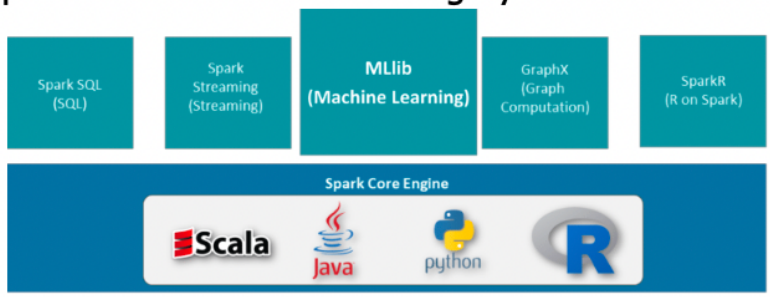
Spark Machine Learning

Spark ML and PySpark

- Spark ML is a machine-learning library
  - Classification: logistic regression, naive Bayes
  - Regression: generalized linear regression, survival regression
  - Decision trees, random forests, and gradient-boosted trees
  - Recommendation: alternating least squares (ALS)
  - Clustering: K-means, Gaussian mixtures (GMMs)
  - Topic modeling: latent Dirichlet allocation (LDA)
  - Frequent item sets, association rules, and sequential pattern mining
- PySpark is an interface for using Python

From [2]



Binary Classification Example

- **Binary Classification** is the task of predicting a binary label
  - Is an email spam or not spam?
  - Should I show this ad to this user or not?
  - Will it rain tomorrow or not?
- The Adult dataset
  - https://archive.ics.uci.edu/ml/datasets/Adult
  - 48842 individuals and their annual income
  - We will use this information to predict if an individual earns **<=50K or >50k** a year

Dataset Information

- Attribute Information:
  - age: continuous
  - workclass: Private,Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
  - fnlwgt: continuous
  - education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc...
  - education-num: continuous
  - marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent...
  - occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners...
  - relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
  - race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
  - sex: Female, Male
  - capital-gain: continuous
  - capital-loss: continuous
  - hours-per-week: continuous
  - native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany...
- Target/Label: - <=50K, >50K

Analyzing Flow

- Load data
- Preprocess Data
- Fit and Evaluate Models
  - Logistic Regression
  - Decision Trees
  - Random Forest
- Make Classification

Spark ML

- Spark ML is a machine-learning library
  - Classification: logistic regression, naive Bayes
  - Regression: generalized linear regression, survival regression
  - Decision trees, random forests, and gradient-boosted trees
  - Recommendation: alternating least squares (ALS)
  - Clustering: K-means, Gaussian mixtures (GMMs)
  - Topic modeling: latent Dirichlet allocation (LDA)
  - Frequent item sets, association rules, and sequential pattern mining

Classification vs Prediction

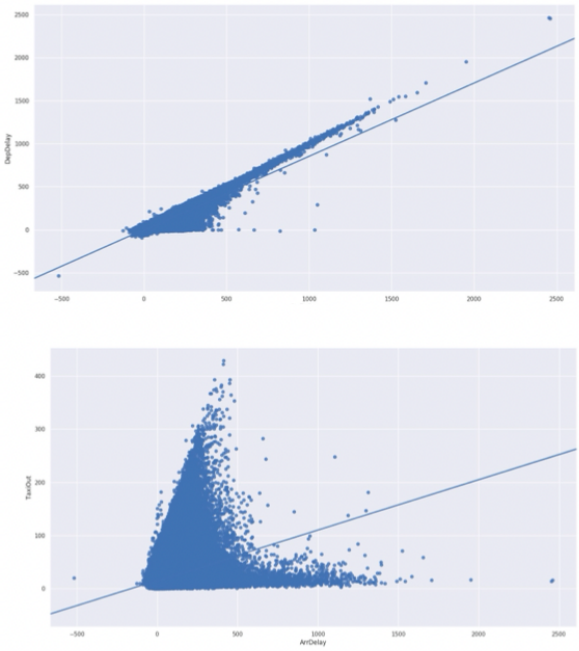
- Classification models predict categorical class labels [2]
  - Binary classification
- Prediction models predict continuous valued functions
  - Regression analysis is a statistical methodology that is most often used for numeric prediction

Predicting the arrival delay of commercial flights

- Problem
  - We want to be able to predict, based on historical data
    - The arrival delay of a flight using only information available before the flight takes off
- Dataset
  - http://stat-computing.org/dataexpo/2009/the-data.html
  - The data used was published by the US Department of Transportation
  - It compromises almost 23 years worth of data
- Approach
  - Using a regression algorithm

Dataset Information

	Name	Description
1	Year	1987-2008
2	Month	1-12
3	DayofMonth	1-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number
12	ActualElapsedTime	in minutes
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes
16	DepDelay	departure delay, in minutes



Analyzing Flow

- Load data
- Preprocess Data
- Train the data and obtain a model
- Evaluate the resulting model
- Make Predictions