Data Analysis Project in BigQuery

Denis Storozhuk, AFPA/Data Science programme, July 2023

Stages of the project

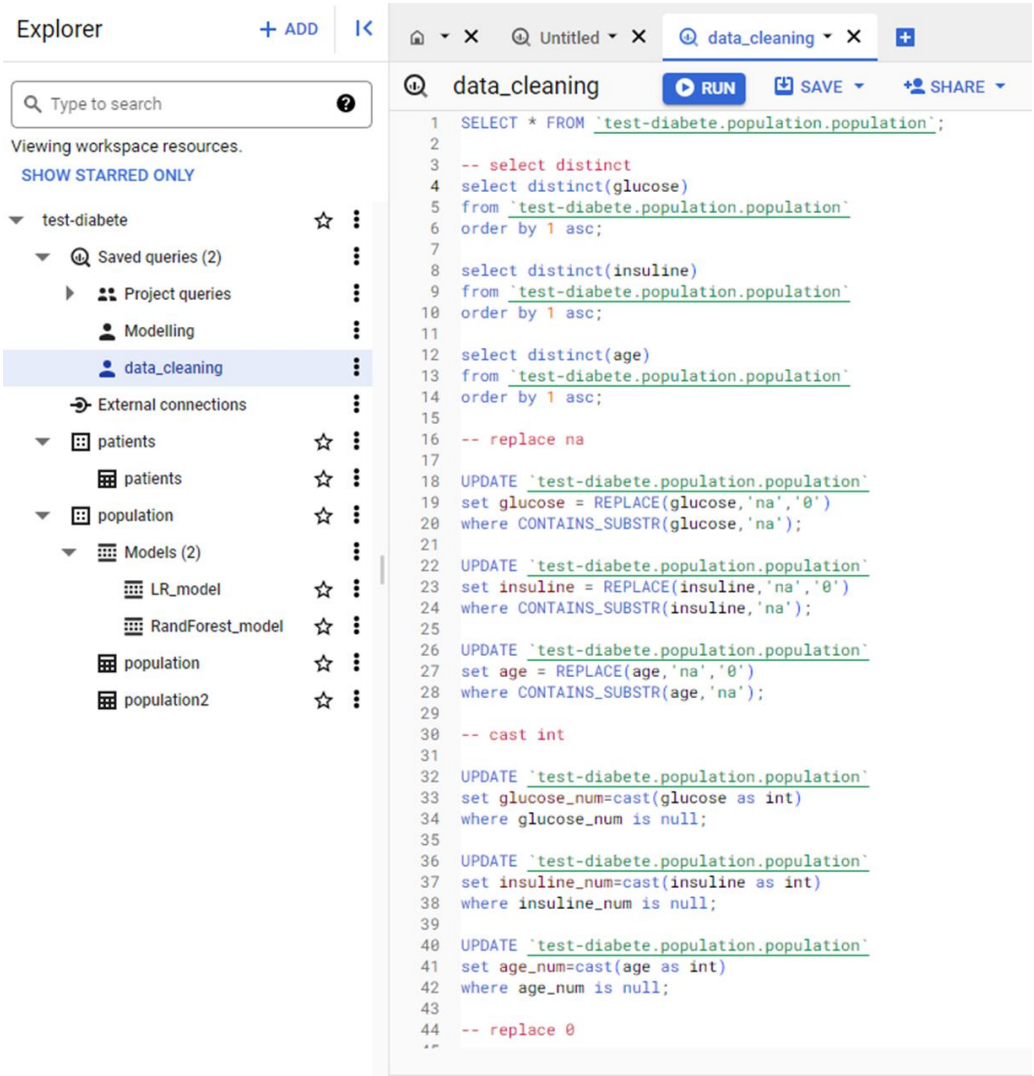
1. Brief / Debriefing
2. Load, Transform, Export data
3. EDA (Exploratory Data Analysis)
4. Modelling
    1. Logistic Regression
    2. Random Forrest Classifier
    3. Deep Neural Networks
    4. Boosted Trees
5. Hyperparameter Tuning
6. Summary / Key Learnings

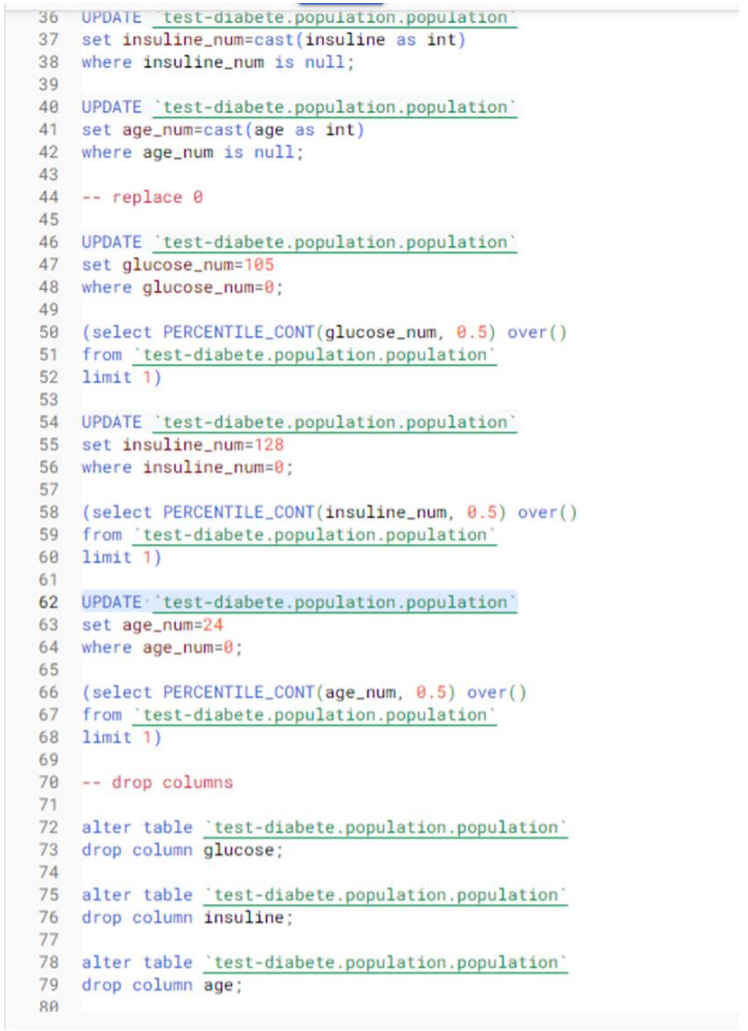
Data Cleaning

Stages of data cleaning

1. Load data
2. Transform data
    1. Replace NANs
    2. Check/change datatypes
    3. Replace Zeros with Medians
    4. Drop non-used columns
3. Ready to export data (e.g. for EDA)

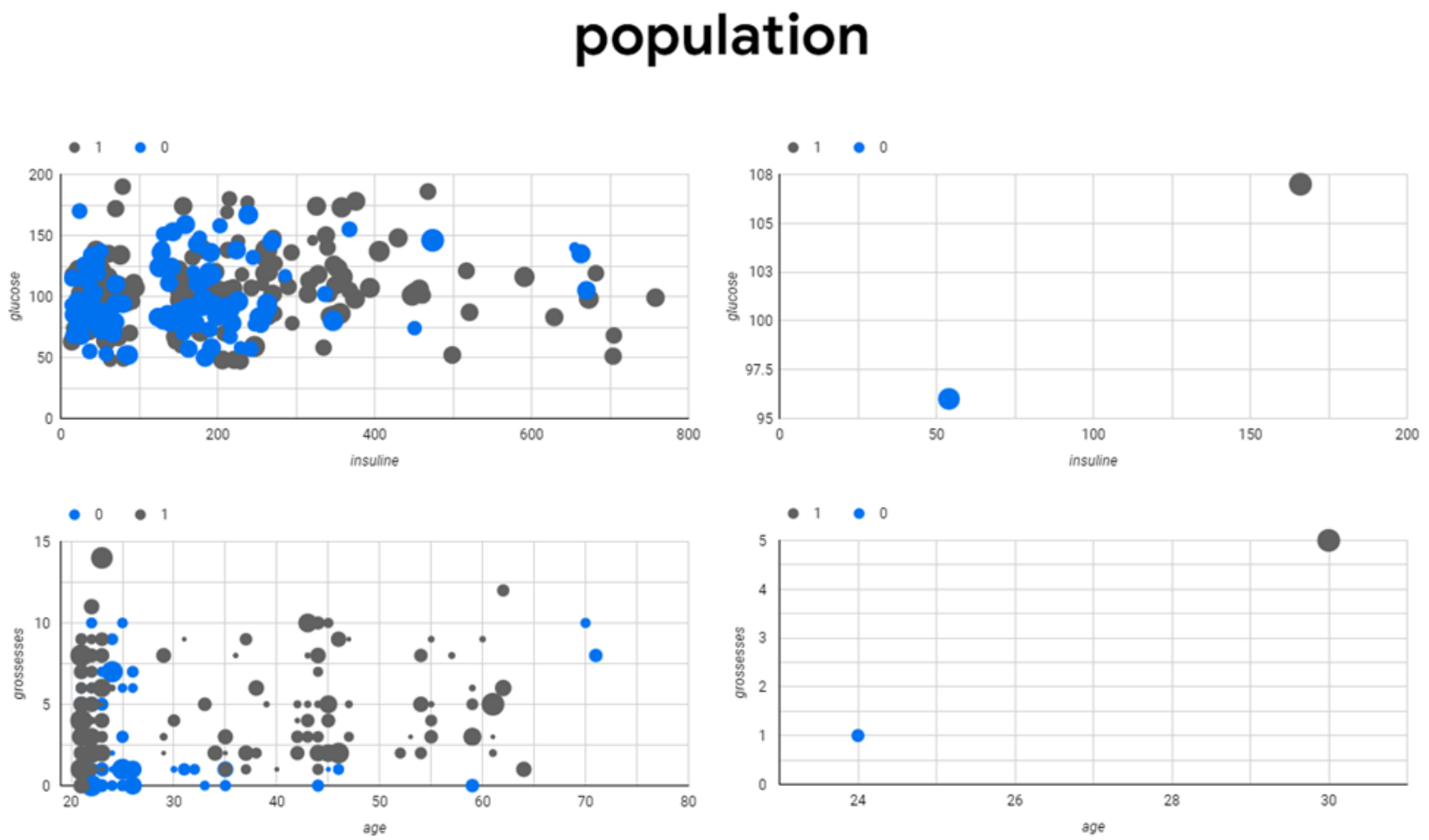
The code in SQL



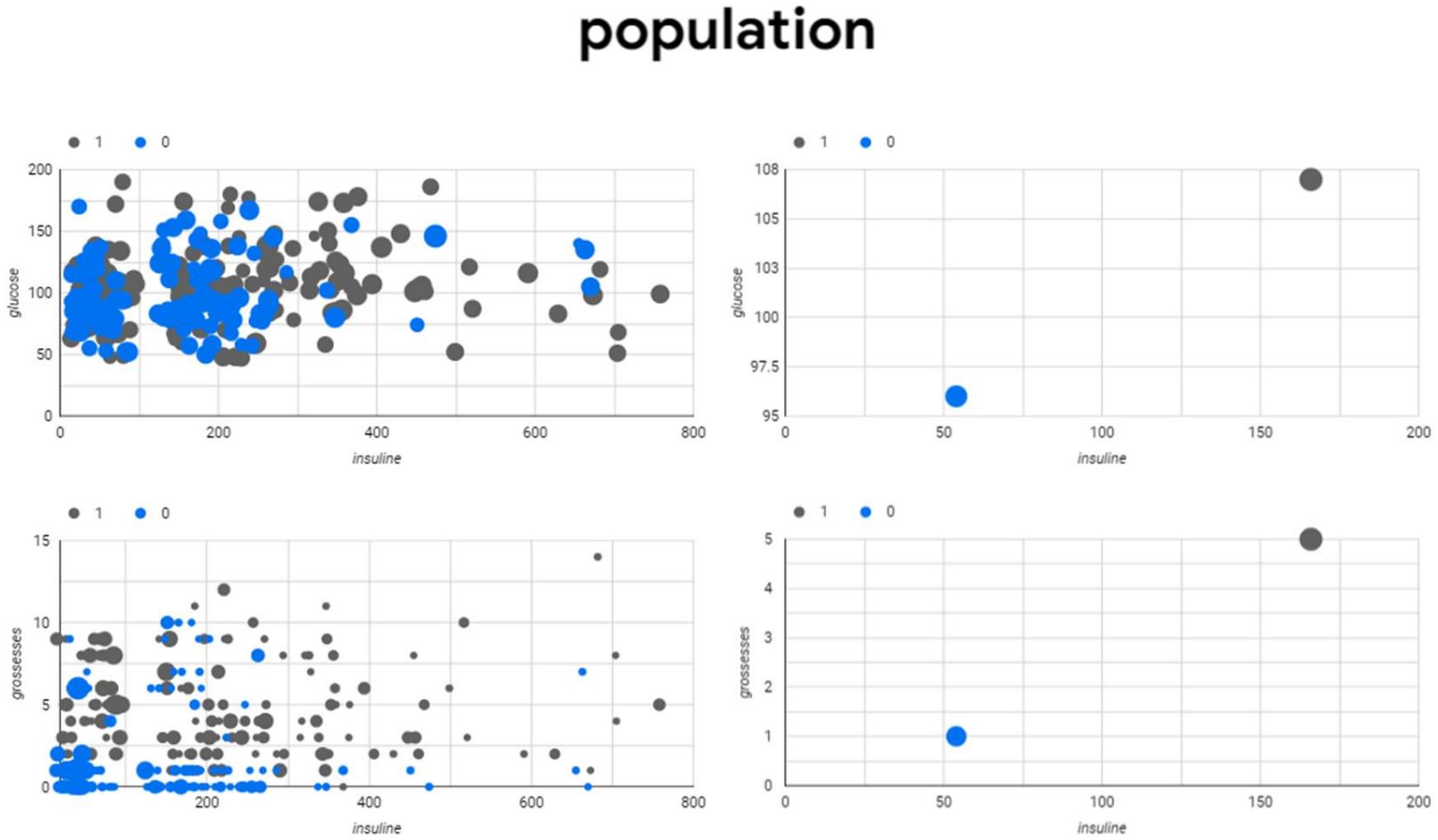


EDA / Exploratory Data Analysis

https://lookerstudio.google.com/s/hutQa9uTOE4



https://lookerstudio.google.com/s/hutQa9uTOE4

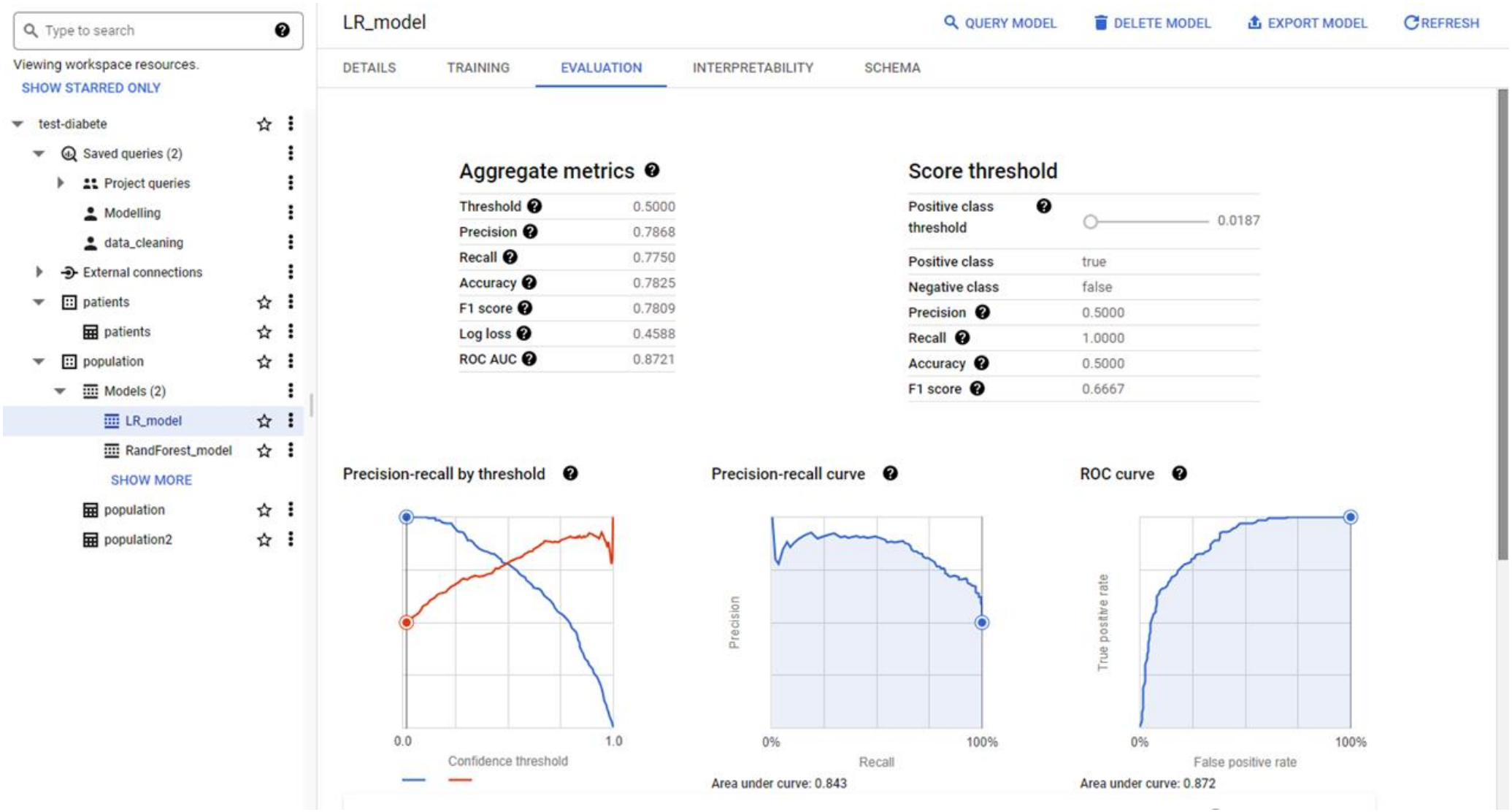


Logistic regression model

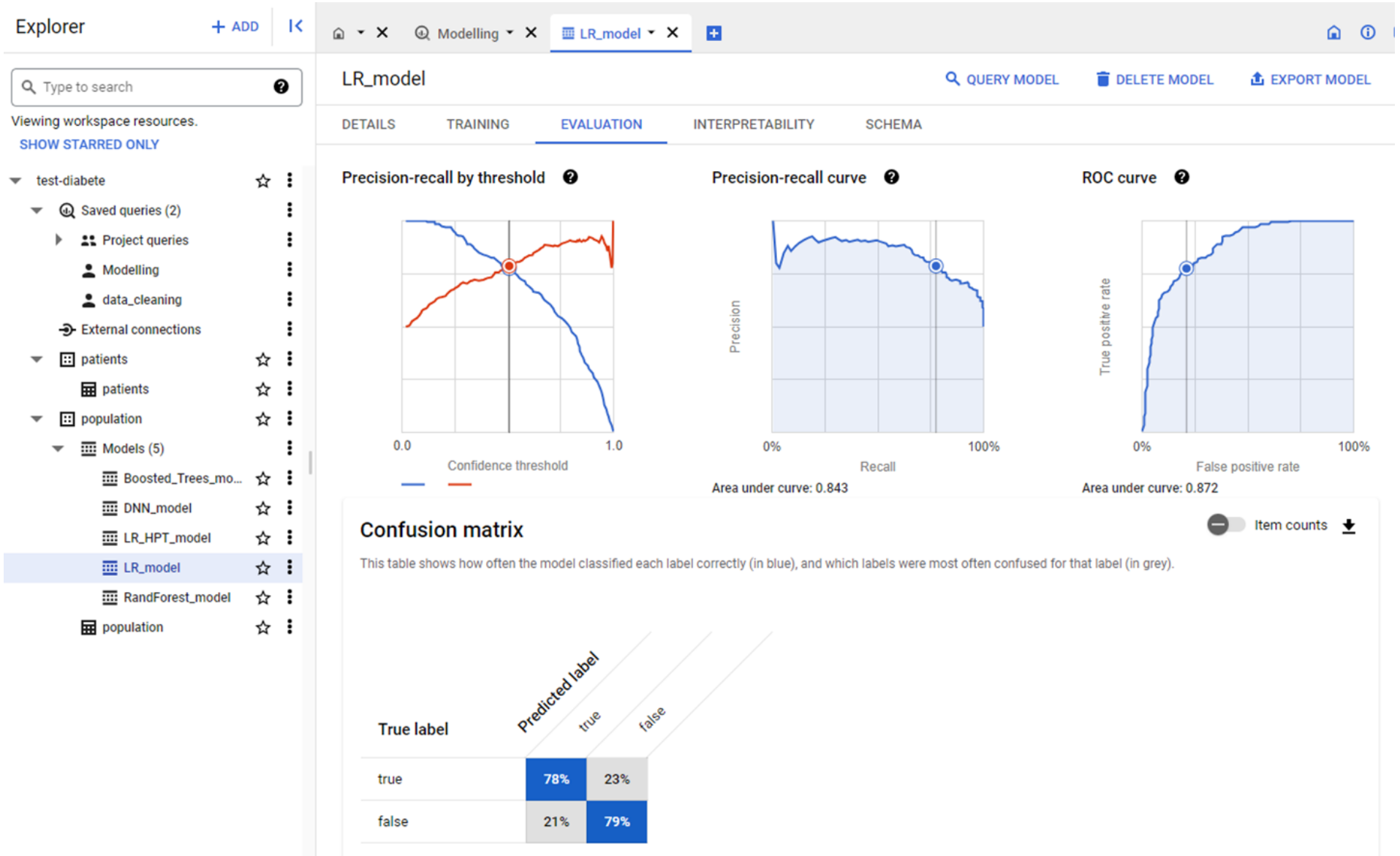
Classification Problem

Supervised Learning

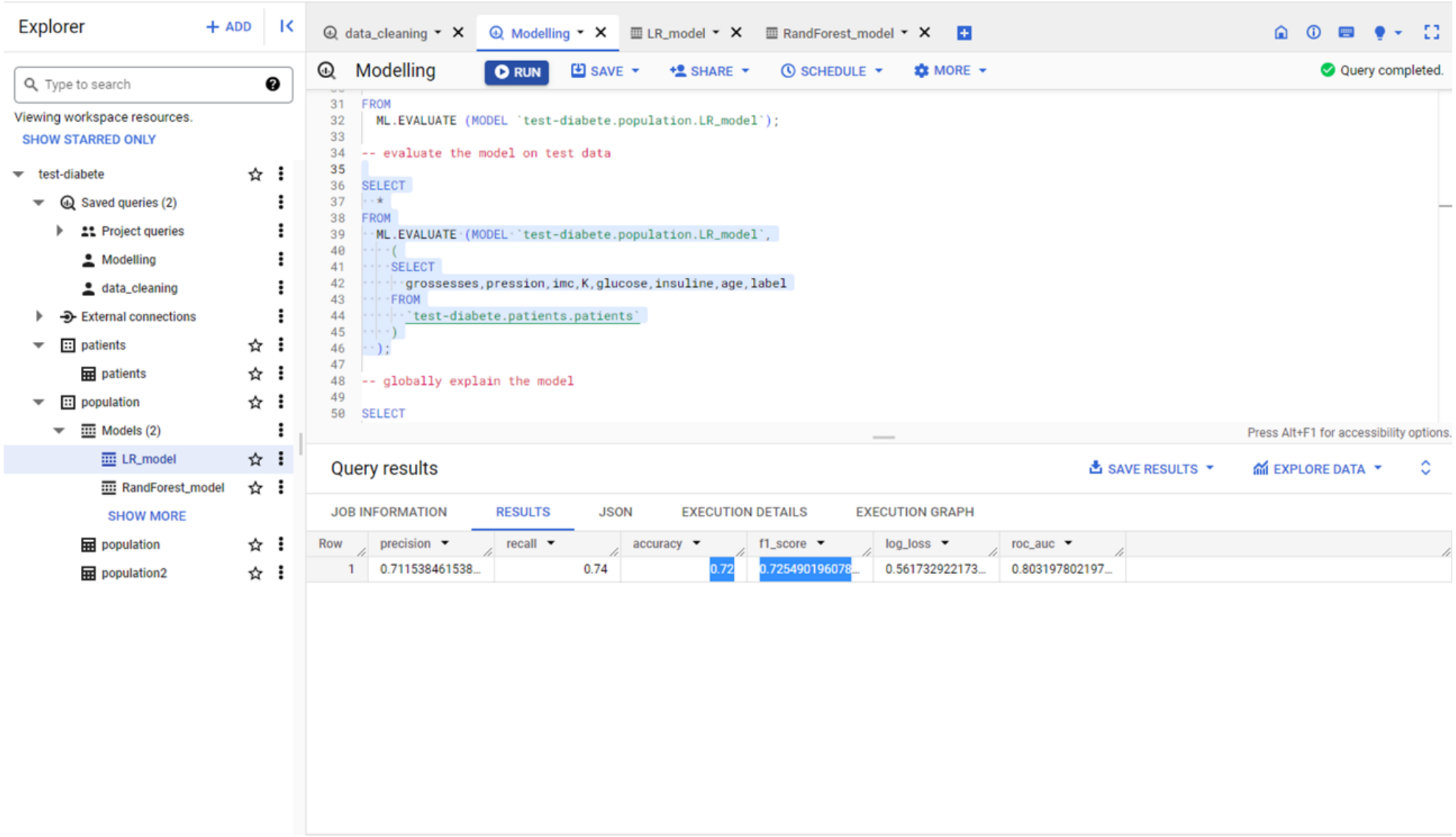
Summary metrics (1)



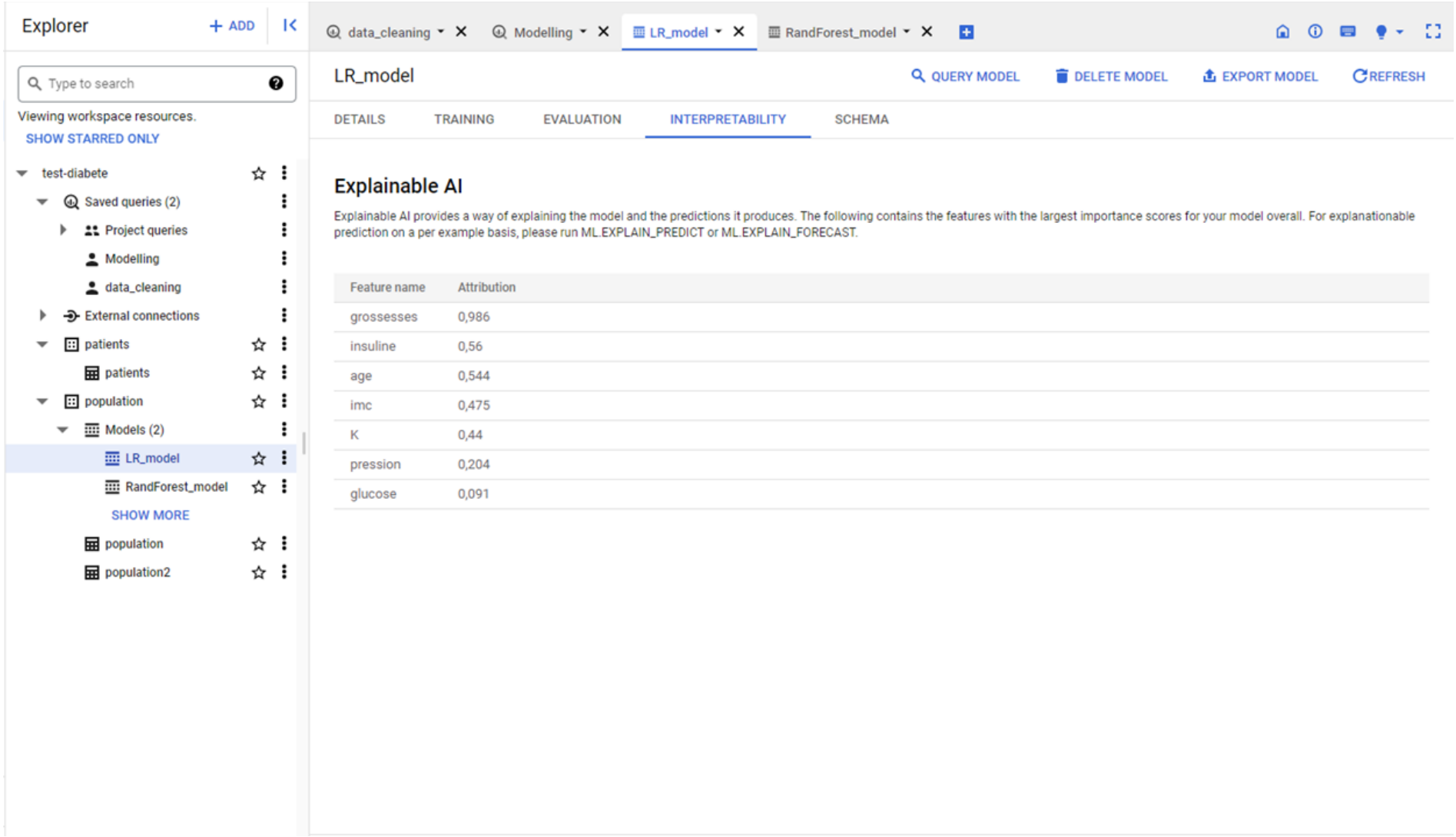
Summary metrics (2)



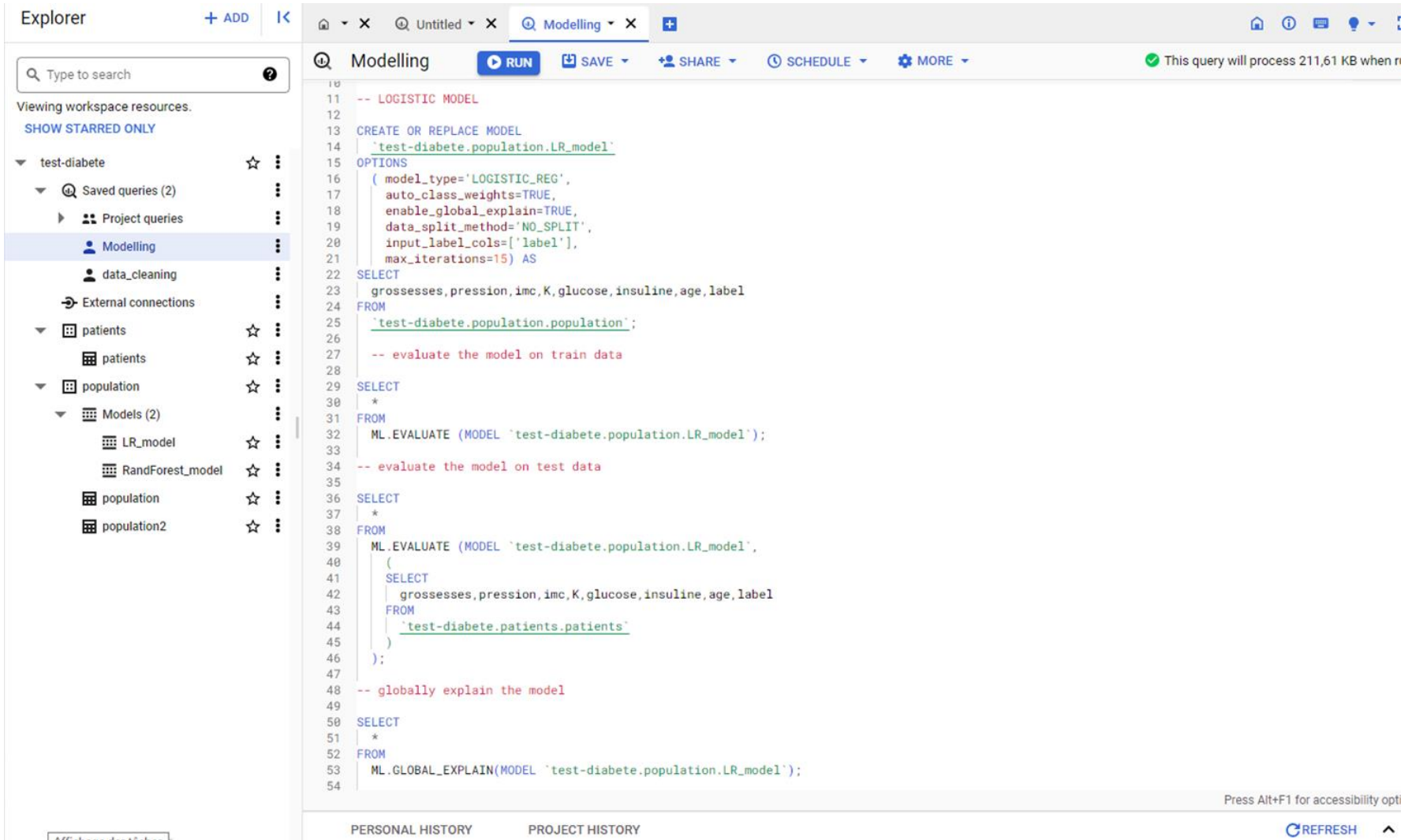
Summary metrics for Test data



Features Importance



SQL code

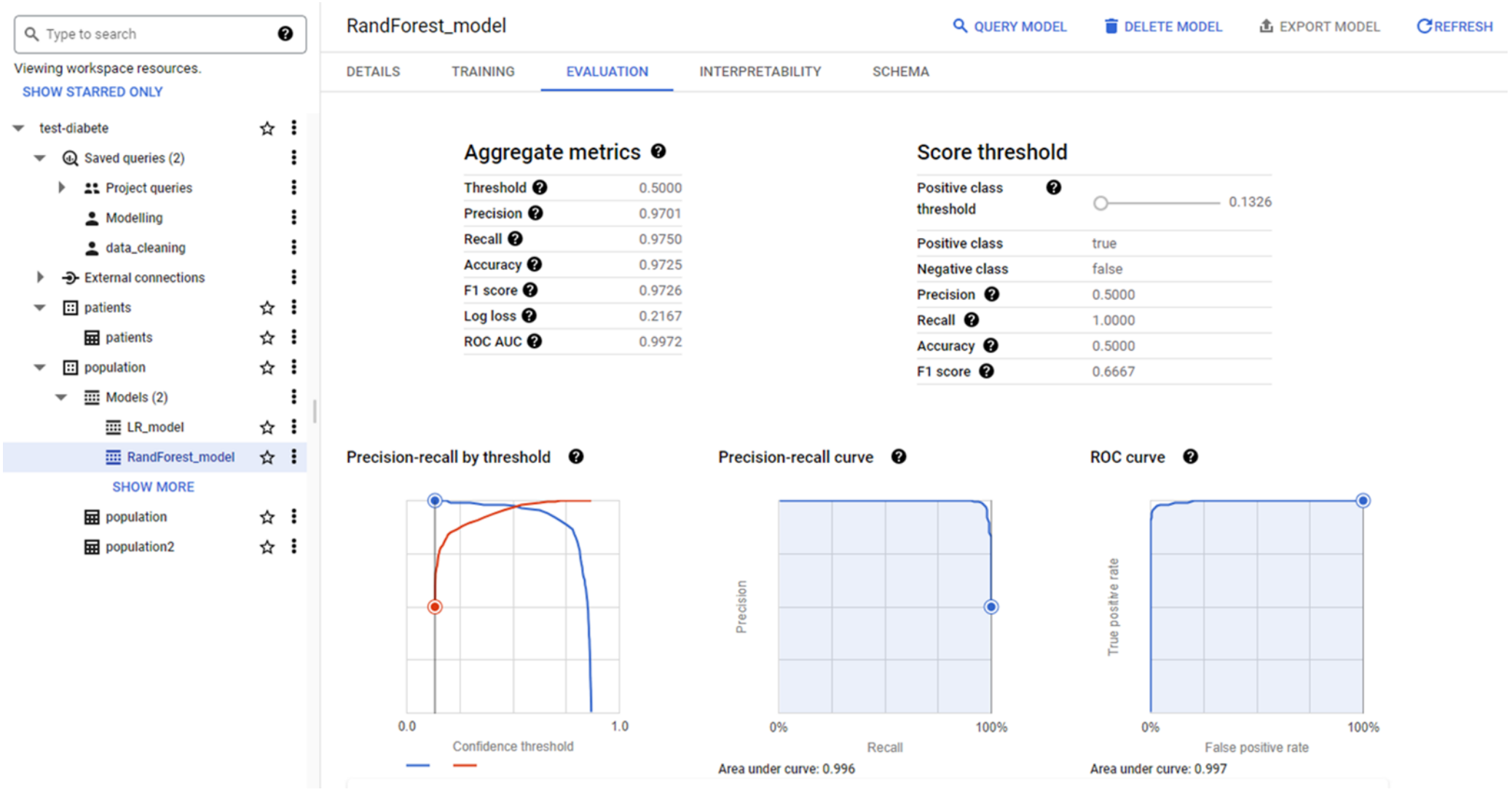

Explorer
+ ADD
Type to search
Viewing workspace resources.
SHOW STARRED ONLY
test-diabete
Saved queries (2)
Project queries
Modelling
data_cleaning
External connections
patients
patients
population
Models (2)
LR_model
RandForest_model
population
population2
Untitled
Modelling
Modelling
RUN
SAVE
SHARE
SCHEDULE
MORE
This query will process 211,61 KB when r
-- LOGISTIC MODEL
CREATE OR REPLACE MODEL
`test-diabete.population.LR_model`
OPTIONS
(model_type='LOGISTIC_REG',
auto_class_weights=TRUE,
enable_global_explain=TRUE,
data_split_method='NO_SPLIT',
input_label_cols=['label'],
max_iterations=15) AS
SELECT
grossesses,pression,imc,K,glucose,insuline,age,label
FROM
`test-diabete.population.population`;
-- evaluate the model on train data
SELECT
*
FROM
ML.EVALUATE (MODEL `test-diabete.population.LR_model`);
-- evaluate the model on test data
SELECT
*
FROM
ML.EVALUATE (MODEL `test-diabete.population.LR_model`,
(
SELECT
grossesses,pression,imc,K,glucose,insuline,age,label
FROM
`test-diabete.patients.patients`
)
);
-- globally explain the model
SELECT
*
FROM
ML.GLOBAL_EXPLAIN(MODEL `test-diabete.population.LR_model`);
Press Alt+F1 for accessibility opti
PERSONAL HISTORY
PROJECT HISTORY
REFRESH

Random Forest Classifier

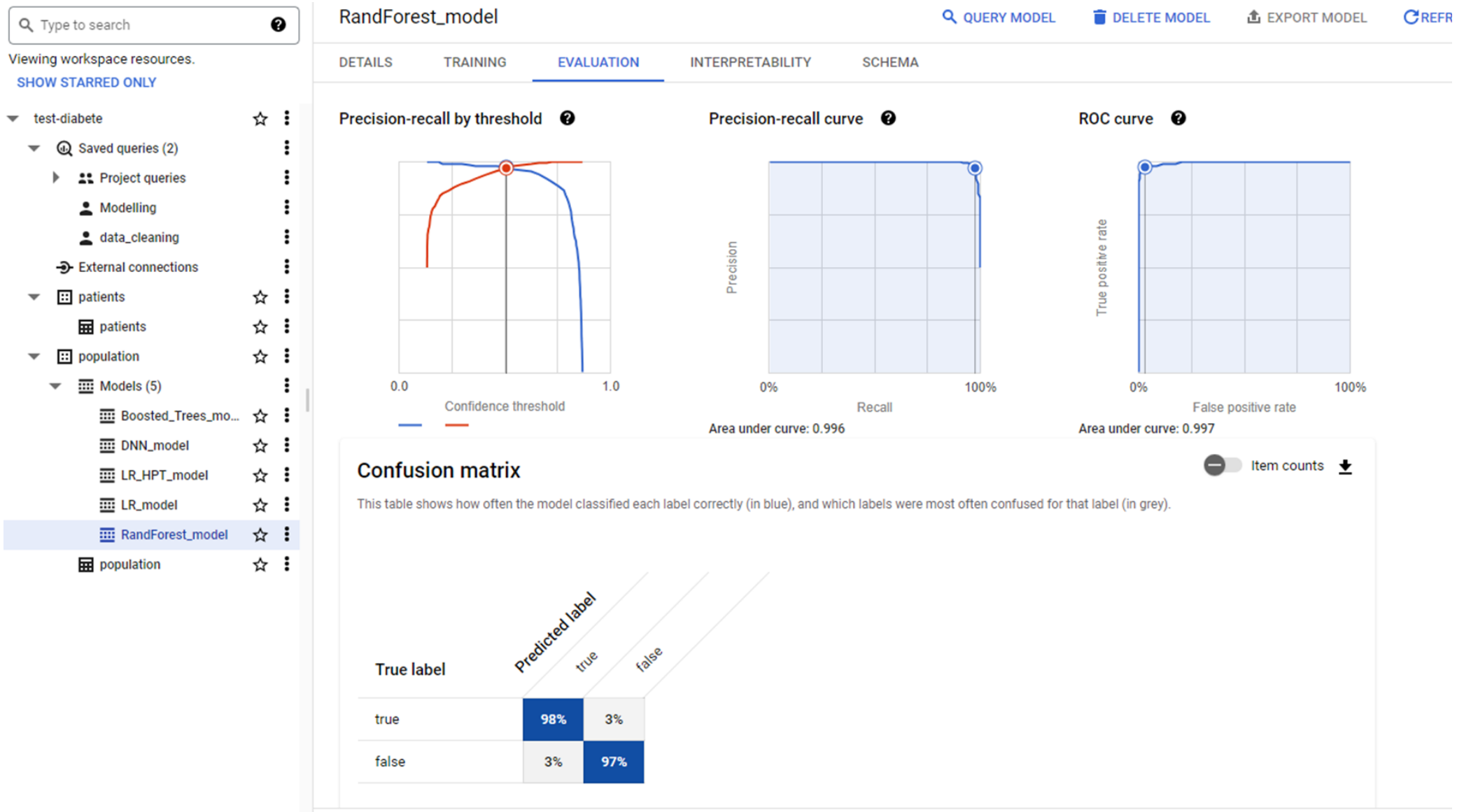
Classification Problem

Supervised Learning

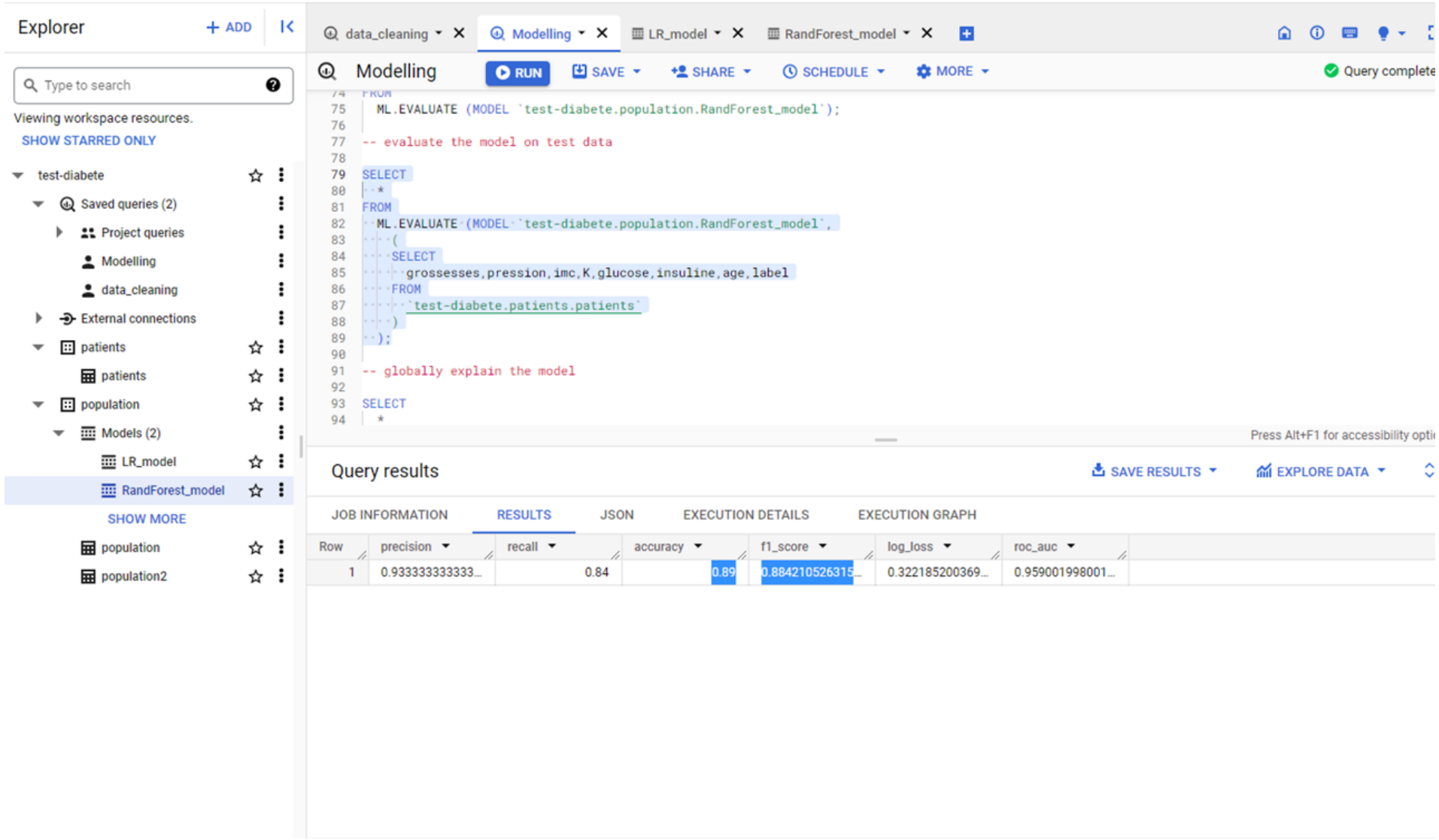
Summary metrics (1)



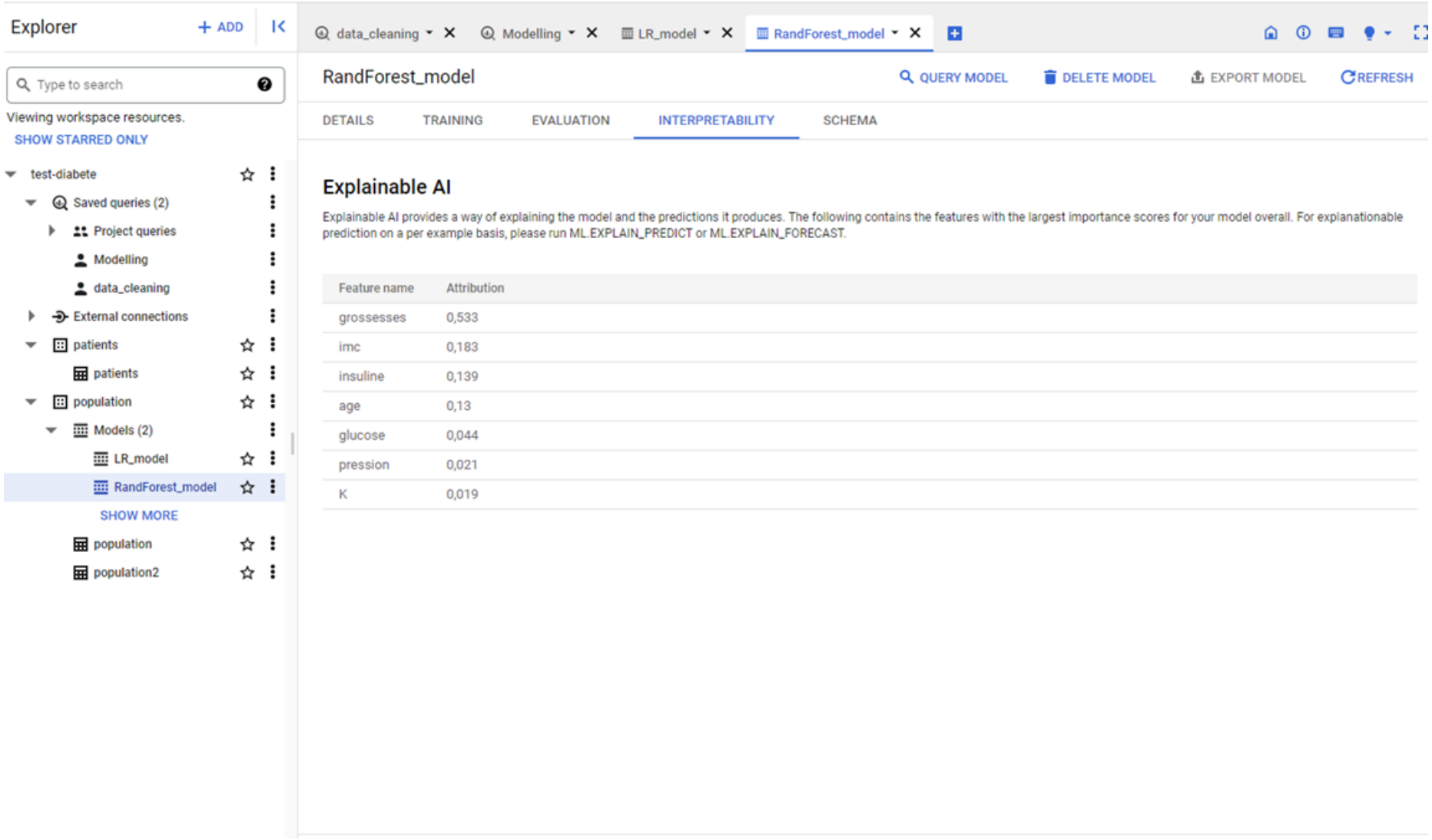
Summary metrics (2)



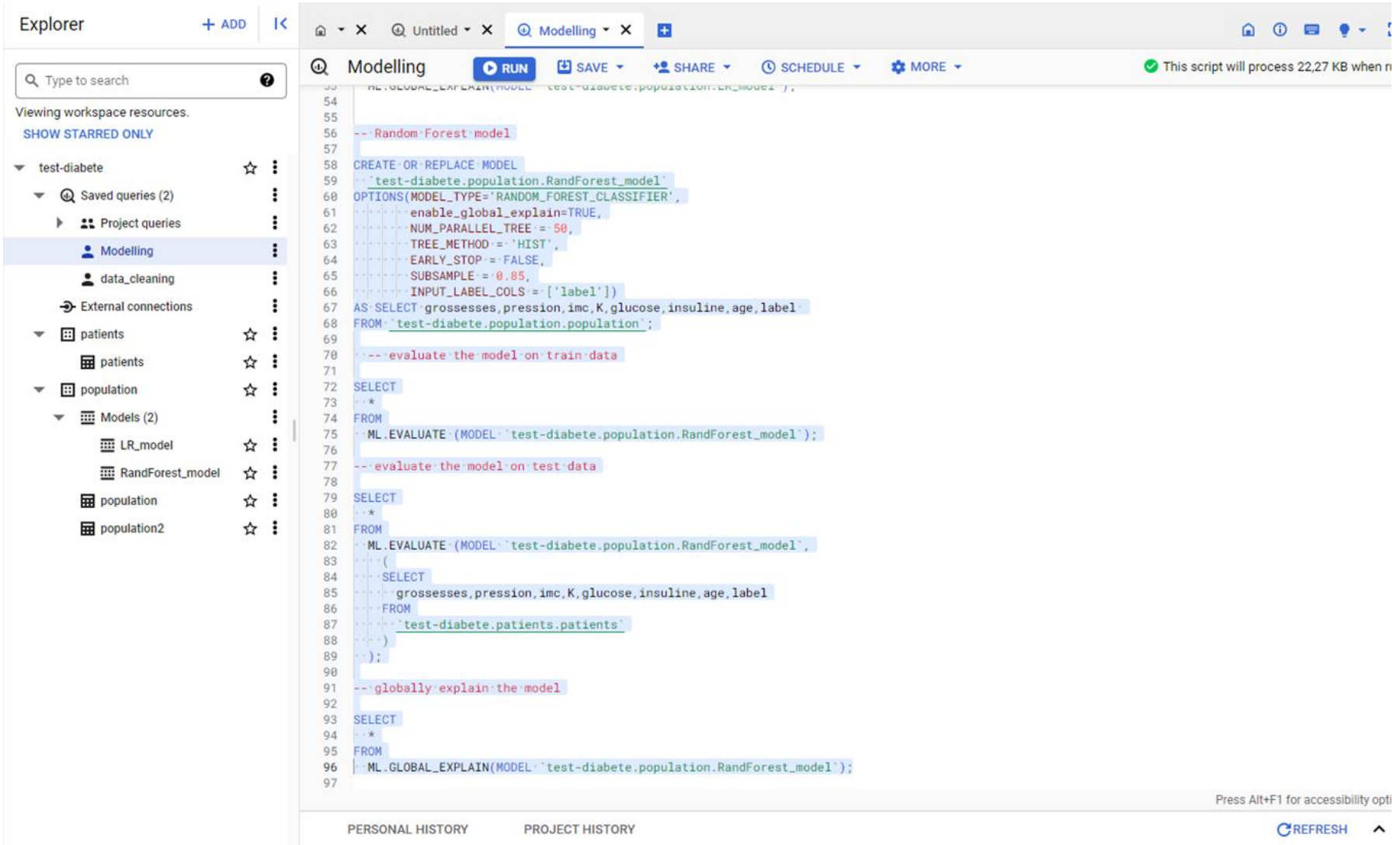
Summary metrics for Test data



Features Importance



SQL code


Explorer
+ ADD
Type to search
Viewing workspace resources.
SHOW STARRED ONLY
test-diabete
Saved queries (2)
Project queries
Modelling
data_cleaning
External connections
patients
patients
population
Models (2)
LR_model
RandForest_model
population
population2
Untitled
Modelling
Modelling
RUN
SAVE
SHARE
SCHEDULE
MORE
This script will process 22,27 KB when r
54
55
56 -- Random Forest model
57
58 CREATE OR REPLACE MODEL
59 `test-diabete.population.RandForest_model`
60 OPTIONS(MODEL_TYPE='RANDOM_FOREST_CLASSIFIER',
61 enable_global_explain=TRUE,
62 NUM_PARALLEL_TREE = 50,
63 TREE_METHOD = 'HIST',
64 EARLY_STOP = FALSE,
65 SUBSAMPLE = 0.85,
66 INPUT_LABEL_COLS = ['label'])
67 AS SELECT grossesses,pression,imc,K,glucose,insuline,age,label
68 FROM `test-diabete.population.population`;
69
70 -- evaluate the model on train data
71
72 SELECT
73 *
74 FROM
75 ML.EVALUATE (MODEL `test-diabete.population.RandForest_model`);
76
77 -- evaluate the model on test data
78
79 SELECT
80 *
81 FROM
82 ML.EVALUATE (MODEL `test-diabete.population.RandForest_model`,
83 (
84 SELECT
85 grossesses,pression,imc,K,glucose,insuline,age,label
86 FROM
87 `test-diabete.patients.patients`
88)
89);
90
91 -- globally explain the model
92
93 SELECT
94 *
95 FROM
96 ML.GLOBAL_EXPLAIN(MODEL `test-diabete.population.RandForest_model`);
97
Press Alt+F1 for accessibility opti
PERSONAL HISTORY
PROJECT HISTORY
REFRESH

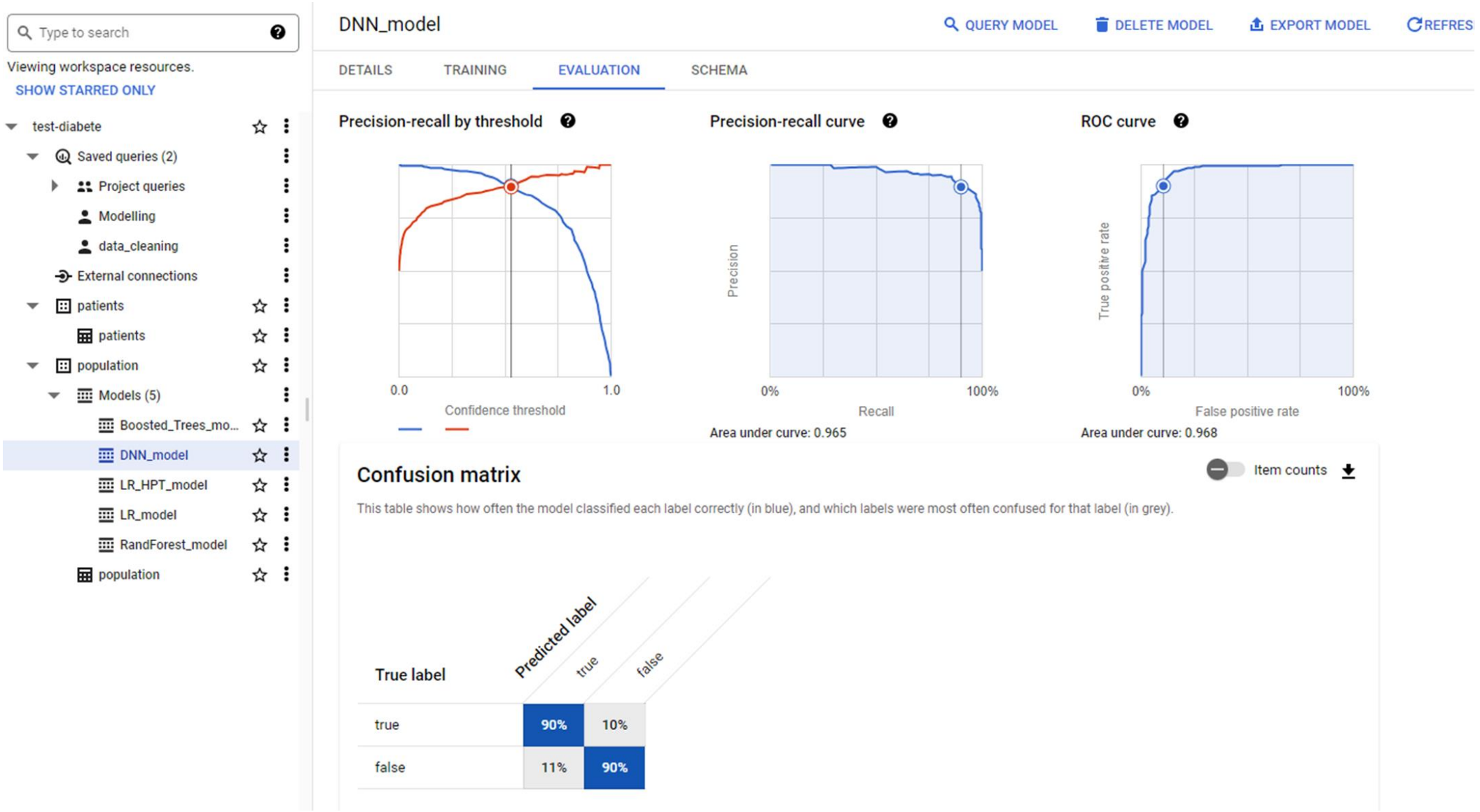
Deep Neural Networks

Classification Problem

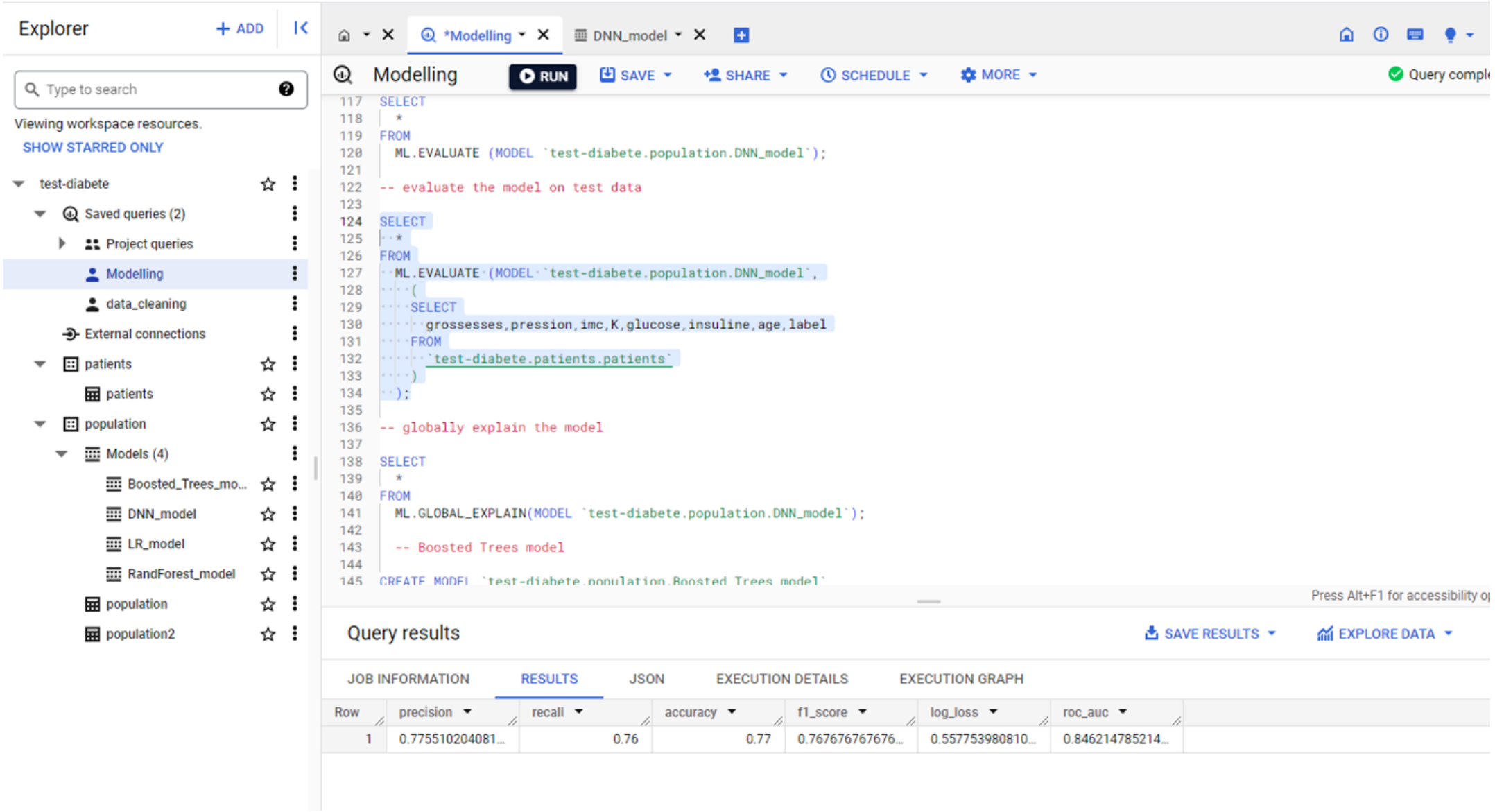
Supervised Learning

Summary metrics (1)

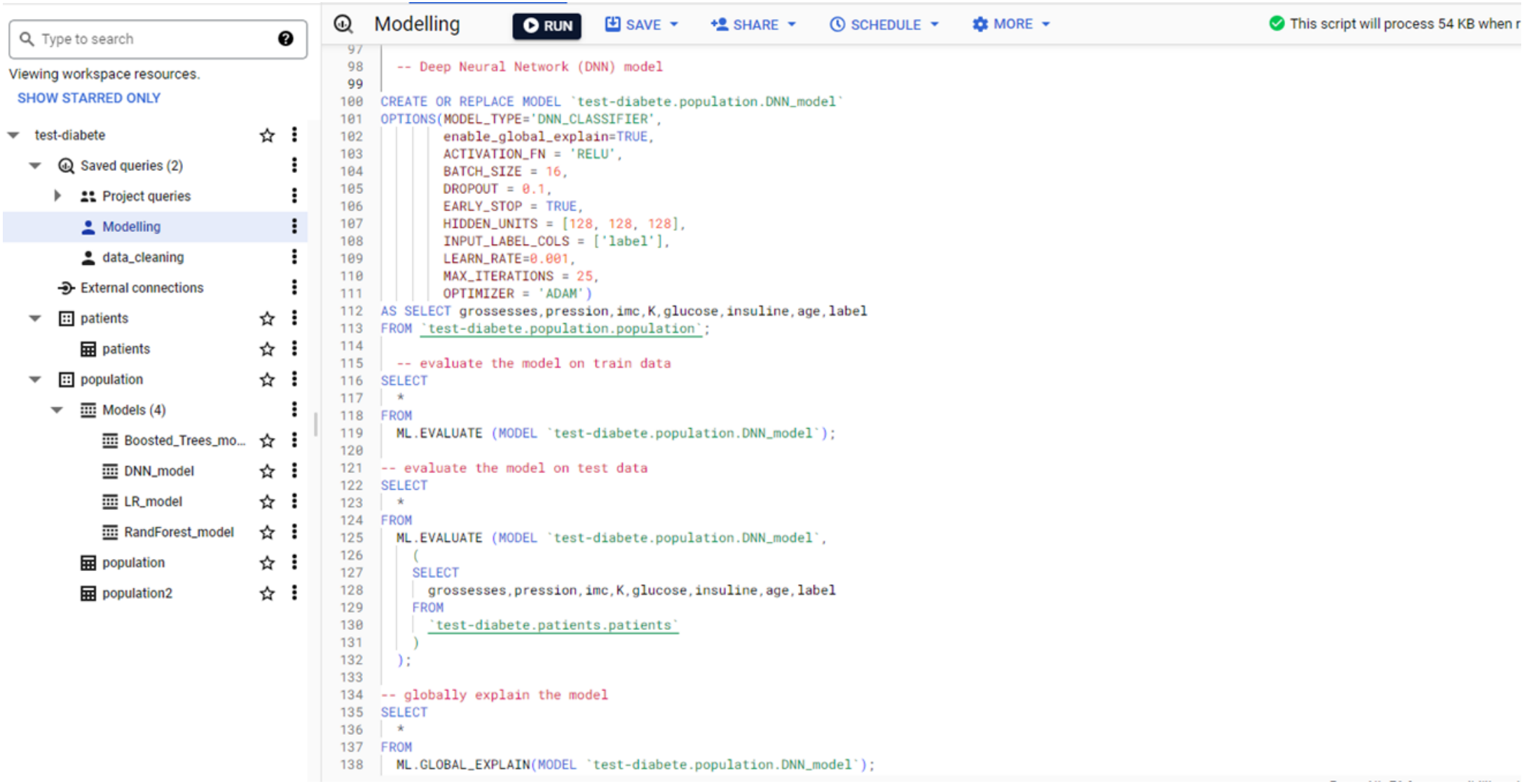
Summary metrics (2)



Summary metrics for Test data



SQL code

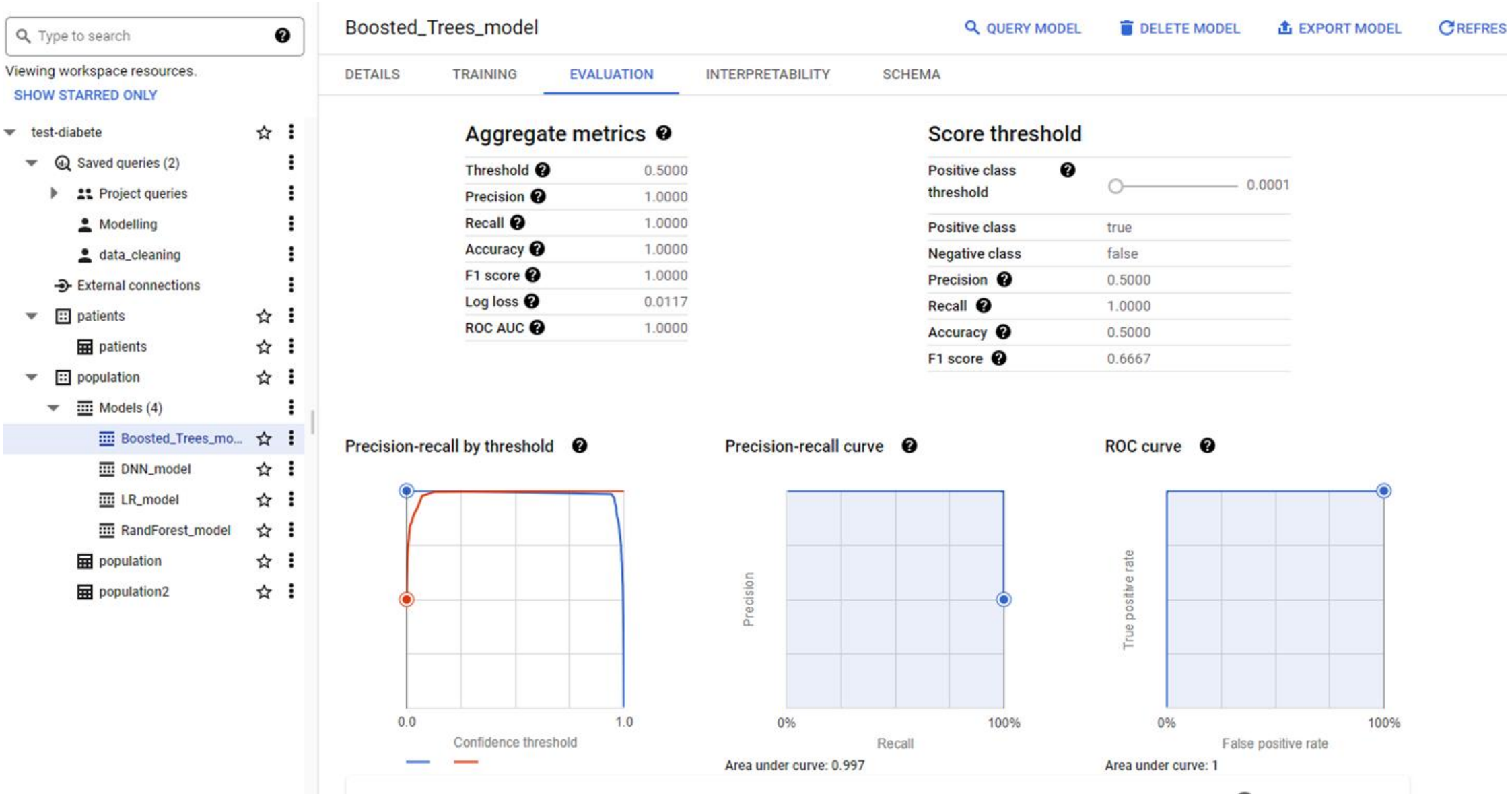

Type to search
Viewing workspace resources.
SHOW STARRED ONLY
test-diabete
Saved queries (2)
Project queries
Modelling
data_cleaning
External connections
patients
patients
population
Models (4)
Boosted_Trees_mo...
DNN_model
LR_model
RandForest_model
population
population2
Modelling
RUN
SAVE
SHARE
SCHEDULE
MORE
This script will process 54 KB when r
-- Deep Neural Network (DNN) model
CREATE OR REPLACE MODEL `test-diabete.population.DNN_model`
OPTIONS(MODEL_TYPE='DNN_CLASSIFIER',
enable_global_explain=TRUE,
ACTIVATION_FN = 'RELU',
BATCH_SIZE = 16,
DROPOUT = 0.1,
EARLY_STOP = TRUE,
HIDDEN_UNITS = [128, 128, 128],
INPUT_LABEL_COLS = ['label'],
LEARN_RATE=0.001,
MAX_ITERATIONS = 25,
OPTIMIZER = 'ADAM')
AS SELECT grossesses,pression,imc,K,glucose,insuline,age,label
FROM `test-diabete.population.population`;
-- evaluate the model on train data
SELECT
*
FROM
ML.EVALUATE (MODEL `test-diabete.population.DNN_model`);
-- evaluate the model on test data
SELECT
*
FROM
ML.EVALUATE (MODEL `test-diabete.population.DNN_model`,
(
SELECT
grossesses,pression,imc,K,glucose,insuline,age,label
FROM
`test-diabete.patients.patients`
)
);
-- globally explain the model
SELECT
*
FROM
ML.GLOBAL_EXPLAIN(MODEL `test-diabete.population.DNN_model`);

Boosted Trees Classifier

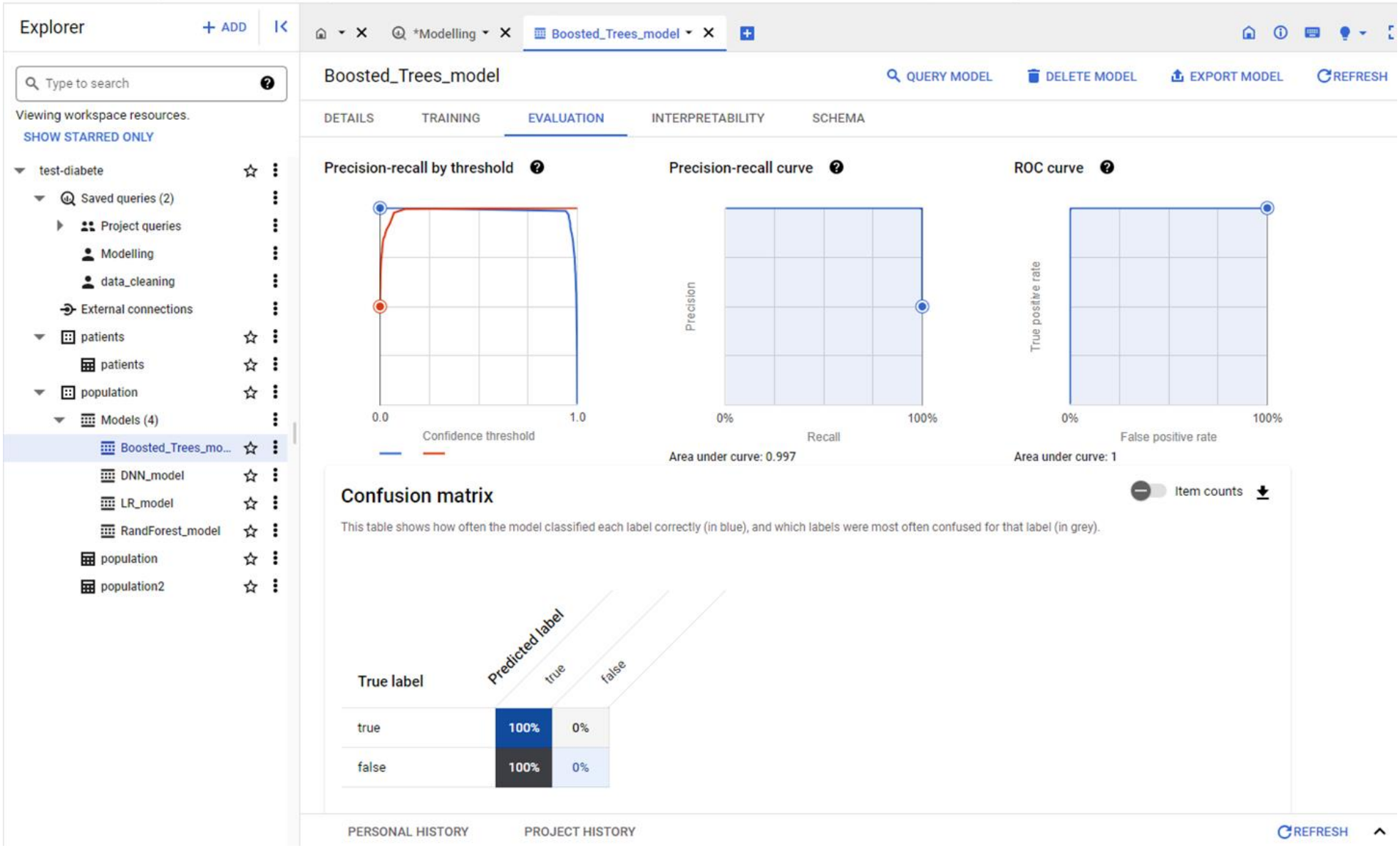
Classification Problem

Supervised Learning

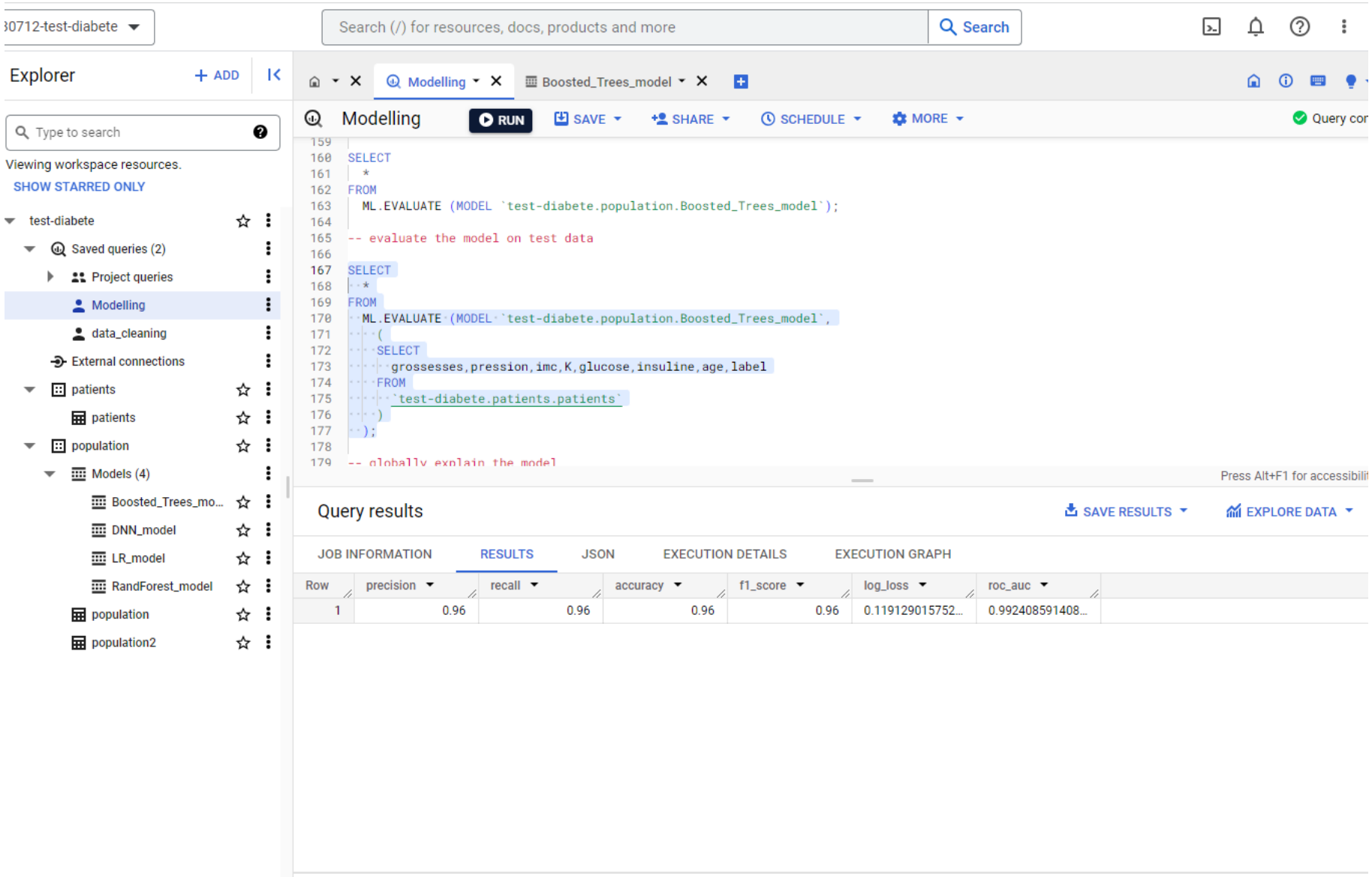
Summary metrics (1)



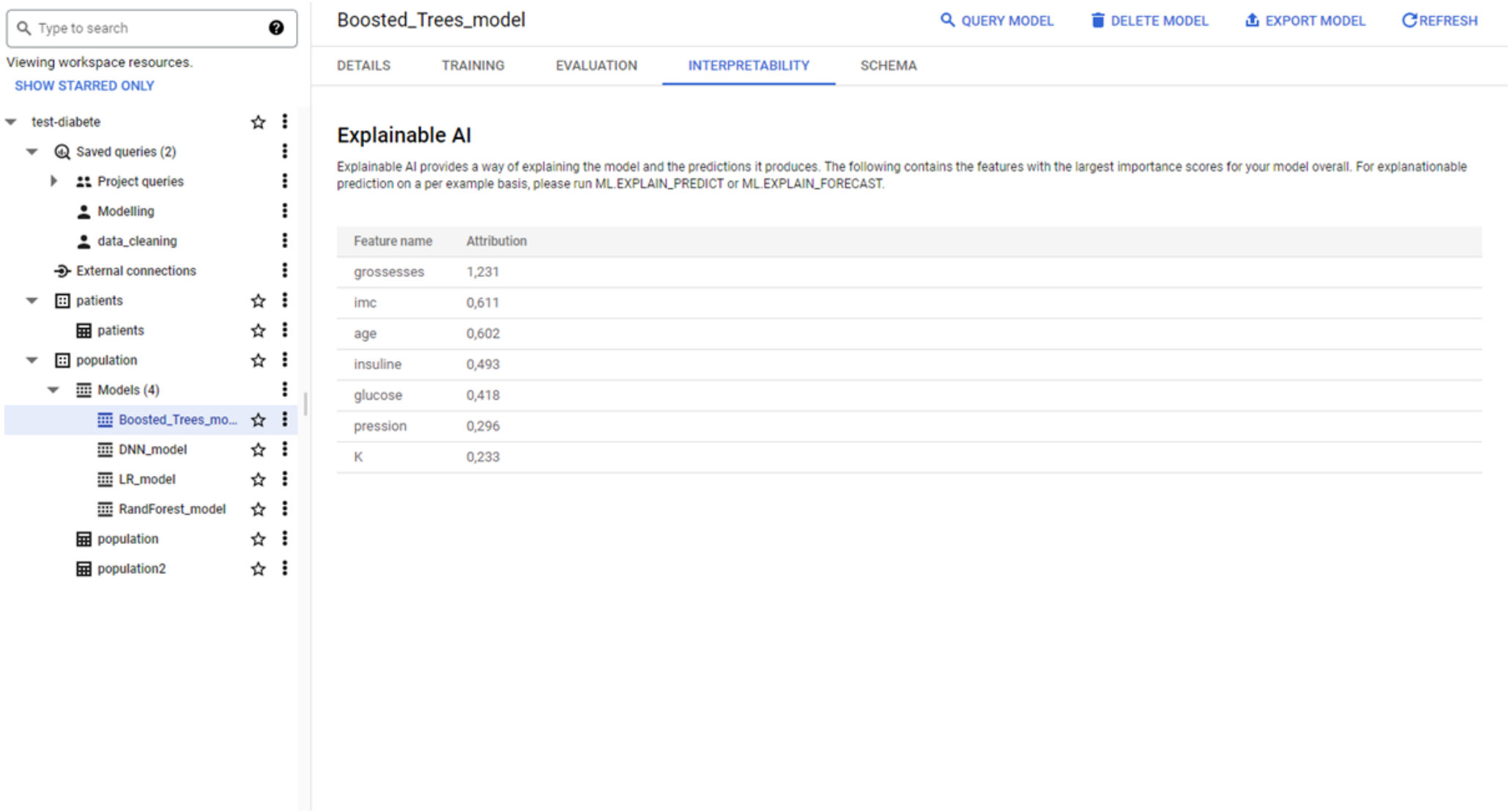
Summary metrics (2)



Summary metrics for Test data



Features Importance



SQL code

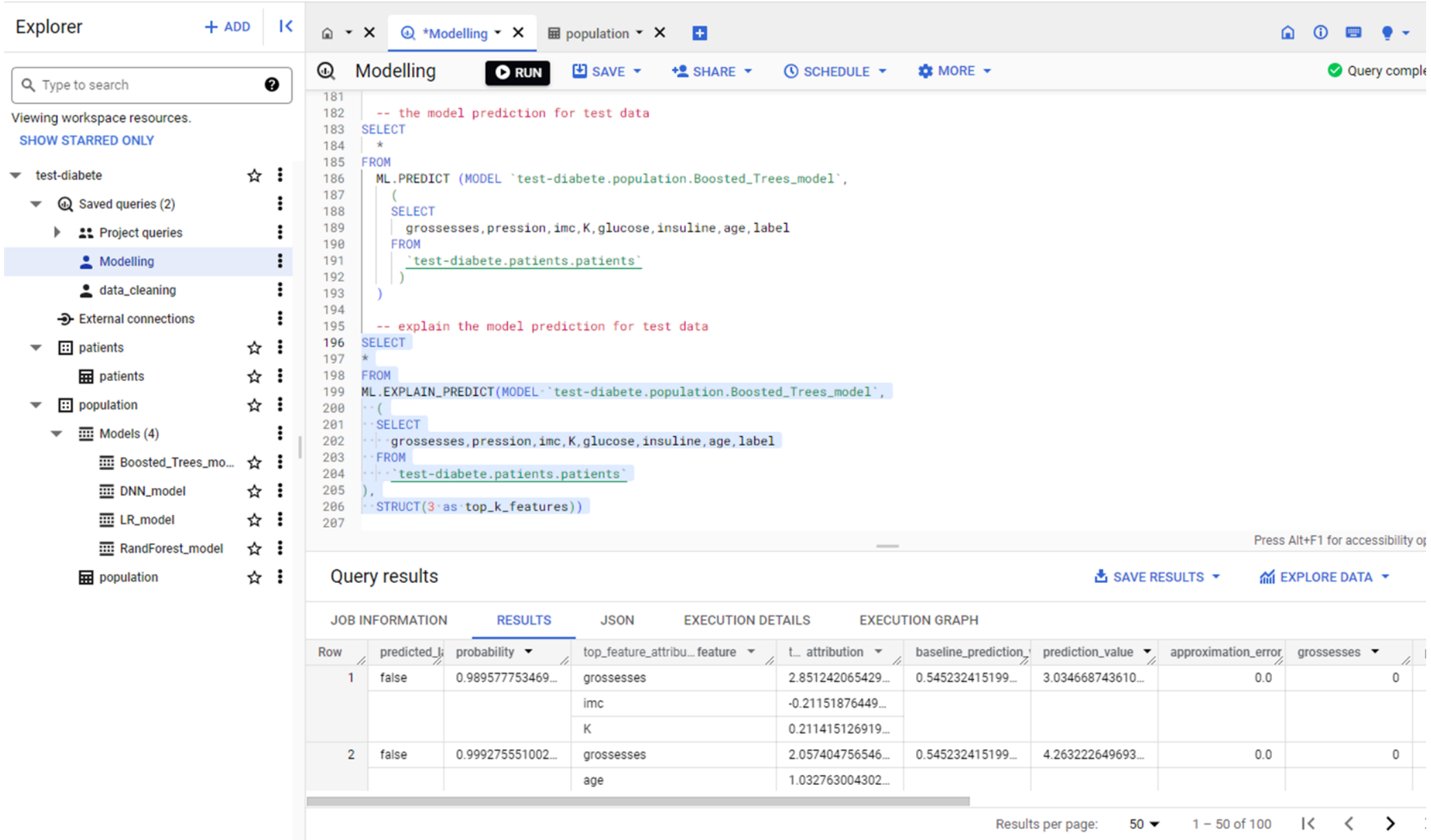
Modelling RUN SAVE SHARE SCHEDULE MORE This script will process 54 KB when run.

```
142
143   -- Boosted Trees model
144
145 CREATE MODEL `test-diabete.population.Boosted_Trees_model`
146 OPTIONS(MODEL_TYPE='BOOSTED_TREE_CLASSIFIER',
147           enable_global_explain=TRUE,
148           BOOSTER_TYPE = 'GBTREE',
149           NUM_PARALLEL_TREE = 1,
150           MAX_ITERATIONS = 50,
151           TREE_METHOD = 'HIST',
152           EARLY_STOP = FALSE,
153           SUBSAMPLE = 0.85,
154           INPUT_LABEL_COLS = ['label'])
155 AS SELECT grossesses,pression,imc,K,glucose,insuline,age,label
156 FROM `test-diabete.population.population`;
157
158   -- evaluate the model on train data
159
160 SELECT
161   *
162 FROM
163   ML.EVALUATE (MODEL `test-diabete.population.Boosted_Trees_model`);
164
165 -- evaluate the model on test data
166
167 SELECT
168   *
169 FROM
170   ML.EVALUATE (MODEL `test-diabete.population.Boosted_Trees_model`,
171     (
172     SELECT
173       grossesses,pression,imc,K,glucose,insuline,age,label
174     FROM
175       `test-diabete.patients.patients`
176     )
177   );
178
179 -- globally explain the model
180 SELECT
181   *
182 FROM
183   ML.GLOBAL_EXPLAIN(MODEL `test-diabete.population.Boosted_Trees_model`);
```

Press Alt+F1 for accessibility options.

Query results SAVE RESULTS EXPLORE DATA

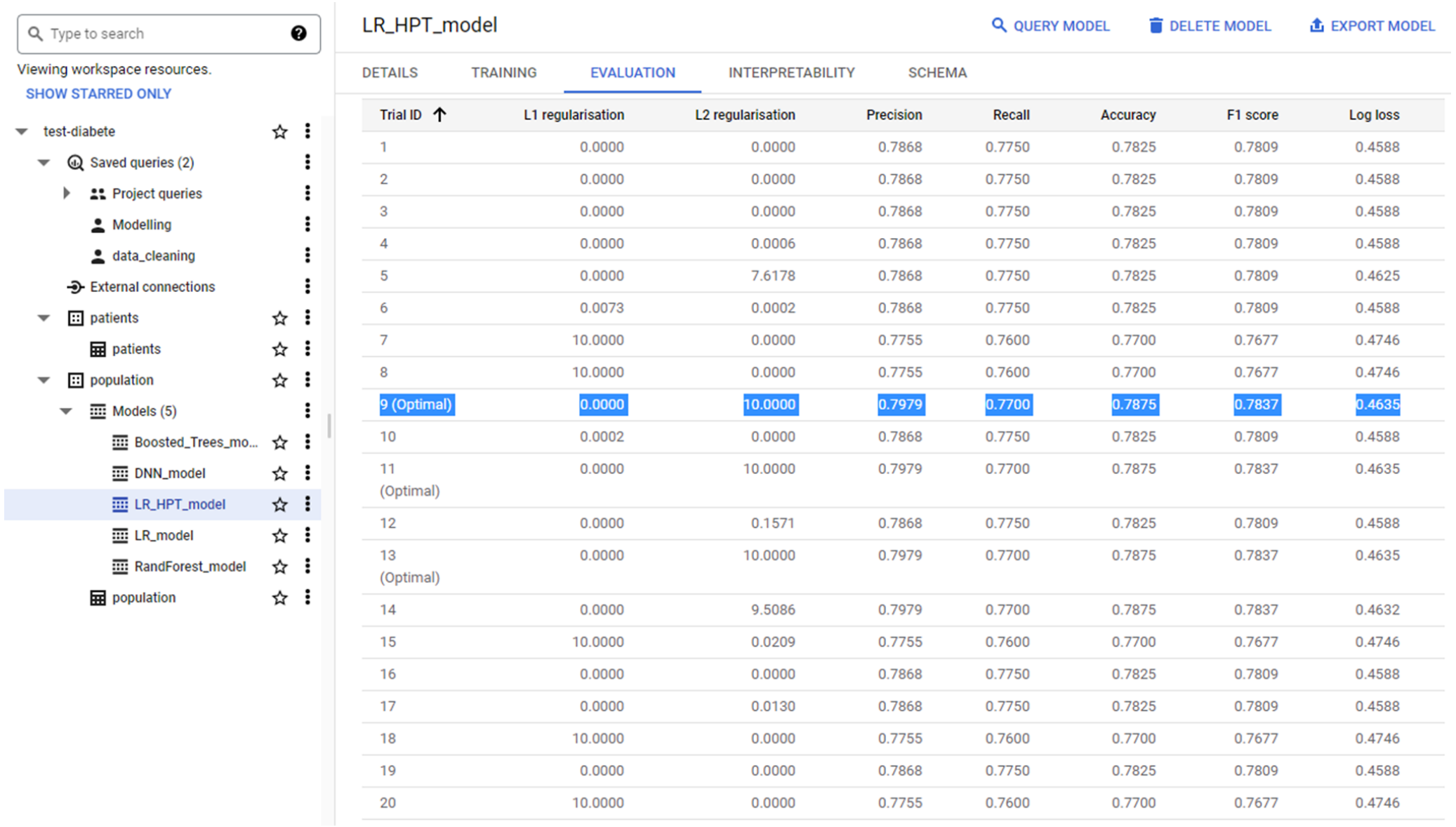
Prediction for test / SQL code



Hyperparameter Tuning

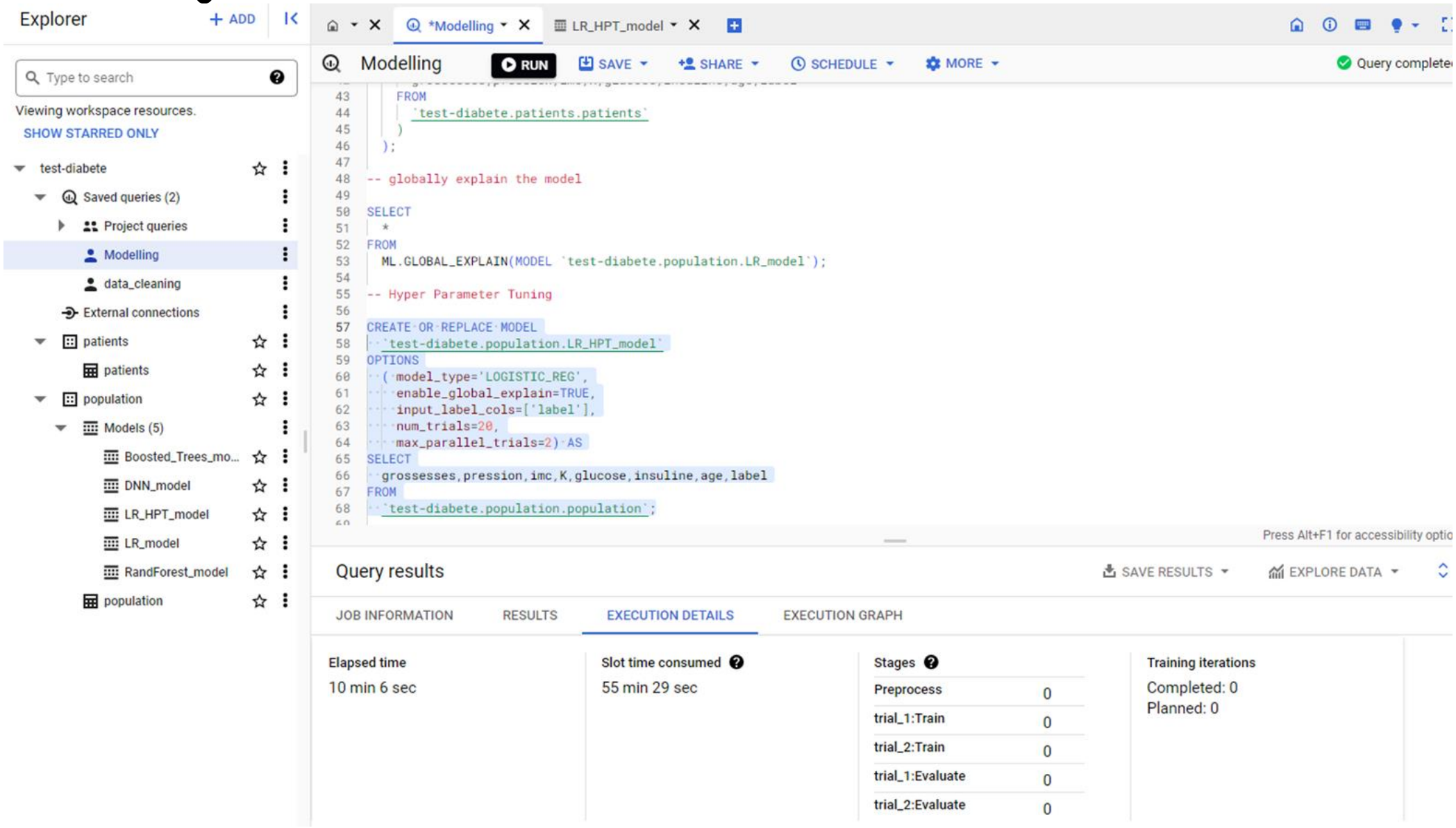
Example for a Logistic Regression model

Evaluation Summary



Trial ID	L1 regularisation	L2 regularisation	Precision	Recall	Accuracy	F1 score	Log loss
1	0.0000	0.0000	0.7868	0.7750	0.7825	0.7809	0.4588
2	0.0000	0.0000	0.7868	0.7750	0.7825	0.7809	0.4588
3	0.0000	0.0000	0.7868	0.7750	0.7825	0.7809	0.4588
4	0.0000	0.0006	0.7868	0.7750	0.7825	0.7809	0.4588
5	0.0000	7.6178	0.7868	0.7750	0.7825	0.7809	0.4625
6	0.0073	0.0002	0.7868	0.7750	0.7825	0.7809	0.4588
7	10.0000	0.0000	0.7755	0.7600	0.7700	0.7677	0.4746
8	10.0000	0.0000	0.7755	0.7600	0.7700	0.7677	0.4746
9 (Optimal)	0.0000	10.0000	0.7979	0.7700	0.7875	0.7837	0.4635
10	0.0002	0.0000	0.7868	0.7750	0.7825	0.7809	0.4588
11 (Optimal)	0.0000	10.0000	0.7979	0.7700	0.7875	0.7837	0.4635
12	0.0000	0.1571	0.7868	0.7750	0.7825	0.7809	0.4588
13 (Optimal)	0.0000	10.0000	0.7979	0.7700	0.7875	0.7837	0.4635
14	0.0000	9.5086	0.7979	0.7700	0.7875	0.7837	0.4632
15	10.0000	0.0209	0.7755	0.7600	0.7700	0.7677	0.4746
16	0.0000	0.0000	0.7868	0.7750	0.7825	0.7809	0.4588
17	0.0000	0.0130	0.7868	0.7750	0.7825	0.7809	0.4588
18	10.0000	0.0000	0.7755	0.7600	0.7700	0.7677	0.4746
19	0.0000	0.0000	0.7868	0.7750	0.7825	0.7809	0.4588
20	10.0000	0.0000	0.7755	0.7600	0.7700	0.7677	0.4746

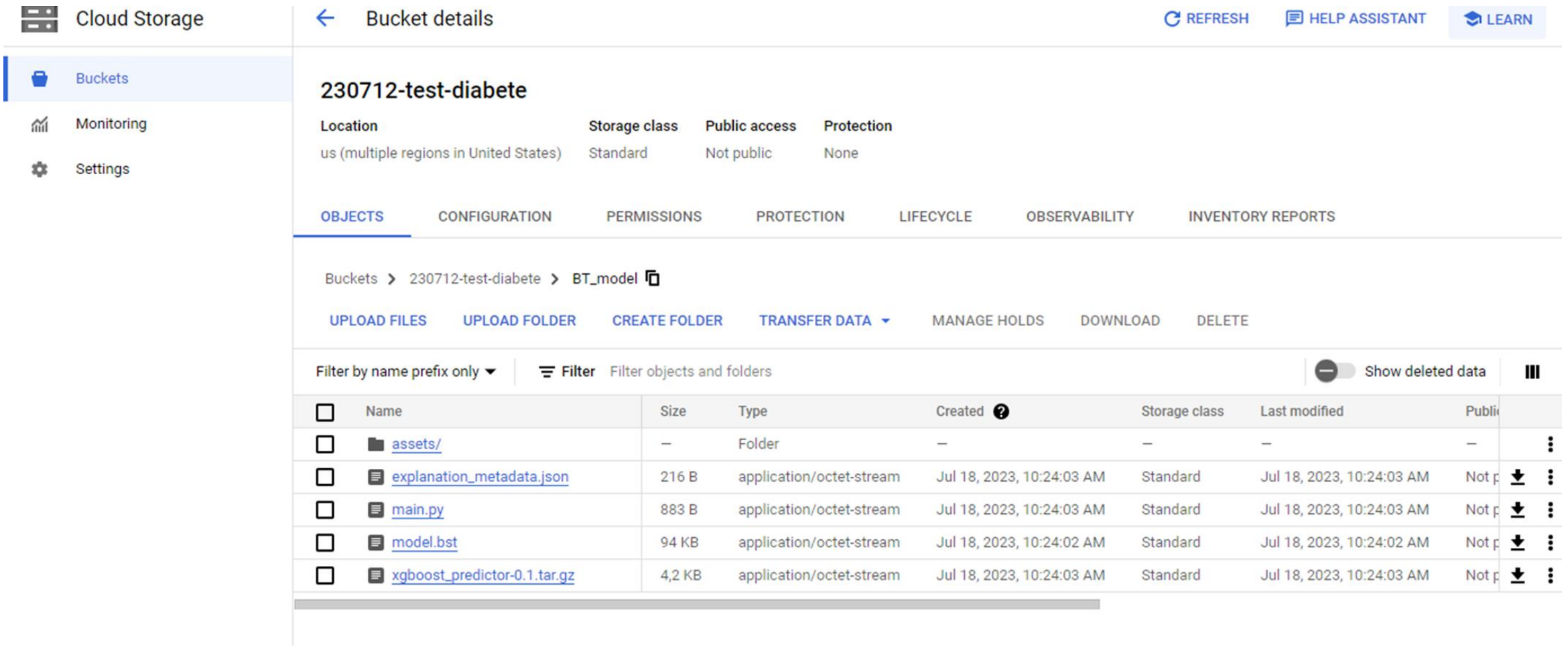
SQL code


Explorer
+ ADD
Type to search
Viewing workspace resources.
SHOW STARRED ONLY
test-diabete
Saved queries (2)
Project queries
Modelling
data_cleaning
External connections
patients
patients
population
Models (5)
Boosted_Trees_mo...
DNN_model
LR_HPT_model
LR_model
RandForest_model
population
*Modelling
LR_HPT_model
Modelling
RUN
SAVE
SHARE
SCHEDULE
MORE
Query complete
FROM
`test-diabete.patients.patients`
);
-- globally explain the model
SELECT
*
FROM
ML.GLOBAL_EXPLAIN(MODEL `test-diabete.population.LR_model`);
-- Hyper Parameter Tuning
CREATE OR REPLACE MODEL
`test-diabete.population.LR_HPT_model`
OPTIONS
(model_type='LOGISTIC_REG',
enable_global_explain=TRUE,
input_label_cols=['label'],
num_trials=20,
max_parallel_trials=2) AS
SELECT
grossesses,pression,imc,K,glucose,insuline,age,label
FROM
`test-diabete.population.population`;
Press Alt+F1 for accessibility optio
Query results
SAVE RESULTS
EXPLORE DATA
JOB INFORMATION
RESULTS
EXECUTION DETAILS
EXECUTION GRAPH
Elapsed time
10 min 6 sec
Slot time consumed
55 min 29 sec
Stages
Preprocess 0
trial_1:Train 0
trial_2:Train 0
trial_1:Evaluate 0
trial_2:Evaluate 0
Training iterations
Completed: 0
Planned: 0

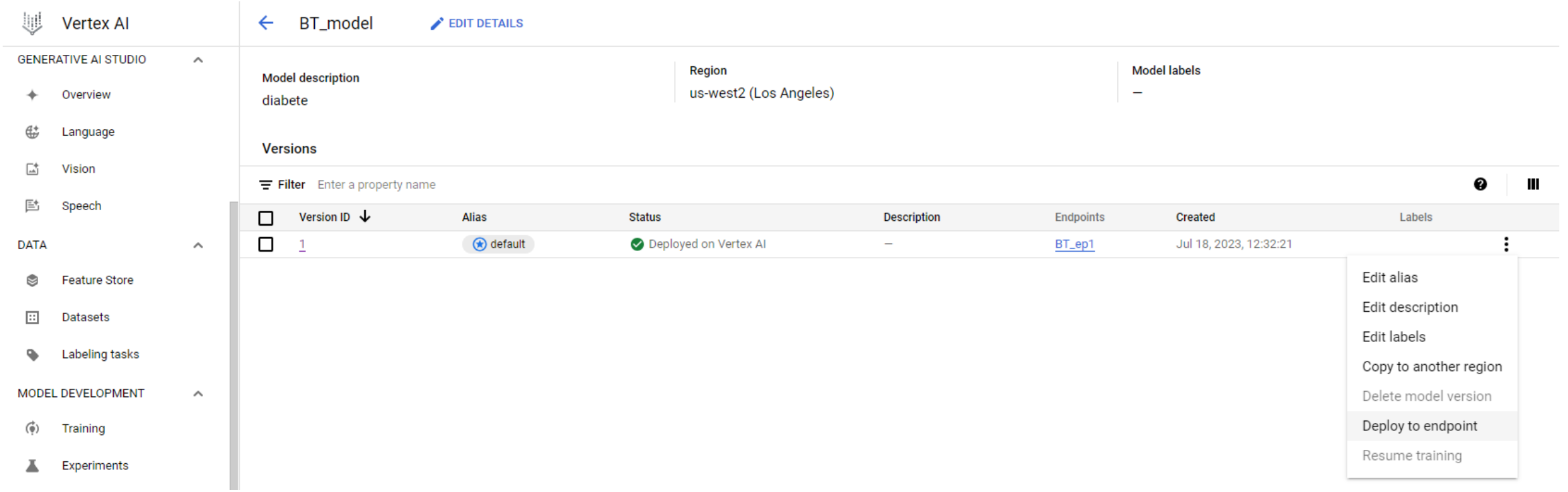
Model deployment

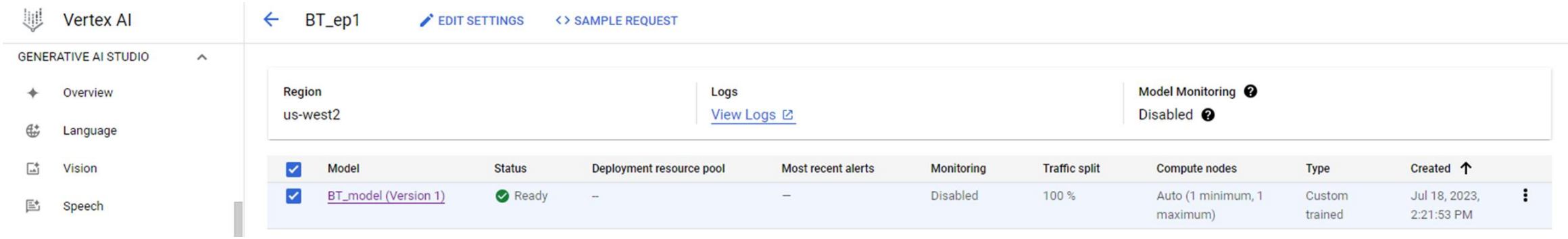
Example for the Boosted Trees model

Export the model to the Bucket

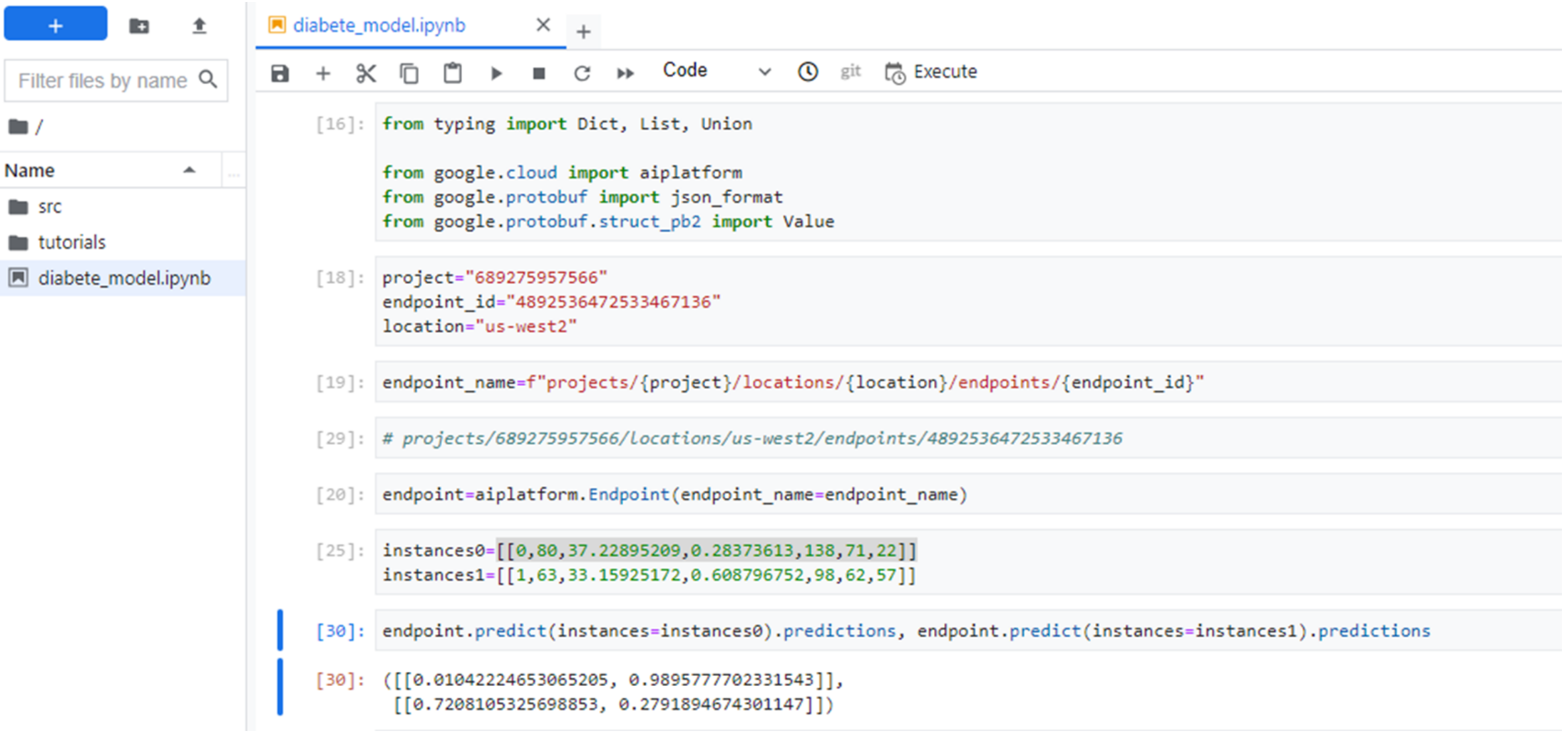


Register and Deploy the model





Workbench request



Summary / Key Learnings

Summary of models performance

Model	Accuracy / Train	Accuracy / Test
Logistic Regression / HPT	0,78 / 0,80	0,72 / 0,71
Random Forrest Trees	0,97	0,89
Deep Neural Network	0,90	0,77
Boosted Trees	1,00	0,96

Major points / technical

1. Data cleaning:
   1. Manual cleaning for modelling is not necessary
   2. BigQuery (BQ) ML automatically preprocesses data cleaning (imputation of missing values & data transformation (standardisation, encoding))
   3. But the type of the missing value is to be NULL
2. BQ Looker gives a lot of possibilities to visualize data & make data stories
3. ML models make data transformation automatically, though it may be done manually
4. There is a possibility for Hyper Parameters Tuning to improve the models' performance