Mini Project 0: Advanced Telco Customer Churn Prediction

🎯 Project Objective

Apply advanced concepts from **Week 2 (EDA)**, **Week 3 & Week 4 (Next Week) (Model Building, Evaluation & Class Imbalance)** to build a comprehensive churn prediction system for Telco customers. This project emphasizes real-world machine learning challenges including class imbalance, ensemble methods, and business-focused evaluation metrics.

📊 Dataset Information

Source: [Telco Customer Churn Dataset (Kaggle)] **File**: `WA_Fn-UseC_-Telco-Customer-Churn.csv` **Size**: 7,043 customers, 21 features **Target Variable**: `Churn` (Yes/No)

Key Features:

- **Customer Demographics**: Gender, SeniorCitizen, Partner, Dependents
- **Account Information**: Tenure, Contract, PaperlessBilling, PaymentMethod
- **Services**: PhoneService, MultipleLines, InternetService, OnlineSecurity, etc.
- **Financial**: MonthlyCharges, TotalCharges

📋 Project Requirements

Deadline: 10th August 2025

Deliverables:

1. **Jupyter Notebook** with comprehensive analysis and modeling
2. **Executive Summary Report** (2-3 pages) with business insights and recommendations
3. **Model Performance Comparison** with proper evaluation metrics
4. **Business Impact Analysis** with actionable recommendations

🔍 Part 1: Advanced Exploratory Data Analysis (EDA)

1.1 Initial Data Assessment

- **Data Quality Check**: Examine data types, missing values, and inconsistencies
- **Target Variable Analysis**: Calculate churn rate and discuss class imbalance implications
- **Feature Overview**: Categorize features into demographic, behavioral, and financial groups

1.2 Class Imbalance Analysis

- **Visualize class distribution** with appropriate charts
- **Calculate imbalance ratio** and discuss impact on model evaluation
- **Analyze churn patterns** across different customer segments
- **Business Context**: Explain why class imbalance matters in churn prediction

1.3 Advanced Univariate Analysis

- **Numerical Features**: Distribution analysis, outlier detection using IQR and Z-score methods
- **Categorical Features**: Frequency analysis and relationship with churn
- **Feature Engineering Opportunities**: Identify potential derived features

1.4 Comprehensive Bivariate Analysis

- **Churn vs Demographics**: Age groups, gender, family status impact
- **Churn vs Services**: Service adoption patterns and churn correlation
- **Churn vs Financial**: Monthly charges, total charges, and payment behavior
- **Statistical Significance**: Use appropriate tests (Chi-square, t-tests) to validate relationships

1.5 Multivariate Analysis

- **Correlation Matrix**: Identify multicollinearity issues
- **Feature Interactions**: Explore combinations that influence churn (e.g., Contract + PaymentMethod)
- **Customer Segmentation**: Group customers by behavior patterns

1.6 Business Insights Generation

- **High-Risk Customer Profiles**: Identify characteristics of customers most likely to churn
- **Retention Opportunities**: Services or contract types that reduce churn

- **Revenue Impact**: Calculate potential revenue loss from churning customers

⚙️ Part 2: Advanced Model Pipeline & Ensemble Methods

2.1 Data Preprocessing Pipeline

- **Data Cleaning**: Handle inconsistencies (e.g., TotalCharges data type issues)
- **Feature Engineering**: Create meaningful derived features
    - Tenure categories (New, Established, Loyal)
    - Service adoption score
    - Average monthly charges per service
    - Payment reliability indicators
- **Encoding Strategies**: Compare different encoding methods for categorical variables
- **Feature Scaling**: Apply appropriate scaling for numerical features

2.2 Ensemble Model Implementation

Implement and compare the following ensemble methods:

2.2.1 Bagging Method: Random Forest

- **Implementation**: Use scikit-learn RandomForestClassifier
- **Hyperparameters to tune**: n_estimators, max_depth, min_samples_split, max_features
- **Analysis**: Feature importance interpretation and business insights

2.2.2 Boosting Method: XGBoost

- **Implementation**: Use XGBoost library
- **Hyperparameters to tune**: learning_rate, max_depth, n_estimators, subsample
- **Analysis**: Feature importance and model interpretation

2.2.3 Advanced Boosting: CatBoost

- **Implementation**: Use CatBoost library for native categorical handling
- **Advantages**: Automatic categorical encoding, reduced overfitting
- **Analysis**: Compare performance with other methods

2.2.4 Baseline Comparison

- **Logistic Regression**: Simple baseline model
- **Decision Tree**: Single tree for interpretability comparison

2.3 Pipeline Construction

- **Scikit-learn Pipelines**: Create modular, reproducible preprocessing and modeling pipelines

- **Cross-Validation Strategy**: Use stratified k-fold to maintain class distribution
- **Hyperparameter Tuning**: Implement GridSearchCV or RandomizedSearchCV

📈 Part 3: Model Evaluation for Imbalanced Data

3.1 Class Imbalance Considerations

- **Why Accuracy Fails**: Demonstrate with concrete examples why accuracy is misleading
- **Business Impact**: Explain cost of false positives vs. false negatives in churn prediction

3.2 Comprehensive Evaluation Metrics

Evaluate all models using the following metrics with detailed interpretation:

- **Precision**: Quality of churn predictions (campaign efficiency)
- **Recall**: Coverage of actual churners (revenue protection)
- **F1-Score**: Balanced performance measure

3.3 Model Comparison Framework

- **Performance Matrix**: Compare all models across all metrics
- **Statistical Significance**: Use appropriate tests to validate performance differences
- **Business Value Analysis**: Translate metrics into business impact (revenue saved, campaign efficiency)

🎯 Part 4: Business Impact Analysis

4.1 Customer Segmentation for Retention

- **High-Risk Segment**: Customers with high churn probability
- **Medium-Risk Segment**: Customers requiring proactive engagement
- **Low-Risk Segment**: Loyal customers for upselling opportunities

4.2 Retention Strategy Recommendations

- **Targeted Interventions**: Specific actions for each risk segment
- **Resource Allocation**: Budget optimization for retention campaigns
- **Expected ROI**: Calculate return on investment for retention efforts

📝 Evaluation Criteria

Technical Excellence (40%)

- Proper implementation of ensemble methods
- Correct evaluation metrics for imbalanced data
- Quality of data preprocessing and feature engineering
- Code organization and documentation

Business Insight (30%)

- Quality of EDA insights and business interpretation
- Actionable recommendations for retention strategies
- Understanding of business impact and ROI
- Clear communication of technical concepts

Methodology (30%)

- Appropriate handling of class imbalance
- Proper cross-validation and hyperparameter tuning
- Statistical rigor in analysis and comparison
- Reproducibility of results

🎓 Learning Outcomes

Upon completion, students will demonstrate:

1. **Advanced EDA Skills**: Ability to extract meaningful business insights from data
2. **Ensemble Method Mastery**: Understanding of bagging, boosting, and their applications
3. **Imbalanced Data Expertise**: Proper evaluation and handling of class imbalance
4. **Business Acumen**: Translation of technical results into business value
5. **Production Readiness**: Consideration of real-world deployment challenges

🔗 Additional Resources

- **Course Materials**: Week 2 & 3 lecture notes and examples
- **Kaggle Dataset**: Telco Customer Churn
- **Ensemble Methods Guide**: Course visual guide on ensemble methods
- **Evaluation Metrics Guide**: Class imbalance evaluation best practices