Machine Learning Model for Customer Churn Prediction

1. Algorithm Selection

Data and Problem Overview

The objective is to predict customer churn by analyzing demographic and transaction data. After excluding the target-leaking feature ("RecencyDays"), the model relies on behavioral metrics (spending, frequency, product diversity) and customer attributes. The prediction task is binary classification, with a focus on understanding drivers of churn for practical business intervention.

Algorithms Considered

- **Random Forest Classifier:**
  Chosen for its robustness, ability to capture nonlinear interactions between variables, and moderate interpretability via feature importance metrics.
- **Alternative options (not selected):**
  - Logistic Regression: High interpretability but limited for complex behavioral data.
  - Gradient Boosting (e.g., XGBoost): Generally higher accuracy, but more complex and less interpretable for business reporting.

Rationale for Random Forest:

- Handles a mix of numerical and categorical variables well.
- Limits overfitting compared to standalone decision trees.
- Identifies the most influential features for business understanding.
- Performs well on tabular, mixed-type data.

2. Model Building and Training

Data Preparation

- Used the combined and cleaned dataset including engineered features:
  - TotalSpent, AvgSpent, TransactionCount, DistinctCategories (product variety), Age, demographic one-hot encodings.
- Target variable: **Churn** (binary: 1 = churned, 0 = retained)
- Test/train split: 80% train, 20% test, stratified to maintain class balance.

Training and Validation

- Applied GridSearchCV to tune n_estimators (100/200) and max_depth (None, 10, 20), optimizing for F1 score.
- Best model selected via 5-fold cross-validation on the training set.
- Evaluated on hold-out test set (metrics below).

3. Model Evaluation

Results (Hold-Out Test Set)

Metric	Value
Precision	0.56
Recall	0.37
F1 Score	0.44
ROC-AUC	0.69

Confusion Matrix:

- **Precision (56%)**: Just over half of customers predicted as churners actually churned.
- **Recall (37%)**: The model identified just over one-third of all actual churners.
- **F1 Score (44%)**: Reflects trade-off between catching churners and minimizing false alarms.
- **ROC-AUC (0.69)**: The model achieves moderate discrimination above random chance.

Feature Importance (Top Features)

Feature	Importance (%)
TotalSpent	27.5
AvgSpent	19.8
TransactionCount	15.2

Feature	Importance (%)
Age	14.2
DistinctCategories	9.6
Demographic Factors*	<3 each

*Demographic factors include one-hot encoded fields for gender, marital status, and income.

Interpretation

- **Spending habits and engagement** are key drivers; customers with higher, more consistent spending and variety are less likely to churn.
- **Age** also materially influences churn risk.
- **Demographics** (income level, marital status, gender) play a secondary role.

4. Business Utilization & Recommendations

Utilizing Model Predictions

- Assign churn risk scores to all active customers in the database.
- **Targeted retention:** Focus outreach on segments with the highest predicted risk, using triggered offers and personalized messaging.
- **Campaign optimization:** Use patterns in spending and engagement (as identified by feature importance) to design prevention programs.
- **Stakeholder insight:** Feature importances and confusion matrix offer transparency for business and compliance teams.

Model Limitations & Improvement Opportunities

- **Current model recall (37%)** may be insufficient for aggressive churn intervention. If avoiding missed churners is more valuable than minimizing false positives, threshold tuning or recall-focused optimization is warranted.
- **Feature engineering:** Add trend-based metrics (e.g., recent spending decline, time since last engagement) and incorporate new data sources (e.g., customer service interactions) for improved accuracy.
- **Class imbalance:** Use resampling (SMOTE/undersampling) or class-weighting to further boost recall if business needs dictate.

5. Conclusion

The Random Forest model provides an actionable baseline for predicting customer churn at Lloyds Banking Group, identifying at-risk segments based on spending behaviours and customer attributes. Ongoing model refinement including the introduction of behavioural trends, fine-tuning cut-offs for precision vs. recall, and ingesting broader interaction data will further enhance business value and retention strategy effectiveness.