Big Query Predictive Models Creation

The task is to create predictive models using different Data Science platforms and compare the results of prediction choosing the best model for this particular task. I used the dataset from Kaggle.com https://www.kaggle.com/tejashvi14/employee-future-prediction
The task is to predict if an employee will leave the company or will keep working in the same company in the next 2 years, the model will be created based on the dataset from Kaggle. The task is a Binary Classification task.

We have the following data for the model creation:

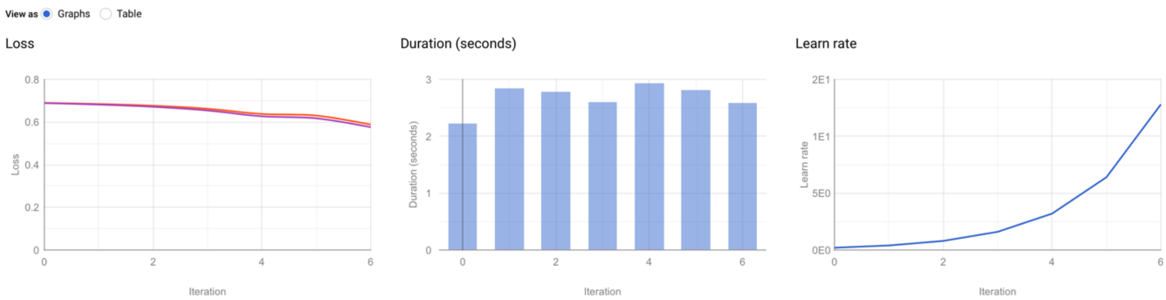
1. Education
2. JoiningYear
3. City
4. PaymentTier
5. Age
6. Gender
7. EverBenched
8. ExperienceInCurrentDomain
9. LeaveOrNot – target field

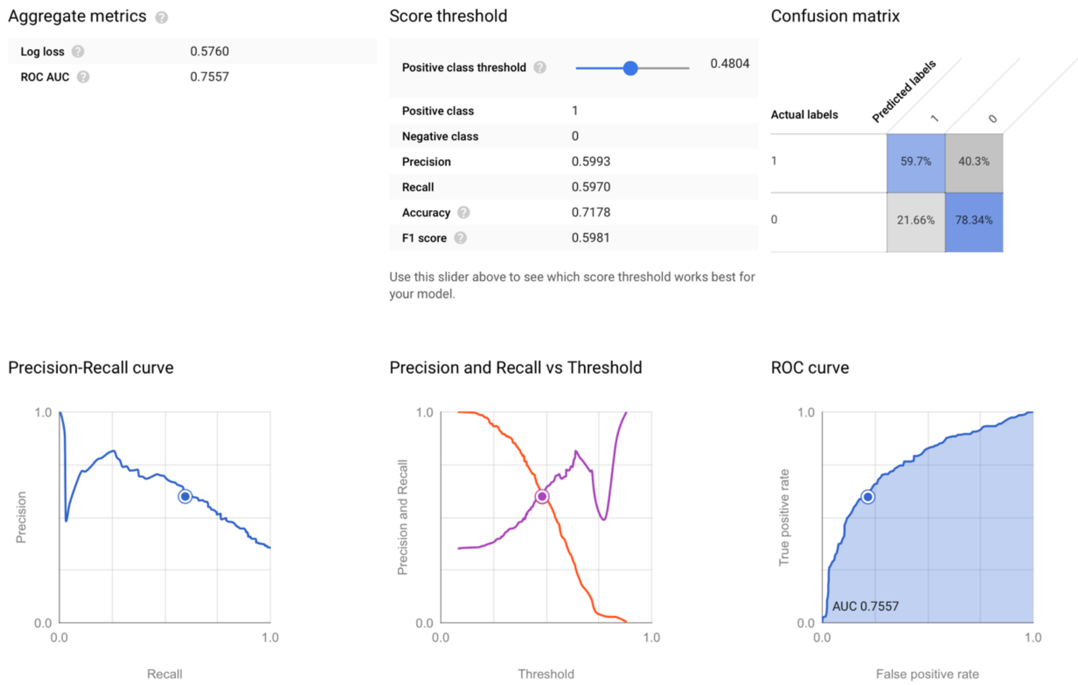
Step 1. Creating Linear Regression model using BiqQuery without EDA – LR_Simple1

```sql
CREATE OR REPLACE MODEL `ML.Employee_LR_simple1`
OPTIONS(MODEL_TYPE = 'LOGISTIC_REG',
    INPUT_LABEL_COLS = ['LeaveOrNot'],
    MAX_ITERATIONS = 30,
    LEARN_RATE_STRATEGY = 'LINE_SEARCH',
    EARLY_STOP = TRUE,
    MIN_REL_PROGRESS = 0.01,
    DATA_SPLIT_METHOD = 'AUTO_SPLIT',
    AUTO_CLASS_WEIGHTS = TRUE,
    ENABLE_GLOBAL_EXPLAIN = TRUE,
    CATEGORY_ENCODING_METHOD = 'DUMMY_ENCODING'
) as

SELECT * FROM `my-project-dec9.ML.Train_Employee`
```

Step 2. Evaluating the LR Simple1 model using quality





Quality Metrics	Ideal Result	LR Simple1 Model's Result
Log Loss (Logarithmic Loss) Log-loss is indicative of how close the prediction probability is to the corresponding actual/true value (0 or 1 in case of binary classification). The more the predicted probability diverges from the actual value, the higher is the log-loss value.	The more the predicted probability diverges from the actual value, the higher is the log-loss value. **Aiming for 0.**	**0.5760**
ROC AUC (area under the curve of the receiver operating characteristic) AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The	The higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. **Aiming to 1.**	0.7557

true-positive rate is also known as sensitivity, recall or probability of detection. The false-positive rate is also known as probability of false alarm and can be calculated as (1 − specificity). It can also be thought of as a plot of the power as a function of the Type I Error of the decision rule (when the performance is calculated from just a sample of the population, it can be thought of as estimators of these quantities). The ROC curve is thus the sensitivity or recall as a function of fall-out. In general, if the probability distributions for both detection and false alarm are known, the ROC curve can be generated by plotting the cumulative distribution function (area under the probability distribution from $-\infty$ to the discrimination threshold) of the detection probability in the y-axis versus the cumulative distribution function of the false-alarm probability on the x-axis.		
Model is not good enough yet, it needs to be fine-tuned.		

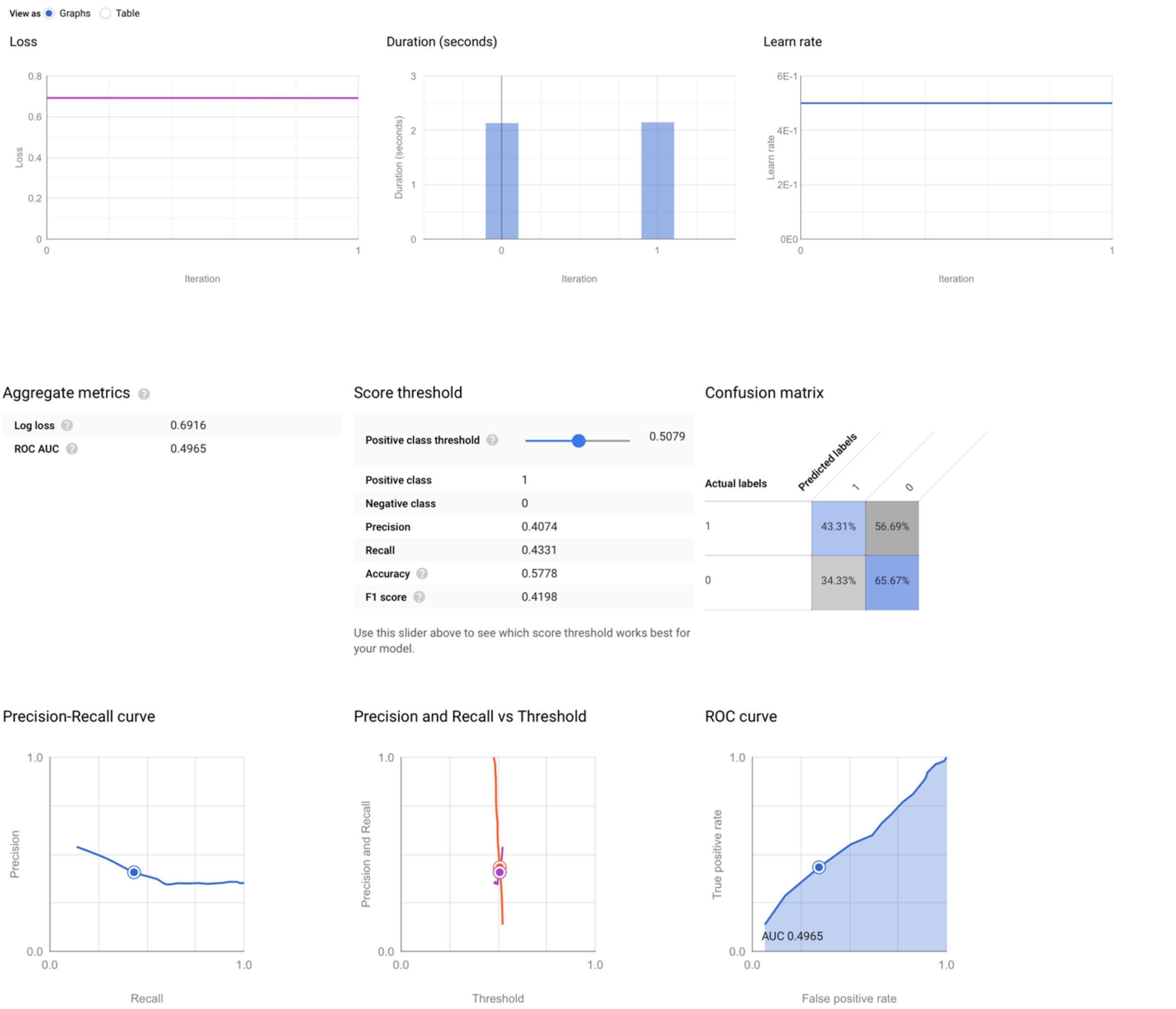
Step 3. Creating and training the Linear Regression Model using BigQuery with EDA - LR_with_features

```sql
CREATE OR REPLACE MODEL `ML.LR_with_features`
TRANSFORM(
CASE WHEN Gender = 'male' THEN 1 ELSE 0 END AS Sex,
CASE WHEN Age BETWEEN 0 AND 21 THEN 1 ELSE 0 END AS Age_VeryYoung,
CASE WHEN Age BETWEEN 21 AND 25 THEN 1 ELSE 0 END AS Age_Young,
CASE WHEN Age BETWEEN 25 AND 28 THEN 1 ELSE 0 END AS Age_Young1,
CASE WHEN Age BETWEEN 28 AND 31 THEN 1 ELSE 0 END AS Age_Young2,
CASE WHEN Age BETWEEN 31 AND 36 THEN 1 ELSE 0 END AS Age_Middle,
CASE WHEN Age > 36 THEN 1 ELSE 0 END AS Age_Adult,
CASE WHEN ExperienceInCurrentDomain < 2 THEN 1 ELSE 0 END As NoExperience,
CASE WHEN ExperienceInCurrentDomain BETWEEN 2 AND 4 THEN 1 ELSE 0 END AS LittleExperience,
CASE WHEN ExperienceInCurrentDomain > 4 THEN 1 ELSE 0 END As PlentyExperience,
* EXCEPT( Gender, Age, ExperienceInCurrentDomain))
OPTIONS
    (MODEL_TYPE = 'LOGISTIC_REG',
    L1_REG = 0.2,
    L2_REG = 0.6,
    LEARN_RATE_STRATEGY = 'CONSTANT',
    LEARN_RATE = 0.5,
    EARLY_STOP = TRUE,
    AUTO_CLASS_WEIGHTS = TRUE,
    ENABLE_GLOBAL_EXPLAIN = TRUE,
    CATEGORY_ENCODING_METHOD = 'ONE_HOT_ENCODING'
) AS
SELECT
  LeaveOrNot AS label, Gender, Age, ExperienceInCurrentDomain
FROM
 `my-project-dec9.ML.Train_Employee`
```

Run | Save query | Save view | Schedule query | More

This query will process 114.3 KB (ML) when run.

Step 4. Evaluating the LR_with_features model



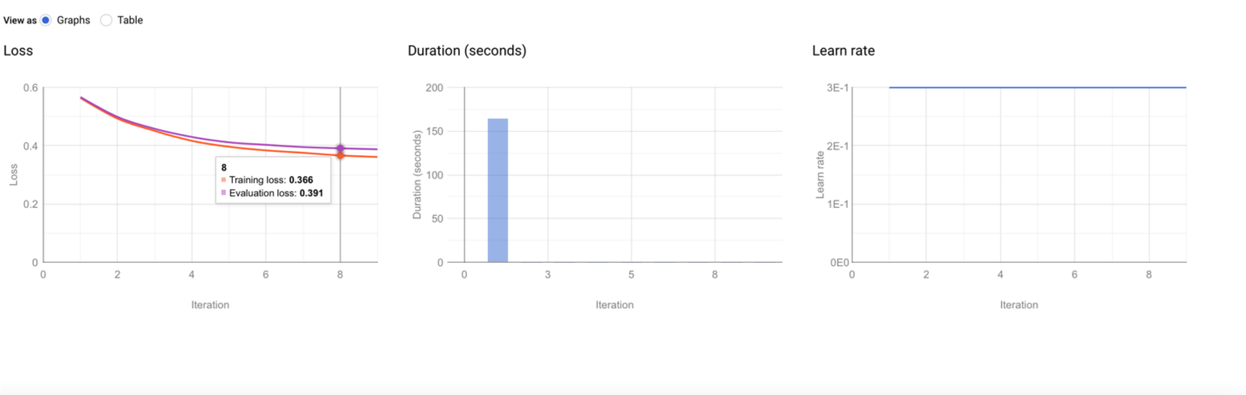
Quality Metrics	Ideal Result	LR with features Model's Result
Log Loss (Logarithmic Loss) Log-loss is indicative of how close the prediction probability is to the corresponding actual/true value (0 or 1 in case of binary classification). The more the predicted probability diverges from the actual value, the higher is the log-loss value.	The more the predicted probability diverges from the actual value, the higher is the log-loss value. **Aiming for 0.**	0.6916
ROC AUC (area under the curve of the receiver operating characteristic) AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It	The higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. **Aiming to 1.**	0.4965

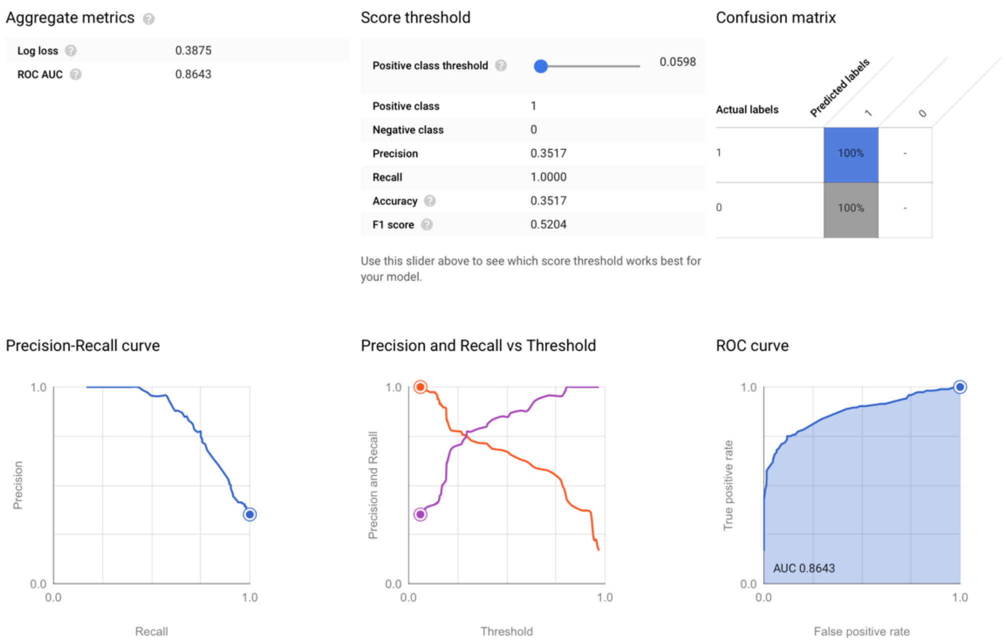
tells how much the model is capable of distinguishing between classes.		
Metrics are worse, that means that the algorithm we selected is not optimal, we need to try different algorithms.		

Step 5 Creating a new model using a different algorithm - Boosted Tree Classifier – XGB.

```
CREATE OR REPLACE MODEL `ML.Employee_XGB`
OPTIONS(MODEL_TYPE = 'BOOSTED_TREE_CLASSIFIER' ,
         BOOSTER_TYPE = 'GBTREE',
         TREE_METHOD = 'HIST',
         MAX_TREE_DEPTH = 5,
         L2_REG = 0.4,
         EARLY_STOP = TRUE,
         INPUT_LABEL_COLS = ['LeaveOrNot'],
         MAX_ITERATIONS = 50,
         ENABLE_GLOBAL_EXPLAIN = TRUE
) as
SELECT *
FROM `my-project-dec9.ML.Train_Employee`
```

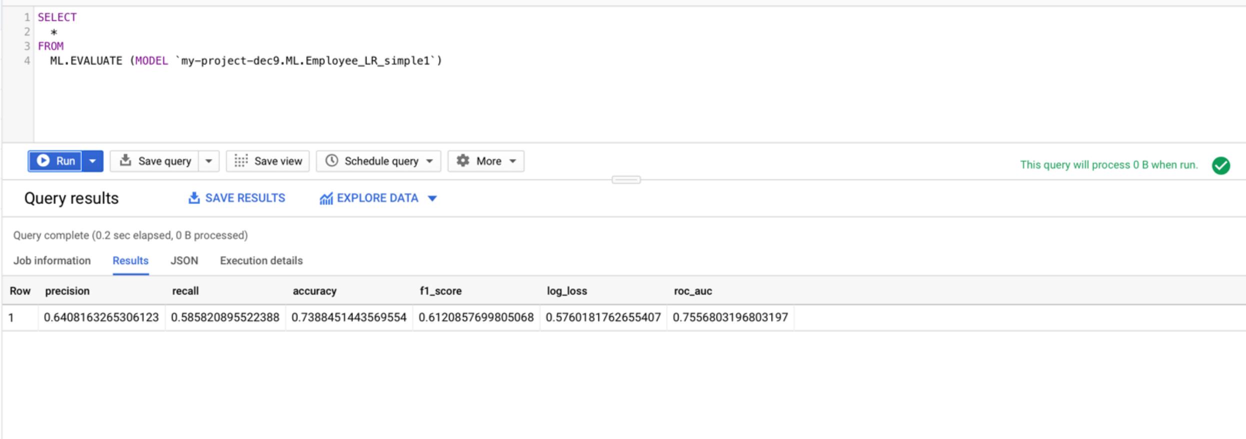
Step 6. Evaluating the model we created – XGB.





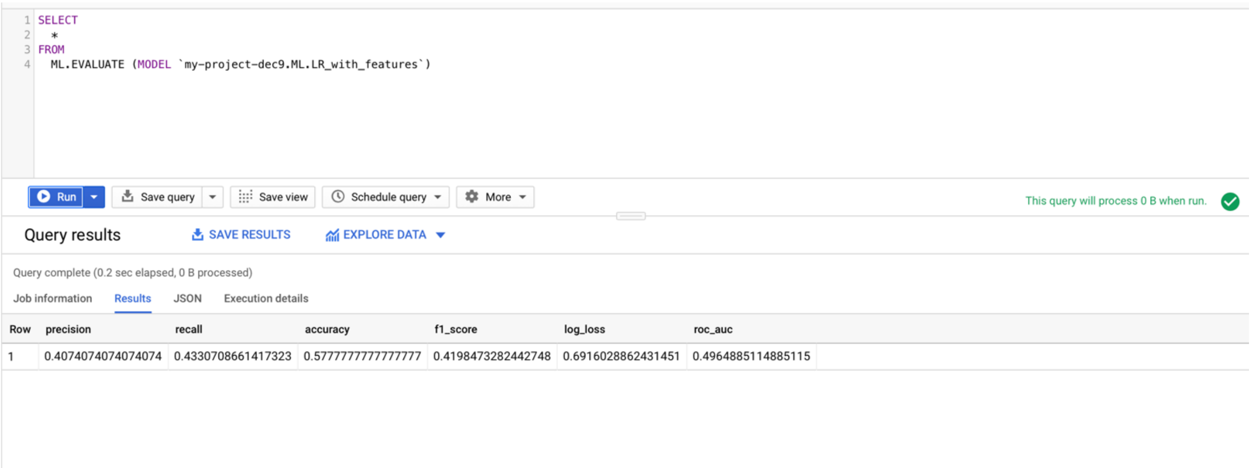
Quality Metrics	LR simple1	LR with features	XGB
Log Loss (Logarithmic Loss) Log-loss is indicative of how close the prediction probability is to the corresponding actual/true value (0 or 1 in case of binary classification). The more the predicted probability diverges from the actual value, the higher is the log-loss value.	0.5760	0.6916	0.3875
ROC AUC (area under the curve of the receiver operating characteristic) AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.	0.7557	0.4965	0.8643
XGB model is demonstrating better results comparing to the Linear Regression models.			

Step 7. Comparing the models (LR simple1, LR with features и XGB) using the ML.EVALUATE function.



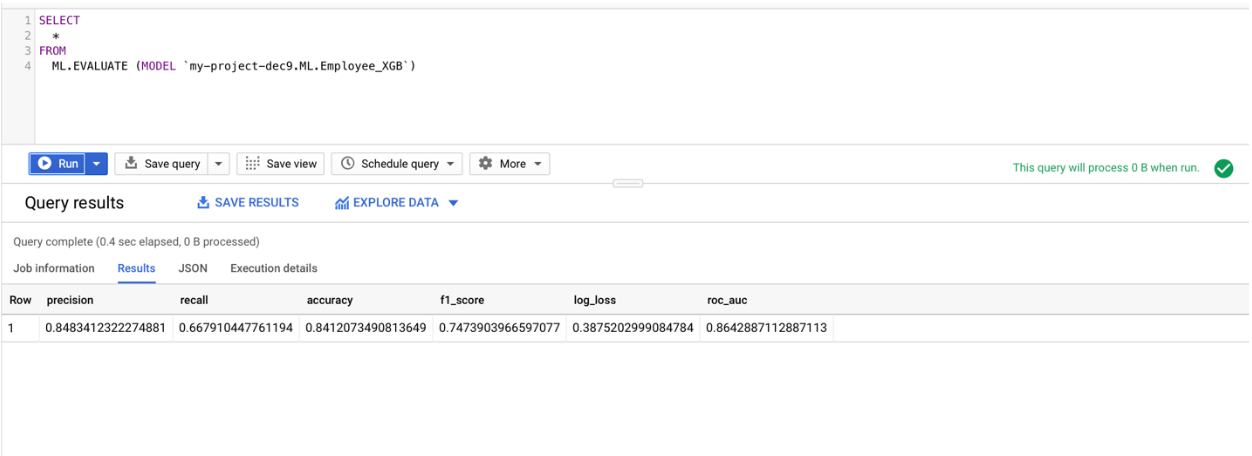
```
SELECT
  *
FROM
  ML.EVALUATE (MODEL `my-project-dec9.ML.Employee_LR_simple1`)
```


Row	precision	recall	accuracy	f1_score	log_loss	roc_auc
1	0.6408163265306123	0.585820895522388	0.7388451443569554	0.6120857699805068	0.5760181762655407	0.7556803196803197



```
SELECT
  *
FROM
  ML.EVALUATE (MODEL `my-project-dec9.ML.LR_with_features`)
```


Row	precision	recall	accuracy	f1_score	log_loss	roc_auc
1	0.4074074074074074	0.4330708661417323	0.5777777777777777	0.4198473282442748	0.6916028862431451	0.4964885114885115



```
SELECT
  *
FROM
  ML.EVALUATE (MODEL `my-project-dec9.ML.Employee_XGB`)
```


Row	precision	recall	accuracy	f1_score	log_loss	roc_auc
1	0.8483412322274881	0.667910447761194	0.8412073490813649	0.7473903966597077	0.3875202999084784	0.8642887112887113

Step 8. Predicting the results using LR simple1, LR with features, XGB models, with the help of ML.PREDICT function.

```sql
SELECT
  *
FROM
  ML.EXPLAIN_PREDICT (MODEL `my-project-dec9.ML.Employee_LR_simple1`,
    (
    SELECT
      *
    FROM
      `my-project-dec9.ML.Test_Employee`
    )
  )
```

Run | Save query | Save view | Schedule query | More

This query will process 56.8 KB when run.

Query results   SAVE RESULTS   EXPLORE DATA

Query complete (0.5 sec elapsed, 56.8 KB processed)

Job information   Results   JSON   Execution details

OrNot	probability	top_feature_attributions.feature	top_feature_attributions.attribution	baseline_prediction_value	prediction_value	approximation_error	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	ExperienceInCurrentDomain	LeaveOrNot
0	0.7617088284513307	JoiningYear	-0.5767646069328181	-0.6820585162640469	-1.162071031985088	0.0	Bachelors	2012	Bangalore	3	22	Male	false	0	0
		Age	0.15315570593429265												
		PaymentTier	-0.0854436288919473												
		ExperienceInCurrentDomain	0.029040014169431628												
		City	0.0												
0	0.7759677869645267	City	-0.8164999823385779	-0.6820585162640469	-1.2423211579682256	0.0	Bachelors	2012	New Delhi	3	22	Female	false	0	0
		Gender	0.7362498563554404												
		JoiningYear	-0.5767646069328181												
		Age	0.15315570593429265												
		PaymentTier	-0.0854436288919473												
0	0.7759677869645267	City	-0.8164999823385779	-0.6820585162640469	-1.2423211579682256	0.0	Bachelors	2012	New Delhi	3	22	Female	false	0	0
		Gender	0.7362498563554404												
		JoiningYear	-0.5767646069328181												
		Age	0.15315570593429265												
		PaymentTier	-0.0854436288919473												

Rows per page: 100   1 - 100 of 852   First page   Last page

```sql
SELECT
  *
FROM
  ML.EXPLAIN_PREDICT (MODEL `my-project-dec9.ML.LR_with_features`,
    (
    SELECT
      *
    FROM
      `my-project-dec9.ML.Test_Employee`
    )
  )
```

Run | Save query | Save view | Schedule query | More

This query will process 56.6 KB when run.

Query results   SAVE RESULTS   EXPLORE DATA

Query complete (0.4 sec elapsed, 56.6 KB processed)

Job information   Results   JSON   Execution details

label	probability	top_feature_attributions.feature	top_feature_attributions.attribution	baseline_prediction_value	prediction_value	approximation_error	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	ExperienceInCurrentDomain	LeaveOrNot
1	0.5098288071161524	Age_Young	0.04734094356555193	-7.172319983207174E-5	0.03932029372613963	0.0	Bachelors	2012	Bangalore	3	22	Male	false	0	0
		Age_Young2	0.012663706725756799												
		LittleExperience	-0.012143226715545234												
		NoExperience	-0.010918303214501978												
		Age_Young1	-0.010529217194136964												
1	0.5098288071161524	Age_Young	0.04734094356555193	-7.172319983207174E-5	0.03932029372613963	0.0	Bachelors	2012	New Delhi	3	22	Female	false	0	0
		Age_Young2	0.012663706725756799												
		LittleExperience	-0.012143226715545234												
		NoExperience	-0.010918303214501978												
		Age_Young1	-0.010529217194136964												
1	0.5098288071161524	Age_Young	0.04734094356555193	-7.172319983207174E-5	0.03932029372613963	0.0	Bachelors	2012	New Delhi	3	22	Female	false	0	0
		Age_Young2	0.012663706725756799												
		LittleExperience	-0.012143226715545234												
		NoExperience	-0.010918303214501978												
		Age_Young1	-0.010529217194136964												

```sql
SELECT
  *
FROM
  ML.EXPLAIN_PREDICT (MODEL `my-project-dec9.ML.Employee_XGB`,
    (
    SELECT
      *
    FROM
      `my-project-dec9.ML.Test_Employee`
    )
  )
```

Run | Save query | Save view | Schedule query | More

This query will process 81.8 KB when run.

Query results   SAVE RESULTS   EXPLORE DATA

Query complete (2.3 sec elapsed, 81.8 KB processed)

Job information   Results   JSON   Execution details

veOrNot	probability	top_feature_attributions.feature	top_feature_attributions.attribution	baseline_prediction_value	prediction_value	approximation_error	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	ExperienceInCurrentDomain	LeaveOrNot
0	0.8093995749950409	PaymentTier	0.20224064588546753	0.8467823266983032	1.223056711256504	0.0	Bachelors	2012	Bangalore	3	22	Male	false	0	0
		City	-0.12933996319770813												
		Education	0.12186403572559357												
		JoiningYear	0.11048183590173721												
		Gender	0.10529474169015884												
0	0.828901007771492	City	0.2866467535495758	0.8467823266983032	1.2889291681349277	0.0	Bachelors	2012	New Delhi	3	22	Female	false	0	0
		Gender	-0.21967868506908417												
		PaymentTier	0.20224064588546753												
		Education	0.12963716685771942												
		JoiningYear	0.09990525990724564												
0	0.828901007771492	City	0.2866467535495758	0.8467823266983032	1.2889291681349277	0.0	Bachelors	2012	New Delhi	3	22	Female	false	0	0
		Gender	-0.21967868506908417												
		PaymentTier	0.20224064588546753												
		Education	0.12963716685771942												
		JoiningYear	0.09990525990724564												

Step 9. XGB model and ML.FEATURE_IMPORTANCE function to see the importance of each field for the created model.

```
1 SELECT * FROM ML.FEATURE_IMPORTANCE (MODEL `my-project-dec9.ML.Employee_XGB`)
```

Run | Save query | Save view | Schedule query | More — This query will process 0 B when run.

Query results — SAVE RESULTS — EXPLORE DATA

Query complete (0.6 sec elapsed, 0 B processed)

Job information | Results | JSON | Execution details

Row	feature	importance_weight	importance_gain	importance_cover
1	Education	64	7.955924218750001	318.4442562499999
2	JoiningYear	42	31.35683533333333	631.0134000000002
3	City	50	10.45532284	250.85576800000007
4	PaymentTier	48	13.505746162499994	434.2616554166666
5	Age	28	2.9830879999999986	102.73426285714284
6	Gender	32	4.452454093750001	364.91915875
7	EverBenched	16	2.3583788750000005	89.7641
8	ExperienceInCurrentDomain	22	3.252678545454546	174.85523636363635

The most important features that have an impact of the employee’s decision to quit are the following features:

- **Education;**
- **Joining Year;**
- **City;**
- **Payment Tier;**
- **Age.**