Predictive Analytics for Customer Retention in E-Commerce and Banking  
  
EMINE UYSAL (ID: 501304049)

Toronto Metropolitan University

CIND 820: Big Data Analytics Project

May 22th, 2025

**REVISED ABSTRACT**

**Customer churn poses a significant challenge in behavior-driven industries such as e-commerce and telecommunications, where long-term profitability relies heavily on customer retention.** This project leverages predictive analytics and explainable machine learning to not only forecast churn but also to understand the behavioral and service-related factors that influence it, with the goal of designing sector-specific retention strategies.

Two real-world datasets sourced from Kaggle are used for this analysis:

1. The **“Telco Customer Churn”** dataset, which includes demographics, service usage, and billing information (Kaggle, 2018);
2. The **“Customer Behavior in E-Commerce”** dataset, featuring transactional behavior such as order frequency, product category engagement, and spending patterns (Imakash, 2023).

These datasets provide a cross-sector view of churn dynamics, allowing for comparative analysis between industries. The project explores the following primary research questions:

* *Which behavioral and service-related features are most predictive of churn in each sector?*
* *Can interpretable machine learning models predict churn before it occurs with sufficient accuracy?*
* *How do churn drivers and mitigation strategies differ between the e-commerce and telecom sectors?*

The methodology includes rigorous data preprocessing—such as one-hot encoding, missing value imputation, and multicollinearity checks using Variance Inflation Factor (VIF). Class imbalance is addressed using **SMOTE** (Chawla et al., 2002) to enhance model performance on minority (churn) classes. Three classification models—**Logistic Regression**, **Random Forest**, and **XGBoost**—are trained and evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics.

To address concerns around model interpretability and overfitting, **SHAP** (Lundberg & Lee, 2017) and **LIME** (Ribeiro et al., 2016) are used to visualize feature importance, while **cross-validation and regularization techniques** are employed to ensure model generalizability. An interactive **Power BI dashboard** visualizes key findings, helping both technical and non-technical stakeholders explore high-risk customer segments and actionable insights.

While both datasets provide valuable behavioral and demographic data, they contain synthetic elements and have limited representativeness. Ethical considerations around real-world applicability and data bias are acknowledged, and future work will incorporate more diverse, high-fidelity datasets for model refinement.

**REFERENCES**

**• Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002).** SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://www.jair.org/index.php/jair/article/view/10302

**• Imakash. (2023).** *Customer Behavior in E-commerce*. Kaggle.  
<https://www.kaggle.com/datasets/uom190346a/e-commerce-customer-behavior-dataset>

• **Kaggle. (2018).** *Telco Customer Churn*.  
https://www.kaggle.com/datasets/blastchar/telco-customer-churn

• Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

**• Ribeiro, M. T., Singh, S., & Guestrin, C. (2016).** "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://dl.acm.org/doi/10.1145/2939672.2939778