**CSCA 5622 Introduction to Machine Learning: Supervised Learning**

**Bank Customer Churn Prediction Using Supervised Learning**

**Slide 1: Title Slide (30 seconds)**

Hello everyone, welcome to my presentation on Customer Segmentation Using Unsupervised Learning. My

**Slide 1: Title Slide (30 seconds)**

Hi, I’m Molly B, and today I’m presenting my final project for CSCA 5622: Bank Customer Churn Prediction using supervised learning. This project tackles a critical issue for banks, and I’ll walk you through the problem, my approach, and the results. Let’s get started.

**Slide 2: The Problem: Customer Churn in Banking (1 minute 15 seconds)**

Customer churn is a major challenge for banks, costing them millions every year. My project aims to predict whether a bank customer will churn—close their account—which is a binary classification problem with the target variable Exited, where 0 means no churn and 1 means churn. Retaining high-value clients can save banks between $10 million and $500 million annually by 2025, according to industry estimates. I used the Kaggle Bank Customer Churn Prediction dataset, which has 10002 samples and 10 features—8 numeric and 2 categorical. However, I faced challenges like class imbalance, with only about 20% of customers churning, and minor missing values, such as 1 in the Geography feature.

**Slide 3: Approach: Data Preprocessing (1 minute 15 seconds)**

To prepare the data for modeling, I followed several preprocessing steps. First, I handled missing values by imputing Geography with the label ‘missing’ and numeric features like Age with their medians. To address the class imbalance, I applied SMOTE to balance the training data to a 50/50 split between churn and no-churn. I then encoded categorical features like Geography and Gender using one-hot encoding and scaled numeric features with StandardScaler to ensure model compatibility, especially for algorithms like Neural Networks. The outcome was a dataset ready for modeling, with balanced classes and normalized features. This plot shows how churn varies by Geography, with Germany having a higher churn rate, which influenced my preprocessing decisions.

**Slide 4: Data Numeric Distributions (1 minute 15 seconds)**

Before modeling, I explored the distributions of numeric features to guide preprocessing. As you can see in these histograms, Age is skewed right, with a peak around 30 to 40, suggesting a potential log-transformation to normalize it for better modeling performance. Balance is bimodal, with a large peak at 0, indicating many customers have no balance, and another peak around 120000. These insights helped me decide to scale all numeric features, ensuring they’re on a comparable range for the models to perform effectively.

**Slide 5: Correlations (1 minute 15 seconds)**

I also analyzed correlations to understand feature relationships. There’s a weak positive correlation of 0.03 between Age and Exited, but the box plot here shows Age’s predictive power more clearly—customers who churn have a median age of around 45, compared to 35 for those who don’t. The correlation matrix reveals no strong multicollinearity among numeric features, with the highest correlation being -0.30 between Balance and NumOfProducts. This means all features can be retained for modeling without concerns of redundancy, ensuring a robust feature set for the next steps.

**Slide 6: Approach: Model Development (1 minute 30 seconds)**

For modeling, I used four supervised learning models on an 80/20 train-test split. Logistic Regression served as a baseline linear model. Random Forest, a non-linear ensemble, was tuned using GridSearchCV for optimal hyperparameters. I implemented a Neural Network with deep learning, using GPU acceleration, structured with 64/32/1 layers and dropout for regularization. Finally, CatBoost, a gradient boosting model also using GPU, was chosen for its strength with categorical data and SHAP for interpretability. My techniques included SMOTE to handle imbalance, GPU acceleration to speed up training for Neural Network and CatBoost, and dropout to reduce overfitting in the Neural Network. This Random Forest feature importance plot highlights Age, NumOfProducts, and Balance as key predictors.

**Slide 7: Comparison: Model (1 minute 30 seconds)**

Let’s compare the models’ performance. Logistic Regression, the baseline, achieved a ROC-AUC of 0.77 and a Class 1 F1-score of 0.29, with high precision at 0.83 but low recall at 0.19, and an accuracy of 0.81. Random Forest improved significantly, with a ROC-AUC of 0.86, F1-score of 0.59, precision of 0.83, recall of 0.45, and accuracy of 0.87. The Neural Network also scored a ROC-AUC of 0.86 and an F1-score of 0.60, with precision at 0.50, a high recall of 0.74, and accuracy of 0.80. Finally, CatBoost performed the best, with a ROC-AUC of 0.88, F1-score of 0.62, precision of 0.54, recall of 0.72, and accuracy of 0.81, showing a strong balance across metrics.

**Slide 8: Results: Model Performance (1 minute 15 seconds)**

Here’s a summary of the results. CatBoost, the best model, achieved a ROC-AUC of 0.87 and a Class 1 F1-score of 0.62, showing strong performance in identifying churners. Random Forest and Neural Network both scored a ROC-AUC of 0.86, with F1-scores of 0.59 and 0.60, respectively. Logistic Regression lagged with a ROC-AUC of 0.77 and F1-score of 0.29. This ROC curve confirms CatBoost’s superior discriminative ability, while the bar chart highlights its lead in both ROC-AUC and F1-score, balancing accuracy and interpretability effectively.

**Slide 9: Insights: Understanding Churn Drivers (1 minute)**

:Using SHAP with CatBoost, I identified the key drivers of churn: Age, NumOfProducts, Tenure, and Geography, particularly Germany, as shown in this plot. Red indicates higher values increasing churn likelihood, while blue decreases it. For example, older age and being in Germany push the model output towards churn, while having more products reduces the risk. This leads to a clear business insight: older customers in Germany with fewer products are at higher risk, allowing banks to target these customers for retention efforts.

**Slide 10: Conclusion and Future Work (1 minute)**

In conclusion, CatBoost excelled with a ROC-AUC of 0.87, as shown in this bar chart, providing actionable insights through SHAP, such as targeting older customers in Germany with fewer products. This can enable banks to save between $10 million and $500 million annually through targeted retention strategies. For future work, I plan to incorporate transaction data using LSTM models for dynamic predictions and deploy CatBoost via a Flask API for real-time churn prediction.

Thank you for listening, and I hope you found this presentation insightful.