UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Anomaly Detection for Predicting Churn in POS Payment Networks

Proyecto de Titulación

José Antonio Asitimbay Zurita

Felipe Grijalva, Ph.D.

Director de Trabajo de Titulación

Trabajo de titulación de posgrado presentado como requisito para la obtención del título de Magíster en Inteligencia Artificial

Quito, 01 de diciembre 2024

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ
COLEGIO DE POSGRADOS

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Anomaly Detection for Predicting Churn in POS Payment Networks

José Antonio Asitimbay Zurita

Nombre del Director del Programa:	Felipe Grijalva
Título académico:	Ph.D.
Director del programa de:	Inteligencia Artificial

Nombre del Decano del colegio Académico:	Eduardo Alba
Título académico:	Doctor en Ciencias Matemáticas
Decano del Colegio:	Ciencias e Ingenierías

Nombre del Decano del Colegio de Posgrados:	Dario Niebieskikwiat
Título académico:	Doctor en Física

Quito, Diciembre 2024

© DERECHOS DE AUTOR

Nombre del estudiante:	José Antonio Asitimbay Zurita
Código de estudiante:	00338524
C.I.:	172289120-5
Lugar y fecha:	Quito, 01 de Diciembre de 2024.

DEDICATORIA

Dedico este trabajo a mi familia, quienes han sido mi mayor pilar a lo largo de este camino. A mis padres, por enseñarme a soñar en grande y trabajar con dedicación, inculcándome valores que han sido esenciales para alcanzar mis metas. A mi hermana, quien con su alegría y constante apoyo ha sido una fuente inagotable de motivación y aliento. Y a mis amigos, cuya amistad y respaldo han dejado una huella imborrable en este proceso de aprendizaje y crecimiento.

AGRADECIMIENTOS

A mis compañeros de clase, quienes no solo compartieron este recorrido académico, sino que también aportaron ideas y conocimientos que enriquecieron cada etapa de la maestría. Su apoyo y camaradería hicieron de esta experiencia algo inolvidable.

También agradezco a mis colegas de trabajo, cuyo apoyo y colaboración ayudaron a dar forma y dirección a este proyecto. Su visión y aportes resultaron clave para alinearlo con las necesidades del ámbito profesional.

Por último, quiero agradecer a mi mamá, cuyo amor, paciencia y respaldo emocional fueron un aliciente constante durante todo este proceso. Su apoyo incondicional fue mi mayor motivación para seguir adelante.

RESUMEN

Para abordar este problema, se realizó una segmentación de negocios basada en el volumen de transacciones y el monto promedio de las mismas, utilizando técnicas no supervisadas como Modelos de Mezcla Gaussiana (GMM) y K-Means. Una vez segmentados los negocios con el mejor método, que resultó ser GMM, se analizaron las transacciones a nivel de red y tipo de tarjeta. Dado el significativo desbalance entre negocios churn y no churn, se optó por métodos de detección de anomalías no supervisados, incluyendo One-Class SVM, Isolation Forest y Local Outlier Factor (LOF).

LOF demostró el mejor rendimiento, logrando un AUC de Precisión-Recall de 0.99 para el segmento Micro y 0.96 para el segmento Pequeño, con valores de recall de 0.99 y 0.95, respectivamente, capturando casi todas las anomalías reales en estas categorías. Para los segmentos Masivo, Grande y Mediano, se observaron valores más bajos de AUC de Precisión-Recall (0.71, 0.52 y 0.53, respectivamente) debido a una mayor cantidad de falsos positivos. Sin embargo, estos falsos positivos son bajos en términos absolutos, ya que los casos de churn en estos segmentos son raros. Esto los convierte en una herramienta valiosa para identificar posibles candidatos a churn futuro, facilitando estrategias de retención proactivas.

Palabras clave: Detección de Anomalías, Predicción de Churn, Redes de Pago POS, Gaussian Mixture Models (GMM), Local Outlier Factor (LOF), One-Class SVM, Isolation Forest

ABSTRACT

To address this issue, a clustering of businesses was performed based on transaction volume and average transaction amounts, using unsupervised techniques such as Gaussian Mixture Models (GMM) and K-Means. Once the businesses were segmented with the best method, which turned out to be GMM, transactions were analyzed at the network and card type level. Given the significant imbalance between churn and non-churn businesses, unsupervised anomaly detection methods were chosen, including One-Class SVM, Isolation Forest, and Local Outlier Factor (LOF).

LOF demonstrated the best performance, achieving a Precision-Recall AUC of 0.99 for Micro and 0.96 for Pequeño segments, with recall values of 0.99 and 0.95, respectively, effectively capturing almost all true anomalies in these categories. For the Masivo, Grande, and Mediano segments, lower Precision-Recall AUC values were observed (0.71, 0.52, and 0.53, respectively) due to higher false positives. However, these false positives are low in absolute terms, as churn cases in these segments are rare. This makes them valuable for identifying potential future churn candidates, aiding in proactive retention strategies.

Key words: Anomaly Detection, Churn Prediction, POS Payment Networks, Gaussian Mixture Models (GMM), Local Outlier Factor (LOF), One-Class SVM, Isolation Forest

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Anomaly Detection for Predicting Churn in POS Payment Networks

José Asitimbay, *Member, IEEE, Felipe Grijalva, Senior Member, IEEE,*

***Abstract*—This project focuses on the early detection of businesses that may stop using the point-of-sale (POS) service or restrict its use, resulting in a reduction in transaction volume, which could be considered as a lost customer. In the competitive POS service environment, retaining customers is essential, as losing a customer and acquiring a new one is more costly than maintaining an existing one. Early identification of businesses at risk of churn enables the application of specific retention strategies, improving user experience, increasing loyalty, and preventing customer loss.**

To address this issue, a clustering of businesses was performed based on transaction volume and average transaction amounts, using unsupervised techniques such as Gaussian Mixture Models (GMM) and K-Means. Once the businesses were segmented with the best method, which turned out to be GMM, transactions were analyzed at the network and card type level. Given the significant imbalance between churn and non-churn businesses, unsupervised anomaly detection methods were chosen, including One-Class SVM, Isolation Forest, and Local Outlier Factor (LOF).

LOF demonstrated the best performance, achieving a Precision-Recall AUC of 0.99 for Micro and 0.96 for Pequeño segments, with recall values of 0.99 and 0.95, respectively, effectively capturing almost all true anomalies in these categories. For the Masivo, Grande, and Mediano segments, lower Precision-Recall AUC values were observed (0.71, 0.52, and 0.53, respectively) due to higher false positives. However, these false positives are low in absolute terms, as churn cases in these segments are rare. This makes them valuable for identifying potential future churn candidates, aiding in proactive retention strategies.

***Index Terms*—Anomaly Detection, Churn Prediction, POS Payment Networks, Gaussian Mixture Models (GMM), Local Outlier Factor (LOF), One-Class SVM, Isolation Forest**

I. Introduction

CURRENTLY, the use of point-of-sale (POS) terminals is essential for a wide variety of businesses. Since the pandemic in 2020, their adoption has increased considerably, as they enable fast and secure transactions that facilitate business operations without the need for handling cash. Additionally, offering card payment options allows businesses to expand their potential customer base and improve user satisfaction. However, in the competitive POS service environment, retaining clients has become a key priority. This is because losing a client (churn) not only results in reduced transaction volume but also involves considerable costs in acquiring new clients. Therefore, the early detection of businesses at risk of stopping their use of POS services or restricting their use is essential for implementing retention strategies that prevent churn and strengthen customer loyalty.

In terms of research, churn prediction has been approached using various methods, from supervised models to unsupervised techniques. In this case, the number of churn records is very limited compared to the large volume of non-churn transactions, so unsupervised methods, such as anomaly detection, were chosen to identify these at-risk businesses.

Before addressing anomaly detection, it is essential to segment businesses into groups with similar transaction patterns. This allows for the analysis of anomalies within groups that share comparable transaction behavior and, once possible churn cases are identified, the application of specific retention strategies for each group. The way a business is retained varies depending on its type; for example, a large chain may require a different approach than a small enterprise.

The objective of this study is to develop an effective model for the early detection of churn in POS payment networks by combining clustering techniques and anomaly detection. To achieve this, a clustering of businesses was first performed based on transaction volume and average transaction amount, using Gaussian Mixture Models (GMM) and K-Means as unsupervised clustering algorithms. Once the businesses were segmented, three anomaly detection models were implemented and evaluated: One-Class SVM, Isolation Forest, and Local Outlier Factor (LOF). These models were selected for their ability to identify anomalous patterns in unlabeled data, which is suitable in this case due to the imbalance between churn and non-churn businesses.

Emphasis was placed on reducing false negatives, as the main goal is to identify all businesses at risk of churn to apply the retention process. At the same time, it is important to manage false positives, as, although they do not represent immediate churn, these businesses could be future churn candidates, allowing corrective actions to be taken in a timely manner.

J. Asitimbay and F. Grijalva are with Universidad San Francisco de Quito USFQ

II. Related Work

Anomaly detection has been extensively studied in diverse contexts, such as fraud detection and the identification of patterns in highly imbalanced datasets, which bear some similarity to our churn detection project. In this work, we integrate clustering techniques to segment businesses, allowing for the analysis of groups of companies with common transactional behaviors, and anomaly detection to identify businesses at risk of churn within payment networks that use POS terminals. This section reviews relevant prior works for the methods and objectives of this study.

A. Anomaly Detection

Anomaly detection has been widely applied in areas such as fraud detection and the analysis of unusual behaviors. Singh et al. [1] and Naaz et al. [2] demonstrated the effectiveness of Isolation Forest (IF) and Local Outlier Factor (LOF) in detecting fraud in credit card transactions, highlighting the importance of tuning hyperparameters such as the contamination level to balance recall and precision. Similarly, Ullah et al. [3] applied LOF and CBLOF for churn prediction in banking systems, emphasizing LOF's capability to identify local density deviations, which is crucial for detecting anomalous customer behaviors.

Isolation Forest, specifically designed for outlier detection, uses isolation trees to iteratively partition data and measure how quickly instances are isolated. This makes it computationally efficient and robust against imbalanced datasets [4], [5], [6]. In contrast, LOF measures the local density of points and compares it to the density of their immediate neighbors, enabling the detection of outliers in dense areas of the space [1], [7], [8]. On the other hand, One-Class SVM explicitly models normal points as a hyperplane in a feature space, classifying any deviations as anomalies. This approach is particularly useful in scenarios where anomalous data points are extremely rare or nonexistent [4], [9], [10].

In unconventional contexts, Usman et al. [4] applied algorithms such as IF and One-Class SVM to detect anomalous patterns in earthquake data, emphasizing the need to balance false positives and negatives. On the other hand, ELHadad et al. [5] and Goldstein et al. [11] compared IF and LOF for anomaly analysis in power consumption and multivariate data, respectively, highlighting the adaptability of these algorithms to different dataset characteristics. Additionally, Naaz et al. [2] explored the joint use of LOF and Isolation Forest for detecting fraudulent credit card transactions, showcasing their effectiveness in high-dimensional datasets.

B. Clustering as a Preprocessing Step for Anomaly Detection

Clustering methods, such as K-Means and Gaussian Mixture Models (GMM), are often employed as preprocessing steps to group data into homogeneous sets, facilitating more focused anomaly detection [12]. These methods allow the segmentation of data into groups that share similar characteristics, which is crucial for tailoring strategies in churn detection. K-Means is a simple and scalable method that minimizes within-cluster variance but may fail with non-spherical data shapes or complex distributions [7]. In contrast, GMM models data as a combination of Gaussian distributions, making it more flexible for representing clusters of varying shapes and sizes [13], [12].

Sugiharto et al. [7] demonstrated the utility of GMM and K-Means for customer segmentation in marketing, while Lipeng et al. [13] employed clustering to address data imbalance issues in churn prediction for telecommunications. Similarly, He et al. [10] introduced a cluster-based outlier detection approach, demonstrating its effectiveness in identifying anomalies in datasets with multiple clusters. Budiarto et al. [14] emphasized the importance of clustering as a foundation for outlier detection, using K-Means to segment data before applying algorithms such as LOF and One-Class SVM. These approaches highlight how segmentation can improve the precision of anomaly detection models by narrowing the analysis to groups with similar behaviors.

C. Techniques for Imbalanced Datasets

Handling imbalanced datasets is a critical challenge in anomaly detection and churn prediction. Sundarkumar and Ravi [9] proposed an approach that combines undersampling with One-Class SVM to address this imbalance, achieving promising results in churn prediction. Singh et al. [1], Forhad et al. [6], and Varmedja et al. [15] also explored resampling techniques and hyperparameter tuning strategies to enhance the performance of models like IF and LOF. Similarly, Goldstein et al. [11] performed a comparative evaluation of unsupervised anomaly detection algorithms on multivariate data, underscoring the importance of selecting appropriate evaluation metrics for imbalanced datasets.

In this project, instead of implementing resampling techniques, we chose to train models directly on imbalanced data, preserving its natural distribution. This approach allows us to evaluate model performance under conditions more representative of the real POS network environment, especially in contexts where churn patterns vary from month to month.

D. Contributions of the Present Study

This project combines GMM-based clustering with anomaly detection algorithms such as LOF, IF, and One-Class SVM, adapting these approaches to the specific context of churn prediction in POS networks. Unlike previous studies, we emphasize the integration of domain-specific features, such as transaction types and payment networks, to improve detection accuracy and retention strategies. Future research could explore temporal analysis and dynamic behavior within clusters to further enhance the identification of businesses at risk.

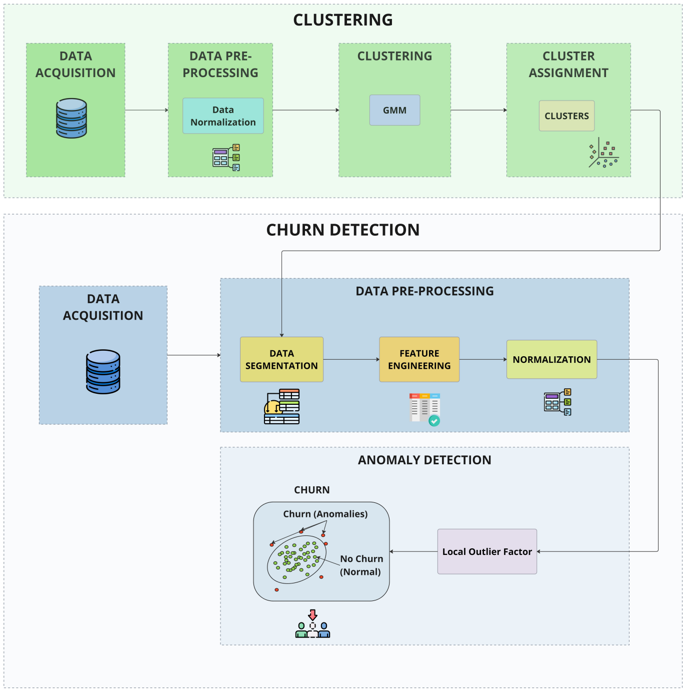

Figure 1. Workflow diagram of the proposed methodology.

III. Methodology

A. *Overview of the Methodology*

The proposed methodology aims to identify businesses that use POS payment networks and are at risk of churn or have already exhibited this behavior. It is important to note that, in this context, churn not only refers to businesses that have ceased operations but also to those that do not fully process transactions with all available cards.

The analysis is divided into two main parts: clustering and anomaly detection, as shown in Figure 1. These two stages are necessary due to the heterogeneity of the businesses analyzed: large chains and hospitals cannot be directly compared to small businesses or startups. For this reason, segmentation (clustering) is first performed to group businesses with similar characteristics. Subsequently, anomaly detection is applied within each segment, allowing each business to be analyzed in the context of its peers.

For the clustering stage, transactional and monthly average amounts of each business are used, considering data from approximately the last two years. It is important to mention that this analysis and process were carried out in collaboration with the commercial departments, following their suggestions and feedback.

B. *Data Preprocessing*

The first step involves data cleaning. In this stage:

- Businesses managed specially by the commercial area for retention purposes are removed, as they could introduce noise into the analysis and bias the results.

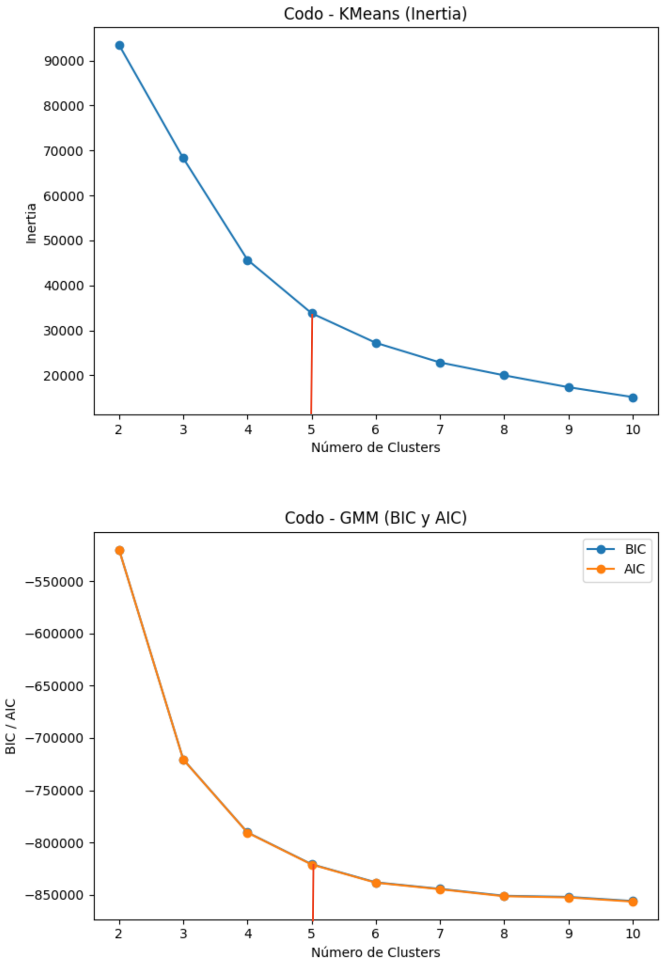

Figure 2. Elbow Method used to determine the optimal number of clusters.

- Data normalization is applied to ensure the compatibility of variables across clustering models.

C. *Clustering*

Clustering algorithms K-Means and Gaussian Mixture Models (GMM) are applied, testing different numbers of clusters, ranging from 2 to 7. The evaluation of these models is carried out using:

- **K-Means**: The elbow method is used, evaluating inertia (intra-cluster distance).
- **GMM**: The BIC and AIC criteria are employed.

Both methods agree that the optimal number of clusters is 5, as shown in Figure 2. Once $K = 5$ is defined, clustering is performed, and businesses are classified into the following categories: massive, large, medium, small, and micro, based on their transactional characteristics.

D. Anomaly Detection

In the second stage, anomaly detection is applied within each cluster using transactional data at the card and network level, considering approximately two years of information. This allows for the identification of businesses that appear to still be operating but do not process transactions with all cards, which is also considered a type of churn.

1) Preprocessing and Feature Engineering:

- Data cleaning and preprocessing are performed.
- Each business is assigned to its corresponding category based on the previous clustering results.
- Feature engineering is conducted, creating new variables such as:
  - Percentage of transactions by card type relative to the total.
  - Percentage of transactions by network.

  These new variables enrich the data and provide more information to the models.

2) Data Splitting:

- The data is divided into training (train) and testing (test) sets:
  - **Train**: Active businesses from the last 18 months (non-churn).
  - **Test**: Data from a recent month, reflecting an imbalanced and realistic scenario.
- **Normalization**: Once the data is split, normalization is applied to ensure that all variables are within the same range and compatible with the models. This is particularly important for algorithms such as One-Class SVM, LOF, and Isolation Forest, which are sensitive to the scale of variables.

E. Anomaly Detection Models

Three anomaly detection models were tested: One-Class SVM, Isolation Forest (IF), and Local Outlier Factor (LOF). Each model was tuned with different hyperparameter combinations to achieve the best possible performance. Below is a brief explanation of the key parameters for each model:

1) **One-Class SVM**:
   - `kernel`: Defines the type of decision boundary used for classifying anomalies. Common options include 'linear' (for linear separability) and 'RBF' (a Gaussian kernel for non-linear separability).
   - ν: A hyperparameter representing an upper bound on the fraction of training errors and a lower bound on the fraction of support vectors. Higher values increase the model's sensitivity to anomalies.
   - γ: Controls the influence of individual data points. Higher values lead to a more localized decision boundary, which is useful for handling complex data distributions.

   For more details, refer to [16].

2) **Isolation Forest (IF)**:
   - `n_estimators`: The number of isolation trees to construct. A higher number generally improves the model's stability but increases computational cost.
   - `max_samples`: The maximum number of samples used to train each isolation tree. It determines the subset size and influences the randomness of the splits.
   - `contamination`: An estimate of the proportion of anomalies in the dataset, used to set a threshold for classifying anomalies.
   - `max_features`: The maximum number of features considered for splitting at each node, helping balance accuracy and runtime.

   For more details, refer to [17].

3) **Local Outlier Factor (LOF)**:
   - `n_neighbors`: Determines the number of neighbors to use when calculating the local density. Larger values result in smoother density estimates.
   - `algorithm`: Defines the algorithm used to compute the nearest neighbors (e.g., 'auto', 'brute', 'ball_tree'), affecting speed and memory efficiency.
   - `leaf_size`: Used for tree-based methods like 'ball_tree' or 'kd_tree', influencing query efficiency.
   - `metric`: Specifies the distance metric (e.g., 'minkowski', 'euclidean') to calculate the neighborhood relationships.

   For more details, refer to [18].

Each model was evaluated using the test set, and key metrics such as the confusion matrix, recall, and area under the precision-recall curve (PR AUC) were obtained. These metrics were chosen due to their relevance in assessing the ability of the models to identify anomalies effectively.

As demonstrated in previous works, such as [1], [4], and [6], a parameter tuning approach was employed to test multiple values for these parameters and select the combination that yielded the best performance. This iterative process ensured that the models were optimized for the specific characteristics of each business segment.

F. Final Evaluation and Selection of the Best Model

The best-performing model is selected and saved for testing with data from a completely new month. This process includes:

1) Scaling the data for the new month.
2) Clustering businesses using the previously adjusted clustering models.

3) Applying feature engineering and scaling the new variables.
4) Using the selected anomaly detection model to identify businesses at risk.

This approach ensures that the models are tuned to a scenario as close to reality as possible within POS networks, improving their ability to detect churn in future scenarios.

IV. Results

A. Clustering

In the clustering stage, the results highlight the effectiveness of Gaussian Mixture Models (GMM) over K-Means. As shown in Figure 3, GMM achieved a clearer separation between clusters, while K-Means showed limitations in differentiating certain groups, especially for the Micro segment, where a large portion of the data was concentrated in a single region (purple cluster).

The analysis identified five main business categories: Masivo, Grande, Mediano, Pequeño, and Micro. These categories were defined based on the average transaction amounts and monthly transaction volume, allowing segmentation that aligned with the characteristics of the analyzed businesses. The results were validated with the commercial team, who confirmed that the groupings reflected significant and consistent differences based on their experience.

B. Anomaly Detection

For the anomaly detection stage, the methods One-Class SVM, Isolation Forest (IF), and Local Outlier Factor (LOF) were evaluated using transactional data for each cluster. Key metrics for evaluation included Precision-Recall AUC, recall, and the confusion matrix, with a strong emphasis on minimizing false negatives, as this is critical for churn detection.

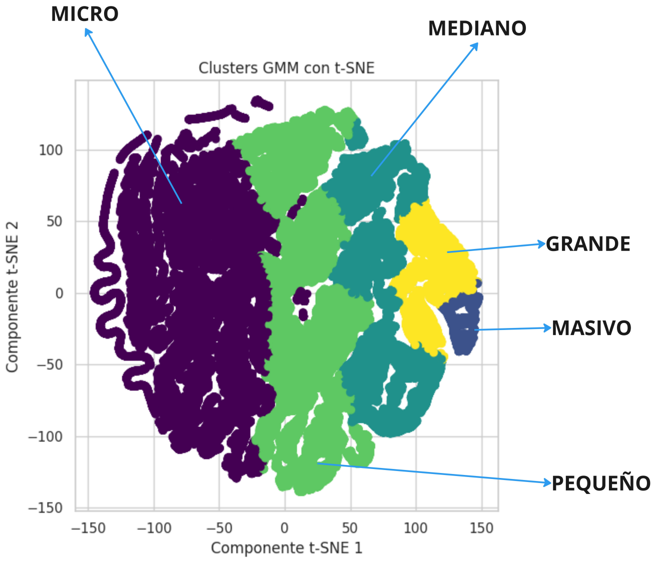

Figure 3. Visualization of clusters obtained with GMM.

Table I
Precision-Recall AUC comparison across anomaly detection methods.

Segment	One-Class SVM	Isolation Forest	LOF
Masivo	0.43	0.12	0.71
Grande	0.71	0.54	0.52
Mediano	0.28	0.21	0.53
Pequeño	0.10	0.23	0.96
Micro	0.15	0.11	0.99

Table II
LOF parameters for each business segment.

Segment	n_neighbors	algorithm	leaf_size	metric	contamination
Masivo	20	auto	30	minkowski	auto
Grande	40	auto	30	minkowski	auto
Mediano	30	brute	10	minkowski	auto
Pequeño	20	auto	30	minkowski	auto
Micro	40	auto	30	minkowski	auto

1) Model Comparison: The results obtained from evaluating the three methods across each business segment are presented in Table I. As shown, LOF outperformed the other models in most segments, achieving an optimal balance between precision and recall. Although One-Class SVM achieved nearly perfect recall (indicating almost no false negatives), the high number of false positives affected its operational efficiency. On the other hand, LOF maintained competitive recall levels while significantly reducing false positives.

2) Final Metrics with LOF: The LOF model was fine-tuned for each business segment by adjusting key parameters such as the number of neighbors and the contamination level. Table II summarizes the final hyperparameters used for each segment. These parameters were carefully selected to balance precision and recall across the various business segments.

Using these optimized parameters, the final performance metrics for each business segment were obtained, as shown in Table III. Below is a detailed breakdown of the results:

- **Micro and Pequeño Segments:**
  - Achieved the highest Precision-Recall AUC values: 0.99 for Micro and 0.96 for Pequeño.
  - Precision: Very high (0.94 for Micro, 0.79 for Pequeño), indicating a high probability that detected anomalies are indeed true anomalies.
  - Recall: Excellent (0.99 for Micro, 0.95 for Pequeño), capturing almost all true anomalies within these segments.
  - F1-Score: High (0.96 for Micro, 0.86 for Pequeño), reflecting a strong balance between precision and recall.
  - Accuracy: Consistently high at 0.99 for both segments, confirming overall model reliability.
- **Masivo, Grande, and Mediano Segments:**
  - Lower Precision-Recall AUC values: 0.71 for Masivo, 0.52 for Grande, and 0.53 for Mediano.
  - Precision: Comparatively lower, ranging from 0.27 to 0.48, indicating a higher rate of false positives in these segments.

Table III
Final LOF metrics for each business segment.

Segment	Precision-Recall AUC	Precision	Recall	F1-Score	Accuracy
Masivo	0.71	0.32	1.00	0.48	0.98
Grande	0.52	0.48	0.96	0.64	0.99
Mediano	0.53	0.27	0.95	0.42	0.98
Pequeño	0.96	0.79	0.95	0.86	0.99
Micro	0.99	0.94	0.99	0.96	0.99

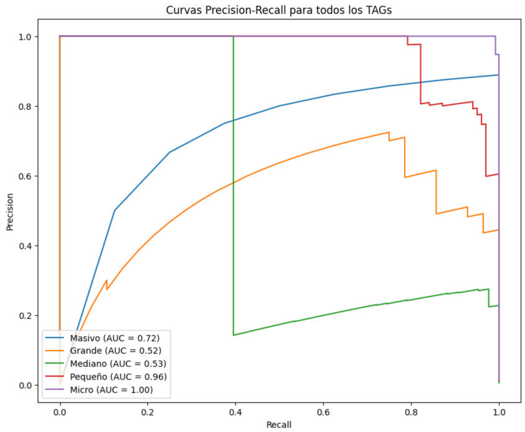

Figure 4. Precision-Recall curve for the LOF model across all business segments.

  - Recall: Consistently high across all three segments (1.00 for Masivo, 0.96 for Grande, 0.95 for Mediano), ensuring that most true anomalies are detected.
  - F1-Score: Lower values (0.48 for Masivo, 0.64 for Grande, 0.42 for Mediano), reflecting the trade-off between precision and recall due to data imbalance.
  - Accuracy: High overall, at 0.98 for Masivo and Mediano, and 0.99 for Grande, showing the robustness of the model despite challenges in precision.

Additionally, the Precision-Recall curve, depicted in Figure 4, provides a comprehensive view of the trade-off between precision and recall across different thresholds. This visualization highlights key insights into the performance of the LOF model across all business segments:

- **Micro - Pequeño:**
  - The curves for Micro and Pequeño segments show a near-optimal balance between precision and recall, with the Micro segment achieving an almost perfect curve (AUC = 1.00) and the Pequeño segment closely following (AUC = 0.96).
  - These results confirm that the LOF model is highly effective at detecting churn in these segments, which comprise the majority of businesses and are critical for operational decision-making.
- **Masivo - Grande:**
  - The Masivo and Grande segments exhibit lower AUC values (0.72 for Masivo and 0.52 for Grande), primarily due to their smaller number of businesses and true churn cases, which leads to a higher proportion of false positives.
  - However, as highlighted in the conclusions, the higher false positive rates in these segments are less critical because the absolute number of businesses in these categories is relatively low. This allows for more targeted and resource-efficient follow-up actions by the retention team.
- **Mediano:**
  - The Mediano segment shows moderate performance with an AUC of 0.53. This segment often includes businesses with varied transactional behaviors, which can make precise anomaly detection more challenging.

V. Conclusions

This project initially explored supervised approaches using neural networks like LSTM and GRU. However, due to the abrupt and non-sequential nature of changes in business behavior, anomaly detection techniques were found to be more effective. These methods enabled the identification of outliers without relying on temporal sequences.

The clustering stage was crucial, not only for improving model performance but also for customizing retention strategies. Segmentation allowed for tailored approaches to large chains and small businesses, adapting the actions to the specific characteristics of each group.

Feature engineering played a key role in improving model metrics. Initially, models without engineered features failed to achieve satisfactory results. However, introducing domain-specific variables, such as the percentage of transactions by card type and network, significantly reduced both false positives and false negatives, particularly in the Micro and Pequeño segments, where most businesses are concentrated.

While the Masivo, Grande, and Mediano segments showed a higher proportion of false positives relative to true churn cases, these absolute values were low due to the stability of these businesses. These false positives enable the implementation of preventive actions, improving customer relationships and loyalty.

Extending the analysis period from 1 year to 2 years helped capture a broader range of transactional behaviors, enhancing the consistency of the model. The final validation with an entirely new month confirmed the effectiveness of the approach, identifying at-risk businesses that could be targeted with specific retention strategies.

VI. Limitations and Future Work

This study demonstrated the effectiveness of anomaly detection in identifying businesses at risk of becoming churners. However, several limitations were encountered during the project.

The first challenge was data acquisition, which required extensive coordination with the commercial and retention departments to align all stakeholders toward a common goal. These are common difficulties in cross-functional projects but required significant effort to overcome.

The second limitation was the imbalance in the dataset. This imbalance made supervised approaches, such as using LSTM and GRU models, less effective, leading us to adopt anomaly detection methods as a more practical solution.

The third limitation was the response capacity of the retention department. As mentioned earlier, One-Class SVM showed better results in minimizing false negatives, but the substantial increase in false positives significantly restricted the department's ability to act effectively. On the other hand, LOF demonstrated not only a more balanced performance but also computational efficiency, which makes it a practical choice for large-scale or time-sensitive analyses.

For future work, we propose evaluating these analyses on a monthly basis to observe the outcomes of the initial implementations and testing the feasibility of transitioning to a weekly analysis. Additionally, future research could focus on analyzing specific business types or industries. This approach would allow us not only to identify at-risk businesses but also to uncover opportunities for expanding into new business segments and acquiring new clients.

References

[1] P. Singh *et al.*, "Anomaly detection classifiers for detecting credit card fraudulent transactions," in *2024 4th International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 2024.

[2] S. Naaz and H. John, "Credit card fraud detection using local outlier factor and isolation forest," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 4, pp. 1060–1064, 2019. [Online]. Available: https://doi.org/10.26438/ijcse/v7i4.10601064

[3] I. Ullah *et al.*, "Churn prediction in banking system using k-means, lof, and cblof," in *International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, 2019.

[4] M. Usman *et al.*, "Comparative analysis of elliptic envelope, isolation forest, one-class svm, and local outlier factor," *International Journal of Computer Science and Information Technology (IJCSIT)*, 2019.

[5] A. ELHadad *et al.*, "Comparison of enhanced isolation forest and enhanced local outlier factor in anomalous power consumption labelling," in *International Conference on Electrical, Communication, and Computer Engineering (ICEECE)*, 2022.

[6] N. Forhad *et al.*, "Churn analysis: Predicting churners," in *Telecom Analytics*, 2014.

[7] N. Sugiharto *et al.*, "Mall customer clustering using gaussian mixture model, k-means, and birch algorithm," in *International Conference on Information and Communication Technology (ICOIACT)*, 2023.

[8] M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," *ACM SIGMOD International Conference on Management of Data*, pp. 93–104, 2000. [Online]. Available: https://doi.org/10.1145/335191.335388

[9] G. G. Sundarkumar *et al.*, "One-class support vector machine based undersampling," *IEEE Transactions on Information Forensics and Security*, 2015.

[10] Z. He, X. Xu, and S. Deng, "Discovering cluster-based local outliers," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 1641–1650, 2003.

[11] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLOS ONE*, vol. 11, no. 4, p. e0152173, 2016.

[12] A. A. Patel, *Hands-On Unsupervised Learning Using Python*, 1st ed. Sebastopol, CA, USA: O'Reilly Media, Inc., 2019.

[13] L. Lipeng *et al.*, "Telecom customer churn prediction based on imbalanced data re-sampling method," in *IMC*, 2013.

[14] E. H. Budiarto *et al.*, "Unsupervised anomaly detection using k-means, local outlier factor and one-class svm," 2019.

[15] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit card fraud detection - machine learning methods," in *2019 18th International Symposium on INFOTEH-JAHORINA (INFOTEH)*. IEEE, 2019, pp. 1–5.

[16] "One-class svm — scikit-learn documentation," https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html, accessed: 2024-11-20.

[17] "Isolation forest — scikit-learn documentation," https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html, accessed: 2024-11-20.

[18] "Local outlier factor — scikit-learn documentation," https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html, accessed: 2024-11-20.