Customer Churn Analysis Report

1. Main Objective

The primary objective of this analysis is to build a predictive model to identify customers who are likely to churn. By accurately predicting churn, the company can implement targeted retention strategies to reduce customer turnover, thereby increasing revenue and customer satisfaction.

2. Dataset Description

The dataset contains 23 features related to customer services, usage patterns, and subscription details, with a total of 2,113 entries. Key features include **months** (customer tenure), **multiple** (whether the customer uses multiple lines), and **gb_mon** (monthly data usage), which help assess the customer's service engagement. Features like **security**, **backup**, **protection**, and **support** indicate subscriptions to additional services, while **unlimited** reflects if the customer has an unlimited data plan. **contract** specifies the contract type (month-to-month, one-year, or two-year), and **paperless** indicates whether the customer uses paperless billing. The **monthly** charge is a crucial feature affecting churn, as is **satisfaction**, representing customer happiness. Payment methods like **credit card** or **mailed check** and internet types such as **DSL** or **fiber optic** are also included, alongside promotional offers (e.g., **offer A** to **offer E**). The target variable, **churn_value**, indicates whether a customer has churned (1 for churned, 0 for retained).

3. Data Cleaning and Preprocessing

Data was cleaned, and missing values were handled. Categorical variables were label encoded, and numerical features were standardized for model training.

4. Model Training and Evaluation

Three models were trained using hyperparameter tuning with GridSearchCV: Logistic Regression, Random Forest, and Gradient Boosting. The models were evaluated based on precision, recall, and accuracy. Also, visual methods of evaluation such as confusion matrices and ROC curves were also utilized

5. Tuned Hyperparameters

5.1 Logistic Regression (Tuned)

Best Hyperparameters: {'C': 1, 'solver': 'liblinear'}

5.2 Random Forest (Tuned)
Best Hyperparameters: {'max_depth': None, 'n_estimators': 100}

5.3 Gradient Boosting (Tuned)
Best Hyperparameters: {'learning_rate': 0.1, 'n_estimators': 100}

6. Results

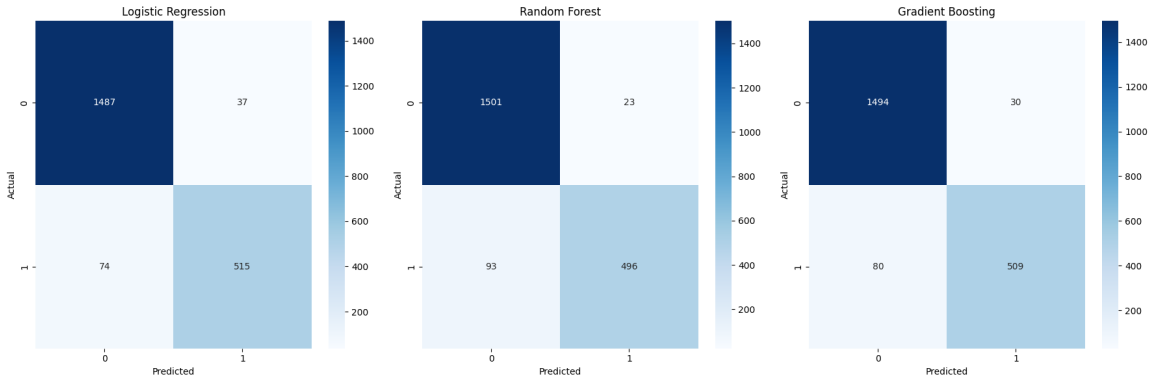
6.1 Classification Report of all models

Models	Accuracy	AUC-ROC	*Precision	*Recall	*F1-Score
Logistic Regression	0.9474	0.9869	0.95	0.95	0.95
Random Forest	0.9451	0.9786	0.95	0.95	0.94
Gradient Boosting	0.9479	0.9865	0.95	0.95	0.95

*These are weighted average scores

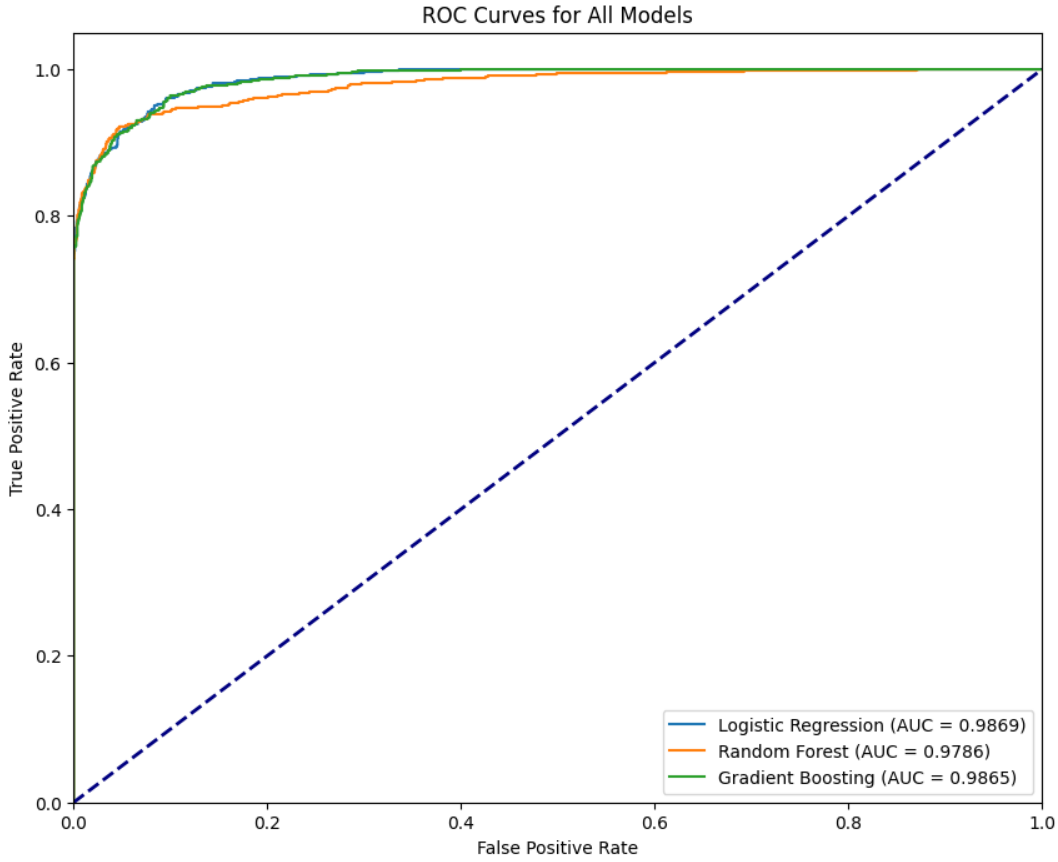
The classification report table summarizes the performance of three models: Logistic Regression, Random Forest, and Gradient Boosting. All models show high accuracy, with Logistic Regression and Gradient Boosting slightly outperforming Random Forest. The AUC-ROC scores, which measure the models' ability to distinguish between classes, are similarly high, with Logistic Regression leading at 0.9869. Precision, recall, and F1-score (weighted averages) are nearly identical across the models, indicating that all models handle both precision (correct positive predictions) and recall (capturing actual positives) well. Overall, all three models perform similarly, with Logistic Regression and Gradient Boosting having a slight edge in accuracy and AUC-ROC.

6.2 Confusion Matrix of all models



The confusion matrices show similar performance across all three models. Logistic Regression correctly predicts 1,487 true negatives and 515 true positives, with 74 false negatives and 37 false positives. Random Forest has 1,501 true negatives and 496 true positives, with 93 false negatives. Gradient Boosting balances well with 1,494 true negatives and 509 true positives, misclassifying 80 false negatives. Overall, the models perform similarly, with Logistic Regression and Gradient Boosting slightly outperforming Random Forest in handling false negatives.

6.2 ROC Curve of all models



The ROC curves for all three models—Logistic Regression, Random Forest, and Gradient Boosting—are nearly identical, indicating excellent model performance in distinguishing between churned and non-churned customers. Logistic Regression and Gradient Boosting have the highest AUC scores of 0.9869 and 0.9865, respectively, while Random Forest has a slightly lower AUC of 0.9786. Overall, all models show strong discriminatory power, with Logistic Regression performing marginally better in terms of AUC.

7. Recommendation

Based on the classification report, confusion matrices, and ROC curves, all three models—Logistic Regression, Random Forest, and Gradient Boosting—perform exceptionally well in predicting customer churn, with high accuracy, AUC-ROC scores, and balanced precision and recall. However, **Logistic Regression** stands out slightly with the highest AUC-ROC (0.9869) and fewer false negatives in the confusion matrix, making it the most reliable model. Given its strong performance and interpretability, Logistic Regression is recommended for this classification task.

8. Key Findings

The most important features driving customer churn are as follows:

• **Satisfaction**: In both Random Forest and Gradient Boosting models, customer satisfaction is by far the most significant feature. Lower satisfaction scores are strongly linked to higher churn, highlighting the importance of keeping customers satisfied to reduce churn.

• **Contract Type**: The length of the contract (month-to-month, one-year, two-year) also plays a crucial role. Shorter contracts tend to increase the likelihood of churn, as customers with month-to-month contracts are more flexible and likely to leave.

• **Monthly Charges**: Although less influential than satisfaction, monthly charges remain an important factor. Customers with higher monthly bills are more prone to churn, possibly due to the perceived lack of value for the cost.

• **Tenure (Months as a Customer)**: The number of months a customer has been with the company affects churn, with newer customers being more likely to leave.

• **GB per Month (Data Usage)**: While less significant, data usage still contributes to churn predictions. Customers who use more data may churn if they feel that their data plan does not meet their needs.

• **Security Services**: In the Gradient Boosting model, security service subscriptions also play a minor role in influencing churn, suggesting that customers without such services might be more likely to leave.

9. Next Steps

1. Feature Enhancement: Adding more customer-related data, such as satisfaction survey scores or support interaction history, could improve the model's predictive power and provide better explanations for why customers churn.

2. Regular Model Recalibration: As new data becomes available, regularly retraining the model will ensure it stays accurate and reflects any changes in customer behaviour or external factors.

3. Advanced Explainability Tools: Utilizing tools like SHAP or LIME can help explain individual predictions, offering deeper insights into why certain customers are at risk of churning, aiding in targeted retention efforts.