



Big Query public dataset

Census Adult Income





Schema


[Link to schema on gdrive](#)


age	INTEGER	NULLABLE	Age
workclass	STRING	NULLABLE	Nature of employment
functional_weight	INTEGER	NULLABLE	Sample weight of the individual from the original Census data. How likely they were to be included in this dataset, based on their demographic characteristics vs. whole-population estimates.
education	STRING	NULLABLE	Level of education completed
education_num	INTEGER	NULLABLE	Estimated years of education completed based on the value of the education field.
marital_status	STRING	NULLABLE	Marital status
occupation	STRING	NULLABLE	Occupation category
relationship	STRING	NULLABLE	Relationship to the household
race	STRING	NULLABLE	Race
sex	STRING	NULLABLE	Gender
capital_gain	INTEGER	NULLABLE	Amount of capital gains
capital_loss	INTEGER	NULLABLE	Amount of capital loss
hours_per_week	INTEGER	NULLABLE	Hours worked per week
native_country	STRING	NULLABLE	Country of birth
income_bracket	STRING	NULLABLE	Either ">50K" or "<=50K" based on income.

Query the dataset and save to csv

 RUN


 SAVE ▾

 SCHEDULE ▾

 MORE ▾

```
1  SELECT * FROM `census-309201.census_destination_dataset.census_adult_income`;
```

Query results

 SAVE RESULTS

Save to csv on gdrive



Explore the dataset in Data Studio

[Link to Data Studio](#)



Problem statement

Given the attributes, predict the binary income bracket of the person



BigQueryML

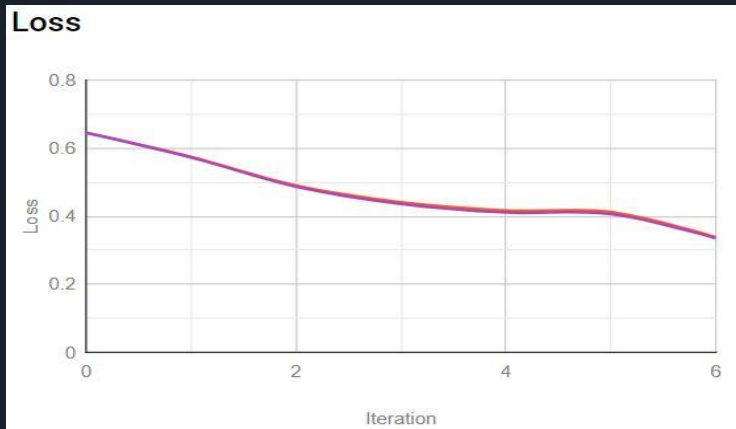
[Link to course slides](#)

[Link to GCP tutorial](#)

```
CREATE OR REPLACE MODEL
  `census_destination_dataset.census_model`
OPTIONS
  ( model_type='LOGISTIC_REG',
    auto_class_weights=TRUE,
    input_label_cols=['income_bracket']) AS
SELECT
  *
FROM
  `census_destination_dataset.census_adult_income`
```

New model named census-309201:census_destination_dataset.census_model

Trained model evaluation



True label	Predicted label	
	>50K	≤50K
>50K	100%	0%
≤50K	100%	0%



Predict using the model

```
#standardsql
SELECT * FROM ML.PREDICT(MODEL census_destination_dataset.census_model,
(
SELECT
*
FROM
`census_destination_dataset.census_adult_income`
))
```




Evaluating the model

```
SELECT * FROM ML.EVALUATE(MODEL census_destination_dataset.census_model)
```

Job information

Results

JSON

Execution details

Row	precision	recall	accuracy	f1_score	log_loss	roc_auc
1	0.6705521472392638	0.6930881420418517	0.8472014366956001	0.6816339257873402	0.3352706343846717	0.9022707292707293



BigQuery ML for benchmarking

BigQuery and BigQuery ML can be used for exploratory data analysis and creating a benchmark model