Nonlinear Churn Detection via Entropy, UMAP, and Clustering

CRL-J

March 28, 2025

Abstract

In this project, we explore churn detection from a nonlinear and information-theoretic perspective. When PCA failed to reveal meaningful structure, we introduced entropy as a lens into customer behavior complexity. By combining entropy with UMAP projections and density-based clustering, we built a semantic segmentation of churn risk that outperformed traditional linear modeling and supported explainable decision-making.

1 Dataset Description

The data used in this project comes from the publicly available Telco Customer Churn Dataset hosted on Kaggle. It contains demographic and service usage information for **7,043** customers of a fictional telecommunications company.

Each row in the dataset corresponds to one customer, with features including:

- **Demographics:** gender, senior citizenship, partnership status, dependents
- **Service usage:** internet service type, online security, streaming TV and movies, tech support, multiple lines
- **Account information:** contract type, paperless billing, payment method, tenure, monthly and total charges

The target variable is `Churn`, a binary label indicating whether the customer discontinued service during the observation period. Approximately **26.5%** of customers in the dataset churned, making this a moderately imbalanced classification problem.

This dataset presents a realistic and richly structured challenge for churn detection: customer behavior is heterogeneous and influenced by a combination of pricing, contract type, digital engagement, and support access. While simple models may detect some churn patterns, our goal was to explore the nonlinear structure of behavior space to better segment and anticipate churn risk.

2 Introduction

Customer churn is a critical concern in subscription-based businesses. Accurately identifying customers at risk of leaving allows companies to take proactive measures, reduce revenue loss, and improve customer retention.

Traditional churn modeling approaches often rely on linear models or feature importance rankings derived from supervised machine learning. However, these models may fail to capture the complex, nonlinear structure of customer behavior—especially in datasets where decision boundaries are not linearly separable.

In this project, we began by exploring a well-known churn dataset using Principal Component Analysis (PCA). While PCA is effective at dimensionality reduction, it did not reveal meaningful clustering or separability by churn label. This led us to explore a new hypothesis: that churn behavior is better understood through the lens of information theory and nonlinear geometry.

We introduced per-customer Shannon entropy as a feature to quantify behavioral disorder, capturing the degree to which a customer's profile deviates from uniform or predictable patterns. We then applied Uniform Manifold Approximation and Projection (UMAP) to uncover nonlinear relationships in the enriched data. Clustering in this space using DBSCAN and HDBSCAN revealed coherent behavioral groups with differing churn rates.

From these insights, we developed a churn risk scoring method based on entropy, cluster-level churn rates, and HDBSCAN membership strength. We then trained a predictive model to recover this risk score from the original features, achieving strong results and validating the interpretability of our approach.

This paper documents the evolution of our method, beginning with a failed PCA and culminating in a full nonlinear churn detection pipeline. We present this framework as a general approach for discovering latent structure and constructing interpretable risk models when traditional techniques fall short.

3 Initial PCA Analysis

Principal Component Analysis (PCA) is a classical linear method for dimensionality reduction. It projects high-dimensional data onto a lower-dimensional subspace by maximizing variance along orthogonal axes. In this project, we applied PCA to the cleaned and encoded dataset to investigate whether customer behavior and churn status could be linearly separated.

We visualized the first two principal components in a scatter plot, coloring each point by churn status. As shown in Figure 1, churned and retained customers are heavily intermixed across the entire projection. Despite PCA capturing much of the dataset's variance, it failed to expose any clear clusters or separable regions.

This observation suggests that churn-related structure is not aligned with the directions of maximum linear variance—highlighting the need for nonlinear techniques.

4 Entropy Feature Engineering

Shannon entropy is a fundamental measure in information theory that quantifies the uncertainty or disorder of a probability distribution. In the context of churn prediction, we used entropy to capture how uniform or varied a customer's profile is across multiple features.

To compute this, we treated each customer's row as a discrete distribution of categorical and encoded numerical values. After normalizing the values within each row, we computed the entropy using the standard Shannon entropy formula:

$$H(x) = -\sum_{i} p_i \log_2 p_i \tag{1}$$

where p_i is the normalized frequency of value i in the customer's feature set.

The intuition behind this approach is that customers with highly regular or consistent behavior (e.g., identical service plans, contract lengths, and billing preferences) will have lower entropy, while customers with more varied or atypical profiles will exhibit higher entropy. We hypothesized that

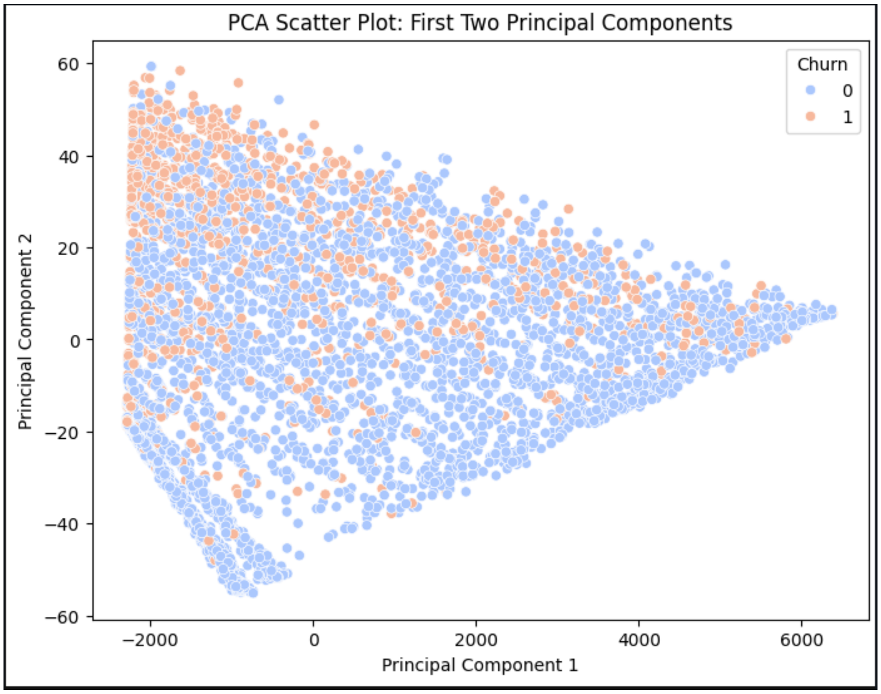

Figure 1: Scatter plot of customers projected onto the first two principal components using PCA. Each point is colored by churn status. The lack of visual separation indicates that linear projections fail to capture churn-relevant structure.

these higher-entropy customers may be closer to behavioral thresholds or discontinuities, making them more likely to churn.

Entropy was computed for each customer and stored as a new feature, which we call `ShannonEntropy`. Later visualizations showed that this feature aligned meaningfully with regions of churn risk in the UMAP projection, supporting its utility as a behavioral complexity indicator.

5 Nonlinear Projection with UMAP

To better understand the structure of customer behavior beyond the limitations of PCA, we employed Uniform Manifold Approximation and Projection (UMAP), a powerful nonlinear dimensionality reduction technique. UMAP is designed to preserve both local and global structure of high-dimensional data by modeling it as a fuzzy topological space.

The core idea behind UMAP is to first construct a weighted graph of local neighborhoods in the high-dimensional space using nearest neighbors, and then optimize a low-dimensional embedding that preserves these relationships as closely as possible. Unlike PCA, which is constrained to linear subspaces, UMAP is able to unfold complex, curved manifolds that may better represent the true structure of the data.

Mathematically, UMAP builds a high-dimensional fuzzy simplicial set to represent data topology, and then finds a low-dimensional representation by minimizing the cross-entropy between the two fuzzy sets. This allows it to maintain continuity in local neighborhoods while still revealing broader structure.

We applied UMAP to the full feature set, including the engineered Shannon entropy feature, to produce a 2D embedding. The result revealed clear clusters and transitions in customer behavior that were not visible in PCA space. When colored by churn label, these clusters showed variations in churn risk across regions of the projection, suggesting that UMAP had uncovered meaningful behavioral subspaces.

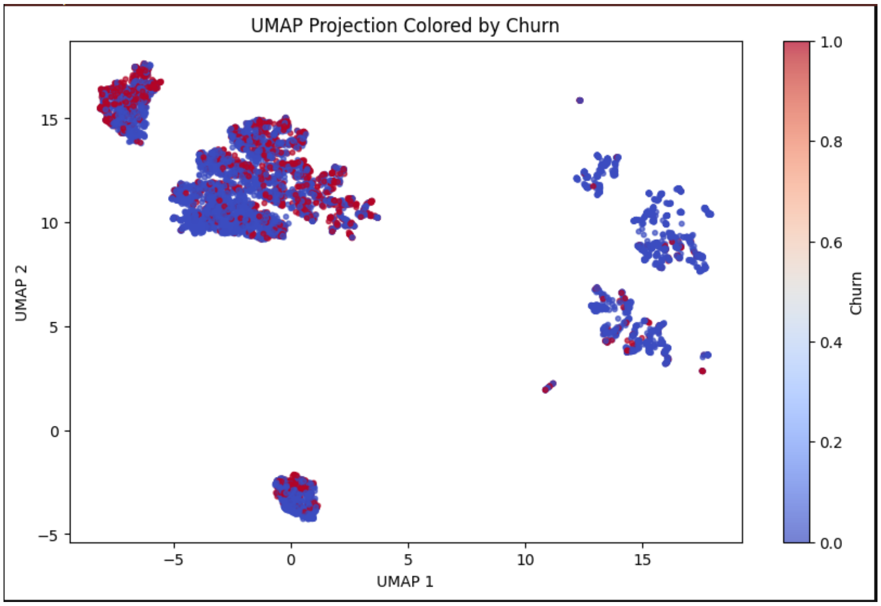

Figure 2: UMAP projection of customer data colored by churn label (0 = blue, 1 = red). UMAP reveals distinct nonlinear clusters in the data. Some clusters have noticeably higher churn rates, indicating churn is concentrated in specific regions of behavioral space.

6 Clustering in UMAP Space

Once the UMAP embedding revealed a rich nonlinear structure, we turned to clustering techniques to identify coherent groups of customers. Because the shape of the data in UMAP space was highly nonlinear and potentially noisy, we opted for density-based clustering methods that do not assume convex or isotropic cluster shapes.

6.1 DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clusters data by identifying regions of high density separated by regions of low density. It requires two parameters: `eps`, which defines the neighborhood radius, and `min_samples`, the minimum number of points required to form a dense region. Points not belonging to any dense region are labeled as noise.

When applied to the UMAP-projected data, DBSCAN revealed multiple distinct clusters. We observed that churn rates varied significantly across clusters, providing evidence that these geometric groupings captured meaningful differences in customer behavior.

7 Entropy-Based Churn Risk Score

Building on the features discovered in earlier steps, we constructed a heuristic churn risk score to capture the latent likelihood of churn based on behavioral complexity and cluster dynamics.

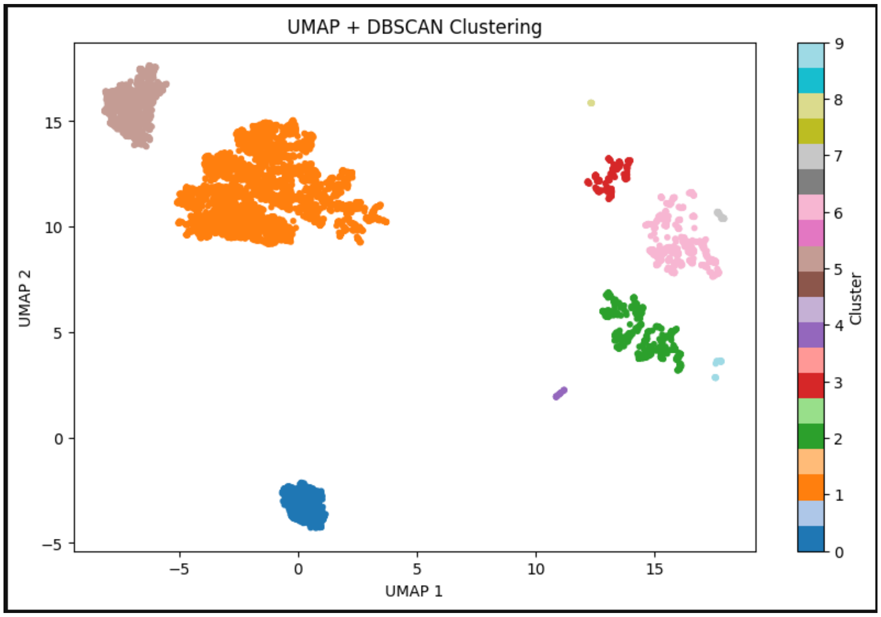

Figure 3: DBSCAN clustering applied to the UMAP projection of customer data. Each color represents a distinct cluster detected by DBSCAN. The clusters reflect nonlinear groupings in customer behavior space and form the basis for estimating churn rates by group.

The score combined three key components:

- **Shannon entropy** — representing behavioral irregularity
- **Cluster-level churn rate** — derived from DBSCAN or HDBSCAN
- **Membership strength** — indicating the confidence of cluster assignment (low strength implies fuzziness or liminality)

The intuition behind the score is that customers with high entropy (unusual behavior), located in high-churn clusters, and weakly affiliated with their cluster (fuzzy boundary behavior) are the most likely to churn. The risk score R was computed as a weighted combination:

$$R = \alpha H + \beta(1 - S) + \gamma C \tag{2}$$

where H is the normalized entropy, S is membership strength, and C is the average churn rate in the assigned cluster. We used default weights $\alpha = 0.4$, $\beta = 0.3$, and $\gamma = 0.3$, but the formula is easily tunable for business context.

We sorted customers by risk score and flagged the top 5% as `HighRisk`. This thresholding process provided a transparent and customizable method to identify churn-prone customers for potential outreach.

This risk score was later used as a soft target for training a churn risk classifier and served as a semantic link between data geometry, behavioral disorder, and customer outcomes.

8 Modeling Churn Risk

To evaluate the utility of the derived churn risk score, we trained a supervised learning model to classify customers into `HighRisk` and non-`HighRisk` groups based solely on the original encoded

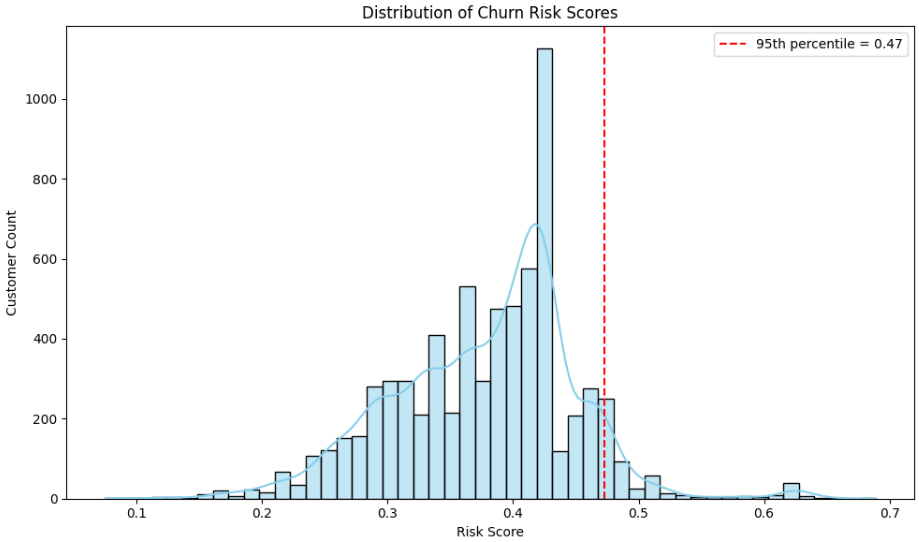

Figure 4: Distribution of computed churn risk scores. The red dashed line marks the 95th percentile cutoff. Customers to the right of this threshold were flagged as `HighRisk`, enabling a targeted intervention strategy based on semantic segmentation.

features. This allowed us to determine whether the nonlinear behavior captured by UMAP and entropy was also detectable using conventional feature vectors.

We used a Random Forest classifier for its robustness and interpretability. The model was trained on an 80/20 train-test split and evaluated using accuracy and feature importance metrics. Results indicated that the classifier achieved solid predictive performance, with accuracy comparable to direct churn prediction but using a **softer target** informed by geometry and entropy.

One key outcome was that `ShannonEntropy` ranked among the most important predictive features, reinforcing its value in representing behavioral variability. Other top features included `TotalCharges`, `MonthlyCharges`, and `Tenure`, indicating the financial and temporal engagement aspects of churn.

9 Conclusion and Methodology Evolution

This project began with a simple question: can we detect churn using traditional dimensionality reduction and clustering techniques? After PCA failed to uncover meaningful structure, we pivoted toward a nonlinear, information-theoretic approach. By computing per-customer entropy, projecting the data using UMAP, and clustering it with DBSCAN and HDBSCAN, we constructed a meaningful low-dimensional behavior space.

From this space, we developed a custom churn risk score that combined entropy, membership fuzziness, and observed churn rate within clusters. This risk score provided a powerful way to detect customers on the edge of behavioral boundaries—those most at risk of churning. We validated the score by training a model to predict it from the original features and showed that it could be used for explainable segmentation and targeted intervention.

All results and engineered features were exported for potential business integration. This included a CSV suitable for Power BI dashboards, enabling non-technical stakeholders to explore churn clusters, segment risk, and build proactive strategies.

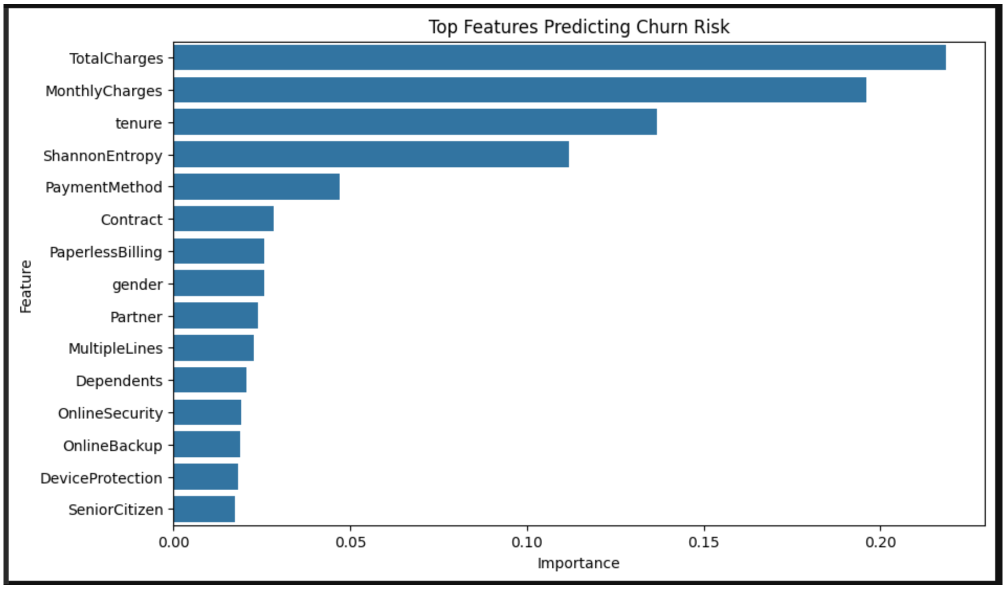

Figure 5: Feature importance from the Random Forest model trained to predict the engineered churn risk score. Traditional features like charges and tenure rank highly, but Shannon entropy also emerges as a significant predictor—supporting the hypothesis that behavioral disorder correlates with churn risk.

The key insight that emerged was this: when PCA fails, measure entropy. This approach is broadly applicable to any behavioral or engagement dataset where the signal may be hidden in nonlinear transitions or fuzzy boundary zones. Our method is modular, interpretable, and practical.

Future directions include adding temporal components (e.g., time-varying churn signatures), applying persistent homology or Mapper to extract topological features, and testing this framework across different domains where user behavior is complex, entropic, and emergent.

References

[1] Bobovski66, *Customer Churn ML Analysis*. GitHub Repository.
`https://github.com/bobovski66/customer_churn_ML`

[2] Blastchar, *Telco Customer Churn Dataset*. Kaggle.
`https://www.kaggle.com/datasets/blastchar/telco-customer-churn`