5. Real-world data

This section extends the experimental results presented in [Chapter 3, Real-world data](#). We first report the ROC and AP curves for the five baseline prediction models on real-world data. We then illustrate the daily card precision with $k = 100$ using a decision tree of depth 2 as the prediction model.

Contents

5.1. Receiver Operating Characteristic curves

Fig. 1 reports the ROC curves for the five baseline classifiers (Logistic regression, decision trees with depth 2 and unlimited depth, random forest, and XGBoost) on real-world data. The curves are similar to those obtained on simulated data in the Section [Receiving Operating Characteristic (ROC) curve](#). Decision trees have the lowest performances, while logistic regression, random forest, and XGBoost have higher performances.
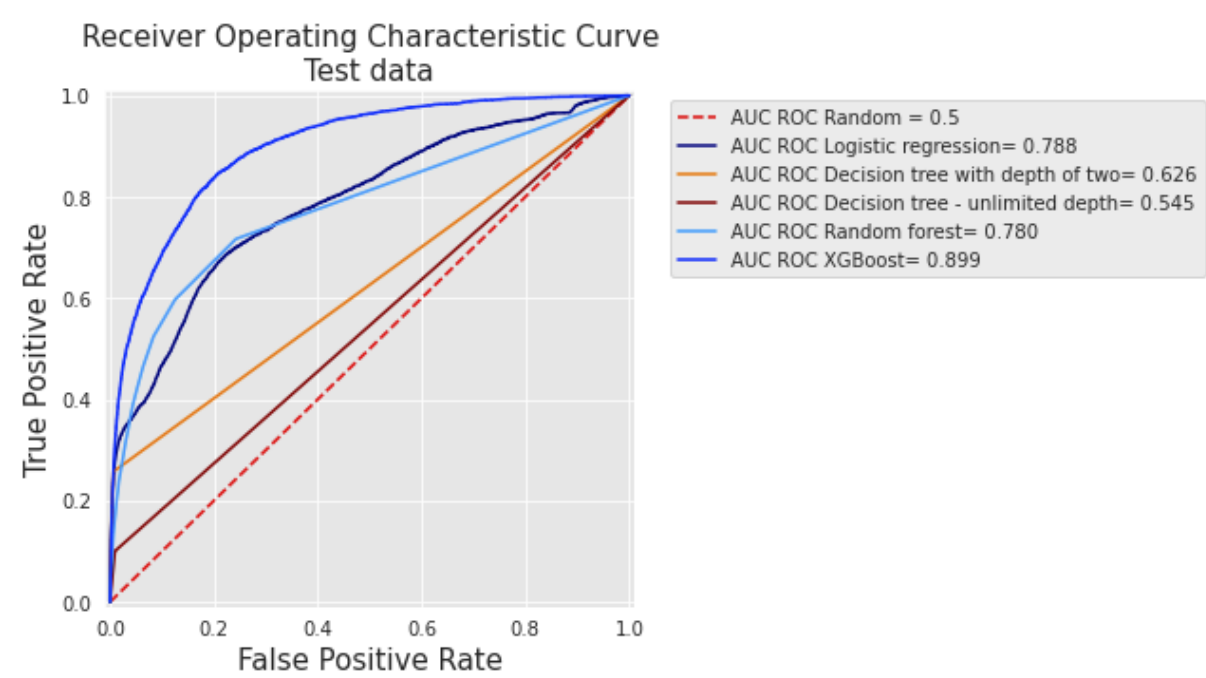

Fig. 1. ROC curves for the five baseline classifiers (Logistic regression, Decision trees with depth 2 and unlimited depth, random forest, and XGBoost) on real-world data.

It can be noted that, contrary to the results obtained with simulated data, XGBoost performs significantly better than random forest and logistic regression (AUC ROC of 0.899 versus AUC ROC of 0.78 and 0.788, respectively).

5.2. Precision-Recall curves

Fig. 2 reports the PR curves for the same five baseline classifiers on real-world data. The curves are at first sight quite different from those obtained on simulated data ([Precision-Recall curve](#)).
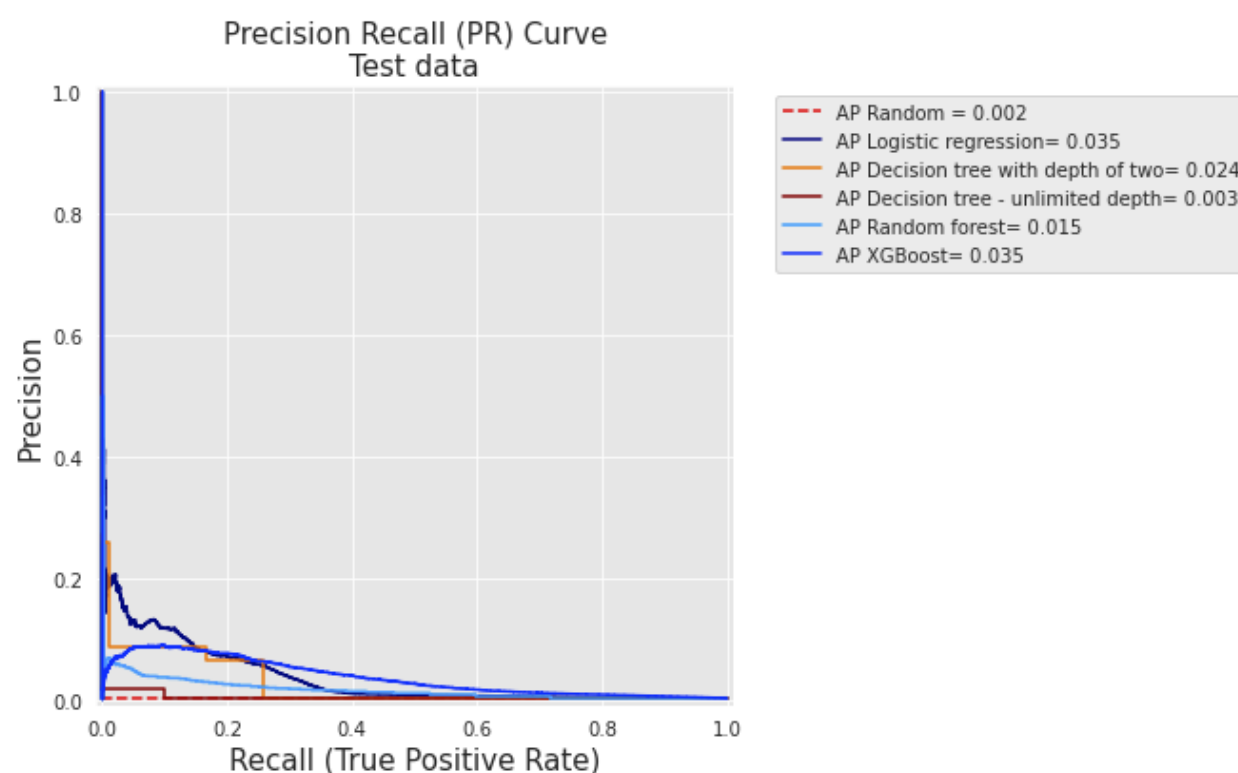
Fig. 2. PR curves for the five baseline classifiers (Logistic regression, Decision trees with depth 2 and unlimited depth, random forest, and XGBoost) on real-world data.

These differences are caused by the performances in terms of average precision which are, overall, much lower than with simulated data. As a result, all curves are shrunk to lower precision values. The reason is twofold. First, the proportion of frauds is lower in the real-world dataset. Second, fraudulent transactions are also harder to detect, which results in more false positives as the detection threshold is increased.

Nonetheless, the relative performances of the different classifiers are similar to those obtained with simulated data. In particular, logistic regression and XGBoost have much higher performances than decision trees. It is worth noting that the performance of the random forest in terms of AP is however poor, and lower than that of a decision tree of depth two. This is the result of overfitting, an issue that will be addressed with model selection in the Chapter 5.

5.3. Card Precision Top-k

We finally report, as in Precision top-k metrics, the number of daily detected frauded cards using real-world data. The results are provided in Fig. 3, for a decision tree of depth 2.
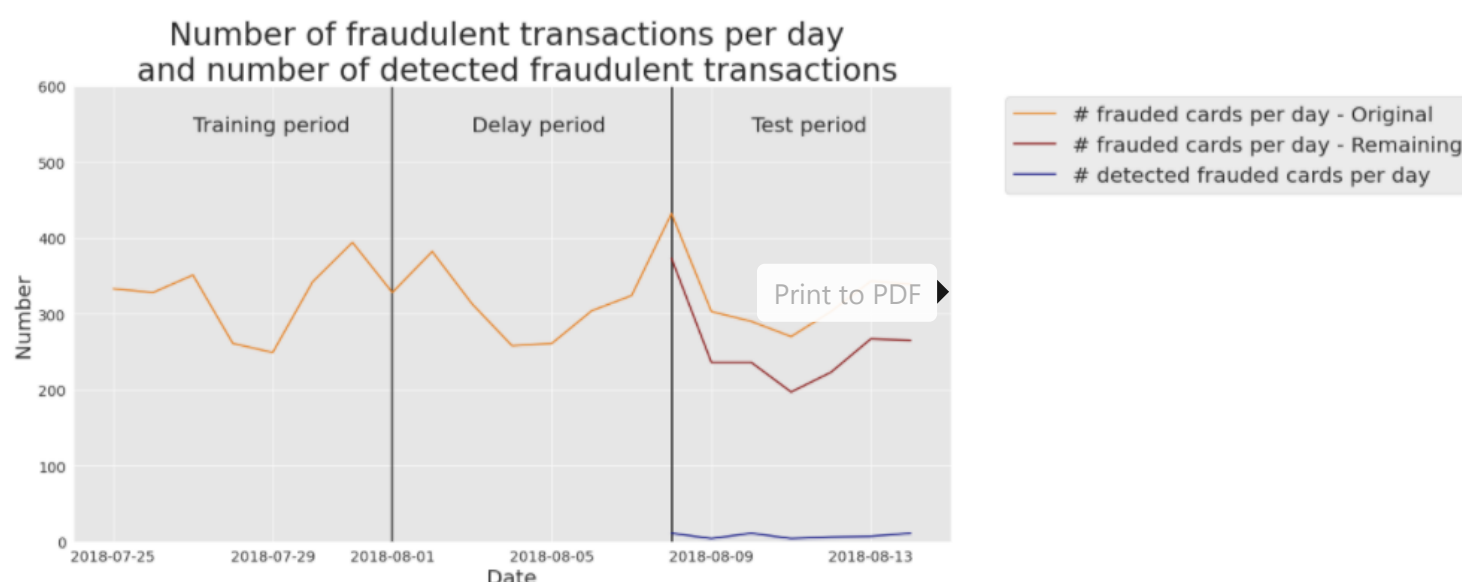

Fig. 3. Number of daily detected frauded cards with real-data, using a decision tree of depth 2.

The real-world dataset contains much more transactions (around 400000 daily transactions) and as a result, the number of daily frauded cards is also higher (ranging between 280 and 420, with an average of around three hundred. Yellow line). Once the known frauded cards are removed, the number of cards that remain to detect ranges between 200 and 350 (Red line). Let us recall that the CP@100 for a decision tree of depth two was 0.1 (Baseline fraud detection system), that is, on average, 10 cards out of the 100 most suspicious are true positives. This is illustrated with the blue line, which represents the number of frauded cards that are effectively detected every day.

Note that the maximum number of correctly detected frauded cards is upper bounded by k. That is, a maximum number of 100 frauded cards may only be detected every day when $k = 100$.