1. Objective
The goal of this task is to predict whether a customer will churn (unsubscribe) or remain subscribed to a service using the provided dataset. Churn prediction is critical for businesses to understand customer behavior, enhance retention strategies, and maximize profitability. This project involves building a machine learning classification model using advanced techniques to achieve high accuracy and predictive power.

2. Exploratory Data Analysis (EDA)
Before proceeding with model building, it is essential to understand the dataset thoroughly. EDA involves inspecting the dataset structure, identifying missing values, and exploring feature distributions. For this dataset:

Columns like RowNumber, CustomerId, and Surname were identified as irrelevant and removed.
The Exited column, the target variable, was analyzed for class distribution. The dataset exhibited an imbalance between churned (Exited = 1) and retained (Exited = 0) customers.
Features such as CreditScore, Age, Tenure, and Balance were inspected for trends and patterns, and categorical variables like Gender and Geography were identified for encoding.
Visualizations like histograms, count plots, and correlation matrices were used to gain insights into the relationships between features and the target variable.

3. Data Preprocessing
Data preprocessing is a critical step to ensure that the dataset is clean and suitable for model training. In this step:

Irrelevant Columns Removal: RowNumber, CustomerId, and Surname were dropped as they do not contribute to customer behavior.
Encoding Categorical Variables:
Gender was label encoded into numerical values.
Geography was one-hot encoded to avoid introducing ordinal bias.
Addressing Class Imbalance: The target variable, Exited, exhibited a class imbalance. SMOTE (Synthetic Minority Over-sampling Technique) was applied to oversample the minority class (Exited = 1) and balance the dataset.
Feature Scaling: Numerical features were scaled using StandardScaler to ensure that all features have the same magnitude, which is critical for models like Logistic Regression and SVM.
4. Model Building
The classification model was built using a stacking ensemble approach, which combines the strengths of multiple machine learning algorithms to achieve better performance. The following models were used:

Base Models:
Logistic Regression: A simple yet effective linear model.
Random Forest Classifier: A robust ensemble model that handles non-linear relationships well.
LightGBM: A gradient boosting framework optimized for speed and performance.
XGBoost: Another powerful gradient boosting algorithm known for its accuracy.
Meta-Model:
Gradient Boosting Classifier: Used as the final model to aggregate predictions from the base models.
The stacking classifier integrates predictions from base models and uses the meta-model for the final classification. Cross-validation ensured that the model was robust and generalizable.

5. Evaluation
The model was evaluated using several metrics to assess its performance comprehensively:

Classification Report: Provides precision, recall, F1-score, and accuracy for both classes (Exited = 0 and Exited = 1).
Confusion Matrix: Visualized using a heatmap to analyze true positive, true negative, false positive, and false negative predictions.
ROC-AUC Score: Measured the overall ability of the model to distinguish between classes, with higher values indicating better performance.

F1-Score: A weighted average of precision and recall, with particular attention to improving the F1-score for the minority class (Exited = 1).
The results demonstrated significant improvement in the F1-score for the churned class due to SMOTE and the advanced stacking ensemble.

6. Feature Importance
Feature importance analysis was conducted to identify which features contributed most to the prediction:

Models like Random Forest, LightGBM, and Gradient Boosting provided insights into feature importance.
A bar plot of feature importance highlighted key drivers of customer churn, such as Age, Balance, and Tenure.
Understanding feature importance not only enhances interpretability but also provides actionable insights for business strategies.

7. Challenges and Solutions
Several challenges were encountered during this project:

Class Imbalance: The dataset exhibited a skewed distribution of the target variable. This was mitigated using SMOTE, which improved the F1-score for the minority class.
Hyperparameter Tuning: Selecting optimal parameters for the base models and meta-model required experimentation to balance model complexity and performance.
Integration of Stacking Ensemble: Ensuring seamless communication between base models and the meta-model required careful implementation and validation.
8. Results and Conclusion
The final model achieved the following results:

Accuracy: 91%
F1-Score for Class 0 (Retained Customers): 94%
F1-Score for Class 1 (Churned Customers): 83%
ROC-AUC Score: 95%
The stacking ensemble demonstrated superior performance, effectively addressing the initial class imbalance and improving predictions for churned customers. The results highlight the power of ensemble techniques in handling complex classification problems and provide actionable insights for reducing customer churn.