

```
! python --version
```

```
Python 3.11.13
```

▼ Connect to Big Query

```
# libraraies
from google.cloud import bigquery
from google.colab import auth

# authenticate
auth.authenticate_user()

#initialize the client
project_id = 'cloud-project-478713'
client = bigquery.Client(project=project_id, location= 'US')
```

```
from google.cloud.bigquery import dataset
# import dataset
dataset_ref = client.dataset('employee_data', project=project_id)
dataset = client.get_dataset(dataset_ref)
table_ref = dataset.table('tbl_hr_data')
table = client.get_table(table_ref)
table.schema
```

```
[SchemaField('satisfaction_level', 'FLOAT', 'NULLABLE', None, None, (), None),
 SchemaField('last_evaluation', 'FLOAT', 'NULLABLE', None, None, (), None),
 SchemaField('number_project', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('average_monthly_hours', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('time_spend_company', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('Work_accident', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('Quit_the_Company', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('promotion_last_5years', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('Departments', 'STRING', 'NULLABLE', None, None, (), None),
 SchemaField('salary', 'STRING', 'NULLABLE', None, None, (), None),
 SchemaField('employee_id', 'STRING', 'NULLABLE', None, None, (), None)]
```

```
new_table_ref = dataset.table('tbl_new_employees')
new_table = client.get_table(new_table_ref)
new_table.schema
```

```
[SchemaField('satisfaction_level', 'FLOAT', 'NULLABLE', None, None, (), None),
 SchemaField('last_evaluation', 'FLOAT', 'NULLABLE', None, None, (), None),
 SchemaField('number_project', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('average_monthly_hours', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('time_spend_company', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('Work_accident', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('Quit_the_Company', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('promotion_last_5years', 'INTEGER', 'NULLABLE', None, None, (), None),
 SchemaField('Departments', 'STRING', 'NULLABLE', None, None, (), None),
 SchemaField('salary', 'STRING', 'NULLABLE', None, None, (), None),
 SchemaField('employee_id', 'STRING', 'NULLABLE', None, None, (), None)]
```

```
# convert to new data frames
df = client.list_rows(table=table).to_dataframe()
df.head()
```

	satisfaction_level	last_evaluation	number_project	average_monthly_hours	time_spend_company	Work_accident	Quit_the_Co
0	0.38	0.53	2	157	3	0	
1	0.80	0.86	5	262	6	0	
2	0.11	0.88	7	272	4	0	
3	0.72	0.87	5	223	5	0	
4	0.37	0.52	2	159	3	0	

```
# convert to data frames
df2 = client.list_rows(table=new_table).to_dataframe()
df2.head()
```

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	Quit_the_Co
0	0.537849	0.122914	2	208	2	0	
1	0.056211	0.322600	2	229	5	1	
2	0.555186	0.555949	2	187	3	0	
3	0.605273	0.713086	2	218	3	0	
4	0.043437	0.162372	2	175	3	0	

Build Model

Install packages

↳ 2 cells hidden

Training Model

```
# get the model
from pycaret.classification import *
```

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15004 entries, 0 to 15003
Data columns (total 11 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   satisfaction_level                    15004 non-null  float64
1   last_evaluation                      15004 non-null  float64
2   number_project                      14999 non-null  Int64
3   average_montly_hours                15004 non-null  Int64
4   time_spend_company                  14999 non-null  Int64
5   Work_accident                       15000 non-null  Int64
6   Quit_the_Company                    15004 non-null  Int64
7   promotion_last_5years               15004 non-null  Int64
8   Departments                         15004 non-null  object
9   salary                              15004 non-null  object
10  employee_id                         15004 non-null  object
dtypes: Int64(6), float64(2), object(3)
memory usage: 1.3+ MB
```

```
# setup or model
setup(df, target='Quit_the_Company', session_id=123,
      ignore_features=['employee_id'],
      categorical_features=['salary', 'Departments'])
```

	Description	Value
0	Session id	123
1	Target	Quit_the_Company
2	Target type	Binary
3	Original data shape	(15004, 11)
4	Transformed data shape	(15004, 21)
5	Transformed train set shape	(10502, 21)
6	Transformed test set shape	(4502, 21)
7	Ignore features	1
8	Numeric features	7
9	Categorical features	2
10	Rows with missing values	0.0%
11	Preprocess	True
12	Imputation type	simple
13	Numeric imputation	mean
14	Categorical imputation	mode
15	Maximum one-hot encoding	25
16	Encoding method	None
17	Fold Generator	StratifiedKFold
18	Fold Number	10
19	CPU Jobs	-1
20	Use GPU	False
21	Log Experiment	False
22	Experiment Name	clf-default-name
23	USI	ac9b

<pycaret.classification.oop.ClassificationExperiment at 0x7811207ffed0>

compare_models()

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.9881	0.9910	0.9585	0.9913	0.9746	0.9668	0.9671	0.9030
lightgbm	Light Gradient Boosting Machine	0.9853	0.9932	0.9505	0.9876	0.9686	0.9591	0.9594	0.7330
xgboost	Extreme Gradient Boosting	0.9852	0.9921	0.9581	0.9797	0.9687	0.9590	0.9592	0.2780
et	Extra Trees Classifier	0.9840	0.9908	0.9505	0.9820	0.9658	0.9554	0.9557	0.9050
gbc	Gradient Boosting Classifier	0.9765	0.9891	0.9313	0.9689	0.9496	0.9343	0.9347	1.1740
dt	Decision Tree Classifier	0.9747	0.9698	0.9605	0.9354	0.9476	0.9310	0.9312	0.2220
ada	Ada Boost Classifier	0.9584	0.9830	0.9085	0.9167	0.9123	0.8851	0.8853	0.4150
knn	K Neighbors Classifier	0.9343	0.9687	0.9205	0.8246	0.8698	0.8260	0.8284	0.2460
qda	Quadratic Discriminant Analysis	0.8749	0.9154	0.8086	0.7169	0.7566	0.6734	0.6784	0.1140
lr	Logistic Regression	0.7932	0.8178	0.3584	0.6129	0.4516	0.3351	0.3536	1.4400
lda	Linear Discriminant Analysis	0.7838	0.8141	0.3428	0.5787	0.4296	0.3071	0.3236	0.1210
ridge	Ridge Classifier	0.7742	0.8142	0.2405	0.5611	0.3361	0.2258	0.2554	0.1090
dummy	Dummy Classifier	0.7617	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1070
svm	SVM - Linear Kernel	0.6963	0.7774	0.2470	0.1084	0.1444	0.0625	0.0679	0.3240
nb	Naive Bayes	0.6826	0.8093	0.8034	0.4147	0.5470	0.3392	0.3832	0.1100

```

RandomForestClassifier
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                        criterion='gini', max_depth=None, max_features='sqrt',
                        max_leaf_nodes=None, max_samples=None,
                        min_impurity_decrease=0.0, min_samples_leaf=1,
                        min_samples_split=2, min_weight_fraction_leaf=0.0,
                        monotonic_cst=None, n_estimators=100, n_jobs=-1,
                        oob_score=False, random_state=123, verbose=0,
                        warm_start=False)

```

```
rf_model = create_model('rf')
```

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.9848	0.9860	0.9442	0.9916	0.9673	0.9574	0.9579
1	0.9819	0.9939	0.9442	0.9793	0.9615	0.9497	0.9499
2	0.9924	0.9959	0.9760	0.9919	0.9839	0.9789	0.9789
3	0.9876	0.9871	0.9600	0.9877	0.9736	0.9655	0.9657
4	0.9848	0.9926	0.9400	0.9958	0.9671	0.9572	0.9578
5	0.9924	0.9932	0.9680	1.0000	0.9837	0.9788	0.9790
6	0.9876	0.9897	0.9640	0.9837	0.9737	0.9656	0.9657
7	0.9924	0.9908	0.9720	0.9959	0.9838	0.9788	0.9789
8	0.9895	0.9891	0.9640	0.9918	0.9777	0.9708	0.9710
9	0.9876	0.9917	0.9522	0.9958	0.9735	0.9654	0.9659
Mean	0.9881	0.9910	0.9585	0.9913	0.9746	0.9668	0.9671
Std	0.0034	0.0029	0.0120	0.0059	0.0074	0.0096	0.0095

```
final_df = predict_model(rf_model)
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Random Forest Classifier	0.9904	0.9931	0.9674	0.9924	0.9797	0.9735	0.9736

```
final_df.head()
```

	satisfaction_level	last_evaluation	number_project	average_monthly_hours	time_spend_company	Work_accident	promotio
1679	0.43	0.55	2	159	3	0	
4665	0.63	0.93	3	236	4	0	
1076	0.09	0.79	6	276	4	0	
1253	0.85	1.00	4	234	5	0	
2570	0.80	0.96	3	257	5	0	

df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15004 entries, 0 to 15003
Data columns (total 11 columns):
#   Column                Non-Null Count  Dtype
---  -
0   satisfaction_level      15004 non-null  float64
1   last_evaluation        15004 non-null  float64
2   number_project         14999 non-null  Int64
3   average_monthly_hours  15004 non-null  Int64
4   time_spend_company     14999 non-null  Int64
5   Work_accident          15000 non-null  Int64
6   Quit_the_Company       15004 non-null  Int64
7   promotion_last_5years  15004 non-null  Int64
8   Departments            15004 non-null  object
9   salary                 15004 non-null  object
10  employee_id            15004 non-null  object
dtypes: Int64(6), float64(2), object(3)
memory usage: 1.3+ MB
```

new_predictions = predict_model(rf_model, data = df2)

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Random Forest Classifier	0.9300	0	0.0000	0.0000	0.0000	0.0000	0.0000

new_predictions.head()

	satisfaction_level	last_evaluation	number_project	average_monthly_hours	time_spend_company	Work_accident	promotion_1
0	0.537849	0.122914	2	208	2	0	
1	0.056211	0.322600	2	229	5	1	
2	0.555186	0.555949	2	187	3	0	
3	0.605273	0.713086	2	218	3	0	
4	0.043437	0.162372	2	175	3	0	

```
new_predictions.to_gbq('employeedata.pilot_predictions',
                        project_id,
                        chunksize=None,
                        if_exists='replace')
```

100%|██████████| 1/1 [00:00<00:00, 7345.54it/s]

plot_model(rf_model, plot='feature')

