Google Cloud

Forecasting Fundamentals

Agenda

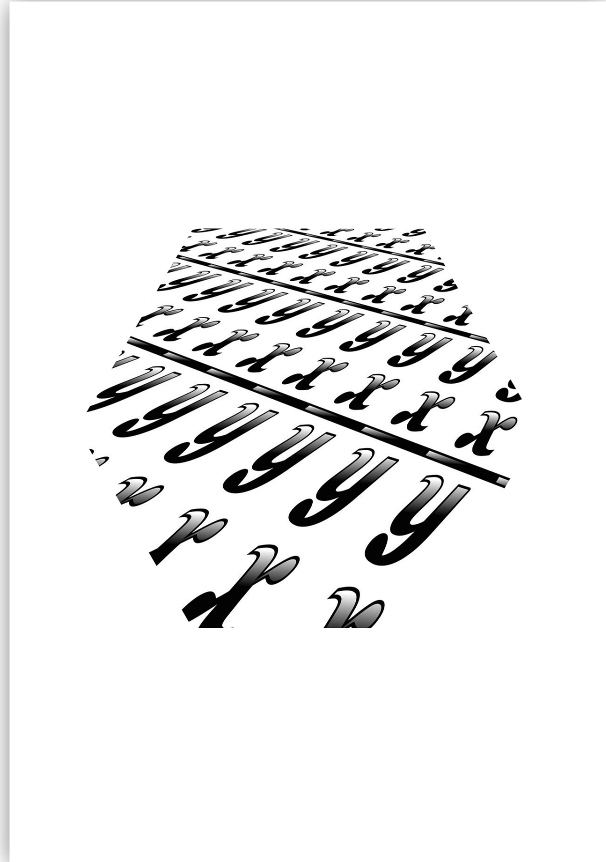
What is Forecasting

Terminology

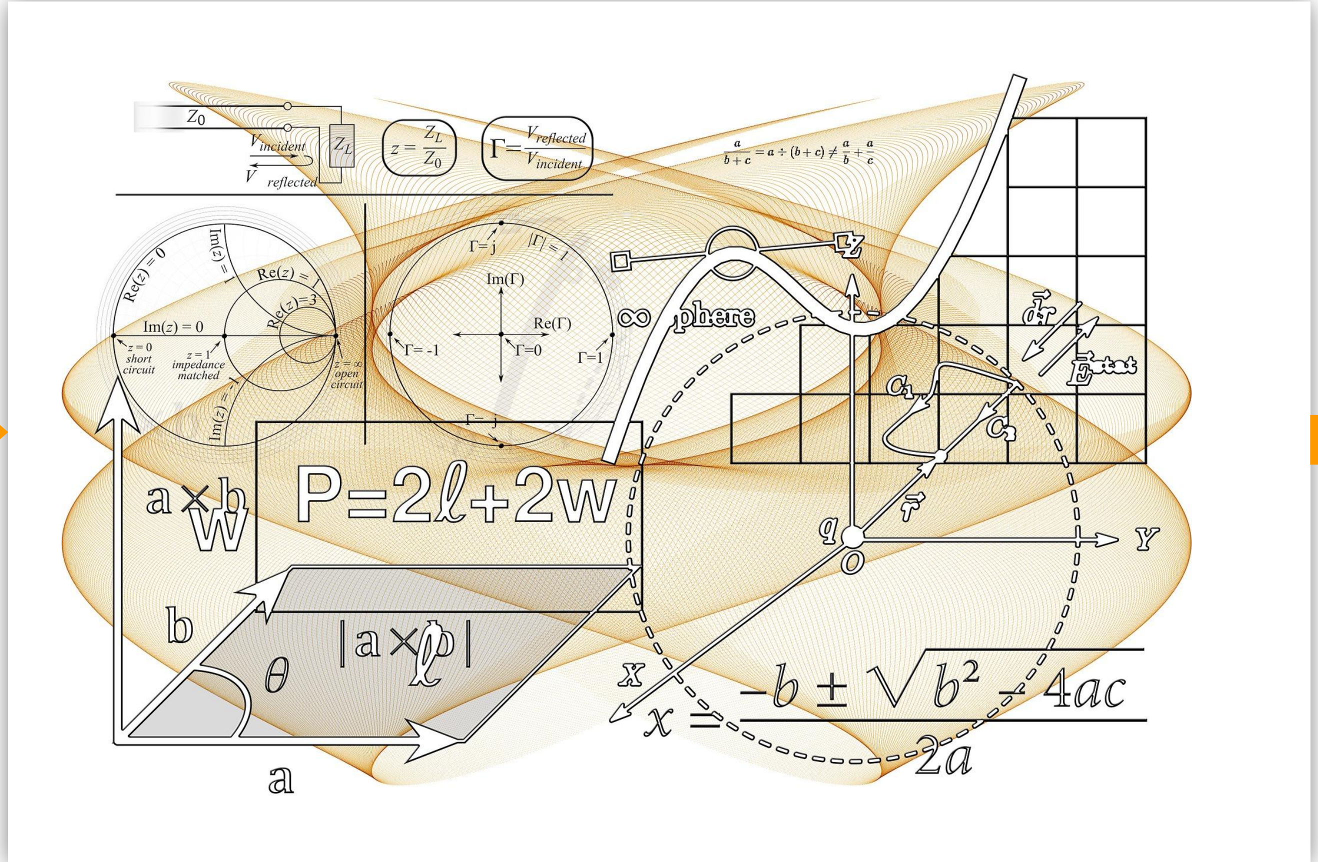
Choosing the right model

Introducing BQML

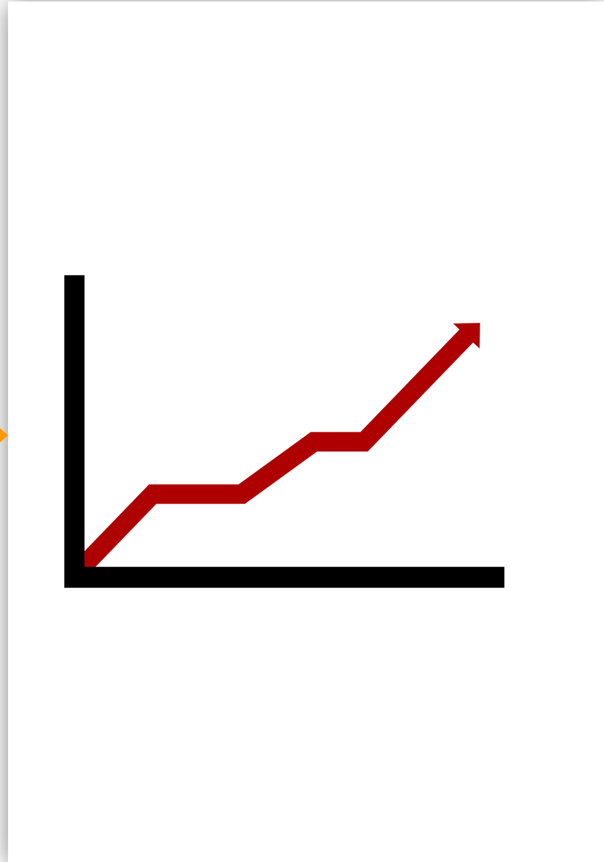
Forecasting



Lots of data



Complex mathematics in multidimensional spaces



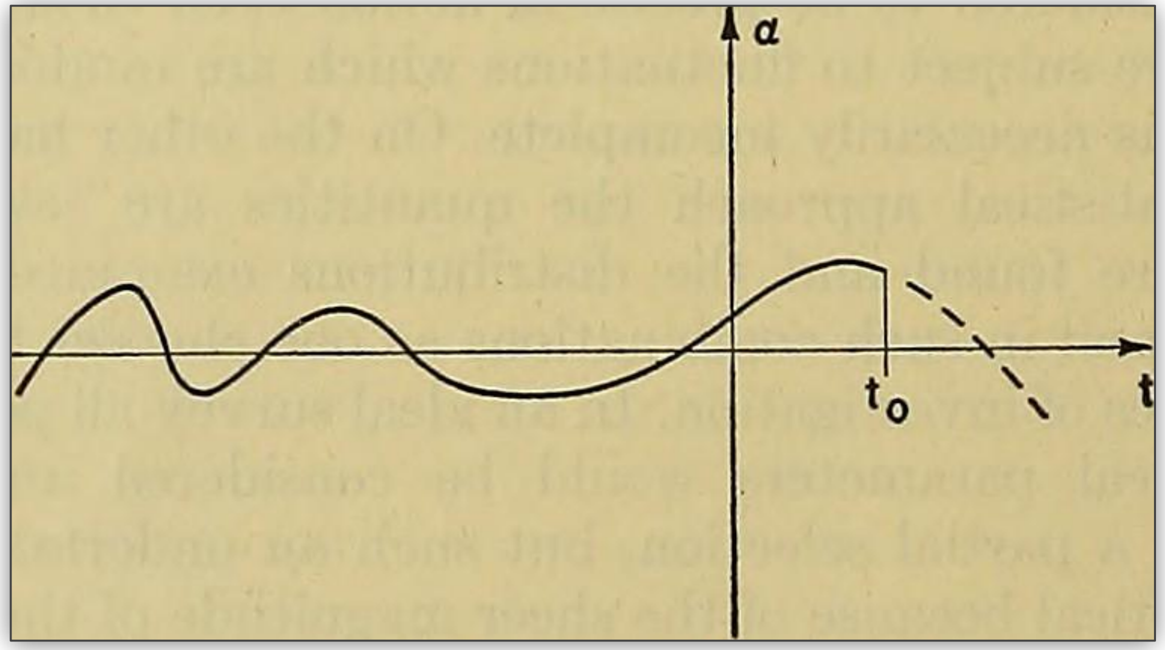
Magical results

Forecasting vs. Regression

Forecasting

Causal → Regression → Regression: Models the relationship between two or more explanatory variables and a response variable

Time Series → MA → Simple Average Uses an average of all past data as a forecast.



Forecasting

Forecasting vs. Regression

Both Regression and Time Series forecast:

- Collect data
- Organize data
- Create model
- Experiment a lot
- Produce Results

Regression:

Use variables to explain the response

Time Series Forecast:

Use past data to predict future

Agenda

What is Forecasting

Terminology

Choosing the right model

Introducing BQML

A quick example

Predicting customer lifetime value with a ML model

Let’s predict the lifetime value of an ecommerce customer using Regression

ABC **hits_product_v2ProductName**

177 Categories

Google Men's Vintage Badge Tee Black
Google Women's 1/4 Zip Performance Pullover Two-Tone Blue
Google Men's Lightweight Microfleece Jacket Black
Google Men's Lightweight Microfleece Jacket Black
Google Women's Quilted Insulated Vest Black
Google Women's Lightweight Microfleece Jacket
Google Women's Lightweight Microfleece Jacket
Women's Weatherblock Shell Jacket Black
Men's Weatherblock Shell Jacket Black

Google Analytics provides us with aggregated site visit metrics

Results | Details

Row	fullVisitorId	distinct_days_visited	ltv_pageviews	ltv_visits	ltv_avg_time_on_site_s	ltv_revenue	ltv_transactions
1	7813149961404844386	79	1395	138	479.63	6245720000	67
2	7713012430069756739	2	514	6	1954.33	181940000	35
3	6760732402251466726	30	868	41	723.55	4812820000	34
4	5526675926038480325	1	466	1	7013.0	87960000	25
5	1957458976293878100	148	4303	284	796.46	77113430000	22
6	4983264713224875783	2	366	4	3807.5	74850000	21
7	2402527199731150932	28	559	31	906.61	3270100000	19

In ML terms, an instance (or observation) is a row of data

Results | Details

Row	fullVisitorId	distinct_days_visited	ltv_pageviews	ltv_visits	ltv_avg_time_on_site_s	ltv_revenue	ltv_transactions	avg_session_quality	first_visit	last_visit	ltv_days
1	6007196403211981721	8	147	11	772.5	null	null	7.5	2016-08-04	2017-07-15	345
2	7587138749751940102	9	94	9	312.33	24380000	1	1.0	2016-08-03	2017-07-14	345
3	0720311197761340948	114	148	146	2118.0	null	null	1.0	2016-08-05	2017-07-15	344
4	9557989866096732580	3	18	3	356.5	null	null	1.0	2016-08-03	2017-07-13	344
5	0824839726118485274	127	3153	282	1520.0	null	null	26.0	2016-08-01	2017-07-10	343
6	2742641486650042668	17	113	20	266.28	387000000	2	23.0	2016-08-02	2017-07-11	343
7	1957458976293878100	148	4303	284	796.46	77113430000	22	1.5	2016-08-04	2017-07-12	342
8	1950585318332186454	6	19	7	51.4	null	null	1.5	2016-08-05	2017-07-11	340

What we are trying to predict for is the label

Results | Details

Row	fullVisitorId	distinct_days_visited	ltv_pageviews	ltv_visits	ltv_avg_time_on_site_s	ltv_revenue	ltv_transactions	avg_session_quality	first_visit	last_visit	ltv_days
1	6007196403211981721	8	147	11	772.5	null	null	7.5	2016-08-04	2017-07-15	345
2	7587138749751940102	9	94	9	312.33	24380000	1	1.0	2016-08-03	2017-07-14	345
3	0720311197761340948	114	148	146	2118.0	null	null	1.0	2016-08-05	2017-07-15	344
4	9557989866096732580	3	18	3	356.5	null	null	1.0	2016-08-03	2017-07-13	344
5	0824839726118485274	127	3153	282	1520.0	null	null	26.0	2016-08-01	2017-07-10	343
6	2742641486650042668	17	113	20	266.28	387000000	2	23.0	2016-08-02	2017-07-11	343
7	1957458976293878100	148	4303	284	796.46	77113430000	22	1.5	2016-08-04	2017-07-12	342
8	1950585318332186454	6	19	7	51.4	null	null	1.5	2016-08-05	2017-07-11	340

Here we are predicting the lifetime revenue (number)

Labels could also be a discrete class of customer like “high value”

Row	fullVisitorId	distinct_days_visited	ltv_pageviews	ltv_visits	ltv_avg_time_on_site_s	ltv_revenue	ltv_transactions	avg_session_quality	first_visit	last_visit	ltv_days	label
1	7587138749751940102	9	94	9	312.33	24380000	1	1.0	2016-08-03	2017-07-14	345	High Value Customer
2	6007196403211981721	8	147	11	772.5	null	null	7.5	2016-08-04	2017-07-15	345	
3	9557989866096732580	3	18	3	356.5	null	null	1.0	2016-08-03	2017-07-13	344	
4	0720311197761340948	114	148	146	2118.0	null	null	1.0	2016-08-05	2017-07-15	344	
5	2742641486650042668	17	113	20	266.28	387000000	2	23.0	2016-08-02	2017-07-11	343	High Value Customer
6	0824839726118485274	127	3153	282	1520.0	null	null	26.0	2016-08-01	2017-07-10	343	
7	1957458976293878100	148	4303	284	796.46	77113430000	22	1.5	2016-08-04	2017-07-12	342	High Value Customer
8	9801276214964695322	79	462	106	219.44	null	null	1.5	2016-08-01	2017-07-07	340	
9	1950585318332186454	6	19	7	51.4	null	null	1.5	2016-08-05	2017-07-11	340	
10	0084834161383601528	7	97	7	258.0	69260000	2	2.0	2016-08-04	2017-07-10	340	High Value Customer
11	928398408398925152	40	553	43	285.37	462190000	2	2.0	2016-08-02	2017-07-07	339	High Value Customer
12	351277725820061611	20	60	20	221.33	null	null	1.0	2016-08-05	2017-07-10	339	
13	4143624098732715494	6	13	7	52.5	null	null	1.0	2016-08-03	2017-07-08	339	
14	1927175312147751345	13	180	14	427.21	44970000	1	2.0	2016-08-03	2017-07-08	339	High Value Customer
15	1315772786660606104	28	272	36	340.3	279320000	3	21.25	2016-08-09	2017-07-14	339	High Value Customer

What you are trying to predict for (number or discrete class) influences the model you will choose

The other columns of data are your potential feature columns

Row	fullVisitorId	distinct_days_visited	ltv_pageviews	ltv_visits	ltv_avg_time_on_site_s	ltv_revenue	ltv_transactions	avg_session_quality	first_visit	last_visit	ltv_days	label
1	7587138749751940102	9	94	9	312.33	24380000	1	1.0	2016-08-03	2017-07-14	345	High Value Customer
2	6007196403211981721	8	147	11	772.5	null	null	7.5	2016-08-04	2017-07-15	345	
3	9557989866096732580	3	18	3	356.5	null	null	1.0	2016-08-03	2017-07-13	344	
4	0720311197761340948	114	148	146	2118.0	null	null	1.0	2016-08-05	2017-07-15	344	
5	2742641486650042668	17	113	20	266.28	387000000	2	23.0	2016-08-02	2017-07-11	343	High Value Customer
6	0824839726118485274	127	3153	282	1520.0	null	null	26.0	2016-08-01	2017-07-10	343	
7	1957458976293878100	148	4303	284	796.46	77113430000	22	1.5	2016-08-04	2017-07-12	342	High Value Customer
8	9801276214964695322	79	[illegible]	[illegible]	[illegible]	[illegible]	null	1.5	2016-08-01	2017-07-07	340	
9	1950585318332186454	6	[illegible]	7	[illegible]	null	null	1.5	2016-08-05	2017-07-11	340	
10	0084834161383601528	7	97	7	258.0	69260000	2	2.0	2016-08-04	2017-07-10	340	High Value Customer
11	9283984083989255152	40	553	43	285.37	462190000	2	2.0	2016-08-02	2017-07-07	339	High Value Customer
12	351277725820061611	20	60	20	221.33	null	null	1.0	2016-08-05	2017-07-10	339	
13	4143624098732715494	6	13	7	52.5	null	null	1.0	2016-08-03	2017-07-08	339	
14	1927175312147751345	13	180	14	427.21	44970000	1	2.0	2016-08-03	2017-07-08	339	High Value Customer
15	1315772786660606104	28	272	36	340.3	279320000	3	21.25	2016-08-09	2017-07-14	339	High Value Customer

Feature Columns

You will try to model the relationship between the features and your label

What if I don't know where a new customer will fit?

Historical Training Data (Known LTV

Row	fullVisitorId	distinct_days_visited	ltv_pageviews	ltv_visits	ltv_avg_time_on_site_s	ltv_revenue	ltv_transactions	avg_session_quality	first_visit	last_visit	ltv_days	label
1	7587138749751940102	9	94	9	312.33	24380000	1	1.0	2016-08-03	2017-07-14	345	High Value Customer
2	6007196403211981721	8	147	11	772.5	null	null	7.5	2016-08-04	2017-07-15	345	
3	9557989866096732580	3	18	3	356.5	null	null	1.0	2016-08-03	2017-07-13	344	
4	0720311197761340948	114	148	146	2118.0	null	null	1.0	2016-08-05	2017-07-15	344	
5	2742641486650042668	17	113	20	266.28	387000000	2	23.0	2016-08-02	2017-07-11	343	High Value Customer
6	0824839726118485274	127	3153	282	1520.0	null	null	26.0	2016-08-01	2017-07-10	343	
7	1957458976293878100	148	4303	284	796.46	77113430000	22	1.5	2016-08-04	2017-07-12	342	High Value Customer
8	9801276214964695322	79	462	106	219.44	null	null	1.5	2016-08-01	2017-07-07	340	
9	1950585318332186454	6	19	7	51.4	null	null	1.5	2016-08-05	2017-07-11	340	
10	0084834161383601528	7	97	7	258.0	69260000	2	2.0	2016-08-04	2017-07-10	340	High Value Customer
11	928398408398925152	40	553	43	285.37	462190000	2	2.0	2016-08-02	2017-07-07	339	High Value Customer
12	351277725820061611	20	60	20	221.33	null	null	1.0	2016-08-05	2017-07-10	339	
13	4143624098732715494	6	13	7	52.5	null	null	1.0	2016-08-03	2017-07-08	339	
14	1927175312147751345	13	180	14	427.21	44970000	1	2.0	2016-08-03	2017-07-08	339	High Value Customer
15	1315772786660606104	28	272	36	340.3	279320000	3	21.25	2016-08-09	2017-07-14	339	High Value Customer

Future Data (Unknown LTV)

17	7904807859681747547	3	42	3	1162.0	null	null	1.0	2016-08-05	2017-07-09	338	??????????????????
18	4405445121320750966	51	358	62	517.36	null	null	1.0	2016-08-08	2017-07-12	338	??????????????????
19	1419607020881916790	5	22	5	711.0	null	null	1.0	2016-08-12	2017-07-15	337	??????????????????
20	3862335714593915688	13	92	16	154.23	238000000	1	2.0	2016-08-09	2017-07-12	337	??????????????????

What if I don't know where a new customer will fit?

Historical Training Data (Known LTV

Row	fullVisitorId	distinct_days_visited	ltv_pageviews	ltv_visits	ltv_avg_time_on_site_s	ltv_revenue	ltv_transactions	avg_session_quality	first_visit	last_visit	ltv_days	label
1	7587138749751940102	9	94	9	312.33	24380000	1	1.0	2016-08-03	2017-07-14	345	High Value Customer
2	6007196403211981721	8	147	11	772.5	null	null	7.5	2016-08-04	2017-07-15	345	
3	9557989866096732580	3	18	3	356.5	null	null	1.0	2016-08-03	2017-07-13	344	
4	0720311197761340948	114	148	146	2118.0	null	null	1.0	2016-08-05	2017-07-15	344	
5	2742641486650042668	17	113	20	266.28	387000000	2	23.0	2016-08-02	2017-07-11	343	High Value Customer
6	0824839726118485274	127	3153	282	1520.0	null	null	26.0	2016-08-01	2017-07-10	343	
7	1957458976293878100	148	4303	284	796.46	77113430000	22	1.5	2016-08-04	2017-07-12	342	High Value Customer
8	9801276214964695322	79	462	106	219.44	null	null	1.5	2016-08-01	2017-07-07	340	
9	1950585318332186454	6	19	7	51.4	null	null	1.5	2016-08-05	2017-07-11	340	
10	0084834161383601528	7	97	7	258.0	69260000	2	2.0	2016-08-04	2017-07-10	340	High Value Customer
11	9283984083989251528	40	553	43	285.37	462190000	2	2.0	2016-08-02	2017-07-07	339	High Value Customer
12	3512777258200616111	20	60	20	221.33	null	null	1.0	2016-08-05	2017-07-10	339	
13	4143624098732715494	6	13	7	52.5	null	null	1.0	2016-08-03	2017-07-08	339	
14	1927175312147751345	13	180	14	427.21	44970000	1	2.0	2016-08-03	2017-07-08	339	High Value Customer
15	1315772786660606104	28	272	36	340.3	279320000	3	21.25	2016-08-09	2017-07-14	339	High Value Customer

Future Data (Unknown LTV)

17	7904807859681747547	3	42	3	1162.0	null	null	1.0	2016-08-05	2017-07-09	338	??????????????????
18	4405445121320750966	51	358	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2017-07-12	338	??????????????????
19	1419607020881916790	5	22	5	711.0	null	null	1.0	2016-08-12	2017-07-15	337	??????????????????
20	3862335714593915688	13	92	16	154.23	238000000	1	2.0	2016-08-09	2017-07-12	337	??????????????????

Infer or predict it with a model! →

Data instead of rules

How do we predict the value in the future?

After the model is trained, you can see the relative importance of each field

Historical Training Data (Known LTV

Row	fullVisitorId	distinct_days_visited	ltv_pageviews	ltv_visits	ltv_avg_time_on_site_s	ltv_revenue	ltv_transactions	avg_session_quality	first_visit	last_visit	ltv_days	label
1	7587138749751940102	9	94	9	312.33	24380000	1	1.0	2016-08-03	2017-07-14	345	High Value Customer
2	6007196403211981721	8	[illegible]	11	772.5	null	null	[illegible]	[illegible]	[illegible]	345	
3	9557989866096732580	3	[illegible]	3	356.5	null	null	[illegible]	[illegible]	[illegible]	344	
4	0720311197761340948	[illegible]	[illegible]	146	2118.0	null	null	[illegible]	[illegible]	[illegible]	344	
5	2742641486650042668	17	113	20	266.28	387000000	2	23.0	2016-08-02	2017-07-11	343	High Value Customer
6	0824839726118485274	127	3153	282	1520.0	null	null	26.0	2016-08-01	2017-07-10	343	
7	1957458976293878100	148	4303	284	796.46	77113430000	22	1.5	2016-08-04	2017-07-12	342	High Value Customer
8	9801276214964695322	79	462	106	219.44	null	null	1.5	2016-08-01	2017-07-07	340	
9	1950585318332186454	6	19	7	51.4	null	null	1.5	2016-08-05	2017-07-11	340	
10	0084834161383601528	7	97	7	258.0	69260000	2	2.0	2016-08-04	2017-07-10	340	High Value Customer
11	928398408398925152	40	553	43	285.37	462190000	2	2.0	2016-08-02	2017-07-07	339	High Value Customer
12	351277725820061611	20	60	20	221.33	null	null	1.0	2016-08-05	2017-07-10	339	
13	4143624098732715494	6	13	7	52.5	null	null	1.0	2016-08-03	2017-07-08	339	
14	1927175312147751345	13	180	14	427.21	44970000	1	2.0	2016-08-03	2017-07-08	339	High Value Customer
15	1315772786660606104	28	272	36	340.3	279320000	3	21.25	2016-08-09	2017-07-14	339	High Value Customer

0.5 0.4 0.1 0.1 0.1

Future Data (Unknown LTV)

17	7904807859681747547	3	42	3	1162.0	null	null	1.0	2016-08-05	2017-07-09	338	??????????????????
18	4405445121320750966	51	358	62	517.36	null	null	1.0	2016-08-08	2017-07-12	338	??????????????????
19	1419607020881916790	5	22	5	711.0	null	null	1.0	2016-08-12	2017-07-15	337	??????????????????
20	3862335714593915688	13	92	16	154.23	238000000	1	2.0	2016-08-09	2017-07-12	337	??????????????????

In Forecasting, how do we use the target or label?

Results | Details

Row	fullVisitorId
1	6007196403211981721
2	7587138749751940102
3	0720311197761340948
4	9557989866096732580
5	0824839726118485274
6	2742641486650042668
7	1957458976293878100
8	1950585318332186454

Here we will be predicting the lifetime revenue (number) in the future for one customer using their past values!

How do we predict the value in the future?

Historical Training Data (Known LTV

Row	fullVisitorId	distinct_days_visited	ltv_pageviews	ltv_visits	ltv_avg_time_on_site_s	ltv_revenue	ltv_transactions	avg_session_quality	first_visit	last_visit	ltv_days	label
1	7587138749751940102	9	94	9	312.33	24380000	1	1.0	2016-08-03	2017-07-14	345	High Value Customer
2	6007196403211981721	8	147	11	772.5	null	null	7.5	2016-08-04	2017-07-15	345	
3	9557989866096732580	3	18	3	356.5	null	null	1.0	2016-08-03	2017-07-13	344	
4	0720311197761340948	114	148	146	2118.0	null	null	1.0	2016-08-05	2017-07-15	344	
5	2742641486650042668	17	113	20	266.28	387000000	2	[illegible]	[illegible]	[illegible]	343	High Value Customer
6	0824839726118485274	127	3153	282	1520.0	null	null	26.0	2016-08-01	2017-07-10	343	
7	1957458976293878100	148	4303	284	796.46	77113430000	22	[illegible]	[illegible]	[illegible]	342	High Value Customer
8	9801276214964695322	79	462	106	219.44	null	null	1.5	2016-08-01	[illegible]	340	
9	1950585318332186454	6	19	7	51.4	null	null	1.5	[illegible]	2017-07-11	340	
10	0084834161383601528	7	97	7	258.0	69260000	2	[illegible]	[illegible]	[illegible]	340	High Value Customer
11	928398408398925152	40	553	43	285.37	462190000	2	2.0	2016-08-02	2017-07-07	339	High Value Customer
12	351277725820061611	20	60	20	221.33	null	null	1.0	2016-08-05	2017-07-10	339	
13	4143624098732715494	6	13	7	52.5	null	null	1.0	2016-08-03	2017-07-08	339	
14	1927175312147751345	13	180	14	427.21	44970000	1	2.0	2016-08-03	2017-07-08	339	High Value Customer
15	1315772786660606104	28	272	36	340.3	279320000	3	21.25	2016-08-09	2017-07-14	339	High Value Customer

We use the past values of LTV for one customer to predict future values of LTV.

Future Data (Unknown LTV)

17	[illegible]	3	42	3	1162.0	null	null	1.0	2016-08-05	2017-07-09	338	[illegible]
18	[illegible]	51	[illegible]	[illegible]	[illegible]	null	null	[illegible]	[illegible]	[illegible]	338	[illegible]
19	[illegible]	[illegible]	22	5	[illegible]	null	null	1.0	2016-08-12	2017-07-15	337	[illegible]
20	[illegible]	13	92	16	154.23	238000000	1	2.0	2016-08-09	2017-07-12	337	[illegible]

We don’t use the past values of explanatory variables to predict future values of LTV.

ML terminology review

- **Label** = the correct answer typically from historical data (can be number, string, etc.)
- **Feature** = other columns of data for the model to learn from
- **Model** = a computer-determined recipe to get from features to label
- **Model Types** = (we will cover soon)
- **Training** = showing the model lots of examples for it to learn the relationship
- **Weight** = Adjustable parameter of a model.
- **Evaluation** = how the model performs on a set of known labels it has not seen before in training
- **Prediction** = using a trained model to predict on unknown labels

Agenda

What is Forecasting

Terminology

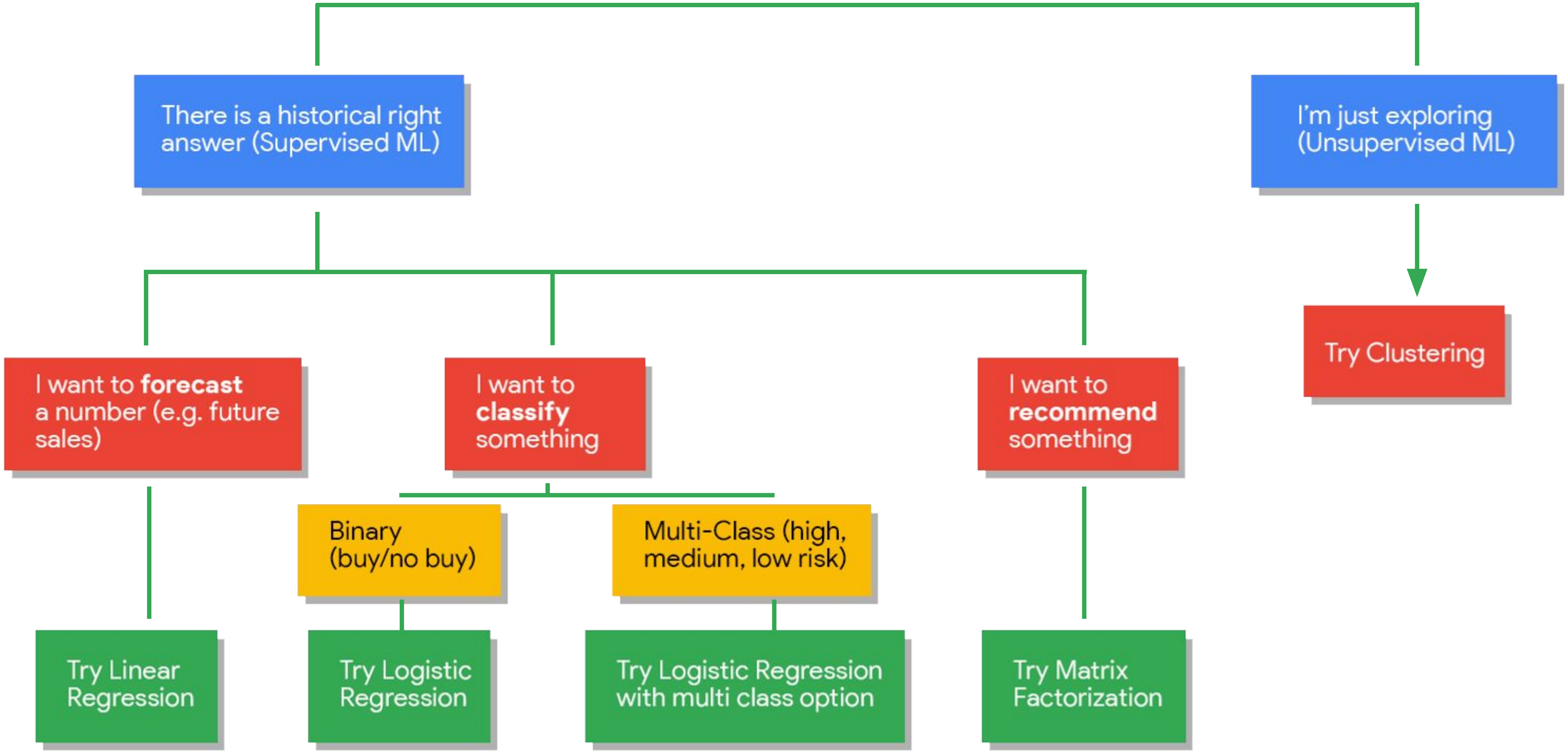
Choosing the right model

Introducing BQML

Okay.. I’ve got data

What model should I use?

Choose the right model type for your structured data use case



Quiz

What model should you use if...
I want to predict ecommerce sales figures for the next quarter

1. Forecasting (linear regression etc..)
2. Classification (logistic regression etc..)
3. Recommendation (matrix factorialization etc..)
4. Unsupervised Learning (clustering etc..)
5. All of the above

Quiz

What model should you use if...
I want to predict ecommerce sales figures for the next quarter

1. Forecasting (linear regression etc..)
2. Classification (logistic regression etc..)
3. Recommendation (matrix factorialization etc..)
4. Unsupervised Learning (clustering etc..)
5. All of the above

Quiz

What model should you use if...
I want to predict whether a user will buy or not buy on a visit

1. Forecasting (linear regression etc..)
2. Classification (logistic regression etc..)
3. Recommendation (matrix factorialization etc..)
4. Unsupervised Learning (clustering etc..)
5. All of the above

Quiz

What model should you use if...
I want to predict whether a user will buy or not buy on a visit

1. Forecasting (linear regression etc..)
2. Classification (logistic regression etc..)
3. Recommendation (matrix factorialization etc..)
4. Unsupervised Learning (clustering etc..)
5. All of the above

Agenda

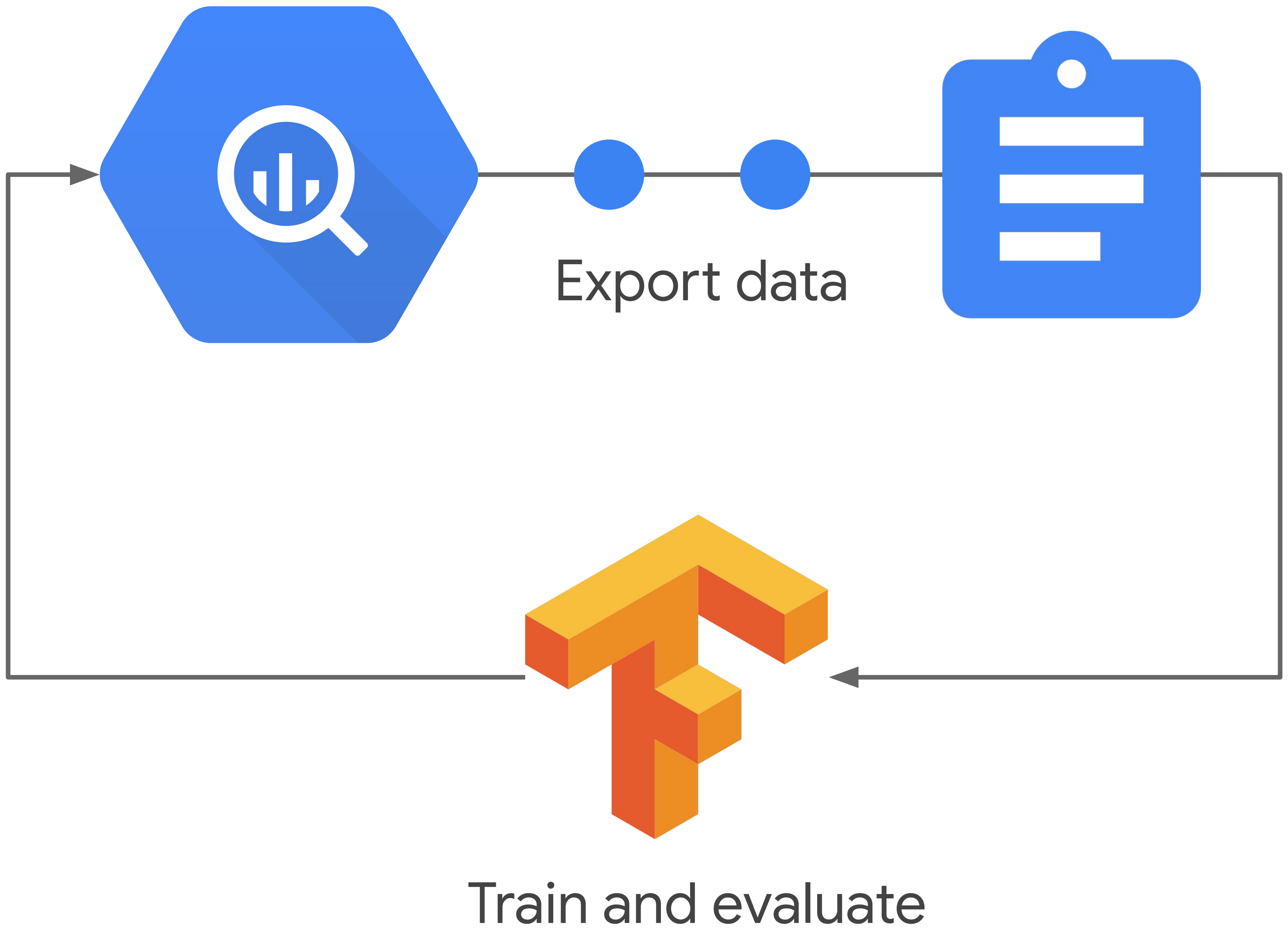
What is Forecasting

Terminology

Choosing the right model

Introducing BQML

It can take days to months to create an ML model



BQML is a way to easily build machine learning models

Perception APIs

Google-trained models for your app

App Developers

BQML

Easy to use, Familiar SQL

Data Analysts

Training & Prediction

End to End Platform for your data & models

Data Engineers / Data Scientists

Cloud ML Engine

Working with BigQuery ML

1 Dataset

2 Create/train

```
CREATE MODEL `bqml_tutorial.sample_model`
OPTIONS(model_type='logistic_reg') AS
SELECT
```

3 Evaluate

```
FROM
  ML.EVALUATE(MODEL
`bqml_tutorial.sample_model`,
    TABLE eval_table)
```

4 Predict/classify

```
FROM
    ML.PREDICT(MODEL
`bqml_tutorial.sample_model`,
      table game_to_predict) )
AS predict
```

BigQuery ML

- Write Machine Learning models with SQL
- Experiment and iterate right where your data lives -- in BigQuery
- Build classification (binary and multi-class) and forecasting models
- Know ML? Inspect model weights and adjust hyperparameters too

Lab

Forecasting Stock Prices using Regression in BQML

Lab Objectives

How to import data into Big Query

How to add features

How to build a model on BQ

How to evaluate results

Demo

Creating a Machine Learning Model with SQL

Predicting stock price

BigQuery ML Cheatsheet

- **Label** = alias a column as ‘label’ or specify column in OPTIONS using input_label_cols

- **Feature** = passed through to the model as part of your SQL SELECT statement
```
SELECT * FROM ML.FEATURE_INFO(MODEL `mydataset.mymodel`)
```

- **Model** = an object created in BigQuery that resides in your BigQuery dataset

- **Model Types** = Linear Regression, Logistic Regression (more coming)
```
CREATE OR REPLACE MODEL <dataset>.<name>
OPTIONS(model_type='<type>') AS
<training dataset>
```

- **Training Progress** = SELECT * FROM ML.TRAINING_INFO(MODEL `mydataset.mymodel`)

- **Inspect Weights** = SELECT * FROM ML.WEIGHTS(MODEL `mydataset.mymodel`, (<query>))

- **Evaluation** = SELECT * FROM ML.EVALUATE(MODEL `mydataset.mymodel`)

- **Prediction** = SELECT * FROM ML.PREDICT(MODEL `mydataset.mymodel`, (<query>))

Lab: Predict Stock prices with BigQuery ML

What is the price of AAPL likely to be in a few days?

BigQuery

Lab

Predicting Stock Prices with BigQuery ML

- Create a ML training dataset
- Select a model to train
- Train, evaluate, and predict
- Improve ML model performance