CUSTOMER CHURN PREDICTION

1. Introduction

Customer churn prediction is critical for businesses to retain customers and increase profitability. This project aims to build a reliable machine learning model to predict which customers are likely to churn based on their profile and behavior data. The analysis focuses on feature engineering, model training, tuning, evaluation, and deployment readiness.

2. Data Overview

The dataset includes demographic, financial, and behavioral attributes of customers such as Credit Score, Geography, Age, Balance, Number of Products, Satisfaction Score, and Exited (target variable indicating churn). Initial exploratory data analysis and visualization helped identify key trends and feature importance.

3. Feature Engineering

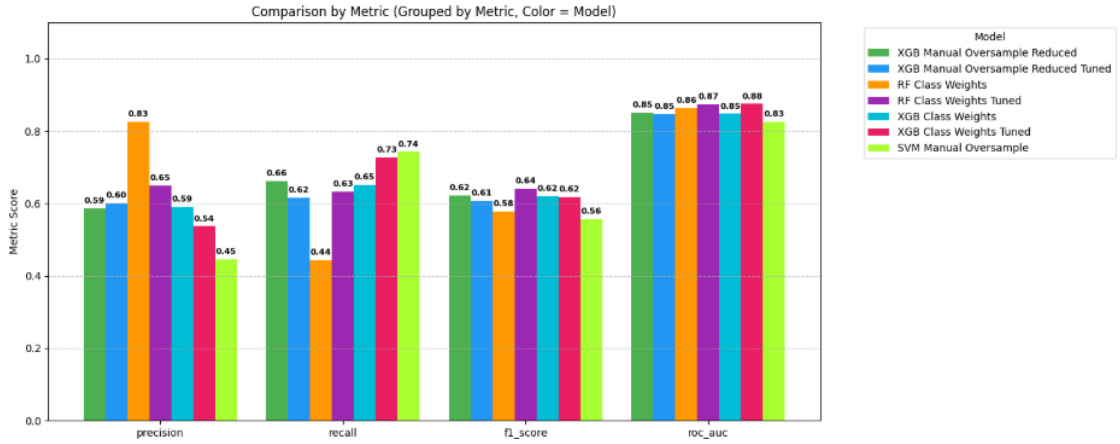
Various features were tested including ratios like balance-to-salary and tenure groups. However, the original features combined with proper preprocessing proved sufficient, as new engineered features did not significantly improve model performance.

4. Data Preprocessing

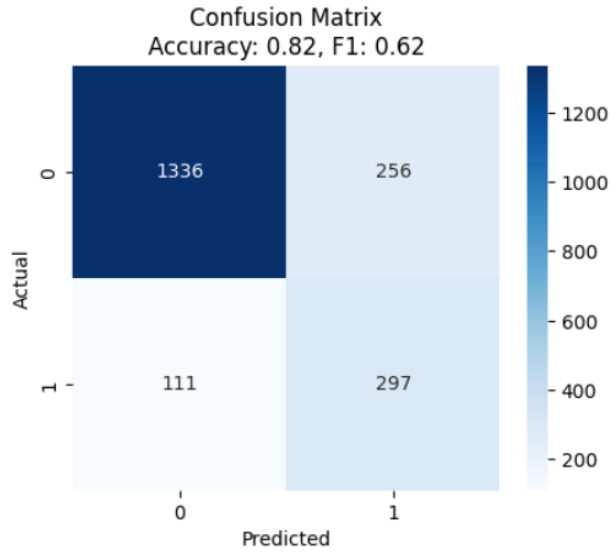
Categorical variables were encoded using One-Hot Encoding with dummy variable trap prevention. Continuous variables were scaled using StandardScaler. The dataset was split into training and test sets, with oversampling of minority classes applied manually to address class imbalance.

5. Model Development and Evaluation

Multiple models were tested including Logistic Regression, Random Forest, XGBoost, and SVM, each trained with different sampling strategies and hyperparameter tuning. The best performing model was XGBoost with class weights and tuned hyperparameters, balancing precision, recall, F1-score, and ROC-AUC.



Model	Accuracy	Precision	Recall	F1-score	ROC AUC
XGB Class Weights Tuned	0.8165	0.5371	0.7279	0.6181	0.8751



6. Model Interpretation

Feature importance analysis revealed the most predictive features such as Geography (Germany), Credit Score, Age, and Satisfaction Score. These insights can guide business strategies to focus on key risk factors affecting churn.

7. Final Model Pipeline

A clean and robust ML pipeline was created that includes preprocessing, feature selection, and the final XGBoost model. This pipeline ensures reproducibility and ease of deployment.

8. Model Validation

On a sample of unseen data, the model demonstrated good predictive performance with a balanced confusion matrix and acceptable classification metrics. Visualizations comparing actual vs predicted churn confirmed model reliability.

9. Next Steps and Deployment

The model is ready for deployment as a serialized pipeline object (pickle file). The deployment will involve setting up an API endpoint for real-time predictions or batch processing. Monitoring the model's performance post-deployment will be essential to catch data drift and update the model accordingly.