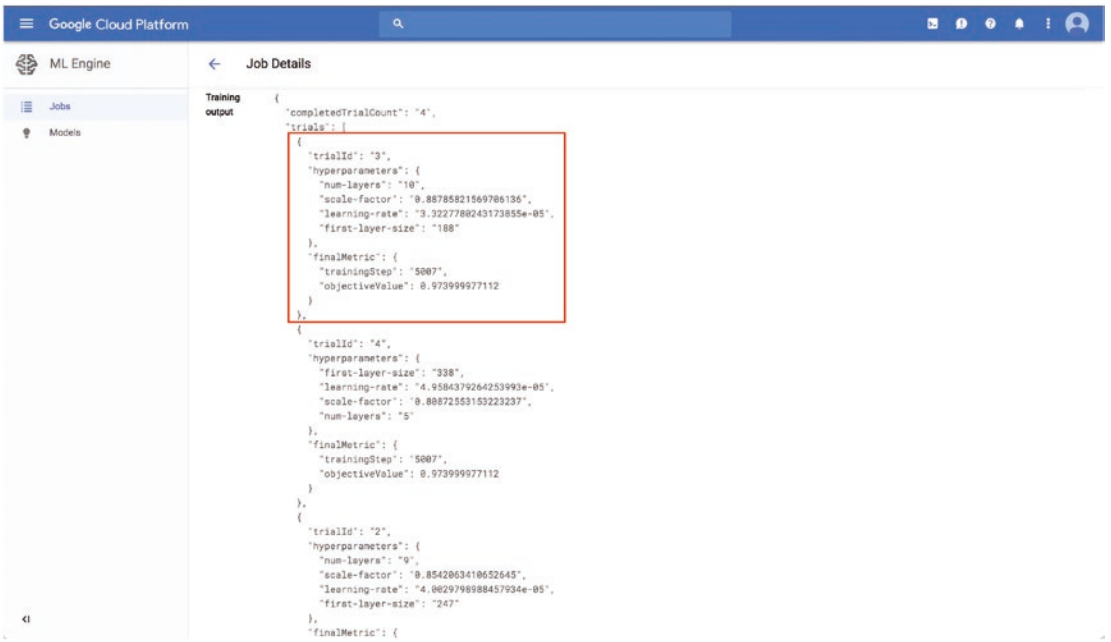

Figure 41-4. *Choosing the best hyper-parameter set*

Making Predictions on Cloud MLE

To make predictions on Cloud MLE, we first create a prediction instance. To do this, run the code in 'create-prediction-service.sh' as shown in the following. The variable 'MODEL_BINARIES' points to the folder location on GCS that stores the trained model for the hyper-parameter setting with '**trialID** = 2'.

```
export MODEL_VERSION=v1
export MODEL_NAME=iris
export MODEL_BINARIES=$GCS_JOB_DIR/3/export/iris/1542241126

# Create a Cloud ML Engine model
gcloud ai-platform models create $MODEL_NAME

# Create a model version
gcloud ai-platform versions create $MODEL_VERSION \
    --model $MODEL_NAME \
    --origin $MODEL_BINARIES \
    --runtime-version 1.8
```

Run the following code to create the prediction service.

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

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

The version details of the created model is as seen in Figure 41-5.

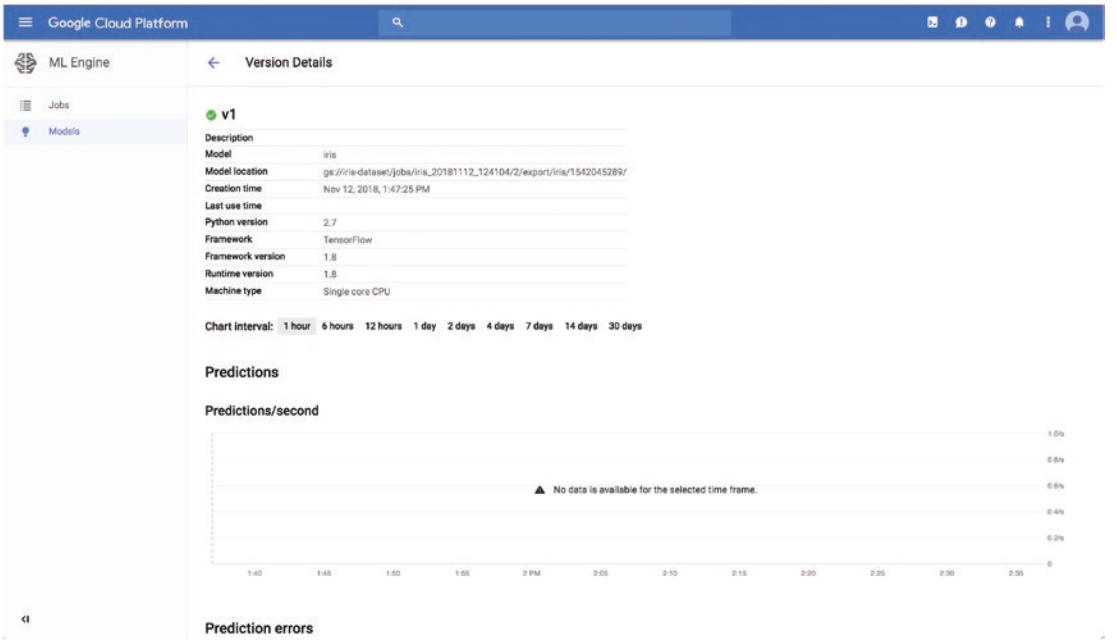

Figure 41-5. *Created model for serving on Cloud MLE*

Run Batch Prediction

Now let's run a batch prediction job on Cloud MLE. The code to execute a batch prediction call on Cloud MLE is provided in the following and stored in 'run-batch-predictions.sh.'

```
export JOB_NAME=iris_prediction
export MODEL_NAME=iris
```