Credit Card Fraud Detection

Group 10

Armaghan Abtahi
Dasaradh Gutta
Kunjal Shah
Priyanka Sharma

Introduction

The Federal Trade Commission's latest report reveals that credit card fraud was the leading type of identity theft reported in the United States in 2021, making up 39% of all reported cases.

Since about two-thirds of all U.S. credit and debit card holders have been victims of fraud, it’s safe to say the average person is well aware that this crime exists

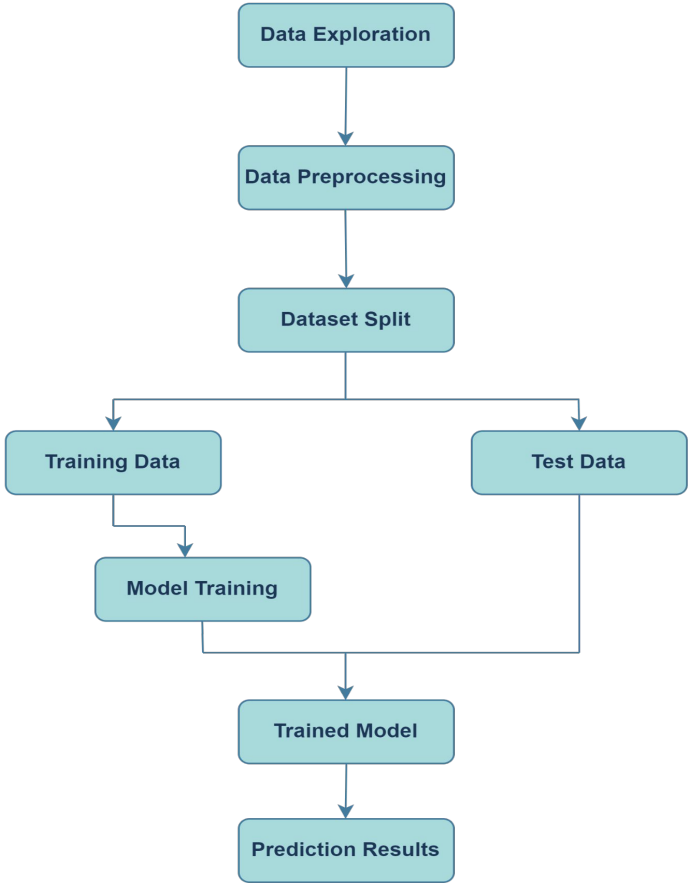
How much was the most recent fraudulent charge on your credit card?

DataSet

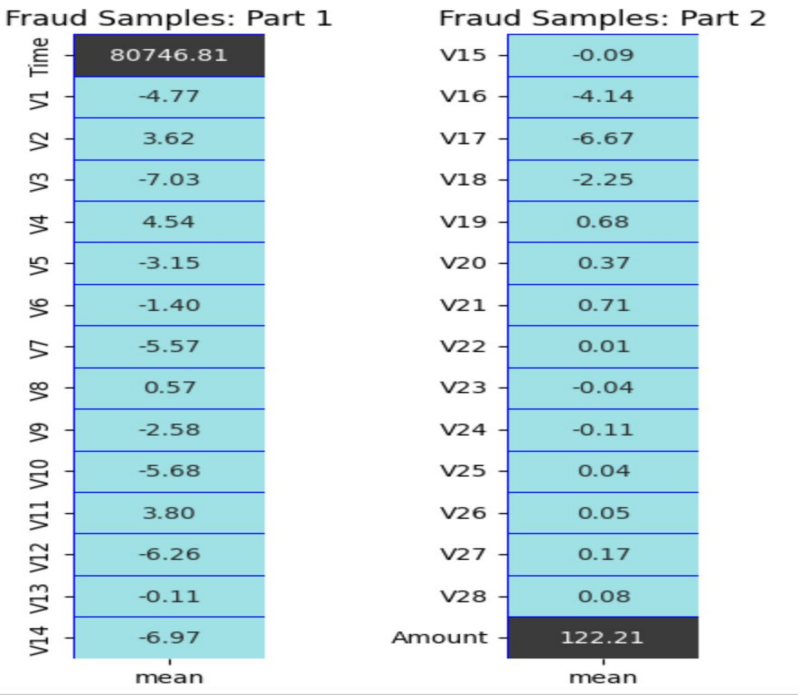
- The dataset contains transactions made by credit cards in September 2013 by European cardholders.
- This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

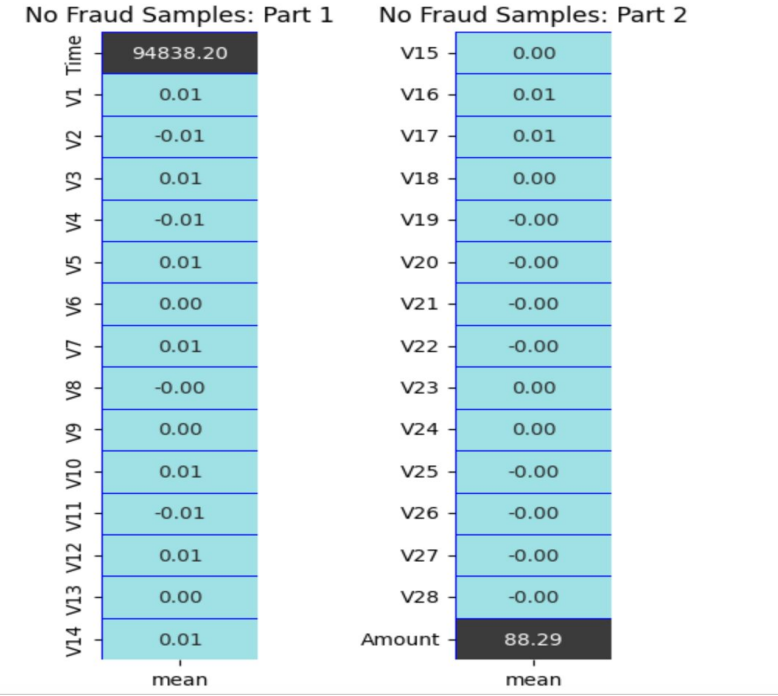
	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
0	0.00	-1.36	-0.07	2.54	1.38	-0.34	0.46	0.24	0.10	0.36	...	-0.02	0.28	-0.11	0.07	0.13	-0.19	0.13	-0.02	149.62	0
1	0.00	1.19	0.27	0.17	0.45	0.06	-0.08	-0.08	0.09	-0.26	...	-0.23	-0.64	0.10	-0.34	0.17	0.13	-0.01	0.01	2.69	0
2	1.00	-1.36	-1.34	1.77	0.38	-0.50	1.80	0.79	0.25	-1.51	...	0.25	0.77	0.91	-0.69	-0.33	-0.14	-0.06	-0.06	378.66	0
3	1.00	-0.97	-0.19	1.79	-0.86	-0.01	1.25	0.24	0.38	-1.39	...	-0.11	0.01	-0.19	-1.18	0.65	-0.22	0.06	0.06	123.50	0
4	2.00	-1.16	0.88	1.55	0.40	-0.41	0.10	0.59	-0.27	0.82	...	-0.01	0.80	-0.14	0.14	-0.21	0.50	0.22	0.22	69.99	0

Methodology

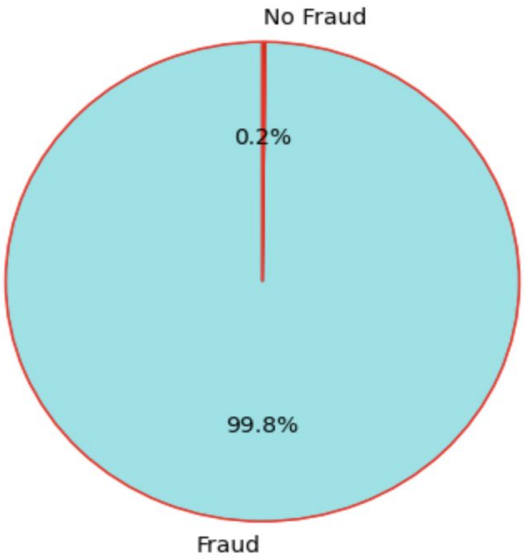


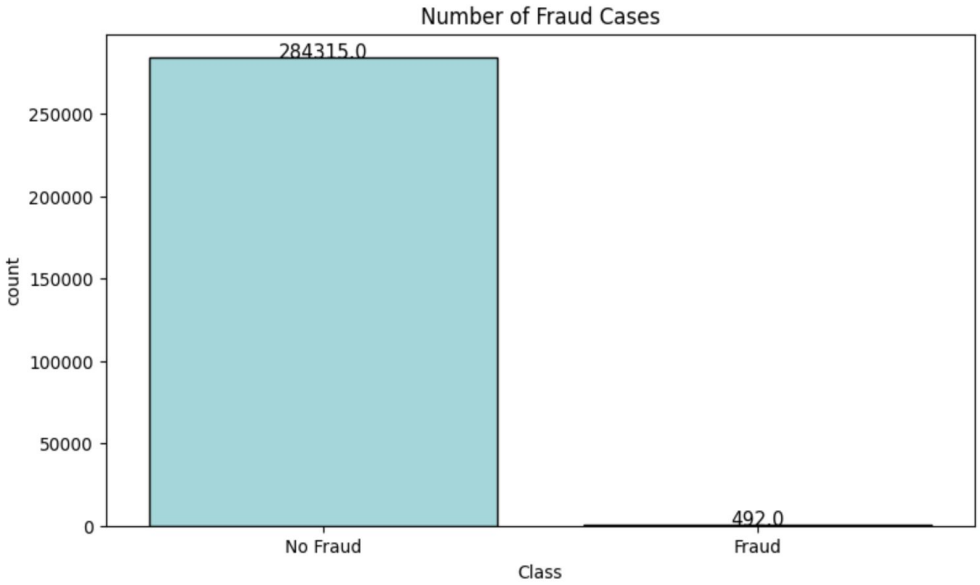
Data Set Visualisation





Data Set Visualization





Data Balancing:

The two common options for handling unbalanced data are:

Undersampling: reduces the majority samples of the target variable,

Oversampling: increases the minority samples to match the majority samples.

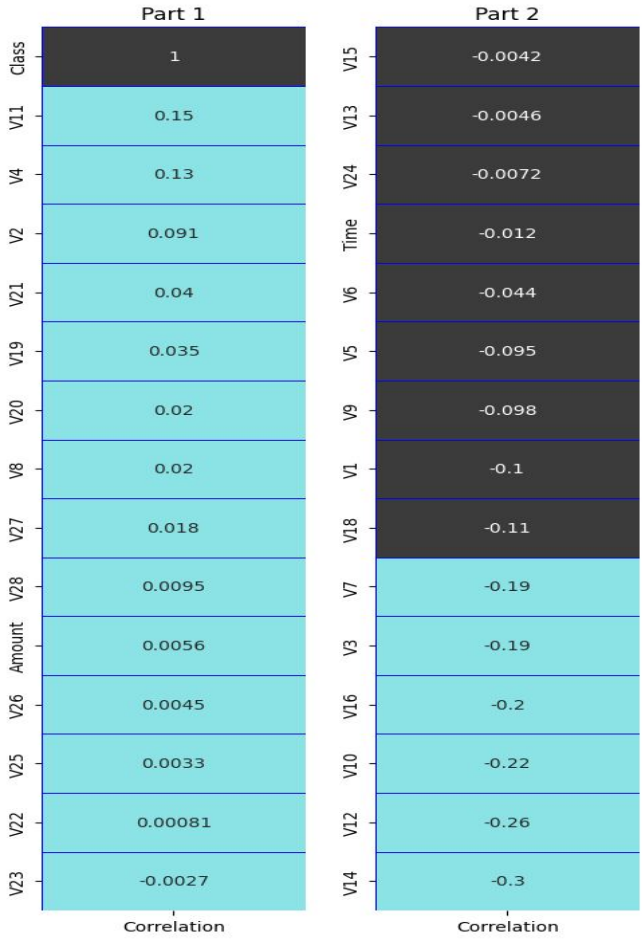
➡ Best approach for optimal performance: combination of undersampling and oversampling. In this case, we first undersample the majority samples and then oversample the minority samples.

Feature Selection:

- The dataset has a large number of features, making it difficult to understand.
- Two models is created based on the selected features from the **correlation map** and the **ANOVA score plot**.

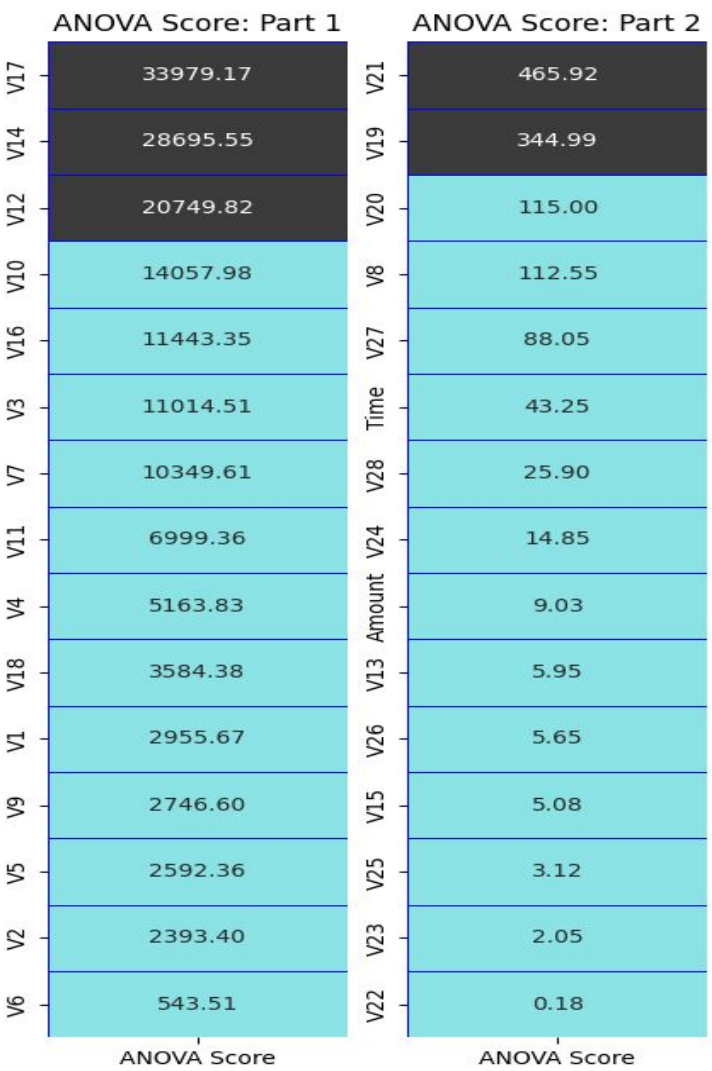
Correlation Map:

- A correlation map is created to show the relationship between each feature and the target variable.
- To select the most relevant features, remove any features that have correlation values falling within the range of -0.1 to 0.1.
- Among the remaining features, V4 and V11 have a positive correlation with the Class feature, while V7, V3, V16, V10, V12, V14, and V17 have a negative correlation with the Class feature.



ANOVA TEST:

- ANOVA score is used to measure the importance of each feature with respect to the target variable.
- Higher the value of the ANOVA score, higher the importance of that feature with the target variable.
- Features with ANOVA scores less than 50 is rejected.



Algorithms

We used several classification algorithms, including:

1. Logistic Regression
2. Support Vector Classifier
3. Decision Tree Classifier
4. Random Forest Classifier
5. K-Nearest Neighbors

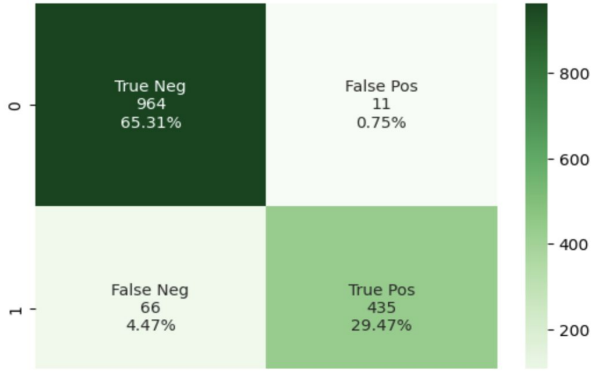
Logistic Regression

Model based on Correlation Plot:

Cross Validation Score: 98.31%

ROC_AUC Score: 92.85%

	precision	recall	f1-score	support
0	0.94	0.99	0.96	975
1	0.98	0.87	0.92	501
accuracy			0.95	1476
macro avg	0.96	0.93	0.94	1476
weighted avg	0.95	0.95	0.95	1476

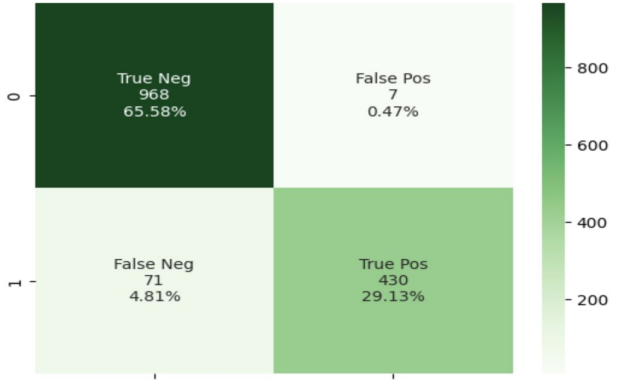


Model based on ANOVA Score:

Cross Validation Score: 98.37%

ROC_AUC Score: 92.56%

	precision	recall	f1-score	support
0	0.93	0.99	0.96	975
1	0.98	0.86	0.92	501
accuracy			0.95	1476
macro avg	0.96	0.93	0.94	1476
weighted avg	0.95	0.95	0.95	1476



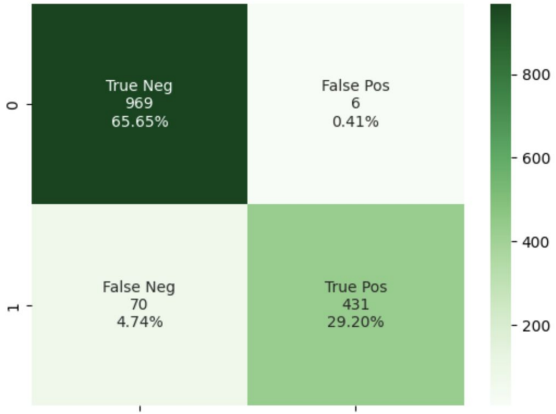
Support Vector Classifier

Model based on Correlation Plot:

Cross Validation Score: 98.32%

ROC_AUC Score: 92.71%

	precision	recall	f1-score	support
0	0.93	0.99	0.96	975
1	0.99	0.86	0.92	501
accuracy			0.95	1476
macro avg	0.96	0.93	0.94	1476
weighted avg	0.95	0.95	0.95	1476

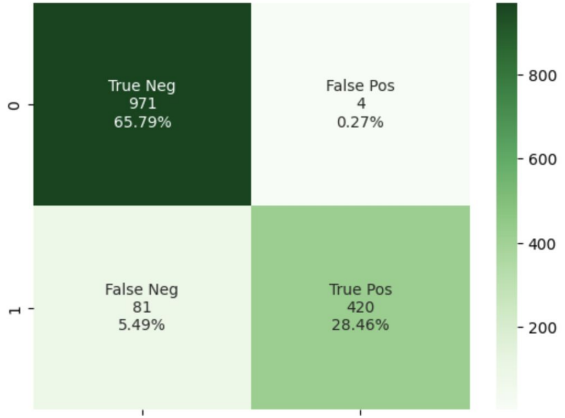


Model based on ANOVA Score:

Cross Validation Score: 98.23%

ROC_AUC Score: 91.71%

	precision	recall	f1-score	support
0	0.92	1.00	0.96	975
1	0.99	0.84	0.91	501
accuracy			0.94	1476
macro avg	0.96	0.92	0.93	1476
weighted avg	0.95	0.94	0.94	1476



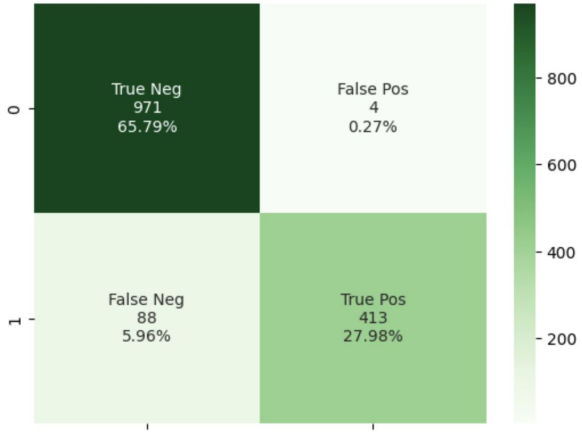
Decision Tree Classifier

Model based on Correlation Plot:

Cross Validation Score: 97.23%

ROC_AUC Score: 93.68%

	precision	recall	f1-score	support
0	0.92	1.00	0.95	975
1	0.99	0.82	0.90	501
accuracy			0.94	1476
macro avg	0.95	0.91	0.93	1476
weighted avg	0.94	0.94	0.94	1476

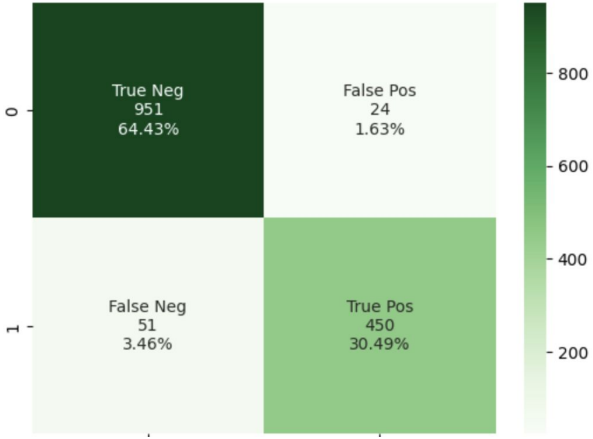


Model based on ANOVA Score:

Cross Validation Score: 96.48%

ROC_AUC Score: 91.01%

	precision	recall	f1-score	support
0	0.95	0.98	0.96	975
1	0.95	0.90	0.92	501
accuracy			0.95	1476
macro avg	0.95	0.94	0.94	1476
weighted avg	0.95	0.95	0.95	1476



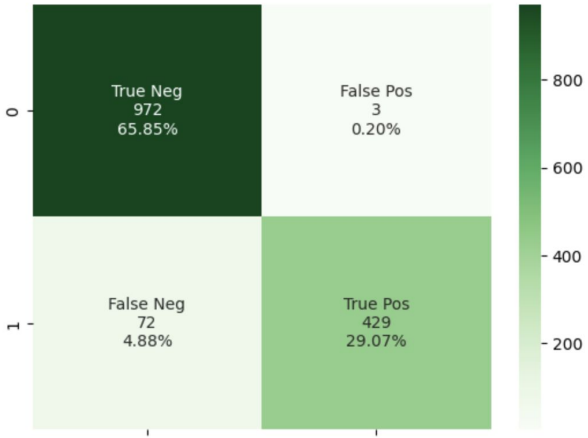
Random Forest Classifier

Model based on Correlation Plot:

Cross Validation Score: 98.35%

ROC AUC Score: 92.66%

	precision	recall	f1-score	support
0	0.93	1.00	0.96	975
1	0.99	0.86	0.92	501
accuracy			0.95	1476
macro avg	0.96	0.93	0.94	1476
weighted avg	0.95	0.95	0.95	1476

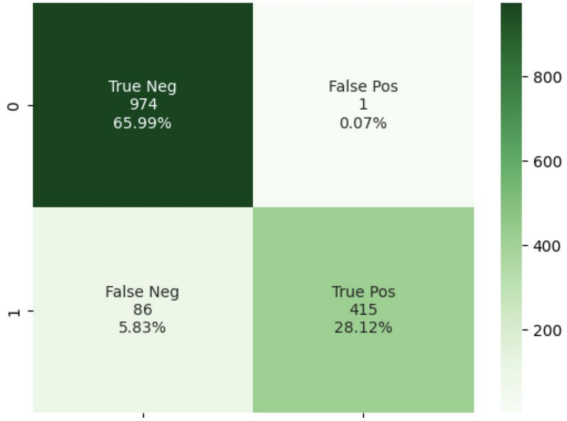


Model based on ANOVA Score:

Cross Validation Score: 98.09%

ROC_AUC Score: 91.37%

	precision	recall	f1-score	support
0	0.92	1.00	0.96	975
1	1.00	0.83	0.91	501
accuracy			0.94	1476
macro avg	0.96	0.91	0.93	1476
weighted avg	0.95	0.94	0.94	1476



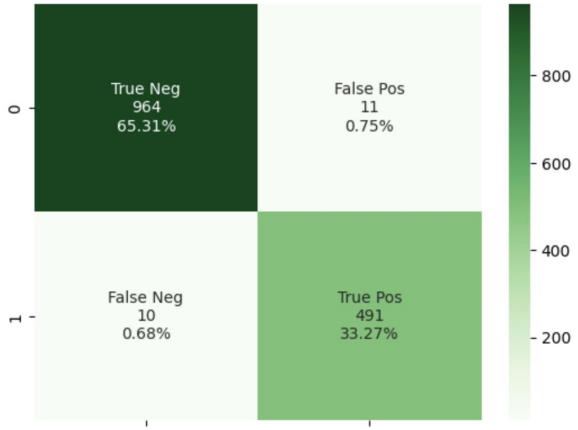
K-Nearest Neighbors

Model based on Correlation Plot:

Cross Validation Score: 99.32%

ROC_AUC Score: 98.44%

	precision	recall	f1-score	support
0	0.99	0.99	0.99	975
1	0.98	0.98	0.98	501
accuracy			0.99	1476
macro avg	0.98	0.98	0.98	1476
weighted avg	0.99	0.99	0.99	1476

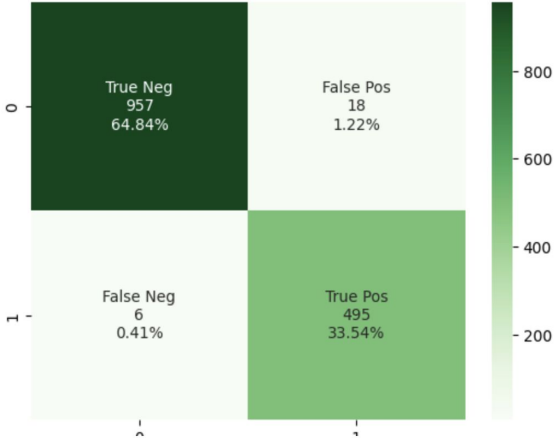


Model based on ANOVA Score:

Cross Validation Score: 99.62%

ROC_AUC Score: 98.48%

	precision	recall	f1-score	support
0	0.99	0.98	0.99	975
1	0.96	0.99	0.98	501
accuracy			0.98	1476
macro avg	0.98	0.98	0.98	1476
weighted avg	0.98	0.98	0.98	1476



Results- Summary of Models

Model	Feature Selection Method	Cross Validation Score	ROC-AUC Score
Logistic Regression	Correlation Plot	98.31%	92.85%
Logistic Regression	ANOVA Score	98.37%	92.56%
Support Vector Classifier	Correlation Plot	98.32%	92.71%
Support Vector Classifier	ANOVA Score	98.23%	91.71%
Decision Tree Classifier	Correlation Plot	97.23%	93.68%
Decision Tree Classifier	ANOVA Score	96.48%	91.01%
Random Forest Classifier	Correlation Plot	98.35%	92.66%
Random Forest Classifier	ANOVA Score	98.09%	91.37%
K-Nearest Neighbors	Correlation Plot	99.32%	98.44%
K-Nearest Neighbors	ANOVA Score	99.62%	98.48%

Thank You