Customer Churn Prediction In Telecom Company

Project Outline

Overview

Methodology

Loading Libraries and Data

Preprocessing

Exploratory Data Analysis

Model Building

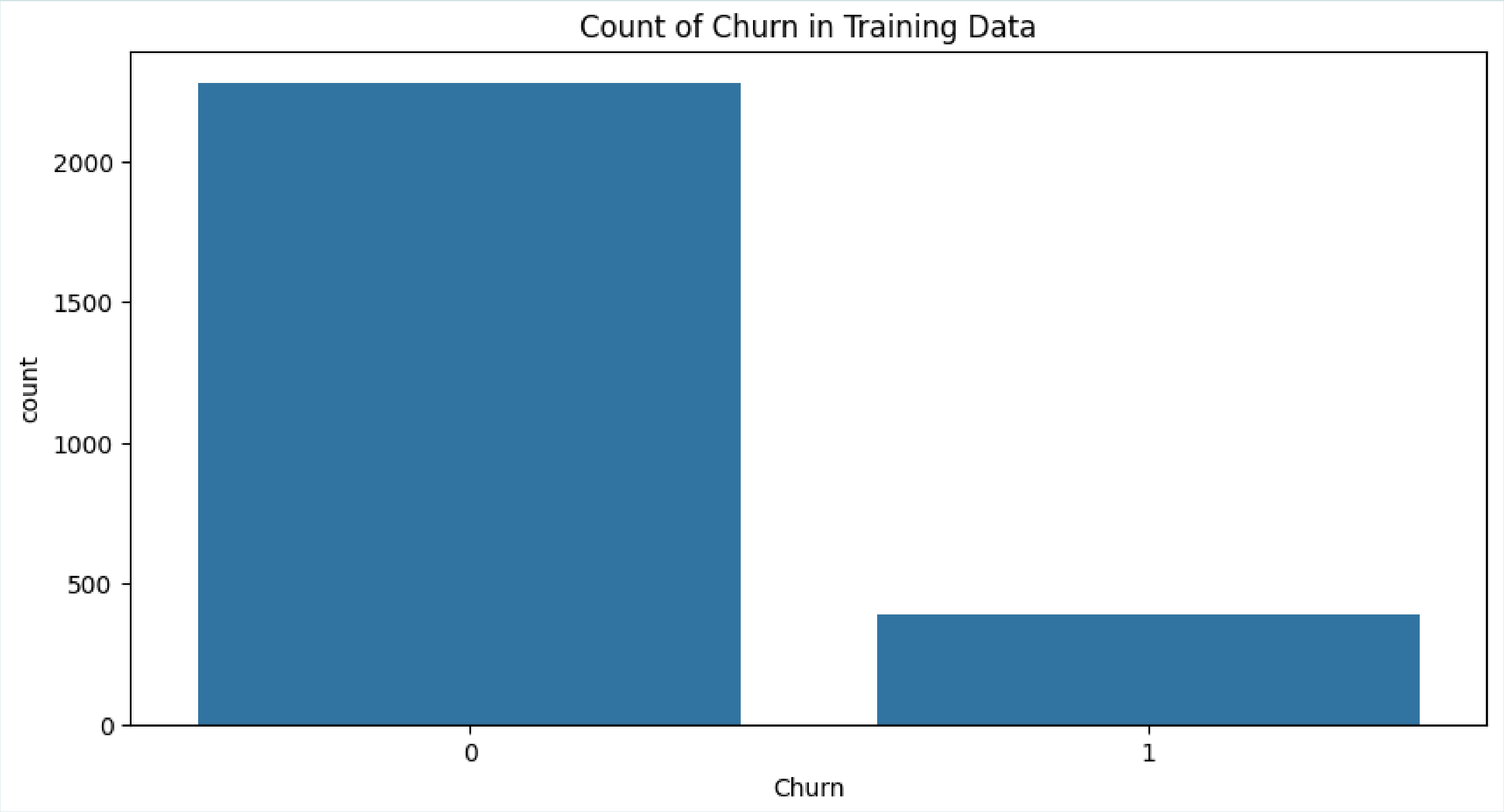
Evaluating Model Performance

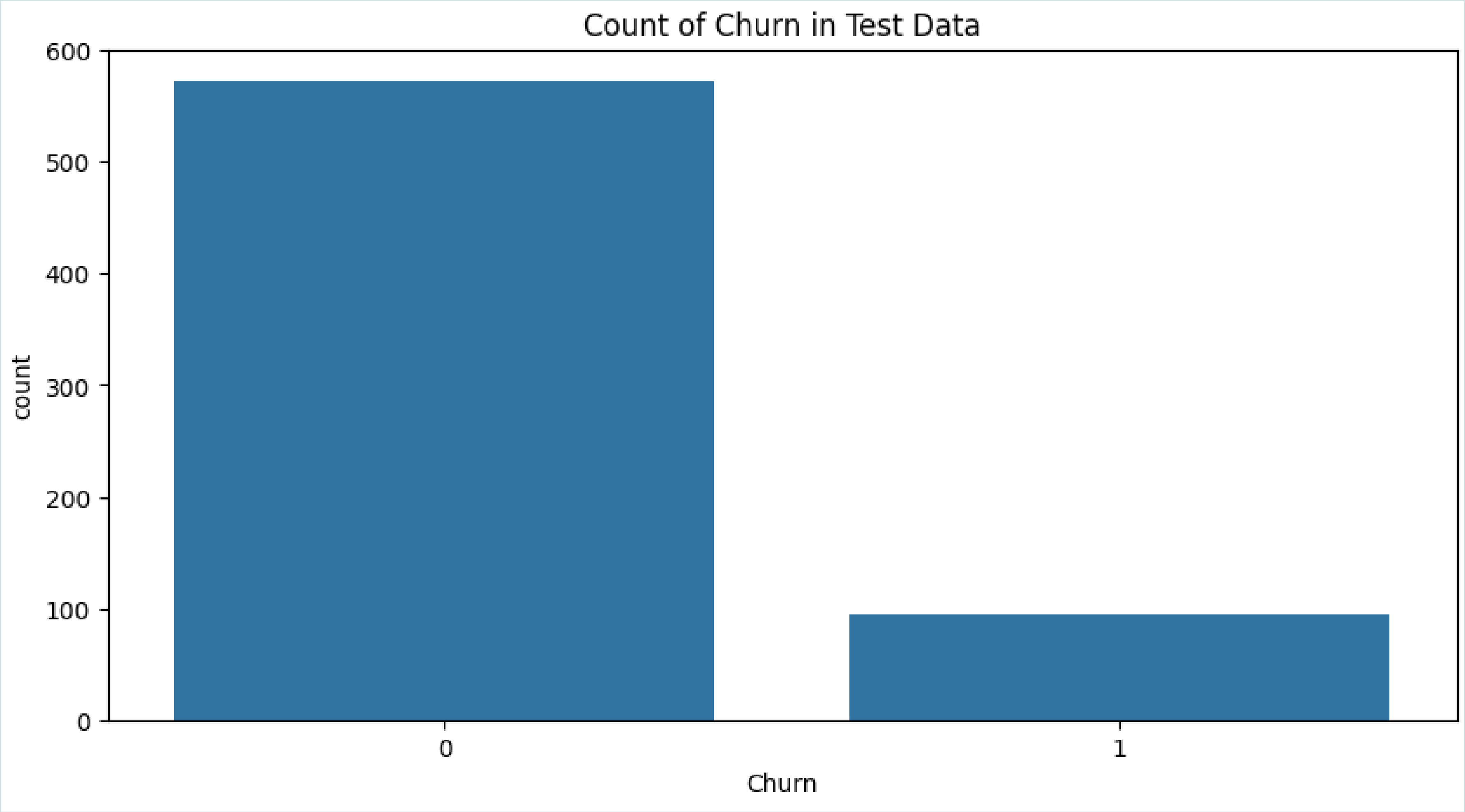
Conclusion

Methodology

- Changing Data Type
- Handling Missing Values
- Dummy Variables
- Handling Outliers by *IQR* method
- Normalization by *Min-Max* Scaling
- Building *Logistic Regression & Random Forest* model

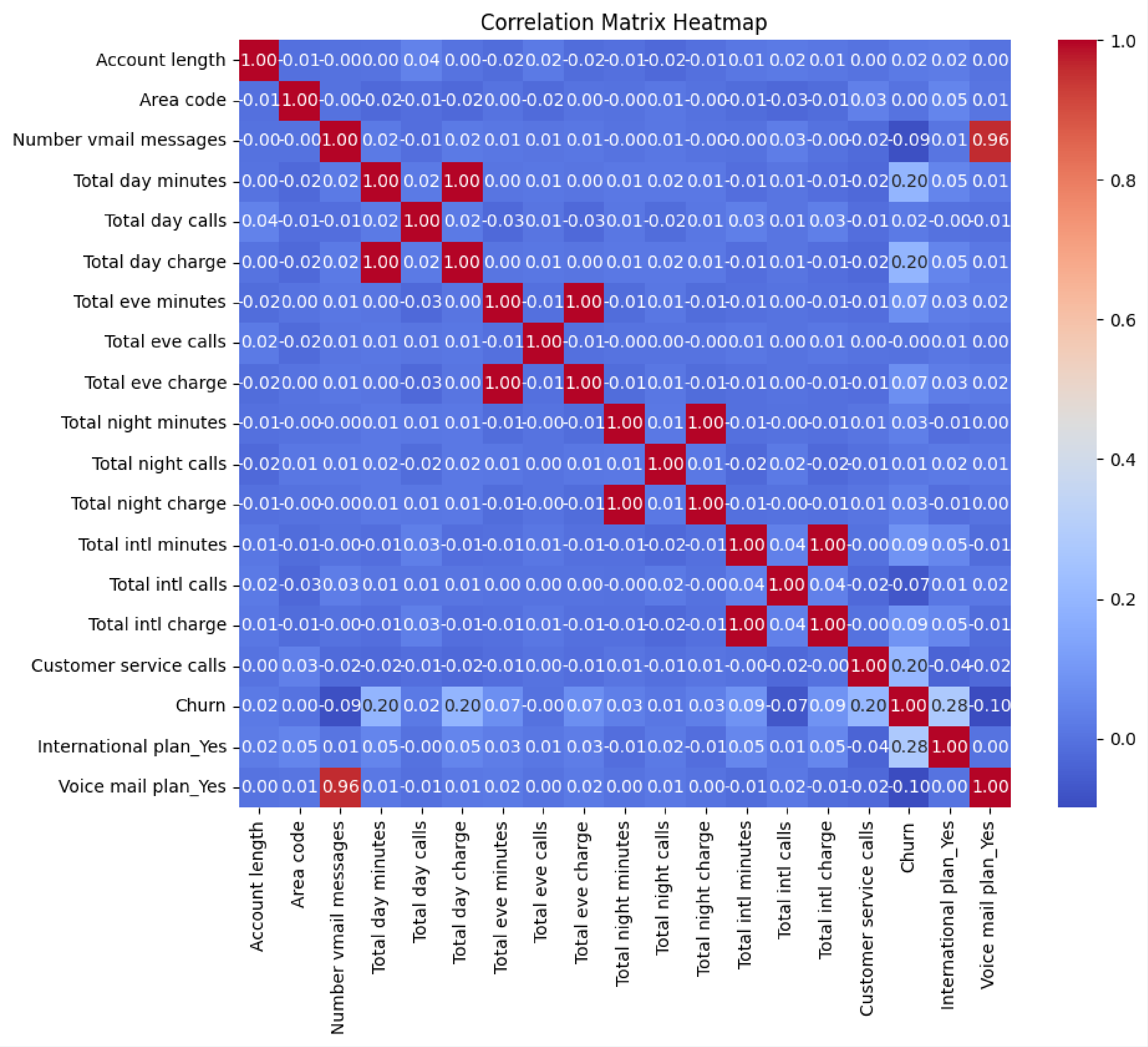
Exploratory Data Analysis





We can observe that there are only few people trying to churn (1) and maximum customers are not planning to churn (0).

Exploratory Data Analysis

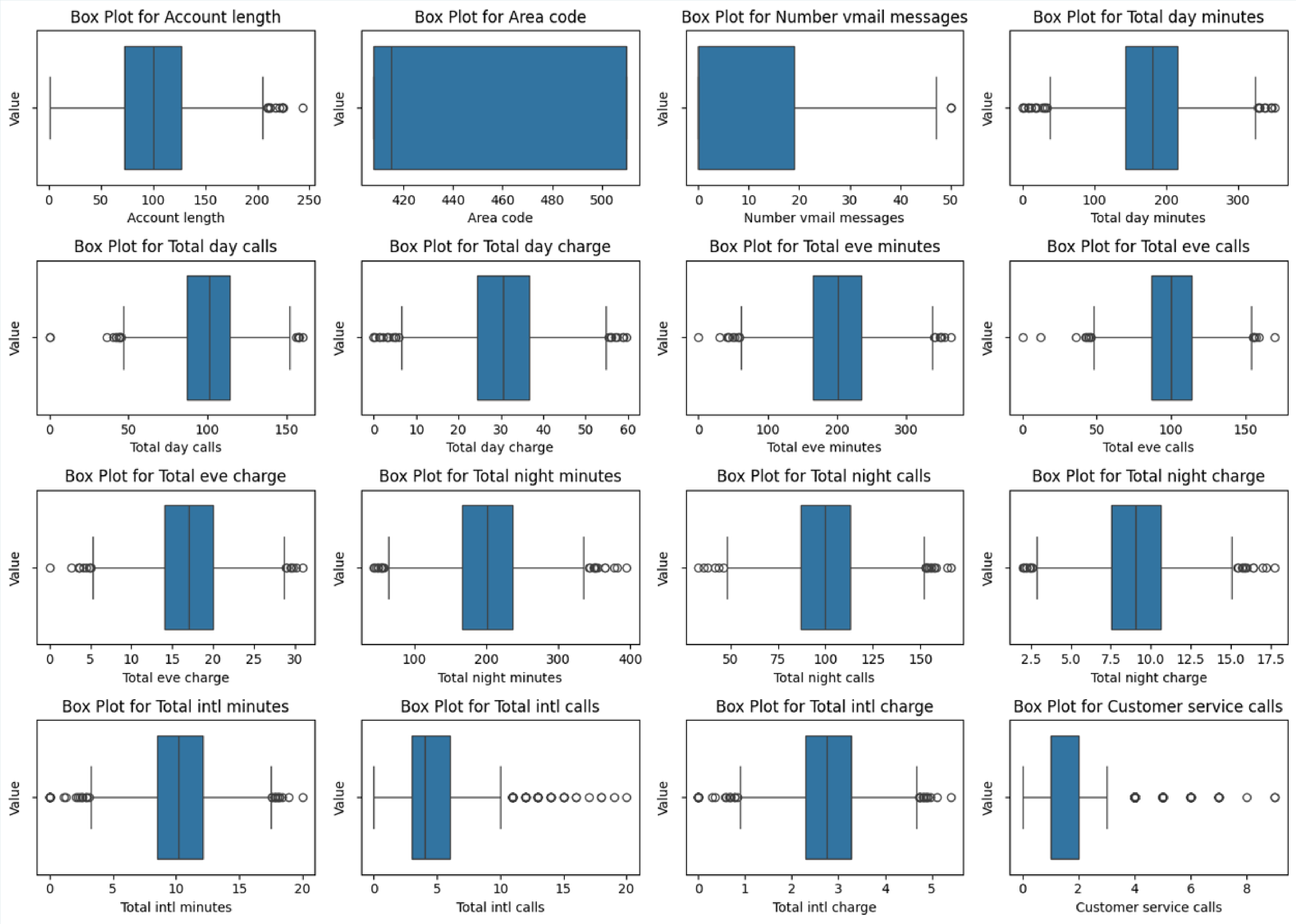


From the correlation matrix, we can see that:

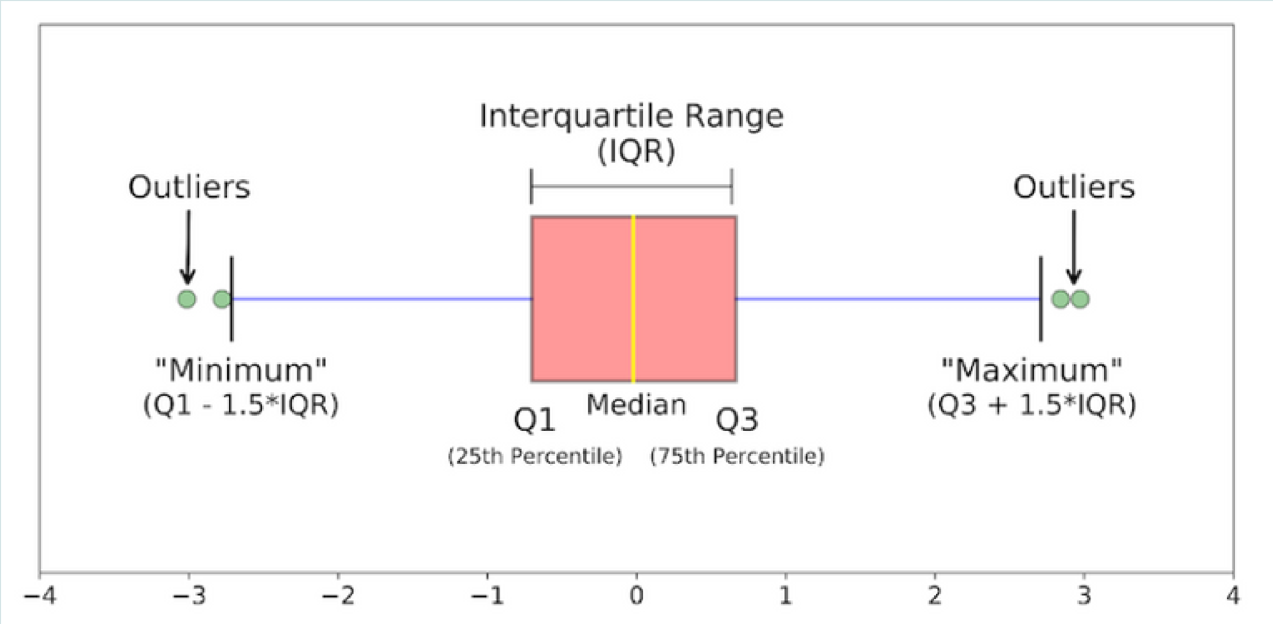
- Churn has no correlation with columns: Area code, Total eve calls
- Churn has highest correlation with columns: Total day charge, Total day minutes, Customer service calls, International plan_Yes

=> Churn is increasing with increase in Total day charge, Total day minutes, Customer service calls, International plan_Yes

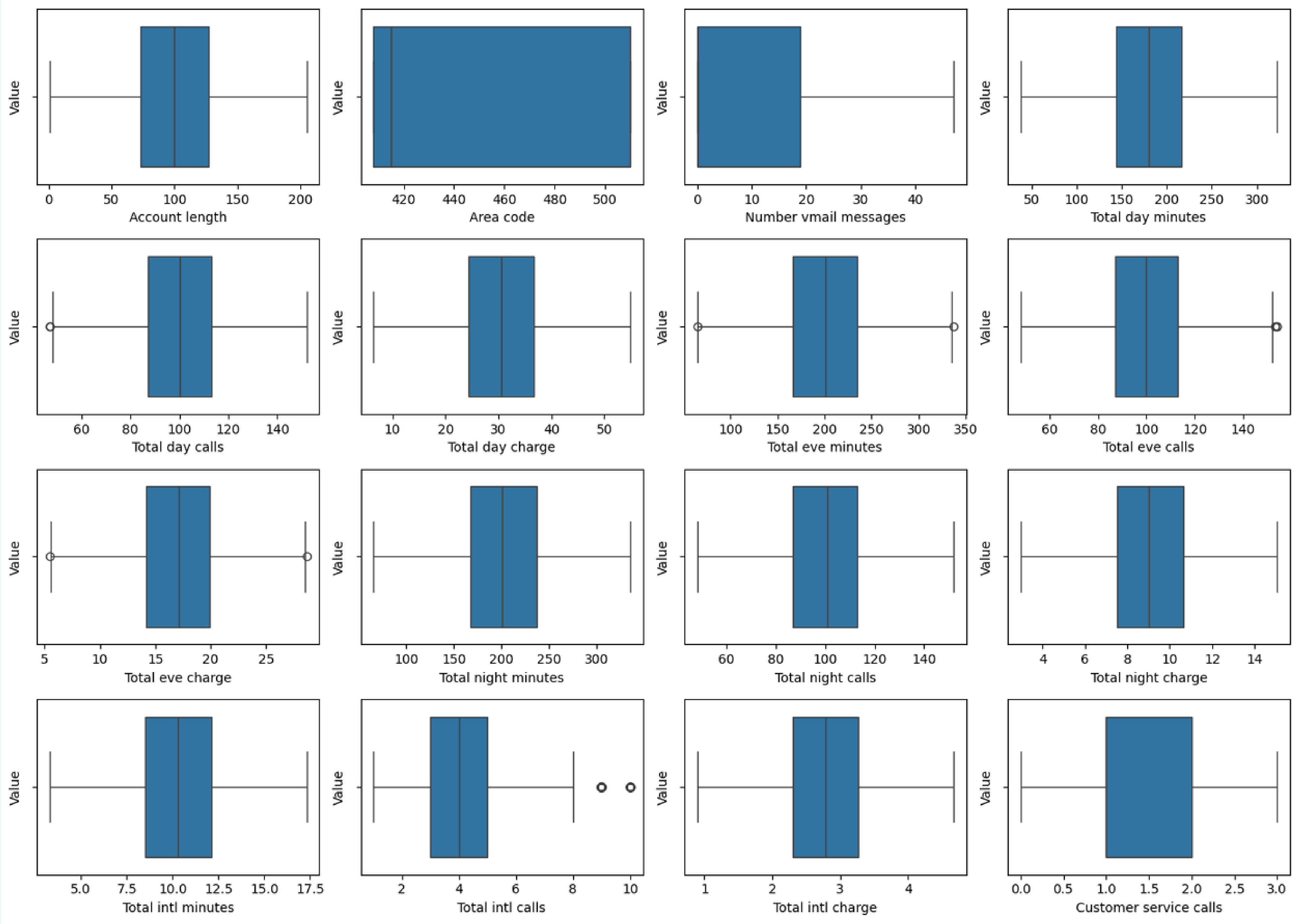
Exploratory Data Analysis



From the plot, we can observe that range of diversity is extreme. That is, even though more data points lie in the average range, there are few datapoints that are at the extreme corners which has an effect on the analysis. Hence, we are removing the outliers to get better results in the analysis.



Exploratory Data Analysis



Here is the plot after removing outliers by Interquartile range (IQR) method.

Model Building

		Precision	Recall	F1-score	Support	Accuracy
Logistic Regression	0	0.92	0.99	0.95	491	0.91
	1	0.68	0.27	0.39	55	
Random Forest	0	0.97	1.00	0.98	491	0.97
	1	1.00	0.73	0.84	55	

Logistic Regression

- **Precision** for class 0 (customers who didn't churn) is high at 0.92, indicating that among all the instances predicted as not churned, 92% were actually not churned. However, the precision for class 1 (customers who churned) is lower at 0.68, indicating that among all the instances predicted as churned, only 68% were actually churned.
- **Recall** for class 0 is high at 0.99, indicating that among all the actual instances of not churned customers, 99% were correctly predicted as not churned. However, the recall for class 1 is low at 0.27, indicating that among all the actual instances of churned customers, only 27% were correctly predicted as churned.
- **F1-score**, which is the harmonic mean of precision and recall, is relatively high for class 0 at 0.95 but low for class 1 at 0.39.
- The **accuracy** of the model is 0.91, indicating that it correctly predicts 91% of the instances in the test set.

Random Forest

- **Precision** for class 0 is very high at 0.97, indicating that among all the instances predicted as not churned, 97% were actually not churned. The precision for class 1 is perfect at 1.0, indicating that among all the instances predicted as churned, 100% were actually churned.
- **Recall** for class 0 is also very high at 1.0, indicating that among all the actual instances of not churned customers, 100% were correctly predicted as not churned. The recall for class 1 is lower at 0.73, indicating that among all the actual instances of churned customers, 73% were correctly predicted as churned.
- **F1-score** for class 0 is extremely high at 0.98 and for class 1 is 0.84.
- The **accuracy** of the model is 0.97, indicating that it correctly predicts 97% of the instances in the test set.

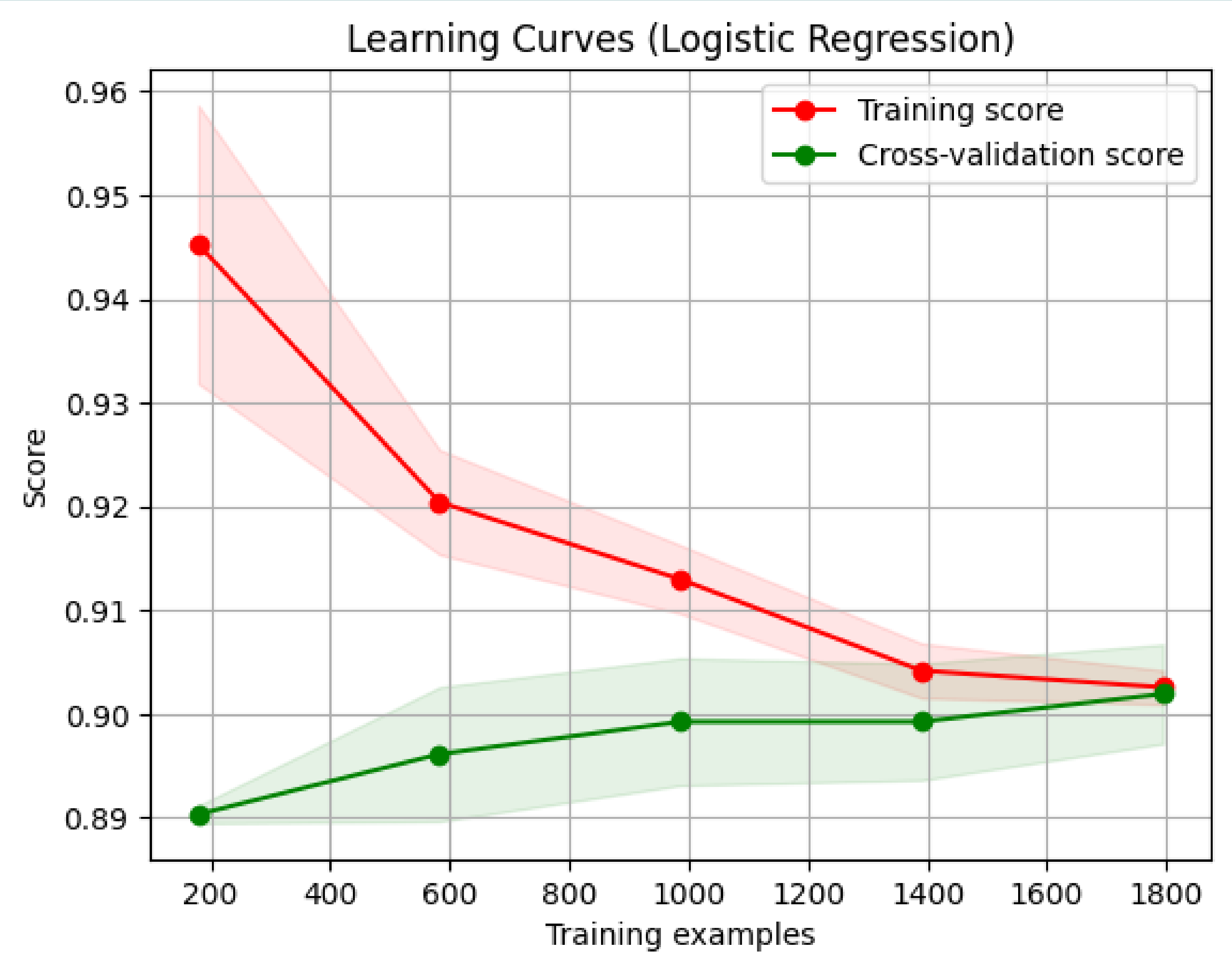
Evaluation & Conclusion

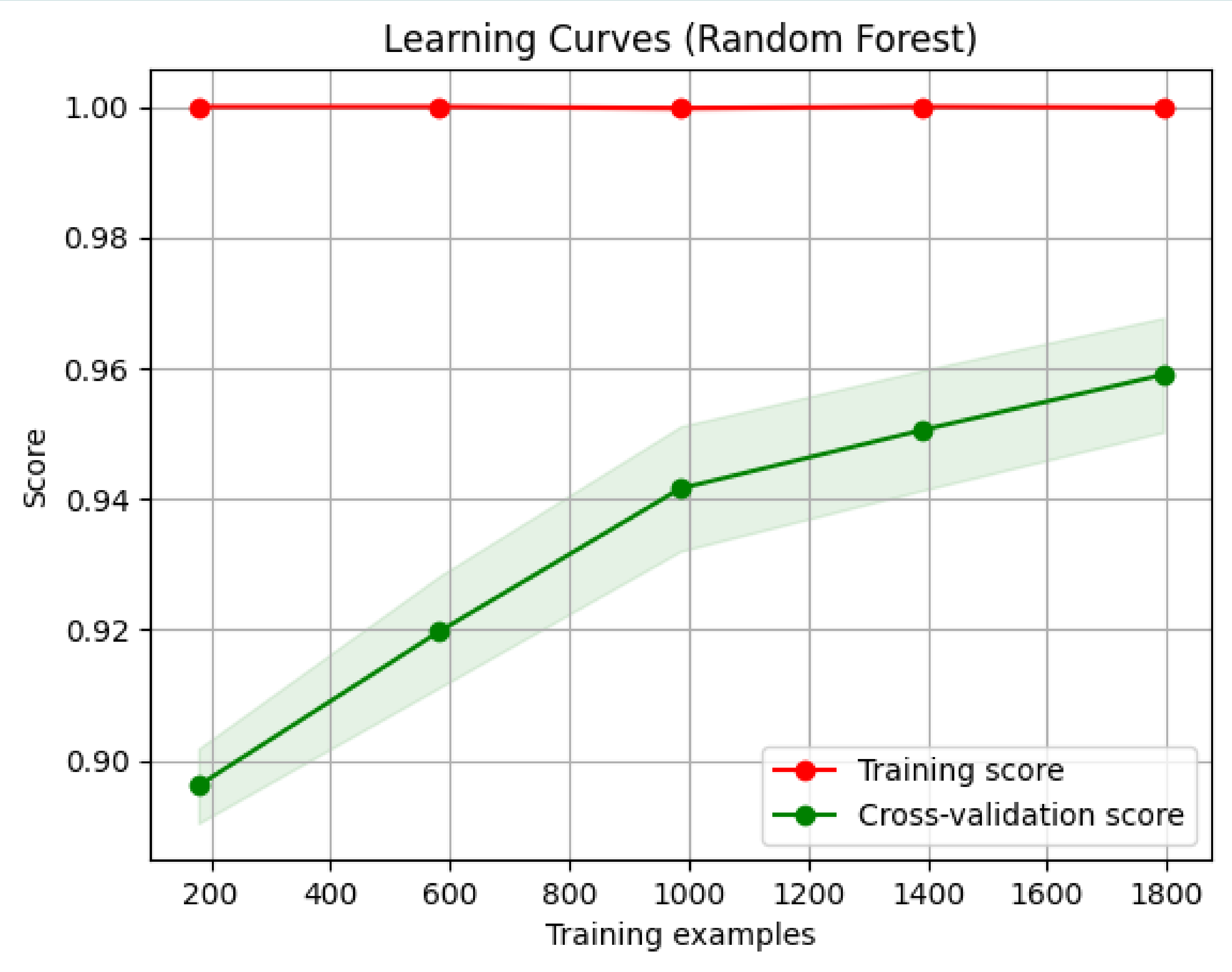
- Overall, the Random Forest model outperforms the Logistic Regression model in terms of precision, recall, and F1-score for both classes.

- The Random Forest model achieves a higher accuracy of 0.97 compared to 0.91 for the Logistic Regression model.

=> This indicates that the **Random Forest model is better** at predicting customer churn based on the provided features.

Evaluation & Conclusion





The number of training examples increases, the training score decreases while the cross-validation score increases, indicating a model that is learning effectively and improving its performance on unseen data.

While the training score remains high and constant, the cross-validation score improves with more training examples, highlighting the model's increasing generalization capability.

THANKS FOR ATTENTION