Create a Scikit-learn Prediction Service on Cloud MLE

The code for creating a prediction service is shown in the following, and is saved in the file 'create-prediction-service.sh'.

```
export MODEL_VERSION=v1
export MODEL_NAME=iris_sklearn
export REGION=us-central1

# Create a Cloud ML Engine model
echo "Creating model..."
gcloud ml-engine models create $MODEL_NAME --regions=$REGION

# Create a model version
echo "Creating model version..."
gcloud ml-engine versions create $MODEL_VERSION \
    --model $MODEL_NAME \
    --config config.yaml
```

The preceding code references a configuration file 'config.yaml'. This file (as shown in the following) holds the configuration for the Scikit-learn model. Let's briefly go through the attributes listed:

- deploymentUri: This points to the bucket location of the Scikit-learn model.
- runtime version: This attribute specifies the Cloud MLE runtime version.
- framework: This attribute is of particular importance as it specifies the model framework in use; this can be SCIKIT_LEARN, XGBOOST, or TENSORFLOW. For this example, it is set to SCIKIT_LEARN.
- pythonVersion: This attribute specifies the Python version in use.

The 'config.yaml' is as defined in the following:

```
deploymentUri: "gs://iris-sklearn/iris_20181119_050517"
runtimeVersion: '1.8'
framework: "SCIKIT_LEARN"
pythonVersion: "3.5"
```

Run the following command to create a prediction service.

```
source ./scripts/create-prediction-service.sh
```

```
Creating model...
Created ml engine model [projects/quantum-ally-219323/models/iris_sklearn].
Creating model version...
Creating version (this might take a few minutes)......done.
```

Make Online Predictions from the Scikit-learn Model

The code to make an online prediction from the Scikit-learn model is shown in the following and is stored in the file 'online-prediction.sh'. In online predictions, the input data is passed directly as a JSON string.

```
export JOB_NAME=iris_sklearn_prediction
export MODEL_NAME=iris_sklearn
export MODEL_VERSION=v1
export TEST_FILE_GCS=gs://iris-sklearn/test-sample.json
export TEST_FILE=./test-sample.json

# download file
gsutil cp $TEST_FILE_GCS .

# submit an online job
gcloud ml-engine predict --model $MODEL_NAME \
        --version $MODEL_VERSION \
        --json-instances $TEST_FILE

echo "0 -> setosa, 1 -> versicolor, 2 -> virginica"
```