CAIO OLIVEIRA QUINAMO
DATA SCIENTIST

ANOMALY DETECTION IN FINANCIAL TRANSACTIONS FOR FRAUD IDENTIFICATION USING AUTOENCODERS AND RANDOM FOREST

CAIO QUINAMO
ANALYTICS SOLUTIONS
MACHINE LEARNING &
DATA SCIENCE FOR BUSINESS IMPACT

DOCUMENTATION OF THE TECHNICAL REPORT

INDEX

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

0. INTRODUCTION

In recent years, online shopping has experienced exponential growth. This new consumption pattern, now fully integrated into the modern market, has driven the emergence of digital business models. Large companies, in response to this shift, have been forced to migrate to digital channels, leaving behind their exclusive reliance on traditional means.

Globally, e-commerce sales increased from $1.34 trillion in 2014 to an estimated $6.33 trillion in 2024, with projections reaching $8.03 trillion by 2027. This represents a nearly sixfold increase in just thirteen years. In 2023, digital wallets accounted for 49% of online transactions, followed by credit cards at 21%. Revenues generated from digital payments are expected to reach $14.79 trillion by 2027. Currently, two-thirds of the global adult population uses this type of payment, with a 95% adoption rate in developed countries.

In the United States, online sales rose from $79.02 billion in Q4 2014 to $308.91 billion in the same quarter of 2024. In 2023, 69% of online adults in the U.S. had used some form of digital payment in the past three months. The most popular platforms include PayPal (40%), Apple Pay (24%), and Venmo (16%). A total of 89% of Americans use digital payments, and 62% use at least two different methods.

This digital boom has also created space for online fraud to grow as a parallel business fueled by cybercriminals. While cybersecurity has advanced, attackers have refined their techniques to bypass detection systems. Among the most common methods are carding (stolen card usage), phishing (data capture on fraudulent websites), synthetic identity fraud (combining real and fictitious data), chargeback fraud (false refund claims after receiving goods), and account takeover.

Fraud in online transactions poses a significant risk for both users and businesses. Consumers may suffer financial loss, identity theft, and reduced trust in the digital ecosystem. Businesses face consequences such as financial losses, reputational damage, increased operational costs due to added security, and customer attrition. As digital transactions increase, so do the opportunities for malicious actors, demanding increasingly sophisticated solutions for fraud detection and prevention.

This project adopts a didactic and representative approach. A transaction history will be analyzed to identify patterns that characterize fraudulent operations, supported by explanatory visualizations and a Machine Learning model capable of highlighting the key factors for classification. The primary goal is to detect as many frauds as possible while minimizing false positives.

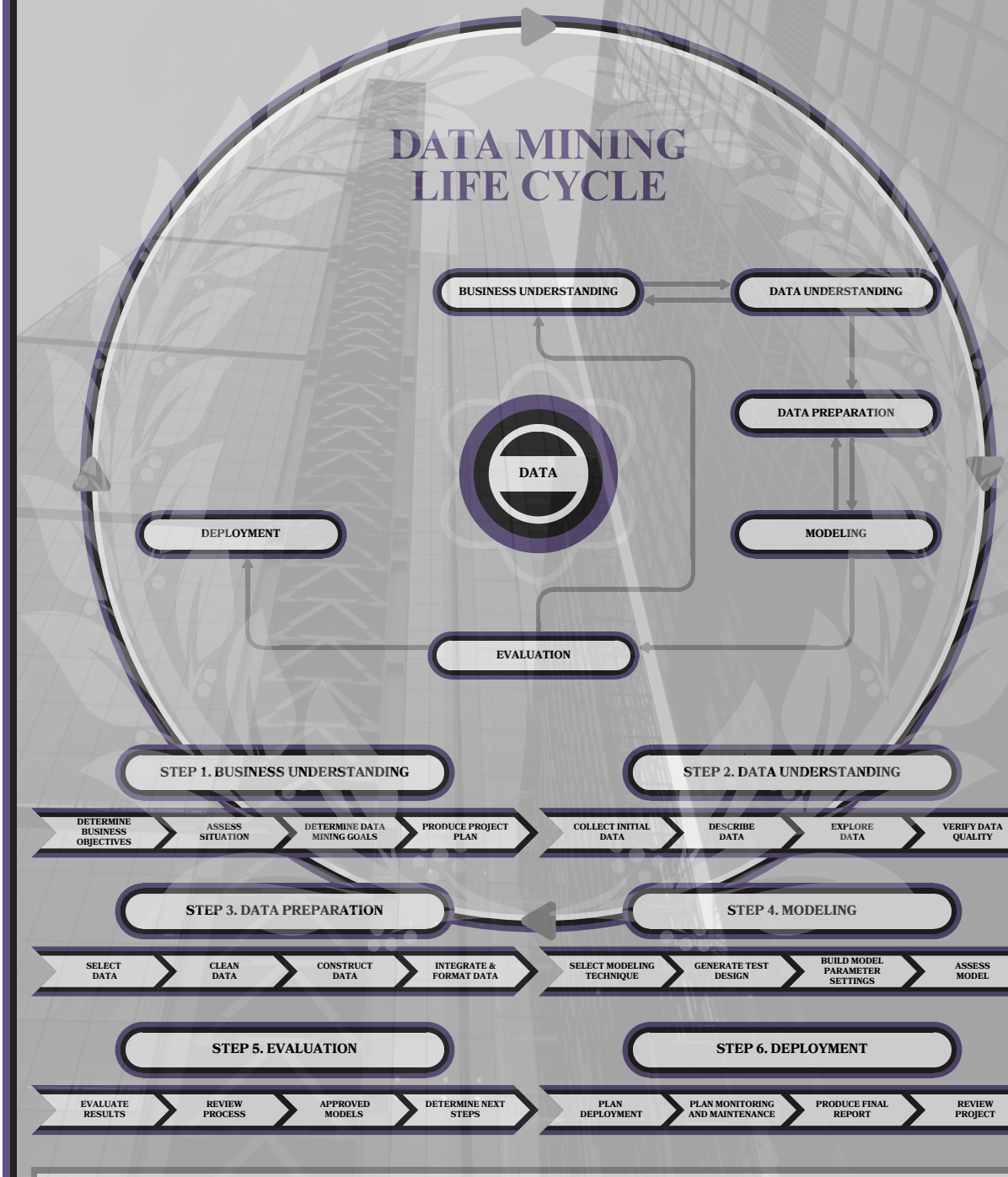
The entire process will be developed following the CRISP-DM methodology, which structures the work into six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The last phase is outside the scope of this project and will not be addressed.

At the core of the approach is an autoencoder trained exclusively on legitimate transactions (the majority class) to learn the latent representation of normal data. Since it is not exposed to fraud during training, it is expected to perform poorly in reconstructing anomalous transactions, thereby producing a high reconstruction error. This error will be used as a new feature to feed a supervised classification model based on Random Forest. This combination aims to unite the strengths of unsupervised learning (anomaly detection via the autoencoder) with the robustness of a supervised classifier, thus improving the overall accuracy of the system.

However, one of the main challenges will be that certain legitimate transactions may exhibit patterns similar to fraudulent ones, potentially confusing the model. Therefore, a proper analysis of outliers and class imbalance (an inherent aspect of this kind of problem) will be critical to the system's performance.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

CRISP-DM METHODOLOGY



CRISP-DM (Cross Industry Standard Process for Data Mining) is a robust and widely recognised methodology for the development of data mining and data science projects. Designed to be industry-neutral, it offers a comprehensive and structured approach that guides practitioners through the entire lifecycle of a data-driven project. The framework is comprised of six iterative phases that ensure deep integration between analytical efforts and business objectives. By emphasising clarity, repeatability and adaptability, CRISP-DM facilitates effective decision-making, continuous improvement and long-term value creation from data assets.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

1. BUSINESS UNDERSTANDING

1.1 DETERMINE BUSINESS OBJECTIVES

Is it possible to detect fraudulent transactions in a digital payment system automatically and accurately, while at the same time minimizing the number of legitimate operations incorrectly flagged as fraud?

This is the central question that drives this project, and whose answer could make a substantial difference for small and medium-sized enterprises (SMEs) in today's digital ecosystem. Nowadays, SMEs operate in an environment where fraudulent electronic transactions are increasing in volume, complexity, and sophistication. Unlike large corporations, these businesses often lack advanced technological infrastructure and specialized personnel to implement effective fraud detection solutions. As a result, they tend to rely on external payment gateways or third-party services, which are typically based on generic models and may not adapt well to the specific operational reality of each business.

This context highlights an urgent need: to have access to an effective, affordable, and context-aware predictive tool that can support real-time decision-making about potential fraud, without harming legitimate users.

The business objectives are as follows: **Improve fraud detection accuracy:** Develop a system capable of automatically identifying fraudulent transactions with high precision, minimizing both false negatives and false positives. **Protect customer experience:** Reduce unjustified rejections of valid operations, helping to maintain customer trust in the platform and avoid losses due to commercial friction. **Minimize economic losses:** Lower the financial impact of undetected fraud, as well as indirect costs such as chargeback fees, penalties, and reputational damage. **Gain autonomy and competitiveness:** Build a solution tailored to the specific needs of the business, allowing for faster and more informed decisions without relying exclusively on generic third-party filters. **Ensure scalability and adaptability:** Implement a system that can grow alongside the business and adapt to emerging fraud patterns as digital threats evolve.

1.2 ASSESS SITUATION

This project has an academic and demonstrative purpose, so it will not work with real business data. Instead, a public dataset, widely referenced in digital fraud research, will be used. This dataset contains credit card transactions carried out by European cardholders over two days in September 2013. Although the data has been anonymized, it preserves representative patterns of the fraud detection problem and allows the simulation of a realistic environment.

A critical feature of this dataset is its severe class imbalance, where fraudulent transactions represent a tiny fraction of the total. This imbalance presents significant challenges in model development, as it requires strategies that preserve sensitivity without dramatically increasing the false positive rate. While this technical limitation complicates the classification task, it also offers a valuable opportunity to test robust approaches in adverse conditions.

Development will be carried out on a mid-to-high-range personal workstation, sufficient for exploratory analysis and lightweight model training. The chosen environment is Jupyter Notebooks with Python, leveraging standard libraries such as Pandas, Numpy, Scikit-Learn, Matplotlib, Seaborn, and TensorFlow/Keras.

Among the main limitations of the project are the absence of company-specific data, the lack of access to real-time transactions, and the exclusion of deployment in a production environment. However, a complete workflow will be simulated, following data science best practices, to demonstrate the viability of the proposed approach and its potential for real-world adaptation.

1.3 DETERMINE DATA MINING GOALS

From a data mining perspective, the primary objective of this project is to build a predictive system capable of identifying anomalous transactions in a digital card payment environment, maximizing the detection of actual frauds while minimizing the misclassification of legitimate transactions.

The solution must be robust enough to handle highly imbalanced class distributions, which are typical in financial fraud detection problems, and offer clear interpretability of the results to support its application in real-world contexts.

This means translating the business problem into specific data mining tasks, such as anomaly detection and binary classification, with a focus on optimizing metrics that better reflect system performance in imbalanced contexts, such as the area under the precision-recall curve (AUPRC) and the F1-score. Therefore, the priority is to develop a model that is not only technically accurate but also operationally functional: one that minimizes false positives that impact legitimate user experience, while maintaining high sensitivity to fraud patterns, even when these occur in marginal volumes.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

To achieve this objective, the data mining strategy will combine supervised and unsupervised learning techniques, allowing the capture of subtle patterns in transactional behavior. In addition, feature selection methods will be incorporated to improve model efficiency, reduce redundancy, and enhance generalization capacity. The evaluation process will be rigorous and focused on metrics appropriate to the problem, seeking an optimal balance between operational accuracy and risk of loss.

1.4 PRODUCE PROJECT PLAN

The project will follow the six phases defined by the CRISP-DM methodology, adapted to the specific context of fraud detection in digital transactions. Each stage will include concrete tasks and well-defined goals to ensure a rigorous and coherent approach.

In the business understanding phase, the impact of digital fraud will be analyzed, with a particular focus on the limitations faced by small and medium-sized enterprises. Based on this analysis, a central question will be defined to guide the entire data mining process. The data understanding phase will include an initial exploration of the dataset, identifying its structure, general patterns, outliers, and, most importantly, the degree of class imbalance, which will strongly influence later modeling decisions.

The data preparation phase will involve cleaning, transformation, and the creation of new relevant features, aiming to maximize data quality for modeling. Dataset partitioning strategies will be established in accordance with the techniques used, ensuring a clear separation between training and evaluation sets.

During the modeling phase, a hybrid system will be developed that combines anomaly detection and supervised classification techniques. Cross-validation and feature selection methods will be applied to optimize predictive performance and reduce the risk of overfitting.

In the evaluation phase, various performance metrics will be measured, including F1-score, AUPRC, and AUROC, with priority given to those most appropriate for imbalanced classification contexts. The model will be validated not only in quantitative terms, but also by assessing the practical impact of errors, especially the proportion of legitimate transactions incorrectly flagged as fraud.

Finally, although the deployment phase is not within the scope of this project, its role is recognized as essential in real-world applications. In practice, the integration of the model into operational systems, real-time monitoring, and adaptability to new fraud patterns are key aspects of its long-term effectiveness and sustainability.

Throughout the project, particular attention will be given to the design of the data workflow, with the goal of avoiding common pitfalls that can compromise model validity, such as data leakage. This phenomenon, which occurs when information from the test set inadvertently influences training, can lead to misleading performance metrics and is especially difficult to detect in workflows with multiple transformation stages.

To mitigate this risk, the dataset will be carefully partitioned into separate training, validation, and test sets, keeping the test set completely isolated until the final evaluation. All preprocessing steps and feature engineering operations will be carried out exclusively within the training and validation sets. This strategy ensures that the performance metrics obtained truly reflect the model's behavior on unseen data, thus safeguarding the integrity of the evaluation process.

Although the detailed workflow will be discussed in later sections, from this stage forward a strict data separation policy and process replicability will be established as fundamental principles, in line with professional best practices in data science.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

2. DATA UNDERSTANDING

2.1 COLLECT INITIAL DATA

Early detection of credit card fraud is essential to protect both users and financial institutions by preventing unauthorized charges and financial losses. For this project, a public dataset widely used in academic research on digital fraud detection has been employed.

The dataset contains 284,807 transactions carried out by European cardholders over two days in September 2013. Among these, only 492 transactions were classified as fraudulent (approximately 0.172% of the total), highlighting the strong class imbalance that is typical in fraud detection problems.

The predictive variables have been transformed using Principal Component Analysis (PCA) for anonymization purposes. This results in 28 principal components, labeled V1 through V28. Only two variables have not been transformed: **Time**, which measures the number of seconds elapsed since the first recorded transaction. **Amount**, which represents the monetary value of each operation. The target variable is Class, which takes the value 1 for fraudulent transactions and 0 for legitimate ones.

This dataset has been chosen for its representativeness in real-world fraud detection scenarios and its suitability for applying advanced modeling techniques in highly imbalanced contexts. Given this imbalance, performance evaluation will emphasize metrics such as AUPRC, which are more appropriate for this type of problem. All usage complies with the applicable licensing (**Open Database License**).
Reference dataset: **Credit Card Fraud Detection**.

2.2 DESCRIBE DATA

The dataset consists of 284,807 records and 31 columns, organized in a DataFrame. Most of the predictive variables are continuous, of type float64, and were transformed using Principal Component Analysis (PCA), resulting in 28 anonymized principal components (V1 to V28) with no direct interpretability. While the exact meaning of these variables is unknown, their statistical relationship with the target class can still be explored.

The three columns not transformed via PCA are:

- **Time:** Indicates the number of seconds since the first recorded transaction. Although temporal in nature, its direct interpretation is limited. Its relevance will be empirically evaluated during analysis and modeling.
- **Amount:** Represents the monetary value of the transaction. It is the only feature with an explicit, interpretable meaning and will be considered relevant for modeling, though it will be rescaled accordingly.
- **Class:** The binary target variable. It is 1 if the transaction is fraudulent, and 0 otherwise. This is the label the model aims to learn and predict.

The dataset has been pre-anonymized and formatted, reducing the need for preprocessing at this stage. The required transformations (mainly related to variable scaling) will be applied during later phases of the workflow. As such, the initial preprocessing does not present significant technical challenges, as the dataset is specifically designed for experimentation in fraud detection tasks.

2.3 EXPLORE DATA

2.3.1 Analysis of the target variable

The original dataset, without any prior modifications, includes 284,807 entries and 31 variables, with no missing values. This allows for immediate work with the data without the need for imputation techniques.

However, 1,081 duplicate records were identified: 1,062 belonging to the negative class (0) and 19 to the positive class (1). Given the strong class imbalance (with the positive class representing less than 0.2% of the total), the decision was made to remove only the duplicates from the negative class, while retaining those from the positive class.

This decision is justified by the need to preserve as much information as possible regarding fraudulent transactions, since their scarcity could limit the model's ability to learn representative fraud patterns. While these duplicates might stem from errors in real-time data capture, it is also possible that they reflect repeated fraud attempts by an attacker over a short period. Therefore, their presence could introduce noise, but also provide critical evidence of anomalous behavior. Given this potential informational value, it was decided to retain the positive duplicates during this analysis phase.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

2.3.2 Analysis of the Correlations with the target variable

To better understand the relationships between the independent variables and the target variable (Class), a correlation analysis was conducted. The five variables with the highest absolute correlation with the positive class are: V17, V14, V12, V10, and V16. All of them show significant negative correlations, suggesting that lower values of these features are associated with a higher likelihood of a transaction being fraudulent.

On the other hand, the variables Time and Amount, which are the only ones with explicit meaning (outside the PCA-transformed feature space), exhibit very low correlation with the target variable. Despite this low linear correlation, their individual behavior was analyzed in more detail, as they may contain relevant patterns not captured by simple correlation measures. The attached notebook explores these variables more thoroughly, from both a statistical and visual perspective, assessing their potential usefulness in the modeling pipeline.

2.3.3 Univariate analysis of «Time» and «Amount»

In terms of distribution, the Time variable displays a relatively symmetrical shape, while Amount shows strong positive skewness, with a skewness coefficient of 16.978, indicating a concentration of low values and a long tail toward high values. This raises a relevant methodological concern: **to what extent does skewness affect the other features in the dataset?** Features with pronounced skewness (either positive or negative) can pose a significant challenge for model performance, particularly in architectures sensitive to scale and distribution shape, such as autoencoders.

To address this issue, an initial preprocessing approach was explored using a Yeo-Johnson transformation and a quantile transformation, aimed at correcting skewness without requiring strictly positive data. Afterwards, a Min-Max scaling was applied to normalize features to the [0, 1] range. This strategy sought to improve model learning by reducing distortion caused by skewed distributions and by establishing a uniform scale.

However, after multiple experiments and evaluations, it was observed that applying Min-Max scaling exclusively to all features, and only applying a quantile transformation to the Amount variable (the one with the highest absolute skewness), yielded better results. This was evident both in terms of autoencoder reconstruction error and the classification metrics of the subsequent Random Forest model. These findings suggest that the autoencoder was able to adapt to the inherent skewness in most variables or, alternatively, that preserving the original shape of distributions helped retain relevant patterns for fraud detection.

The choice of MinMaxScaler over alternatives such as StandardScaler or RobustScaler was based on two main reasons:

1. Its ability to preserve the original distribution shape, which is critical in contexts where aggressive transformations may degrade model performance.

2. Its suitability for neural network models, since autoencoders are sensitive to input scale and benefit from working with normalized data within a fixed range.

Although Min-Max scaling is sensitive to extreme values, this study chose to retain outliers rather than clip or remove them. The rationale behind this decision is that outliers, far from being errors, may represent genuine anomalous behaviors, which are crucial for model learning. In anomaly detection tasks, removing rare patterns could undermine the model's ability to detect precisely those exceptional cases.

A comparative descriptive analysis was performed between legitimate and fraudulent transactions based on the Time and Amount variables. A noteworthy finding was that the mean value of fraudulent transactions is 38.23% higher than that of legitimate ones. However, 50% of fraudulent transactions have a value below €9.25, confirming the strong skewness of the Amount variable. Furthermore, a considerable difference was observed between the maximum values of each class: the non-fraudulent class (0) reaches up to €25,691, while the fraudulent class (1) does not exceed €2,126.

From a temporal perspective, the first fraudulent transaction occurred 406 seconds after the dataset's start, and the last one happened 2,444 seconds before its end. This suggests a non-uniform temporal distribution of fraud events. However, no clear time-based patterns were identified that could add value to the model.

Additionally, it was found that in both classes, the 99th percentile of Amount is significantly lower than the maximum value, indicating the presence of extreme outliers that disproportionately affect the mean, especially in the non-fraudulent class.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

Following experiments involving the inclusion of the Time variable as a model input, it was concluded that its presence degraded overall performance, and it was therefore excluded from the workflow. Its behavior appeared to introduce more noise than useful information for the classification task.

Although skewness coefficients were not used as a direct criterion for preprocessing decisions (except for the Amount variable), a complementary exploratory analysis was performed. This analysis revealed that approximately 60% of the variables exhibit extremely high skewness, four variables fall within a moderate skewness range, and the remainder show acceptable levels. The figure below illustrates these results along with the corresponding skewness coefficients for each variable.

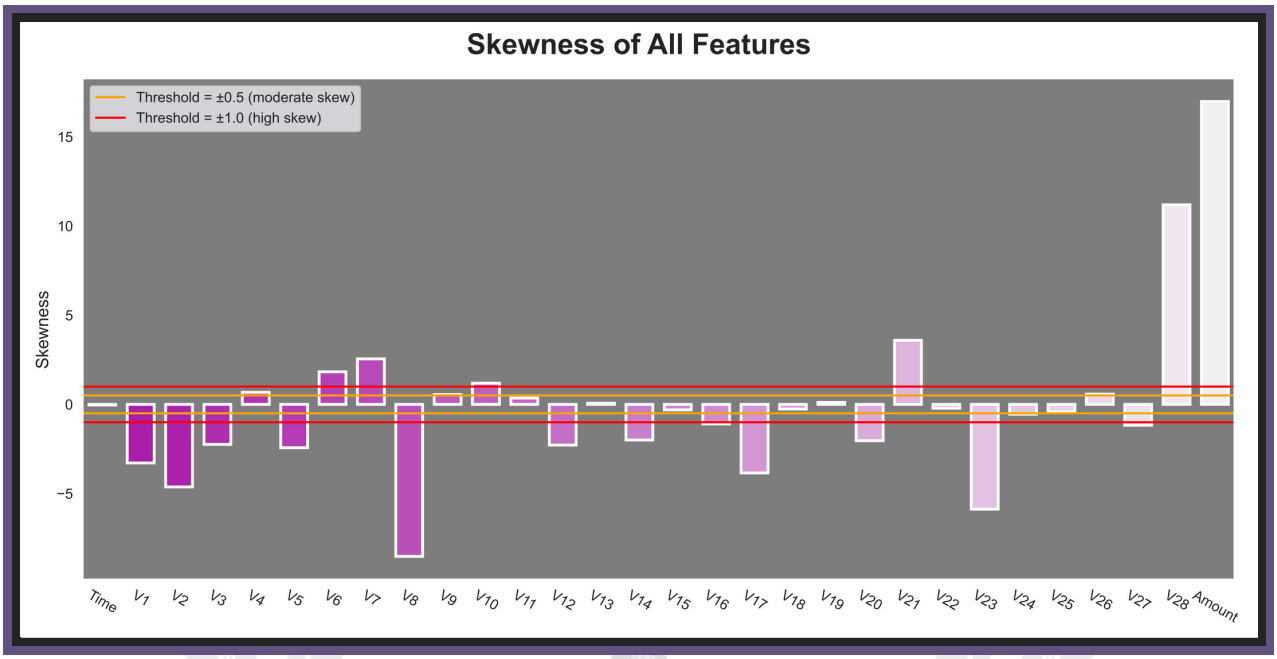


Figure 1. Skewness Coefficient of all features
This bar chart shows the skewness values of each feature in the dataset. Positive skewness indicates a longer right tail, while negative skewness indicates a longer left tail. Features with high absolute skewness may benefit from transformation to improve model performance.

2.3.4 Data visualization (before autoencoder reconstruction)

Given the large number of features in this dataset, it is not feasible to directly visualize their relationship with the target variable using simple 2D plots. The most appropriate alternative is to apply a dimensionality reduction technique that transforms the original feature space into two dimensions, thereby enabling a visual exploration of the dataset's internal structure.

Among the available techniques (such as PCA – Principal Component Analysis, UMAP – Uniform Manifold Approximation and Projection, and LLE – Locally Linear Embedding), t-SNE (t-distributed Stochastic Neighbor Embedding) was selected due to its superior ability to preserve the local structure of the data, even when projected into a low-dimensional space. Unlike PCA, which performs a linear transformation and tends to preserve global variance, t-SNE focuses on maintaining proximity between similar observations. This property is particularly relevant in anomaly detection or when analyzing minority classes, where small differences can be critical.

Although t-SNE is not suitable as a direct classification technique and does not allow reconstruction of the original variables, it is a powerful visual exploration tool that can reveal potential clusters, overlaps, or underlying structures. In this case, t-SNE was applied to a random sample of 10% of the legitimate transactions, while all available fraudulent cases were included. This strategy aims to reduce computational load without compromising the representativeness of the overall data distribution.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

The purpose of this visualization is to explore whether there is any natural separation or clustering between legitimate and fraudulent transactions. This information is valuable for justifying the use of more sophisticated techniques, such as autoencoders. In the resulting visualization, although a clearly defined separation between classes is not observed, regions with higher densities of fraudulent transactions can be identified, as well as more dispersed areas that may indicate subtle differences in the internal data structure. Nevertheless, many transactions from both classes appear overlapping and chaotically distributed, which motivated the design of an autoencoder aimed at learning a more discriminative internal representation of the dataset.

In this context, it becomes appropriate to generate the same visualization after training the autoencoder, using the learned latent space. This will allow for a visual inspection of the model's ability to improve class separation by transforming the internal structure of the data, thus offering an additional criterion to evaluate its effectiveness.

2.3.5 Univariate outlier detection via Z-Score standardization

From the exploratory analysis of the Time and Amount variables, interest arose in investigating the presence of outliers. Although outliers were ultimately not removed from the final model, this decision followed a series of systematic tests, including adjustments to Z-score thresholds to remove only the most extreme outliers, and experiments that retained certain lower-magnitude outliers within the non-fraudulent class. No interventions were made on outliers in the fraudulent class, as removing observations from the minority class could significantly compromise the model's representativeness.

The results showed that excluding outliers increased the sensitivity of the combined model (Autoencoder + Random Forest), particularly in the presence of data with atypical behavior. Specifically, when the model was trained solely on "typical" data, a significant increase in false positives was observed when applied to legitimate but atypical transactions. This suggests a loss in generalization capacity to the legitimate variability found in real-world settings. While retaining outliers may slightly increase the number of false negatives, it leads to a substantial reduction in false positives, which is preferable in a context where false positives may carry a high operational cost.

Moreover, training the autoencoder without outliers resulted in a more distinct latent separation between classes, although at the expense of reduced coverage of the actual non-fraudulent behavior spectrum. This trade-off between structural separation and generalization led to the decision to retain moderate outliers within the training set.

The outlier analysis was mainly aimed at identifying characteristic patterns within the positive class. Since fraudulent transactions aim to mimic legitimate ones while often exhibiting localized deviations, it was hypothesized that many of them present at least one univariate outlier.

To detect these, Z-score standardization was applied, computing the distance of each observation from the variable's mean, normalized by its standard deviation. A threshold of ±3 Z-score units was set, classifying values exceeding this limit as outliers. However, removing all transactions containing at least one univariate outlier would have resulted in excessive loss of legitimate data, while also introducing significant sampling bias. This limitation stems from the univariate approach, which analyzes each feature independently and ignores inter-feature correlations.

To facilitate visual analysis of these outliers, a dimensionality reduction technique using PCA (Principal Component Analysis) was applied. PCA was chosen for its ability to preserve the global variance of the dataset. Unlike methods like t-SNE, which prioritize local structure, PCA provides a representation consistent with the original statistical distances, and is also more computationally efficient.

The most fraud-predictive variables (V14, V17, V10, V16, V3, and V12), previously identified based on their high correlation with the target variable, showed a marked concentration of outlier values within the fraudulent class. This behavior suggests that fraudulent transactions tend to deviate from normal statistical patterns, generating anomalies in key features. At the same time, these variables also exhibited a moderate number of outliers in the non-fraudulent class, indicating that not all outliers are fraudulent, but that many frauds involve outliers. This duality justifies the use of these variables not only as direct predictors, but also for feature engineering, such as creating binary outlier indicators or extreme Z-score count features, while ensuring that legitimate transactions with valid atypical behavior are not excluded.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

2.3.6 Multivariate outlier detection via Mahalanobis distance

Following the univariate detection of outliers using the Z-Score, a complementary approach was implemented to identify multivariate outliers through Mahalanobis distance. Unlike univariate analysis, this method captures relationships between multiple variables by considering their covariance. It evaluates how far an observation lies from the multivariate center of the dataset.

The choice of this technique lies in its ability to identify observations that, while not standing out in any single variable, are atypical when variables are considered jointly. In the context of financial fraud, this is especially useful: fraudulent transactions may appear normal on each individual variable but deviate when their combined interaction is considered.

To establish a decision threshold, a percentile-based criterion was adopted—specifically, the value corresponding to the 99.9th percentile of Mahalanobis distances calculated on the Z-Score standardized data. This threshold focuses detection on the most extreme cases (approximately 0.1% of the observations), balancing sensitivity to anomalies without excessively compromising the volume of legitimate data. This is crucial in imbalanced datasets, where overly removing non-fraudulent but unusual observations could introduce bias.

It is important to note that, in both univariate and multivariate detection, the identification of outliers is largely subjective. It depends on the context, the objectives of the analysis, and the analyst's criteria. The threshold selection should be adapted to the specific use case, considering its impact on model interpretation and generalization. A stricter threshold may favor the detection of high-impact fraud, while a more lenient one may be preferable in environments prioritizing the reduction of false positives.

Among the main strengths of Mahalanobis distance are:

- Its consideration of the multivariate structure of the dataset, integrating correlations between variables.
- Its statistically sound nature under the assumption of multivariate normality.
- Its computational simplicity and reasonable scalability in medium-sized datasets.

Nonetheless, it has limitations. It is sensitive to the presence of extreme values (though this was mitigated through standardization) and depends on a well-conditioned covariance matrix, which can be problematic in data with high multicollinearity or many dimensions.

Other techniques for multivariate outlier detection exist, such as Isolation Forest, a tree-based method that does not require distributional assumptions and performs well with large data volumes; One-Class SVM, useful when only a single majority class is available and the goal is to model its boundary; and LOF (Local Outlier Factor), which measures the local density of each observation compared to its neighbors, making it suitable for detecting complex structures. However, for this analysis, Mahalanobis was chosen due to its statistical interpretability, its consistency with prior standardization, and its effectiveness in a well-structured dataset following PCA reduction.

Applying this approach, approximately 22% of the fraud cases were detected as multivariate outliers. This result is expected, as many fraudulent transactions do not behave as global outliers: attackers often camouflage themselves within general behavior, especially when the fraud does not involve large amounts or extreme deviations. Mahalanobis is effective in detecting global outliers, but not those that depend on individual context, break personal patterns, or present non-linear relationships. Thus, certain transactions may not deviate from the multivariate center yet still be anomalous under other perspectives that this linear method does not capture.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

2.4 VERIFY DATA QUALITY

The analyzed dataset presents a valid and suitable structure for academic and demonstrative analyses. Although it contains a considerable number of features, most are encoded and lack explicit contextual meaning. The absence of relevant contextual information, such as IP address, country of origin, or device type, limits the potential for conducting deeper behavioral analyses of fraud, particularly from a user behavior perspective.

From a technical standpoint, the dataset's structural quality is high. No missing values were detected, which simplified the workflow design by eliminating the need for imputation techniques. Regarding duplicates, repeated rows in the negative class (non-fraud) were removed, while duplicate records in the positive class (fraud) were retained due to their scarcity and the importance of capturing as much variability as possible within a minority class.

The variables exhibit highly skewed distributions, some with a bias toward extreme values. However, no transformations were applied, as the models used (Autoencoder and Random Forest) proved to be robust to these characteristics. Outliers were not removed either, since they were considered valuable to the Autoencoder's learning process by including both common patterns and infrequent legitimate transactions.

Given the high number of variables, feature selection was performed using the SelectKBest method with the mutual_info_classif scoring function. This technique reduced the dataset to the 25 most relevant variables for the Random Forest model. The selection was conducted after encoding with the Autoencoder, considering that this architecture transforms the latent representation of the data. Therefore, the final model was allowed to determine which encoded variables were most useful for classification. During this process, 6 variables were discarded—including Time—and the Reconstruction Error was incorporated as a new key feature.

Although the strong class imbalance realistically represents typical fraud behavior in practice, the low proportion of the positive class limited the model's ability to learn deeply from it. Nevertheless, the original distribution was maintained and adaptive adjustments were applied, such as using the class_weight='balanced' strategy and customizing the decision threshold, instead of relying on oversampling techniques.

In general, the dataset is particularly useful for educational purposes: it allows experimentation with different approaches, application of various feature engineering techniques, and evaluation of multiple fraud detection strategies. However, for operational use, its applicability is limited due to the lack of real user behavior context, which restricts both exploratory analysis and the design of more specialized models.

According to the automated evaluation provided by Kaggle, this dataset received a usability score of 8.53 out of 10. This score was based on the following criteria:

Completeness · 100%
Check. Subtitle, Check: Tag, Check: Description, Check: Cover Image

Credibility · 33%
Close: Source/Provenance, Check: Public Notebook, Close: Update Frequency

Compatibility · 100%
Check: License, Check: File Format, Check: File Description, Check: Column Description

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

3. DATA PREPARATION

3.1 SELECT DATA

Before training the Autoencoder model, all available variables were considered, with the exception of the Time variable, which was discarded as it does not represent an interpretable temporal metric nor provide relevant information to the problem. This variable possesses a cumulative scale without a direct correspondence to actual chronological units, limiting its usefulness for modeling transactional patterns over time.

The target variable used was Class, encoded as 0 for legitimate transactions and 1 for fraudulent transactions. The dataset exhibits a strong class imbalance, with the positive class representing only 0.18% of the total observations. To address this imbalance, the partition into training, validation, and test sets was carried out through stratification on the target variable, ensuring the presence of fraud examples in each subset. This decision entailed forgoing temporal order in the partitioning, prioritizing statistical representativeness over chronological sequence, since maintaining that sequence could have resulted in subsets without the minority class, negatively affecting the calculation of sensitive metrics such as the Area Under the Precision-Recall Curve (AUPRC).

Before the final training of the Random Forest model, a feature selection process was applied using SelectKBest, optimizing the value of k through stratified cross-validation, using AUPRC as the target metric. The optimal value was k = 20, which was used as a fixed hyperparameter in the integrated pipeline of SelectKBest + RandomForestClassifier.

Alternatives such as RFE (Recursive Feature Elimination) were also explored, but SelectKBest was chosen for its greater simplicity, its direct interpretability regarding the original variables (without the need for introspection on an estimator), and its native integration in scikit-learn, which facilitated its use in reproducible and auditable workflows. These properties are particularly relevant in contexts such as fraud detection, where transparency in the variable selection process is a critical requirement.

For a visual description of this preparation flow, including the variable selection criteria and the partitioning logic, please refer to section 3 of the notebook, where a schematic diagram summarizing the data architecture prior to modeling is included.

3.2 CLEAN DATA

Duplicate rows were identified within the non-fraudulent class, and these were removed to avoid unnecessary biases during training. Their statistical impact was marginal, as they represented a very low percentage of the total. In contrast, in the fraudulent class all records were retained, including duplicates, due to the extreme scarcity of observations, where each sample is valuable for the model's learning.

No missing values were found in any variable. A high presence of outliers was detected in multiple features, which were not removed since they could correspond to valid extreme behaviors, particularly relevant in the context of financial fraud. These outliers were utilized during the training of the Autoencoder, which used only the legitimate class, as they provided greater variability within what is considered "normal", thus enhancing the model's ability to detect anomalies.

Different scaling methods were evaluated: RobustScaler, PowerTransformer, and MinMaxScaler. The latter was selected for offering the best performance in terms of reconstruction, for the following reasons: the activation functions used (tanh in the hidden layers and ReLU in the output layer) operate more efficiently with data in the [0, 1] range. MinMaxScaler preserves the original proportions between values, which favors the structural fidelity of the learning. Unlike other transformations that generate negative values, MinMaxScaler avoids conflicts with ReLU, whose output is restricted to non-negative values.

Additionally, a QuantileTransformer was applied only to the Amount variable, which exhibited extreme skewness. Under scaling with MinMaxScaler, its distribution was flattened, generating very low relative values and reconstructions close to zero for all rows. This nullified its predictive capacity by eliminating any correlation with the target variable. The quantile transformation better preserved its informative relevance.

Both the scaler and the transformer were trained exclusively on the data from the legitimate class (X0train) to avoid data leakage. Consequently, some validation and test values fell outside the [0, 1] range, although the model demonstrated a good capacity to generalize in these situations.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

Finally, although most variables exhibited high skewness, no transformations such as logarithmic or Box-Cox were applied, as their impact was either null or detrimental to the overall performance of the model—with the sole exception previously noted for the Amount variable.

3.3 CONSTRUCT DATA

To enrich the dataset prior to supervised training, a new variable was constructed based on the reconstruction error generated by an autoencoder trained exclusively on legitimate transactions. This variable captures the degree of deviation of each transaction from the normal patterns learned by the model, and it was added as an additional feature to the original dataset.

The usefulness of this new variable was evaluated using Pearson's correlation coefficient with respect to the target variable, yielding a moderately positive value. This suggests that the reconstruction error provides useful information to discriminate between legitimate and fraudulent transactions, although it was not used directly as a decision threshold.

Additionally, the t-SNE dimensionality reduction technique was applied to visualize the feature space before and after the transformation performed by the autoencoder. The comparison showed greater structural separation between classes after reconstruction, supporting the utility of the autoencoder as a prior stage for compression and enrichment of the representation space, thereby facilitating subsequent supervised learning.

3.4 INTEGRATE & FORMAT DATA

A data integration strategy was defined to ensure independence between the training, validation, and test subsets, preventing any kind of information leakage throughout the workflow. Figure 1 in the code (section 3) illustrates this general partitioning scheme, preserving experimental integrity from the initial preparation to the final evaluation.

During model development, no automated pipeline was implemented using sklearn. Instead, a manual and progressive integration of the various transformation and cleaning steps was chosen, allowing for greater control and traceability in each phase of the process. This decision was especially relevant given the sequential nature of the workflow: the dataset was iteratively enriched and adapted as different techniques, such as feature engineering or compression via autoencoders, were applied.

Although a unified automated flow was not implemented during the model construction phase, a modular function was later designed to reproduce this complete preprocessing pipeline. This function is described in section 5 (Evaluation), where it is used to transform new input data and apply the trained model in a consistent and secure manner.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

4. MODELING

4.1 SELECT MODELING TECHNIQUE

The choice of modeling technique emerged during the exploratory data analysis (EDA), particularly when attempting to visualize the structure of the dataset in two dimensions. Given the large volume of data and limited computational resources, dimensionality reduction was applied using the t-SNE technique on a sample of the dataset. A random 10% of non-fraudulent transactions were selected, along with all fraudulent transactions.

The resulting plot revealed that the transactions followed a spiral-shaped distribution, with fraudulent instances completely overlapping non-fraudulent ones. This observation raised a key question: is there a way to transform the feature space to facilitate the separation between both classes, even if this separation is not visually evident in the reduced dimension?

As a response, an autoencoder was implemented with the goal of transforming the data in a way that would emphasize subtle differences between the classes. The idea was to train the autoencoder exclusively on the majority class (non-fraud) so that it could learn its structure in as much detail as possible. This way, when presented with fraudulent data (unseen during training), the autoencoder would produce poor reconstructions, increasing the reconstruction error for these cases. Later, this error would be used as a new variable, combined with the autoencoder's reconstructed output, as input to a classifier based on Random Forest.

During the autoencoder's experimental phase, the impact of removing univariate outliers from the negative class was evaluated. This removal was done using Z-score standardization, with a more conservative threshold of ±5.5 z-units, aiming to exclude only the most extreme outliers without discarding relevant information. It was observed that by removing these values, the reconstruction error range for fraudulent transactions increased, theoretically enhancing the model's discriminative power.

However, further tests showed that training the autoencoder without these outliers led the Random Forest classifier to label many atypical yet legitimate cases as fraud, negatively affecting its generalization capability. Although fraud detection improved, the model generated too many false positives, compromising its practical utility. Therefore, it was decided to retain outliers in the training set, in order to build a more balanced model, one that could identify more fraud cases without misclassifying atypical legitimate cases.

Once it was identified that including outliers in the training set slightly reduced the discriminative power of the autoencoder, it became necessary to compensate for this loss by maximizing the classification model's performance. To this end, several binary classification algorithms were evaluated, aiming to optimize predictive power on both the training and unseen (test) data.

Among the models considered, the two main candidates were XGBoost and Random Forest Classifier (RFC). In initial tests, XGBoost achieved a slightly higher score in the area under the precision-recall curve (AUPRC) metric, with an absolute improvement of 0.33% over RFC, and also showed a slightly lower standard deviation (though not statistically significant). Despite this marginal advantage, Random Forest was chosen as the primary classifier. This decision was based on the analyst's judgment, valuing RFC's stability, interpretability, and ease of integration into the workflow, without dismissing XGBoost as a valid alternative for this problem.

The proposed approach integrates three complementary algorithms that operate together to optimize classification system performance:

1. Autoencoder: trained exclusively on data from the negative class (non-fraud), its purpose is to reconstruct the inputs and generate the reconstruction error as an additional feature. This allows quantifying class discrepancies in terms of reconstruction, aiding anomaly detection.

2. SelectKBest: a feature selection algorithm based on the mutual information metric for classification. This technique identifies the most relevant variables for the final model, optimizing performance without introducing unnecessary complexity.

3. Random Forest Classifier: a robust and efficient classification model used to generate the final predictions based on the data processed in the previous stages. This classifier was later fine-tuned through hyperparameter optimization, with a specific focus on improving performance under the AUPRC metric.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

Initially, SelectKBest was used without hyperparameter tuning, aiming to select the top features before building an initial estimator. This preliminary setup provided a baseline upon which a full workflow was later developed. Although Recursive Feature Elimination (RFE) with Random Forest was considered as an alternative selection method, SelectKBest was ultimately preferred due to its low computational cost, ease of integration within scikit-learn's sequential processing pipelines, and its solid performance observed during cross-validation.

In summary, the complete model consists of three interdependent components: the autoencoder for transforming and enriching the data, the feature selection method to effectively reduce dimensionality, and the Random Forest classifier as the core predictive engine. This architecture was designed to maximize the model's efficiency and detection capability, particularly in scenarios where the positive class (fraud) is rare and hard to distinguish.

4.2 GENERATE TEST DESIGN

To approach the problem rigorously, it was established that the primary metric to optimize would be the Area Under the Precision-Recall Curve (AUPRC), since the goal is to effectively predict the positive class (fraud). This metric is especially suitable for imbalanced classification scenarios, as it focuses exclusively on the model's performance with respect to true positives, without being influenced by the large proportion of true negatives, a known limitation of metrics such as accuracy or AUROC.

AUPRC directly penalizes false positives through the precision component, which is crucial in this context where an incorrect prediction can lead to significant consequences (e.g., legitimate transactions being blocked). Moreover, AUPRC provides a detailed assessment of the model's ranking quality: a high AUPRC not only implies effective detection of positives but also that these instances are correctly ranked according to their estimated probability. This property is particularly valuable in applications where a dynamic threshold or alert prioritization is required.

For these reasons, AUPRC was chosen as the objective metric for both feature selection and classifier hyperparameter tuning, in order to maximize the model's discriminative power on the minority class.

Since AUPRC does not require a classification threshold for its computation, a post-training threshold optimization stage was carried out using the F1-score as the selection criterion. As the harmonic mean between precision and recall, the F1-score is well-suited for imbalanced datasets, as it simultaneously captures the model's ability to correctly detect positives (recall) and avoid false alarms (precision).

Threshold optimization plays a corrective role against the dataset's inherent asymmetry, enabling the probabilistic output of the model to be translated into a more balanced binary decision. This tuning step helps achieve a more effective trade-off between type I (false positives) and type II (false negatives) errors, without compromising the overall evaluation provided by the AUPRC.

To ensure the robustness and reliability of the model's evaluation metrics, multiple model selection strategies were implemented, this being one of the most computationally demanding phases. In total, validation techniques were applied at three key points in the pipeline.

The first corresponded to the feature selection process, which used stratified cross-validation to maintain the original class distribution across folds. A 10-fold scheme was applied, evaluating the AUPRC metric for various k values in the SelectKBest algorithm. The optimal number of features was selected based on performance using a baseline Random Forest model, allowing for the identification of the features that best contributed to detecting the positive class.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

The second validation instance focused on hyperparameter optimization of the Random Forest classifier, using RandomizedSearchCV with 50 iterations. Each iteration was evaluated through 10-fold stratified cross-validation, with AUPRC once again chosen as the optimization metric. This ensured that the selected hyperparameters maximized the model's ability to correctly distinguish the minority class. A fixed random seed was maintained throughout the process to guarantee reproducibility and comparability across runs. It is important to note that stratification was consistently applied across all evaluation stages, due to the high class imbalance and the critical need for the model to learn meaningful representations of both classes.

For fine-tuning the key hyperparameters specifically, the maximum tree depth and the minimum number of samples required to split a node, the GridSearchCV method was used, again with the same stratified 10-fold cross-validation setup.

In a third stage, with the model already optimized, a final cross-validation was performed to determine the optimal decision threshold. This step relied on the model's output probabilities on the training set, selecting the threshold that maximized the F1-score. Since this metric is directly influenced by the decision threshold, it served as a suitable complement to AUPRC, enabling the translation of the model's probabilistic output into a binary prediction that better balances precision and recall, which is critical in this context.

Finally, an integrated visual diagram was developed to summarize all stages of the modeling and validation process, including the specific points where each evaluation technique was applied. This schematic representation clearly illustrates the pipeline architecture and the logic behind each methodological decision.

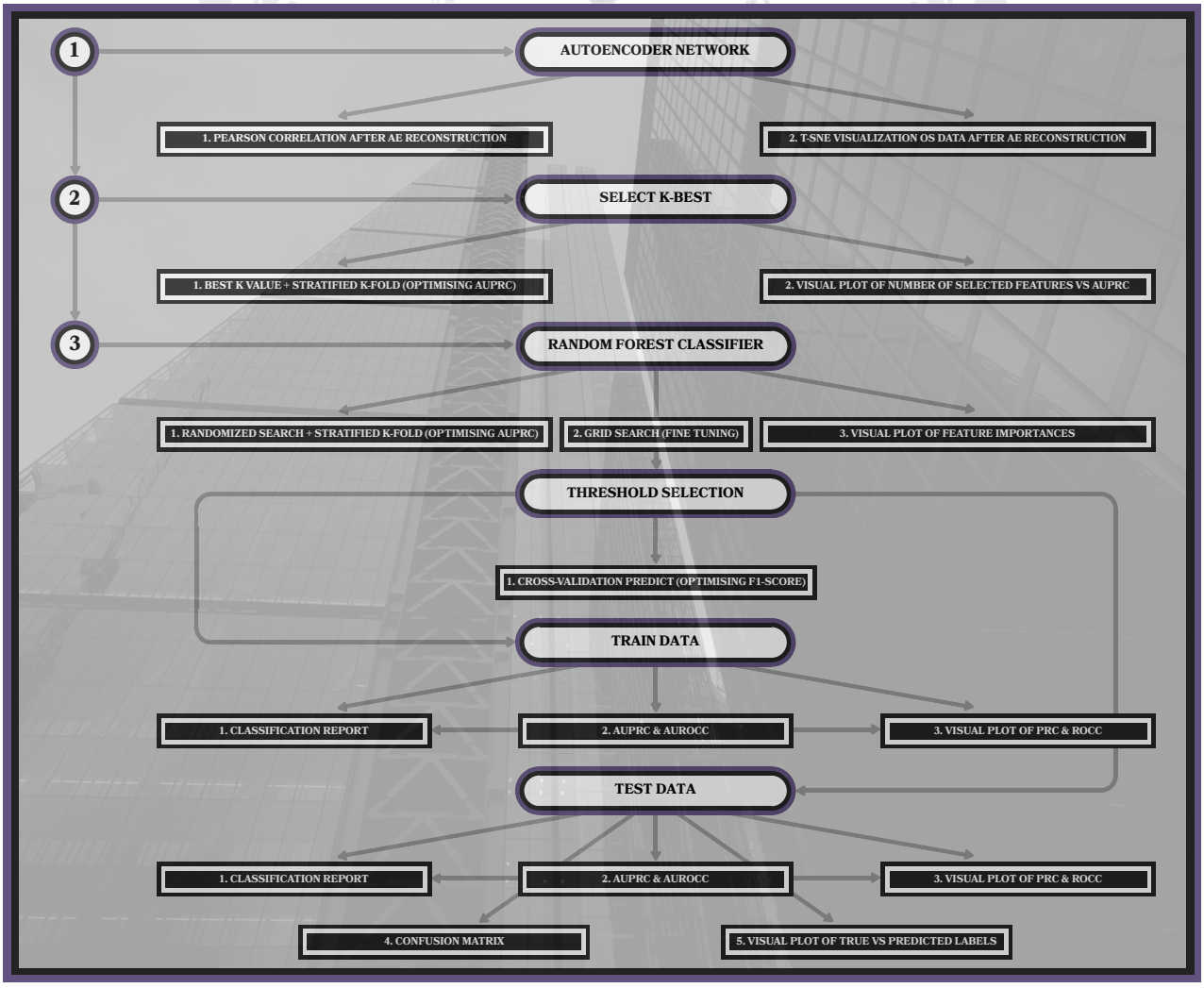


Figure 2. Test design of the proposed modelling process
Diagram of the test flow applied at each stage of the modelling process, including visualisations, metrics and validations used to evaluate the performance of the autoencoder, feature selection and Random Forest classifier.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

4.3 BUILD MODEL PARAMETER SETTINGS

AUTOENCODER ARCHITECTURE

The autoencoder was implemented using the functional API of Keras with TensorFlow as backend, adopting a symmetric encoder-decoder architecture. This structure (128→64→128→64) was designed for progressive compression followed by mirrored reconstruction, enabling the learning of compact latent representations without significantly compromising the integrity of the original information.

Tanh activation functions were used in all hidden layers due to their zero-centered nature, which enhances training stability. Although Tanh outputs lie within the range [-1, 1], its symmetry complements the input data scaled to [0, 1] using MinMaxScaler. In contrast, the output layer employed ReLU, as the data is constrained to positive values. This choice prevents the model from reconstructing negative values outside the expected domain, reinforcing the role of the autoencoder as an anomaly filter.

Dropout regularization (30%) was introduced after each encoder layer, which proved crucial given the unilateral training (only on legitimate class data). This technique mitigates overfitting and forces the model to learn robust representations by randomly deactivating neurons during training, thus promoting generalization to out-of-distribution inputs.

Training was performed using the Adam optimizer, chosen for its adaptability and efficiency in regression settings. The loss function was the mean squared error (MSE) between the input and its reconstruction, allowing a direct quantification of reconstruction error. 30% of the legitimate training data was held out for validation, enabling monitoring of model behavior on unseen examples. After several trials, the following training parameters were set: batch size of 256 and 25 epochs, with convergence achieved without signs of overfitting. Fixed random seeds ensured full reproducibility.

The reconstruction error produced by the autoencoder was used as a key feature for fraud detection. It emerged as the most important variable according to the feature importance ranking provided by the subsequent Random Forest model, reinforcing the effectiveness of the autoencoder as an initial anomaly-based filter.

Regarding generalization beyond the [0, 1] input range, the model maintained consistent performance on validation and test data, even though the scaler was fit exclusively on X0train. Early stopping was not applied; instead, overfitting was controlled via explicit regularization and cross-validation.

Overall, the autoencoder design was grounded in principles of robustness, interpretability, and generalization capacity, fully aligned with the anomaly detection objective. Every component of the architecture was empirically validated and integrated coherently into the model's broader workflow.

FEATURE SELECTION WITH SELECTKBEST

To reduce dataset dimensionality and enhance model efficiency, univariate feature selection was performed using SelectKBest, with mutual information (mutual_info_classif) as the scoring function. While this metric is primarily designed for discrete variables, it has proven effective with continuous features as well, due to its ability to capture non-linear relationships with the target variable.

The optimal number of features was selected empirically through 10-fold stratified cross-validation, with the average precision score (AUPRC) as the optimization metric. This selection process employed a simple Random Forest model, with no hyperparameter tuning aside from fixing a random seed and setting the class weight to "balanced." The best performance was achieved using 20 features, thus defining the final input dimensionality of the model.

RANDOM FOREST, HYPERPARAMETER OPTIMIZATION, AND FINAL ARCHITECTURE

The Random Forest classifier was selected for its suitability in datasets with high dimensionality, non-linear relationships, and inherent noise, all of which are common in fraud detection scenarios. This algorithm does not require prior normalization and is known for its robustness to overfitting, its capacity to handle imbalanced distributions, and its interpretability through feature importance evaluation.

The model was trained on an enriched dataset, including the reconstruction error generated by the autoencoder as an additional variable. This new feature proved to be key, as it added an extra dimension capturing each observation's deviation from the dominant learned pattern, thereby enhancing the model's ability to detect outliers.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

To optimize classifier performance, RandomizedSearchCV was applied over a predefined hyperparameter space using stratified cross-validation. The following dimensions were explored: **n_estimators:** The number of trees in the forest was tuned to strike a balance between predictive performance and computational cost. While more trees increase stability, they also yield diminishing marginal returns. **max_depth:** The maximum tree depth was limited to prevent the model from overfitting noise, especially in the minority class. **min_samples_split** and **min_samples_leaf:** These parameters helped avoid overly specific splits that could capture noise or outliers, thereby improving generalization. **max_features:** The number of features considered at each split was optimized to increase tree diversity and reduce internal correlation, improving generalization.

Following the initial optimization phase using random search, a focused fine-tuning was performed on two key hyperparameters: the maximum depth of each tree **(max_depth)** and the minimum number of samples required to split a node **(min_samples_split)**. A search grid was defined around the optimal values identified by RandomizedSearchCV, exploring five values above and below the initial estimate. This second tuning phase was conducted using GridSearchCV with stratified cross-validation, keeping all other hyperparameters fixed at their previously selected values.

The use of RandomizedSearchCV over an exhaustive grid search enabled efficient exploration of the search space, reducing computational time without compromising performance. This process was critical not only for optimizing key metrics such as AUPRC but also for preserving the model's interpretability, an essential requirement in environments where system traceability and validation are paramount.

The final model architecture was implemented as a classification pipeline, centered on a Random Forest classifier specifically configured for the challenge of fraud detection in a highly imbalanced context. The final configuration, the result of optimization through RandomizedSearchCV and stratified cross-validation, includes the following key parameters: **n_estimators = 300:** A total of 300 trees was set to ensure robust prediction and reduce variance via aggregated voting from multiple estimators. **max_depth = 18:** This depth limit restricts each tree's complexity, helping prevent overfitting—particularly to the rare positive (fraud) class. **min_samples_split = 8:** A minimum of 8 samples is required for a split, improving the model's generalization to infrequent patterns. **class_weight = 'balanced':** This setting allows the algorithm to assign weights inversely proportional to class frequencies, enhancing sensitivity to fraud. **random_state = 4:** A fixed seed ensures reproducibility of results. **n_jobs = -1:** All available cores are utilized to accelerate training.

This hyperparameter combination results in a robust and efficient architecture, capable of adapting to the inherent challenges of learning from imbalanced classes. The final model not only achieves strong predictive performance, but also maintains an adequate level of interpretability and reproducibility for deployment in critical environments such as financial fraud detection.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

4.4 ASSESS MODEL

AUTOENCODER BEHAVIOR

The selected autoencoder, after multiple empirical tests, exhibited behavior consistent with its purpose as an unsupervised detection component. Although outliers were retained during training—which in some cases reduced the reconstruction error gap between classes, a consistent trend was observed: fraudulent transactions tend to yield higher reconstruction errors compared to legitimate ones.

This pattern supports the core hypothesis of the approach: being trained exclusively on genuine data, the autoencoder poorly reconstructs observations that deviate from the dominant pattern, making reconstruction error an effective signal for anomaly detection. Beyond its role as an explicit feature, the autoencoder also had a positive impact on the structure of the transformed data. Notably, variables such as Amount, which had lost discriminative power after normalization, regained relevance after reconstruction. An increase in their absolute correlation with the target variable was detected, suggesting that the reconstruction process not only preserves but can also enhance useful patterns for downstream classification.

To evaluate the internal structure generated by the autoencoder, dimensionality reduction using t-SNE was applied to the reconstructions. The results showed the emergence of two distinct clusters: one predominantly composed of legitimate transactions, and another grouping a significant proportion of fraud cases. This segmentation is particularly relevant given that the model was not trained on fraudulent data. The fact that frauds are projected into a separate region of the latent space reinforces the idea that the autoencoder has learned a representation that is structurally sensitive to anomalous deviations, implying an implicit form of unsupervised detection.

However, some fraud cases were also found to overlap with legitimate transactions in the reduced space, indicating that while the separation is strong, it is not perfect. This result aligns with the autoencoder's design goals and demonstrates its effectiveness as an initial filter that significantly contributes to the overall model's performance without requiring direct supervision. Overall, the autoencoder not only fulfills its primary role as a generator of a useful reconstruction error metric, but also acts as a feature transformer, promoting more informative and useful latent structures for final classification. Its integration into the workflow strengthens the overall detection system architecture by adding robustness, interpretability, and generalization capability.

BEHAVIOR OF SELECTKBEST

Feature selection was carried out using SelectKBest with mutual information scoring and stratified cross-validation (10 folds), using a Random Forest as the base estimator. The best AUPRC performance (0.873) was achieved when selecting 20 features.

Performance was poor when using fewer than 10 variables, but stabilized significantly from that point onward. This indicates that the model remains robust even with many variables, although the optimal subset helps reduce noise, improve generalization, and facilitate subsequent hyperparameter tuning. Overall, SelectKBest enhanced both the model's performance and efficiency without compromising its predictive power.

RANDOM FOREST CLASSIFIER BEHAVIOR

After the hyperparameter optimization process, the final model selected was a Random Forest classifier configured with balanced class weighting, consisting of 300 trees, a maximum depth of 18, and a minimum split of 8 samples. To accelerate training, the use of all available CPU cores was enabled (n_jobs = -1), and a fixed random seed (value 4) was set to ensure reproducibility of the experiment.

Before the final prediction, the decision threshold was fine-tuned to optimize the F1-score. This step involved exploring multiple threshold values and selecting the one that maximized the balance between precision and recall. Initially, the model showed nearly perfect metrics on the training set, suggesting potential overfitting. However, after applying the optimal threshold, more realistic and generalizable results were obtained:

- **F1-score (fraud): 0.8913**
- **Precision (fraud): 0.9567**
- **Recall (fraud): 0.8343**

It is important to note that AUPRC and AUROCC metrics, being based on model behavior across multiple thresholds, are not affected by threshold adjustment. The values obtained were:

- **AUPRC: 0.8726**
- **AUROCC: 0.9631**

These results reflect excellent discriminative power in a highly imbalanced setting. The AUPRC value, far exceeding the positive class proportion (frauds), confirms that the model maintains high precision even as recall increases, crucial in contexts where false positives carry a manageable cost compared to false negatives.

While the model detects approximately 83% of fraudulent transactions, representing good recall, some frauds still go undetected. This limitation may be critical depending on the business-defined risk tolerance. Consequently, if exhaustive detection is prioritized, complementary strategies such as model ensembling or using more sensitive thresholds could be considered.

Regarding feature importance, the analysis revealed that the most influential variable was V14, followed by the derived reconstruction error from the autoencoder, and then V10, V4, and V12. These five variables form the predictive core of the model. While the Amount variable was expected to have a stronger impact, its contribution, though smaller, was still meaningful, especially after transformation with QuantileTransformer, which improved its correlation with the target class.

In summary, the Random Forest classifier proved to be robust, accurate, and highly interpretable, with an excellent balance between performance and traceability. Nonetheless, as will be discussed in Section 5, it will be essential to validate these metrics on the test set to assess the model's generalization capability and rule out any residual overfitting effects.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

5. EVALUATION

5.1 EVALUATE RESULTS

In this final stage, the model's performance was evaluated on the test set, which had never been used during training or hyperparameter tuning. This allows for a more accurate estimation of its generalization capacity and its potential behavior in a real-world scenario.

The data were transformed following the previously defined preprocessing flow, ensuring structural compatibility with the training set. Subsequently, class probabilities were generated, and the optimized decision threshold was applied, discarding the default value (0.5) in order to achieve a better balance between precision and recall.

The results obtained were consistent with those from training and validation, confirming the model's robustness. Precision for the fraud class reached 93.89%, reflecting a very low false positive rate, while recall was 83.11%, indicating that the model detects more than 8 out of every 10 frauds. The corresponding F1-score was 0.8817, confirming an excellent balance between the two metrics. At a global level, the model showed very high discriminative ability, with an AUPRC of 0.8941 and an AUROCC of 0.9771, metrics that are particularly relevant in imbalanced class contexts. The confusion matrix supports this conclusion, showing a minimal error rate for legitimate transactions: approximately one error every 3,185 cases.

Additionally, a 2D projection via t-SNE of the latent space revealed a partial segmentation of fraud cases, with many grouped in a clearly differentiated region. The model was particularly effective at detecting the majority of frauds that were structurally separated from legitimate transactions. However, it showed limitations in identifying frauds that remained overlapped with genuine cases scattered across the latent space. This suggests an opportunity for improvement in the autoencoder's architecture or hyperparameters. Addressing this "gray zone" in the latent space could lead to better detection of subtle or structurally camouflaged frauds.

Conclusion: The model demonstrates solid performance, generalizes well, and maintains consistency between validation and test results. Although there is room to improve sensitivity toward harder-to-distinguish frauds, the system is viable for operational deployment, provided it aligns with the business-defined risk threshold. The complete workflow—from early detection via autoencoder to classification with a calibrated Random Forest—has proven effective. Moving forward, we recommend iterating on the autoencoder to enhance its ability to separate overlapping cases in the latent space, thus improving sensitivity without sacrificing precision.

5.2 REVIEW PROCESS

During the evaluation phase, a comprehensive review was carried out to verify compliance with the objectives defined in the early stages of the project. This review was not limited to quantitative metric analysis but also included a qualitative assessment of the model's behavior under different operational scenarios, with particular focus on robustness, interpretability, and generalization capability.

To mitigate the risk of overfitting, stratified cross-validation was used during training, maintaining consistent class proportions across all folds. Additionally, a completely independent test set was used—one that had not participated in any previous phase. This set underwent the full preprocessing flow, including normalization, data reconstruction via the autoencoder, computation of the reconstruction error, and concatenation of original and engineered features, thereby ensuring structural consistency with the validation environment.

The results obtained on this set were consistent with the expected values. Global metrics such as precision, recall, F1-score, AUPRC, and AUROCC reflected a high level of generalization, with no signs of performance deterioration compared to training data. Complementarily, a visual analysis of the latent space using t-SNE showed a reasonable separation between classes. This visualization confirmed the model's ability to capture useful structural patterns, qualitatively validating the numerical results.

However, an error analysis was conducted to better understand the model's limitations. It was found that frauds overlapping with legitimate cases in diffuse regions of the latent space continue to represent a challenge, as they tend to evade the classifier. This finding suggests a potential area for improvement in the autoencoder or in methods to enrich the representation space.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

From a practical standpoint, key aspects related to production deployment were also considered. Although robust, the model requires periodic monitoring to ensure its stability in the face of changing transaction patterns or the emergence of concept drift. Updating decision thresholds or retraining with new data could become part of a responsible maintenance plan.

In conclusion, the review process demonstrated that the model not only meets but exceeds the success criteria defined in the initial stages. It is a reliable system, with a well-balanced trade-off between precision and recall, and strong foundations both in design and validation. Nonetheless, there is still room for improvement, particularly in detecting subtle or well-camouflaged fraud cases. Future iterations could focus on refining the autoencoder's capabilities or exploring ensemble methods to capture these residual cases.

5.3 APPROVED MODELS

Upon completing the validation process on the test set, the final approved solution consisted of a Random Forest classifier trained on the representations reconstructed by the autoencoder, including the reconstruction error as an additional explanatory feature. This architecture was selected not only for its outstanding quantitative performance but also for its robustness, generalization ability, and interpretability when compared to more complex alternatives.

The model selection was grounded in its consistent performance on unseen data, demonstrating alignment with the objectives established at the project's outset. During the evaluation phase, the full preprocessing pipeline was applied, including normalization of the original features, transformation through the autoencoder, calculation of reconstruction error, and feature selection via SelectKBest, prior to classification. This methodological pipeline was consolidated into an effective, end-to-end integrated solution.

The decision threshold was precisely tuned to maximize the F1-score, optimizing the balance between precision and recall. This calibration stabilized the model's behavior in operational settings where misclassification errors can entail substantially different costs.

5.4 DETERMINE NEXT STEPS

The project concluded successfully with the development of a robust and generalizable fraud detection model, thoroughly validated on an independent test set. Although an operational deployment phase was not in scope, several improvement paths were identified to guide future development and implementation efforts.

A key future priority is to revisit the autoencoder architecture. While it enabled effective separation of many fraudulent observations in the latent space, some cases remain overlapped with the legitimate class. Optimizing this representation may enhance the system's sensitivity to subtler frauds. Additionally, adapting the workflow for real-time detection environments is proposed, which entails challenges such as minimizing latency, dynamically recalibrating the decision threshold, and continuously monitoring production performance.

During development, significant challenges emerged that proved instructive. One of the most critical was data leakage, which occurred due to applying transformations prior to data partitioning. This was resolved by encapsulating each preprocessing step (including normalization, the autoencoder, and error calculation) within the training set and subsequently replicating it during validation and testing. Overfitting was also addressed, initially identified through discrepancies between training and validation performance, and mitigated via regularization, stratified cross-validation, and feature selection with SelectKBest.
Moreover, a key conceptual error was corrected by moving away from a fixed threshold of 0.5, which is unsuitable for imbalanced datasets. Instead, a precision-recall curve analysis was used to identify the point that maximized the F1-score, thereby enhancing fraud detection without compromising precision. This adjustment was essential to improve the system's sensitivity.

Structurally, a robust and reproducible workflow was designed to ensure consistency between training and inference. This design mitigated issues related to making predictions outside the complete processing pipeline. Finally, challenges inherent to the class imbalance were successfully addressed by ensuring a representative distribution across all datasets and by employing metrics such as AUPRC, AUROCC, and t-SNE visualizations to better understand the latent space structure and identify improvement opportunities.

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

6. REFERENCES

GROWTH OF GLOBAL E-COMMERCE

- Retail e-commerce sales worldwide from 2014 to 2027
- 51 eCommerce Statistics In 2025 (Global and U.S. Data)

USE OF DIGITAL WALLETS AND ONLINE PAYMENT METHODS

- Digital Wallet Statistics
- Top US Payment Methods (2023–2027)

TYPES OF FRAUD IN ONLINE TRANSACTIONS

- What Is Carding? How It Works, Prevention Methods, and Examples
- 6 Types of Credit Card Fraud & How Businesses Can Stop Them
- The Ultimate Guide to Fraud Detection and Prevention

CRISP-DM METHODOLOGY

- The CRISP-DM Process: A Comprehensive Guide

AUTOENCODERS AND FRAUD DETECTION

- Sehrawat, D., & Singh, Y. (2023). Auto-Encoder and LSTM-Based Credit Card Fraud Detection. SN Computer Science, 4(557)
- Anónimo. (2023). Credit Card Fraud Detection Using an Autoencoder Model with New Loss Function. OAJI
- Semi Supervised Classification using AutoEncoders
- How Autoencoders Work: Intro and UseCases

RANDOM FOREST AND SUPERVISED CLASSIFICATION

- Anónimo. (2023). Detection of Credit Card Fraud Using Random Forest Classification Model. ResearchGate
- Anónimo. (2023). A Random Forest Classifier Approach to Payment Fraud Detection. IJISRT

FEATURE SELECTION WITH MUTUAL INFORMATION

- Anónimo. (2023). The Effect of Feature Selection on the Accuracy of X-Platform User Identification. Electronics, 13(1), 205
- Bellani, C., Kraev, E., & Shestopaloff, A. (2023). Feature Selection with Neural Estimation of Mutual Information. ICLR 2024 Submission

VISUALISATION AND SEPARATION OF CLASSES WITH T-SNE

- Anónimo. (2023). Credit Card Fraud Detection using Logistic Regression Compared with t-SNE to Improve Accuracy. ResearchGate

OTHER WORKS OF REFERENCE

- Credit Fraud || Dealing with Imbalanced Datasets.
- Best techniques and metrics for Imbalanced Dataset
- SMOTE with Imbalance Data

CAIO OLIVEIRA QUINAMO
DATA SCIENTIST
caioquinamocontact@gmail.com

7. LICENCE AND FINAL NOTE FROM THE AUTHOR

LICENCE

FINAL NOTE FROM THE AUTHOR

The author is sincerely grateful for the time and effort of all those who have had the opportunity to review and comment on this work. Their engagement is invaluable, as it contributes to the growth of the ideas presented and fosters a meaningful exchange of knowledge.

The author values open and constructive debate and believes in the power of sharing diverse ideas and perspectives. Whether you agree or disagree with the points raised in this work, your opinions and comments are always welcome, as they help to refine understanding of the subject matter and encourage intellectual growth.

Please do not hesitate to contact us by email with any questions, suggestions for improvement or to engage in discussions about the work. Sharing your ideas and opinions, whether they are criticisms, alternative viewpoints or suggestions for further research, is always appreciated and welcomed.

Your genuine participation in these discussions contributes to enriching the body of knowledge and to the collective advancement of the field. Thank you for your valuable contributions.

— CAIO O. QUINAMO

caioquinamocontact@gmail.com

"THE GOAL IS TO TURN DATA INTO INFORMATION, AND INFORMATION INTO INSIGHT."
— CARLY FIORINA

"TELLING THE TRUTH WITH DATA."
— HANS ROSLING

LinkedIn GitHub kaggle Gmail

CONTACT AND SOCIAL MEDIA

INTERNAL CODE: NDSPP/2025/0001-01-1

RATIO ET INTEGRITAS DUCUNT AD PROSPERITATEM