|  |  |
| --- | --- |
| **ASSIGNMENT 2: RESEARCH OUTLINE** | |
| **1) Business Understanding** | |
| 1.1. Thesis (As Question): | How well do the sale of automobile data in the past forecast predict the future SAARTOT index returns in different time intervals (monthly, quarterly, yearly) |
| 1.2. Expectation (Yes/No): | We expect the monthly prediction gives best forecast |
| 1.3. Baseline: | External: F1-score, ROC, AUC |
| 1.4. Expected Value-Add Metrics: | Find relationship between sale auto data and forecast SAARTOT index |
| **2) Data Understanding** | |
| 2.1. Dataset: | Size (#27 Rows x #229 Cells): Source(s): Bloomberg |
| 2.2. List Categorical (CF) & Numerical Features (NF): | total import auto sales(NF), total domestic auto sales(NF), total sales per brand(NF), total sales per auto size(NF) |
| 2.3. Transforming CF/NF into one another: | No |
| 2.4. Assigning Independent & Dependent Variables: | DV: SAARTOT index  IV: total sales per brand(NF), total sales per size(NF), total sales per auto size(NF) |
| 2.5. Changing Dimensionality: | No |
| 2.6. Time series: | No |
| 2.7. List Sampling and Resampling Methods: | k-fold, cross validation |
| 2.8. List other data manipulation / prep techniques: | PCA is used for Logistic Regression. |
| **3) Model Understanding** | |
| 3.1. How are you generating the baseline: | We will use SAARTOT index price in the past to test our predictions |
| 3.2. Are your forecasting: | Yes. Time horizon: 2000.01-2019.01 |
| 3.3. Model Selection (circle or add): | Logistic Regression  XGBoost Tree classification |
| 3.4. For each technique, list key parameters and settings: | Logistic Regression: penalty, objective function, features  XGBoost Tree classification: learning rate, depth, number of boosted trees, leaf nodes, loss function, values of regularization, minimum loss reduction, |
| 3.5. Will you change parameter settings to improve fit: | Yes. Changes: penalty, objective function, tree depth, loss function, number of boosted trees |
| **4) Performance / Fit** | |
| 4.1. Do you expect to outperform on the baseline: | Yes |
| 4.2. What fit metrics will you use (circle or add): | Type 1/2 Error, ROC, AUC, F1-score |