AutoML Using H2O Sparkling Water

Speaker,
Parvez Khan

You Don’t Need A Phd in Data Science To Train a ML Model

Agenda

- Introduction
- Why AutoML
- Environment Setup
- Model Training Using Sparkling Water
- Auto Sklearn vs H2O3 vs Sparkling Water
- Export AutoML Model
- Summary

Introduction

AutoML - Automatic Machine Learning.

H2O Sparkling Water is a powerful open-source ML platform that combines the scalability of Apache Spark with the ease-of-use and flexibility of H2O.ai machine learning libraries.

Democratize machine learning and make it more accessible to individuals/organizations with limited expertise in data science.

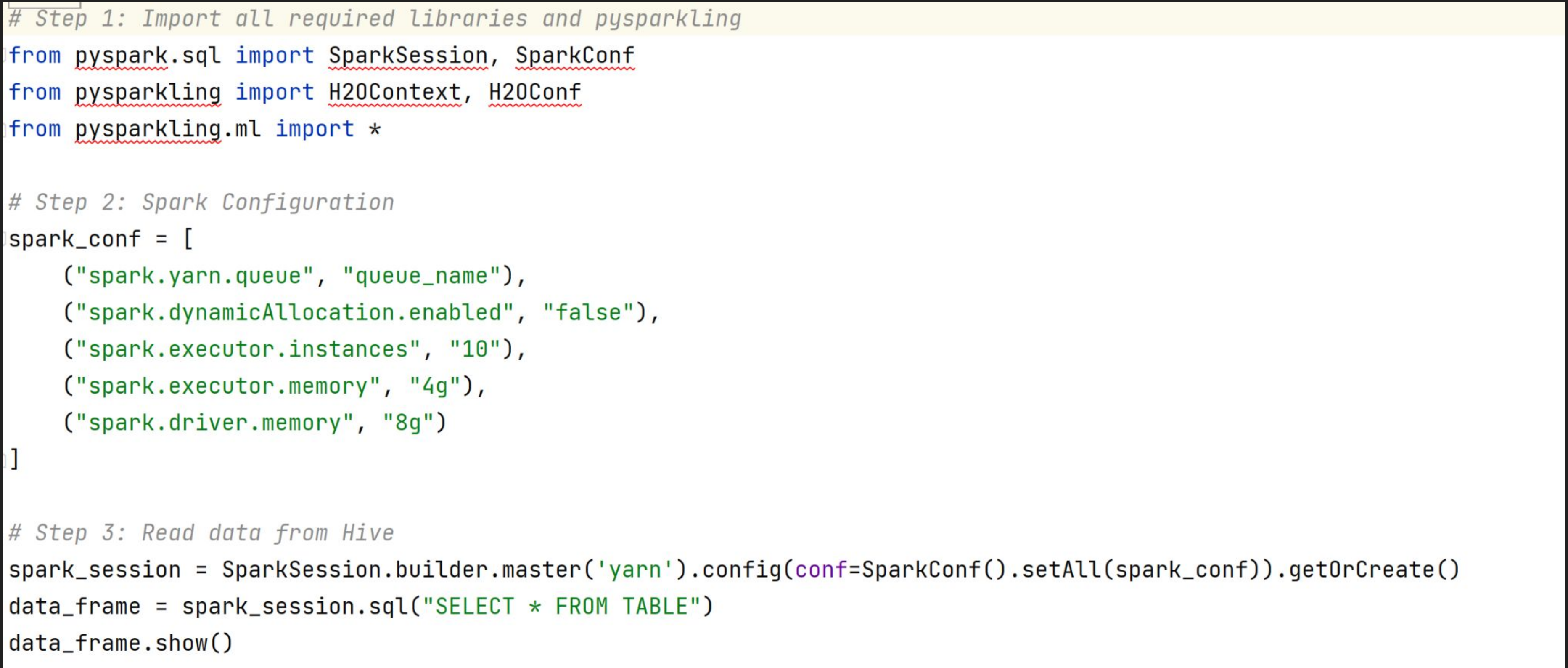
Scalable, Time efficient, Optimal Model Selection etc

Deploy the model cross origin platform.

Environment Setup

1. Python>3.5
2. JAVA
3. Pyspark 2.4
4. Install h2o sparkling water `pip install h2o-sparkling-2.4`
5. Jupyter Notebook or Any IDE

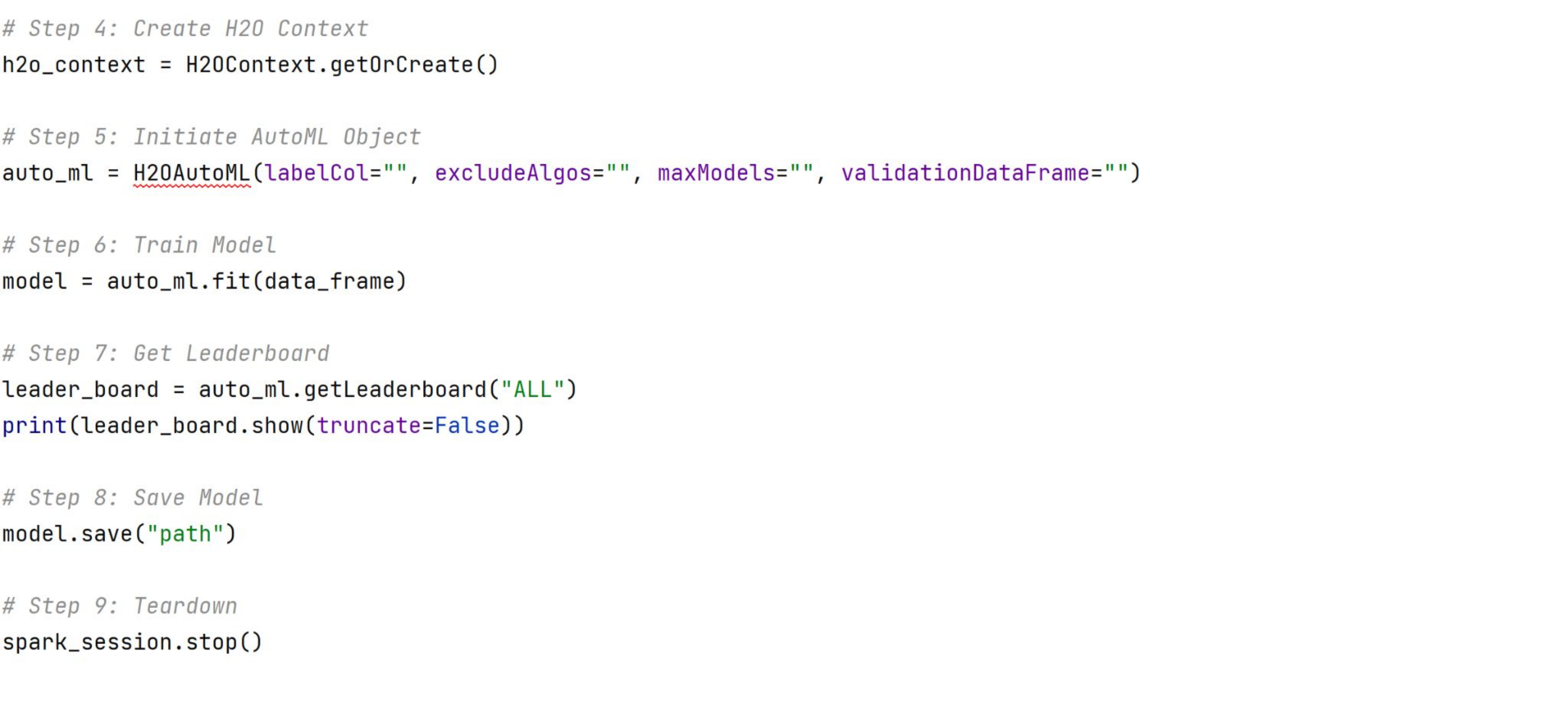
Model Training Using Sparkling Water



```
# Step 1: Import all required libraries and pysparkling
from pyspark.sql import SparkSession, SparkConf
from pysparkling import H2OContext, H2OConf
from pysparkling.ml import *

# Step 2: Spark Configuration
spark_conf = [
    ("spark.yarn.queue", "queue_name"),
    ("spark.dynamicAllocation.enabled", "false"),
    ("spark.executor.instances", "10"),
    ("spark.executor.memory", "4g"),
    ("spark.driver.memory", "8g")
]

# Step 3: Read data from Hive
spark_session = SparkSession.builder.master('yarn').config(conf=SparkConf().setAll(spark_conf)).getOrCreate()
data_frame = spark_session.sql("SELECT * FROM TABLE")
data_frame.show()
```



```
# Step 4: Create H2O Context
h2o_context = H2OContext.getOrCreate()

# Step 5: Initiate AutoML Object
auto_ml = H2OAutoML(labelCol="", excludeAlgos="", maxModels="", validationDataFrame="")

# Step 6: Train Model
model = auto_ml.fit(data_frame)

# Step 7: Get Leaderboard
leader_board = auto_ml.getLeaderboard("ALL")
print(leader_board.show(truncate=False))

# Step 8: Save Model
model.save("path")

# Step 9: Teardown
spark_session.stop()
```

Link to complete notebook here: https://

Auto Sklearn vs H2O AutoML

	AutoSklearn	**H2O AutoML**
Accuracy		
Training Time		
Leaderboard Time		
Boosting Algo Support	Yes	Yes
Deep Learning Support	No	Yes
Platform Dependent	Yes	Yes
Exaplainability	No	Yes
Scalability	No	Yes

Export Model As MOJO

Questions?

Thanks!!