CUSTOMER CHURN

CUSTOMER CHURN-:

Customer churn is the rate at which customers terminate services

Gender
Age
Tenure
Usage Frequency
Support Calls
Payment Delay
Subscription Type
Contract Length
Total Spend
Last Interaction
Churn

Features

Churn

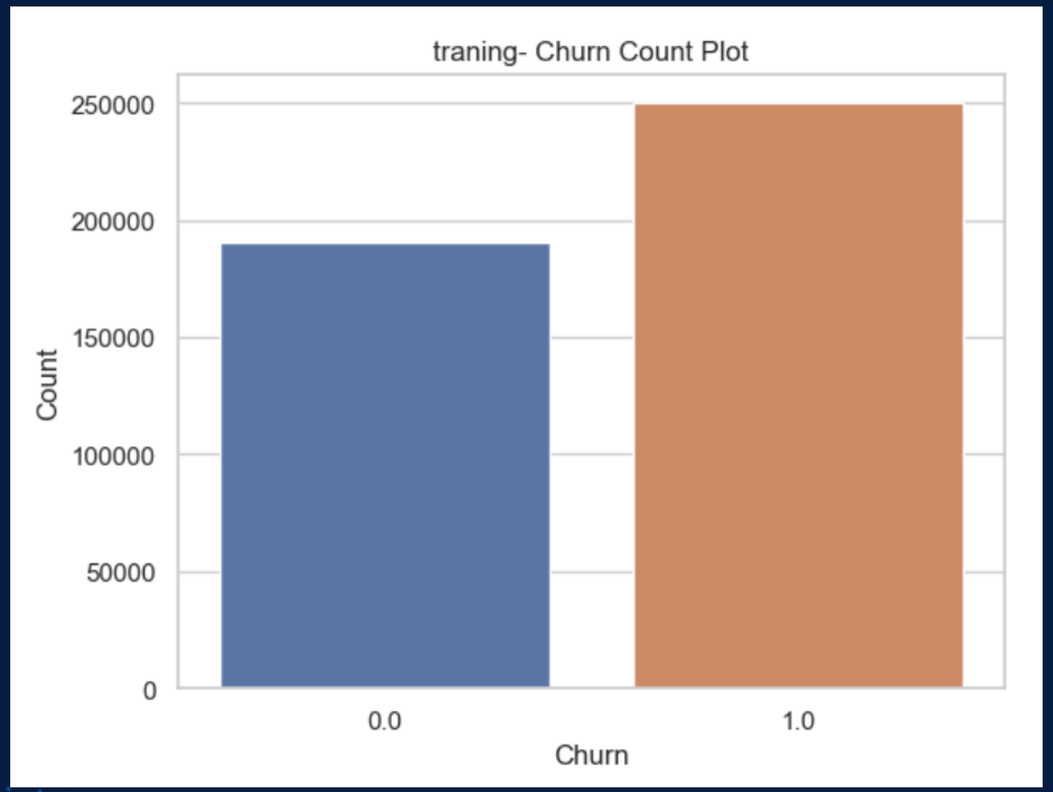
To Predict

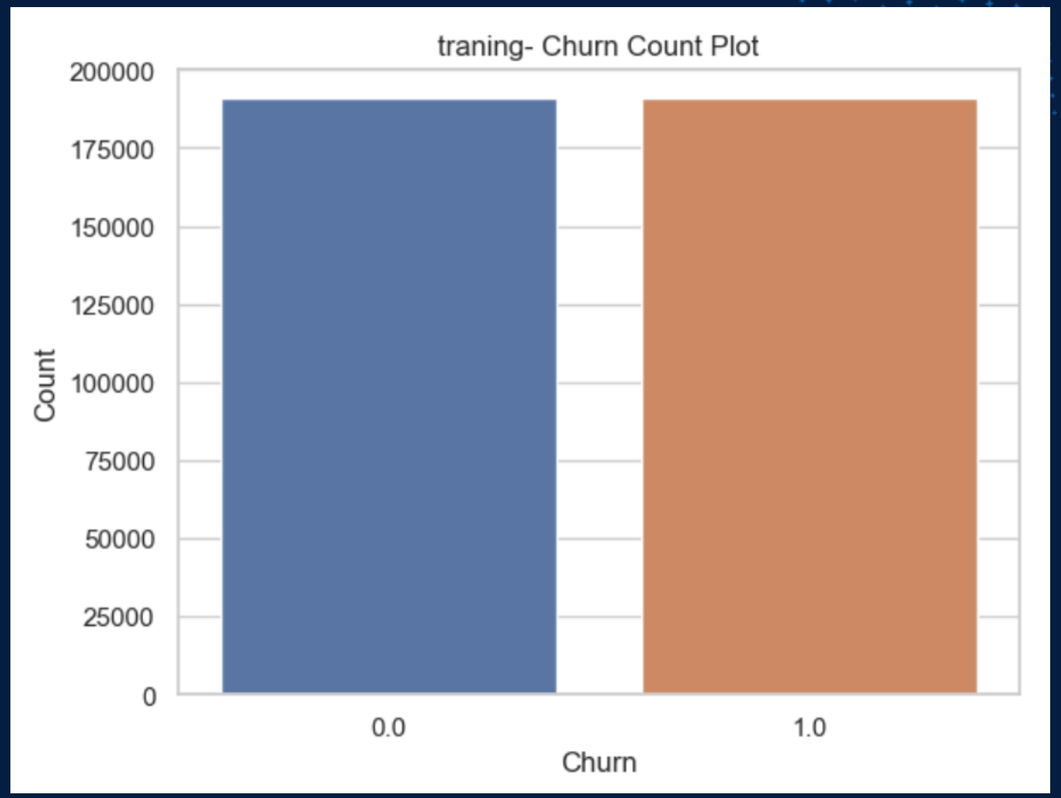
Data Overview

Age	Gender	Tenure	Usage_Frequency	Support_Calls	Payment_Delay	Subscription_Type	Contract_Length
35.0	Female	43.0	25.0	1.0	20.0	Basic	Annual
46.0	Female	60.0	19.0	4.0	13.0	Standard	Quarterly
36.0	Male	46.0	12.0	1.0	8.0	Standard	Annual
22.0	Female	12.0	25.0	2.0	7.0	Standard	Quarterly
58.0	Male	54.0	28.0	7.0	29.0	Basic	Annual
25.0	Female	2.0	15.0	8.0	8.0	Standard	Monthly
58.0	Male	33.0	25.0	6.0	9.0	Basic	Quarterly

Total_Spend	Last_Interaction	Churn
483.40	9.0	1.0
730.75	3.0	0.0
849.98	9.0	0.0
981.51	3.0	0.0
847.00	12.0	1.0
634.00	15.0	1.0
132.00	29.0	0.0

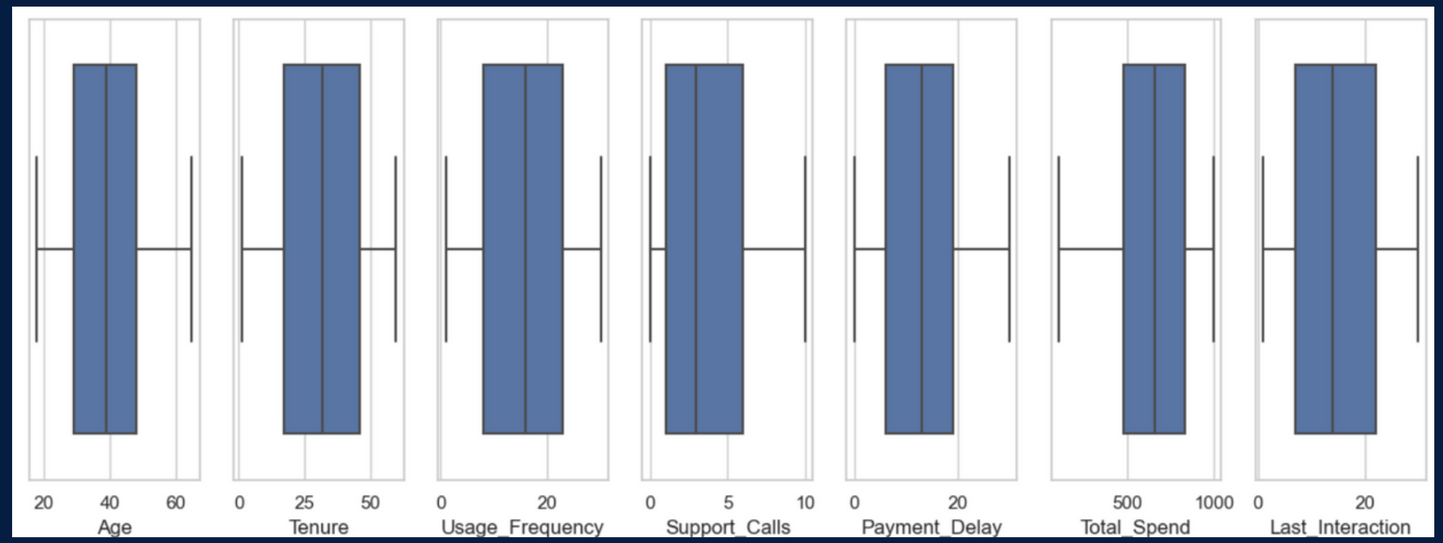
Down sampling





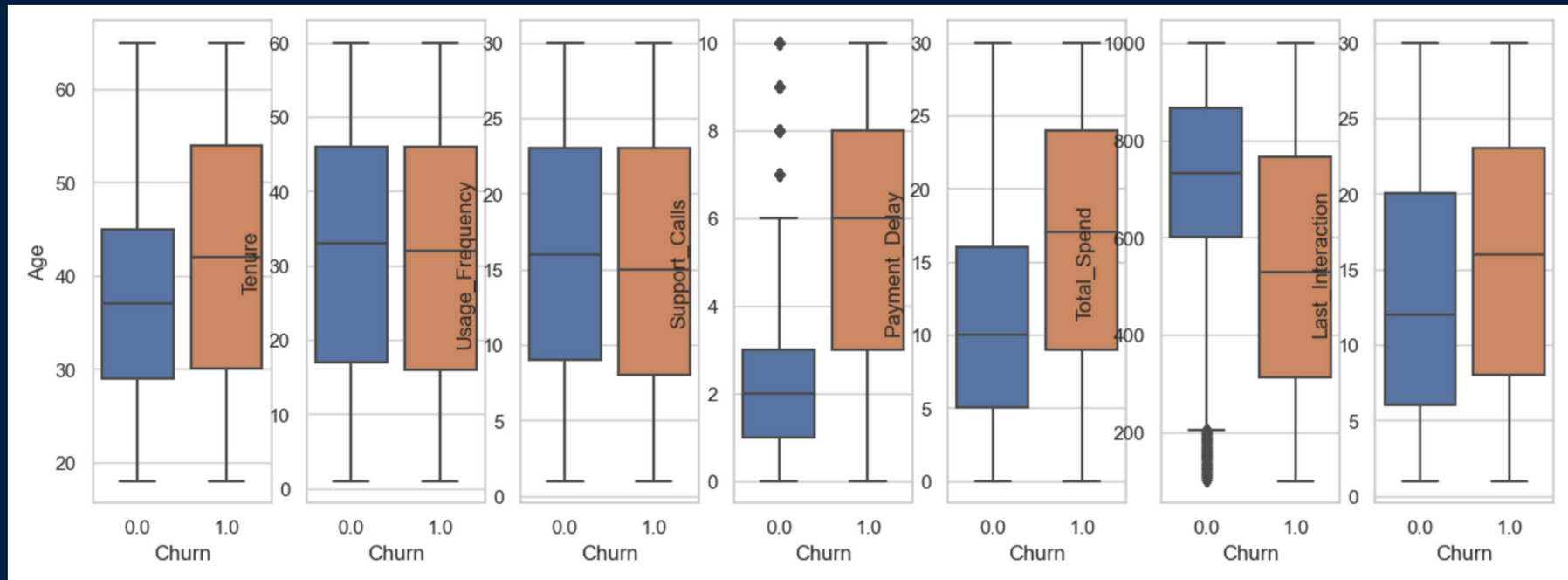
Down sampling by removing observations

Outlier Analysis

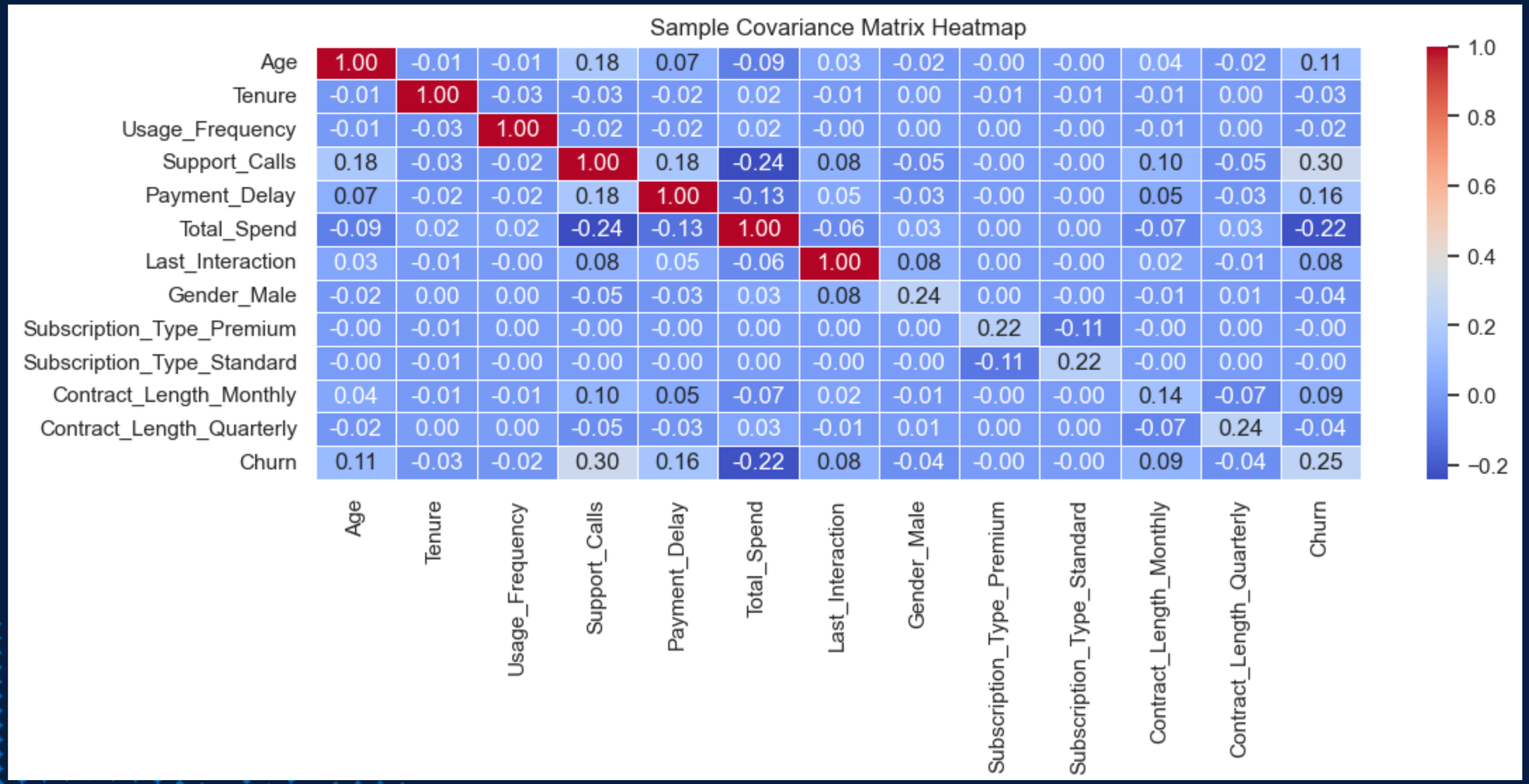


Boxplot -Feature

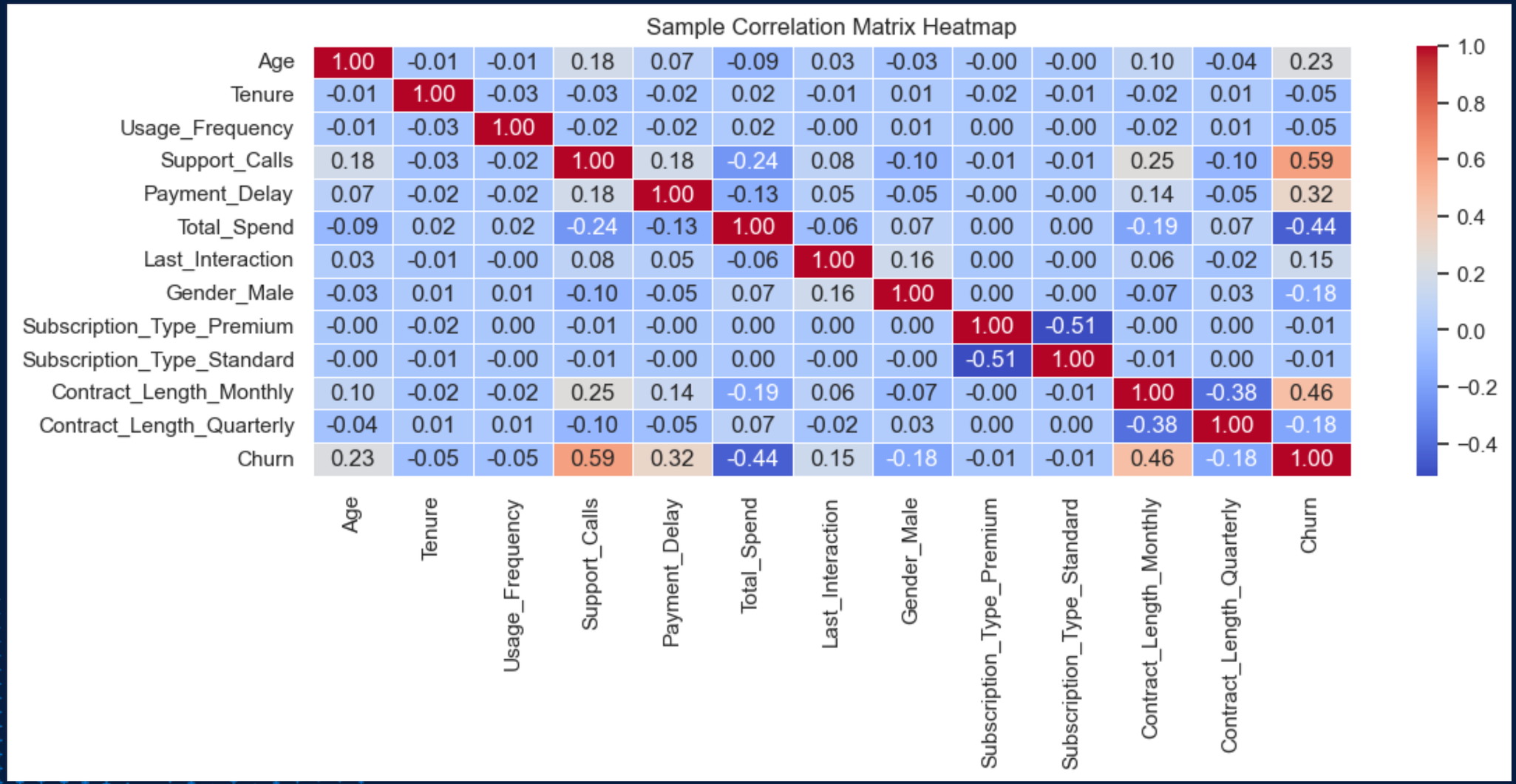
Boxplot -Feature VS Target



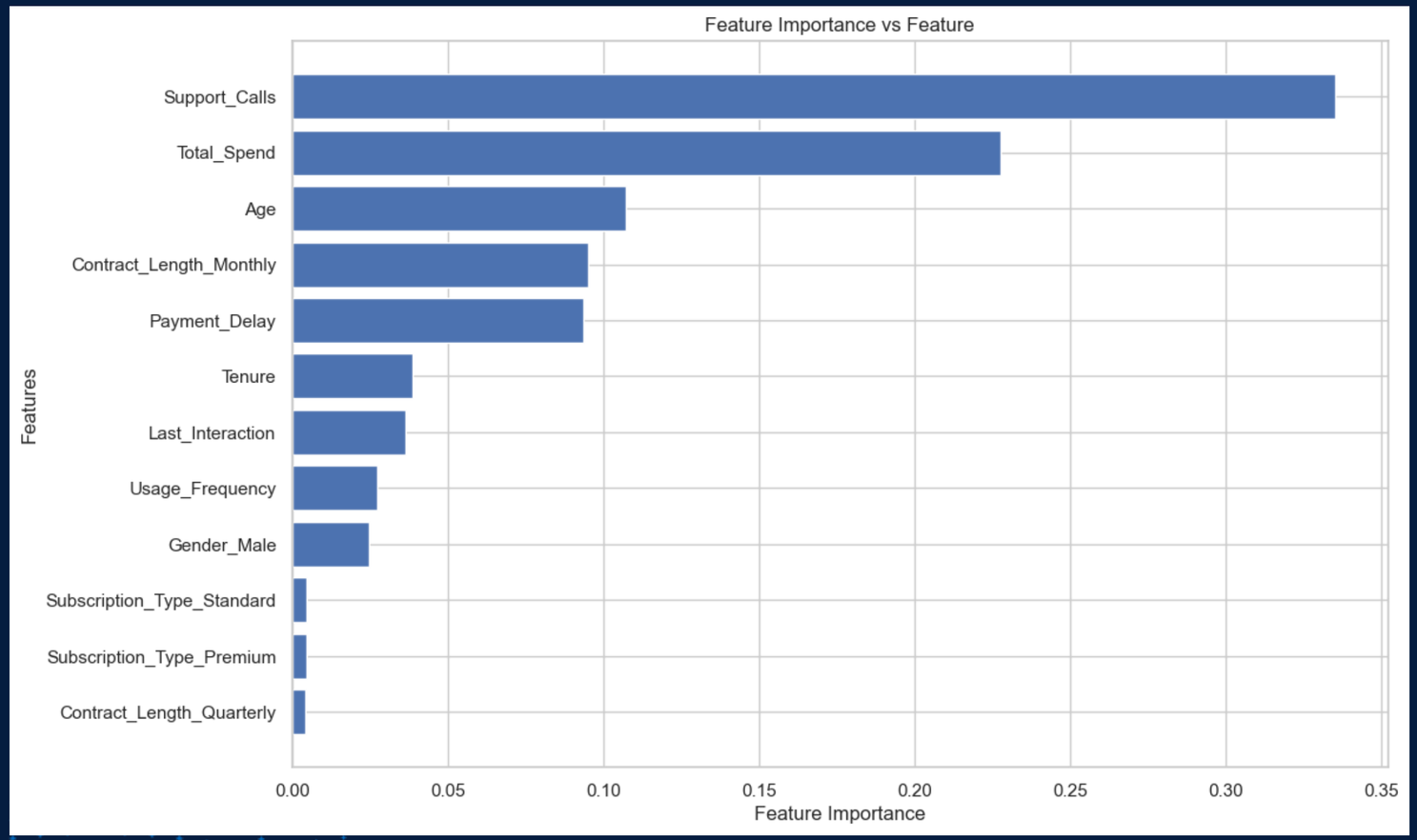
Covariance Matrix



Correlation Matrix



Random Forest Analysis

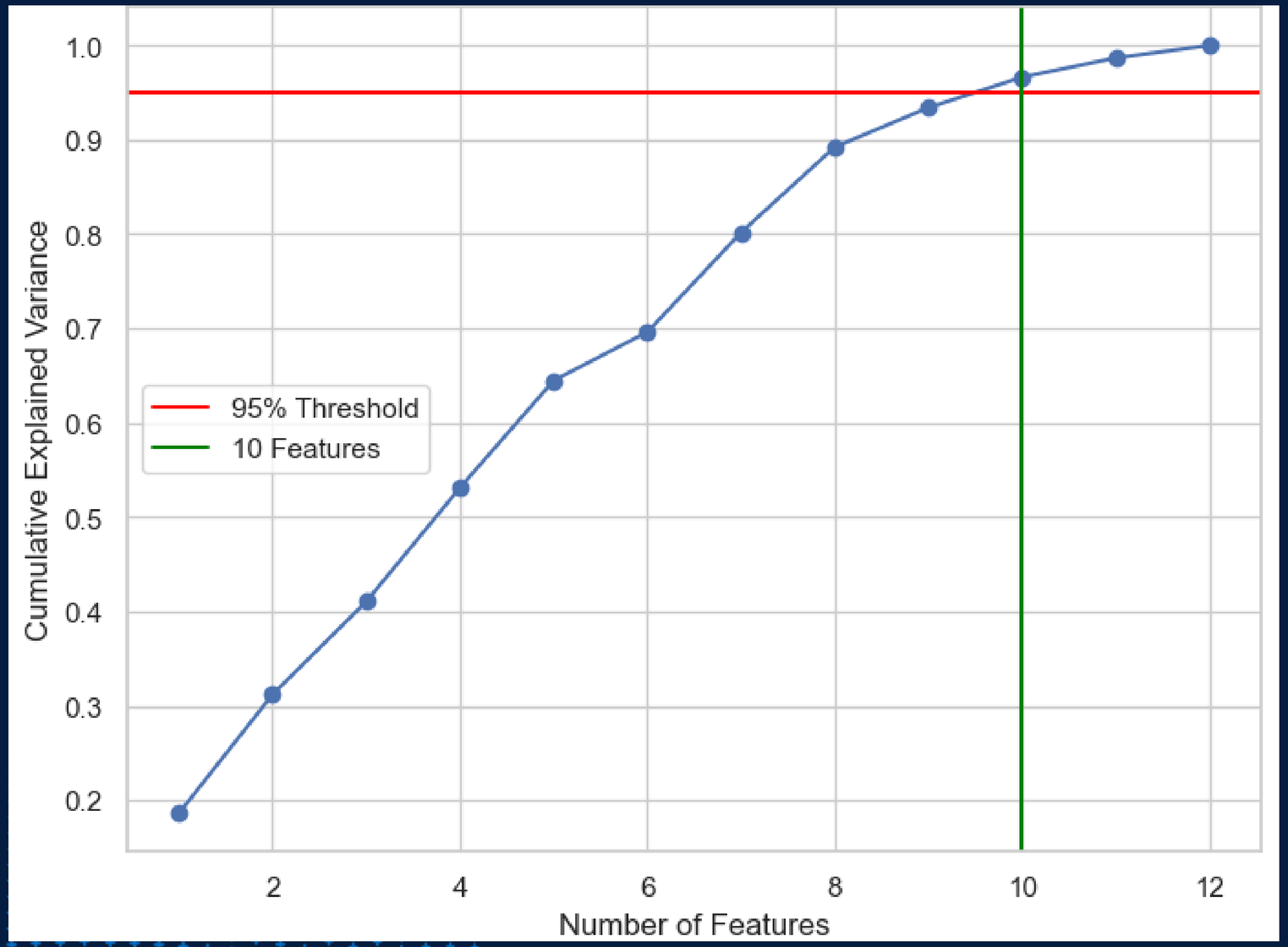


Threshold= 0.005

Contract_Length_Quarterly= 0.0044
Subscription_Type_Premium= 0.0046
Subscription_Type_Standard= 0.0046
Gender_Male= 0.0246
Usage_Frequency= 0.0275
Last_Interaction= 0.0364
Tenure= 0.0389
Payment_Delay= 0.0935
Contract_Length_Monthly= 0.095
Age= 0.1074
Total_Spend= 0.2275
Support_Calls= 0.3354

Eliminated Features
Contract_Length_Quarterly
Subscription_Type_Premium
Subscription_Type_Standard

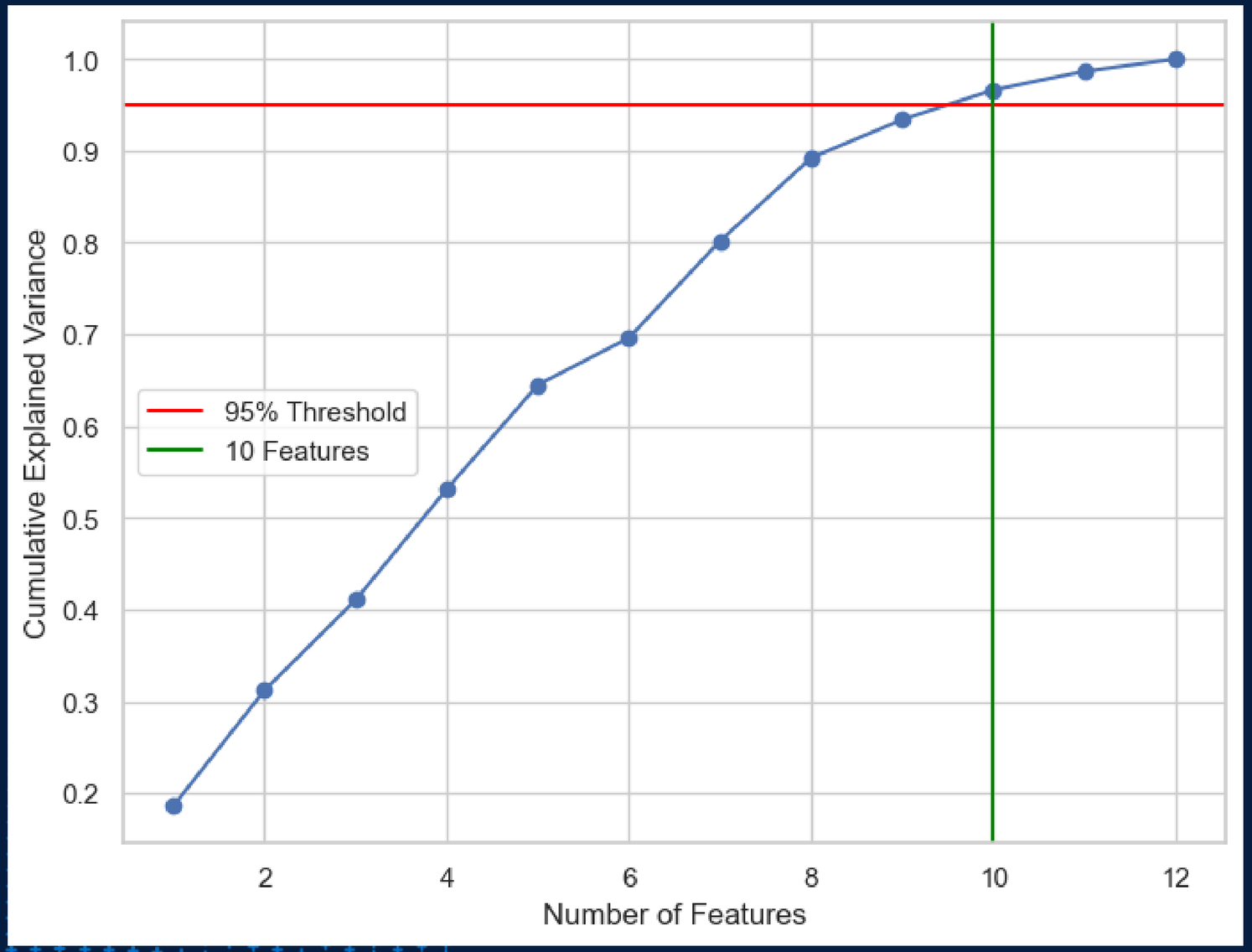
Principal Component Analysis



Threshold= 0.95

Original Data Conditional Number: 3.75
Transformed Data Conditional Number: 2.54
Number of Principal Components: 10
Threshold for Retained Variance: 0.95

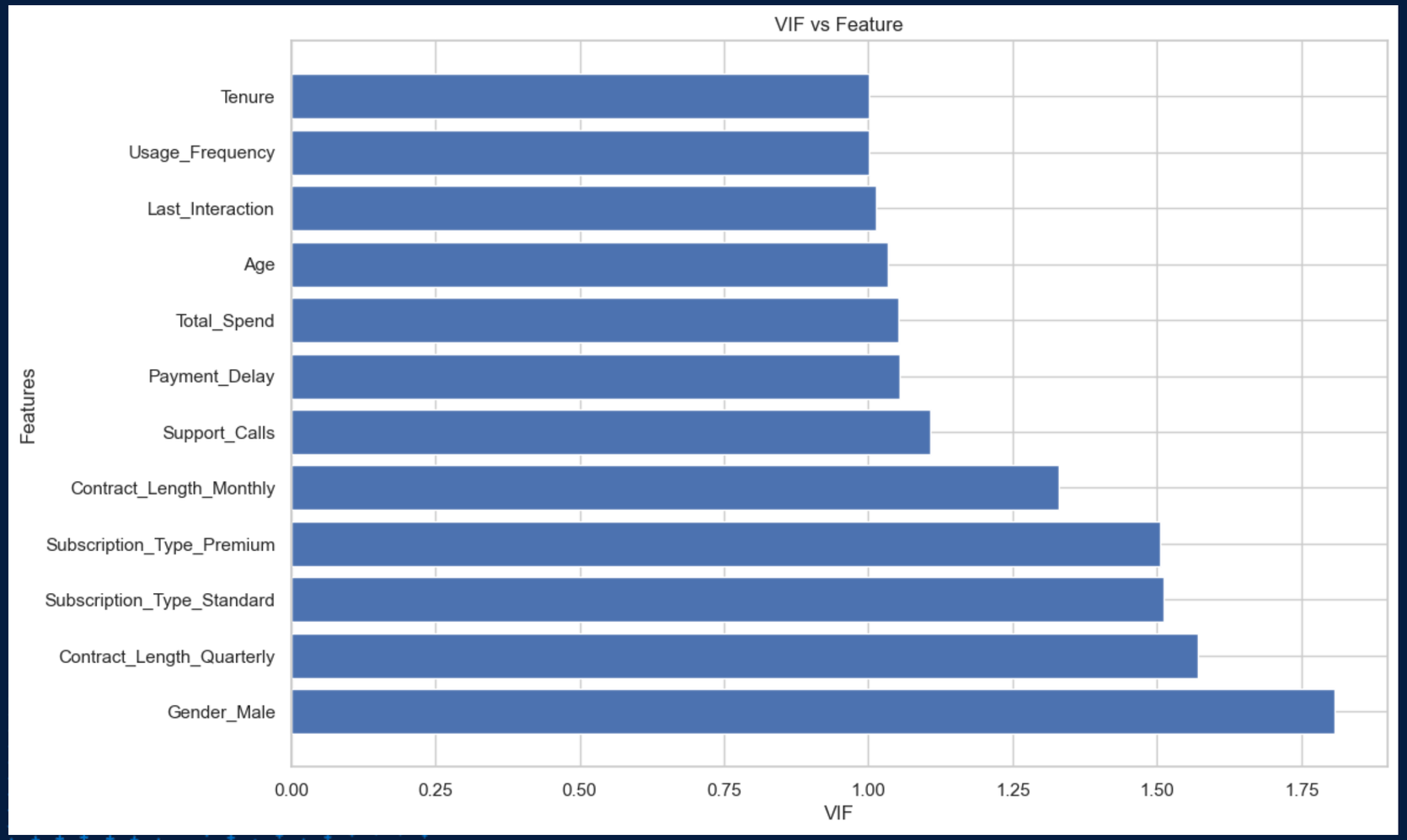
Singular Value Decomposition



Threshold= 0.95

Number of Components: 10
Singular Values: [274.975, 225.973, 225.148, 221.028, 217.179, 211.472, 207.416, 191.7, 129.983, 114.453]

VIF



VIF values for all features are close to 1, indicating low multicollinearity

Regression Analysis

T-Test Analysis

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.5141	0.010	-49.568	0.000	-0.534	-0.494
Age	0.0604	0.004	15.793	0.000	0.053	0.068
Tenure	0.0022	0.004	0.587	0.557	-0.005	0.010
Usage_Frequency	-0.0008	0.004	-0.203	0.839	-0.008	0.007
Payment_Delay	0.0165	0.004	4.130	0.000	0.009	0.024
Last_Interaction	0.0056	0.004	1.470	0.142	-0.002	0.013
Total_Spend	-0.0201	0.004	-4.977	0.000	-0.028	-0.012
Gender_Male	0.0075	0.008	0.972	0.331	-0.008	0.023
Subscription_Type_Premium	-0.0132	0.009	-1.431	0.152	-0.031	0.005
Subscription_Type_Standard	-0.0039	0.009	-0.428	0.669	-0.022	0.014
Contract_Length_Monthly	0.0531	0.011	4.823	0.000	0.032	0.075
Contract_Length_Quarterly	-0.0093	0.008	-1.113	0.266	-0.026	0.007
Churn	1.0179	0.009	108.479	0.000	0.999	1.036

-Age has a significant positive relationship with support calls, and the coefficient is significantly different from zero.
-Tenure is not a statistically significant predictor of support calls, as the p-value is greater than the significance level
-Usage frequency is not a statistically significant predictor of support calls, as the p-value is greater than the significance level.
-Payment delay, Contract length (Monthly),Churn has a significant positive relationship with support calls.
-Total spend has a significant negative relationship with support calls
-Gender (Male), Subscription type (Premium),Subscription type (Standard), Contract length (Quarterly) is not a statistically significant predictor of support calls.

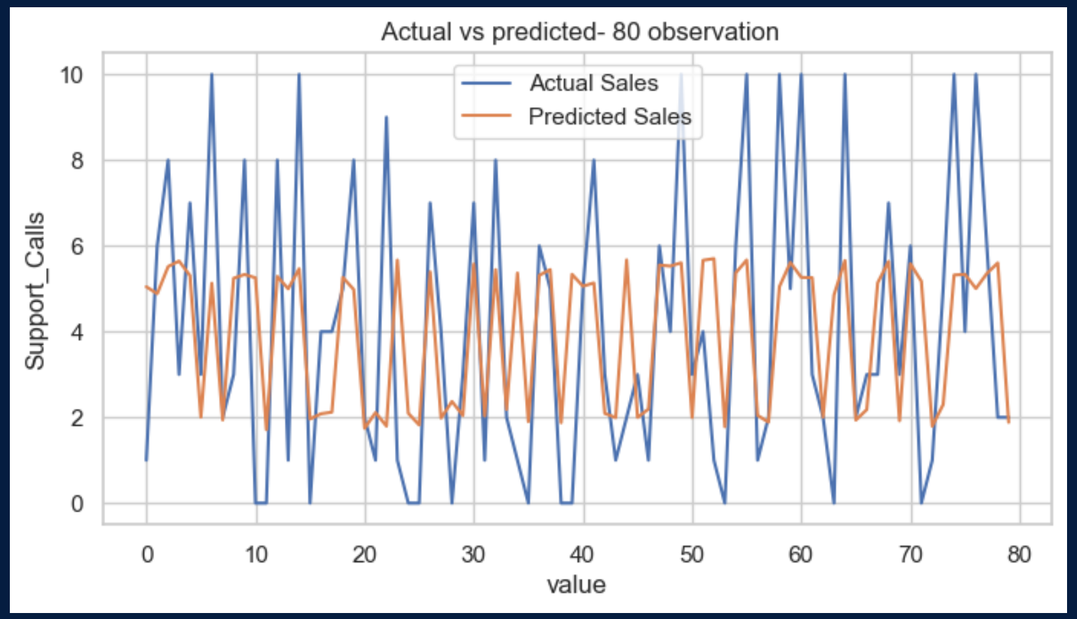
F-Test Analysis

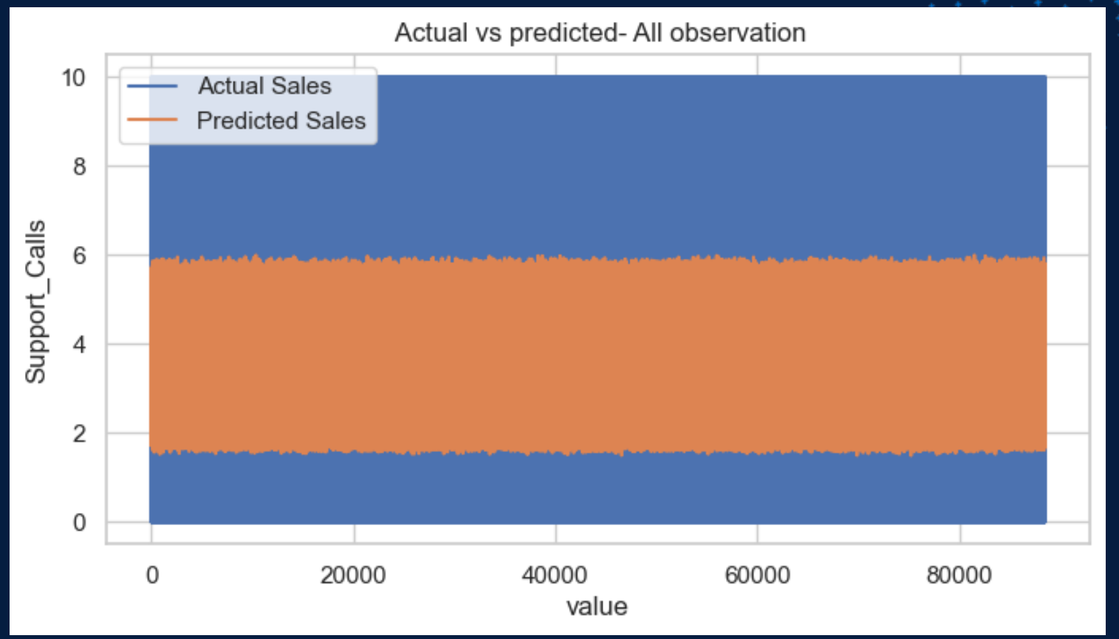
R-squared:	0.300
Adj. R-squared:	0.300
F-statistic:	1789.
Prob (F-statistic):	0.00
Log-Likelihood:	-62012.
AIC:	1.240e+05
BIC:	1.242e+05

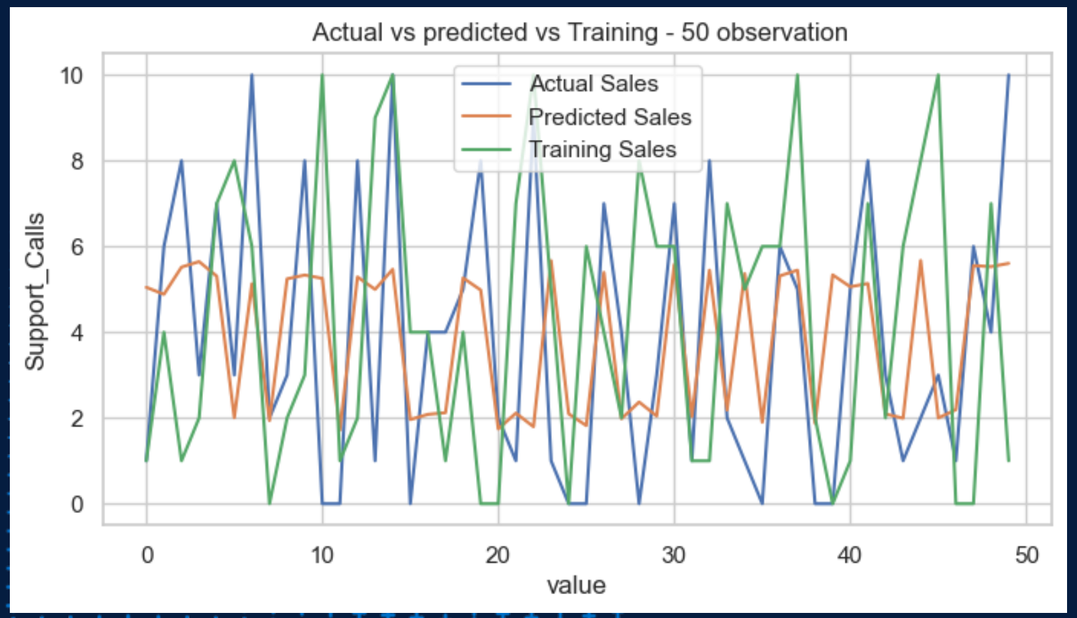
F-Statistic: 1789.0
Critical F-Value:1.75

- A higher F-statistic suggests a better fit of the model to the data.
-Coefficients in the model are non zero, No Null Hypothesis

Linear Regression

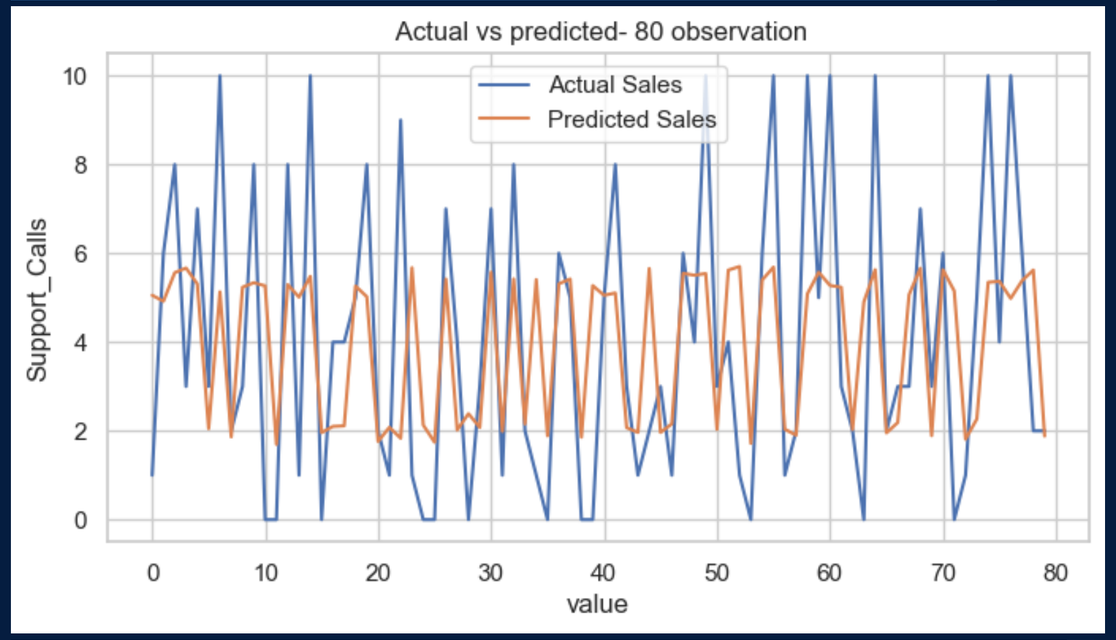


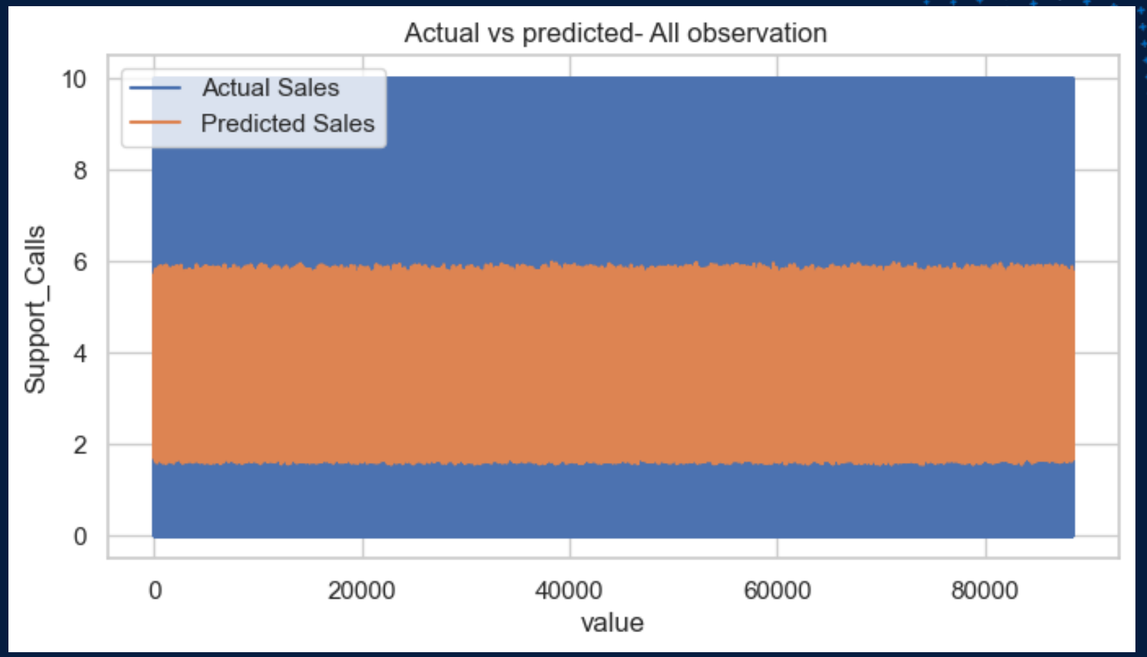


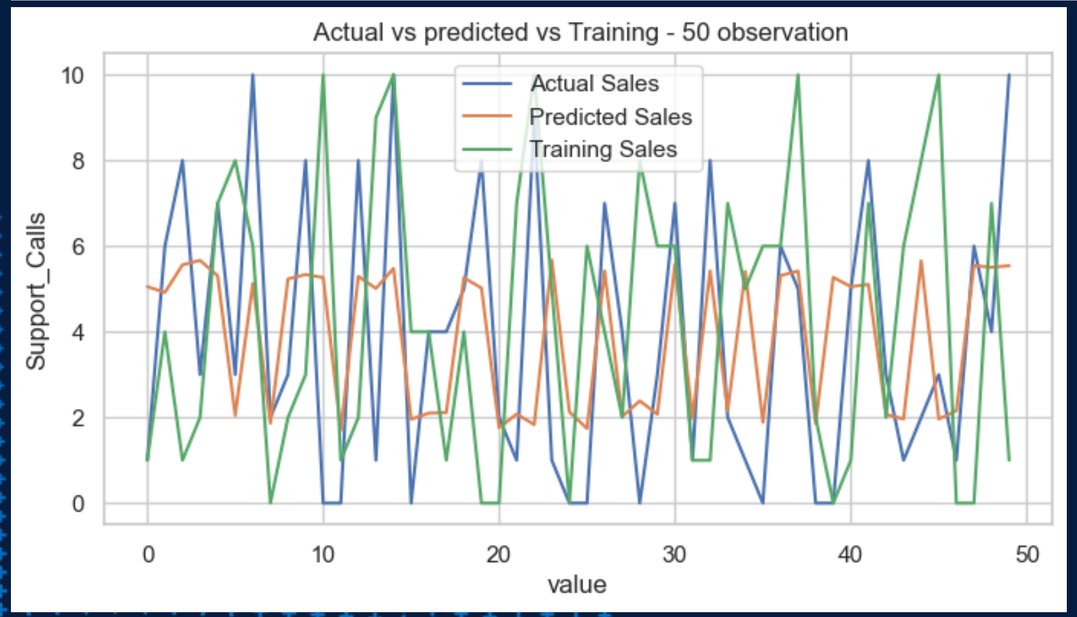


Model Equation:- Support_Calls = -0.514 + 0.060 Age + 0.002 Tenure -0.001 Usage_Frequency + 0.017 Payment_Delay + 0.006 Last_Interaction -0.020 Total_Spend + 0.008 Gender_Male -0.013 Subscription_Type_Premium -0.004 Subscription_Type_Standard + 0.053 Contract_Length_Monthly -0.009 Contract_Length_Quarterly + 1.018 Churn

OLS

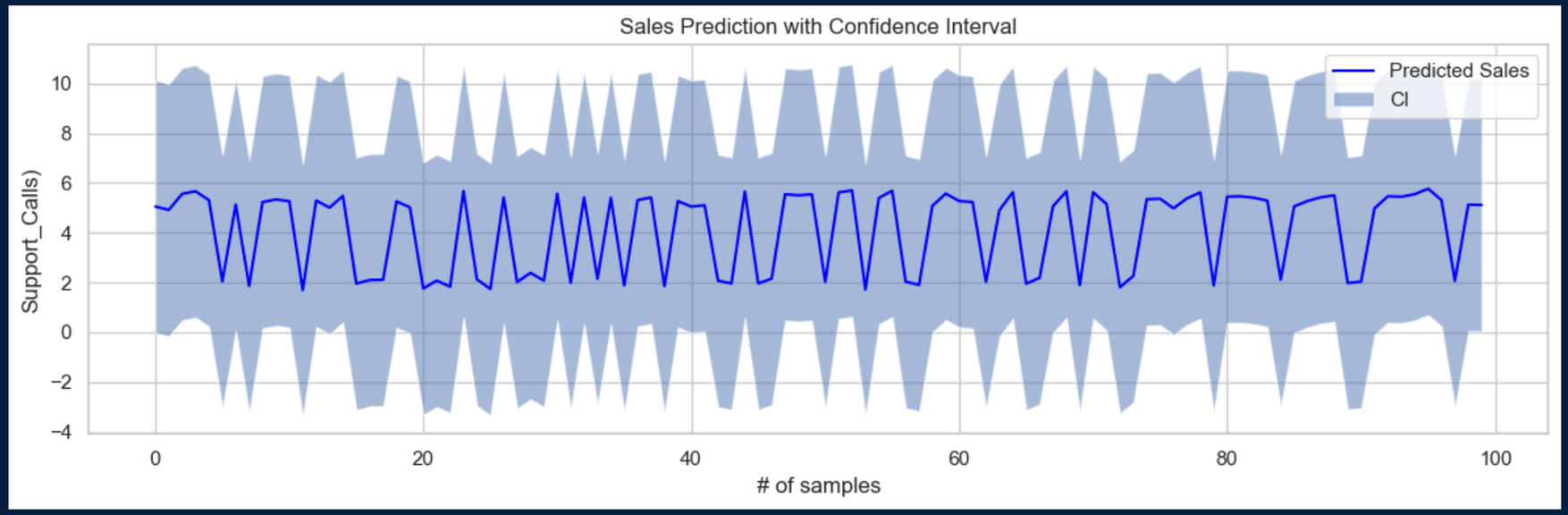


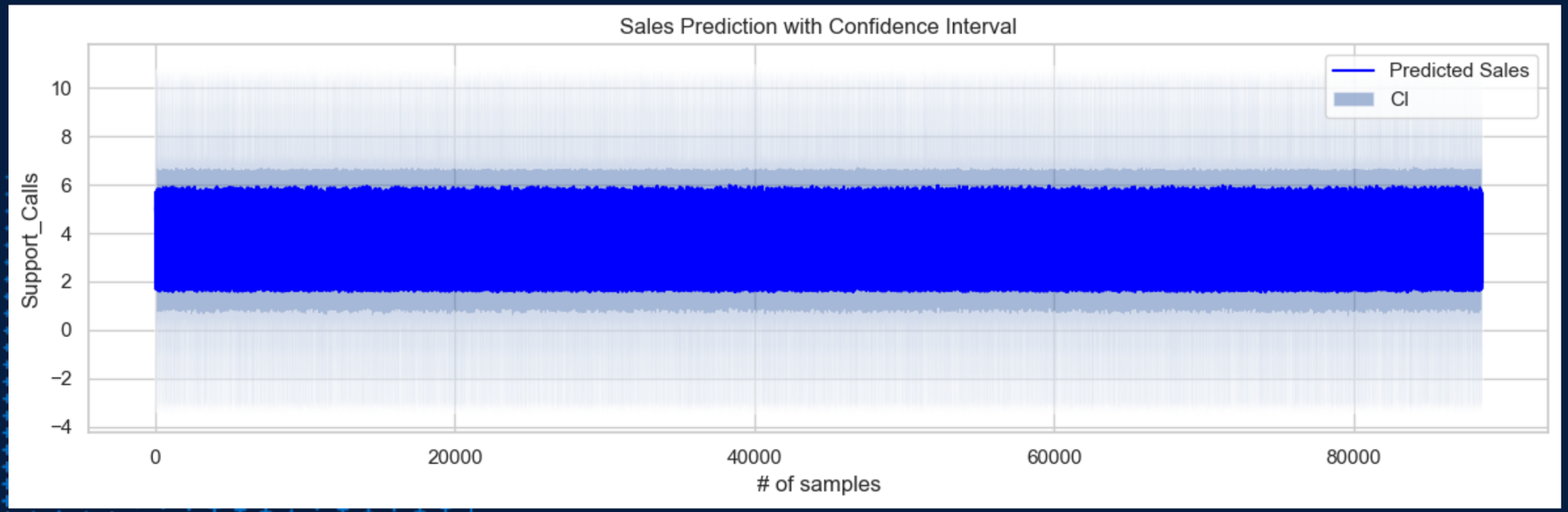




Model Equation:- Support_Calls= -0.52 + 0.06 Age + 0.017 Payment_Delay - 0.02 Total_Spend + 0.058 Contract_Length_Monthly + 1.018 Churn

OLS


Sales Prediction with Confidence Interval
Predicted Sales
CI
Support_Calls)
10
8
6
4
2
0
−2
−4
0
20
40
60
80
100
# of samples


Sales Prediction with Confidence Interval
Predicted Sales
CI
Support_Calls
10
8
6
4
2
0
−2
−4
0
20000
40000
60000
80000
# of samples

Linear Regression vs OLS

Metric	OLS Model	Linear Regression Model
R-squared	0.300	0.300
Adjusted R-squared	0.300	0.300
AIC	124043.354	95482.038
BIC	124096.273	95587.875
MSE	6.747	6.747

-Same R-squared and Adjusted R-squared
-The Linear Regression Model has a lower AIC value (95482.038) compared to the OLS Model (124043.354).
-The Linear Regression Model has a lower BIC value (95587.875) compared to the OLS Model (124096.273).
-Same MSE

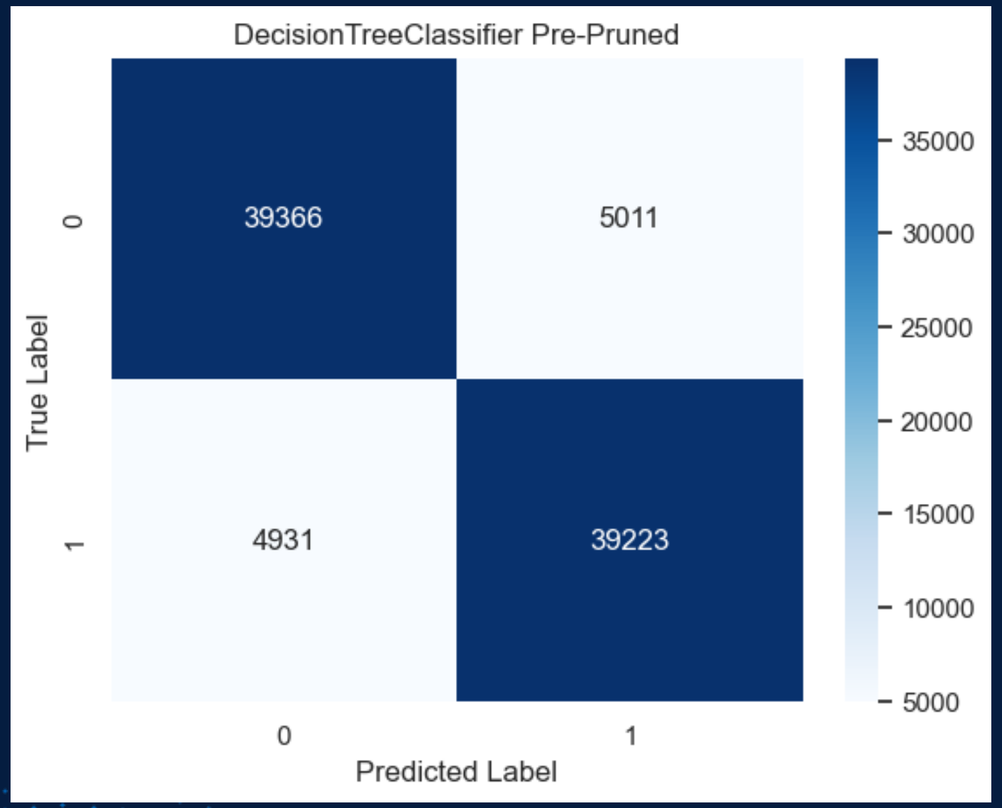
Classification Analysis

DecisionTree Classifier

feature	Feature Importances
Subscription_Type_Standard	0.00436
Contract_Length_Quarterly	0.0047
Subscription_Type_Premium	0.00547
Usage_Frequency	0.02544
Gender_Male	0.02955
Last_Interaction	0.03163
Tenure	0.03966
Payment_Delay	0.09388
Contract_Length_Monthly	0.10105
Age	0.10417
Total_Spend	0.22538
Support_Calls	0.33471

Feature removed:- ['Contract_Length_Quarterly', 'Subscription_Type_Premium', 'Subscription_Type_Standard']

DecisionTree Pre-Pruned

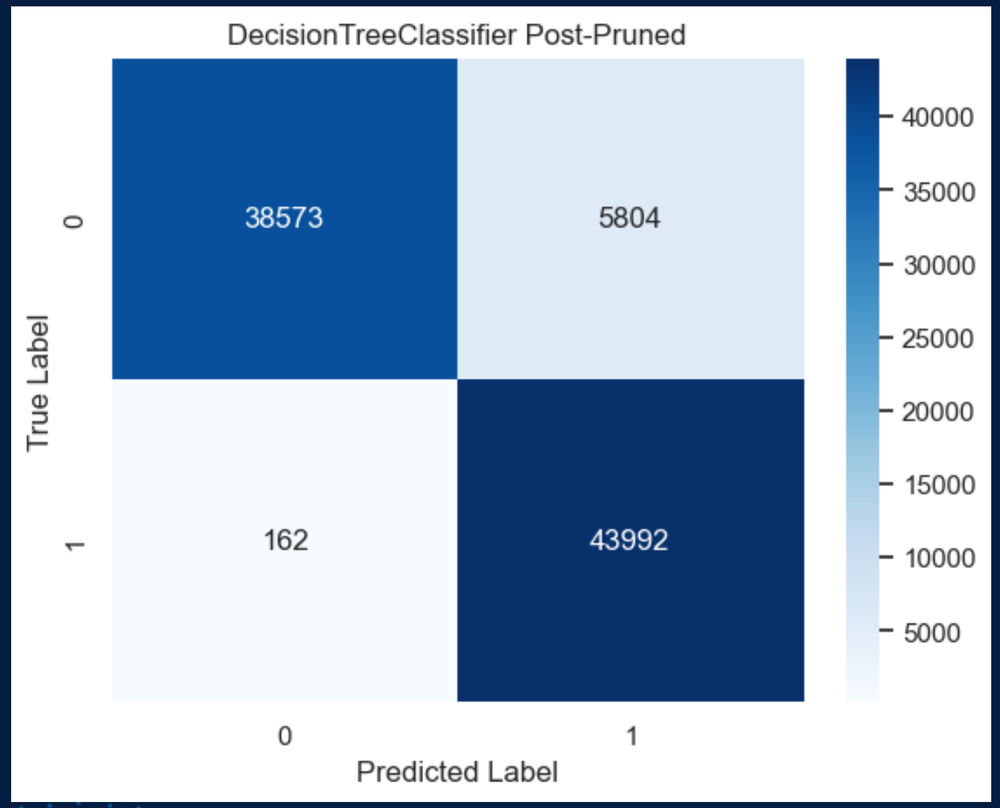


Best parameters found:

'criterion': 'entropy'
'max_depth': 20
'max_features': 'sqrt'
'min_samples_leaf': 3
'min_samples_split': 10
'splitter': 'best'

	Accuracy	confusion Matrix	recall	AUC	Specificity	F-score
Pre-Pruned	0.89	[[39366 5011] [4931 39223]]	0.89	0.93	0.887	0.888

DecisionTree Post-Pruned

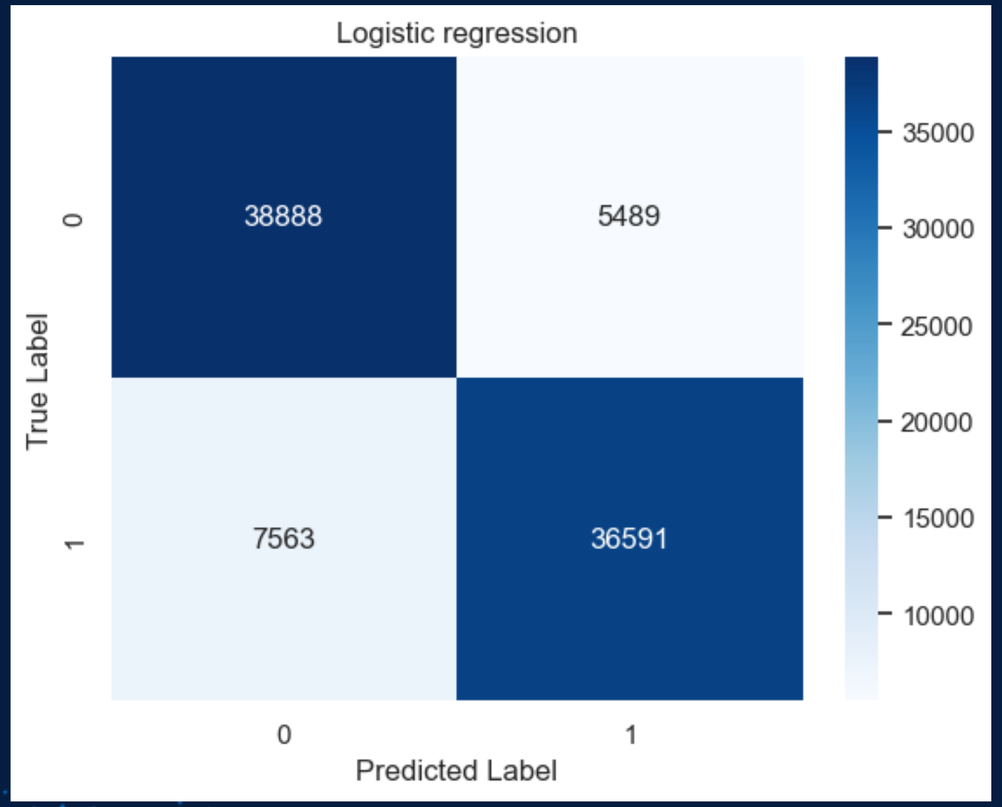


Best parameters found:

ccp_alpha: 0.00007
criterion: gini
min_impurity_decrease: 0.0
min_samples_leaf: 1
min_samples_split: 2
min_weight_fraction_leaf: 0.0
random_state: 5805
splitter: best

	Accuracy	confusion Matrix	recall	AUC	Specificity	F-score
Post-Pruned	0.93	[[38573 5804] [162 43992]]	1.0	0.96	0.869	0.936

Logistic regression

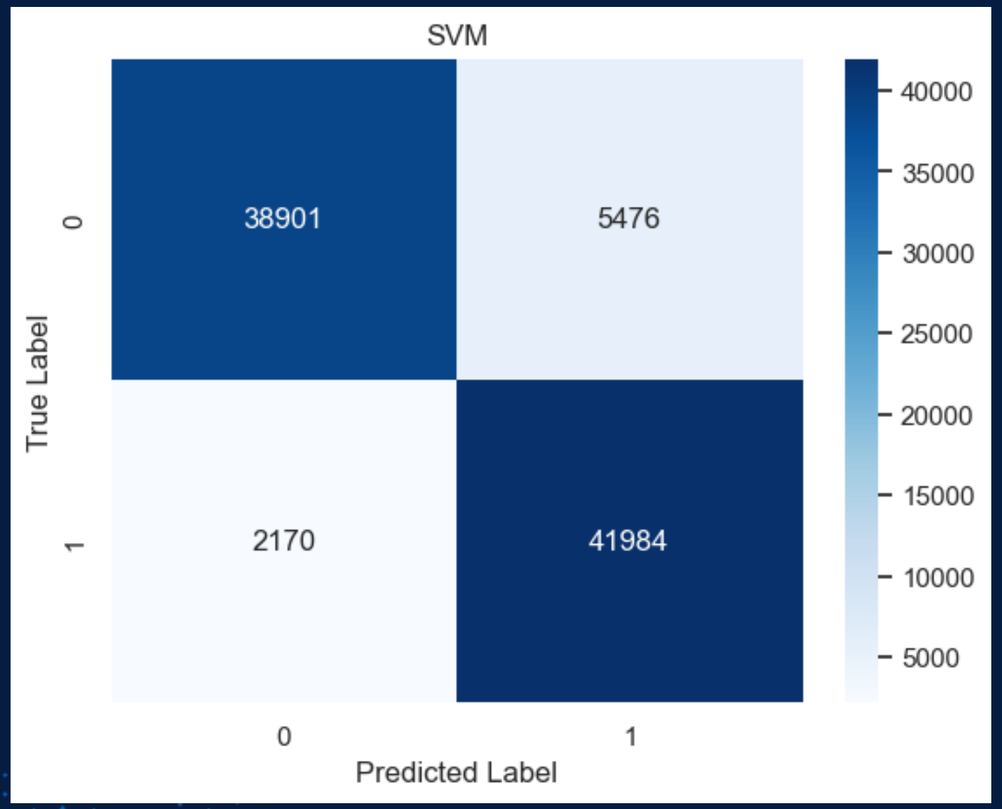


Best parameters found:

'C': 1.0,
'penalty': 'l2',
'solver': 'liblinear'

	Accuracy	confusion Matrix	recall	AUC	Specificity	F-Score
logistic regression	0.85	[[38888 5489] [7563 36591]]	0.83	0.91	0.88	0.85

Support Vector Machine

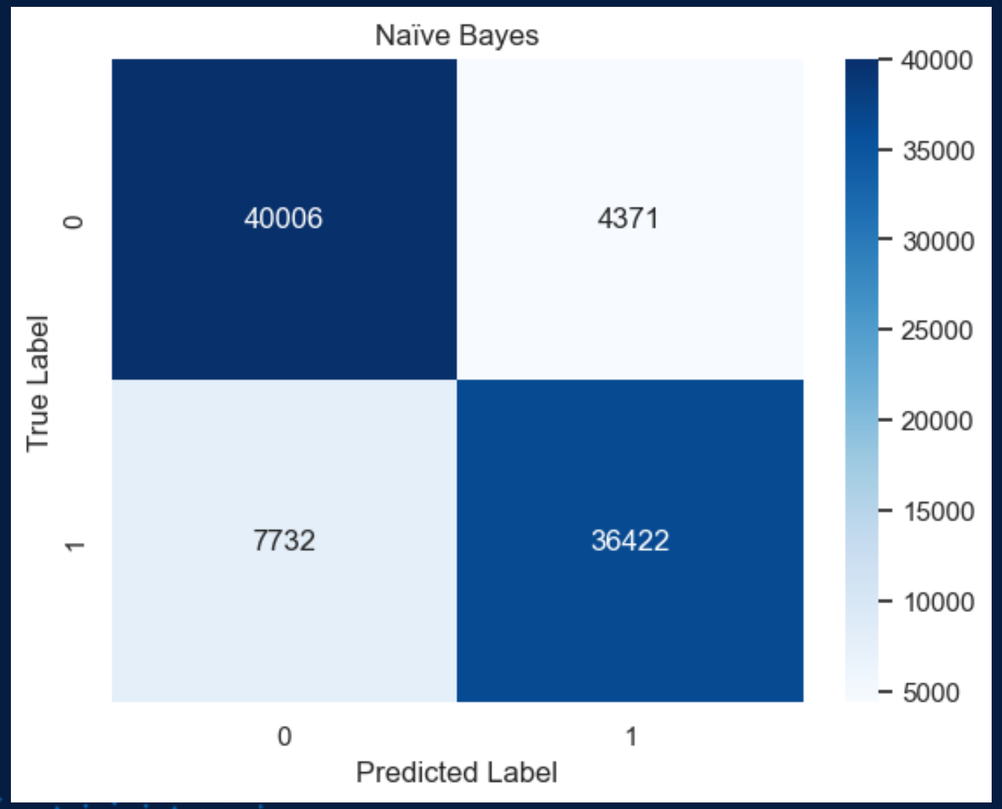


Best parameters found:

'C': 10
'gamma': 'scale',
'kernel': 'rbf'

	Accuracy	confusion Matrix	recall	AUC	Specificity	F-Score
SVM	0.91	[[38901 5476] [2170 41984]]	0.95	0.94	0.877	0.917

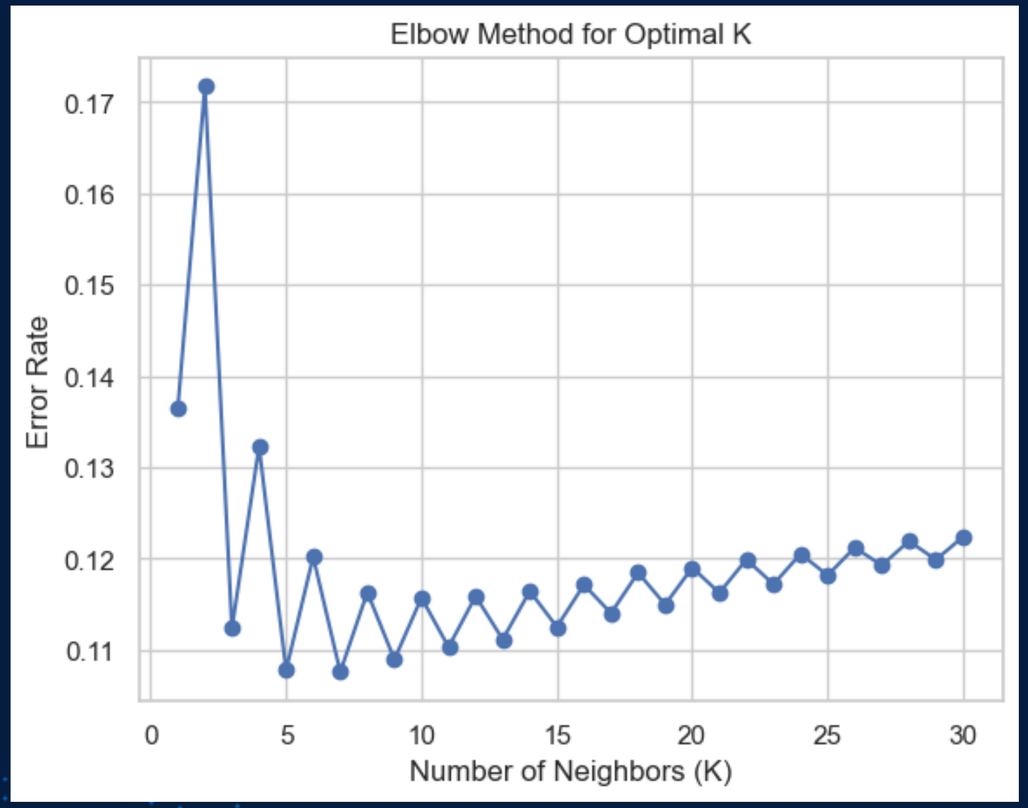
Naïve Bayes



-Using GaussianNB
-Stratified K-fold(5) Cross-Validation Accuracy: 86.40%

	Accuracy	confusion Matrix	recall	AUC	Specificity	F-Score
Naïve Bayes	0.86	[[40006 4371] [7732 36422]]	0.82	0.93	0.902	0.858

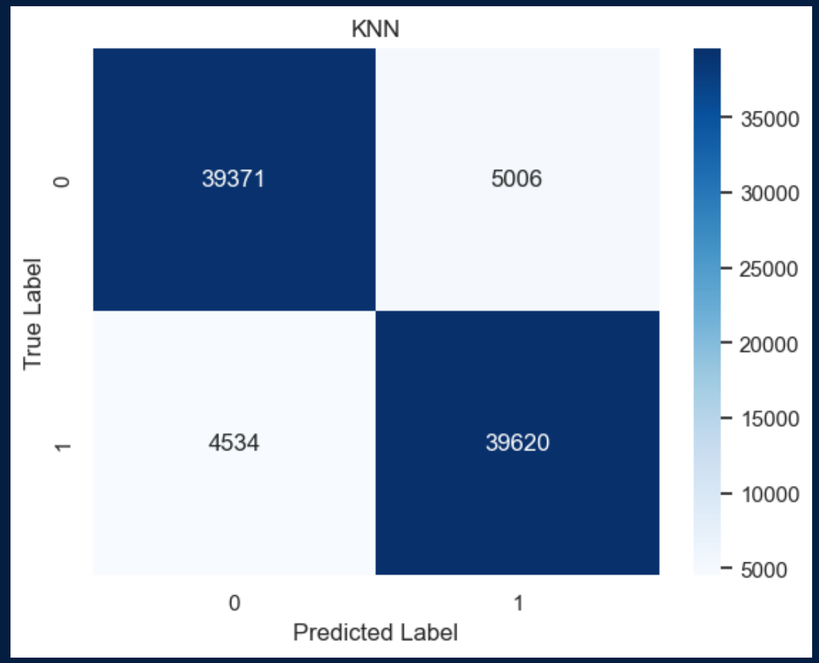
K-Nearest Neighbors



Best parameters found:

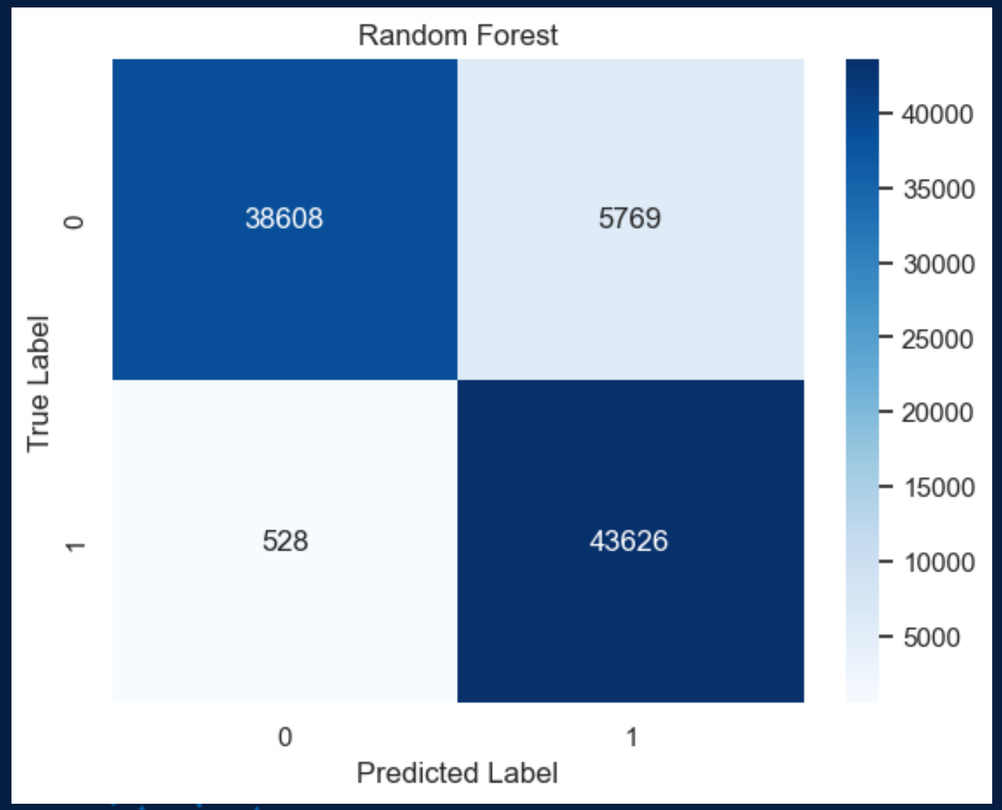
n_neighbors =5

Observing Elbow method and output from grid search.



	Accuracy	confusion Matrix	recall	AUC	Specificity	F-Score
KNN	0.89	[[39371 5006] [4534 39620]]	0.9	0.93	0.887	0.893

Random Forest

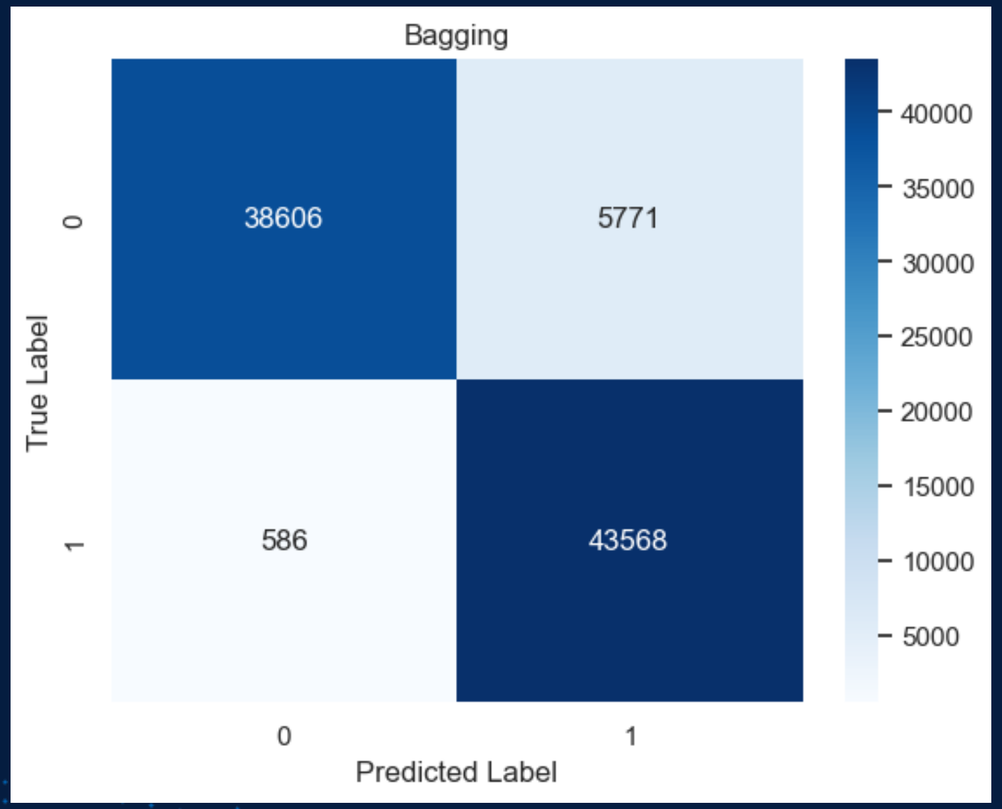


Best parameters found:

'max_depth': 13,
'n_estimators': 100

	Accuracy	confusion Matrix	recall	AUC	Specificity	F-Score
Random Forest	0.93	[4534 39620]] [[38608 5769]	0.99	0.95	0.87	0.933

Bagging



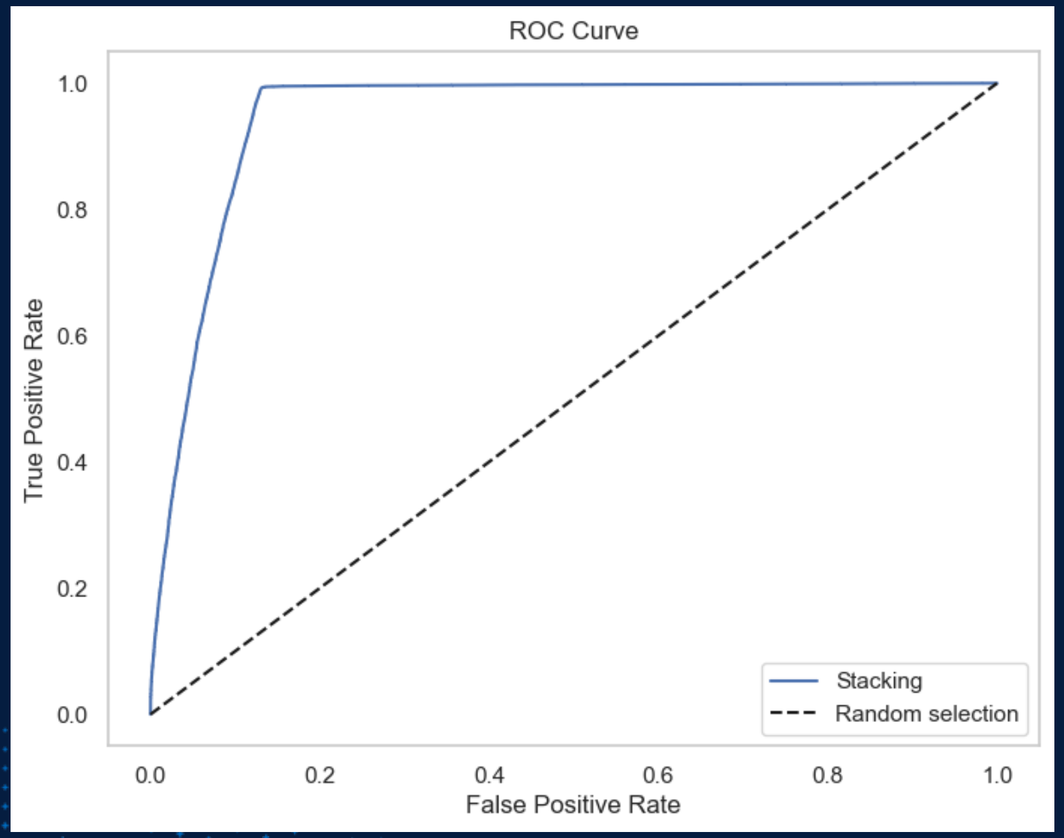
Best parameters found:

'-n_estimators': 15

-Bagging was implemented using a base estimator of RandomForestClassifier
-RandomForestClassifier with 100 trees and a maximum depth of 13

	Accuracy	confusion Matrix	recall	AUC	Specificity	F-Score
Bagging	0.93	[[38606 5771] [586 43568]]	0.99	0.95	0.87	0.932

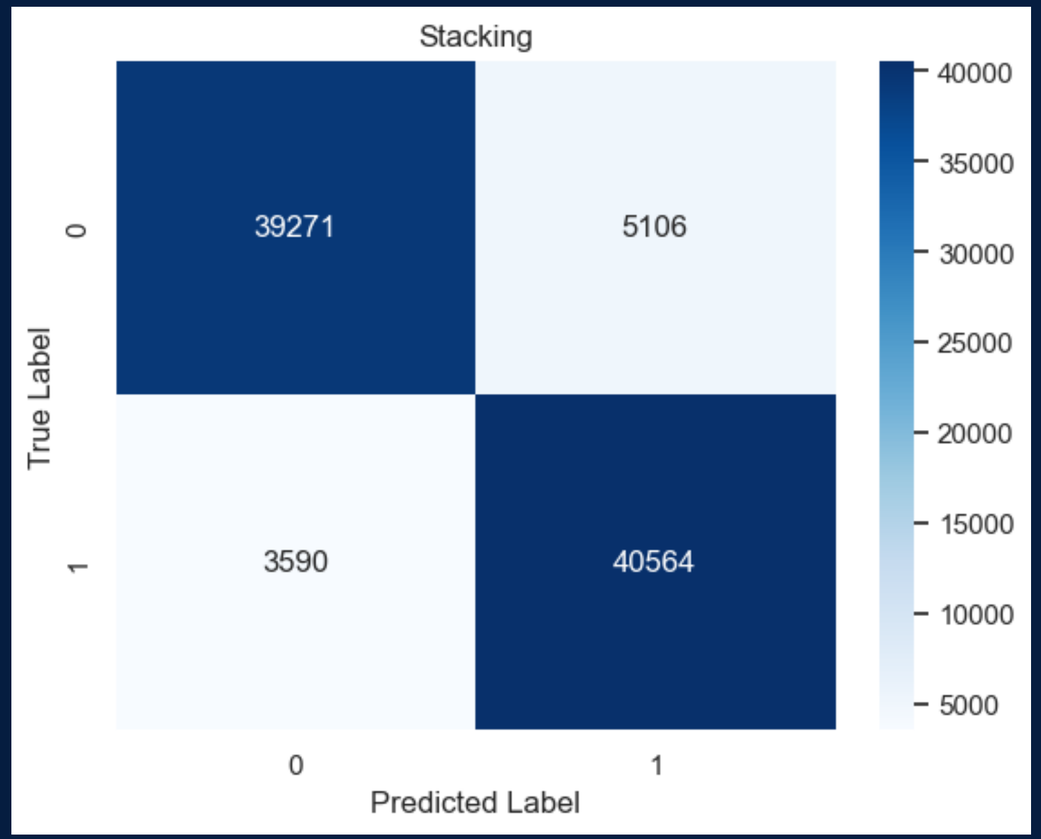
Stacking



meta-classifier used for aggregation - Logistic Regression

meta-classifier-: Naïve Bayes and Random Forest

Stacking

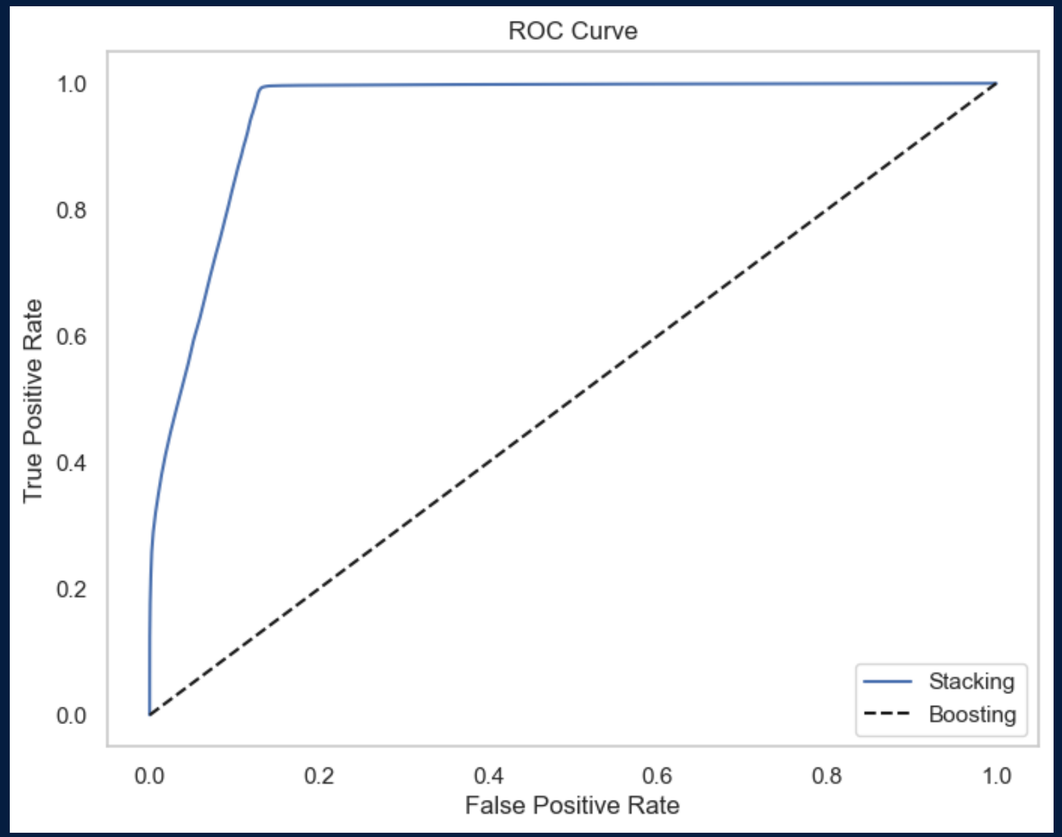


	Accuracy	confusion Matrix	recall	AUC	Specificity	F-Score
Stacking	0.9	[[39271 5106] [3590 40564]]	0.92	0.95	0.885	0.903

Stacking, Naïve Bayes, Random Forest

Model	Accuracy	Recall	AUC	Specificity	F-Score
Naïve Bayes	0.86	0.82	0.93	0.902	0.858
Random Forest	0.93	0.99	0.95	0.87	0.933
Stacking	0.9	0.92	0.95	0.885	0.903

Boosting

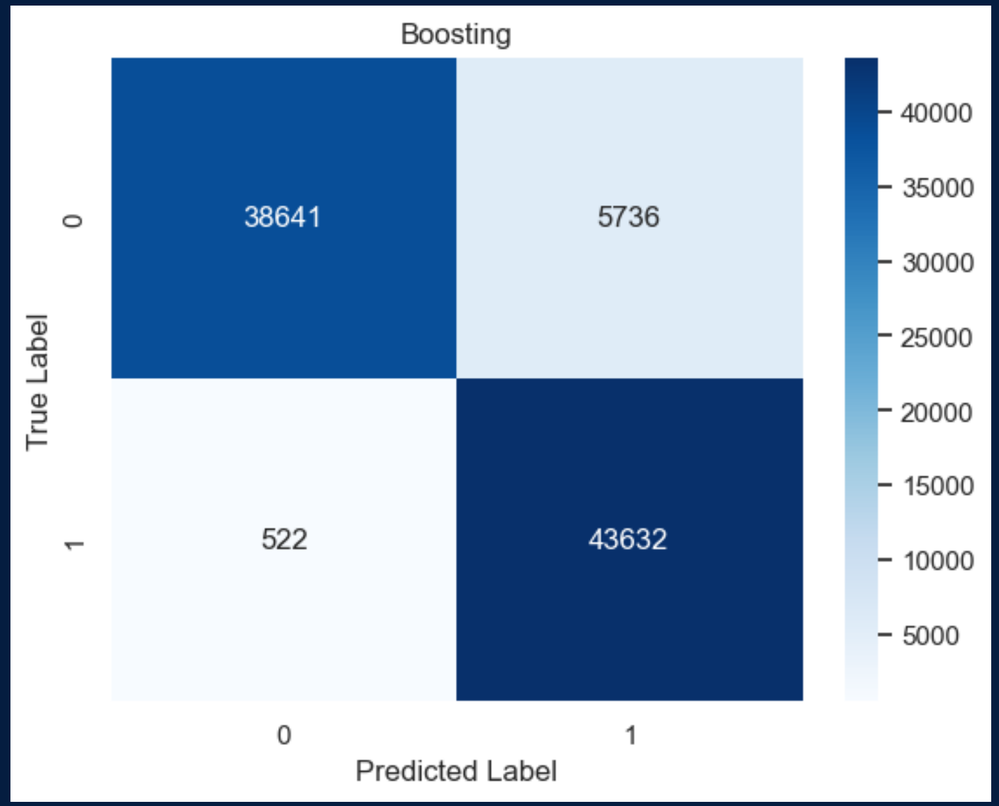


'base_estimator__n_estimators': 100

-AdaBoostClassifier was utilized with RandomForestClassifier as the base estimator

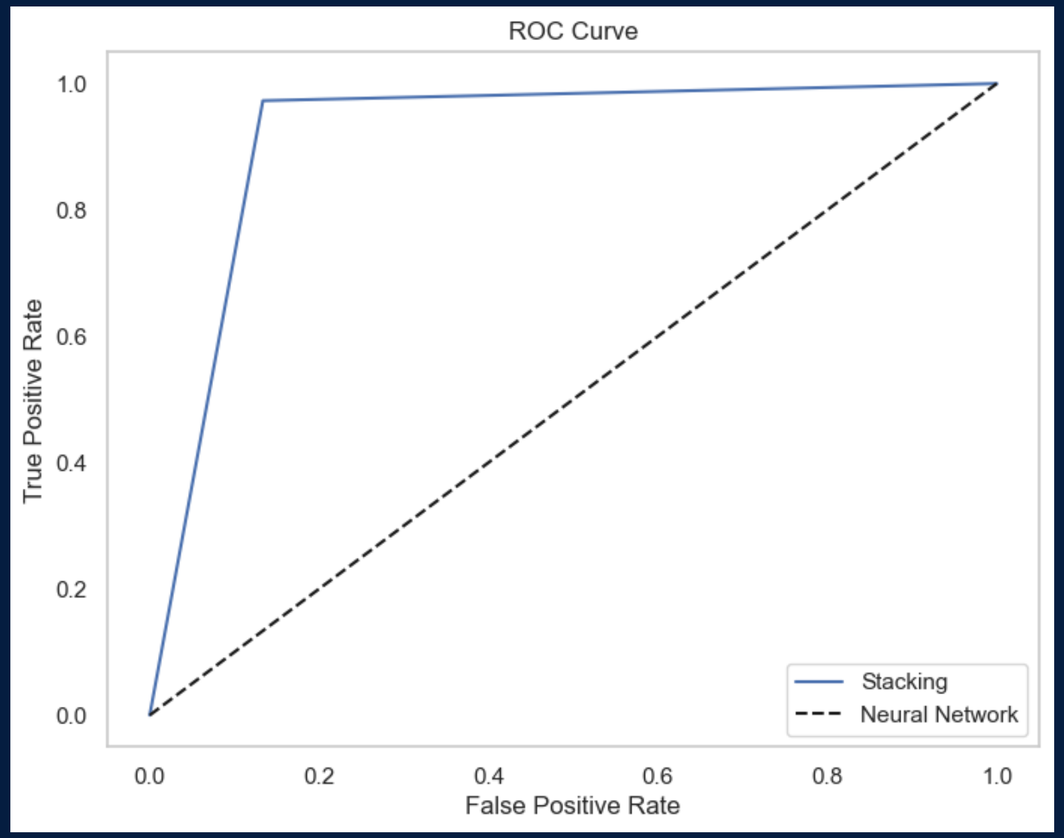
-Boosting model achieved an impressive accuracy of 0.93

Boosting



	Accuracy	confusion Matrix	recall	AUC	Specificity	F-Score
Boosting	0.93	[[38641 5736] [522 43632]]	0.99	0.95	0.871	0.933

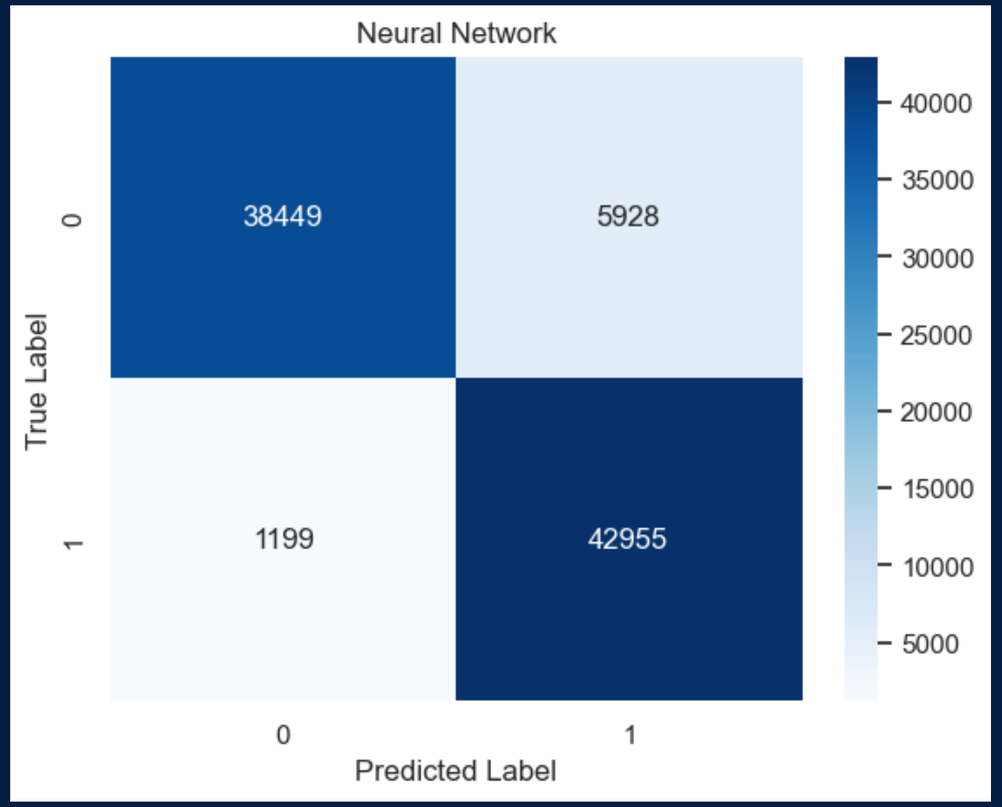
Neural Network



'hidden_layer_sizes': (50, 25, 10)
'max_iter': 100

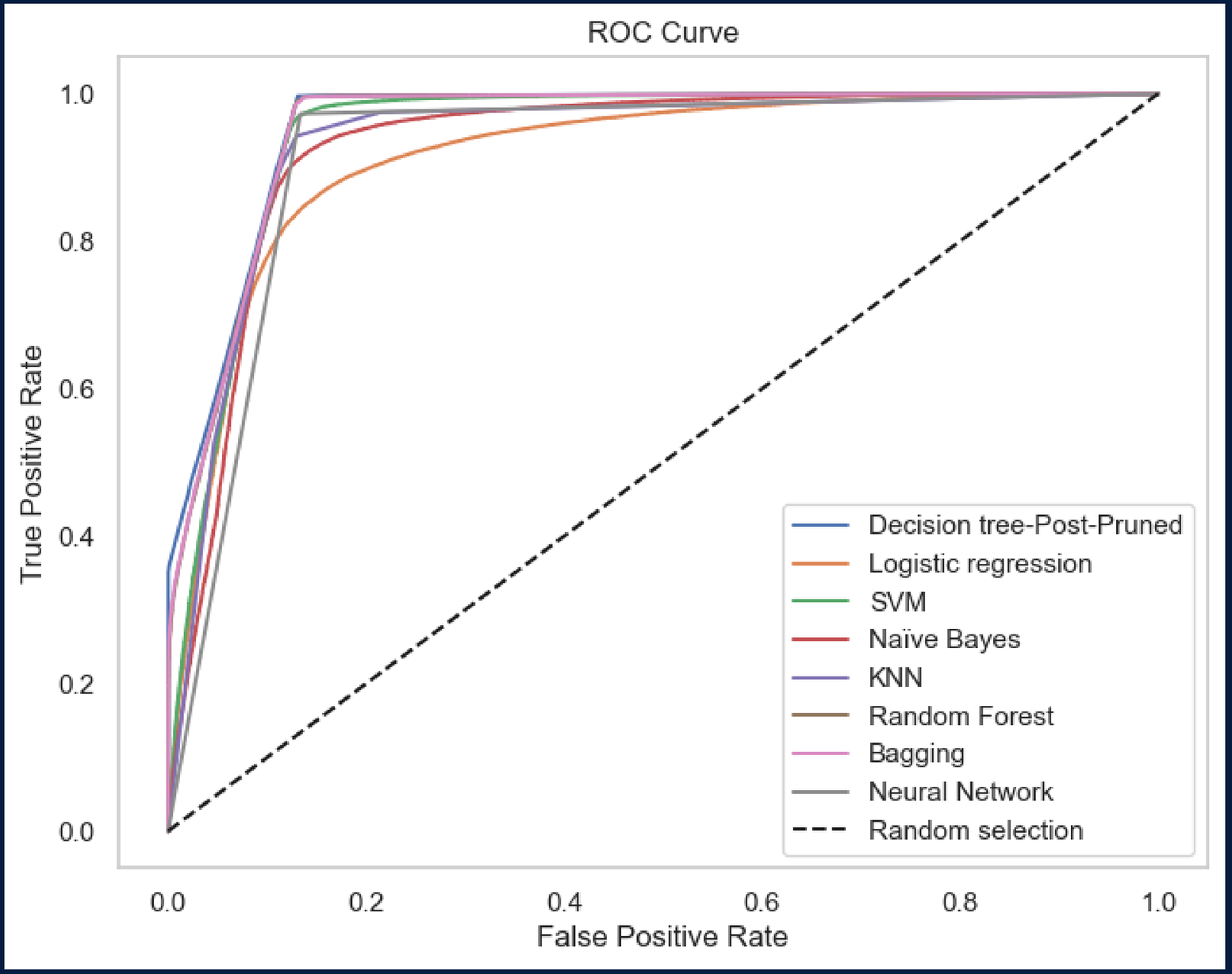
AUC:- 0.92

Neural Network



	Accuracy	confusion Matrix	recall	AUC	Specificity	F-Score
Neural Network	0.92	[[38449 5928] [1199 42955]]	0.97	0.92	0.866	0.923

ROC curve with all classification model



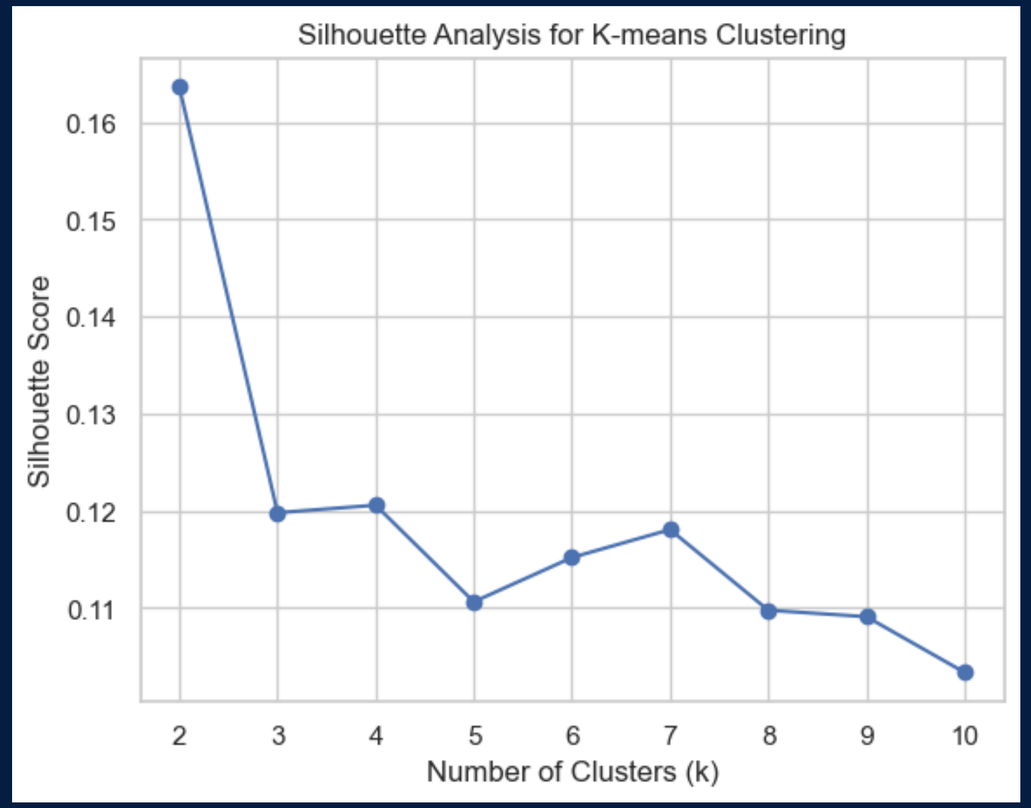
-All models demonstrated AUC values above that of a random classifier (AUC = 0.5),

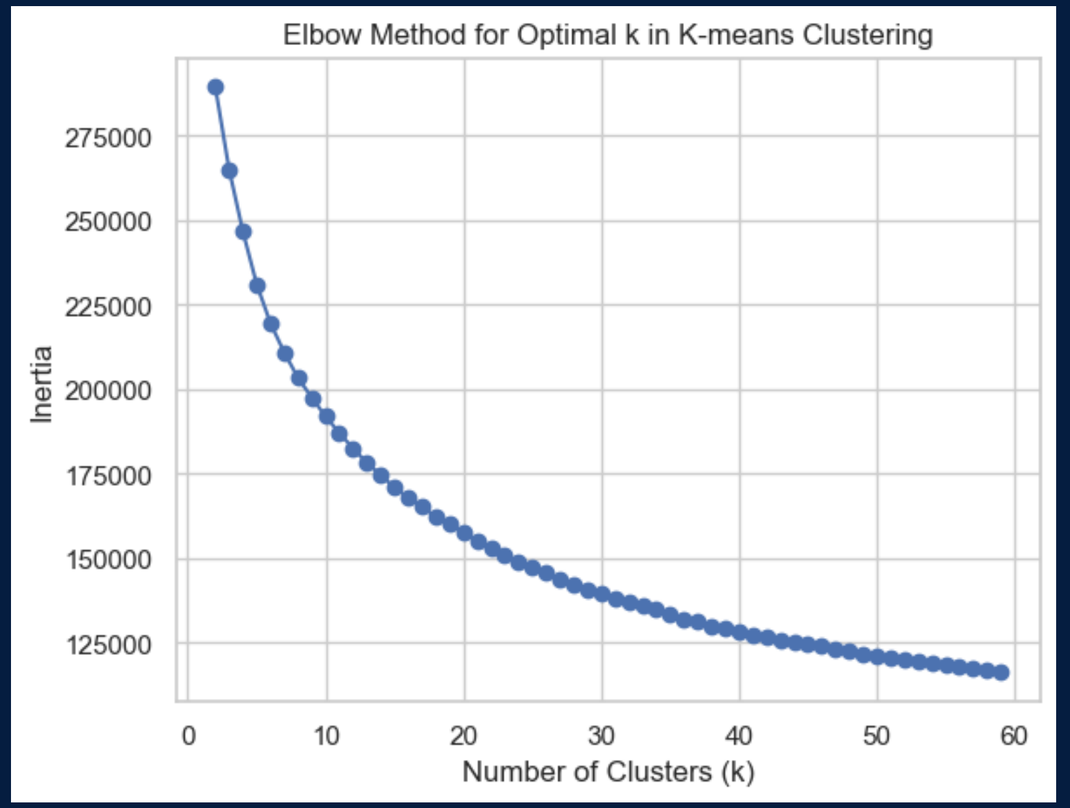
-Decision Tree Post-Pruned has better discriminative performance

All classification model

	AUC	Accuracy	confusion Matrix	recall	Specificity	F-Score
Decision Tree Post-Pruned	0.96	0.93	[[38573 5804] [162 43992]]	1.0	0.869	0.936
logistic regression	0.91	0.85	[[38888 5489] [7563 36591]]	0.83	0.876	0.849
SVM	0.94	0.91	[[38901 5476] [2170 41984]]	0.95	0.877	0.917
Naïve Bayes	0.93	0.86	[[40006 4371] [7732 36422]]	0.82	0.902	0.858
KNN	0.93	0.89	[[39371 5006] [4534 39620]]	0.9	0.887	0.893
Random Forest	0.95	0.93	[[38608 5769] [528 43626]]	0.99	0.87	0.933
Bagging	0.95	0.93	[[38606 5771] [586 43568]]	0.99	0.87	0.932
Stacking	0.95	0.9	[[39271 5106] [3590 40564]]	0.92	0.885	0.903
Boosting	0.95	0.93	[[38641 5736] [522 43632]]	0.99	0.871	0.933
Neural Network	0.92	0.92	[[38449 5928] [1199 42955]]	0.97	0.866	0.923

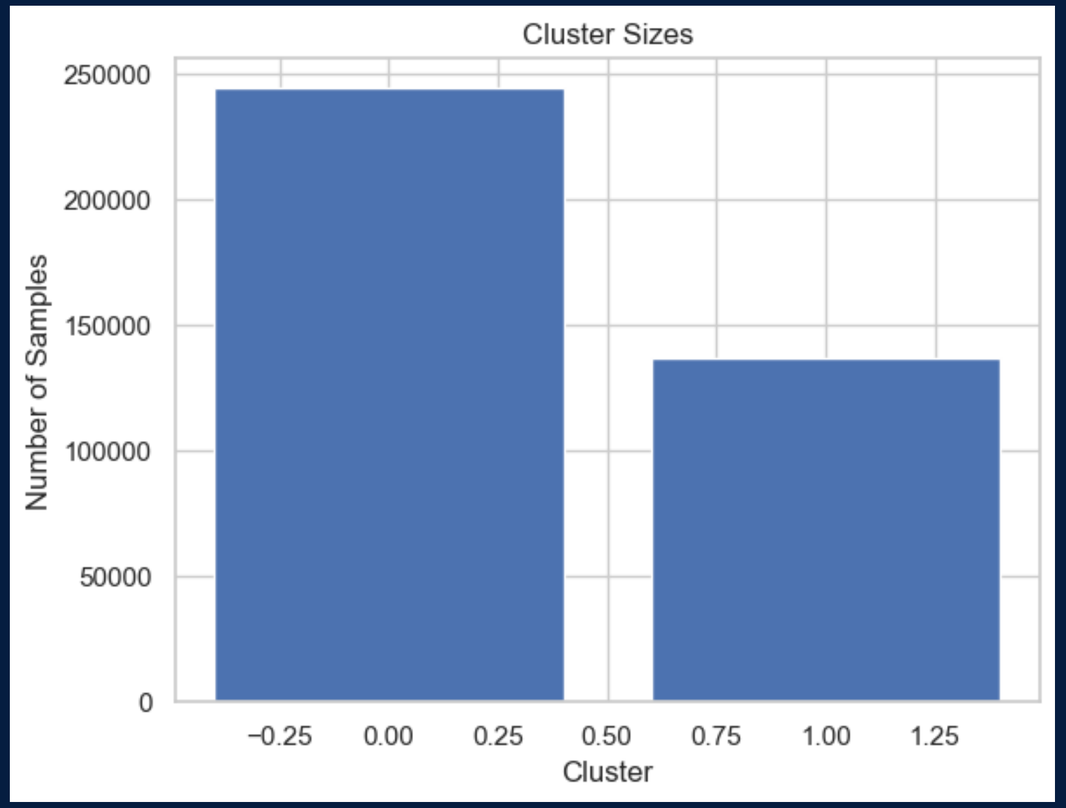
K-means clustering

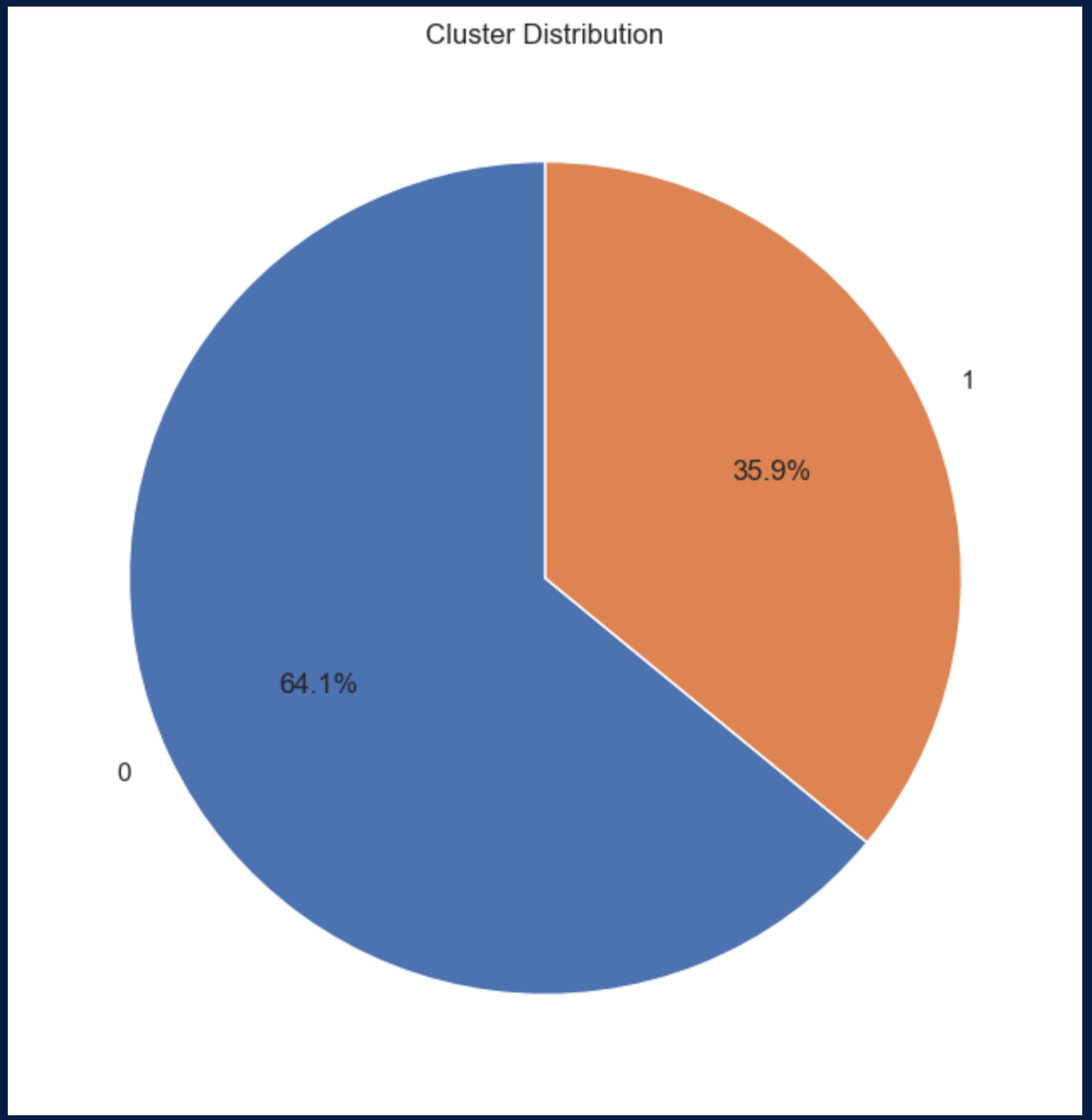




- Fram Silhouette score graph, vakue of k= 2

K-means clustering





Apriori-Rules

antecedents	consequents	support	confidence	lift	conviction
(NO_Churn)	(Support_Calls_Low)	0.42292	0.845840	1.463138	2.736767
(Support_Calls_Low)	(NO_Churn)	0.42292	0.731569	1.463138	1.862676
(Child, NO_Churn)	(Support_Calls_Low)	0.28822	0.913650	1.580436	4.885924
(Child, Support_Calls_Low)	(NO_Churn)	0.28822	0.794301	1.588602	2.430734
(Male, NO_Churn)	(Support_Calls_Low)	0.27912	0.859095	1.486067	2.994218
(Male, Support_Calls_Low)	(NO_Churn)	0.27912	0.800734	1.601469	2.509214
(Spend_High, NO_Churn)	(Support_Calls_Low)	0.24424	0.869181	1.503514	3.225079
(Spend_High, Support_Calls_Low)	(NO_Churn)	0.24424	0.844011	1.688023	3.205361
(NO_Churn, Usage_Frequency_High)	(Support_Calls_Low)	0.22508	0.847376	1.465795	2.764309
(Usage_Frequency_High, Support_Calls_Low)	(NO_Churn)	0.22508	0.744854	1.489708	1.959663

Apriori-Rules- Item Set

0.57810	(Support_Calls_Low)
0.56498	(Male)
0.55892	(Child)
0.51268	(Usage_Frequency_High)
0.50000	(Churn)
0.50000	(NO_Churn)
0.48732	(Usage_Frequency_Less)
0.44108	(Adult)

support	itemsets
0.43502	(Female)
0.42292	(NO_Churn, Support_Calls_Low)
0.40468	(Contract_Length_Annual)
0.40462	(No_Payment_Delay)
0.39898	(Contract_Length_Quarterly)
0.38758	(Minor_Payment_Delay)
0.37964	(Interaction_Low)
0.36350	(Spend_Medium)

support	itemsets
0.22038	(Spend_Medium, Support_Calls_Low)
0.22264	(Child, Male, Support_Calls_Low)
0.22270	(Usage_Frequency_High, Female)
0.22280	(Adult, Usage_Frequency_High)
0.22286	(Minor_Payment_Delay, Child)
0.22342	(Support_Calls_High)
0.22508	(NO_Churn, Usage_Frequency_High, Support_Calls_Low)