MLlib/SKLearn

The Chosen
Ethan Tolson
Samuel Rios-Lazo
Josue Flores
Pun Tangsiriruangrit
Esther Tang
Jacob Farr

Learning Goals

- What is MLlib
  - Use Cases
  - Estimators and Transformers (Featurization)
- What is SKLearn
  - Use Cases
- MLlib vs. SKLearn
  - Data Format
  - Workflow
- Classification in MLlib vs SKLearn
  - Wrangling Data
  - Simple Model Creation
  - Model Evaluation
- Regression in MLlib vs SKlearn
  - Using Pipelines to Create Models
  - Hyperparameter Tuning
  - Model Evaluation- R2 and RMSE

Introduction - Pun

What is MLlib?

How does MLlib store data?

When would you use MLlib?

What is Sklearn?

How does Sklearn store data?

When would you use Sklearn?

Classification

MLlib Classification - Esther

Sklearn Classification - Josue

Scikit-learn, focuses on single-machine processing and is well-suited for smaller datasets that can fit into memory on a single machine.

Is a standalone library and does not have built-in integration with a big data processing framework.

Scikit-learn is primarily Python-based and provides a consistent API across all its algorithms

Regression

MLlib Regression - Pipelines and Hyperparameters - Jacob and Ethan

See Regression Notebook

Sklearn Regression and Hyperparameters - Samuel

```
model = GradientBoostingRegressor()
```

n_estimators	30	70
max_depth	1	5

Sklearn includes a pool of regression models to choose from, such as Gradient Boosting Regression, Linear Regression, XGBoost Regression, etc

Grid search is also a popular tool when searching for the best hyperparameters

It helps you find the best hyperparameters from the ones included in your grid by testing all the combinations and selecting the ones that performed the best

Thank You!