BIG DATA MANAGEMENT SYSTEMS & TOOLS

FRAUD DETECTION REPORT

TABLE OF CONTENTS

INTRODUCTION

OVERVIEW

Two factors played an important role in picking a topic for our project. Firstly, I looked for topics relevant to our day-to-day lives. Secondly, I intended to find a fascinating data challenge that would be unique, while leveraging the knowledge built from the Big Data Management Systems & Tools course. After some exploration, I settled on building a model to detect fraudulent credit card transactions.
Consumer credit cards are a vital part of the economic infrastructure in the modern world and are ubiquitous in usage in most countries. In the United States alone, over $3 trillion dollars moved through credit cards in 2021, with over 72% of adults having at least one credit card.

Fraudulent transactions pose a risk to both customers and credit card companies, and their impact on the business is significant. Fraud across all customer cards worldwide is estimated to be 6.81 cents per $100 in total volume in 2020, with a total of $10.24 billion fraud losses in 2020 for the United States alone. As such, credit card companies are faced with the challenge of processing and identifying transactions as valid or fraudulent in real time to provide the best customer experience and maintain their reputation.

Furthermore, I've also experienced situations where valid transactions were incorrectly identified as fraudulent by credit card companies. The process of validating those transactions, making the purchase, and sometimes even replacing the credit card that was incorrectly canceled by the company is also tedious and leads to negative customer experience.
With those experiences in mind, I am keenly aware of the importance of fraud detection and the negative business impact of both false positives and false negatives within the predictive algorithm.

This thinking made me curious to try resolving the problem for myself and explore a dataset where predicting credit card transactions as fraudulent or valid would be possible. This report details our experience and outcomes.

DATASET

A dataset of European cardholders' credit card transactions is available through Kaggle, and I determined that it was suitable for our purposes. The transactions spanned over two days of September 2013. There were 492 fraudulent transactions out of 284,807 transactions in the dataset. As expected, fraudulent transactions only occur a small percentage of the time, for a total of 0.172% of all transactions, meaning that our dataset was highly imbalanced. This presented us with a technical challenge that is realistic to a real-world scenario and different enough from the challenges faced in class assignments, finding our problem to be appropriate for the purpose of this group assignment and report.

The dataset was obtained here in November 2022. The data itself was collected and put together during a research collaboration of Worldline and the Machine Learning Group from the Université Libre de Bruxelles on big data mining and fraud detection.

TECHNICAL APPROACH

To effectively use the tools and techniques introduced in this course, I built pipelines and improved our models using pySparkML package. I believe this challenge to be sufficiently different from the material covered in the course, given that I am exploring the topic of classification in a deeper and more thorough way.

My goal was to build a machine learning model that properly identifies fraudulent credit card transactions.
To do so, I had to understand the features available in the dataset and derive any additional features that would improve the strength of our model. My hypothesis was that I could be effective at classifying fraudulent transactions based on this dataset by exploring different classification algorithms and metrics available in the pySparkML package.

DATA EXPLORATION

For confidentiality reasons, this dataset contained only numerical variables which are the result of PCA transformation, and the original features could not be provided, nor did I have any additional background information about the data.

- The features V1, V2, through to V28 are the principal components provided in this dataset
- The feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset
- The feature 'Amount' is the transaction's dollar amount
- The feature 'Class' is the response variable, and it takes value 1 in case of fraud and 0 otherwise. The percentage of fraudulent transactions in this dataset was 0.172%, which is highly imbalanced

EXPLORATORY ANALYSIS

As mentioned, features V1 through V28 provide the principal components of the dataset without providing the specific context for privacy and user data protection reasons. We can see from the Features Distribution graph below (image 1), that each feature presents a unique and reasonable distribution that is intuitively acceptable for machine learning modeling.

The *Time* variable shows dates and times over the span of two days, and was not used in our models, given the short time span. The *Class* variable is binary, with fraudulent transactions given the value of 1 and non-fraudulent transactions the value of 0. The *Amount* variable shows an interesting distribution that merits further exploration.

Features Distribution

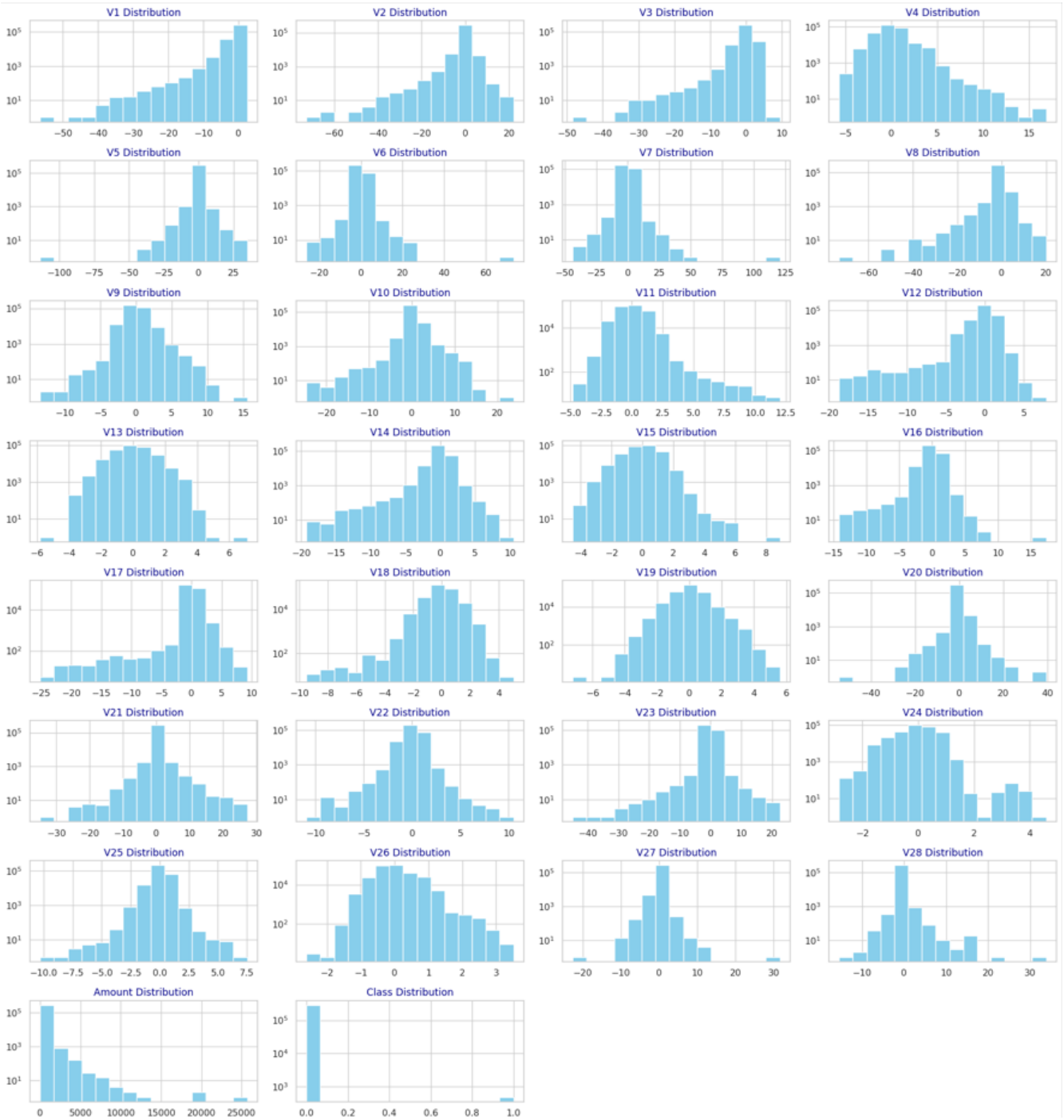

image 1

Looking at the box plot of the *Amount* feature for cases identified as fraudulent (image 2), we can see that the data skews right, with some clear outliers for higher amounts, meaning that the transaction dollar amount for fraudulent transactions is usually low. The maximum fraud amount is USD 2,125, while 95% of fraudulent transactions involve amounts below USD 641. This could be because lower amounts are less likely to get flagged as fraudulent, so fraudsters make multiple lower amount transactions, as opposed to one big fraudulent transaction. The extent to which the transaction amount will have predictive value for this classification is an interesting consideration.

Box Plot of Fraudulent Transactions by Amount

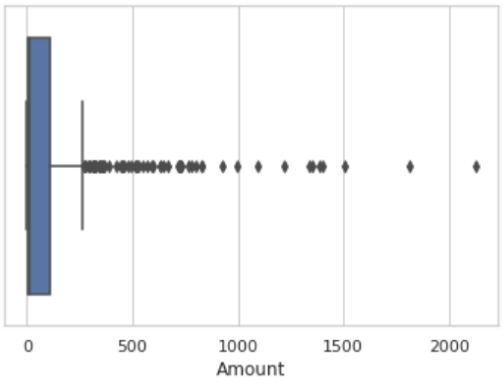

image 2

Finally, I generated a correlation matrix of all features (image 3). Some features had relatively stronger correlations with the *Class* feature, notably *V12*, *V14* and *V17*. The knowledge we have gathered from exploration data analysis built the foundation for our predictive modeling.

Correlation Matrix

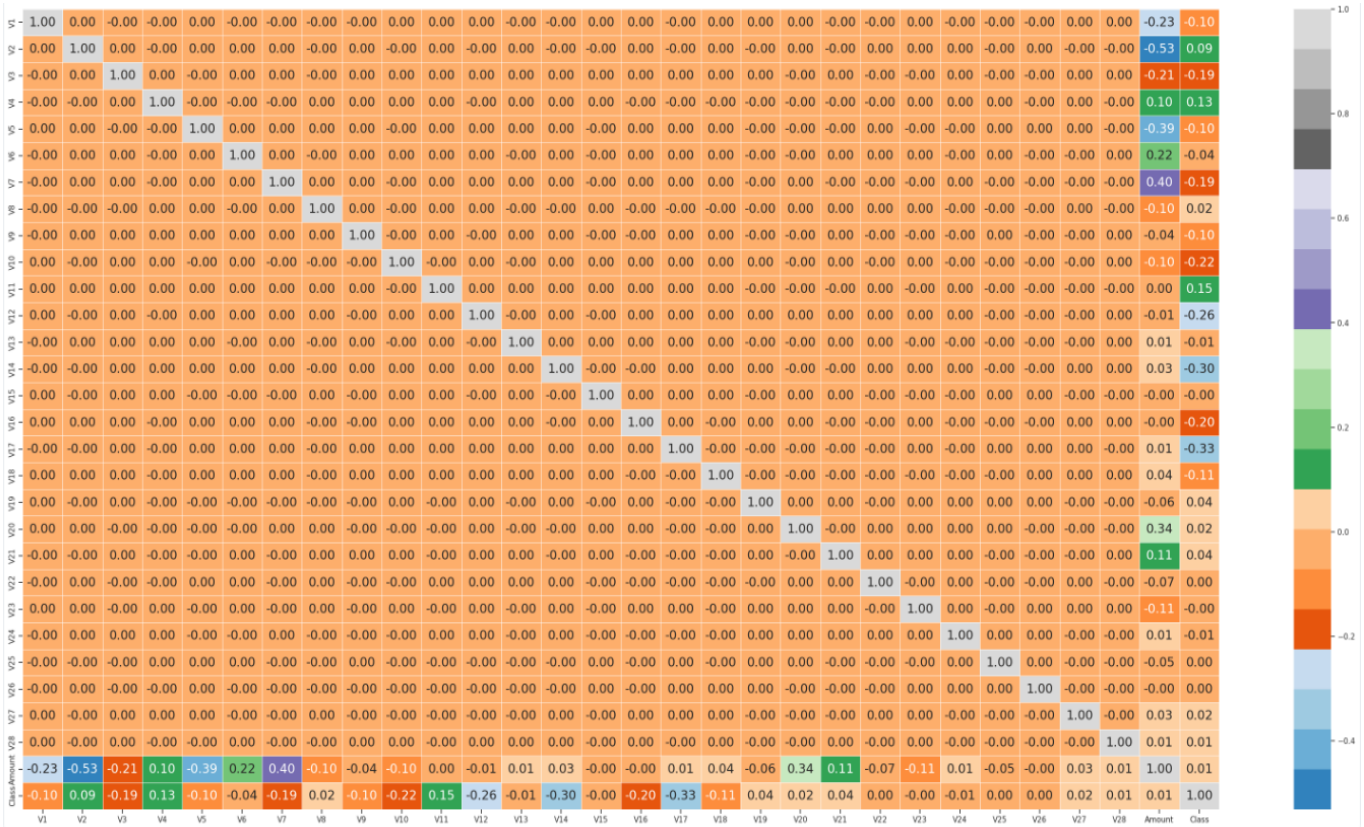

image 3

DATA PREPARATION

DATA CLEANING

I deemed the data cleaning step of the process done, after reducing the number of features to only include the relevant ones, checking for missing values and making sure all our data was numerical. To do so, I started with an explorative data analysis to better understand the features available, before removing the field 'Time' since it was not one of the features, we were interested in exploring, and since it did not seem relevant to our model. There were no missing values in our dataset, and all features were numerical. The data was split into a training and a test set. The training set, which consisted of 75% of the data, was used to build our classification models. Then, the classification models were run against the test set, which consisted of the remainder of the data, to assess the model's strength.

DATA BALANCING

I found the dataset to be highly imbalanced, with 577 non-fraudulent transactions for every 1 fraudulent transaction. While this represents a hopeful reality with only a small minority of transactions being fraudulent, the machine learning problem would be greatly challenged by using this dataset as is. To move forward, we used the oversampling method to correct for the data imbalance, which consists of synthetically duplicating fraudulent transactions, to increase the counts.

An alternative solution to handle imbalanced datasets is called undersampling, which consists of reducing the number of non-fraudulent transactions to improve the ratio of fraudulent to non-fraudulent transactions. I did not use this method, since our dataset had a small number of fraudulent transactions (577) and undersampling would have left us with a very small dataset, resulting in the loss of a large portion of information. A dataset of ~1000 transactions would have led to weak models, which is why oversampling was preferred for this use case. Oversampling was only applied to the training set, used to build our classification models.

VECTORIZING

The data was prepared by pulling all the features into one column of type vector using the MLLib library in Spark. This made it easy to embed a prediction right in a DataFrame and made it very clear as to what was getting passed into our models and what wasn't. This also made it easy to incrementally add new features, simply by adding to the vector.

PREDICTIVE MODELING

MODEL DESIGN

I built multiple classification models to detect fraudulent transactions. To assess the models' strength, I compared their accuracy, precision and recall scores. Below are the classification techniques used and their definition.

Logistic Regression Model: is a regression algorithm that takes the log-odds of a linear combination of the available variables. A binary logistic regression was an option for this classification problem for cases where "1" labeled fraudulent cases and "0" labeled the non-fraudulent cases.

Random Forest: is a classification algorithm which consists of many decision trees. It uses feature randomness to build each individual tree, and the amalgamation of trees creates a forest with stronger predictive power than any individual tree.

Gradient Boosted Tree: is an ensemble learning technique that uses a collection of weaker models to come up with a strong classifier model.

MODEL EVALUATION

Ideally, a model would correctly predict all non-fraudulent transactions as non-fraudulent and would flag all fraudulent transactions as such. However, in reality, these behaviors are complex and there is a trade-off between correctly predicting fraudulent transactions and missing fraudulent transactions.

As mentioned earlier in the overview section of our report, there is a loyalty, reputational and monetary cost to wrongfully flagging non-fraudulent credit card transactions as fraudulent or missing to flag fraudulent transactions on a client's account. Therefore, when building a predictive classification model, we need to be mindful of this trade-off. Metrics used to compare models are defined as follows:

Accuracy: is the proportion of overall predictions that the model classified correctly.

Precision: is the proportion of positive instances that were actually correct.

Recall: It is the proportion of actual positive instances that were identified correctly.

Area Under the Receiver Operating Characteristic Curve (ROC_AUC): tells us how good at ranking predictions the model is. It tells us the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance.

PR_AUC: It is the area under the precision recall curve. The curve helps us to see the precision and -recall tradeoff across different threshold values. The PR_AUC score summarizes that trade-off for easy comparison with other classifiers.

F1 score: F1 is a measure of models accuracy on a dataset. It is used to evaluate binary classification systems. It is defined as the harmonic mean of the model's precision and recall.

Confusion matrices were also created to visualize the accuracy of each model.

A confusion matrix is a table that is used to describe the performance of the classifier on the test set. It helps us analyze the correct and incorrect predictions of the classification models.

The left diagonal indicates the number of correct predictions made by the model. The observations on the upper left corner of the array indicates the number of non-fraudulent transactions classified correctly. The observations on bottom right of the array represent the number of fraudulent transactions classified correctly. The observations on the right diagonal represent the incorrect predictions made by the model. The upper right cell indicates the number of non-fraudulent transactions identified incorrectly and the bottom left corner represents the number of fraudulent transactions identified incorrectly.

Logistic Regression Confusion Matrix

prediction	0.0	1.0
Class		
0	70560	52
1	19	102

Random Forest Confusion Matrix

prediction	0.0	1.0
Class		
0	70590	22
1	21	100

Gradient Boost Confusion Matrix

prediction	0.0	1.0
Class		
0	70565	47
1	29	92

image 4

ANALYSIS

The table below summarizes the performance of the three models on different metrics:

Metric	Logistic Regression Model	Random Forest Model	Gradient Boosted Model
Recall	84.3%	82.6%	76.0%
Precision	66.2%	82.0%	66.2%
F1	74.2%	82.3%	70.7%
PR_AUC	76.7%	84.0%	77.1%
ROC_AUC	97.6%	97.8%	94.3%

Recall and precision are relevant metrics when determining the performance of a fraud detection model. Accuracy is a metric that describes how well a model performs across all classes. It informs us about how many predictions were correct out of the total number of predictions.

Since our main objective is to predict fraudulent transactions we will focus on precision, recall and F1 scores. Recall measures how many transactions were correctly identified as fraudulent out of all the actual fraudulent transactions. Precision measures how many transactions were actually fraudulent out of all the fraudulent transactions predicted. F1 score is a measure of model accuracy and it is the harmonic mean of both recall and precision. Thus it concerns only the positive class.

The table above shows that the logistic regression model has the highest recall but quite low precision leading to low F1 score. The random forest model has the highest precision score and a reasonably high recall value leading to high F1 score. The ROC_AUC of both the logistic regression and the random forest models are almost similar but the random forest model has the highest PR_AUC. Thus, the best model for credit card fraud detection is the random forest model because it has reasonably high values for both precision and recall. Improving recall for fraud detection means that we are detecting more fraudulent transactions, which is the best scenario for our model and business objectives. Although there is a cost to wrongfully identifying non-fraudulent transactions as fraudulent, I believe, that the reputational and financial cost of not detecting fraudulent activities is more detrimental to a financial institution, than wrongfully flagging non-fraudulent as fraudulent. And therefore, I suggest erring on the side of caution, by picking a model that maximizes recall, while still having a decent precision score.

LIMITATIONS

The limitation of this dataset is that no additional background information was provided about the features. All the features were in principal component form. Thus, we missed lots of information about the features that could have enabled us to determine good predictors of fraud. To build a stronger model, we could transform the features, perform more hyper-parameter tuning and try other classification models.

CONCLUSIONS

The goal of my project was to predict fraudulent credit card transactions using Spark's MLLib. I used a dataset of European cardholders' credit card transactions made over two days in 2013. We leveraged various classification techniques to predict fraudulent activities and measured various metrics to determine the models' performances, including recall and precision. Credit card companies need to have reasonably high recall and precision. If a financial institution uses a model with low precision, many legitimate clients might be blocked from making transactions that were incorrectly identified as fraudulent. This mistake may have reputation, loyalty, and monetary repercussions on the financial institution.

On the other hand, if recall is low then the company may miss many fraudulent transactions and incur the costs of settling them with vendors and clients. Thus, choosing the right predictive algorithm depends on business policy and on strategically picking the appropriate trade-off threshold between blocking non-fraudulent transactions, and

missing fraudulent ones. Our random forest model was the strongest model built, with both high precision and high recall.

In conclusion, the threat of fraudulent activities isn't going anywhere and creating strongly predictive fraud detections models is extremely important for financial institutions and their clients. With fraudsters becoming more and more imaginative, financial institutions need to heavily rely on machine learning and the advancements in big data management systems and tools to face these threats.

ACKNOWLEDGEMENTS

Background research sources included:

- Nilson Report. Payment Cards in the U.S. Projected. Published Oct 31, 2022. Accessed Nov 29, 2022.
- Gabrielle, Natasha. Credit and Debit Card Market Share by Network and Issuer. The Ascent. Published April 19, 2022. Accessed Nov 29, 2022.
- Nilson Report. Card Fraud Losses Dip to $28.58 Billion. Published Dec 7, 2021. Accessed Nov 29, 2022.
- Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015
- Dal Pozzolo, Andrea; Caelen, Olivier; Le Borgne, Yann-Ael; Waterschoot, Serge; Bontempi, Gianluca. Learned lessons in credit card fraud detection from a practitioner perspective, Expert systems with applications,41,10,4915-4928,2014, Pergamon
- Dal Pozzolo, Andrea; Boracchi, Giacomo; Caelen, Olivier; Alippi, Cesare; Bontempi, Gianluca. Credit card fraud detection: a realistic modeling and a novel learning strategy, IEEE transactions on neural networks and learning systems,29,8,3784-3797,2018,IEEE
- Dal Pozzolo, Andrea Adaptive Machine learning for credit card fraud detection ULB MLG PhD thesis (supervised by G. Bontempi)
- Carcillo, Fabrizio; Dal Pozzolo, Andrea; Le Borgne, Yann-Aël; Caelen, Olivier; Mazzer, Yannis; Bontempi, Gianluca. Scarff: a scalable framework for streaming credit card fraud detection with Spark, Information fusion,41, 182-194,2018,Elsevier
- Carcillo, Fabrizio; Le Borgne, Yann-Aël; Caelen, Olivier; Bontempi, Gianluca. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization, International Journal of Data Science and Analytics, 5,4,285-300,2018,Springer International Publishing
- Bertrand Lebichot, Yann-Aël Le Borgne, Liyun He, Frederic Oblé, Gianluca Bontempi Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection, INNSBDDL 2019: Recent Advances in Big Data and Deep Learning, pp 78-88, 2019
- Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Frederic Oblé, Gianluca Bontempi Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection Information Sciences, 2019
- Yann-Aël Le Borgne, Gianluca Bontempi Reproducible machine Learning for Credit Card Fraud Detection - Practical Handbook
- Bertrand Lebichot, Gianmarco Paldino, Wissam Siblini, Liyun He, Frederic Oblé, Gianluca Bontempi Incremental learning strategies for credit cards fraud detection, IInternational Journal of Data Science and Analytics