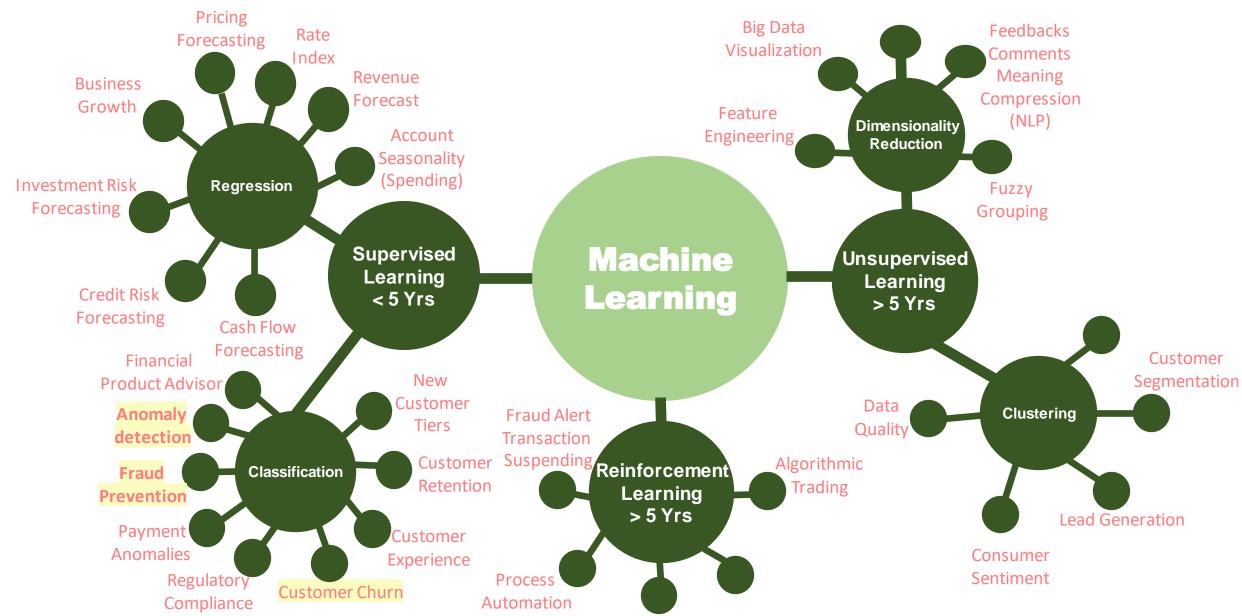
Topics to Discuss

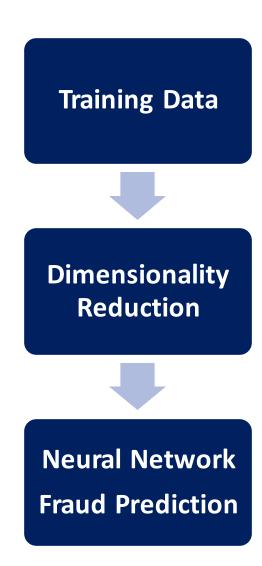
- ML pipeline from POC to Production
- Bring the current ML models to AWS,
 - Data Pipeline
 - Data Models (Training/Testing/Forecast Data) & ML Database
 - Gitlab Python Code Control & (Jupyter Notebook)
 - Kubeflow/Dataiku deployment
 - ML Data consumption (PostgreSQL/AWS Redshift) :Tableau Visualization & MLaas API
 - ML Wiki
- ML K8 & SparkSQL Architecture Scalability
- Data Compliance & Strategy
 - Coordinate between Data Centers on AWS
 - Data Masking
- ML Platform:
 - Platform Tools: SageMaker/Kubeflow/H2O/Data Robot/Dataiku
 - Dataiku to orchestrate(CI/CD/CT) the pipeline flow of ML projects
 - Templates for different project size
- ML Projects:
 - See Next Slide
- ML Roadmap: Showcase & ML workshops to communicate with businesses, Peers & Leaders
- ML Leadership: Communication, Problem Solving Skills, Mentor/Guideline/Goal with team

TAXONOMY OF MACHINE LEARNING METHODOLOGIES IN BANKING

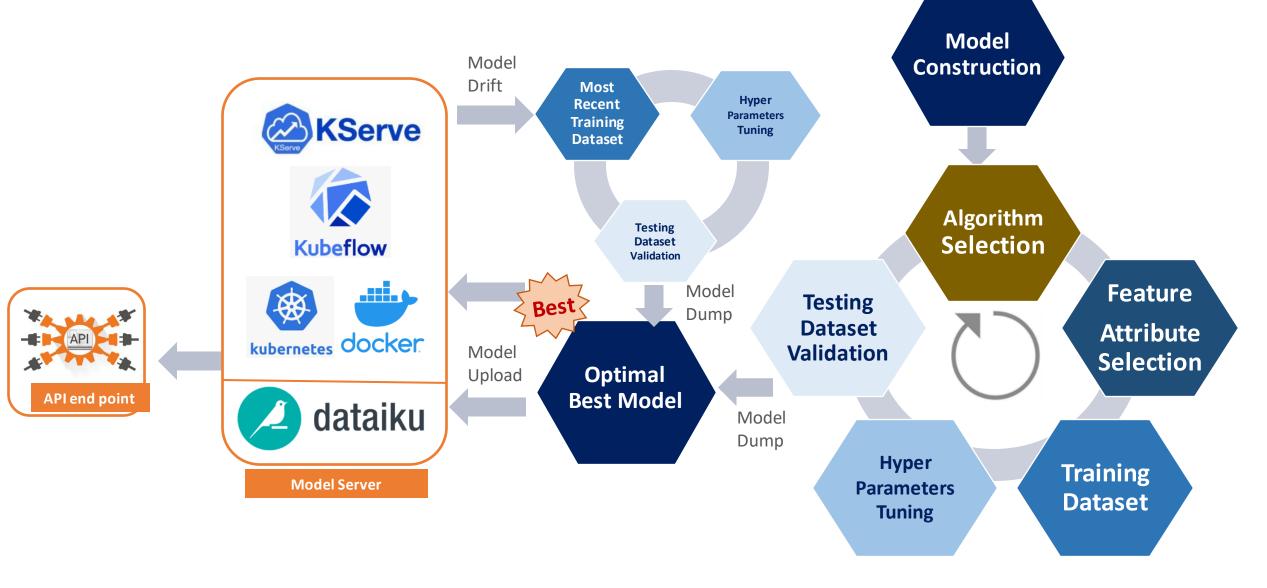


Fraud Prediction Risk Score

- Features include:
 - Transaction Type
 - Personal Information(Age, Race, Education, etc)
 - Data (\$\$, debt, credit)
 - IP address
 - Location
 - Credit Score
 - Payment History Score
 - Current Income
 - Credit History
- Measures:
 - Fraud Risk Score
- Date:
 - N/A
- Frequency:
 - N/A

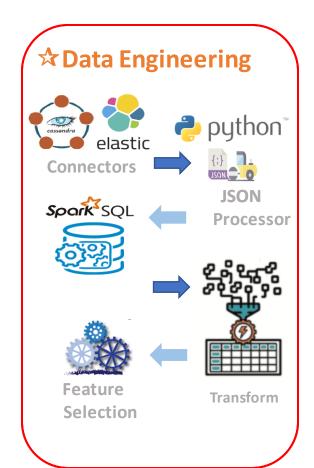


ML Model Validation/Deployment

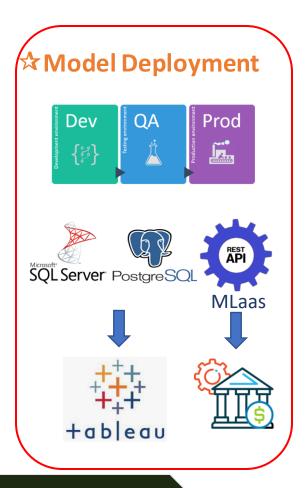


Big Data Machine Learning Data Architecture







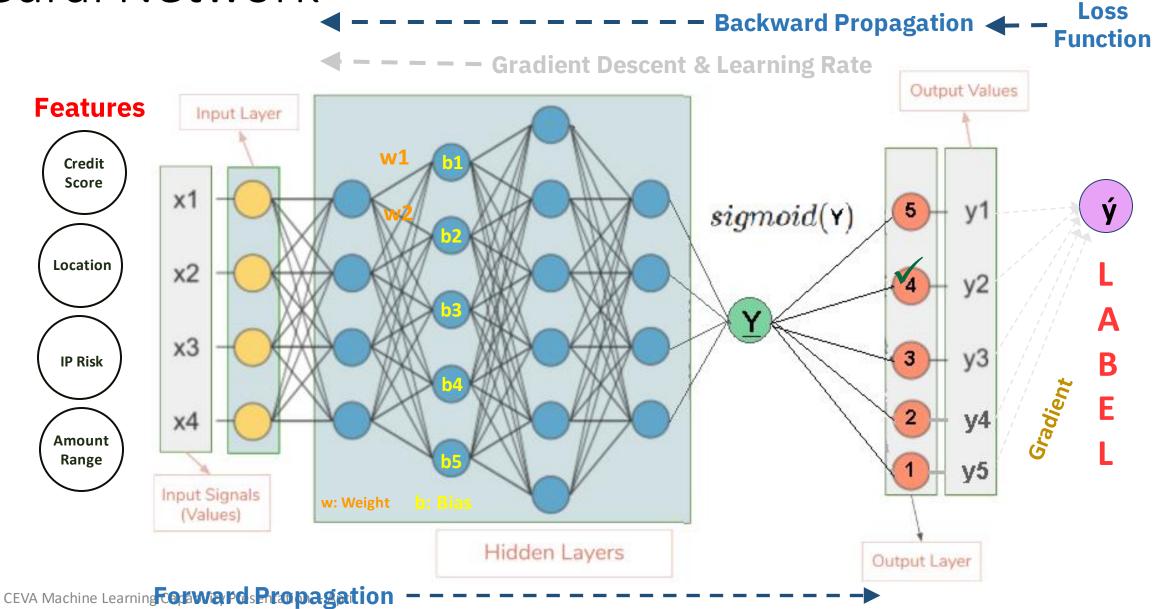


Big Data Parallel Processing Multi-Threaded Bulk Data Loading

Big Data Training Process

Trained Machine Learning Models

Neural Network



2020

Enterprise ML Platform Project Management

Data Model

Gold

- Clean Data/Denormalized
- Data Source: Structured/Unstructured
- MDM Mapped
- Data Size in > 0.5 TB
- Refresh Automatically

Silver

- Clean Data/Denormalized
- DataSource: Structured/Unstructured
- Data Size > 10 GB
- Refresh Automatically

Brown

- Clean Data Denormalized Form
- Data Size < 10 GB
- Refresh Automatically

RAW

- Various Forms
- Data Size in MB
- CEVA Machine Refresh Manually entation April 2020

Algorithm

Neuro-Linguistic Programming

- •Sentiment Analysis (+/-)
- NER (Reason Code)
- •Stemming & Lemmatization(Reason Code)

Classification

- Neural Network
- XGBoost
- Logistic Regression

Optimization

• Gradient Decent

Regression

- Linear
- Time Series

ML Solution Deployment

Large

- Multiple Algorithms
- Cross Data Centers
- Multiple Data Sources
- Dashboard/API Access

Medium

- Multiple Data Sources
- Multiple Algorithms
- Dashboard
- API Access

Small

- •Single Data Source
- Single Algorithm
- Tableau Dashboard Delivery

Adhoc

- Data not in Data Lake
- Single Algorithm
- Result in text file format(S3)