Credit Card Churn

Classification Analysis

Main Objective of Analysis

Using various classification models, the goal of this analysis was to determine which model provided the best results based on standard metrics to determine credit card churn. This kind of classification analysis would be useful to a bank manager who wants to know why certain customers close their accounts and assess what kinds of factors would prevent future customers from churning. The primary focus of this analysis was prediction rather than interpretation and after comparing the predictions of several different classification models, utilize an ensemble method to combine them and assess which model overall performed the best. The model that scored the best on standard metrics for classification can later on be handed over to the bank so they can utilize it to better understand credit card churn.

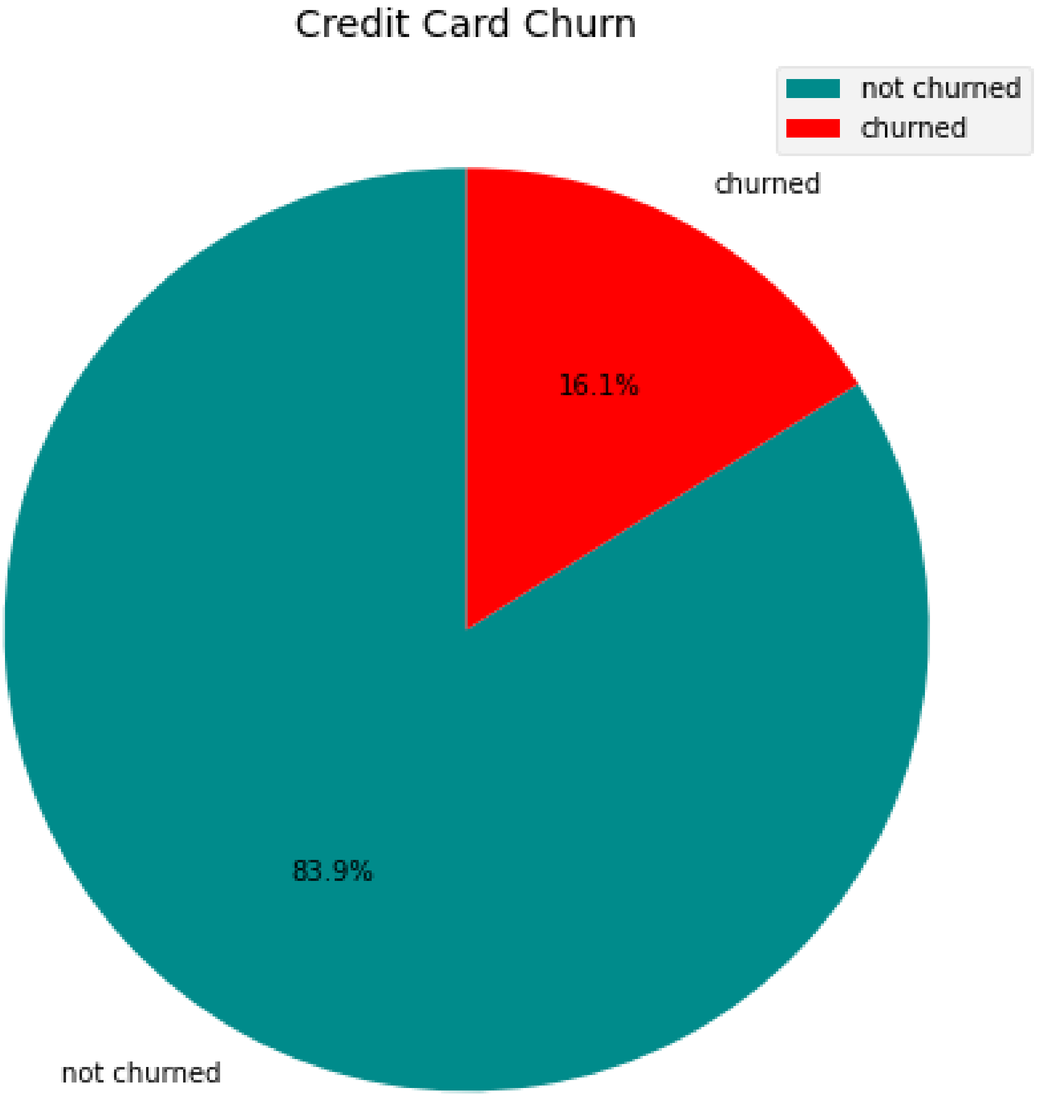
Description of Dataset

The dataset is from an anonymous source with sensitive personal information like name, birthday, and credit card number excluded. It is a sample of 10127 accounts with the following 23 attributes:

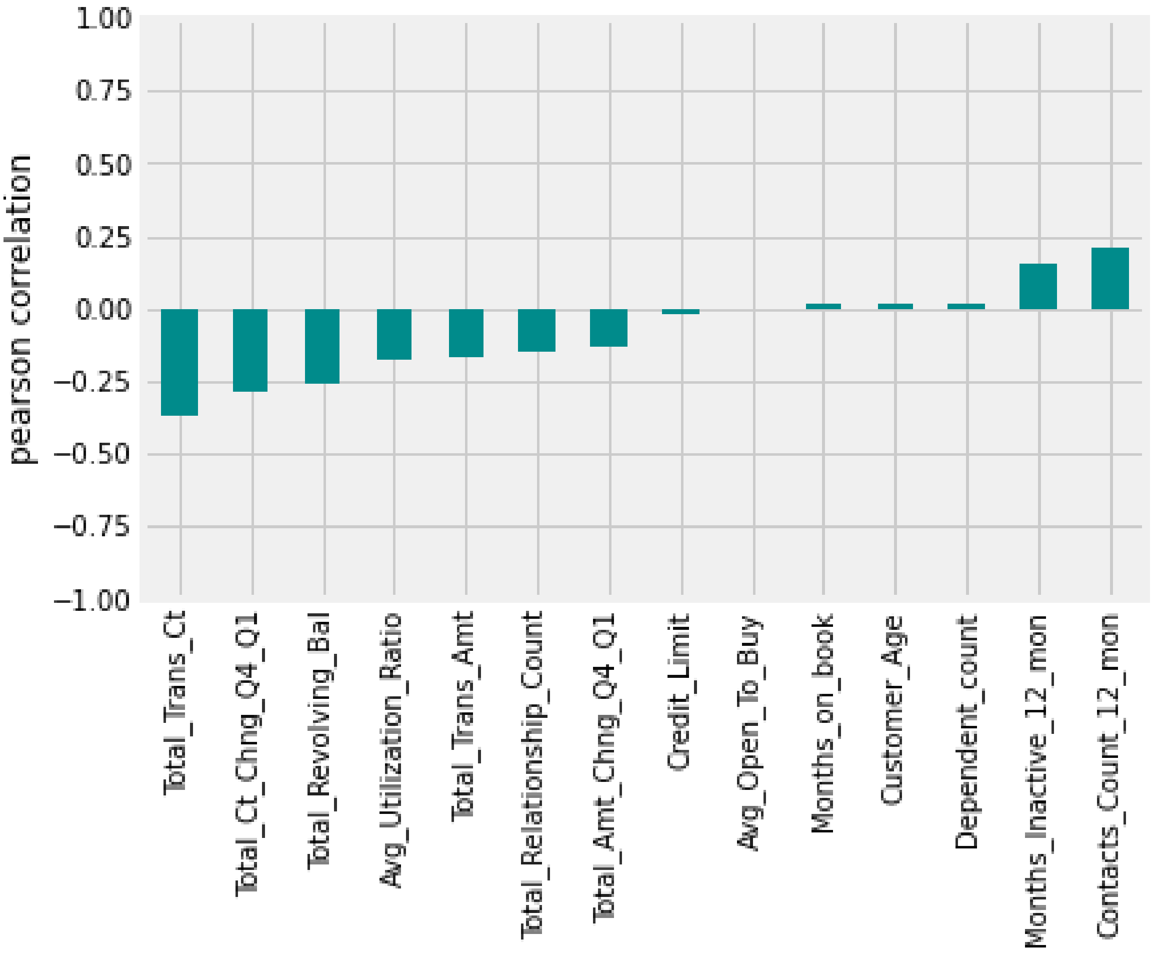
1. Client number (int64)
2. Attrition flag (object)
3. Customer age (int64)
4. Gender (object)
5. Dependent count (int64)
6. Education level (object)
7. Marital status (object)
8. Income category (object)
9. Card category (object)
10. Months on book (int64)
11. Total relationship count (int64)
12. Months inactive 12 months (int64)
13. Contacts count 12 months (int64)
14. Credit limit (float64)
15. Total revolving balance (int64)
16. Average open to buy (float64)
17. Total amount changed Q4 Q1 (float64)
18. Total transaction amount (int64)
19. Total transaction count (int64)
20. Total count changed Q4 Q1 (float64)
21. Average utilization ratio (float64)
22. Data column from previous project
23. Data column from previous project

Exploratory Data Analysis, Data Cleaning, and Feature Engineering

The target variable for the present analysis is the churn rate, which is indicated by the attrition flag in the dataset. Plotting the churned versus not-churned data points, the data is clearly unbalanced with not-churned data points outnumbering the churned data points and may result in some bias and variance issues, but for the purposes of this analysis, this facet of the data was not altered:



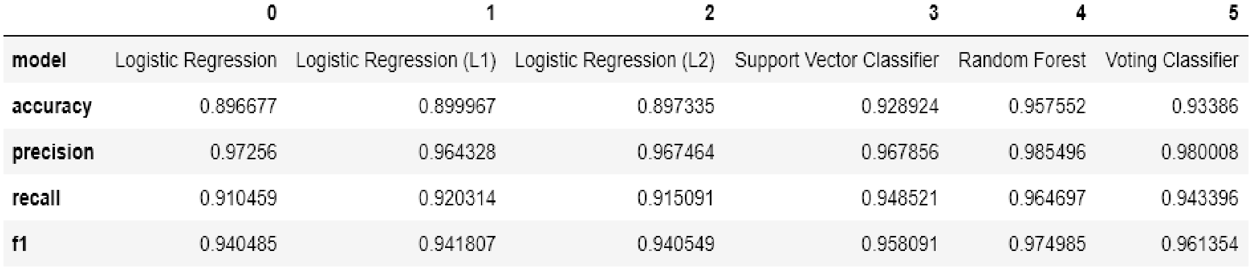
A correlation bar chart was also plotted which can be used for interpretive analysis in upcoming projects. Higher total transaction counts reduce the tendency for a customer to churn, more so than the total transaction amount.



Some numerical variables were plotted as density curves and distributions varied depending on the type of variable. No null values were present in the dataset. Several of the dependent categorical variables had an 'Unknown' value which was kept and treated as its own category rather than dropping those observations entirely or replacing them with other values. The client number column was dropped as it did not provide meaningful numerical information, as well as the last two columns which were remnants of a previous project and not relevant here. For some models used in this analysis, catergorical variables must be encoded to integers, so variables with two unique values such as gender were binarized using `LabelBinarizer()` and variables with more than two unique values such as income category were encoded using `LabelEncoder()`. To split the data into train and test sets, `StratifiedShuffleSplit()` was used with one split and a test size ratio of 0.3. Finally, because variables such as age and total transaction amount were on widely different ranges of numerical values, `MinMaxScaler()` was utilized to scale the data so certain models used in this analysis would work better.

Classification Models

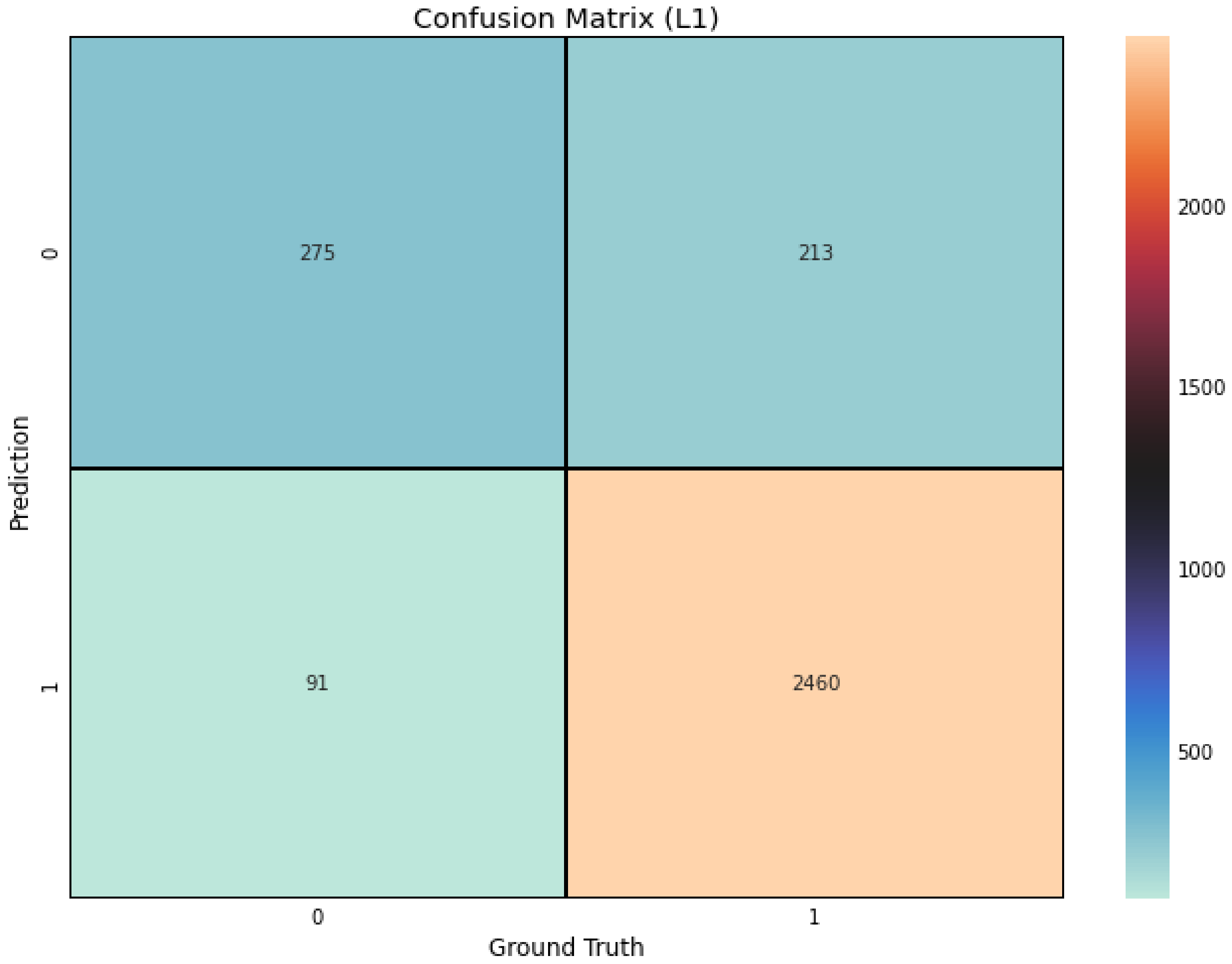
For the puposes of this analysis, the following classification models were used: logistic regression, support vector classifier (SVC), random forest, voting classifier. Simple logistic regression was used and then the scores were compared to L1 and L2 regularized logistic regression. After running `LogisticRegressionCV()` with `Cs=10`, `cv=4`, and `solver=liblinear`, L1 logistic regression resulted in the best accuracy score. `GridSearchCV()` was used to determine the optimal parameters for the SVC model, which were: `C=10.0`, `gamma=0.5`, and `kernel='rbf'`. `GridSearchCV()` was used to determine the optimal parameters for the random forest model, which were: `n_estimators=100` and `criterion='gini'`. The L1 logistic classifier, which resulted in the best accuracy score among the logistic regression models in this analysis, as well as SVC and random forest were combined in the ensemble voting classifier. The following are the classification scores for the predictions of these models:

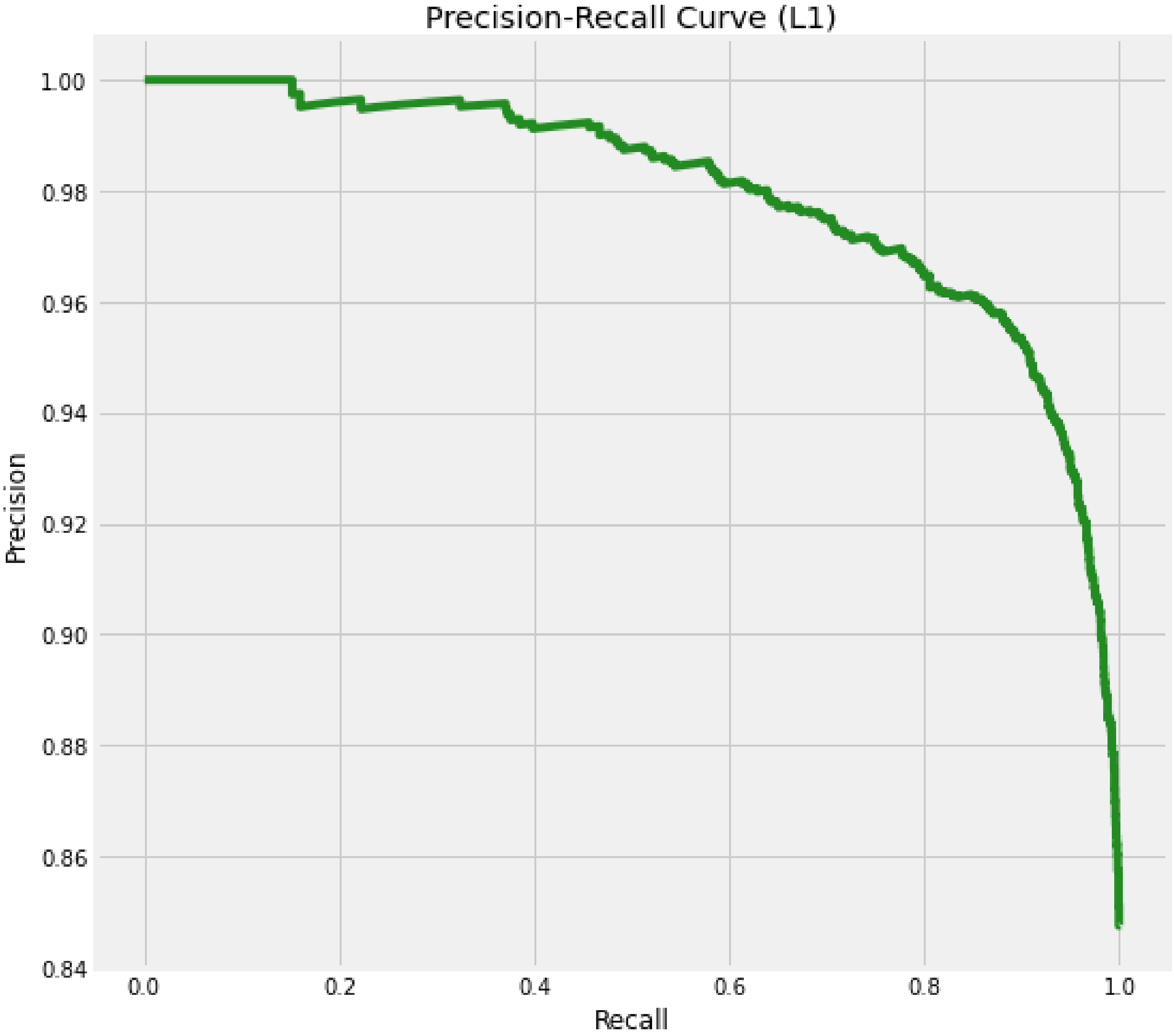


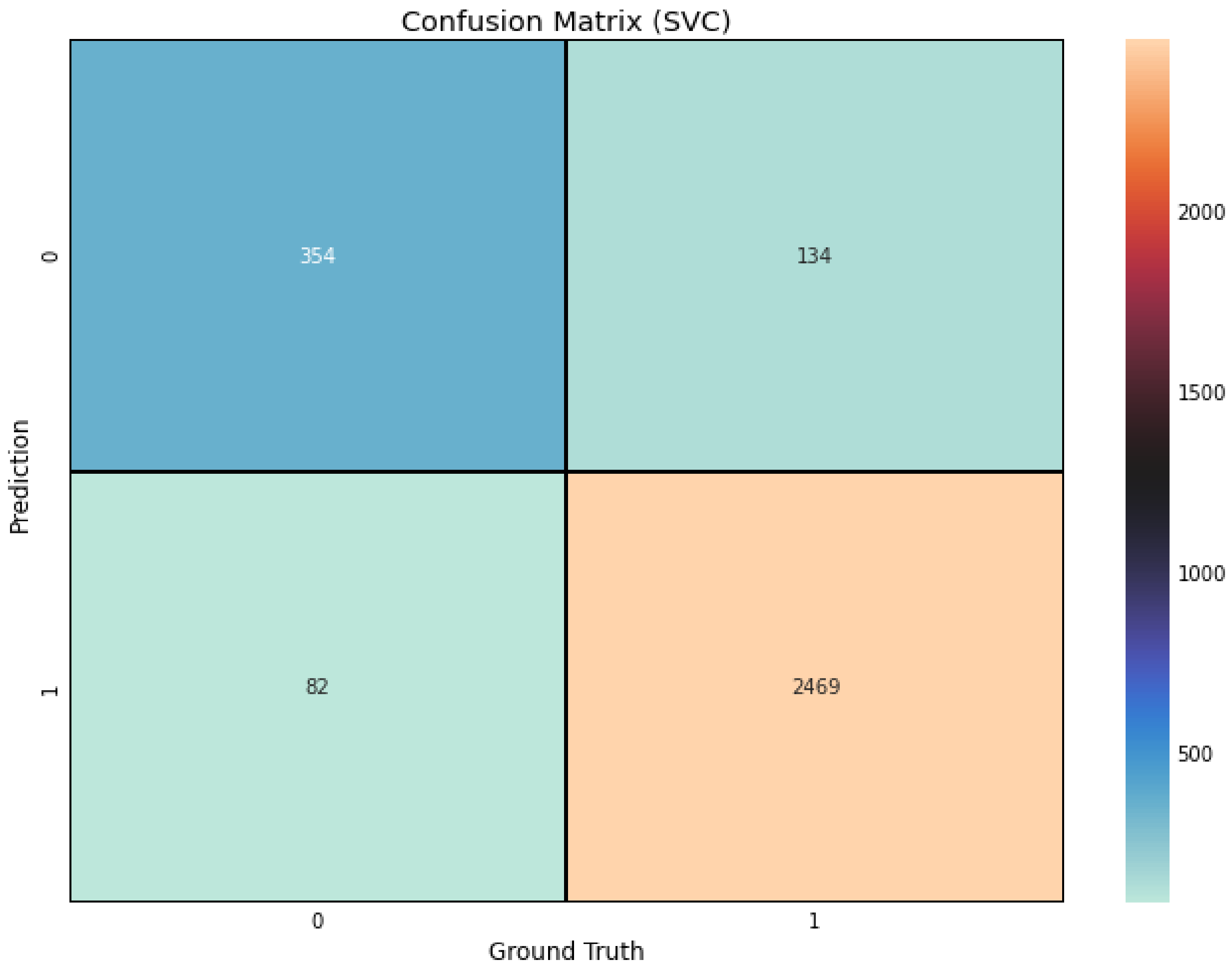
	0	1	2	3	4	5
model	Logistic Regression	Logistic Regression (L1)	Logistic Regression (L2)	Support Vector Classifier	Random Forest	Voting Classifier
accuracy	0.896677	0.899967	0.897335	0.928924	0.957552	0.93386
precision	0.97256	0.964328	0.967464	0.967856	0.985496	0.980008
recall	0.910459	0.920314	0.915091	0.948521	0.964697	0.943396
f1	0.940485	0.941807	0.940549	0.958091	0.974985	0.961354

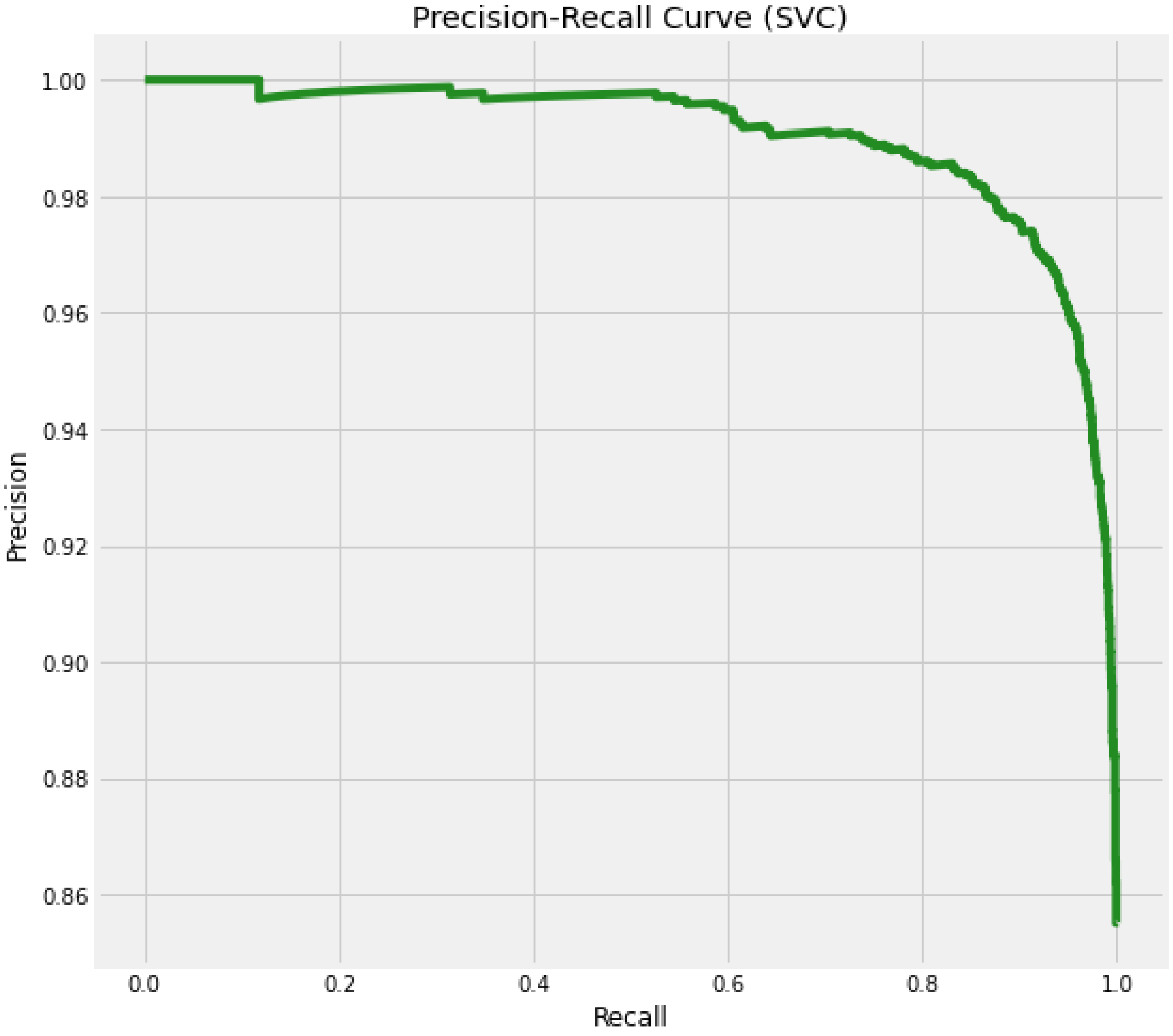
The random forest classifier produced the best scores in all of the metrics measures here, accuracy, precision, recall, and F1. It is clearly the most appropriate model for this credit card churn dataset. The voting classifier came in second and while it had the random forest model in its ensemble, bias and variance issues from the L1 logistic regression model and SVC model interfered with scores and didn't perform as well as random forest alone. All three variations of the logistic regression classifiers proved to be inferior to the other models trained and tested.

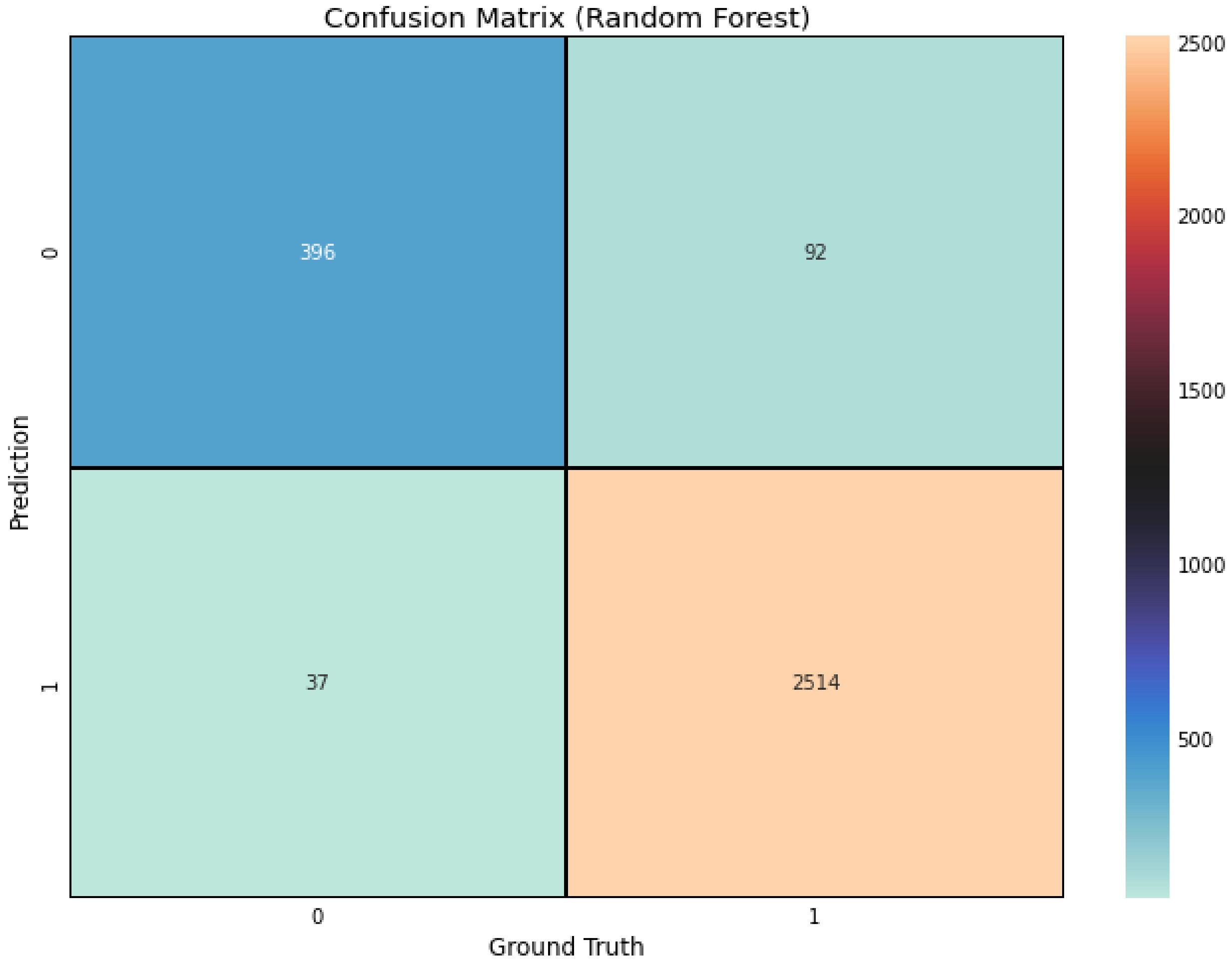
The following graphs are confusion matrices and precision-recall curves for L1 logistic regression, SVC, random forest, and voting classifier. Precision-recall curves were used rather than ROC curves because the dataset was unbalanced when comparing churned and not-churned observations. In the confusion matrices, all classification models had higher numbers of false positives compared to false negatives. This is likely due to the fact that the dataset was unbalanced. There are marked differences in the precision-recall curves for the classification models and they indicate that random forest is the best classifier for this dataset with superior scores in accuracy, precison, recall, and F1.

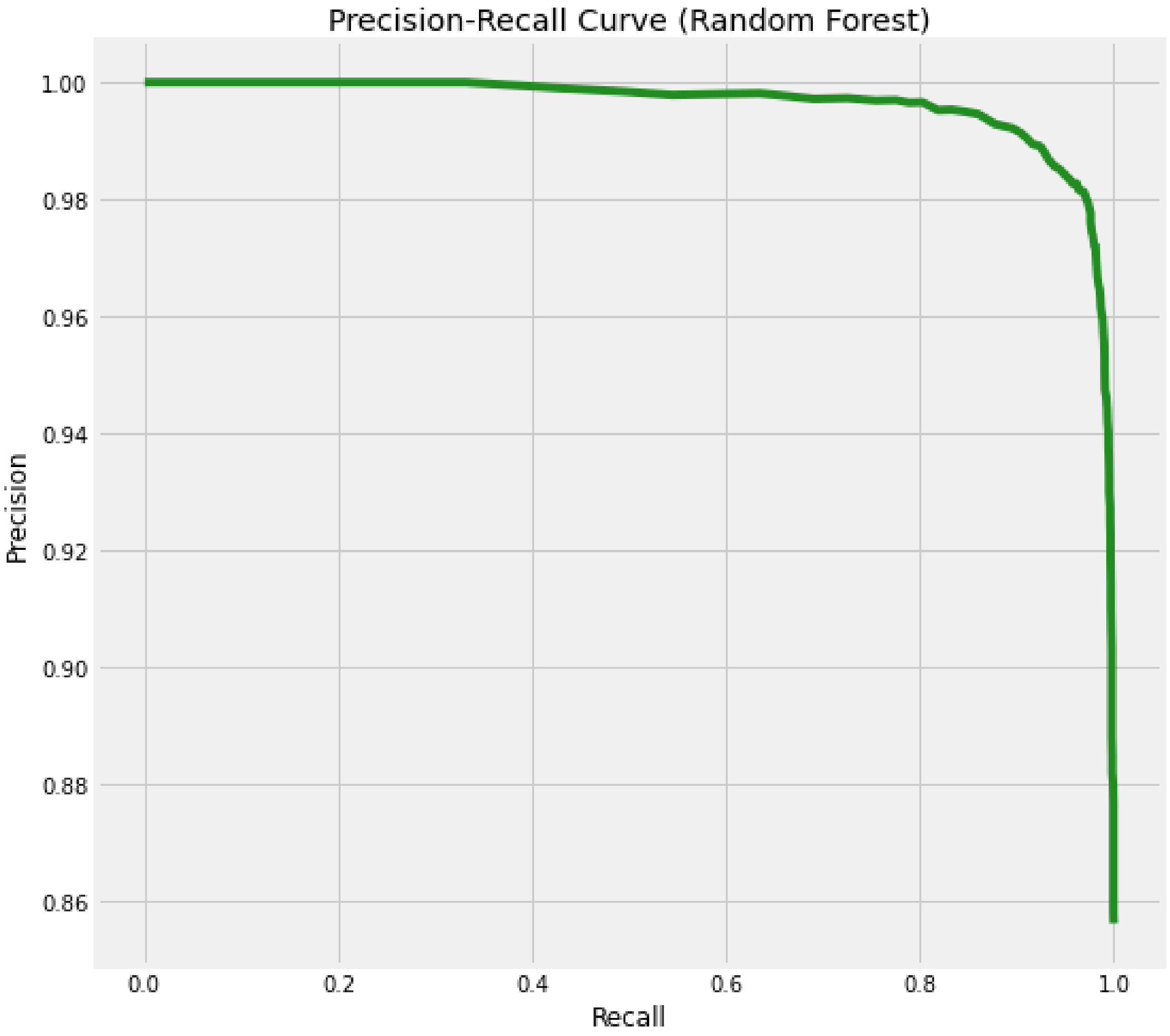


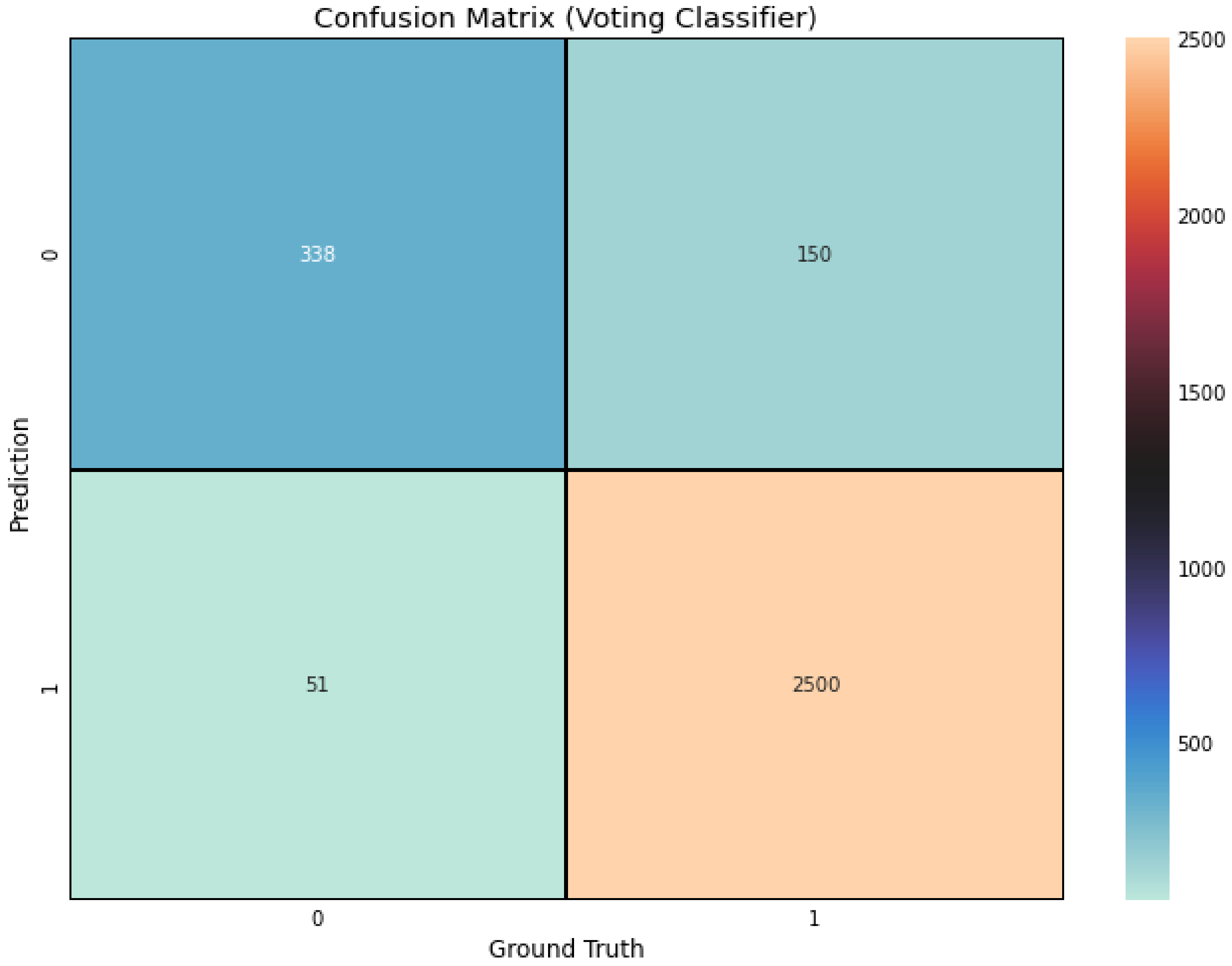

Precision-Recall Curve (L1)
Precision
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.0
0.2
0.4
0.6
0.8
1.0
Recall

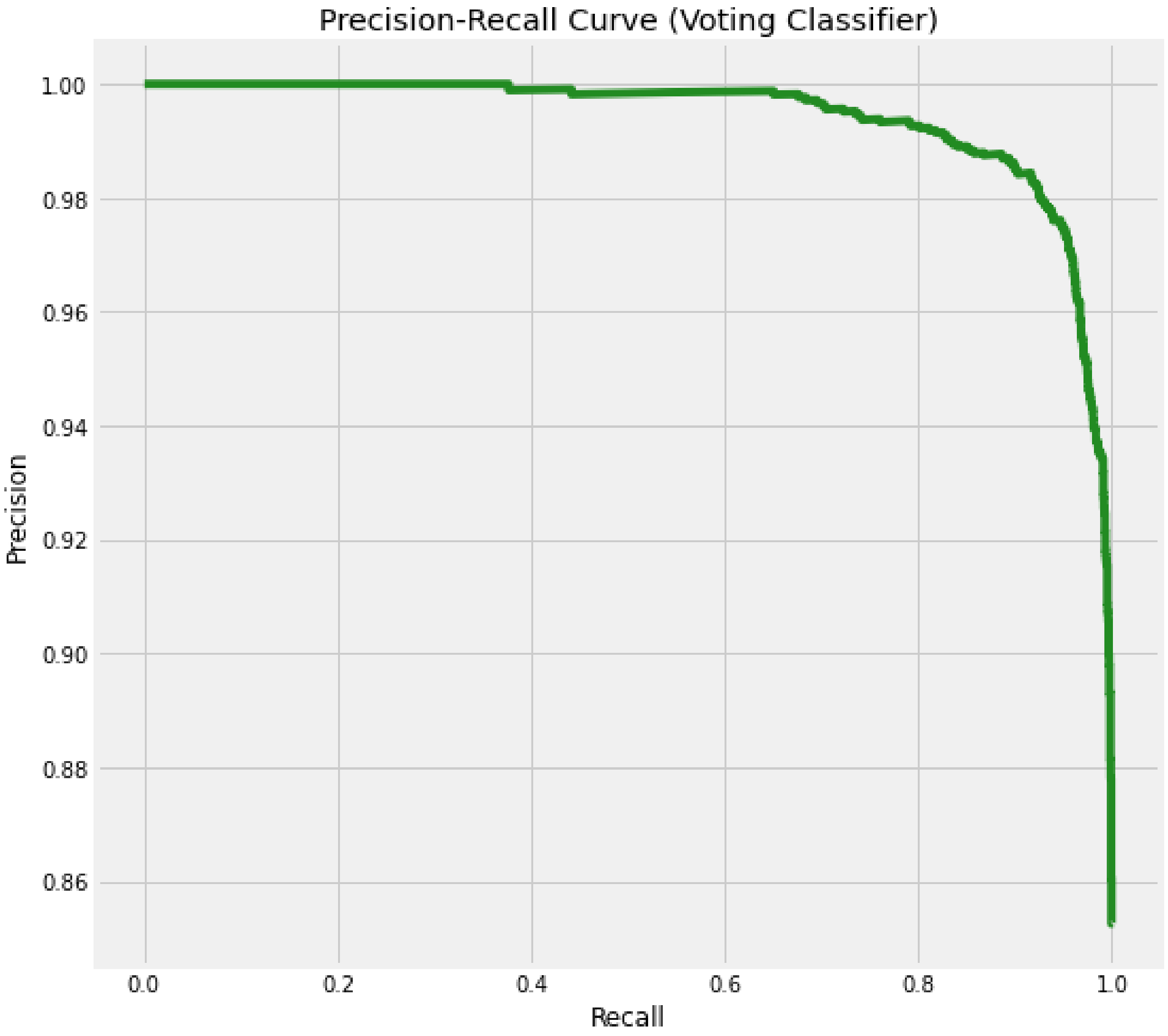

Confusion Matrix (SVC)
354
134
82
2469
0
1
Prediction
0
1
Ground Truth
2000
1500
1000
500


Precision-Recall Curve (SVC)
Precision
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
Recall
0.0
0.2
0.4
0.6
0.8
1.0


Confusion Matrix (Random Forest)
Prediction
0
1
396
92
37
2514
0
1
Ground Truth
2500
2000
1500
1000
500


Precision-Recall Curve (Random Forest)
Precision
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
Recall
0.0
0.2
0.4
0.6
0.8
1.0


Confusion Matrix (Voting Classifier)
Prediction
Ground Truth
0
1
0
1
338
150
51
2500
2500
2000
1500
1000
500


Precision-Recall Curve (Voting Classifier)
Precision
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
Recall
0.0
0.2
0.4
0.6
0.8
1.0

Key Findings and Insights

Taking into consideration the accuracy, precision, recall, and F1 scores as well as the visualizations for the confusion matrices and the precision-recall curves, the various classification metrics differ and the purely numeric models such as logistic regression performed more poorly than a meta-classifier like random forest which doesn't require feature encoding or scaling. The fact that the random forest classifier de-correlates trees and creates a random subset of features for each tree, thus reducing errors, may have played a key role in making it the superior classification model in this analysis of credit card churn, which included a fair amount of categorical variables. Interestingly, the voting classifier which used the random forest model as part of its ensemble didn't perform as well as random forest alone, indicating that the L1 logistic regression and SVC models also used in the voting classifier served to be confounding and shows that a more complex model isn't necessarily a better one.

Next Steps

There were two issues in this analysis that were not addressed. First, the fact that the dataset was unbalanced with not-churned observations outnumbering the churned observations. The dataset can be either upsampled or downsampled to create more of a balance. Second, many of the categorical variables in the dataset contained "Unknown" as values, which could mean a lot of different things. The dataset could be cleaned further in preprocessing to either replace "Unknown" with another value, or drop observations completely, although that might result in a reduction of the dataset significant enough to decrease the performance of classification models. Finally, for a more interpretive approach to the issue of credit card churn, a decision tree classifier could be built and then visualized.