Customer Behavior Prediction Model Using Random Forest Classifier

1. Introduction

This project aims to develop a predictive model to forecast customer behavior based on their demographic and transactional data. By analyzing patterns in the data, we aim to predict whether a customer will engage with a product or service based on various features, such as income, age, and past transaction history.

2. Methodology

2.1 Data Preprocessing and Segmentation

To make accurate predictions, the data was preprocessed and segmented in the following steps:

1. **Data Cleaning:**
   - Missing values were imputed using suitable techniques (mean for continuous variables and mode for categorical variables).
   - Categorical variables were encoded into numerical formats using label encoding and one-hot encoding where necessary.
2. **Feature Engineering:**
   - Features like `Annual_Income`, `Age Group`, and `Spending Habits` were transformed for better model accuracy.
   - Created new features to improve predictive power, such as `Income Groups` and `Age Segments`.
3. **Segmentation:**
   - The dataset was divided into different segments using K-Means clustering to understand customer groups based on income and spending patterns. This segmentation allowed for tailored marketing strategies to be developed.

2.2 Model Selection and Training

1. **Model Choice:**
   - Several models were tested, including Random Forest Classifier, Logistic Regression, and Support Vector Machines.
   - The Random Forest Classifier was chosen for its ability to handle high-dimensional data and its robustness against overfitting.
2. **Hyperparameter Tuning:**
   - Hyperparameters such as `n_estimators` and `max_depth` were tuned using grid search and cross-validation to enhance model performance.

    - Random Forest was fine-tuned to achieve the best accuracy and minimize overfitting.
3. **Cross-Validation:**
    - Cross-validation was performed to ensure that the model generalizes well to unseen data and to evaluate its performance across different subsets of the dataset.

---

3. Performance Metrics

3.1 Evaluation Metrics

The model was evaluated using various metrics to assess its performance:

- **Accuracy:** The overall proportion of correct predictions.
- **Precision:** The proportion of true positive predictions among all positive predictions made.
- **Recall:** The proportion of actual positives correctly identified.
- **F1-Score:** The harmonic mean of precision and recall, used when classes are imbalanced.
- **Confusion Matrix:** Helps in understanding the distribution of predicted vs. actual labels.

The Random Forest model achieved the following results:

- **Accuracy**: 87%
- **Precision (Class 0)**: 0.89
- **Precision (Class 1)**: 0.66
- **Recall (Class 0)**: 0.96
- **Recall (Class 1)**: 0.37
- **F1-Score (Class 0)**: 0.93
- **F1-Score (Class 1)**: 0.47

The **ROC Curve** and **AUC (Area Under Curve)** were also computed to evaluate the model's performance across different classification thresholds.

3.2 Visualizations

- **Confusion Matrix Visualization:** A heatmap was generated to visualize the confusion matrix, allowing us to assess model performance in terms of both false positives and false negatives.
- **Feature Importance:** The importance of each feature was plotted to show which factors most significantly influenced the prediction.

---

4. Insights and Business Recommendations

4.1 Insights:

- **Customer Behavior Patterns:**
  - Customers with higher incomes tend to engage less frequently but spend more per transaction, while customers with lower incomes exhibit higher transaction frequency but lower average spend.
  - Age groups showed distinct buying behaviors, with younger consumers spending less frequently on higher-end products.

4.2 Business Recommendations:

1. **Targeted Marketing:**
   - Implement tailored marketing strategies for different customer segments. For high-income customers, use loyalty programs and exclusive offers, whereas for younger, lower-income segments, offer frequent small incentives.
2. **Improving Customer Retention:**
   - Use the predictions to identify at-risk customers who may need additional attention to boost engagement and reduce churn.
3. **Upselling and Cross-Selling:**
   - Focus on upselling to higher-income customers by recommending complementary or premium products based on their spending history.

---

5. Conclusion

The project successfully implemented a Random Forest Classifier to predict customer behavior based on demographic and transactional data. The model showed a good balance between precision and recall, particularly for the majority class, with potential improvements in targeting the minority class. The insights and recommendations from this project can help businesses optimize marketing strategies and improve customer engagement and retention.
