|  |  |
| --- | --- |
| **ASSIGNMENT 2: RESEARCH OUTLINE** | |
| **1) Business Understanding** | |
| 1.1. Thesis (As Question): | Can we predict the direction of movement of SAARTOT index in future using past sales figures of major car manufacturer and manufacturer details such as P/E ratio etc.? |
| 1.2. Expectation (Yes/No): | Yes |
| 1.3. Baseline: | External: Support Vector Machine, Metrics: Accuracy, ROC AUC score |
| 1.4. Expected Value-Add Metrics: | Identify if we can leverage cyclical patterns in the Auto industry to predict future direction of movement of auto index |
| **2) Data Understanding** | |
| 2.1. Dataset: | Size (#Rows x #Cells): 227 x 27 Source(s): Bloomberg |
| 2.2. List Categorical (CF) & Numerical Features (NF): | NF: Sales per auto-manufacturer per month, Factors such as PE Ratio, Price to Sales Ratio, Price to Book value ratio etc. |
| 2.3. Transforming CF/NF into one another: | No |
| 2.4. Assigning Independent & Dependent Variables: | Yes. ID: Sales per auto-manufacturer per month DV: SAARTOT index for the next month |
| 2.5. Changing Dimensionality: | No |
| 2.6. Time series: | No: Implicitly time series, only using the fact for splitting the training and test data, otherwise treating as a classification task |
| 2.7. List Sampling and Resampling Methods: | Split testing and training on the basis of timestamp |
| 2.8. List other data manipulation / prep techniques: | Join function performed on data obtained, data preprocessing using imputation methods |
| **3) Model Understanding** | |
| 3.1. How are you generating the baseline : | Yes. Techniques: Support Vector Machine with RBF kernel and penalty term = 1 |
| 3.2. Are your forecasting: | Yes. If yes, time horizon: 2000 January till 2019 February |
| 3.3. Model Selection (circle or add): | Gradient Boosted Trees (also XGBOOST), Support Vector Machine |
| 3.4. For each technique, list key parameters and settings: | Support Vector Machine: penalty, # of iterations, Kernel function  Gradient Boosted Trees (and XGBOOST): Loss, Learning Rate,# of Estimators, depth |
| 3.5. Will you change parameter settings to improve fit: | Yes, Hyper parameter tuning using grid search |
| **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): | Accuracy, ROC AUC score |