Customer-Consistency-Prediction-using-Machine-Learning

This project focused on **predicting customer consistency**, a critical task for identifying and retaining most valuable users.

The **problem statement**, is to build a machine learning model that can accurately predict a customer's **Consistency Tag**. This tag categorizes users into groups like 'High-Value Consistent', 'Medium-Value Degrowth', and 'Low-Value Consistent'. By predicting this status, we can proactively intervene with retention strategies tailored to each group.

1. Data Overview & Preparation

A dataset of **100 customer records**, which is relatively small, but rich in detail. The initial data had about **12 columns**, primarily numerical features covering:

- **Revenue Metrics:** tot_rev, tot_voice_rev, total_data_rev.
- **Usage Metrics:** total_data_volume, total_calls_m_new.
- **Categorical Tags:** voip_usage_tag and target, Consistency_Tag.

2. Data Cleaning Steps

1. **Missing Data:** Imputed missing values in numerical columns using the **median**. Chose the median over the mean because usage and revenue data are often skewed, and the median is much more resistant to those extreme values.
2. **Handling Outliers:** Identified significant outliers in features like total revenue and call volume. Applied the **Interquartile Range (IQR) capping method**. Did not drop these data points because, with only 100 rows, every piece of data is precious. Capping allows us to keep the information while neutralizing the outlier's damaging effect on models.
3. **Encoding:** Used label encoding for consistency_tag and one – hot encoding for voip_usage_tag

3. Exploratory Data Analysis (EDA)

1. **Target Imbalance:** Observed that the target variable was imbalanced, with 'MV_Degrowth_New' being the largest group. This means must pay close attention to the **F1-Score**, not just raw accuracy, to ensure model performs well across all classes.
2. **Revenue is King:** Bivariate analysis showed a strong separation:
    - **High-Value Consistent** customers consistently show **significantly higher total revenue** (tot_rev) and **call volume** (total_calls_m_new) compared to all other groups. This feature is clearly primary predictor.
3. **Multicollinearity:** Found high correlation between revenue-related features (e.g., tot_rev and total_data_rev). This indicated that it needed a strategic approach to feature selection.

4. Feature Engineering & Selection

To maximize predictive power, performed two key feature enhancement steps:

Feature Engineering

Created two new ratio features to capture spending habits:

1. **Revenue per Call:** tot_rev / (total_calls_m_new + voip_tot_calls). This is a direct measure of customer spending efficiency.
2. **Data to Voice Revenue Ratio:** total_data_rev / tot_voice_rev. This tells us if a user is more data-centric or voice-centric.

Feature Selection

Used a combined approach for robust feature selection:

- **Random Forest Feature Importance** (an *embedded* method)
- **SelectKBest** using the ANOVA F-value (a *filter* method)

Final set of **Top 7 Features**—which included **tot_rev**, **total_calls_m_new**, and the engineered **rev_per_call**—were selected as they offered the highest predictive signal while mitigating multicollinearity.

5. Modeling & Hyperparameter Tuning

split data (70% Train, 30% Test) and used **Standard Scaling** to prepare the numerical features, which is essential for models like Logistic Regression and Deep Learning model.

Models Used

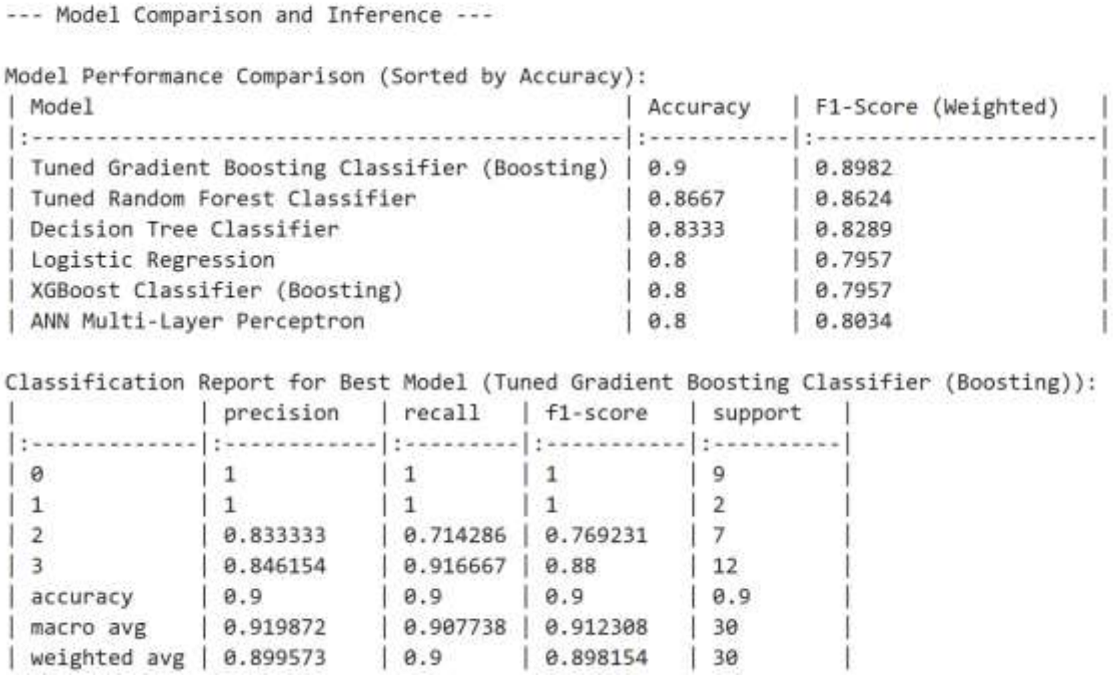
Tested a diverse set of models:

1. **Baseline Models:** Logistic Regression, Decision Tree Classifier.
2. **Ensemble Models (Tuned):**
   - **Random Forest Classifier**
   - **Gradient Boosting Classifier** (A powerful **Boosting Model**)
3. **Advanced Models:**
   - **XGBoost Classifier** (Another state-of-the-art **Boosting Model**)
   - **ANN Multi-Layer Perceptron (MLP)** (A **Deep Learning Model** with three hidden layers: 100, 50, and 25 neurons)

Hyperparameter Tuning

Used **GridSearchCV** on top ensemble models (Random Forest and Gradient Boosting) to systematically find the optimal combination of parameters (like n_estimators and max_depth), ensuring squeeze the best performance from each.

6. Results and Conclusion



```
--- Model Comparison and Inference ---

Model Performance Comparison (Sorted by Accuracy):
| Model                                       | Accuracy   | F1-Score (Weighted)   |
|:--------------------------------------------|:-----------|:----------------------|
| Tuned Gradient Boosting Classifier (Boosting) | 0.9        | 0.8982                |
| Tuned Random Forest Classifier              | 0.8667     | 0.8624                |
| Decision Tree Classifier                    | 0.8333     | 0.8289                |
| Logistic Regression                         | 0.8        | 0.7957                |
| XGBoost Classifier (Boosting)               | 0.8        | 0.7957                |
| ANN Multi-Layer Perceptron                  | 0.8        | 0.8034                |

Classification Report for Best Model (Tuned Gradient Boosting Classifier (Boosting)):
|              | precision   | recall   | f1-score   | support   |
|:-------------|:------------|:---------|:-----------|:----------|
| 0            | 1           | 1        | 1          | 9         |
| 1            | 1           | 1        | 1          | 2         |
| 2            | 0.833333    | 0.714286 | 0.769231   | 7         |
| 3            | 0.846154    | 0.916667 | 0.88       | 12        |
| accuracy     | 0.9         | 0.9      | 0.9        | 0.9       |
| macro avg    | 0.919872    | 0.907738 | 0.912308   | 30        |
| weighted avg | 0.899573    | 0.9      | 0.898154   | 30        |
```

Conclusion:

The most successful model in analysis is the Tuned Gradient Boosting Classifier (Boosting), achieving a robust accuracy of 0.90 (90%).

1. **Boosting Superiority:** The strong performance of the Gradient Boosting model confirms that the relationship between customer features and their 'Consistency_Tag' is complex and non-linear. These advanced ensemble models are better suited than simpler models (like Logistic Regression) to capture the subtle interactions driving customer consistency.
2. **Top Features Validate Strategy:** The model's success is heavily reliant on features derived from customer revenue and usage volume (tot_rev, total_calls_m_new, rev_per_call). This scientifically validates that total spending and spending efficiency are the most impactful characteristics for predicting high-value consistency.

Business Insight

Based on the **Tuned Gradient Boosting Classifier model**:

- **Focus Retention:** Strategy should center on customers exhibiting high revenue or increasing data/voice usage, as these are the top indicators of the desired HV_Consistent_New status.
- **Proactive Intervention:** Implement the Gradient Boosting model to proactively identify customers predicted to fall into the 'Degrowth' categories so that targeted retention efforts can be deployed before revenue loss occurs.