BIG HOMEWORK MERGING LAZY LEARNING AND FORMAL CONCEPT ANALYSIS

Takweh Cedric Fonguh
Data Science Master programme
HIGHER SCHOOL OF ECONOMICS
Moscow, Russia
cedricfonguh@gmail.com

December 13, 2022

1 INTRODUCTION

This homework serves as an entry point for students into the data science world. So it introduces students to three main topics:

1)Typical machine learning project: The pipeline of loading a dataset, feature engineering, designing a new predictive algorithm, results comparison;
2)Lazy learning:Predicting labels for small or rapidly changing data;
3)Rule-learning (on part with FCA): Viewing data as binary descriptions of objects (instead of points in a space of real numbers).

2 ABOUT THE DATASET

The context here is an Audit Risk Dataset for classifying Fraudulent Firms .The goal of the dataset is to help the auditors by building a classification model that can predict the fraudulent firm on the basis the present and historical risk factors. There are total 777 firms data from 46 different cities of a state that are listed by the auditors for targeting the next field-audit work.The information about the sectors and the counts of firms are listed respectively as Irrigation (114), Public Health (77), Buildings and Roads (82), Forest (70), Corporate (47), Animal Husbandry (95), Communication (1), Electrical (4), Land (5), Science and Technology (3), Tourism (1), Fisheries (41), Industries (37), Agriculture (200).

3 METHODS USED–

3.1 Loading and preparing the data

To prepare data for binarization, we first make sure that it is clean and organized. This involves removing any missing or incorrect values, and ensuring that the data is in a consistent format. Here we deleted the columns that hold no weight in the analysis and replace missing values with the average of the values.

3.2 Scaling(Binirization) –

Scaling, also known as binarization, is a preprocessing step used in machine learning to standardize the range and distribution of features in a dataset. This is done to ensure that all features are on the same scale, which can make it easier for the model to learn from the data.

3.3 Data shuffling

We shuffle the data to drop any initial ordering of rows.This is important because the model is trained on a sample of the data, and it is essential that this sample be representative of the entire population. If the data is not shuffled, there is a risk that the model will be trained on a non-representative sample, which can lead to poor performance on new, unseen data.

3.4 Data Representation

Representing data as a list of subsets of attributes can be useful in certain machine learning algorithms because it can help capture the underlying structure of the data. For example, if the data has a hierarchical structure, where certain attributes are more important than others, representing the data as a list of subsets of attributes can help to better capture this hierarchy. This can in turn improve the performance of the machine learning algorithm, because it can better learn the important relationships between the different attributes. To better suit the theory, we represent data X and y as a list of subsets of attributes.

3.5 Making Predictions and Analyzation

Here, we Make predictions and measure time required to obtain these predictions. In general, it is important to consider both accuracy and F1 score (or another relevant metric) when evaluating the performance of a predictive model, as they can provide complementary information about the model's ability to make correct predictions. Accuracy is a measure of how often the model makes correct predictions. It is calculated as the total number of correct predictions divided by the total number of predictions made. So, if a model makes 90 correct predictions out of 100 total predictions, its accuracy would be 90 percent.

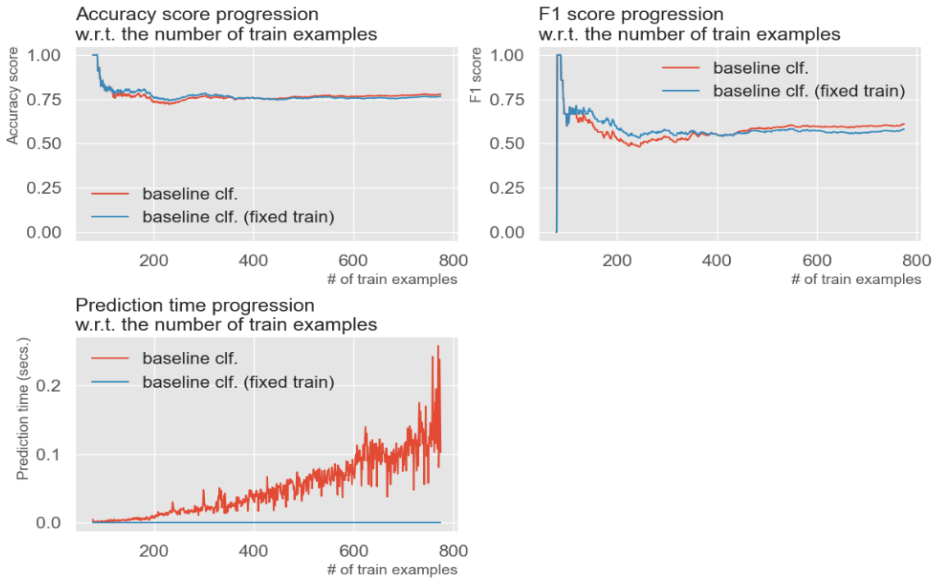
F1 score, on the other hand, is a measure of the balance between precision and recall. Precision is a measure of how many of the predicted positive examples are actually positive, while recall is a measure of how many of the actual positive examples were predicted as positive. The F1 score is the harmonic mean of precision and recall, and it is often used as a more comprehensive measure of prediction quality. Our F1 score is 57.8 percent while the accuracy score is 76.8 percent.

4 RESULTS–

project link is :https://github.com/DrGero20/OSDA-big-hw

`https://github.com/DrGero20/OSDA-big-hw`

4.1 FIGURES-



5 DISCUSSIONS and CONCLUSIONS

We carry out two popular rule - based models which include random forest and XGBoost.The two models out do the lazy learning in both runtime and prediction quality. XGBoost trains a model by fitting many decision trees to the data and then combining their predictions to make a final prediction. A random forest is trained by fitting many decision trees to the data and then combining their predictions to make a final prediction. This means that a random forest is a supervised learning algorithm, while lazy learning is an unsupervised learning algorithm. In conclusion, we realise that our Lazy learning algorithm when compared to other popular rule based models is overall slower and has a lower prediction quality.