Dániel Galacz

Louis Urbin

First Project

Classe di Scienze
Scuola Normale Superiore
Pisa, april 2025

Prof. Fosca Giannotti
Prof. Roberto Pellungrini

Contents

Introduction

This project uses the *Telco Customer Churn* dataset from Kaggle. We discuss data understanding, preprocessing strategies, clustering, classification, and pattern mining techniques to analyze customer churn.

The goal of this project is to explore the dataset and apply clustering and classification techniques with the aim of identifying customers who are likely to churn. In the context of this dataset, the commercial objective is to reduce the churn rate by identifying customers at risk of leaving. The most relevant metric for this task is recall, as it measures how well the model identifies actual churners.

Recall is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

In the case of customer churn prediction, it is more important to correctly identify customers who will churn (True Positives) than to avoid false alarms (False Positives). A high recall ensures that most of the customers at risk of leaving are flagged, allowing the company to intervene and reduce churn. However, this may come at the cost of a higher number of false positives, which could lead to unnecessary marketing efforts for some customers.

Thus, our focus is on maximizing recall to minimize the risk of losing valuable customers.

1

Data Exploration and Preprocessing

We loaded data for over 7,000 customers, including both numerical and categorical features, as well as some missing values. The `TotalCharges` column, initially of type `object`, was converted to numeric using median imputation. We chose the median over the mean because the mean is more sensitive to extreme values.

There are no duplicate values in the `customerID` column, ensuring that each customer is uniquely represented. The other categorical variables appear logically consistent and do not contain missing values.

To explore the structure of the data, we began by plotting feature distributions:

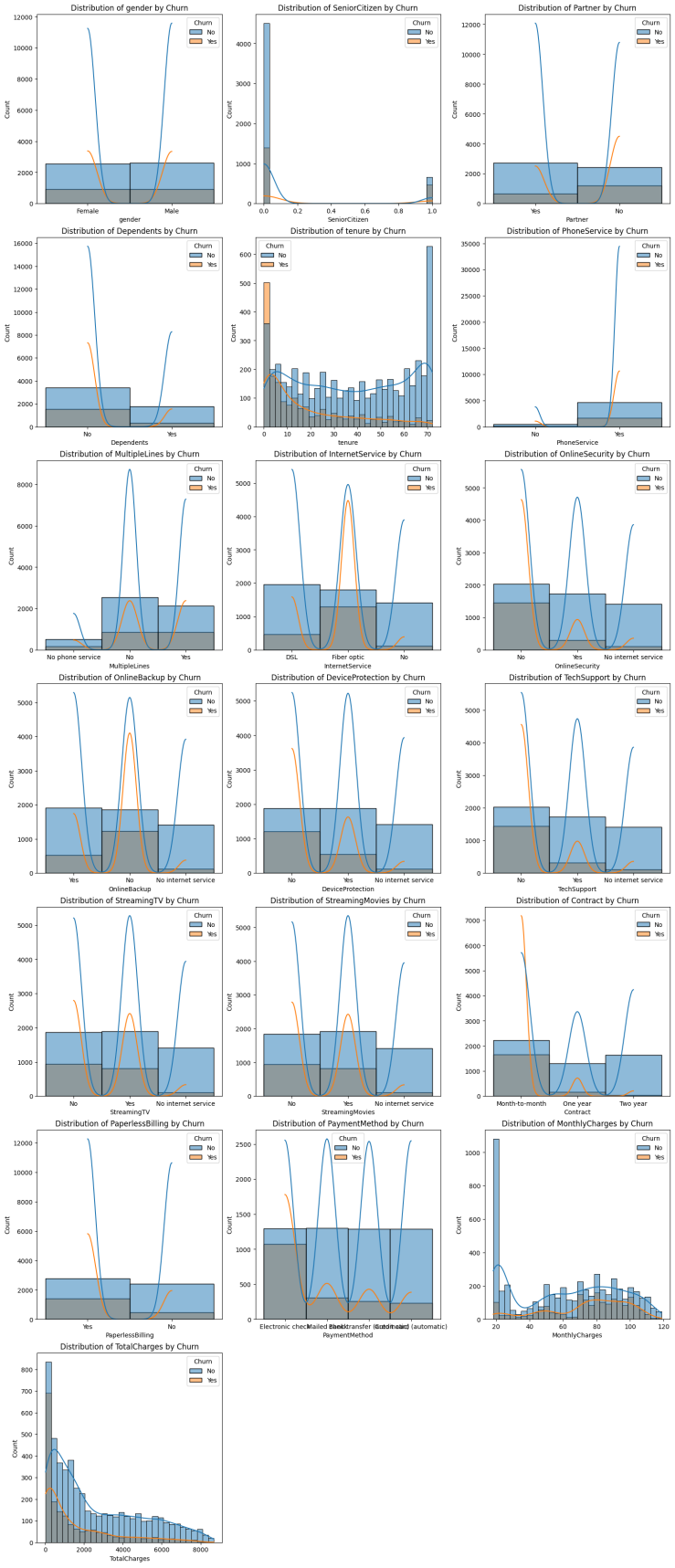
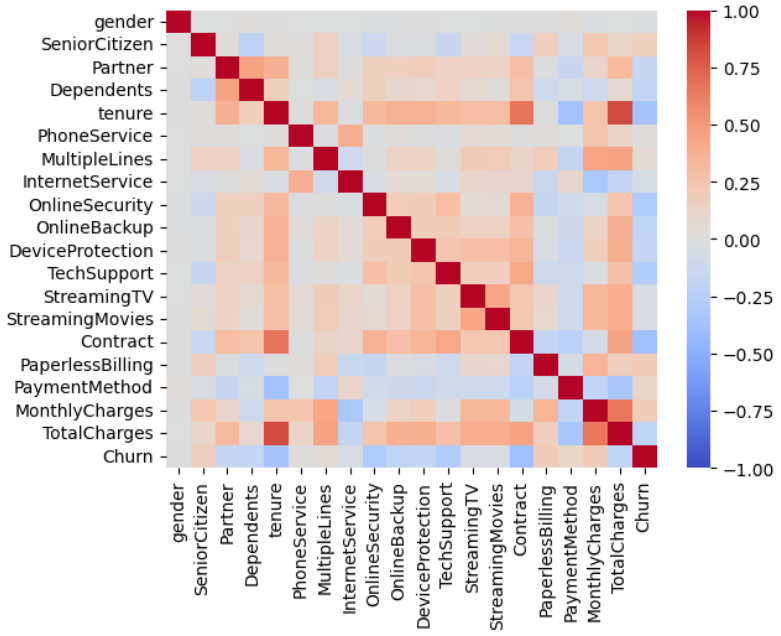


Figure 1.1: Distribution of features

Certain features clearly stand out as more discriminative. Notably, `tenure`, `InternetService`, `OnlineSecurity`, `OnlineBackup`, `DeviceProtection`, `TechSupport`, `StreamingTV`, `StreamingMovies`, `Contract`, and `PaymentMethod` appear to be the most informative in distinguishing between customers who churn and those who stay.

We further examined the relationships between numerical features by visualizing their correlation matrix. Overall, we observed only weak correlations.

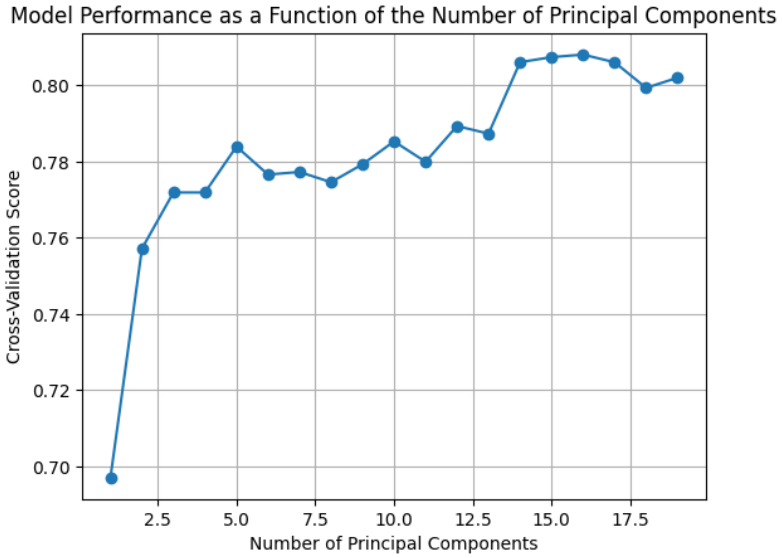


Before diving into classification tasks, we wanted to reduce dimensionality while preserving as much information as possible. To do this, we applied Principal Component Analysis (PCA). But first, a series of preprocessing steps was necessary:

- Categorical features were encoded using `LabelEncoder`.

- All features were scaled using `MinMaxScaler` to bring values into the [0, 1] range.

With this preparation complete, we applied PCA and evaluated the optimal number of components by observing the recall score of a logistic regression model using cross-validation. This choice is motivated by the fact that logistic regression turns out to be one of the most effective models for our classification task—as shown later in the report.



The results suggest that retaining around 14 principal components achieves a good balance between dimensionality reduction and predictive performance. Notably, the first step at 3 components already yields very good results. However, the low computational complexity allows us to select 14 components, gaining an additional 4 points in recall.

2

Clustering

We applied several clustering algorithms (KMeans, DBSCAN, and Hierarchical clustering) on the scaled dataset, using PCA for 2D visualizations.

We began with KMeans. The silhouette score was used to determine the optimal number of clusters.

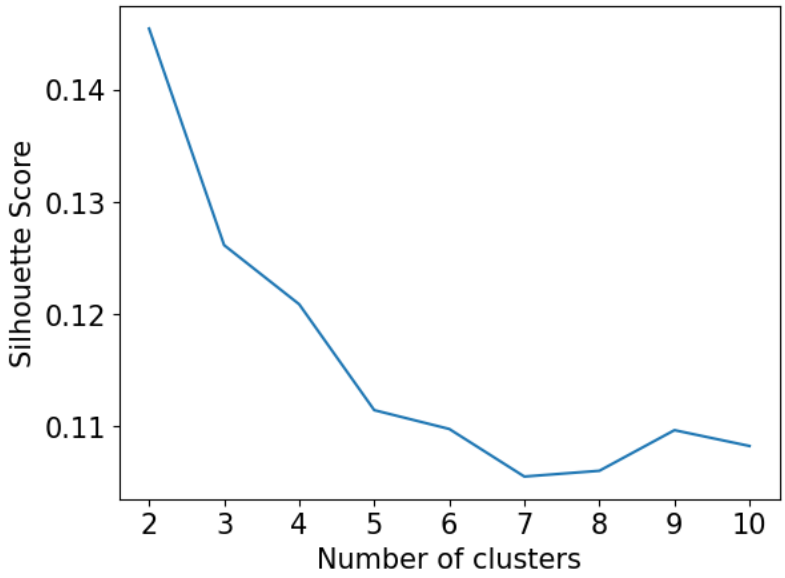
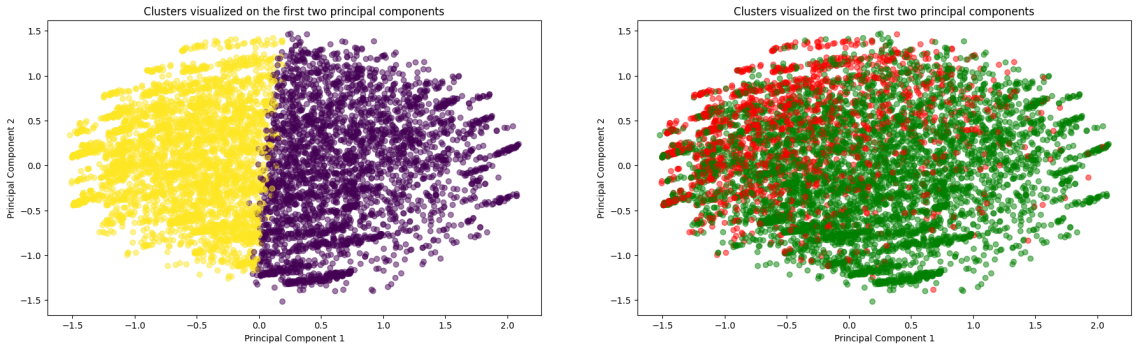


Figure 2.1: Silhouette score by number of clusters (KMeans)

From the above, the optimal number of clusters appears to be either 3 or 4. However, the overall silhouette score remains low, suggesting poorly defined clusters. This impression is confirmed by visualizing the clustering output on the first two principal components.



(a) (b)

Figure 2.2: KMeans clustering results and churn distribution (PCA projection)

Even when projecting into additional PCA dimensions, the clusters remain indistinct and overlapping. This confirms that KMeans fails to reveal meaningful structure in our data.

Next, we experimented with DBSCAN. The ε parameter was set to approximately 1.1 based on the k-distance plot below:

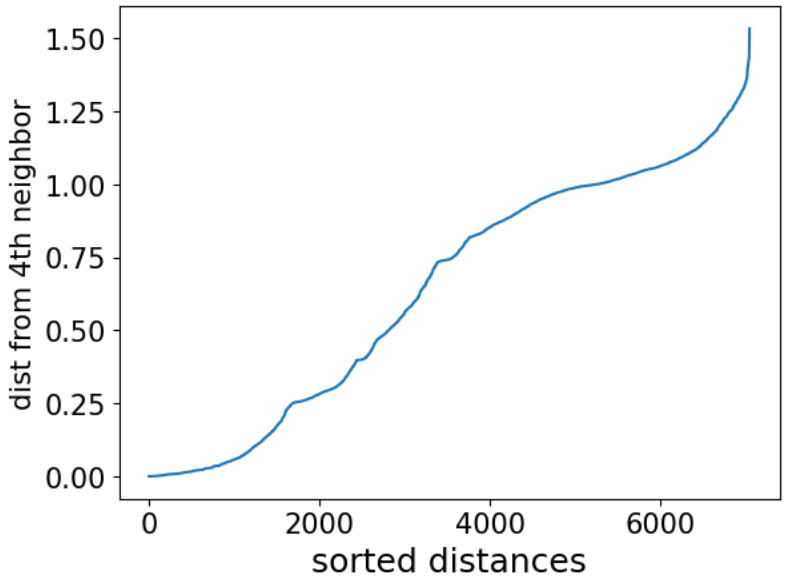
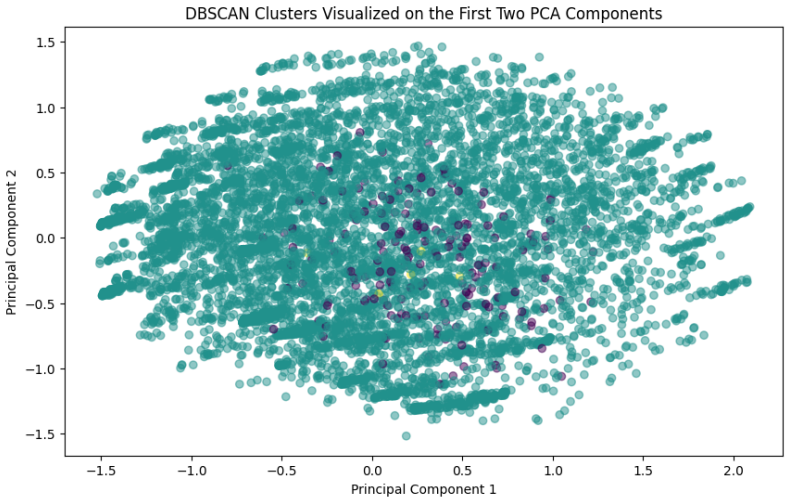


Figure 2.3: k-distance plot used to select ε for DBSCAN

With this setting, DBSCAN identified three clusters, shown below:



The associated silhouette score was only 0.009, indicating extremely poor clustering performance.

Finally, we turned to hierarchical clustering. We tested several linkage criteria, with `complete` and `average` performing slightly better than others. Using the complete linkage method, we generated the following dendrogram:

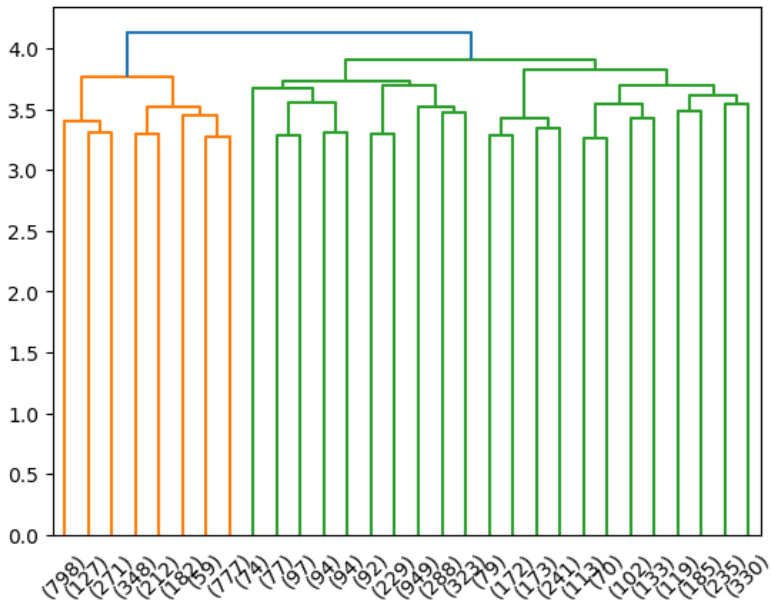
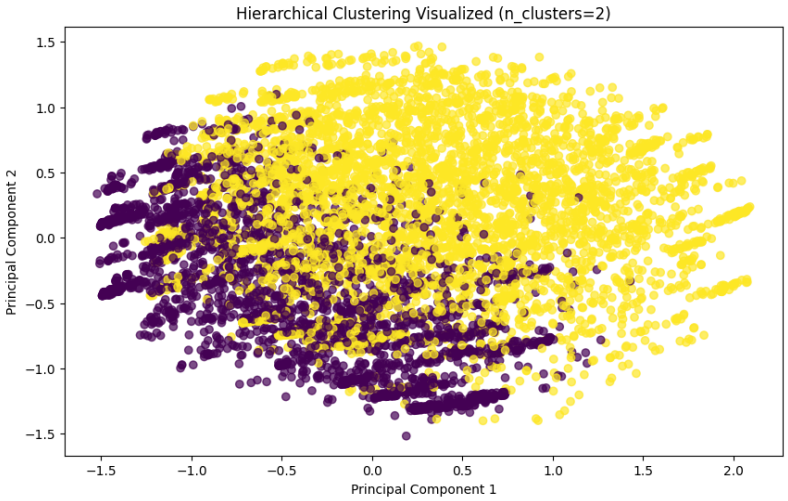


Figure 2.4: Dendrogram (complete linkage)

And obtained the following cluster distribution:



The silhouette score was 0.06—still very low.

In conclusion, clustering does not seem to be a suitable approach for this dataset. None of the methods applied (KMeans, DBSCAN, or Hierarchical clustering) were able to produce well-separated, meaningful clusters that distinguish churned customers from others.

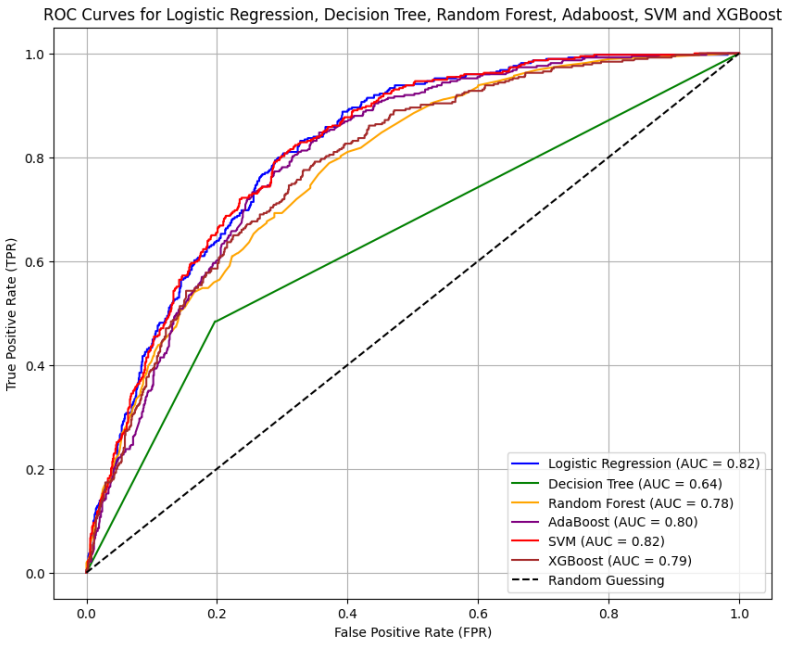
3

Classification

We now consider the problem as a classification task, and for that, we will compare the performance of several models. Note that throughout the following, we will use `class_weight='balanced'` to balance the classes.

To evaluate the different models, we will rely on the confusion matrix, F1, recall, and precision scores, as well as the ROC curve and AUC scores.

First, here are the ROC curves for the 6 models tested:



As we can see, two models stand out: Logistic Regression and SVM. The boosted algorithms are not as effective here. We also note that the Random Forest model is behind the other models, but considering it within a forest greatly improves the result.

Below, we detail the scores and confusion matrices for the two best-performing mod-

els:

	precision	recall	f1-score	support
0	0.91	0.70	0.79	1035
1	0.49	0.80	0.61	374
accuracy			0.73	1409
macro avg	0.70	0.75	0.70	1409
weighted avg	0.80	0.73	0.74	1409

```
[[724 311]
 [ 73 301]]
```

(a) Logistic Regression

	precision	recall	f1-score	support
0	0.91	0.68	0.78	1035
1	0.48	0.82	0.61	374
accuracy			0.71	1409
macro avg	0.70	0.75	0.69	1409
weighted avg	0.80	0.71	0.73	1409

```
[[699 336]
 [ 66 308]]
```

(b) SVM

Figure 3.1: Classification report and confusion matrix for Logistic Regression and SVM

Logistic Regression and SVM appear to be the better models: we achieve a recall score of 0.82, which is great! We don't need the boosted algorithms to get the best results.

Another idea is to create new features from the continuous features, for instance, by creating buckets—splitting the values into several ranges. However, this does not improve performance. To test this, simply execute the last block of code and then rerun the previous classification blocks.

4

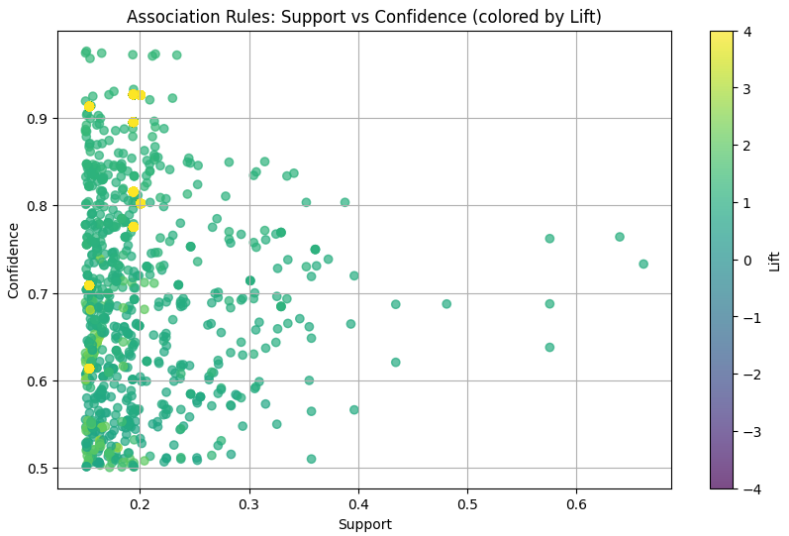
Pattern Mining

We applied association rule mining to uncover meaningful patterns related to customer churn.

First, we converted the dataset into a transactional format, where each row represented a customer and each column corresponded to a feature or service. Using this format, we applied the `Apriori` algorithm to identify frequent itemsets.

From these frequent itemsets, we generated association rules based on *support*, *confidence*, and *lift*. These rules helped highlight combinations of features that were often observed together and related to churn behavior.

We visualized the rules in a support–confidence scatter plot, with lift represented by color.



To focus on the most relevant patterns, we selected the top rules based on a combination of high lift, confidence, and support.

Antécédent	Conséquent	Support	Confiance	Lift
(InternetService=2)	(DeviceProtection=1, Churn=0)	0.200625	0.925950	4.615334
(DeviceProtection=1)	(MonthlyCharges_bucket=1, Churn=0)	0.200625	0.925950	4.169736
(MonthlyCharges_bucket=1)	(DeviceProtection=1, Churn=0)	0.200625	0.801930	3.997162
(InternetService=2, MultipleLines=0)	(DeviceProtection=1, Churn=0)	0.153486	0.913007	4.550819
(InternetService=2, SeniorCitizen=0)	(DeviceProtection=1, Churn=0)	0.193951	0.926730	4.619221
(InternetService=2)	(DeviceProtection=1, Churn=0, SeniorCitizen=0)	0.193951	0.895151	4.615334
(DeviceProtection=1, MultipleLines=0)	(MonthlyCharges_bucket=1, Churn=0)	0.153486	0.913007	4.111449
(DeviceProtection=1, SeniorCitizen=0)	(MonthlyCharges_bucket=1, Churn=0)	0.193951	0.926730	4.173248
(MonthlyCharges_bucket=1, SeniorCitizen=0)	(DeviceProtection=1, Churn=0)	0.193951	0.815522	4.064914
(DeviceProtection=1)	(MonthlyCharges_bucket=1, Churn=0, SeniorCitiz...	0.193951	0.895151	4.200231

Figure 4.1: Top association rules by Lift, Confidence, and Support

Many rules were overlapping or redundant, but one insight stood out:

> Among fiber optic customers (InternetService = 2), those who had Device Protection were significantly less likely to churn.

This suggests that offering Device Protection to other fiber customers could be an effective strategy to reduce churn in this segment.

Conclusion

In this study, we approached the customer churn prediction problem as a classification task, applying various machine learning models. Logistic Regression and Support Vector Machine (SVM) emerged as the top performers, achieving high recall scores—critical for identifying customers at risk of churning. Despite evaluating several more complex models, including ensemble and boosting methods, these simpler approaches consistently delivered better results.

Dimensionality reduction with PCA and clustering techniques did not yield interpretable groupings, and engineered features based on discretizing continuous variables failed to improve model performance.

Beyond predictive modeling, we employed association rule mining to uncover behavioral patterns linked to churn. While many rules were redundant, one insight stood out: fiber optic customers with Device Protection were significantly less likely to churn. This finding highlights a potential business strategy—promoting Device Protection to other fiber customers—as a targeted churn mitigation measure.

Overall, Logistic Regression and SVM proved to be the most effective models for predicting customer churn.