Data Science Institute

MIDDLE TENNESSEE STATE UNIVERSITY

Process for Operationalizing a Machine Learning Problem with the CRISP Model

Business Understanding

1. Framing the problem

- What is the expectation of analyzing the data?
- Is there a question to be answered?
- Is it completely exploratory? (a lot of data and no questions)
- Is it a machine learning problem? (What type of prediction?)
- A visualization or report may be all that is needed.

Data Understanding

2. Setup the workspace

- Use an editor - Jupyter
- Folder management
  - Root folder
    - data folder
    - raw folder
    - WIP folder
    - images folder
    - docs folder
- Import and pip install libraries
  - Numpy, Pandas, Scikit-learn, MatPlotLib, Seaborn, Statsmodels

Prepared by Dr. Charlie H. Apigian

3. Get the Data

- Import from:
    - csv or xls/xlsx, URL, SQL, txt, other files/connections

4. Explore the Data

- Visualize the data
    - histograms, bar charts, scatter plots, correlation matrix
- Group by
- Value counts
- Info()
- Head()
- Describe()

Data Preparation

5. Cleanse the Data

- Cleaning NaN values
    - fillna with value, median, mean, grouped mean
    - Drop NaN

6. Transform the Data

- Change categorical data to numerical – binary, ordinal, or dummy variables
    - Use a function
    - Label Encoding
    - One Hot Encoding (dummy variables)
- Standardize and normalize the data
- Create a pipeline

7. Feature Engineering

- Create new variables based on other features

8. Create your X and y datasets

- Create a dataset for your target variable (y) and your features (X)

Modeling

9. Split the Data

- Train test split
- Standardize X_train and X_test (separately)

10. Select the Model/Test

Supervised Learning

- Numerical target - Regression
  - Lasso
  - Ridge
  - Backwards model building
- Categorical target – Classification
- Probabilistic
  - Logistic regression
  - Naive Bayes
- Decision tree modeling
- Ensemble
  - Random forest
- SVM

Unsupervised learning

- Clustering
  - K-means
  - Hierarchal
- Dimension Reduction
  - PCA

Evaluation

11. Fine tune the Model

- K-folds
- For loop alpha scores
- Grid Search

12. Evaluate the Final Model

- Accuracy Scores (RMSE, etc.)
- Confusion matrix

Deployment

13. Identify and deploy the model for testing and production