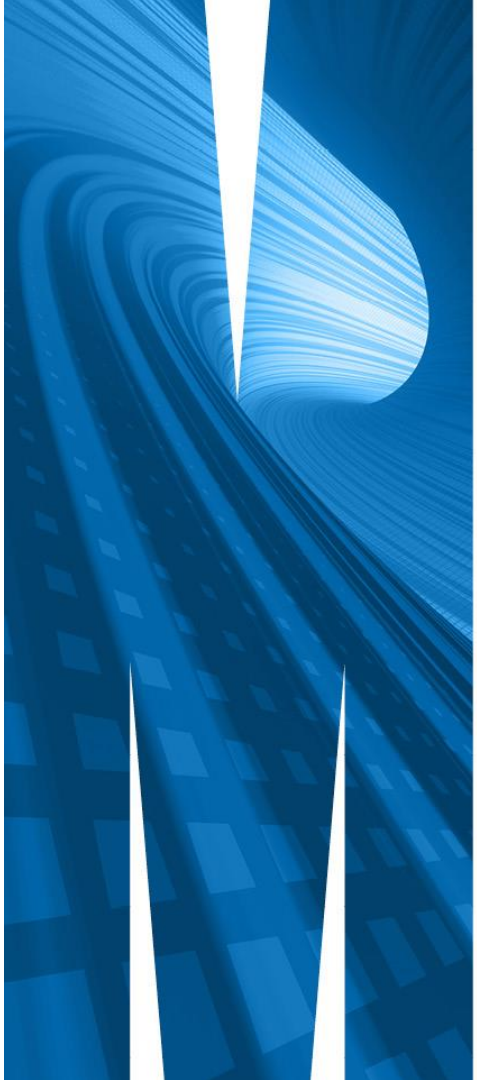
MONASH University

Session 05

FIT5202 Big Data Processing

Transformers, Estimators and Pipeline



Week 5 Agenda

- Session 1-4 Review
- Spark for Machine Learning
- Typical ML Workflow
- Understanding Transformers and Estimators
- Pipeline API
- Tutorial Use Case
    - Adult Income Prediction ML Workflow

Week 1-4 Review

- Introduction
- VM Installation and Setup
- Python Basics
- Spark Introduction
- RDDs and DataFrames

Week 1

- SparkSession vs SparkContext
- Data Partitioning
- RDD vs DataFrame
- Searching in RDDs and DataFrames
- Spark SQL

Week 2

- Spark Join Strategies
  - Broadcast Hash Join
  - Sort Merge Join
  - Shuffled Hash Join
- Parallel Joins
  - Inner, Outer, Left, Right, Left Anti, Left Semi
- Execution plans

Week 3

- Dataframe operations
  - Sort
  - Distinct
  - Groupby
- UDFs

Week 4

Why Spark for ML?

- Unified Analytics Engine
  - Ecosystem for Data Ingestion, Feature Engineering, Model Training and deployment
- No need to downsample data to fit in a single machine

Distributed Framework (Spark MLLib) vs Single Node Framework(sklearn)

A typical ML workflow

1. Data Preparation
2. Feature Engineering
3. ML Pipelines
    1. Selecting top features for our machine learning models
        1. ChiSqSelector API : Perform ChiSquare tests and select the most significant features that influence our target variable
4. Training Models
    1. We split the data into training and test data
5. Model Validation and Selection
    1. Using evaluation metrics
6. Exporting/Deploying the model

ML Pipeline

- Transformer
  - Apply rule-based transformations
  - Prepare data for model training
  - .transform() method
- Estimator
  - Learns parameters from your DataFrame
  - .fit() method,
- Pipeline API : to organize machine learning workflow
  - Organizes a series of transformers and estimators into a single model

Feature Engineering in SparkML

Feature Extractors

- TF-IDF
- Word2Vec
- CountVectorizer
- FeatureHasher

Feature Transformers

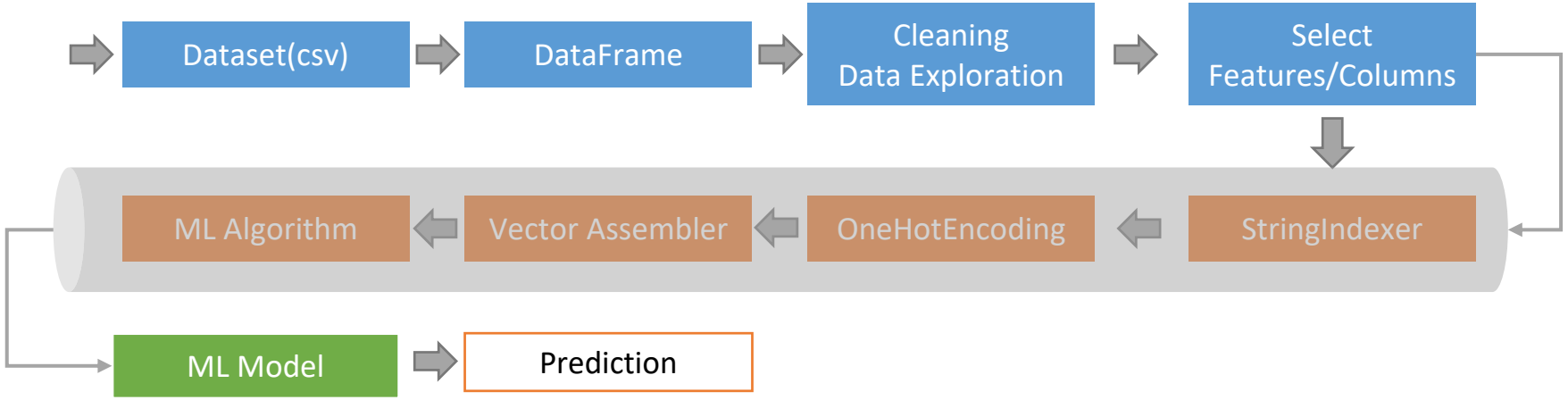
- Tokenizer
- StopWordsRemover
- n-gram
- Binarizer
- PCA
- PolynomialExpansion
- Discrete Cosine Transform (DCT)
- StringIndexer
- IndexToString

Feature Selectors

- VectorSlicer
- RFormula
- ChiSqSelector

https://spark.apache.org/docs/latest/ml-features.html

Adult Income Prediction ML Workflow



Thank You!

See you in the next session.