Telecom Customer Churn Dataset

Comprehensive Machine Learning Pipeline

Mostafa Latifian - 40122193
Parsa Alaviniko - 40120993

Fundamental to Intelligent Systems
Dr. Aliyari

K. N. Toosi University of Technology

Outline

1. Dataset Introduction
2. Data Preprocessing
3. Dimensionality Reduction
4. Model Training
    - SVM
    - Random Forest
    - XGBoost
    - LightGBM
    - Decision Tree
    - Conclusion (Model Training)
5. Hyperparameter Tuning
6. Final Conclusion

Dataset Introduction

Source & Objective

- **Source:** IBM Sample Data Sets
- **Goal:** Predict customer churn

Data Structure

- **Dimensions:** 7,043 rows & 21 features
- **Categories:**
  - Demographics: Gender, seniority...
  - Account: Tenure, contract, charges...
  - Services: Internet, security...

Initial Insights & Quality

- **Class Imbalance:**
  - Stayed: 73.5%
  - Churned: 26.5%
- **Evaluation Strategy:**
  - Focus on **Recall**
  - Avoid Accuracy bias

Data Preprocessing: Methodology

Data Cleaning

- **Handling Missing Values:** Fixed 11 "hidden" nulls in `TotalCharges` by replacing them with 0.
- **Noise Reduction:** Dropped `customerID` (non-predictive feature).

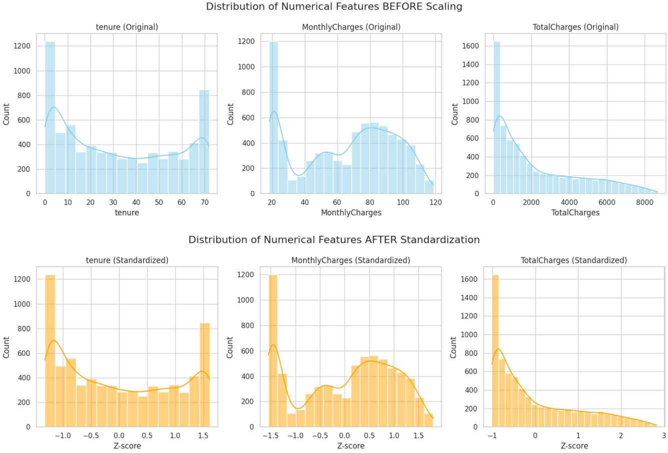
Feature Engineering & Encoding

- **Target Transformation:** Converted `Churn` to binary format (0/1).
- **Categorical Encoding:** Applied One-Hot Encoding.
- **Dimensionality:** Features expanded from 21 to 31.

Feature Scaling

- **Standardization:** Applied Z-Score Scaling to numerical columns (`tenure`, `charges`).
- **Purpose:** Ensured model convergence and prevented feature dominance.

Data Preprocessing: Scaling Visualization



All features including both `tenure` and `TotalCharges` are now centered around 0 (the mean). Most values fall within approximately -2 to $+2$ standard deviations.

Dimensionality Reduction: PCA vs. LDA

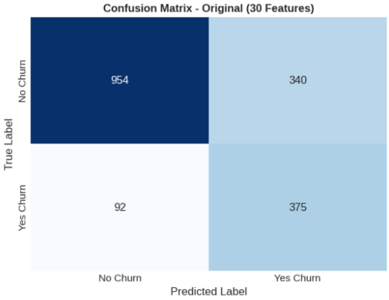
Objective: Reduce complexity and noise while maintaining accuracy and improving efficiency.

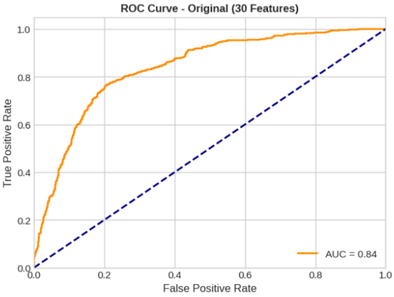
- **Method 1: PCA (Unsupervised)**
  - **Result:** Reduced features from 31 to 17 ($\approx$ 43% reduction).
  - **Outcome:** Accuracy stable at 75.24%; redundant noise removed.
- **Method 2: LDA (Supervised)**
  - **Result:** Compressed 31 features into 1 single dimension.
  - **Outcome:** Accuracy increased to 76.26% (Strong linear separability).

Final Decision: The Original dataset and PCA are prioritized for further modeling to ensure no critical non-linear information is lost, despite LDA's high compression.

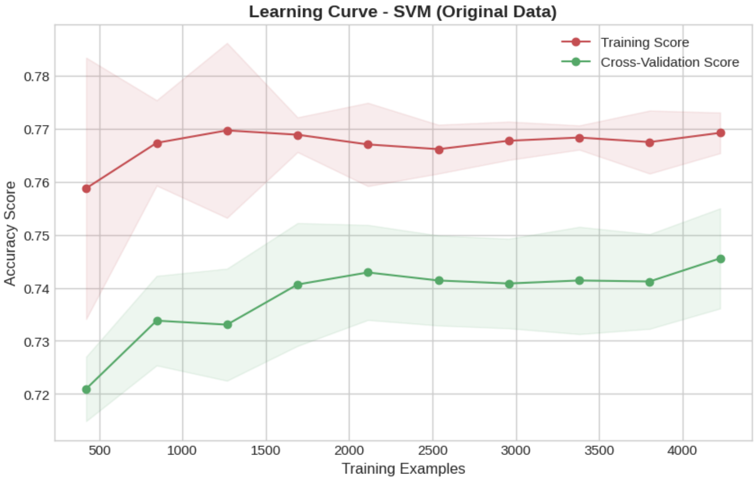
Model Training: SVM Performance

Dataset	Accuracy	Precision	Recall	F1-Score	Hinge Loss
Original	75.47%	52.45%	80.30%	63.45%	0.525
PCA	75.41%	52.39%	79.87%	63.27%	0.527



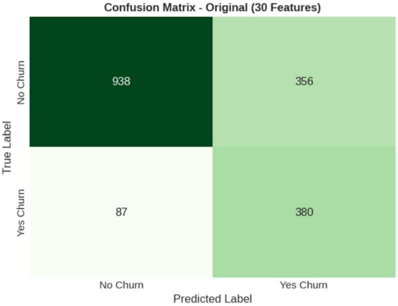


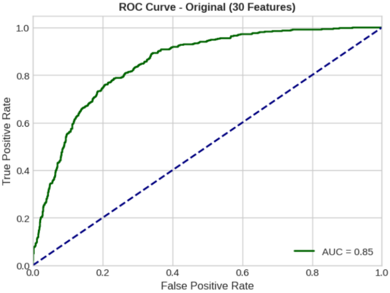
Model Training: SVM Learning Curve



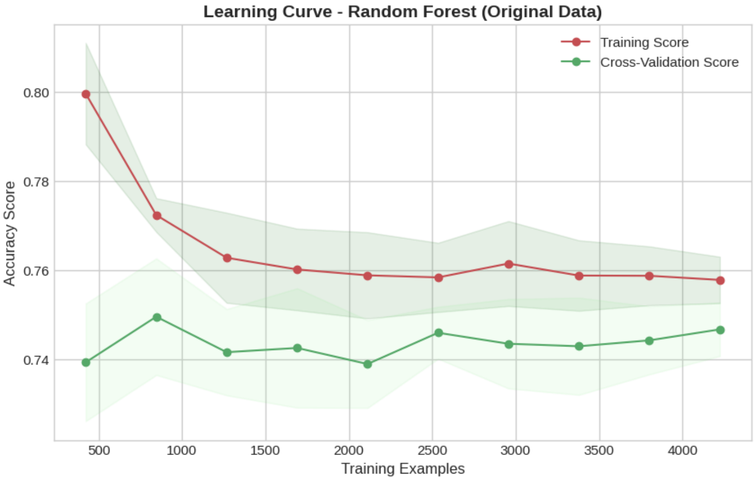
Model Training: Random Forest Performance

Dataset	Accuracy	Precision	Recall	F1-Score	Hinge Loss
Original	74.84%	51.63%	81.37%	63.17%	0.498
PCA	76.83%	54.30%	79.65%	64.58%	0.493



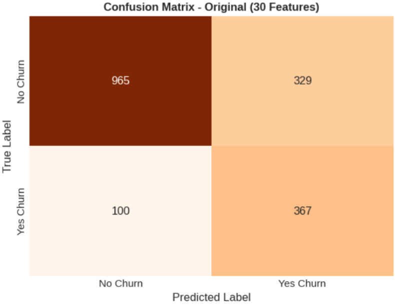


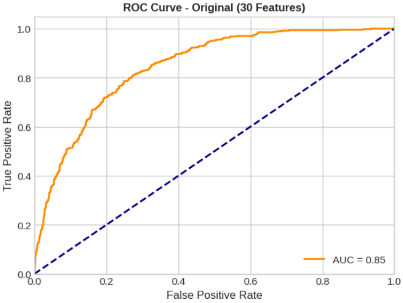
Model Training: Random Forest Learning Curve



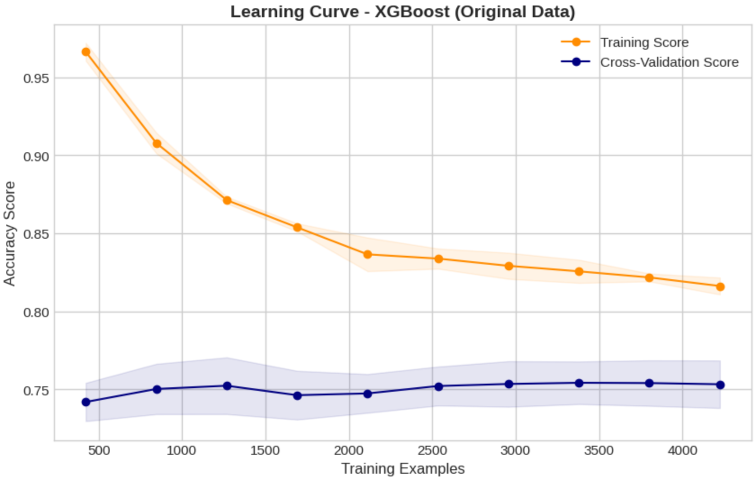
Model Training: XGBoost Performance

Dataset	Accuracy	Precision	Recall	F1-Score	Hinge Loss
Original	76.63%	52.71%	81.79%	64.04%	0.474
PCA	76.09%	53.28%	79.87%	63.92%	0.470



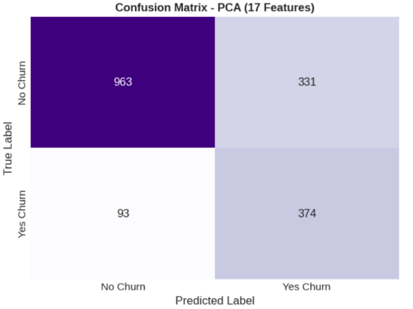


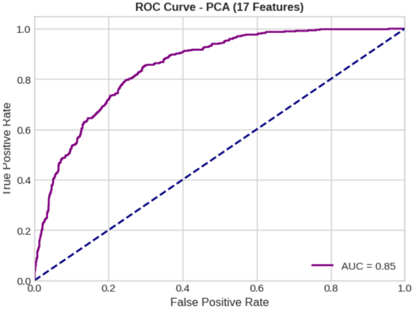
Model Training: XGBoost Learning Curve



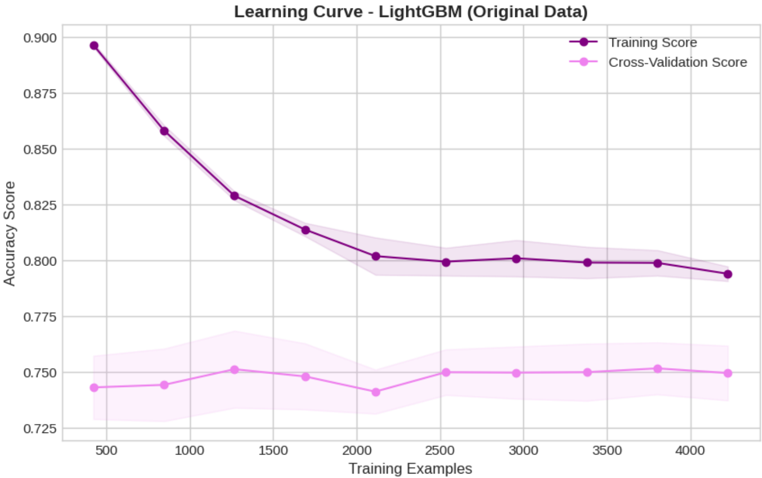
Model Training: LightGBM Performance

Dataset	Accuracy	Precision	Recall	F1-Score	Hinge Loss
Original	75.80%	52.82%	82.01%	64.26%	0.475
PCA	75.92%	53.04%	80.08%	63.82%	0.471



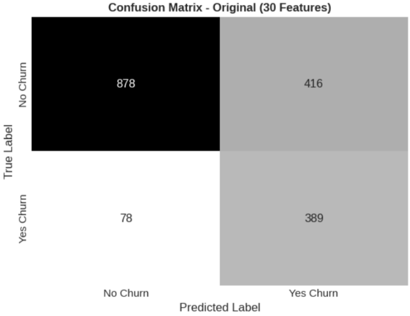


Model Training: LightGBM Learning Curve

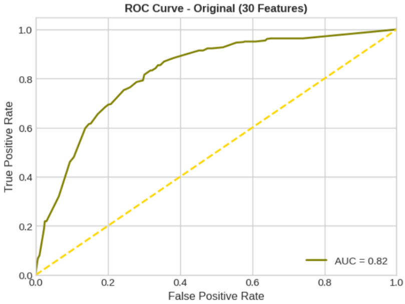


Model Training: Decision Tree

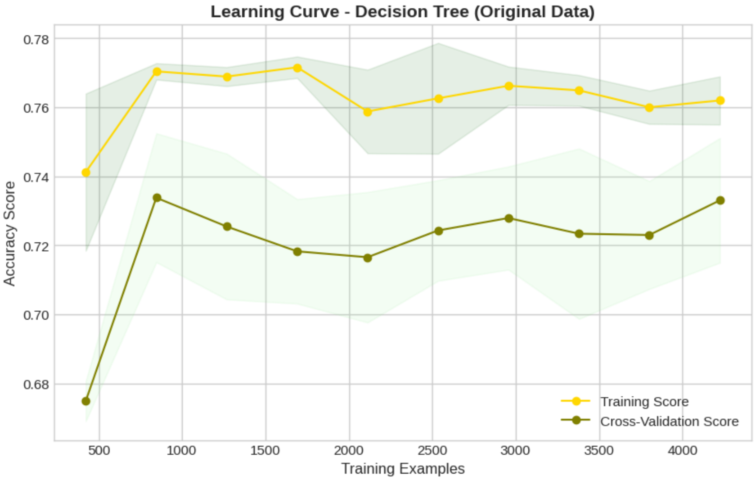
Dataset	Accuracy	Precision	Recall	F1-Score	Hinge Loss
Original	71.94%	48.32%	83.29%	61.16%	0.828
PCA	73.53%	50.06%	78.58%	61.16%	0.675


Confusion Matrix - Original (30 Features)

	No Churn	Yes Churn
No Churn	878	416
Yes Churn	78	389


ROC Curve - Original (30 Features)

Model Training: Decision Tree Learning Curve



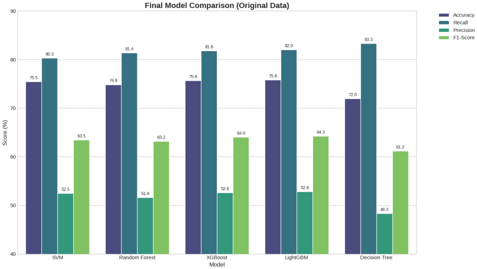
Model Evaluation & Comparison: Quantitative Results

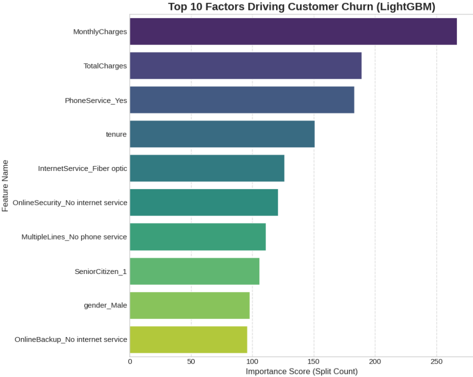
Model	Accuracy	Recall	Precision	F1-Score	Loss
LightGBM	75.81%	82.01%	52.83%	64.26%	0.4754
gray!10 **Decision Tree**	71.95%	83.30%	48.32%	61.16%	0.828
XGBoost	75.64%	81.80%	52.62%	64.04%	0.473
gray!10 **Random Forest**	74.84%	81.37%	51.63%	63.17%	0.498
SVM	75.47%	80.30%	52.45%	63.45%	0.525

tableFinal comparison of all classification models

- **Decision Tree:** Highest Recall (83.3%) but high "False Alarm" rate and instability.
- **LightGBM:** Best balance with high precision and probabilistic stability.

Model Training: Comparison and Top 10 Features





Hyperparameter Tuning: Optimization Results

Model	Accuracy	Recall	Precision	F1-Score	Loss
LightGBM (Grid Search)	74.46%	84.36%	51.30%	63.80%	0.493
LightGBM (Rand Search)	73.25%	86.08%	49.75%	63.05%	0.512
Decision Tree	71.89%	84.15%	48.28%	61.35%	0.560
SVM	74.90%	81.15%	51.70%	63.16%	0.425

- **Insight: LightGBM** using Random Search achieved the highest Recall (86.08%), which is crucial for identifying potential churners.
- **Efficiency:** Tuning significantly improved the models' ability to handle class imbalance compared to baseline versions.

Final Conclusion

Model Selection

After completing all the previous stages, the best-performing model was selected as LightGBM (Grid Search).

We successfully developed a model capable of identifying 84.4% of customers who are likely to leave the company before the churn event occurs.

Thank You!

Thank you for your time and attention.

Presenters:
Mostafa Latifian & Parsa Alaviniko

K. N. Toosi University of Technology