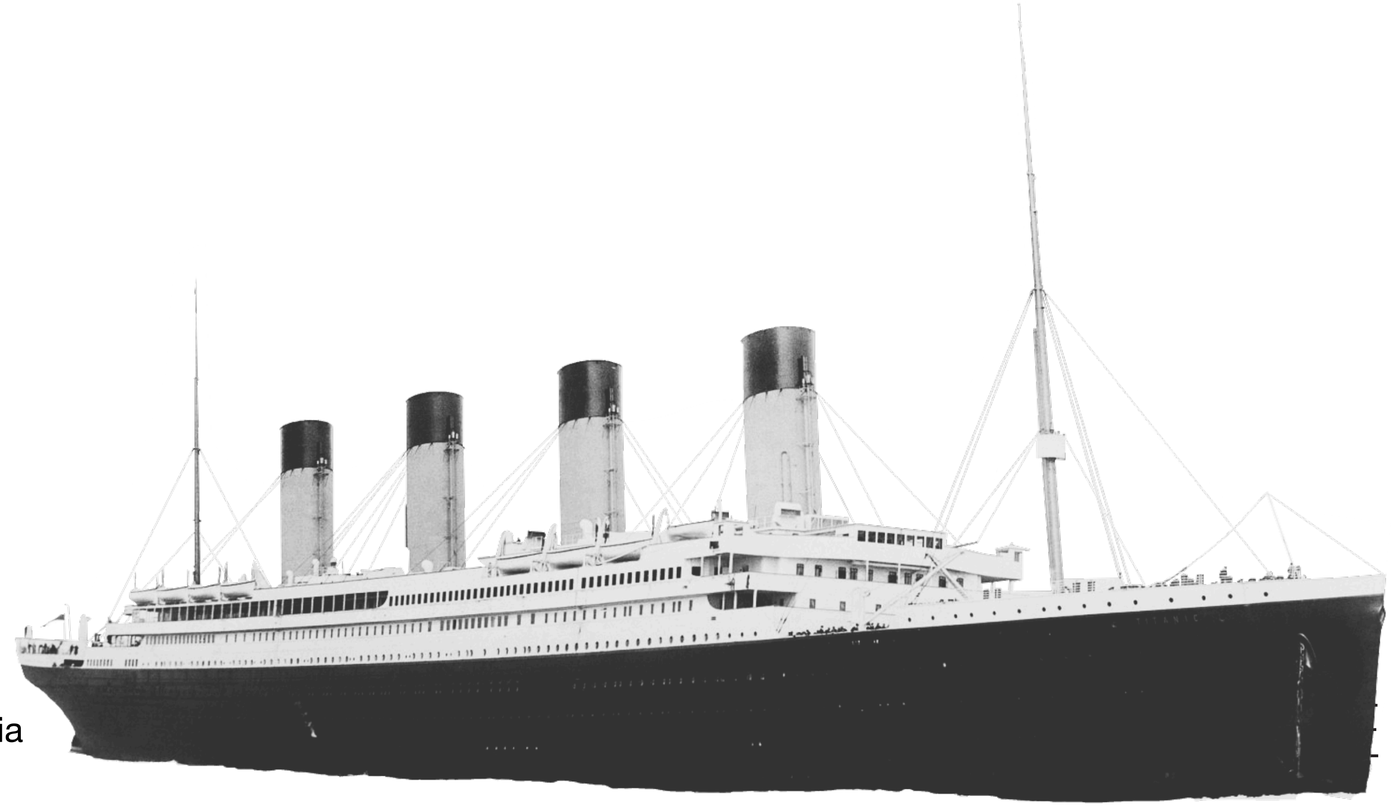
rockPLACE

Machine Learning : 엔드게임
BigQuery의 진화

김하제

kaggle

" Predict which passengers survived the tragedy "



이미지 출처 : Library of Virginia

In [1]:

```python
# 파이썬의 import를 활용해 데이터 분석용 패키지인 판다스(Pandas)를 읽어옵니다.
import pandas as pd

# train.csv 파일을 읽어옵니다. 여기서 PassengerId를 index로 지정해줍니다.
# (=이제 PassengerId를 통해 승객 정보를 읽어올 수 있습니다)
# 읽어온 데이터를 train이라는 이름의 변수에 할당합니다.
train = pd.read_csv("../input/train.csv", index_col="PassengerId")

# train 변수에 할당된 데이터의 행렬 사이즈를 출력합니다.
# 출력은 (row, column) 으로 표시됩니다.
print(train.shape)

# head()로 train 데이터의 상위 5개를 띄웁니다.
train.head()
```

```
(891, 11)
```

Out[1]:

	Survived	Pclass	Name	Sex
PassengerId				
1	0	3	Braund, Mr. Owen Harris	male
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	fema

Sex

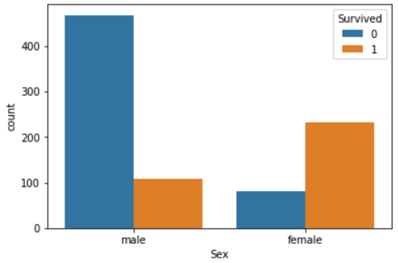
먼저 분석할 컬럼은 성별(Sex)입니다. 우리가 이전에 분석한대로, 성별(Sex) 컬럼은 타이타닉 데이터셋에서 가장 중요한 컬럼입니다. seaborn에서는 countplot이라는 간단한 기능을 통해 성별 컬럼을 시각화 할 수 있습니다.

In [4]:

```python
# 타이타닉의 train 데이터를 바탕으로 성별 컬럼을 시각화합니다.
# 크게 1) 남성 생존자, 2) 남성 사망자, 3) 여성 생존자, 4) 여성 사망자
sns.countplot(data=train, x="Sex", hue="Survived")
```

Out[4]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5ecb43bfd0
```

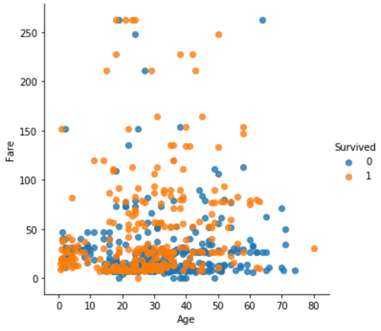


In [12]:

```python
# lmplot을 통해 나이(Age)와 운임요금(Fare)의 상관관계를 분석합니다.
# 다만 이전과는 달리 이번에는 train 데이터가 아닌 low_fare 데이터를 시각화 합니다.
sns.lmplot(data=low_fare, x="Age", y="Fare", hue="Survived", fit_reg=False)
```

Out[12]:

```
<seaborn.axisgrid.FacetGrid at 0x7f5ec7a26400>
```




```python
from sklearn.tree import export_graphviz

# export_graphviz로 Decision Tree를 시각화합니다. 시각화 할 때는 다음의 옵션이 들어갑니다.
# 1) model. 시각화할 트리(Decision Tree) 입니다.
# 2) feature_names. 트리를 만들 때 사용한 feature들의 이름입니다.
# 3) class_names. 살았을 경우 시각화에서 어떻게 표현할 것인지(Survived), 반대로 죽었을 경우 시각화에서 어떻게 표현할 것인지(Perish)를 알려줍니다.
# 4) out_file. 시각화 겨롸를 저장할 파일명입니다. 이번에는 파일로 저장하지 않고 바로 쥬피터 노트북에 띄울 생각이므로 None을 주면 됩니다.
# 마지막으로 시각화한 결과를 dot_tree라는 이름의 변수에 저장합니다.
dot_tree = export_graphviz(model,
                           feature_names=feature_names,
                           class_names=["Perish", "Survived"],
                           out_file=None)

# graphviz에서 Source라는 기능을 통해 Decision Tree를 시각화합니다.
graphviz.Source(dot_tree)
```

Out[49]:

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

SUPERVISED

Predict a category

CLASSIFICATION

«Divide the socks by color»

Predict a number

REGRESSION

«Divide the ties by length»

Data is not labeled in any way

UNSUPERVISED

Divide by similarity

CLUSTERING

«Split up similar clothing into stacks»

Find hidden dependencies

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

Identify sequences

ASSOCIATION

«Find what clothes I often wear together»

rockPLACE

Contents

- Introduce BigQuery ML
- Let's make ML!

rockPLACE

BigQuery

Google Cloud Platform
Enterprise Data Warehouse
for Analytics

Convenience of
Standard SQL

Fully Managed & Serverless

Petabyte-scale Storage
& Queries

Encrypted, Durable
& Highly Available

Real-time Analytics
on Streaming Data

rockPLACE

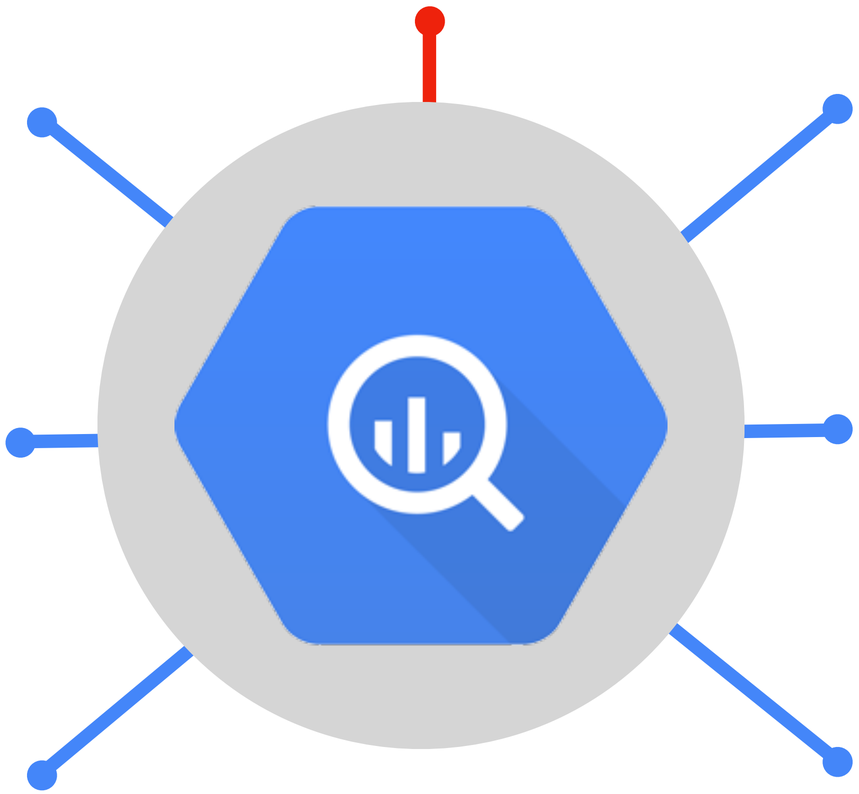
BigQuery ML

Machine Learning

Google Cloud Platform
Enterprise Data Warehouse
for Analytics

Convenience of
Standard SQL

Fully Managed & Serverless



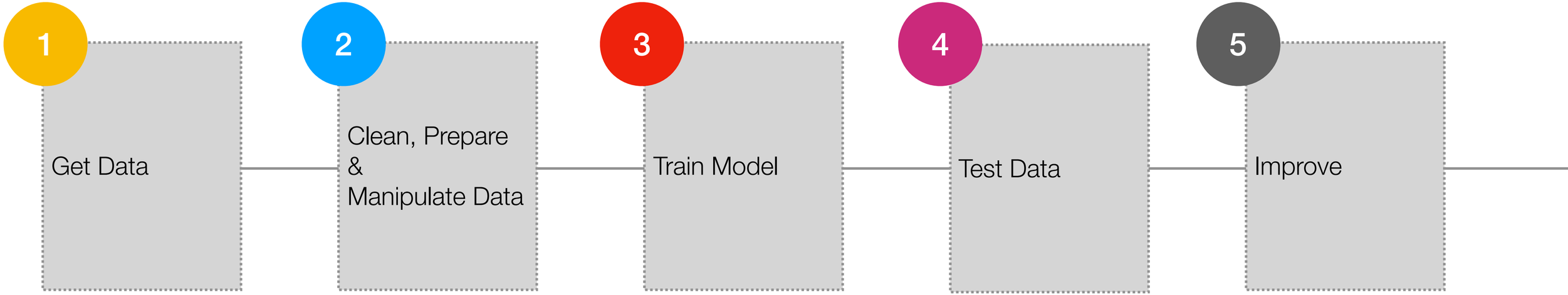
Petabyte-scale Storage
& Queries

Encrypted, Durable
& Highly Available

Real-time Analytics
on Streaming Data

rockPLACE

Flow of Machine Learning



rockPLACE

Flow of BigQuery ML

1. Use familiar SQL for Machine Learning without moving data from BigQuery
2. Train models over all their data in BigQuery to increase development speed
3. Not worry about hypertuning or feature transformations

Supported features

- Standard SQL and UDFs within the ML queries
- Linear Regression (Forecasting)
- Binary Logistic Regression (Classification)
- Model evaluation functions for standard metrics, including the ROC curve
- Model weight inspection
- Feature distribution analysis through standard functions

The possibilities are endless

Marketing

Predict customer lifetime value

Predict funnel conversion

Personalize ads, email, webpage content

Retail

Optimize inventory

Forecast revenue

Enable product recommendations

Optimize staff promotions

Industrial / IoT

Forecast demand for parking, traffic, utilities, personnel

Predict maintenance needs

Prevent equipment downtime

Media / Gaming

Personalize content

Predict game difficulty

Predict player lifetime value

rockPLACE

```
1 CREATE MODEL numbikes.model
2 OPTIONS
3  (model_type='linear_reg', labels=['num_trips']) AS
4 WITH bike_data AS
5 (
6 SELECT
7  COUNT(*) as num_trips,
```

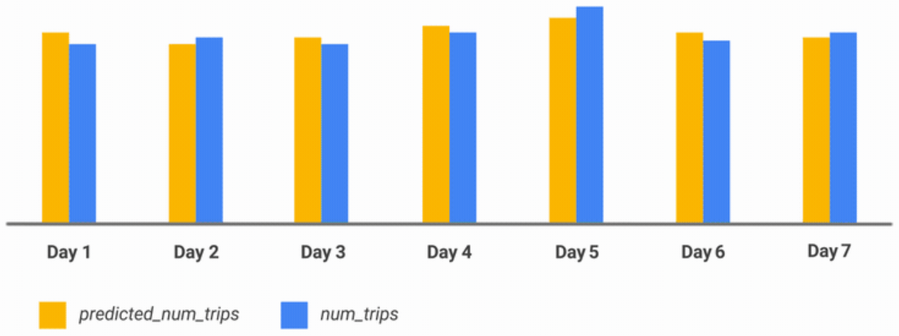
Run Query

```
1 SELECT
2  predicted_num_trips, num_trips, trip_date
3 FROM
4  ml.PREDICT(MODEL `numbikes.model`,(WITH bike_data AS
5 (
6 SELECT
7  COUNT(*) as num_trips,
```

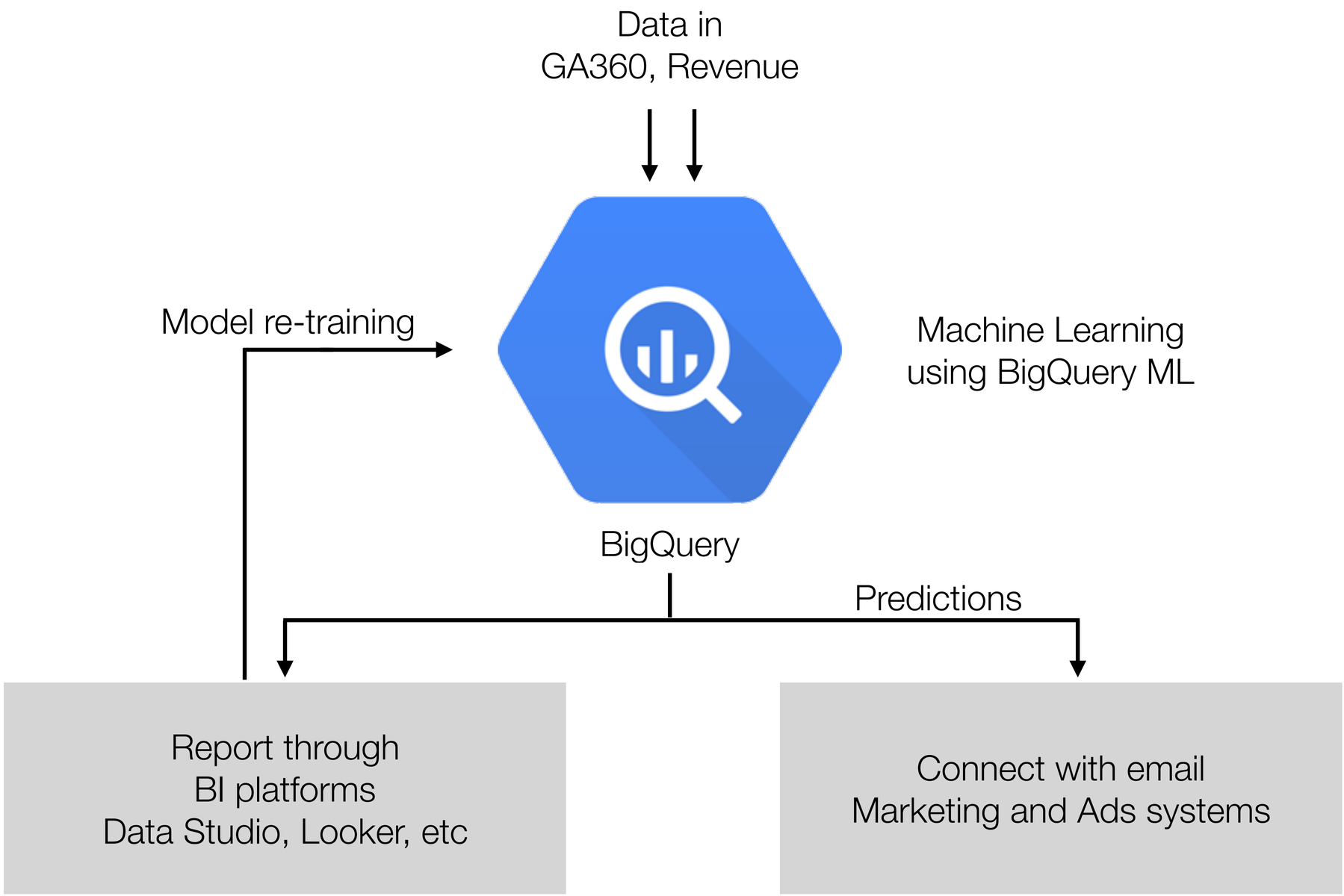
Run Query

Export

predicted_num_trips	num_trips	trip_date



rockPLACE


Data in
GA360, Revenue
Model re-training
Machine Learning
using BigQuery ML
BigQuery
Predictions
Report through
BI platforms
Data Studio, Looker, etc
Connect with email
Marketing and Ads systems

rockPLACE

“CREATE MODEL”

```
CREATE MODEL `[MODEL-NAME]`
OPTIONS(model_type='[TYPE]')

# Model Type
# linear_reg   : Linear Regression
# logistic_reg: Logistic Regression
# kmeans       : K-means Clustering
```

rockPLACE

“ML.EVALUATE”

```
SELECT
  *
FROM
  ML.EVALUATE(MODEL `[MODEL-NAME]`, (
SELECT
  IF(totals.transactions …

# Evaluate Type
# ML.EVALUATE : Precision, Recall, Accuracy, F1 Score,
#               ROC AUC, MAE, MSE, MSLE, R2 Score
# ML.ROC_CURVE : Threshold, Recall, FP Rate, TP, FP, TN,
#                FN
# ML.CONFUSION_MATRIX : Confusion Matrix
```

“ML.PREDICT”

```
SELECT
  *
FROM
  ML.PREDICT(MODEL `[MODEL-NAME]`, (
SELECT
  label,
  …
```

rockPLACE

" Let's make ML! "

Tutorial 1 - 웹쇼핑몰 방문자 거래 예측하기 (GUI & Jupyter Notebook)
Tutorial 2 - 농구 게임 스코어 예측하기 (Jupyter Notebook)
Tutorial 3 - 출생 체중 예측하기 (Jupyter Notebook)

rockPLACE

- Tutorial 1 -

웹쇼핑몰 방문자 구매 예측하기

Logistic Regression

Row	visitorId	visitNumber	visitId	visitStartTime	date	totals.visits	totals.hits	totals.pageviews	totals.timeOnSite	totals.bounces	totals.transactions	totals.transactionRevenue	totals.newVisits	totals.screenviews	totals.uniqueScreenviews	totals.timeOn
1	null	21	1498434758	1498434758	20170625	1	68	48	3343	null	1	44580000	null	null	null	

hits.product.productCouponCode	hits.promotion.promoId	hits.promotion.promoName	hits.promotion.promoCreative	hits.promotion.promoPosition	hits.promotionActionInfo.promoIsView	hits.promotionActionInfo.promoIsClick	hits.refund.refundAmount	hits.refund.loca
	Apparel Row 1	Apparel	home_main_link_apparel.jpg	Row 1	true	null		
	Backpacks Row 2 Combo	Backpacks	home_bags_google_2.jpg	Row 2 Combo				
	Mens T-Shirts Row 3-1	Mens T-Shirts	mens-tshirts.jpg	Row 3-1				
	Womens T-Shirts Row 3-2	Womens T-Shirts	womens-tshirts.jpg	Row 3-2				
	Office Row 5 Color Combo	Office	green_row_link_to_office.jpg	Row 5 Color Combo				
	Drinkware Row 4 Color Combo	Drinkware	red_row_hydrate.jpg	Row 4 Color Combo				
	Google Brand Row 7-1	Google Brand	home_lower_google_500.jpg	Brand Row 7-1				
	YouTube Brand Row 7-2	YouTube Brand	home_lower_youtube_500.jpg	Brand Row 7-2				
	Android Brand Row 7-3	Andriod Brand	home_lower_android_500.jpg	Brand Row 7-3				
null								

- Tutorial 2 -
농구 게임 스코어 예측하기

Linear Regression

Row	game_id	season	scheduled_date	team_id	opponent_id	total_three_points_made	total_three_points_att	team_points_avg_last_1	team_points_avg_last_
1	54396c2c-5a5b-4892-afc8-fa69b49dd97b	2013	2014-03-18	2013:8736b67f-5924-400b-bf94-3bd804c36600	2013:f770d3a2-221c-49ce-9c7a-680f7417fe9e	13	40	56.0	67.
2	1a468fb6-8545-4e8d-87d8-ef0f7d494365	2013	2014-03-19	2013:0d8a328f-20ce-410b-bed1-15fc01308aaf	2013:52df1e19-b142-4a76-a439-ad68455d0581	6	28	78.0	70.
3	c8ff45e3-3ef2-4241-9c5c-65cca4a4b60c	2013	2014-03-20	2013:9fccbf28-2858-4263-821c-fdefb3c7efa3	2013:24051034-96bb-4f78-a3a6-312f3258780f	8	30	43.0	57.
4	51d7ce42-bfc7-41cc-b995-48fdb8fbbc35	2013	2014-03-20	2013:b827dbdf-230f-4916-9703-3bb9004e26eb	2013:c10544de-e3bd-4776-ba2e-83df8c017fd1	9	37	49.0	67.
5	37902a80-a982-4df7-b5a4-89b54ca662b2	2013	2014-03-20	2013:1e42cce3-236b-4f3e-9650-2c45e6332e1e	2013:52df1e19-b142-4a76-a439-ad68455d0581	13	27	68.0	63.
6	ae2be8bc-de3b-49d8-b22a-464d92455f54	2013	2014-03-21	2013:1b1977ec-b32f-4ce3-95a4-d07bab4816cd	2013:683ab61f-546f-44da-b085-c3a5740554aa	12	36	70.0	71.
7	17c101d9-1da9-4c65-8599-9cf758429088	2013	2014-03-21	2013:db6e1cab-3fa3-4a93-a673-8b2a358ff4bf	2013:f8b1bf00-5b73-4ac4-98c0-ec554027ae32	6	34	65.0	76.
8	668626a5-76c4-4194-a1f4-65e85a157e33	2013	2014-03-21	2013:2f4d21f8-6d5f-48a5-abca-52a30583871a	2013:8ab00d43-840a-4c96-bdee-bf88fa6e3f11	12	33	66.0	65.
9	4a3f1890-ce8d-48cc-ba9c-107fc3e58fc4	2013	2014-03-21	2013:e9ca48b2-00ba-41c0-a02b-6885a2da1ff1	2013:7d3b9c72-cbdd-4b64-9aa3-a5bde9b75fdc	8	31	75.0	72.
10	6676d427-0ba6-4215-bf8e-381e08748614	2013	2014-03-21	2013:88104678-e53b-43b3-82f7-efb3a11cedb9	2013:b827dbdf-230f-4916-9703-3bb9004e26eb	5	25	62.0	68.
11	9dd937a4-2b43-47d0-991d-a0dbd7ac1cfc	2013	2014-03-21	2013:6ed15092-2670-450a-99c2-61d861e87644	2013:ad4bc983-8d2e-4e6f-a8f9-80840a786c64	13	27	69.0	63.
12	32a8c6b0-df2d-4869-9625-e192b91ad82d	2013	2014-03-22	2013:2267a1f4-68f6-418b-aaf6-2aa0c4b291f1	2013:d203f38a-a166-4258-bca2-e161b591ecfb	9	36	60.0	67.

rockPLACE

- Tutorial 3 -
출생 체중 예측하기

Linear Regression

	source_year	year	month	day	wday	state	is_male	child_race	weight_pounds	plurality	...	alcohol_use	drinks_per_week	weight_gain_pounds	bor
0	2007	2007	3	NaN	2.0	None	True	NaN	9.060999	1.0	...	None	NaN	52.0	
1	2005	2005	7	NaN	2.0	None	True	NaN	9.625382	1.0	...	False	NaN	15.0	
2	1993	1993	6	NaN	2.0	CA	False	1.0	7.292892	1.0	...	None	NaN	99.0	
3	1970	1970	6	11.0	NaN	KS	False	1.0	6.686620	NaN	...	None	NaN	NaN	
4	1990	1990	4	NaN	5.0	KS	True	1.0	8.999270	1.0	...	False	NaN	36.0	
5	1970	1970	3	26.0	NaN	KY	False	1.0	5.061814	NaN	...	None	NaN	NaN	
6	1982	1982	1	4.0	NaN	MD	True	1.0	9.374055	1.0	...	None	NaN	NaN	
7	1976	1976	7	21.0	NaN	MI	False	1.0	8.437091	1.0	...	None	NaN	NaN	
8	1972	1972	9	18.0	NaN	NE	True	1.0	8.062305	1.0	...	None	NaN	NaN	
9	1974	1974	9	30.0	NaN	TX	False	1.0	7.749249	1.0	...	None	NaN	NaN	
10	1983	1983	1	16.0	NaN	TX	True	1.0	8.624484	1.0	...	None	NaN	NaN	
11	1980	1980	2	8.0	NaN	WA	True	1.0	10.813674	1.0	...	None	NaN	NaN	
12	1969	1969	2	4.0	NaN	CA	True	1.0	7.813183	NaN	...	None	NaN	NaN	
13	2003	2003	12	NaN	5.0	CA	False	NaN	7.251004	1.0	...	None	NaN	99.0	
14	1970	1970	12	27.0	NaN	NY	True	1.0	3.875727	NaN	...	None	NaN	NaN	
15	1986	1986	9	28.0	NaN	OH	False	1.0	0.562179	1.0	...	None	NaN	NaN	

rockPLACE

" It's the end game. "

Google BigQuery ML tutorials
https://cloud.google.com/bigquery-ml/docs/tutorials

BigQuery ML - Machine Learning using SQL in BigQuery
https://www.youtube.com/watch?v=BanOYQVl30I

How to Do Predictive Analytics in BigQuery (Cloud Next '18)
https://www.youtube.com/watch?v=Ml2aCTn7kFY

What's New with BigQuery ML and Using it to Assess Data Quality (Cloud Next '19)
https://www.youtube.com/watch?v=DnlG4frLKmw

2019 Google Bigdata Showcase in Seoul
https://github.com/hajekim/endgame

rockPLACE