**CRISP-DM ml project development framework (modified IBM-modified)**

1. **BUSINESS UNDERSTANDING**
   1. **Business Situation**
      * Needs and Pain Points
      * Opportunities
        + - Increase revenue
          - Decrease costs
          - Increase profits
          - Optimize resource allocation
          - Improve measurement
          - Reduce process time
          - Reduce risk
      * Data assessment
      * Readiness assessment (see project candidate selection process)
      * Define Roadmap
   2. **Business Problem Definition**
      * Define business objectives
      * Define problem and solution requirements
      * Hypothesis generation and definition
      * Define project scope
      * Define success metrics and measurement plan
      * Discuss Critical Success Factors (CSFs)
      * Discuss key features and missing features
      * Discuss data enrichment requirements/costs
      * External data enrichment requirements/costs
      * Define WIP/deployment plan
      * Engage critical business sponsors and stakeholders
      * Discuss confusion matrix costs
      * Discuss security/privacy/AMS (access move store)
   3. **Analytics Approach**
   * Define the analysis
   * Define the dependent variable
   * Define independent variable(s)
   * Discuss unstructured data
   * Express problem in context of ML techniques
   * Discuss sampling requirements
   * Identify most appropriate technique(s) considering
     + - * Project budget
         * Deployment complexity
         * Deployment costs
         * Missing variable acquisition costs
         * Workstream integration
         * Data security
         * Data variety, velocity, volume, veracity
         * Model opacity e.g., black box
         * Scoring requirements e.g., frequency, amounts
2. **DATA UNDERSTANDING**
   1. **Data Requirements**

* Formats
* Structured data
* Unstructured data
* Date/time data
* Readily available data
* Are less accessible data required?
* Are external data required?
* Are new data capture mechanisms required?
  1. **Data Structure**
     + Record counts
     + Missing variables
     + Frequencies - values
     + Frequencies - invalid values
     + Frequencies - missing values
     + Descriptives - basic
     + Descriptives - invalid values
     + Descriptives - missing values
     + Descriptives - outliers
     + Descriptives - normality
     + Date/time - ranges
  2. **Data Audit Report**

1. **DATA PREPARATION**
2. **General**

* Pre-segment instances
* DEFINE - date ranges
* DERIVE - Target variable candidates
* DERIVE - Features/variables
* Prepare Time Series data
* Discard unusable fields

1. **Sampling**

* DEFINE - sampling (e.g., 100%, 10%, 1%)
* DEFINE - sampling (e.g., calibration, validation, test)

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

* Univariate plots (frequencies, histograms)
* Multi-variate plots (scatters)
* Detection - Linearity
* Detection - Outliers
* Detection - Skewness
* Detection - Missing values
* Detection - Target imbalance
* Detection - Unmanageable classes
* Detection - Valid negatives (scale values)
* Detection - Scaling (variables with different scales)
* Detection - Interactions! (trees)
* Detection - Correlation (among features)
* Detection - Correlation (with target)
* Text analytics / sentiment
* Clustering / auto-cluster
* Quick and dirty models
* Leakage check (target variables)

1. **Transformations**

* Handling - Linearity
* Handling - Outliers (Coerce/Drop)
* Handling - Skewness (Binning, Transformations)
* Handling - Missing values (impute: mean med regression)
* Handling - Target imbalance (SMOTE)
* Handling - Unmanageable classes (Binning)
* Handling - Valid negatives (rescale: e.g., X + 10)
* Handling - Scaling (z-scores, standardization)
* Handling - Interactions! (DEFINE new features)
* Handling - Correlations (among features) - combine (ratio)
* Handling - Correlations (among features) - drop
* Handling - Correlations (among features) - Data Reduction
* Auto data preparation

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

* Feature - Selection
* Feature - Independence check
* Feature - Target leakage check
* Feature - Essential missing variable check
* Data reduction

1. **Modeling & Analysis (iterative process)**

* Identify most appropriate techniques
* Build Quick and Dirty models
* Refine models
* Select 3 -5 model candidates

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

* Check - LEAKAGE
* Check - Model performance (classification)
* Check - Model performance (regression)
* Check - Overfitting
* 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)
* Check - Classification cost analysis

1. **DEPLOYMENT**
2. Roadmap (revisit)
3. Testing
4. Production
5. Measurement
6. Monitoring
7. ROI Assessment

END