MACHINE LEARNING AS A SERVICE: APACHE SPARK MLLIB ENRICHMENT AND WEB-BASED CODELESS MODELING

Zhengyi Le

Suning R&D Center at Palo Alto

Who I am?

Zhengyi (Jennifer) Le

Deputy Director, Big Data Lab
Suning R&D Center at Palo Alto

✉ Zhengyi.Le@ussuning.com

SPARK SUMMIT 2017

苏宁易购 suning.com

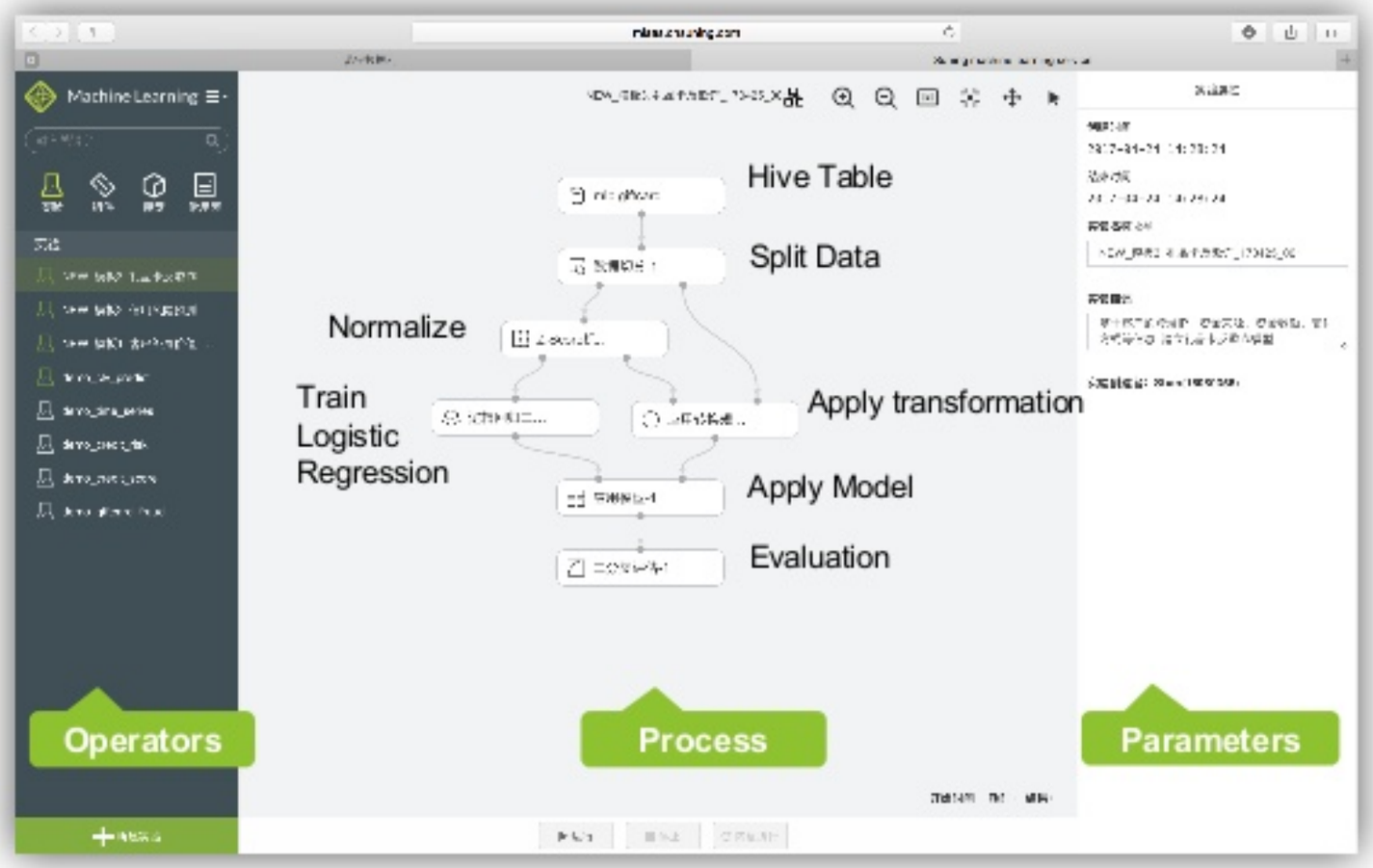
What you want as a Data Scientist ?

- Clean Data
  - And, labeled ~
- Powerful Computing
  - Lightening fast
- Easy to Use
  - No hard grammar or long hidden/default parameters

What we do for you

- **Spark-backend**
- **Codeless**
- **Drag-N-Play**
- **Team workspace**

Machine Learning Service



SPARK SUMMIT 2017

It is NOT EASY to make your work EASY

Spark 2.1.0 Overview Programming Guides API Docs Deploying More

MLlib DataFrame-Based API

Extracting, transforming and selecting features		Classification and regression	Clustering
Feature Transformers Tokenizer StopWordsRemover n-gram Binarizer PCA PolynomialExpansion Discrete Cosine Transform (DCT) StringIndexer IndexToString OneHotEncoder VectorIndexer Interaction Normalizer StandardScaler MinMaxScaler MaxAbsScaler Bucketizer ElementwiseProduct SQLTransformer VectorAssembler QuantileDiscretizer	Feature Extractors TF-IDF Word2Vec CountVectorizer Feature Selectors VectorSlicer RFormula ChiSqSelector Locality Sensitive Hashing LSH Operations Feature Transformation Approximate Similarity Join Approximate Nearest Neighbor Search LSH Algorithms Bucketed Random Projection for Euclidean Distance MinHash for Jaccard Distance	Classification Logistic regression Binomial logistic regression Multinomial logistic regression Decision tree classifier Random forest classifier Gradient-boosted tree classifier Multilayer perceptron classifier One-vs-Rest classifier (a.k.a. One-vs-All) Naive Bayes Regression Linear regression Generalized linear regression Available families Decision tree regression Random forest regression Gradient-boosted tree regression Survival regression Isotonic regression Linear methods Decision trees Tree Ensembles Random Forests Gradient-Boosted Trees (GBTs)	K-means Latent Dirichlet allocation (LDA) Bisecting k-means Gaussian Mixture Model (GMM) Collaborative Filtering Collaborative filteringExplicit vs. implicit feedback Scaling of the regularization parameter ML Tuning: model selection and hyperparameter tuning Model selection (a.k.a. hyperparameter tuning) Cross-Validation Train-Validation Split Advanced topics Optimization of linear methods (developer)Limited-memory BFGS (L-BFGS) Normal equation solver for weighted least squares Iteratively reweighted least squares (IRLS)

SPARK SUMMIT 2017

That is not enough

➡ Our Spark MLlib Extensions

Data Processing

Time Series

Domain-Specific Library for Finance

Processing (our new add-on)

Sampling

- SMOTE Sampling
- Weighted Sampling
- Cluster Sampling
- Random Sampling
- Upper/Down Sampling

Stats

- t-Test
- z-Test
- f-Test
- Distance Correlation

GraphX

- K-core
- Modularity

Feature Selection / Transformation

- Box-Cox Transformation
- Weight of Evidence
- Information Value
- Murmur/Sim Hash
- Random Forest Feature Importance
- Filter Feature Selection
- Forward/Backward Feature Selection

Operations

- Single Column Operations
- Multi Column Operations
- Linear Combination

Classifiers

- XGboost

Binning

- Monotonic Binning
- Equal Width
- Equal Frequency
- Decision Tree Binning

Time Series

- Decomposition
- Simple MA
- AR/MA/ARMA/ARIMA
- Auto-ARIMA
- AIC/BIC/AICc
- ACF/PACF

Evaluation

- Confusion Matrix

NLP

- Auto Tagging
- Self-Define Dictionary Library
- Distance
  - Hamming,
  - (Squared/Weighted) Euclidean
  - (Weighted) Manhattan
  - Chebyshev
  - Cosine
  - Minkowski
  - Tanimoto
- Similarity
  - (adjusted) Cosine
  - Pearson's Rank
  - Jaccard
- Hidden Markov Model
- Conditional Random Fields
- Auto Key Word
- Auto Abstracting

Time Series

date	pce	pop	psavert	uempmed	unemploy
1967-06-30	507.8	198712	9.8	4.5	2944
1967-07-31	510.9	198911	9.8	4.7	2945
1967-08-31	516.7	199113	9	4.6	2958

Spark-ts RDD-Based

DateTimeIndex:
[10/31/06, 11/30/06, 12/31/06, 1/31/07, …]

Key	Series
PCE	[9411, 9479, 9540, 9611,…]
POP	[300836,301070,301296,301481, …]
PSAVERT	[-0.9, -1.1, -0/9, -1,…]

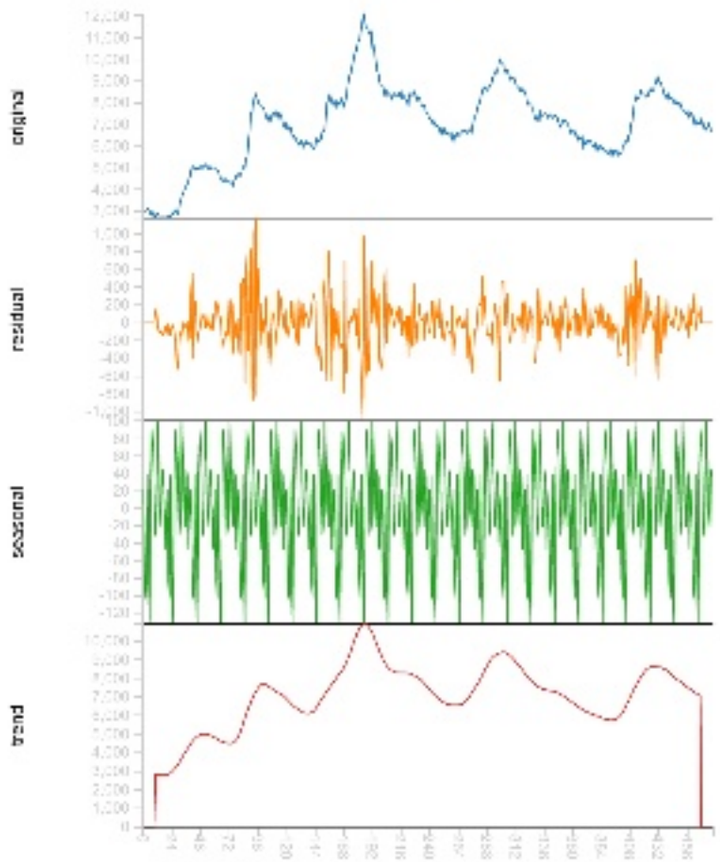
TimeSeriesRDD[K] extends RDD[(K, Vector[Double])]

Our-ts DataFrame-Based

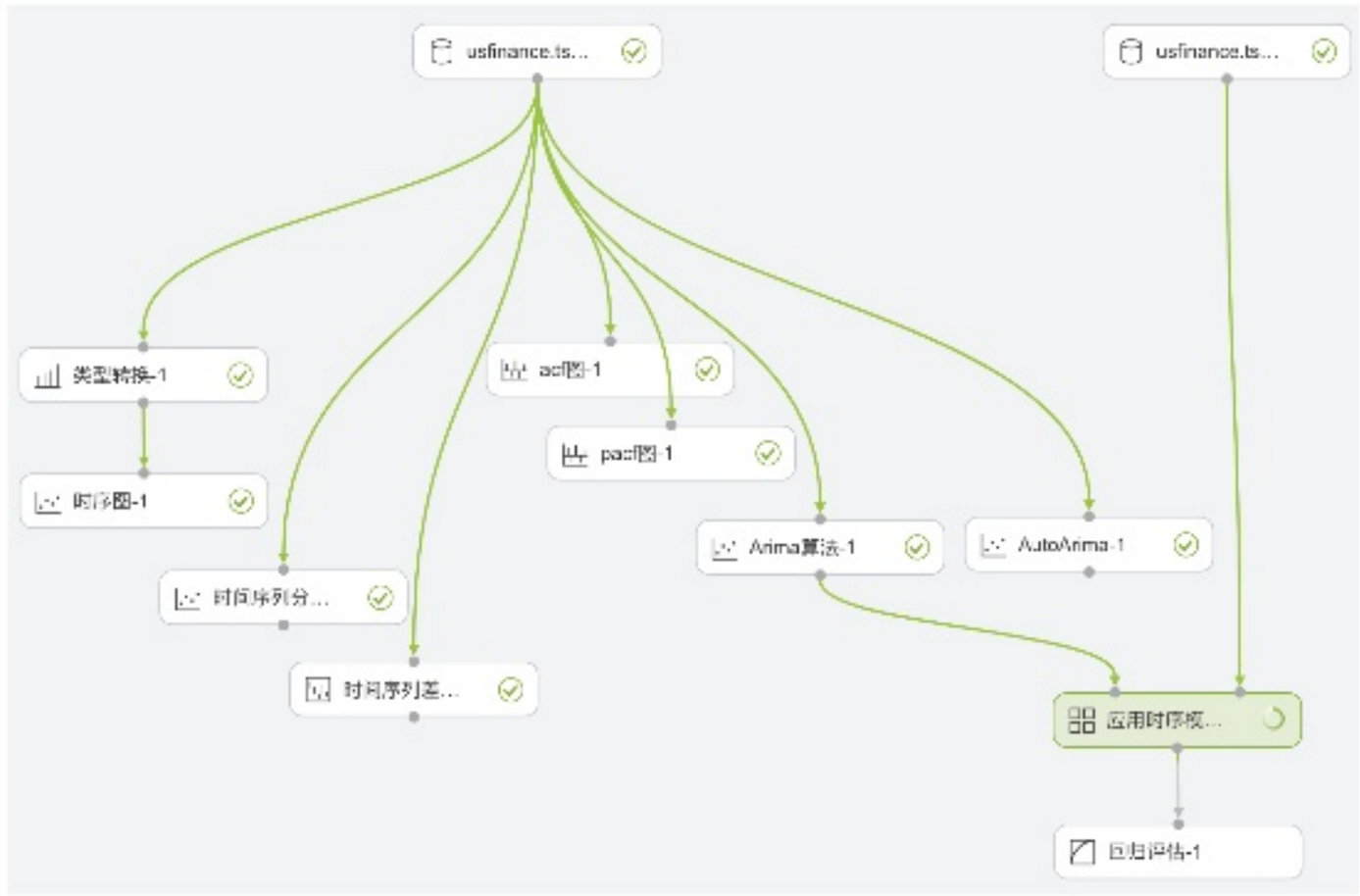
Date	PCE ($B)	POP (K)	PSAVERT (%)	UEMPMED (weeks)	UNEMPLOY (K)
10/31/06	9411	300836	-0.9	8.2	6826
11/30/06	9479	301070	-1.1	7.3	6849
12/31/06	9540	301296	-0.9	8.1	7017
1/31/07	9611	301481	-1	8.1	6865
2/28/07	9653	301684	-0.7	8.5	6724
3/31/07	9705	301913	-1.3	8.7	6801

Time Series

Functionality	Our-ts	Spark-ts	R
DataFrame	Yes	No	Yes (R data.frame)
Time Series Decomposition	Yes	No	Yes
Simple MA	Yes	No	Yes
AR	OLS/ Yule-Walker	OLS	OLS/Yule-Walker/MLE
MA	OLS/Yule-Walker	OLS	OLS/Yule-Walker/MLE
ARMA	OLS/Yule-Walker	OLS	OLS/Yule-Walker/MLE
ARIMA	OLS/Yule-Walker	OLS	OLS/Yule-Walker/MLE
Auto-ARIMA	OLS/Yule-Walker	OLS	OLS/Yule-Walker/MLE
AIC/BIC/AICc	Yes	Yes	Yes
ACF/PACF	Yes	Yes	Yes



Time Series



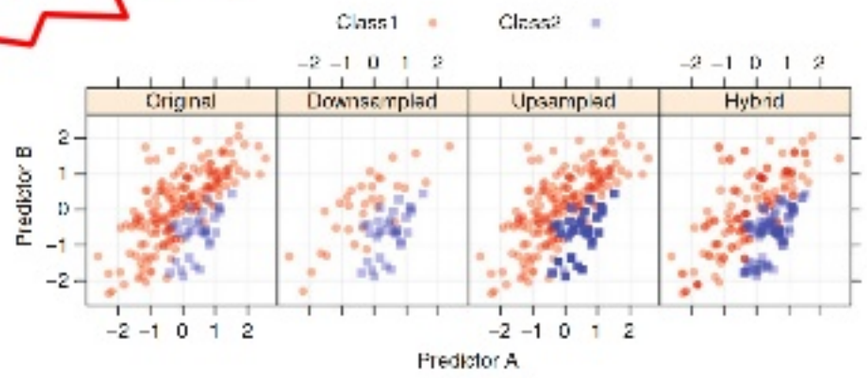
# of Data Rows	0.5 Million	2 Million	16 Million
Our ts	4 *sec*	34 *sec*	180 *sec*
Spark-ts	1 *sec*	2 *sec*	*Fail*
R	140 *sec*	*Fail*	*Fail*

Finance Domain

Example:
Fraud Detection

- SMOTE Sampling
- Cost Sensitive Decision Tree
- IP Mapping
- Mobile Number Grouping
- Bank Card Decoding
- National ID Decoding
- K-core (GraphX)
- Modularity (GraphX)
- Hypergraph (GraphX)

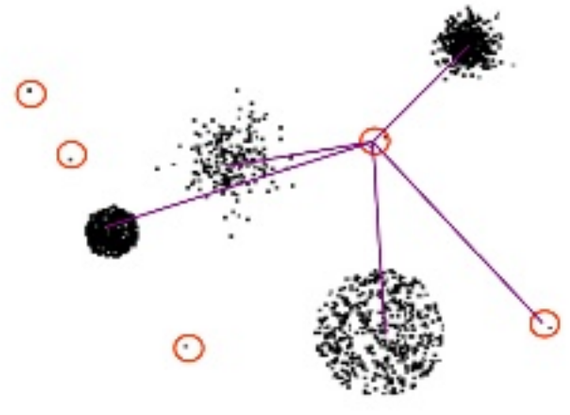
1. Sampling



2. Weight Assignment

	Actual Normal	Actual Fraud
Pred Normal	0	1000
Pred Fraud	5	0

3. Anomaly Detection (Unsupervised Learning)



SPARK SUMMIT 2017

Thank You.

Zhengyi.Le@ussuning.com

Our Demo:
https://youtu.be/pmN_-f-yXos