**CRISP-DM Custom ML Project Development Framework 11-26-20**

(by Chris Engstrom)

1. **BUSINESS UNDERSTANDING**
   1. **Business Situation**
      * *(stakeholder meetings and workshop)*
      * Needs and pain points
      * Opportunities
        + - (increase revenues, decrease expenses) 🡪 Increase Profit
          - Customers (acquire, develop, retain, increase CLV)
          - Reduce costs (supply chain, process time, etc.)
          - Optimize resource allocation (decision optimization)
          - Measure and manage risk
          - Improve decision-making
          - Improve measurement
      * Data assessment (high-level)
      * Readiness assessment (see readiness assessment)
      * Project selection (see project selection process)
      * Roadmap (initial multi-project plan: 3, 6, 12, 24, 36 months)
      * SUMMARY
   2. **Define Business Problem**
      * *(within scope of a project)*
      * Business objectives
      * Problem and solution requirements
      * Project scope
      * Hypothesis generation (specific)
      * Data enrichment requirements/costs (internal and external)
      * Critical Success Factors (CSFs)
      * Critical and missing features
      * Confusion matrix costs (classification)
      * Success metrics and measurement plan
      * Critical business sponsors and stakeholders
      * Security/compliance/privacy/AMSU (access movement storage updates)
      * WIP/deployment/pipeline plan
      * SUMMARY (update)
   3. **Define Analytic Approach**
   * Analysis (pre-modeling)
   * Target (extremely important!)
   * Features
   * Unstructured data (approach)
   * Sampling (if required)
   * Test design (if required)
   * Most appropriate ML technique(s) considering
     + - * Budget
         * Deployment complexity
         * Deployment costs
         * Missing feature acquisition costs
         * Workstream integration
         * Data security
         * Data variety, velocity, volume, veracity
         * Model opacity e.g., black box
         * Scoring requirements e.g., scale, velocity, frequency, volume
         * SUMMARY (update)
2. **DATA UNDERSTANDING**
   1. **Data Requirements**

* Data sources (structured, unstructured)
* Data scope (sources, stratifications, timeframe, etc.)
* Formats
* Datetime data
* Time-series data
* Integration (e.g., primary keys)
* Compliance (e.g., access, anonymization, storage, movement, updates)
* Data access and availability (internal)
* New data capture required?
* External data required?
* Other?
* SUMMARY (update)
  1. **Data Structure**
     + Data source review (columns, counts, primary keys, etc.)
     + Column review and understanding (by data source)
     + Frequencies (values, distinct, invalid, missing, atomized and inconsistent values e.g., ‘MN’ ‘Minn’, etc.)
     + Descriptives -Numeric (basic, missing, invalid, outliers, skewness, etc.)
     + Descriptives - Datetime (basic, format, missing, invalid, outliers, skewness, format)
     + Primary keys (values, distinct, invalid values, missing values, inconsistent values, etc.)
     + Unstructured (if available)
     + Other
  2. **Data Audit Report**

1. General comments
2. Data sources (hygiene, quality, consistency, reliability, sparsity, sensitivity, dimensionality, volume, etc.)
3. Frequencies (values, distinct, invalid, missing, atomized and inconsistent values e.g., ‘MN’ ‘Minn’, etc.)
4. Descriptives -Numeric (basic, missing, invalid, outliers, skewness, etc.)
5. Descriptives - Datetime (basic, format, missing, invalid, outliers, skewness, format)
6. Primary keys (values, distinct, invalid values, missing values, inconsistent, etc.)
7. Biases (systematic and random)
8. Compliance discussion (if necessary)
9. Recommendations
10. SUMMARY (update)
11. **DATA CONSTRUCTION, ANALYSIS, & PREPARATION**
12. **Data Construction and Hygiene**

* Data selection by Date Range defined in project scope
* Construction
  + Select data
  + Integrate data
  + DERIVE - Target (extremely important!)
  + Stratification (if required) - define relevant homogenic groups
  + Sampling (if required) e.g., 1%, 10%, 100%
  + Prepare time series data
  + Prepare unstructured data (if available)
  + Other
  + Data hygiene
* Drop rows (bad, invalid, missing, NaN, etc.)
* Drop columns (irrelevant, bad, sparse, etc.)
* Format data (convert types e.g., int🡪str, int🡪float, int🡪datetime, NaN, other)
* Format data (recode inconsistent data types e.g., ‘MN’ vs. ‘Minn’, etc.)
* Missing values (drop or impute with mean, median, regression, etc.)
* Negative values (invalid) (drop rows or impute with mean, median, regression, etc.)
* Outliers (coerce, drop, bin) \*\*DO FIRST\*\*
  + Other

1. **Exploratory Data Analysis (EDA)**

* Descriptives (basic, missing, invalid, outliers, skewness, etc.)
* Frequencies (values, distinct, invalid, missing, atomized, etc.)
* Plots - Univariate (bar, histograms, line, etc.)
* Plots - Multivariate (scatter, etc.)
* Correlation - among variables
* Correlation - Target and variables
* Crosstabs/ANOVA (Target and variables)
* Linearity - among variables
* Linearity - Target and variables
* Skewness - Target and variables
* Biases - systematic and random
* Target imbalance
* Unmanageable value sets
* Valid negatives (scale values)
* Scaling (variables with different scales)
* Interactions detection (use trees to detect)
* Text analytics / sentiment / tone
* Clustering analysis
* Quick and dirty models
* Other
* SUMMARY (update)

1. **Data Preparation**

* Optimization
* BALANCE - Target imbalance (SMOTE)
  + RECODE - Unmanageable value sets
  + RESCALE - Feature scaling (z-scores, standardization)
  + RESCALE - Negative values (valid) (rescale: e.g., X + 10)
  + Interactions (multivariate) (DEFINE new features)
  + Correlations (among variables) - drop variables
  + Transformations - Normalize skewed variables
  + Transformations - Linearity - among variables
  + Transformations - Linearity - Target and variables
  + Correct biases (if possible)
  + Other
* Feature Engineering
  + DERIVE - Target (updated)(very important!)
  + DERIVE - Features
  + DERIVE - Interaction features (e.g., ratios)
  + DERIVE - Other
  + RECODE - Binning
  + RECODE - Dummies
  + RECODE - Other
  + DATA REDUCTION - PCA - correlation among features
  + Auto data preparation
  + Other

1. **ANALYSIS & MODELING**
2. **Feature Selection**

* Feature Selection
* Feature independence
* Target leakage
* Critical missing features
* Data reduction
* Other
* SUMMARY (update)

1. **Modeling & Analysis**

* Identify most appropriate modeling techniques
* Define test design
* Partitioning (calibration, validation, test)
* Build “Quick and Dirty” models
* Assess and refine models
* Select 3-5 model candidates
* Other

1. **EVALUTION**
2. **Model Evaluation**

* Technical
  + Classification - Confusion matrix - TP TN FP FN accuracy, costs, etc.
  + Regression – R2 and MSE (plots by # features, orders etc.)
  + Regression – Plots (distribution and scatters etc.)
  + Number of features
  + Technique
  + Complexity
  + Leakage
  + Overfitting
  + Other
* Business
  + Complexity
  + Feature weighting
  + Compliance
  + Costs (e.g., data)
  + ROI measureability
  + Marginal value $ (revenue, cost reduction, etc.)
  + Check - Quality (do variables make sense)
  + Check - Costs (are variables too expensive for production)
  + Check - Is output useful?
  + Check - Can the output be acted on?
  + Check - Workstream Integration Plan (WIP)
  + Other
* Deployment
* Complexity
* Pipeline development costs (time, headcount, budget)
* Check - Costs (are variables too expensive for production)
* Check - Is output useful?
* Check - Can the output be acted on?
* Check - Workstream Integration Plan (WIP)
* Check - Classification cost analysis
* Check - Scale (can solution be implemented at scale)
* Other
* Re-calibrate winning models on entire datasets
* Determine next steps
* SUMMRY (update)

1. **DEPLOYMENT**
2. Roadmap (update)
3. Deployment test (small scale)
4. Pipeline - develop
5. Pipeline - test
6. Pipeline - production
7. Pipeline - monitoring and maintenance
8. Testing and Production
9. Measurement and monitoring
10. ROI Assessment
11. Final report
12. Other

END