Data Stream Processing - Rapport TP1

Par Marius Ortega et Louis Teillet

I. TP – PART 2 : Classifiers

1. Classifiers' Presentation

1.1. AdaBoost

Adaboost, or Adaptative Boosting is a part of the class of ensemble methods. Which means it uses a series of weak classifiers in its computation. In addition, given that it is a boosting algorithm, each classifier uses the results of the previous ones to make its prediction.

More Precisely, AdaBoost first associates a uniform weight $\left(\frac{1}{N}\right)$ to each dataset's individual. Then after the computation of each classifier, misclassified points will see their weight increase while well classified points' weights will decrease, giving more importance to misclassified points.

1.2. AMF

Conversely to AdaBoost, AMF or Aggregated Mondrian Forest is a native online learning method. To better understand the algorithm let us define a Mondrian Tree. It is a tree where every node r has a split time τ_r which stochastically increases with the depth of the node. For instance, $\tau_t(root) = 0$ and $\tau_r(leaves) = +\infty$. A Mondrian Forest is an ensemble of Mondrian Trees.

Each node in a tree predicts according to the distribution of the labels it contains. This distribution is regularized using a "Jeffreys" prior with parameter $dirichlet$. For each class with $count$ labels in the node and $n\_samples$ samples in it, the prediction of a node is given by :

$$\frac{count + dirichlet}{n_{samples} + dirichlet \times n_{classes}}$$

Finally, the prediction for a sample is found by aggregating the predictions of sub-trees.

1.3. Hoeffding Tree

In the case of Hoeffding trees considering N of a random variable r. r is information gain. If we compute the mean, r’, of this sample, the Hoeffding bound states that the true mean of r is at least $r' - \varepsilon$, with probability $1 - \delta$, where δ is user-specified :

$$\epsilon = \sqrt{\frac{R^2 \ln\left(\frac{1}{\delta}\right)}{2N}}$$

The Hoeffding Tree algorithm uses the Hoeffding bound to determine, with high probability, the smallest number, N, of examples needed at a node when selecting a splitting attribute.

1.4. KNN

KNN or K-Nearest Neighbors is a typical and extremely famous supervised algorithm. KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points which are closer to the test data to predict the class.

1.5. Drift Retraining

Drift retaining is a wrapper for any classifier. It monitors the incoming data for concept drifts and warnings in the model's accuracy. In case a warning is detected, a background model starts to train. If a drift is detected, the model will be replaced by the background model, and the background model will be reset. In our case, we combined it with the Hoeffding Tree model.

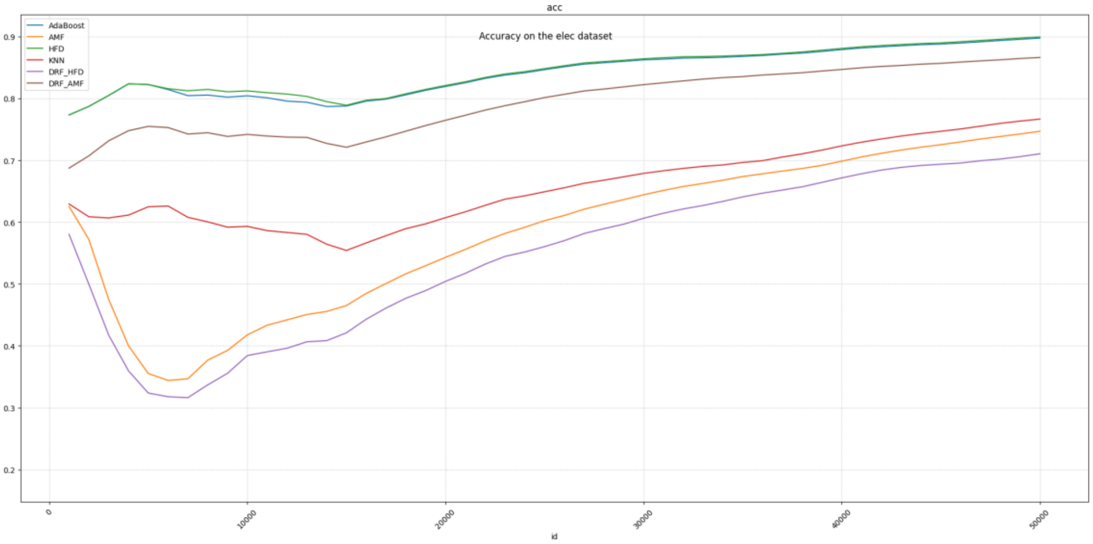
2. Results presentation and conclusion (50 000 samples)

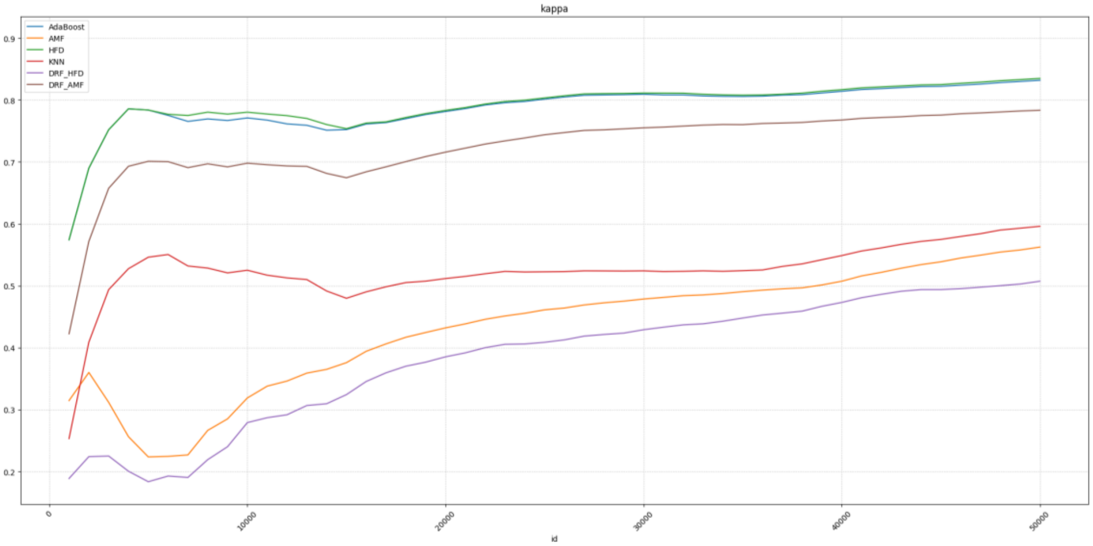
Model	Accuracy	Kappa Cohen	F1	Precision	Recall	Runtime
AdaBoost	0.75	0.56	0.75	0.75	0.75	8min13s
AMF	**0.90**	**0.83**	**0.90**	**0.90**	**0.90**	**8min49s**
Hoeffding Tree	0.71	0.51	0.71	0.71	0.50	52s
KNN	0.87	0.78	0.87	0.87	0.87	7min20s
Drift Retraining (Hoeffding Tree)	0.77	0.60	0.77	0.77	0.77	45s
Drift Retraining (AMF)	**0.90**	**0.83**	**0.90**	**0.90**	**0.90**	**9min14s**

From reading the results above, the best overall model is AMF Classifier.

However, we also notice that Drift Retrainer increased Hoeffding tree classifier’s performance (0.71 without Drift Retrainer and 0.77 with Drift Retrainer for Accuracy). Thus, we are curious to see AMF performances with Drift Retrainer. After experimenting with Drift Retrainer on AMF, we denote that we didn't increase performance considering 50 000 samples.

As a conclusion, we recommend using AMF Classifier as it is our best overall model in performance. Considering runtime (computed on 50 000 elements), it is the worst one while Hoeffding Tree offers the best time complexity results. It makes sense as AMF is a Forest model (ensemble of trees) while Hoeffding is a single tree. At the first glance, the use of Drift Retrainer doesn't seem necessary as we didn't improve metrics and the runtime increased from 8min49s to 9min14s. However, after at looking the plots below, we notice that Drift Retraining AMF is much more stable than the classical AMF. Consequently, the use of Drift Retraining could be important depending on the application (if the result need to be stable fast or not).





II. PART 3 : Clustering

3. Regressors’ Presentation

3.1. Linear Regression (Lasso Penalization)

Linear regression assumes a linear relationship between the variables, making it a simple yet effective approach for regression analysis.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \; + \; \epsilon$$

Multiple linear regression allows us to model and understand the relationship between a dependent variable and multiple independent variables.

In statistics and machine learning, lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

Instead of minimizing : $\sum_{i=1}^{N}\left(\widehat{\beta_0} + \widehat{\beta^T} X_i - Y_i\right)^2$

We minimize : $\sum_{i=1}^{N}\left(\widehat{\beta_0} + \widehat{\beta^T} X_i - Y_i\right)^2 + \lambda \sum_{j=1}^{p} |\widehat{\beta_J}|$

3.2. Hoeffding Tree Regressor

See 1.3

3.3. Hoeffding Adaptive Tree Regressor

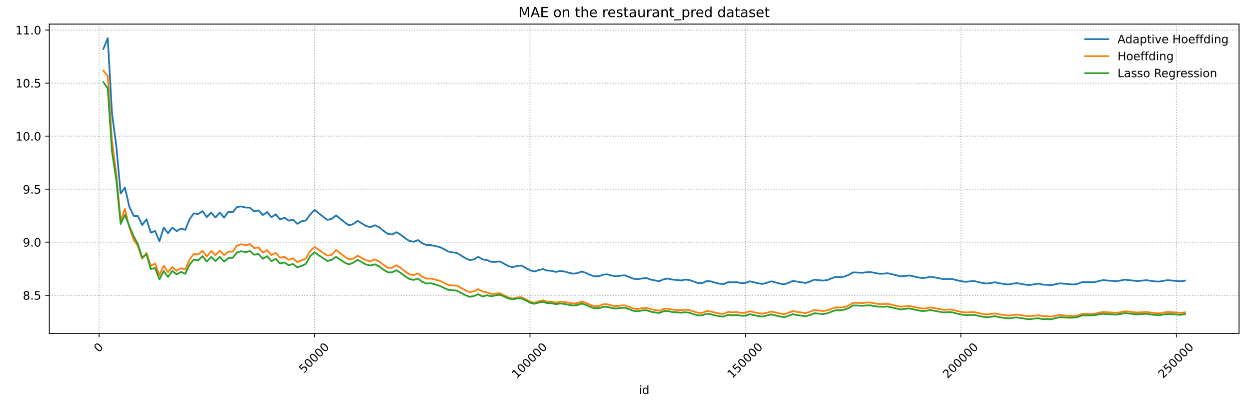
The Hoeffding Adaptive Tree uses drift detectors to monitor performance of branches in the tree and to replace them with new branches when their accuracy decreases.

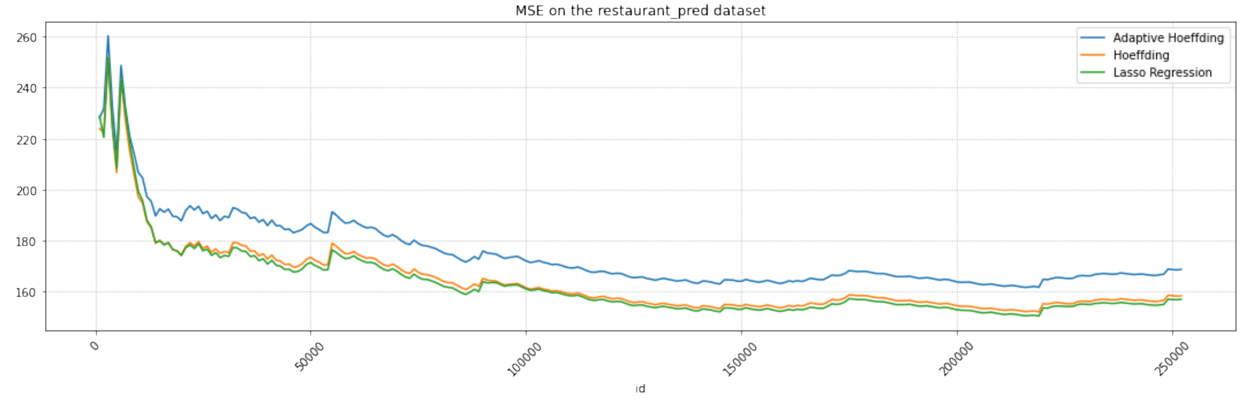
The bootstrap sampling strategy is an improvement over the original Hoeffding Adaptive Tree algorithm. It is enabled by default since, in general, it results in better performance.

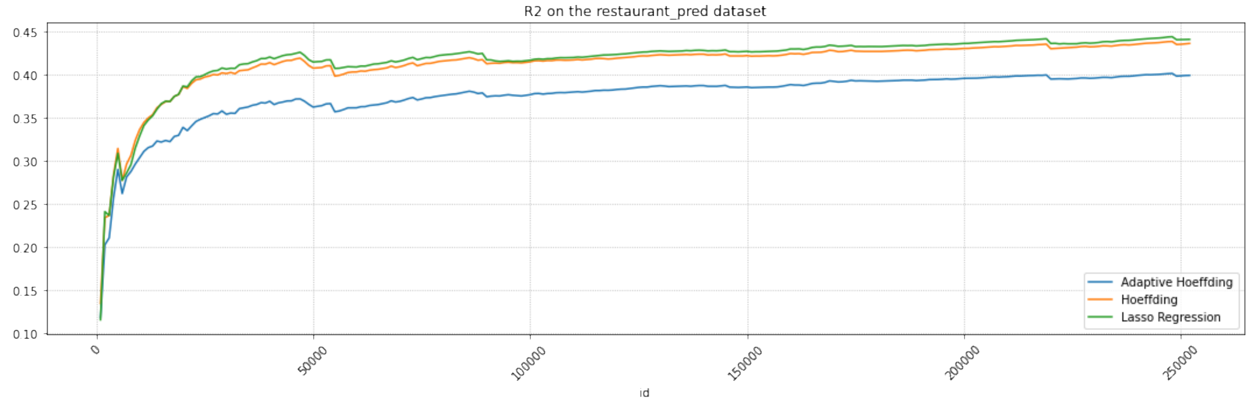
ADWIN (ADaptive WINdowing) is a popular drift detection method with mathematical guarantees. ADWIN efficiently keeps a variable-length window of recent items; such that it holds that there has not been changes in the data distribution. This window is further divided into two sub-windows (W0,W1) used to determine if a change has happened. ADWIN compares the average of W0 and W1 to confirm that they correspond to the same distribution. Concept drift is detected if the distribution equality no longer holds. Upon detecting a drift, W0 is replaced by W1 and a new W1 is initialized. ADWIN uses a significance value δ=∈(0,1) to determine if the two sub-windows correspond to the same distribution.

4. Results presentation and conclusion

Model	MSE	MAE	R2	Runtime
Lasso Regression (lamba = 1)	**157.05**	**8.32**	**0.44**	**1m08s**
Hoeffding Tree	158.37	8.34	0.43	1m18s
Hoeffding Adaptive Tree	174.5	8.80	0.37	2m07s







From reading the results above, the best overall model is Lasso, but its performances are close to Hoeffding Tree ones.

Moreover, Lasso is an easily interpretable model, with low complexity, selecting the most important features. Thus, for an industrial process, it would be the best model with no doubt.

As a conclusion, we recommend using Lasso Regression as it is our best overall model in performance. Considering runtime (computed on 252000 elements), it is also the best one. Hoeffding Adaptive tree is the slowest as expected. Indeed, as an evolutive version of Hoeffding Tree, it has more condition in the parameters optimizations and predictions and consequently takes more time to train.

Bibliography

Crowley, M. (2021). *Anomaly Detection Isolation Forests and Mondrian Forests.* Retrieved from Youtube: https://www.youtube.com/watch?v=XAkXUSxJNlM

Ginni. (2021, November 25). *What is Hoeffding Tree Algorithm.* Retrieved from Tutorialpoints: https://www.tutorialspoint.com/what-is-hoeffding-tree-algorithm

River Documentation. (2023). Retrieved from https://riverml.xyz

Saini, A. (2023, Semptember 21). *AdaBoost Algorithm: Understand, Implement and Master AdaBoost.* Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/