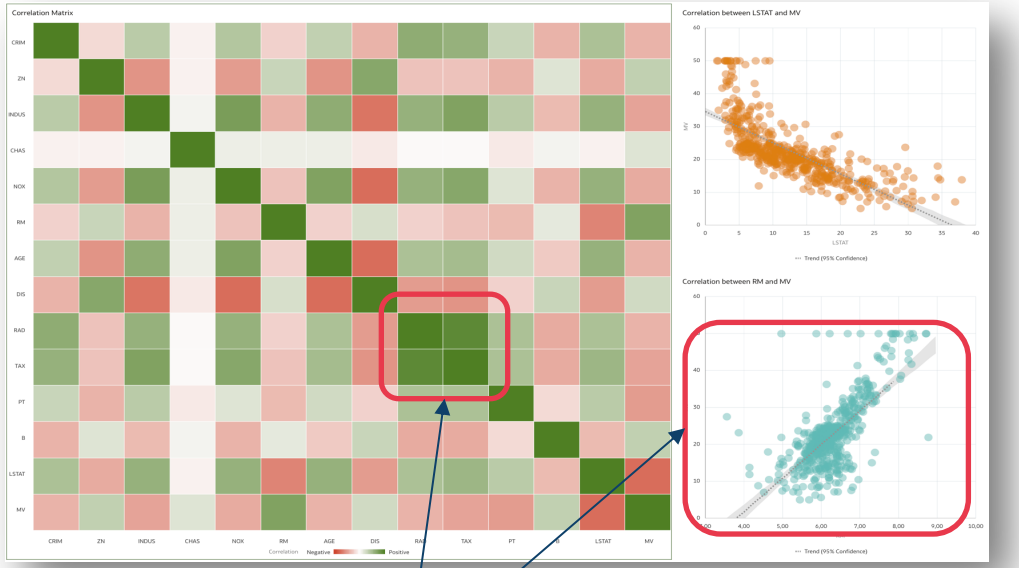
Which value should replace missing values?



- Remove row with missing value or
- Replace missing value with
  - some generic constant
  - some attribute statistics like average or median
  - some stitistic value that belongs to the same specific group. For example, customer segment average.
  - the most probable value.
  - a value that retains characteristics of the sample. For example, calculate a value that retains variability.

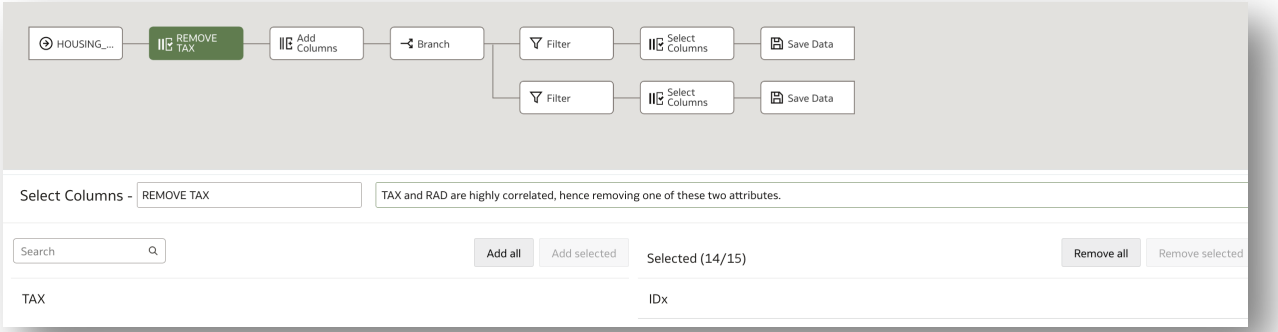
Qubix

… and don‘t forget, correlations won‘t be removed automatically.



High correlation identified between two columns

Data flows is a perfect self-service tool to perform data preparation!



Data Scientist will follow the best practices

- ... using SQL and/or python for required data preparation transformations.
- There might be other data preparation challenges to deal with which would require:
  - data sampling
  - dimensionality reduction
  - aggregations
  - etc.
- When running OML algorithms, in most cases, PREP_AUTO parameter will be set to ON. This will automatically perform data transformations required by specific algorithm.

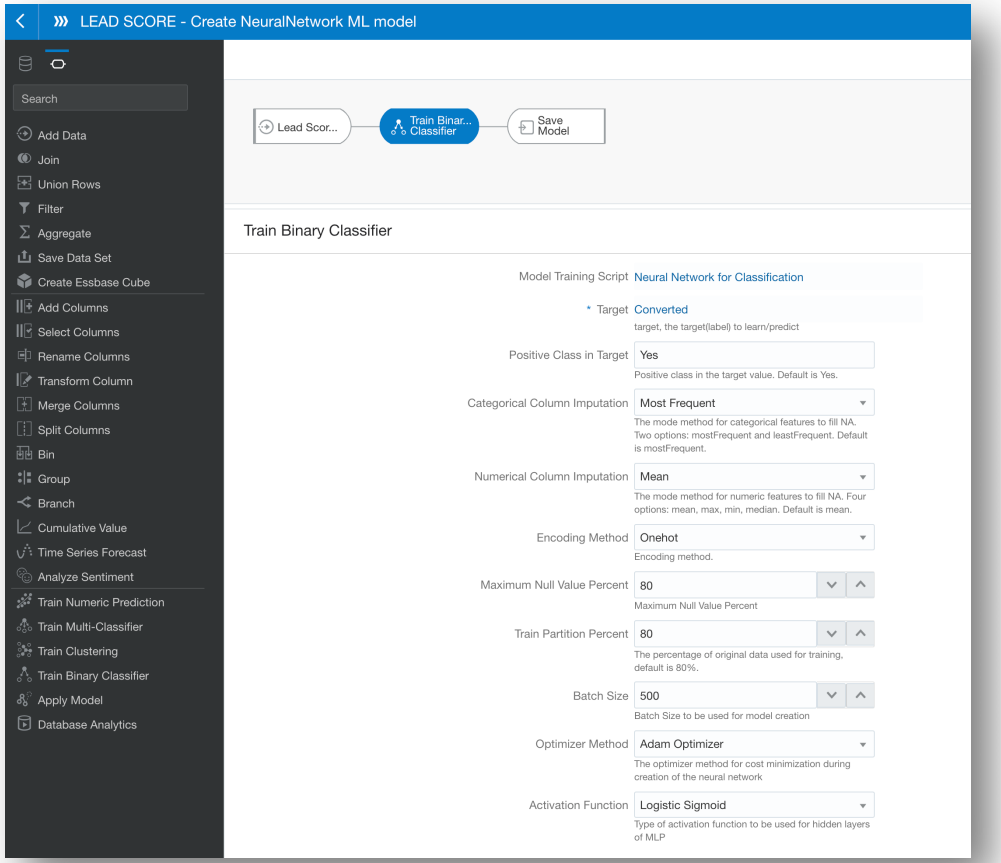
```
%python

try:
    oml.drop(model="ANOMALY_DETECTION_MODEL_CLAIMS0")
except:
    print("Model does not exist")

odm_settings = {'prep_auto' : 'ON'}

svm_mod = oml.svm("anomaly_detection", **odm_settings)
svm_mod.fit(CLAIMS0, None, model_name = 'ANOMALY_DETECTION_MODEL
```

```
Model does not exist

Model Name: ANOMALY_DETECTION_MODEL_CLAIMS0

Model Owner: ML_USER

Algorithm Name: Support Vector Machine

Mining Function: ANOMALY_DETECTION

Settings:
                    setting name                 setting value
0                      ALGO_NAME  ALGO_SUPPORT_VECTOR_MACHINES
1                   ODMS_DETAILS                  ODMS_ENABLE
2  ODMS_MISSING_VALUE_TREATMENT       ODMS_MISSING_VALUE_AUTO
3                  ODMS_SAMPLING        ODMS_SAMPLING_DISABLE
4                      PREP_AUTO                           ON
5             SVMS_CONV_TOLERANCE                        .0001
6             SVMS_KERNEL_FUNCTION                  SVMS_LINEAR
7                SVMS_OUTLIER_RATE                          .01
```

Qubix

ML model training in Oracle Analytics

- Oracle Analytics offers several machine learning algorithms for supervised and unsupervised machine learning
  - Binary and Multivariate Classification
  - Regression
  - Clustering
  - Anomaly Detection*
  - Frequent Itemsets & Association Rules*
  - Time Series
- Limited parametrisation
- Train/test split and evaluation is done automatically
- Additional datasets are created for better explainability



* Requires Oracle Database

Modeling in database and things that come very handy

```
CREATE TABLE DATA_4_MPO_HOUR_BAND
      PARTITION BY LIST(CLANICA)
      (
        PARTITION PO VALUES ('651187'),
        PARTITION VU VALUES ('651188'),
        PARTITION BB VALUES ('651189'),
        PARTITION SE VALUES ('651190'),
        PARTITION EU VALUES ('651191'),
        PARTITION VP VALUES ('651192'),
        PARTITION KA VALUES ('651193'),
        PARTITION ST VALUES ('651194'),
        PARTITION MO VALUES ('651195'),
        PARTITION TE VALUES ('651196'),
        PARTITION ME VALUES ('651197'),
        PARTITION LE VALUES ('651198'),
        PARTITION PR VALUES ('651199')
      )
      AS
    (SELECT DISTINCT p.k_clanica CLANICA,
                     p.k_clanica || p.k_ob
                     o.objekt_l0_clanicao_
                     c.clanica_l0_id || o.
                     a.artikal_l4_subcat_i
                     t.TIME_PART_OF_DAY HO
                    FROM BI_DW_PROD.F_PROD
                         BI_DW_PROD.D_ARTI
                         BI_DW_PROD.D_CLAN
                         BI_DW_PROD.D_OBJE
                         BI_DW_PROD.D_DATE
                         BI_DW_PROD.D_TIME
                    WHERE c.k_clanica = p.
                    AND p.K_OBJEKT = o.K_O
                    AND a.k_artikal = p.k_
                    AND p.K_INV_DATE = d.K
                    AND p.K_INV_TIME = t.K
                    AND d.DATE_DATE > TO_C
                    AND a.ARTIKAL_L4_SUBCA
    )
```

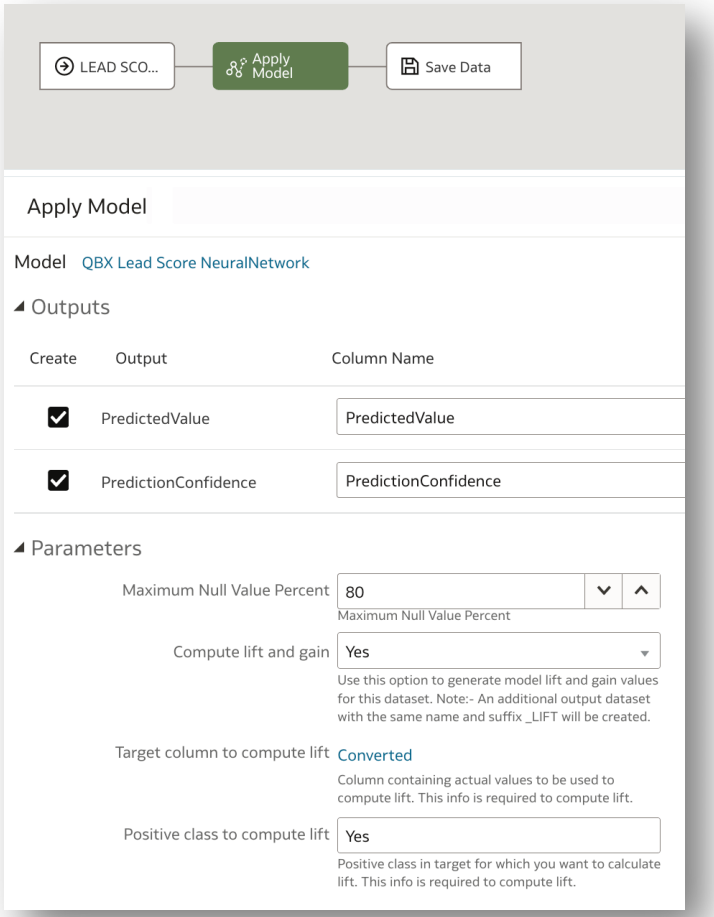
Create Association Rules model

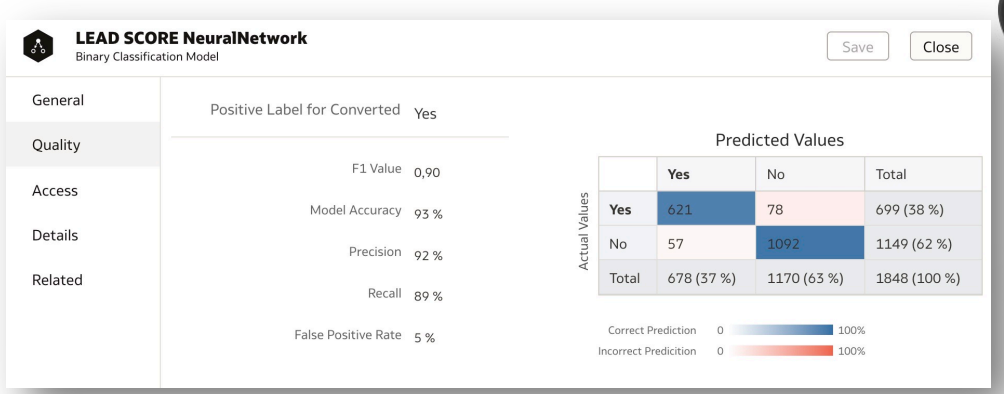
```
%script

/* Drop model AR_F_PRODAJA */

BEGIN DBMS_DATA_MINING.DROP_MODEL('AR_4_MPO_HOUR_BAND');
EXCEPTION WHEN OTHERS THEN NULL; END;
/


DECLARE
    v_setlst DBMS_DATA_MINING.SETTING_LIST;
BEGIN

    v_setlst('ALGO_NAME') := 'ALGO_APRIORI_ASSOCIATION_RULES';
    v_setlst('PREP_AUTO') := 'ON';
    v_setlst('ASSO_MIN_SUPPORT') := '0.0001';
    v_setlst('ASSO_MIN_CONFIDENCE') := '0.0001';
    v_setlst('ASSO_MAX_RULE_LENGTH'):= '2';
    v_setlst('ODMS_ITEM_ID_COLUMN_NAME'):= 'ARTIKAL';
    v_setlst('ODMS_PARTITION_COLUMNS'):= 'MPO,HOUR_BAND';
    v_setlst('ODMS_MAX_PARTITIONS') := '3000';

    DBMS_DATA_MINING.CREATE_MODEL2(
        MODEL_NAME => 'AR_4_MPO_HOUR_BAND',
        MINING_FUNCTION => 'ASSOCIATION',
        DATA_QUERY => 'SELECT * FROM DATA_4_MPO_HOUR_BAND',
        SET_LIST => v_setlst,
        CASE_ID_COLUMN_NAME => 'RACUN');
END;
```

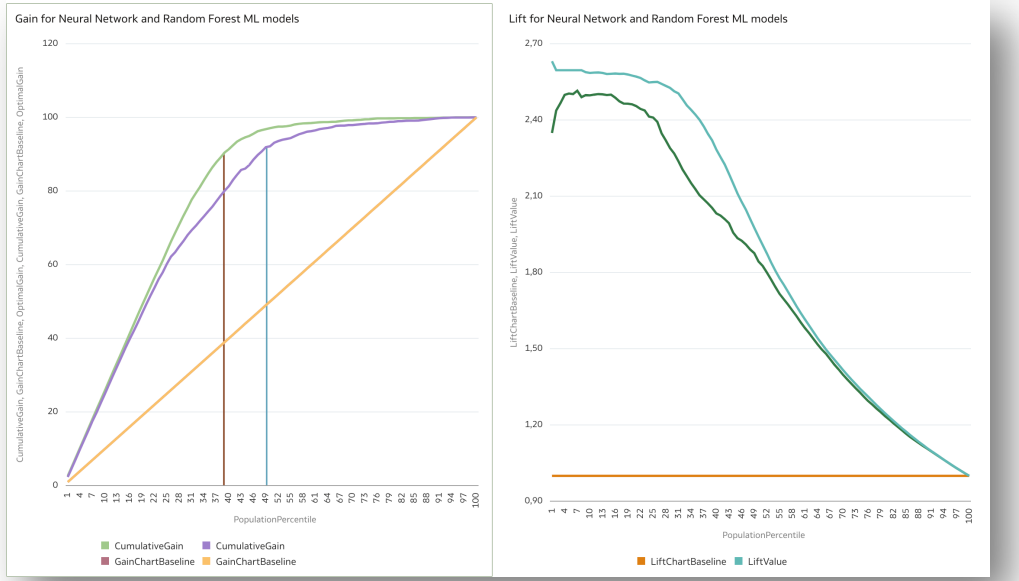
- Data stays where it is, in database
- SQL can be used at what it is the best
- Automatic data preparation is supported out-of-the-box
- Model partitioning allows building several models at the same time

Qubix

Oracle Analytics: scoring is done automatically







OML: scoring is done in traditional way – using SQL

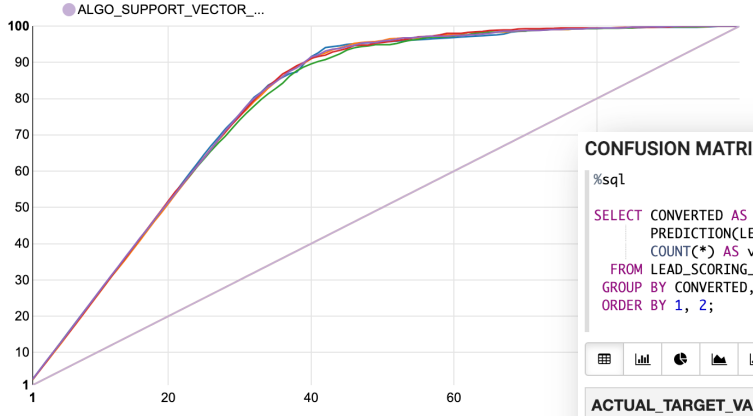
```
-- Score the data and compute lift
BEGIN
DBMS_DATA_MINING.APPLY('LEAD_SCORING_NN_MODEL',
                       'LEAD_SCORING_TEST_DATA',
                       'PROSPECT_ID',
                       'LEAD_SCORING_NN_RESULT');
DBMS_DATA_MINING.COMPUTE_LIFT('LEAD_SCORING_NN_RESULT',
                              'LEAD_SCORING_TEST_DATA',
                              'PROSPECT_ID',
                              'CONVERTED',
                              'LEAD_SCORING_NN_LIFT',
                              'Yes',
                              'PREDICTION',
                              'PROBABILITY',100);
```

```
%sql

SELECT A.ALGO_NAME, A.QUANTILE_NUMBER, ROUND(A.GAIN_CUMULATIVE*100,2) GAIN_CUMULATIVE, B
    .RANDOM_GUESS
FROM LEAD_SCORING_ALL_LIFT_DATA A
, (SELECT ALGO_NAME, QUANTILE_NUMBER, QUANTILE_NUMBER RANDOM_GUESS FROM LEAD_SCORING_ALL_LIFT_DATA)
    B
WHERE A.QUANTILE_NUMBER = B.QUANTILE_NUMBER AND A.ALGO_NAME = B.ALGO_NAME
```

FINISHED

settings

ALGO_RANDOM_FOREST.G... ALGO_RANDOM_FOREST.R... ALGO_GENERALIZED_LIN...
ALGO_GENERALIZED_LIN... ALGO_NEURAL_NETWORK_... ALGO_NEURAL_NETWORK_...
ALGO_NAIVE_BAYES.GAI... ALGO_NAIVE_BAYES.RAN... ALGO_SUPPORT_VECTOR_...
ALGO_SUPPORT_VECTOR_...



CONFUSION MATRIX

```
%sql

SELECT CONVERTED AS actual_target_value,
       PREDICTION(LEAD_SCORING_NN_MODEL USING *) AS predicted_target_value,
       COUNT(*) AS value
  FROM LEAD_SCORING_TEST_DATA
 GROUP BY CONVERTED, PREDICTION(LEAD_SCORING_NN_MODEL USING *)
 ORDER BY 1, 2;
```

settings

ACTUAL_TARGET_VALUE	PREDICTED_TARGET_VALUE	VALUE
No	No	1670
No	Yes	129
Yes	No	147
Yes	Yes	1000

```
%sql

SELECT 'LEAD_SCORING_NN_MODEL' Model,
    (TP+TN)/(TP+TN+FP+FN) Accuracy,
    TP/(TP+FN) Precision,
    TP/(TP+FP) Recall,
    2 * (TP/(TP+FN)*TP/(TP+FP))/(TP/(TP+FN)+TP/(TP+FP)) F1_Score,
    FN/(FN+TP) False_Positive_Rate
FROM (
with
tp as (SELECT tp_value FROM (SELECT CONVERTED AS actual_target_value,
       PREDICTION(LEAD_SCORING_NN_MODEL USING *) AS predicted_target_value,
       COUNT(*) AS tp_value
  FROM LEAD_SCORING_TEST_DATA
 GROUP BY CONVERTED, PREDICTION(LEAD_SCORING_NN_MODEL U
 ORDER BY 1, 2)
 WHERE actual_target_value = 'Yes' and predicted_target
,
fp as (SELECT fp_value FROM (SELECT CONVERTED AS actual
       PREDICTION(LEAD_SCORING_NN_MODEL USING *) AS pre
       COUNT(*) AS fp_value
  FROM LEAD_SCORING_TEST_DATA
 GROUP BY CONVERTED, PREDICTION(LEAD_SCORING_NN_MODEL U
 ORDER BY 1, 2)
 WHERE actual_target_value = 'No' and predicted_target_
,
fn as (SELECT fn_value FROM (SELECT CONVERTED AS actual
       PREDICTION(LEAD_SCORING_NN_MODEL USING *) AS pre
       COUNT(*) AS fn_value
  FROM LEAD_SCORING_TEST_DATA
 GROUP BY CONVERTED, PREDICTION(LEAD_SCORING_NN_MODEL U
 ORDER BY 1, 2)
 WHERE actual_target_value = 'Yes' and predicted_target
,
tn as (SELECT tn_value FROM (SELECT CONVERTED AS actual
       PREDICTION(LEAD_SCORING_NN_MODEL USING *) AS pre
       COUNT(*) AS tn_value
  FROM LEAD_SCORING_TEST_DATA
 GROUP BY CONVERTED, PREDICTION(LEAD_SCORING_NN_MODEL USING *)
 ORDER BY 1, 2)
 WHERE actual_target_value = 'No' and predicted_target_value = 'No')
SELECT tp.tp_value TP, fp.fp_value FP, fn.fn_value FN, tn.tn_value TN
FROM tp, fp, fn, tn)
```

```
%sql

SELECT VIEW_NAME, VIEW_TYPE
  FROM USER_MINING_MODEL_VIEWS
  WHERE MODEL_NAME = 'LEAD_SCORING_RF_MODEL'
  ORDER BY VIEW_NAME;
```

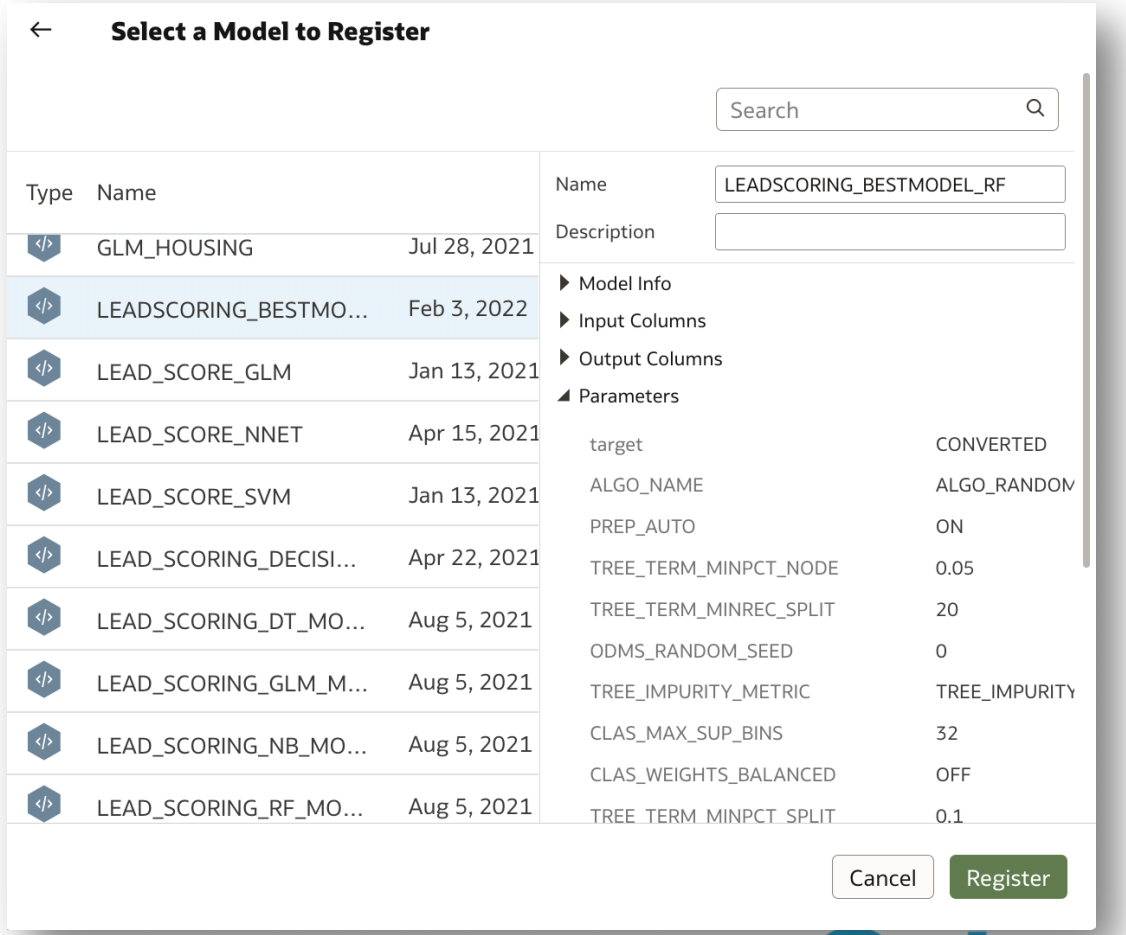
settings

VIEW_NAME	VIEW_TYPE
DM$VALEAD_SCORING_RF_MODEL	Variable Importance
DM$VCLEAD_SCORING_RF_MODEL	Scoring Cost Matrix
DM$VGLEAD_SCORING_RF_MODEL	Global Name-Value Pairs
DM$VSLEAD_SCORING_RF_MODEL	Computed Settings
DM$VTLEAD_SCORING_RF_MODEL	Classification Targets
DM$VWLEAD_SCORING_RF_MODEL	Model Build Alerts

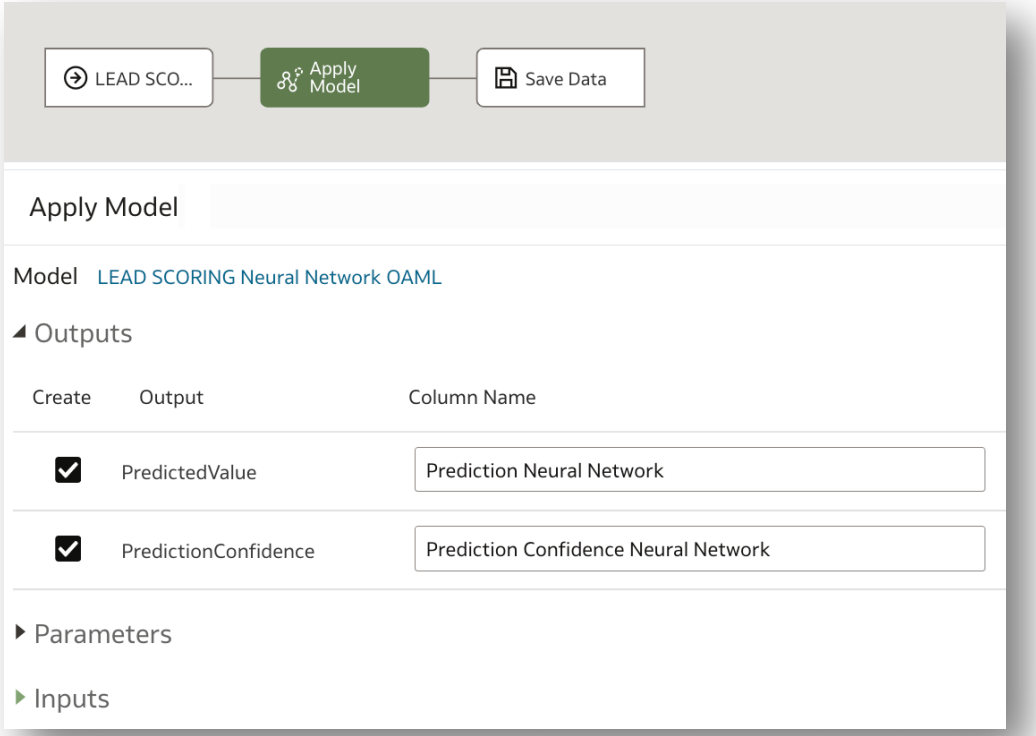
Qubix

Deployment

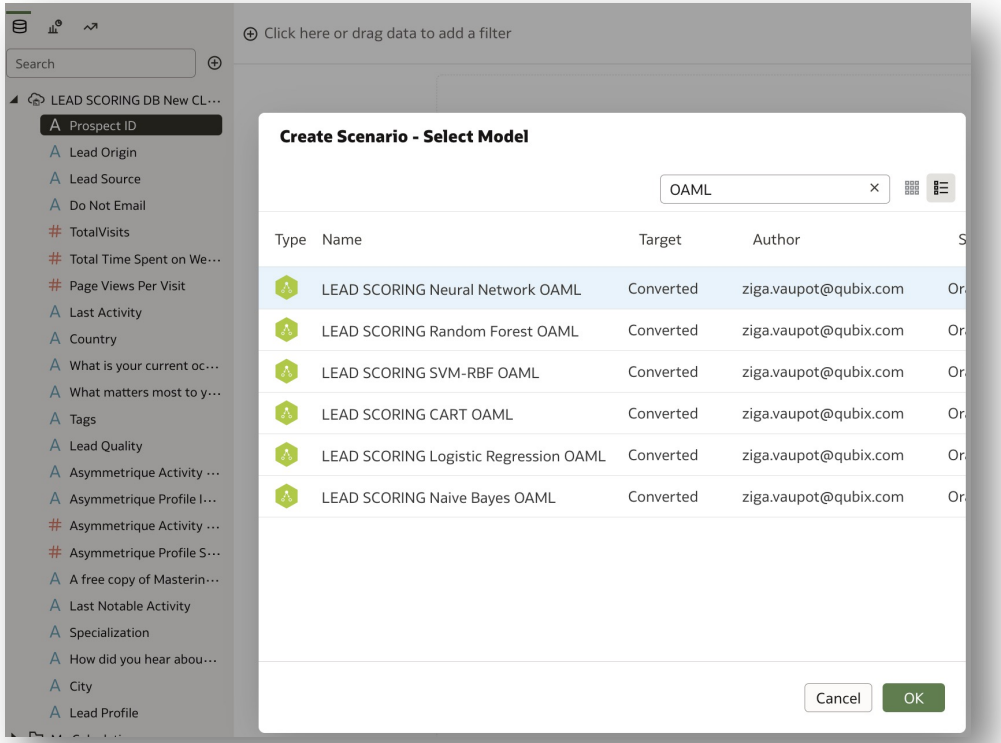
- In Oracle Analytics, models are ready to deploy instantly
- ML models that were created in database must be registered with Oracle Analytics first



From here on, the steps are the same ...



- Create a new data flow and Apply Model to generate a new "predicted data" dataset
- Use „predicted data" dataset in a new workbook



- Create a new workbook and add Scenario, which is ML model application on-the-fly
- This works only with Oracle Analytics ML models

Qubix

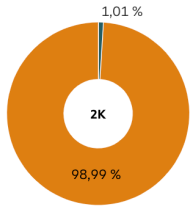
End results

Suspicious Claims

Data | Visualize | Present

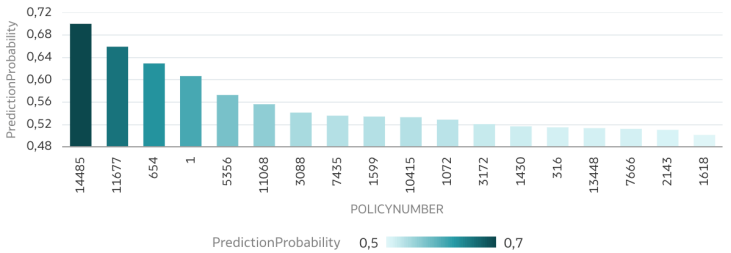
Click here or drag data to add a filter

# of Claims by Prediction



PredictionProbability by POLICYNUMBER, PredictionProbability

Prediction: 0



Prediction: 0

# of Claims

18

PredictionProbability, POLICYNUMBER, Prediction, Attribute_1, Attribute_1_Weight, Attribute_1_Condition, Attribute_2, Attribute_2_Weight, Attribute_2_Condition, Attribute_3, Attribute_3_Weight, Attribute_3_Condition, Attribute_4, Attribute_4_We...

Prediction: 0

POLICYNUMBER	Prediction	PredictionProbability	Attribute_1	Attribute_1_Weight	Attribute_1_Condition	Attribute_2	Attribute_2_Weight	Attribute_2_Condition	Attribute_3	Attribute_3_Weight	Attribute_3_Condition	Attribute_4
14485	0	0,7002	DAYSPOLICYCLAIM	0,12	: Value = 15to30	DAYSPOLICYACCIDENT	0,12	: Value = 8to15	ADDRESSCHANGECLAIM	0,09	: Value = 2to3years	SEX
11677	0	0,6591	DAYSPOLICYCLAIM	0,12	: Value = 15to30	DAYSPOLICYACCIDENT	0,12	: Value = 15to30	NUMBEROFCARS	0,09	: Value = 2vehicles	ADDRESSCHANGE
654	0	0,6293	DAYSPOLICYCLAIM	0,13	: Value = 15to30	DAYSPOLICYACCIDENT	0,13	: Value = 15to30	ACCIDENTAREA	0,07	: Value = Rural	VEHICLEPRICE
1	0	0,6067	ADDRESSCHANGECLAIM	0,10	: Value = 1year	NUMBEROFCARS	0,10	: Value = 3to4	SEX	0,07	: Value = Female	VEHICLECATEGOR
5356	0	0,5729	ADDRESSCHANGECLAIM	0,10	: Value = 2to3years	NUMBEROFCARS	0,10	: Value = 2vehicles	SEX	0,07	: Value = Female	FAULT
11068	0	0,5564	POLICEREPORTFILED	0,12	: Value = Yes	NUMBEROFCARS	0,09	: Value = 2vehicles	ADDRESSCHANGECLAIM	0,09	: Value = 4to8years	ACCIDENTAREA
3088	0	0,5414	NUMBEROFCARS	0,09	: Value = 2vehicles	ADDRESSCHANGECLAIM	0,09	: Value = 4to8years	ACCIDENTAREA	0,07	: Value = Rural	SEX
7435	0	0,5359	NUMBEROFCARS	0,10	: Value = 2vehicles	ADDRESSCHANGECLAIM	0,10	: Value = 4to8years	ACCIDENTAREA	0,07	: Value = Rural	VEHICLEPRICE
1599	0	0,5343	DAYSPOLICYCLAIM	0,13	: Value = 15to30	DAYSPOLICYACCIDENT	0,13	: Value = 8to15	VEHICLECATEGORY	0,04	: Value = Sport	AGEOFPOLICYHO
10415	0	0,5332	DAYSPOLICYCLAIM	0,13	: Value = 8to15	DAYSPOLICYACCIDENT	0,13	: Value = none	NUMBEROFCARS	0,10	: Value = 3to4	MARITALSTATUS
1072	0	0,5287	ADDRESSCHANGECLAIM	0,10	: Value = 2to3years	NUMBEROFCARS	0,10	: Value = 2vehicles	ACCIDENTAREA	0,07	: Value = Rural	AGEOFPOLICYHO

PredictionProbability 0,5 0,7

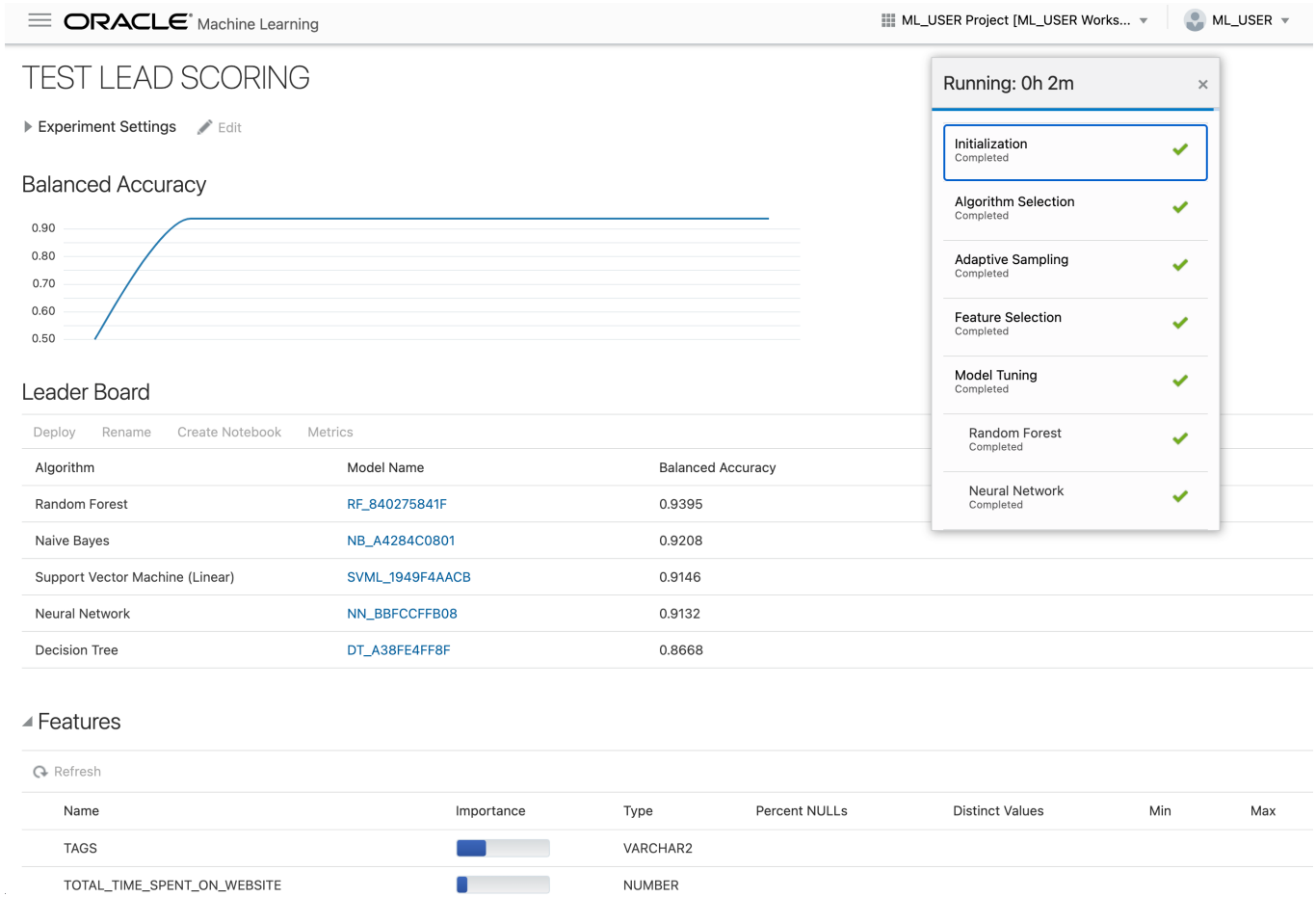
Suspicious Claims Summary | Suspicious Claims Details

2 Slices

Qubix

AutoML

Bringing machine learning self-service to database (?)



- Automated ML model creation
  - Algorithm selection
  - Adaptive sampling
  - Feature extraction
  - Hyperparameter optimisation
- Basic understanding of machine learning process
  - However data (training dataset) preparation is required!
- Model registration with Oracle Analytics

Qubix

Final Thoughts

- Oracle Analytics can support business users/analysts by generating „good enough“ ML models. For more, in-database OML is required.
- Business users are often prepared to compensate speed of development with accuracy of predictions
- AutoML will bring ML in Oracle Analytics even closer to Business users

- ... but still: data needs proper preparation, collected, and the right question has to be asked

Qubix

Enabling the Predictive Enterprise