CUSTOMER CHURN PREDICTION

Capstone Project

Submitted By
Rejin Raveendran
Rejincr001@gmail.com

Submitted On:16 Mar 2025

CUSTOMER CHURN PREDICTION

Model Building

The model can make wrong predictions as:

When evaluating a customer churn prediction model, it is crucial to understand the types of errors the model can make and their implications. The model can make wrong predictions in two primary ways:

1. False Positives (Type I Error):

- The model predicts that a customer will churn, but in reality, they do not.
- Implication: The business may spend resources (e.g., discounts, offers) on retaining customers who were not going to leave, leading to unnecessary costs.

2. False Negatives (Type II Error):

- The model predicts that a customer will not churn, but in reality, they do.
- Implication: The business loses customers without attempting to retain them, resulting in lost revenue and potential damage to customer relationships.

Which case is more important?

False Negatives are Often More Critical:

- Losing a customer (FN) is typically more costly than spending resources on retaining a customer who wasn't going to leave (FP).
- Retaining customers is usually cheaper than acquiring new ones, so missing a churning customer (FN) can have a higher long-term impact.

Which metric to optimize?

Optimize Recall when minimizing FNs is critical (e.g., when losing customers is more costly than retaining non-churning customers). We would want Recall-Score to be maximized, the greater the Recall-Score higher the chances of predicting FN classes correctly.

Recall (Sensitivity):

- **Formula:** Recall = TP / (TP + FN)
- **Implication:** Ensures that most churning customers are identified, even if it means some non-churning customers are flagged.

Validation and Test data preparation for Modelling

We will need to transform the Validation and Test data sets with the same Imputers, Encoders and Scaler methods that we fitted on the Training set for data consistency and to avoid data leakage

Initial Model Building

We built the base models for Training Data set:

Training Performance for Various Models

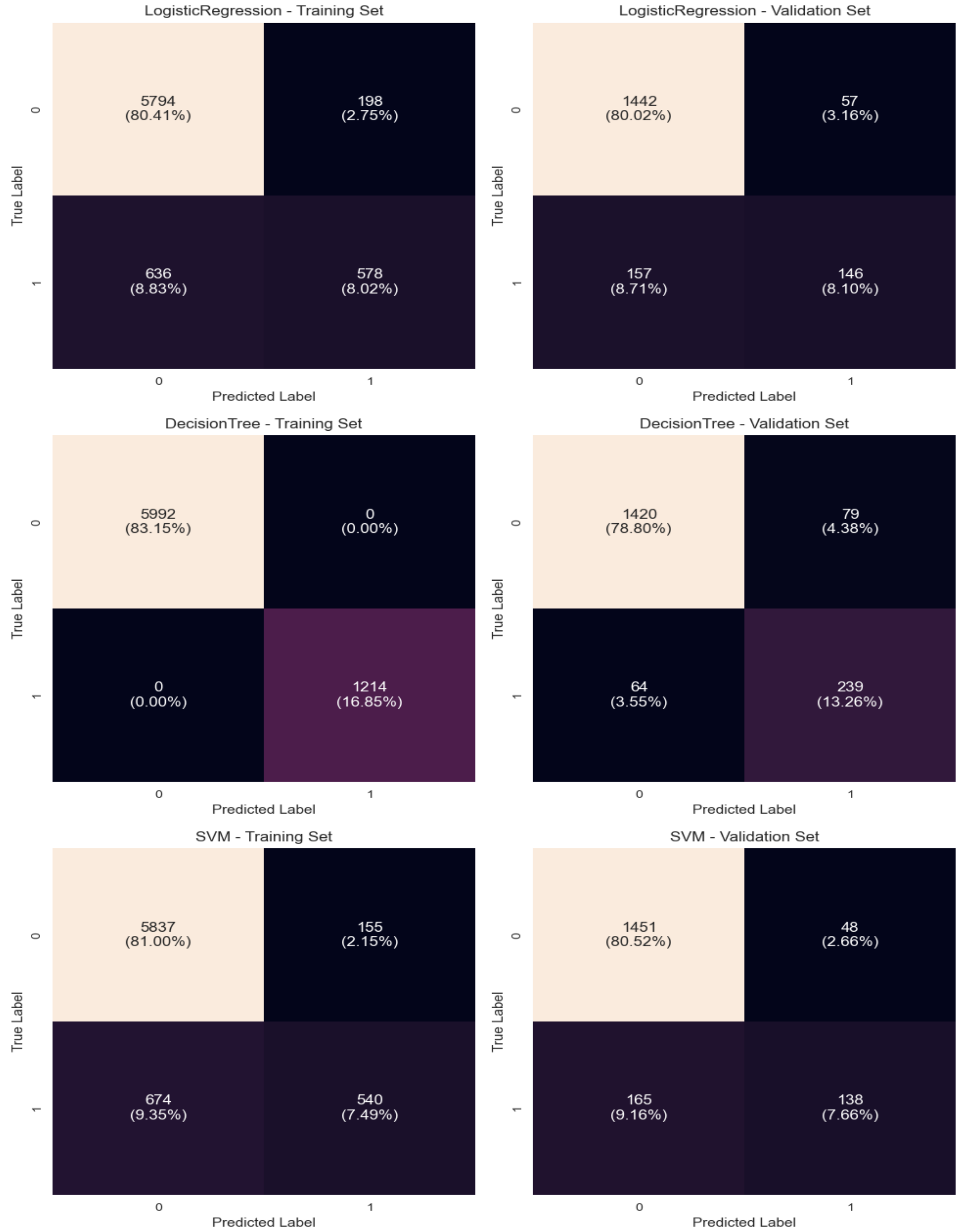
Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
DecisionTree	1	1	1	1	1
RandomForest	1	1	1	1	1
XGBoost	0.999445	1	0.996705	0.99835	1
Bagging	0.996669	0.999161	0.981054	0.990025	0.999953
GBM	0.913128	0.849169	0.588962	0.695525	0.942098
AdaBoost	0.889259	0.747031	0.518122	0.611868	0.888891
LogisticRegression	0.884263	0.744845	0.476112	0.580905	0.870041
SVM	0.884957	0.776978	0.444811	0.565741	0.869692

Validation Performance for Various Models

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
DecisionTree	1	1	1	1	1
XGBoost	1	1	1	1	1
RandomForest	0.999445	1	0.9967	0.998347	1
Bagging	0.991676	0.996552	0.953795	0.974705	0.999877
GBM	0.934517	0.896996	0.689769	0.779851	0.973584

AdaBoost	0.895117	0.774038	0.531353	0.630137	0.908511
LogisticRegression	0.880133	0.723077	0.465347	0.566265	0.876296
SVM	0.880688	0.761905	0.422442	0.543524	0.87319

Confusion Matrices for Training and Validation Sets



LogisticRegression - Training Set

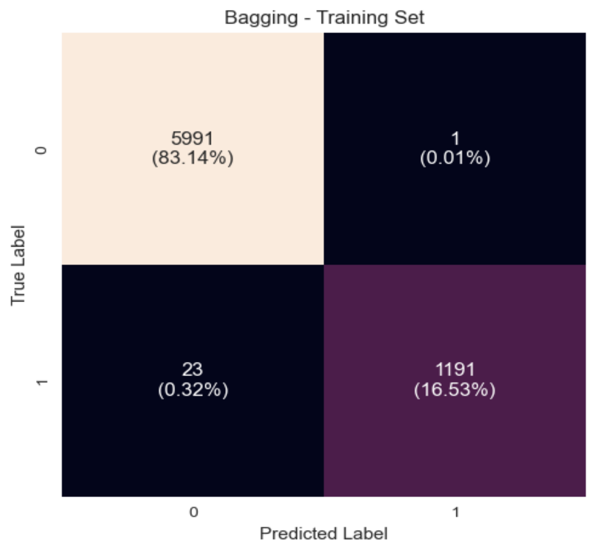
LogisticRegression - Validation Set

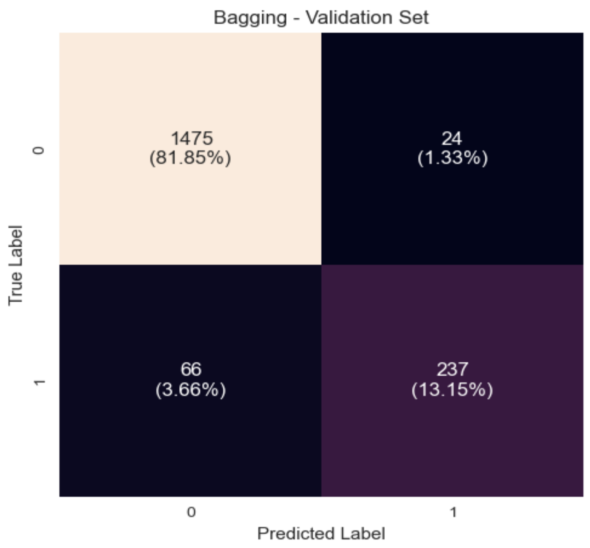
DecisionTree - Training Set

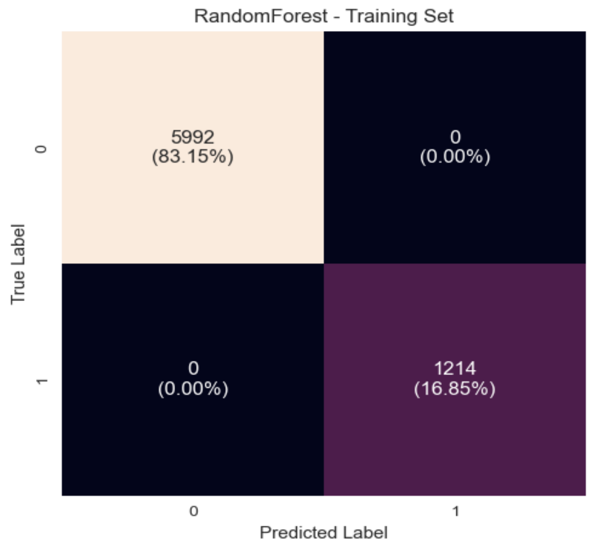
DecisionTree - Validation Set

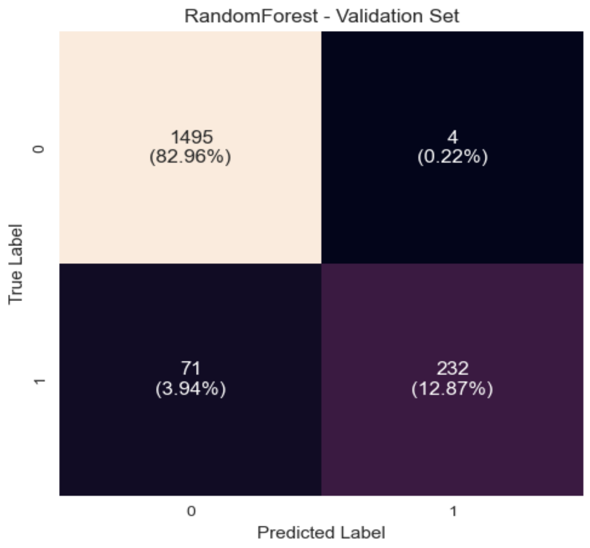
SVM - Training Set

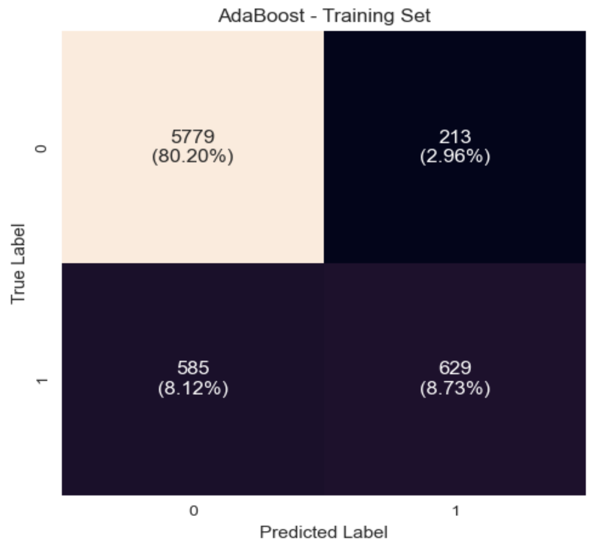
SVM - Validation Set

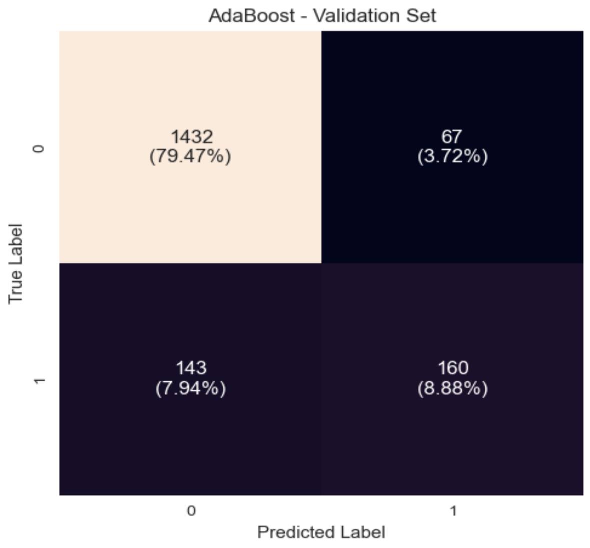

Bagging - Training Set

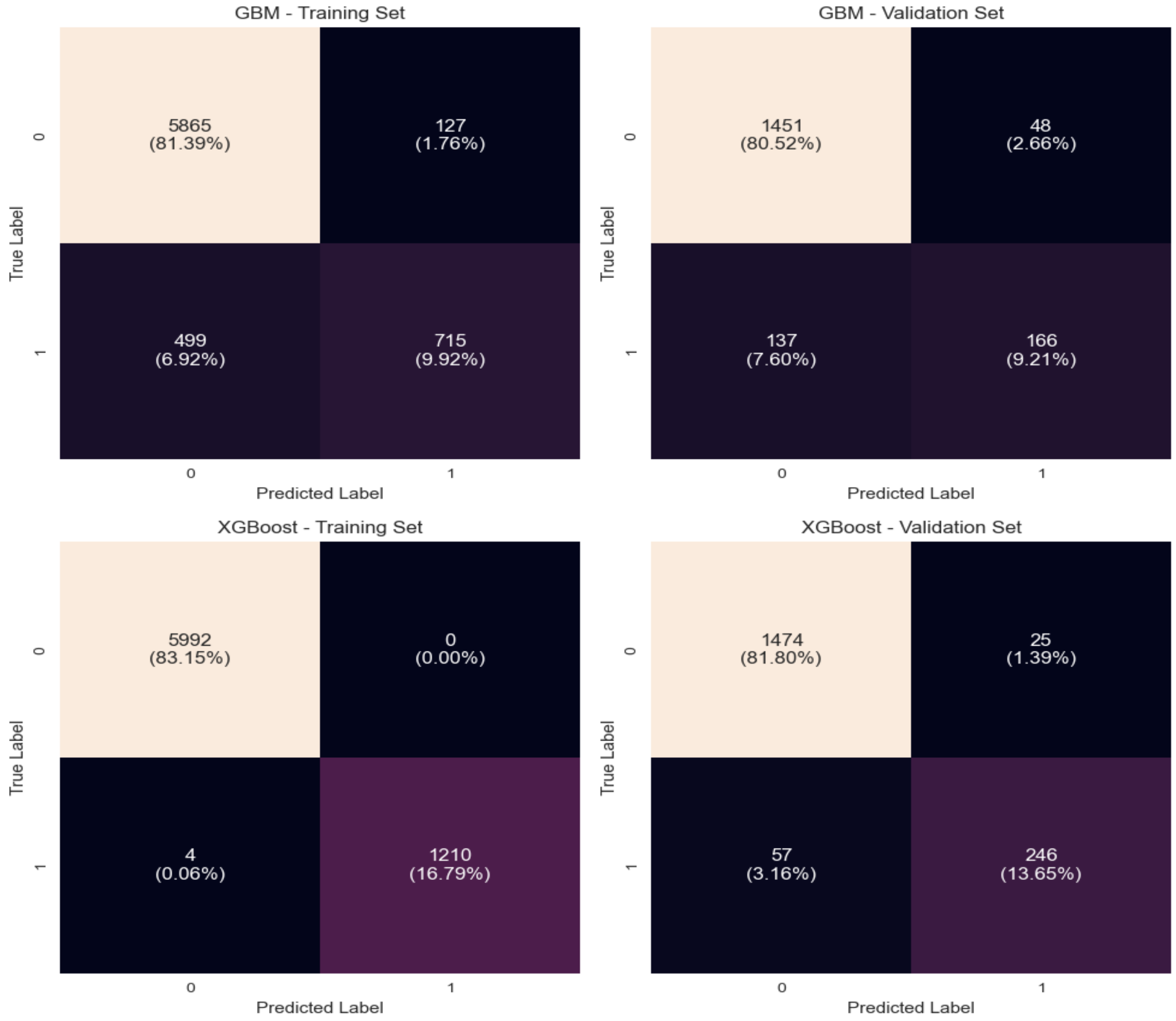

Bagging - Validation Set


RandomForest - Training Set


RandomForest - Validation Set


AdaBoost - Training Set


AdaBoost - Validation Set



Performance Comparison for Various Models

Model	Training Score	Validation Score	Difference	F1-Score	AUC-ROC
SVM	0.444811	0.455446	-0.010635	1	1
AdaBoost	0.518122	0.528053	-0.009931	1	1
LogisticRegression	0.476112	0.481848	-0.005736	0.998347	1

GBM	0.588962	0.547855	0.041107	0.974705	0.999877
XGBoost	0.996705	0.811881	0.184824	0.779851	0.973584
Bagging	0.981054	0.782178	0.198876	0.630137	0.908511
DecisionTree	1	0.788779	0.211221	0.566265	0.876296
RandomForest	1	0.765677	0.234323	0.543524	0.87319

From the above results:

- DecisionTree, RandomForest, Bagging, and XGBoost have near-perfect training scores (Accuracy, Precision, Recall, and F1-score all close to 1). However, their validation recall scores drop significantly (e.g., DecisionTree from 1.00 → 0.788 and RandomForest from 1.00 → 0.766), indicating potential overfitting.

- SVM, AdaBoost, LogisticRegression, and GBM have relatively smaller gaps between training and validation scores, indicating better generalization. But the recall scores for these except GBM are near to 0.50 which will cause random predictions.

- Comparatively, GBM has a reasonable validation recall (0.5479) and a moderate training score (0.5889), showing a balance between learning and avoiding overfitting.

In churn prediction, recall is critical as we need to minimize false negatives (missed churn cases). Models with high recall and low overfitting risk are preferable.

Recommended Models for Further Tuning:

- **Gradient Boosting Machine (GBM)**
  - Balanced training-validation recall (0.5889 → 0.5479).
  - Less prone to overfitting compared to DecisionTree/RandomForest.
  - Can improve with hyperparameter tuning (e.g., learning rate, number of estimators).

- **XGBoost**
  - High recall on training data (0.9967) but a notable drop in validation (0.8052).
  - Still performs better than Bagging/RandomForest in validation.
  - With proper regularization (tree depth, learning rate tuning), it can be optimized.

Final Recommendation For customer churn prediction:

- We will primarily focus on GBM for its balance and generalization ability.
- Also we will consider XGBoost as a secondary model if tuned properly to reduce overfitting.

Applying SMOTE for Oversampling the Training data

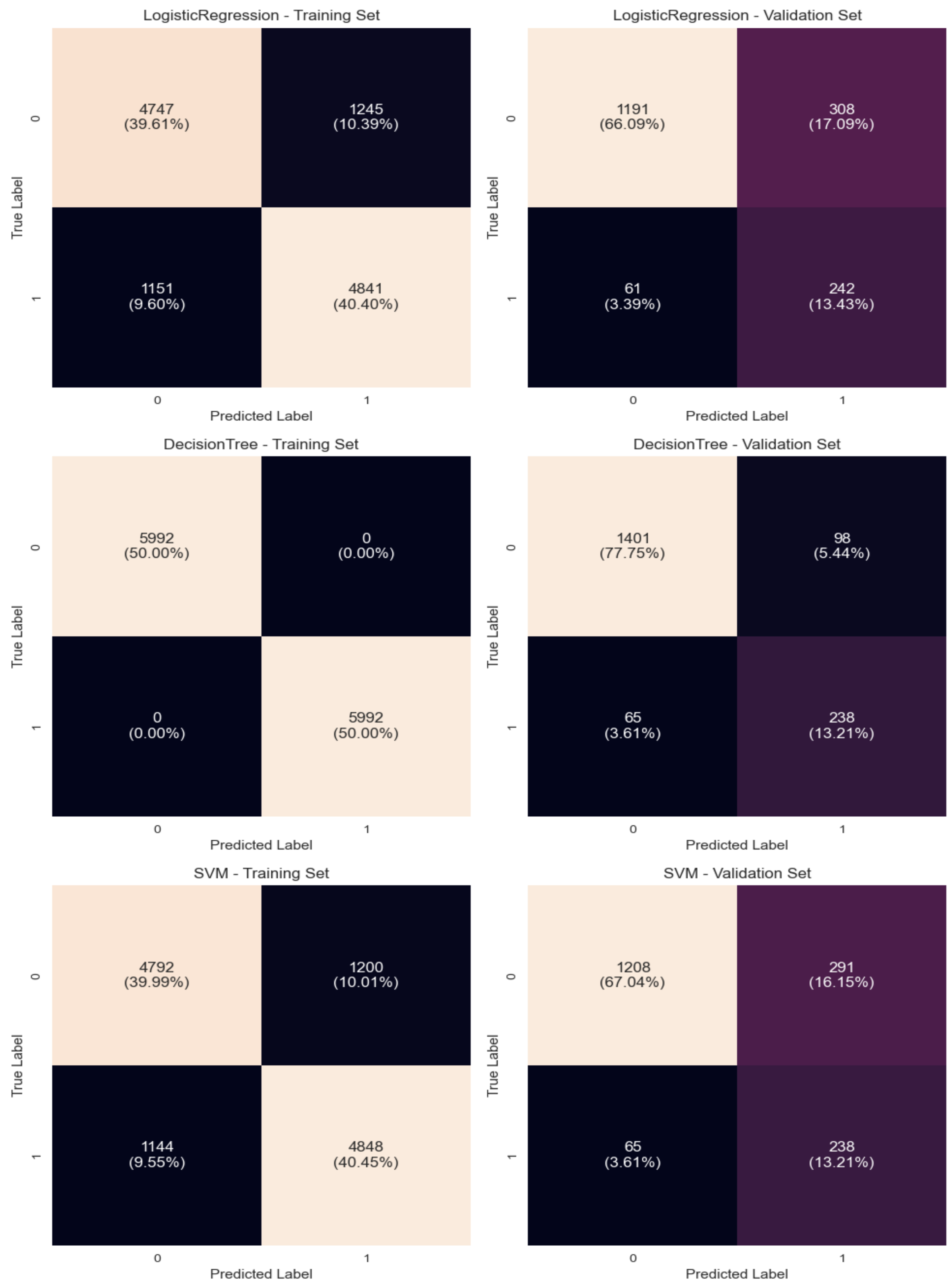
As the training data set was imbalanced, we will apply Oversampling technique to balance the X:y ratio and rebuild these models to see if that improves the Recall score.

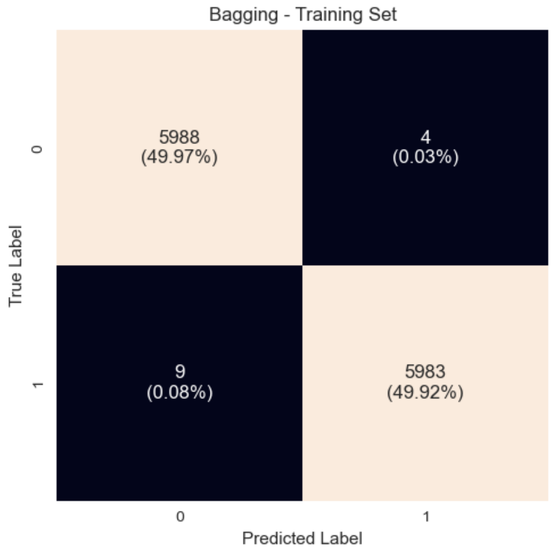
	counts of label: Yes	counts of label: No
Before Oversampling	1214	5992
After Oversampling	5992	5992

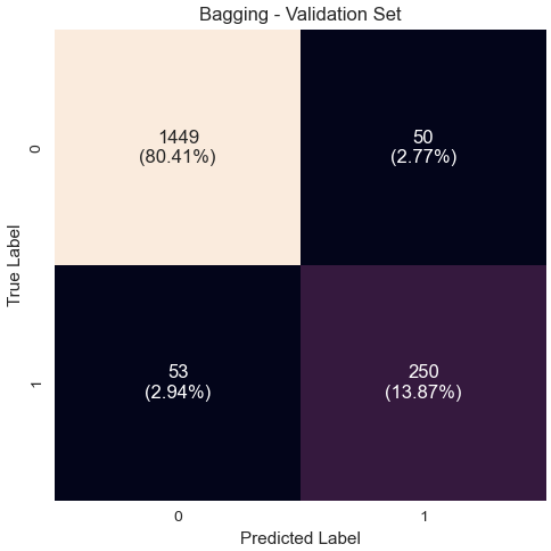
Performance Comparison for Various Models after Oversampling

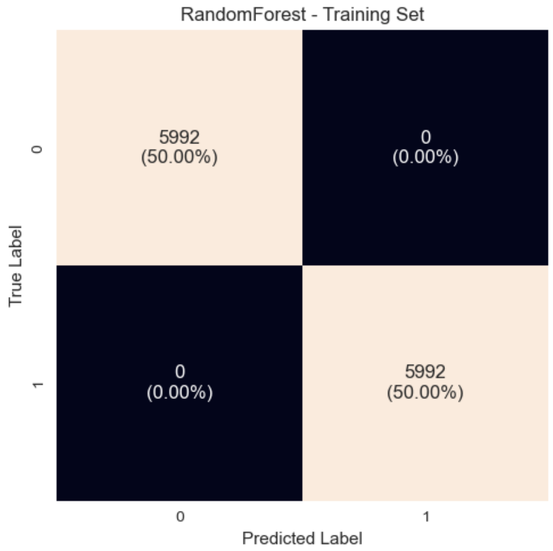
Model	Training Score	Validation Score	Difference	F1-Score	AUC-ROC
LogisticRegression	0.807911	0.79868	0.009231	1	1
SVM	0.809079	0.785479	0.0236	1	1
AdaBoost	0.87016	0.749175	0.120985	0.998347	1
RandomForest	1	0.851485	0.148515	0.974705	0.999877
XGBoost	0.999332	0.844884	0.154448	0.779851	0.973584
Bagging	0.998498	0.825083	0.173415	0.630137	0.908511
GBM	0.92006	0.709571	0.210489	0.566265	0.876296
DecisionTree	1	0.785479	0.214521	0.543524	0.87319

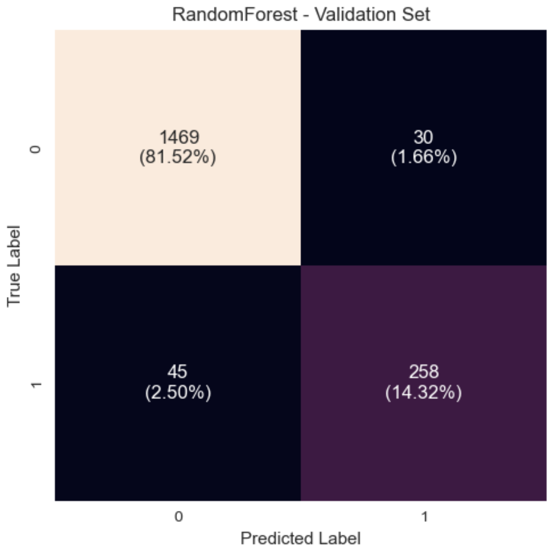
Confusion Matrices for Training and Validation Sets after Oversampling

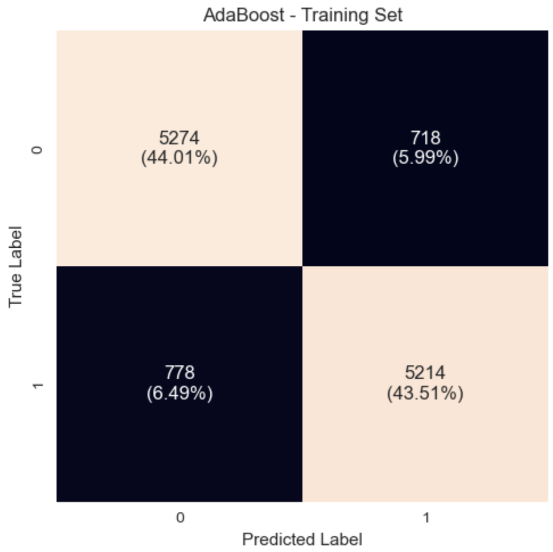


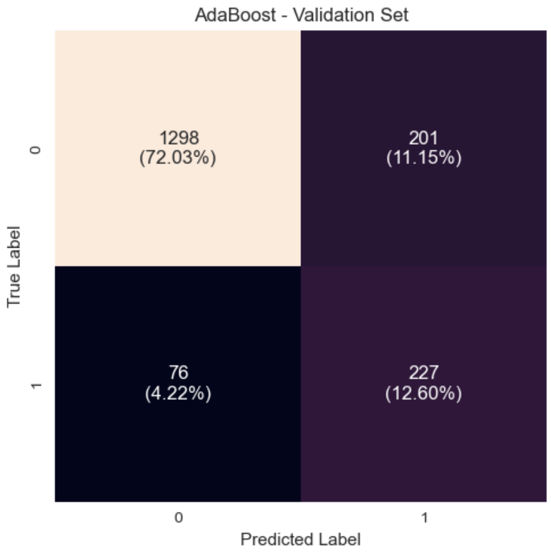

Bagging - Training Set

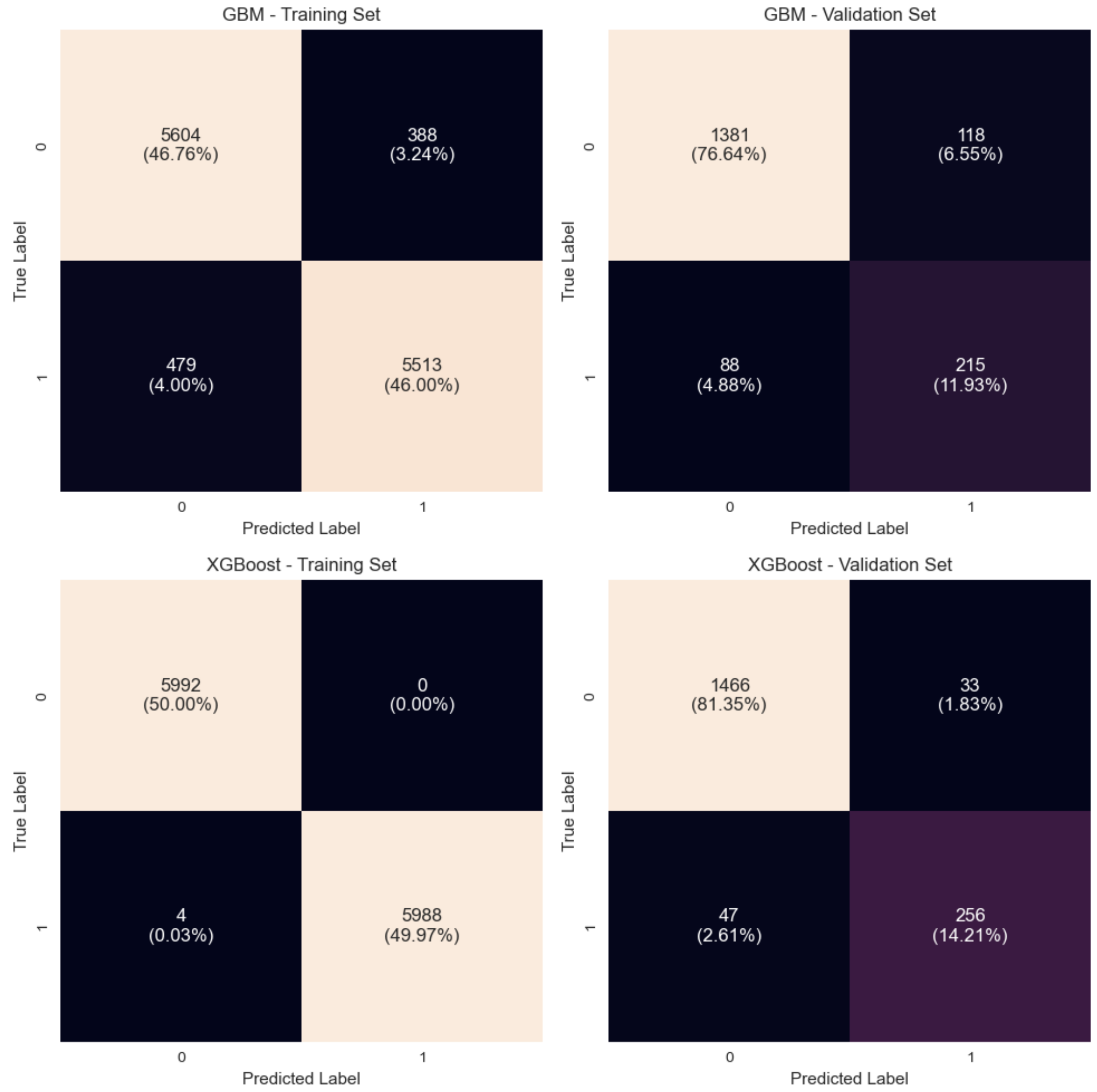

Bagging - Validation Set


RandomForest - Training Set


RandomForest - Validation Set


AdaBoost - Training Set


AdaBoost - Validation Set



Before Oversampling

- AdaBoost (0.528053) and GBM (0.547855) have the highest validation scores. The differences between training and validation scores are small, indicating less overfitting.
- DecisionTree and RandomForest have high training scores (1.000000) but significantly lower validation scores, indicating severe overfitting.
- The Best Trade-off was GBM (Gradient Boosting Machine) which has a reasonable validation score (0.547855) and a moderate difference between training and validation scores (0.588962).

After Oversampling

- LogisticRegression (0.798680) and SVM (0.785479) have the highest validation scores.
- Both models show small differences between training and validation scores, indicating good generalization.

- DecisionTree, RandomForest, and XGBoost have high training scores (1.000000 or close) but significantly lower validation scores, indicating overfitting.

So the best Trade-off is LogisticRegression and SVM show the best balance between training and validation performance.

Recommendations for Hyperparameter Tuning

i. Logistic Regression

- After oversampling, it has the highest validation score (0.798680).
- The difference between training and validation scores is small (0.009231), indicating good generalization.
- Logistic Regression is simple, interpretable, and less prone to overfitting compared to complex models like Decision Trees or Random Forests.

ii. SVM

- After oversampling, it has the second-highest validation score (0.785479).
- The difference between training and validation scores is moderate (0.023433), indicating reasonable generalization.
- SVM is effective for high-dimensional data and can handle non-linear decision boundaries using kernel functions.

iii. Random Forest

- After oversampling, it has a decent validation score (0.851485).
- While it shows some overfitting (difference of 0.148515), it is one of the better-performing ensemble models.
- RandomForest is robust, handles non-linear data well, and provides feature importance.**

iv. XGBoost

- After oversampling, it has a good validation score (0.844884).
- While it shows some overfitting (difference of 0.154448), it is one of the best-performing boosting models.
- XGBoost is highly effective for structured/tabular data and provides excellent performance with proper tuning.

Hyperparameter Tuning

Tuning LogisticRegression Model

Best Parameters for LogisticRegression:

solver	liblinear

penalty	l1
C	0.01

Best Recall Score for LogisticRegression: **0.818257702230991**

Checking model's performance on Training set

Model	Accuracy	Precision	Recall	F1-Score
LogisticRegression	0.790971	0.77723	0.815754	0.796026

Checking model's performance on Validation set

Model	Accuracy	Precision	Recall	F1-Score
LogisticRegression	0.779134	0.418803	0.808581	0.551802

Tuning Support Vector Machine (SVM) Model

```
Fitting 3 folds for each of 10 candidates totalling 30 fits
```

Best Parameters for SVM

kernel	rbf
gamma	scale
class_weight	balanced
C	10

Best Recall Score for SVM: **0.9958274081125024**

Checking model's performance on training set

Model	Accuracy	Precision	Recall	F1-Score
SVC	0.994993	0.992364	0.997664	0.995007

Checking model's performance on Validation set

Model	Accuracy	Precision	Recall	F1-Score
SVC	0.963374	0.881029	0.90429	0.892508

Tuning RandomForest Model

Best Parameters for RandomForest

n_estimators	100
min_samples_split	2
min_samples_leaf	1
max_features	log2
max_depth	30

Best Recall Score for RandomForest: **0.9729553425851538**

Checking model's performance on training set

Model	Accuracy	Precision	Recall	F1-Score
RandomForestClassifier	1	1	1	1

Checking model's performance on Validation set

Model	Accuracy	Precision	Recall	F1-Score
RandomForestClassifier	0.955605	0.888502	0.841584	0.864407

Tuning XGBoost Model

Fitting 5 folds for each of 20 candidates totalling 100 fits

Best Parameters for XGBoost

subsample	0.8
reg_lambda	0.5
reg_alpha	0.5
n_estimators	50
max_depth	7
learning_rate	0.1
gamma	0.1
colsample_bytree	1

Best Recall Score for XGBoost: **0.9330569018979367**

Checking model's performance on training set

Model	Accuracy	Precision	Recall	F1-Score
XGBClassifier	0.977887	0.977648	0.978138	0.977893

Checking model's performance on Validation set

Model	Accuracy	Precision	Recall	F1-Score
XGBClassifier	0.924528	0.761755	0.80198	0.78135

Model Comparison and Final Model Selection

Performance Comparison of Tuned models on Training Set

Model	Accuracy	Precision	Recall	F1-Score
LogisticRegression	0.790971	0.77723	0.815754	0.796026
SVM	0.994993	0.992364	0.997664	0.995007
RandomForest	1	1	1	1
XGBoost	0.977887	0.977648	0.978138	0.977893

Performance Comparison of Tuned models on Validation Set

Model	Accuracy	Precision	Recall	F1-Score
LogisticRegression	0.779134	0.418803	0.808581	0.551802

SVM	0.963374	0.881029	0.90429	0.892508
RandomForest	0.955605	0.888502	0.841584	0.864407
XGBoost	0.924528	0.761755	0.80198	0.78135

Based on these Performance metrics, SVC (Support Vector Classifier) model is the most optimal choice.

Model Interpretation

- **Highest Recall on Validation Set**

  Recall: 0.9043 (highest among all models)

  Since recall is crucial in churn prediction (to minimize false negatives and correctly identify churned customers), SVC performs best.

- **Good Precision & F1-Score**

  SVC has a high precision (0.8810) and F1-score (0.8925), showing a good balance between precision and recall.

  RandomForest and XGBoost have decent recall but lower F1-scores, indicating slightly lower overall performance.

So we use SVC as the final model for customer churn prediction.

Testing the Predictive Model Against the Test Set

Model	Accuracy	Precision	Recall	F1-Score
SVC	0.968028	0.911528	0.897098	0.904255

Since recall is the priority (to correctly identify churned customers), the SVC model remains the best choice as it maintains a high recall (0.8971) even on the test set.

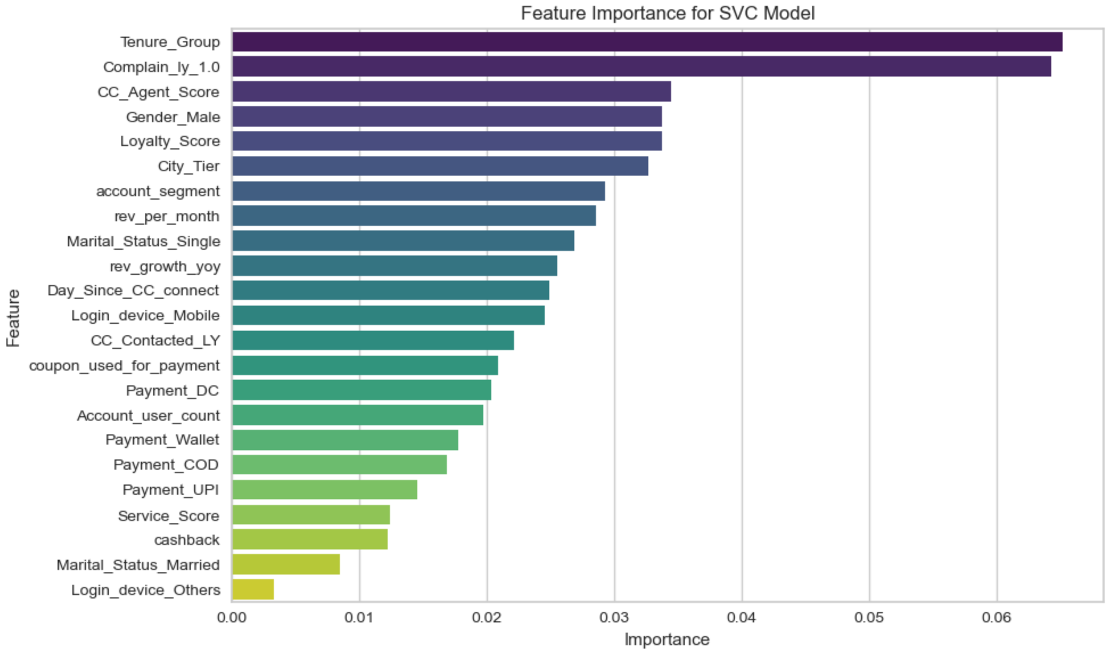
This confirms its generalizability and robustness across datasets.

Feature Importance

For the Support Vector Classifier (SVC), performing feature importance and confusion matrix analysis requires specific techniques because SVC does not inherently provide feature importance like tree-based models (e.g., Random Forest).

As we used non-linear kernels (RBF) as the best parameter, we can use permutation importance, which measures the drop in model performance when a feature's values are shuffled.

Feature	Importance
Tenure_Group	0.065142
Complain_ly_1.0	0.064298
CC_Agent_Score	0.034503
Gender_Male	0.033792
Loyalty_Score	0.033748
City_Tier	0.032682
account_segment	0.029307
rev_per_month	0.028552
Marital_Status_Single	0.026909
rev_growth_yoy	0.025577
Day_Since_CC_connect	0.024911
Login_device_Mobile	0.024556
CC_Contacted_LY	0.022202
coupon_used_for_payment	0.02087
Payment_DC	0.020382
Account_user_count	0.019716
Payment_Wallet	0.017806
Payment_COD	0.016874
Payment_UPI	0.014565
Service_Score	0.012389
cashback	0.012211
Marital_Status_Married	0.008481
Login_device_Others	0.003286



Observations

Top 3 Most Important Features:

- **Tenure_Group**: Customer tenure is the most influential factor in predicting churn. This suggests that how long a customer has been associated with the company significantly impacts churn probability.
- **Complain_ly_1.0**: Customers who complained in the last year are highly likely to churn, indicating dissatisfaction.
- **CC_Agent_Score**: The quality of customer care interactions (e.g., service ratings) plays a critical role in retention.

Demographics and Loyalty Factors:

- **Gender_Male**: Male customers appear to have a higher correlation with churn compared to females.
- **Loyalty_Score**: Customer loyalty score significantly influences retention.
- **Marital_Status_Single**: Single customers may have a higher churn rate compared to married ones.

Revenue and Payment Patterns:

- **rev_per_month & rev_growth_yoy**: Monthly revenue and year-over-year revenue growth are key indicators of customer stability.
- **Payment_Wallet & Payment_UPI**: Certain payment methods are linked to churn behavior, possibly due to ease of transaction or subscription preferences.

Customer Support & Engagement:

- **CC_Contacted_LY**: Customers who interacted with customer support in the last year have a noticeable impact on churn, possibly reflecting past issues.
- **Day_Since_CC_Connect**: The number of days since the last contact with customer care also contributes to churn risk.

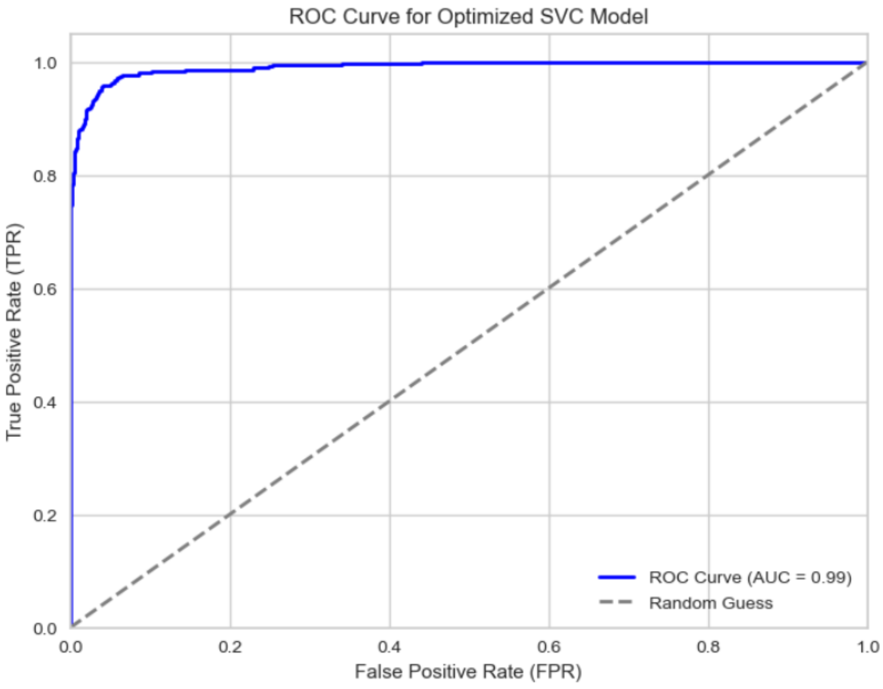
Least Important Features:

- **Marital_Status_Married & Login_device_Others** contribute minimally to churn prediction.
- **cashback**: Offering cashback might not be as influential in reducing churn as expected.

Business Implications & Actions:

- **Focus on customer tenure management**: Implement loyalty programs targeting long-term customers.
- **Enhance customer service**: Address complaints quickly to improve satisfaction.
- **Monitor revenue growth**: Identify declining revenue trends early.
- **Personalized engagement**: Use targeted offers based on customer demographics and past interactions.

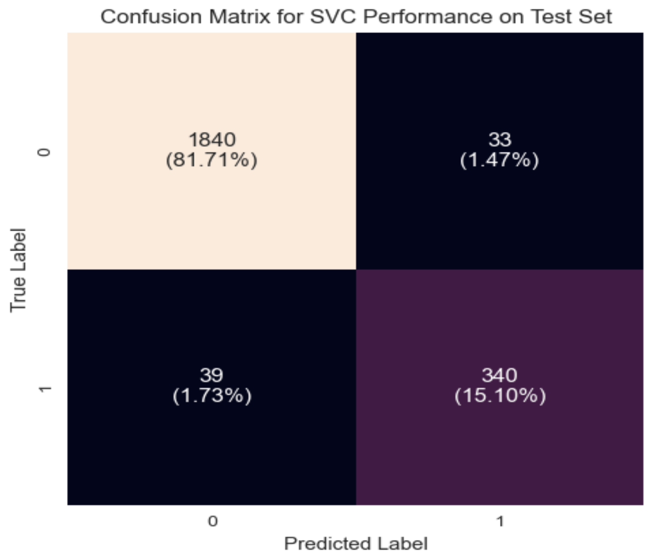
ROC-AUC Curve



Observations

- **High AUC Score (0.99)**
  - The model achieves an AUC of 0.99, indicating exceptional classification performance.
  - This suggests that the model can effectively distinguish between positive and negative classes.
- **Near-Perfect Performance**
  - The ROC curve is close to the top-left corner, demonstrating high sensitivity (True Positive Rate) with very low False Positives.
  - This means the model correctly identifies most positive cases while minimizing incorrect predictions.
- **Minimal False Positives**
  - The False Positive Rate (FPR) remains close to zero for most threshold values.
  - This indicates that the model makes very few incorrect positive predictions.
- **Strong Separation of Classes**
  - The ROC curve (blue line) is significantly above the random classifier line (grey dashed line).
  - This confirms that the model performs far better than random guessing.

Confusion Matrix



Observations

- **High Recall (89.71%)**: The model successfully captures most churn cases, which is crucial for customer retention strategies.
- **Low False Negatives**: Only 39 actual churn cases were missed, meaning the model effectively identifies customers at risk of leaving.
- **Low False Positives**: With just 33 non-churn customers misclassified as churners, unnecessary interventions are minimal.

Key Insights and Business Recommendations

Model Implication on Business

1. **High Recall (89.71%) Ensures Most Churners Are Identified**
   - The model successfully flags customers likely to churn, enabling proactive retention strategies like personalized offers or engagement campaigns.
2. **False Negatives (39 Customers) Need Attention**
   - These customers were actual churners but weren't identified as such.

   - Analyzing their profiles can help refine the model and improve targeting.
3. **False Positives (33 Customers Misclassified as Churners) Have a Minimal Business Impact**
   - While some non-churners might receive unnecessary interventions, this is preferable to missing actual churners.

   - The business can optimize retention strategies to ensure cost-effectiveness.

Business Insights

1. **Tenure Group & Complaints Are Key Indicators of Churn**
   - Customers with shorter tenure are more likely to churn. Retention strategies should focus on new customers.
   - Complaints in the last year strongly influence churn. Improving customer service and addressing complaints proactively can reduce churn.
2. **Customer Service Score Matters**
   - CC_Agent_Score and Loyalty_Score are among the top features. Customers who rate support poorly or have low loyalty scores are at high risk.

   - Providing personalized offers and enhancing support quality can help retain customers.
3. **Revenue and Payment Methods Influence Churn**
   - Higher revenue per month correlates with retention. Understanding spending patterns and incentivizing consistent spending can enhance retention.

   - Payment methods like Wallet, UPI, and COD have lower importance but still contribute to churn. Offering more flexible and preferred payment options can improve customer experience.
4. **Demographics Play a Role**
   - Gender (Male) and Marital Status (Single) appear as significant predictors. Tailored marketing campaigns targeting these demographics could improve engagement.

   - City_Tier and Account Segment indicate that location and customer type influence churn behavior.

Business Recommendations

1. **Improve Customer Support & Issue Resolution**
   Since complaints and CC_Agent_Score impact churn, a dedicated effort to resolve issues quickly will enhance customer satisfaction.

2. **Target At-Risk Segments with Personalized Offers**
   Short-tenure, single, and male customers need engagement strategies such as loyalty rewards, discounts, and exclusive offers.

3. **Enhance Customer Onboarding & Early Retention**
   Since tenure group is a key churn factor, improving onboarding, providing early engagement benefits, and implementing proactive check-ins can boost retention.

4. **Use Predictive Insights to Prioritize Interventions**
   Focus retention efforts on customers with high churn probability based on key features like complaints, loyalty score, and revenue trends.

5. **Refine Marketing and Payment Strategies**
   Offering preferred payment options and understanding revenue per month behavior can help reduce churn and increase customer lifetime value.

By acting on these insights, the business can significantly reduce churn, improve customer retention, and drive long-term profitability.