Counterfeit Currency Classification

Thomas Powell
DSCI-6003 Machine Learning

Tools Used

Libraries

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Sklearn

Programming

- Jupyter Notebook

Database

1,000,000 features

Columns

- Country
- Denomination
- Counterfeit
- Serial Number
- Security Features
- Weight
- Length
- Width
- Thickness

	Country	Denomination	Counterfeit	SerialNumber	SecurityFeatures	Weight	Length	Width	Thickness
0	USA	$100	1	25973198	Hologram	1.731759	130.243185	66.537999	0.098488
1	USA	$20	1	95903230	Security Thread	1.002179	152.596364	76.135834	0.094119
2	EU	€10	0	82937914	Hologram	2.306713	152.857126	66.772442	0.061393
3	USA	€20	1	23612989	Microprint	1.366965	143.133672	78.377052	0.053114
4	EU	€20	1	56025342	Watermark	1.796075	129.664777	75.916093	0.051438
...	...	...	...	...	...	...	...	...	...
999995	EU	$100	1	24436622	Watermark	1.472511	134.888731	75.425943	0.093939
999996	EU	£20	1	82654212	Hologram	2.355633	147.830149	65.232274	0.097358
999997	USA	$5	0	59174754	Microprint	1.393764	150.050308	69.273269	0.068363
999998	EU	£10	0	55268089	Watermark	2.026417	142.852137	77.878841	0.081160
999999	EU	£10	0	59464296	Watermark	0.867139	127.645125	72.608513	0.083379

1000000 rows × 9 columns

Data Preprocessing

Feature Cleaning

- Removed monetary symbol from "Denomination" column
- Renamed "Denomination" column to "Value"
- Added a new column for "Denomination"

	Country	Value	Counterfeit	SerialNumber	SecurityFeatures	Weight	Length	Width	Thickness	Denomination
0	USA	100	1	25973198	Hologram	1.731759	130.243185	66.537999	0.098488	USD
1	USA	20	1	95903230	Security Thread	1.002179	152.596364	76.135834	0.094119	USD
2	EU	10	0	82937914	Hologram	2.306713	152.857126	66.772442	0.061393	EUR
3	USA	20	1	23612989	Microprint	1.366965	143.133672	78.377052	0.053114	EUR
4	EU	20	1	56025342	Watermark	1.796075	129.664777	75.916093	0.051438	EUR
...	...	...	...	...	...	...	...	...	...	...
999995	EU	100	1	24436622	Watermark	1.472511	134.888731	75.425943	0.093939	USD
999996	EU	20	1	82654212	Hologram	2.355633	147.830149	65.232274	0.097358	GBP
999997	USA	5	0	59174754	Microprint	1.393764	150.050308	69.273269	0.068363	USD
999998	EU	10	0	55268089	Watermark	2.026417	142.852137	77.878841	0.081160	GBP
999999	EU	10	0	59464296	Watermark	0.867139	127.645125	72.608513	0.083379	GBP

1000000 rows × 10 columns

Encoding

One Hot Encoded Columns

- Country
- Denomination
- Security Features

SecurityFeatures_Hologram	SecurityFeatures_Microprint	SecurityFeatures_Security Thread	SecurityFeatures_Watermark	Country_EU	Country_UK	Country_USA
1	0	0	0	0	0	1
0	0	1	0	0	0	1
1	0	0	0	1	0	0
0	1	0	0	0	0	1
0	0	0	1	1	0	0
...	...	...	...	...	...	...
0	0	0	1	1	0	0
1	0	0	0	1	0	0
0	1	0	0	0	0	1
0	0	0	1	1	0	0
0	0	0	1	1	0	0

Exploratory Data Analysis

Univariate Analysis (Continuous)

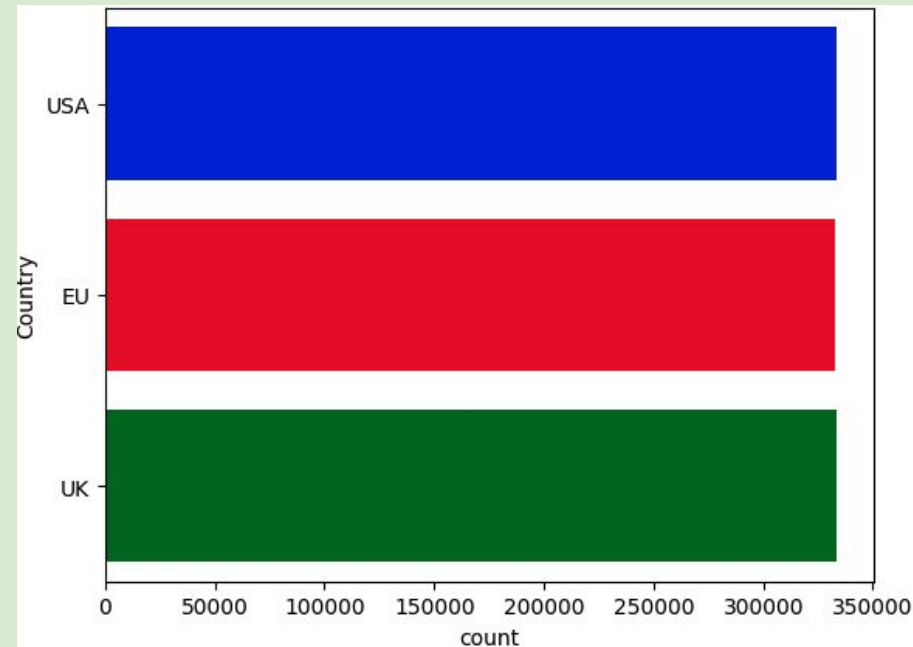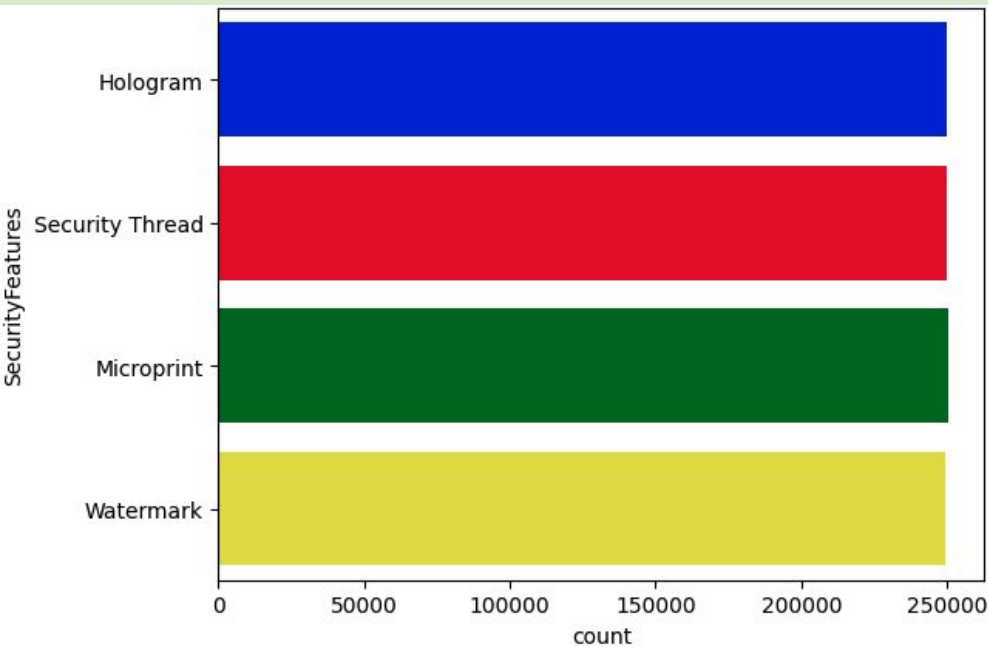|	Value	SerialNumber	Weight	Length	Width	Thickness
count	1000000.000000	1.000000e+06	1000000.000000	1000000.000000	1000000.000000	1000000.000000
mean	21.367511	5.502259e+07	1.649766	140.020542	70.003944	0.074995
std	26.828397	2.598490e+07	0.490712	11.544293	5.772709	0.014442
min	1.000000	1.000015e+07	0.800003	120.000073	60.000005	0.050000
25%	5.000000	3.249784e+07	1.224855	130.034878	64.999762	0.062487
50%	10.000000	5.506594e+07	1.649137	140.032496	70.008440	0.074992
75%	20.000000	7.751115e+07	2.074540	150.022309	75.006372	0.087499
max	100.000000	9.999994e+07	2.499999	159.999961	79.999983	0.100000
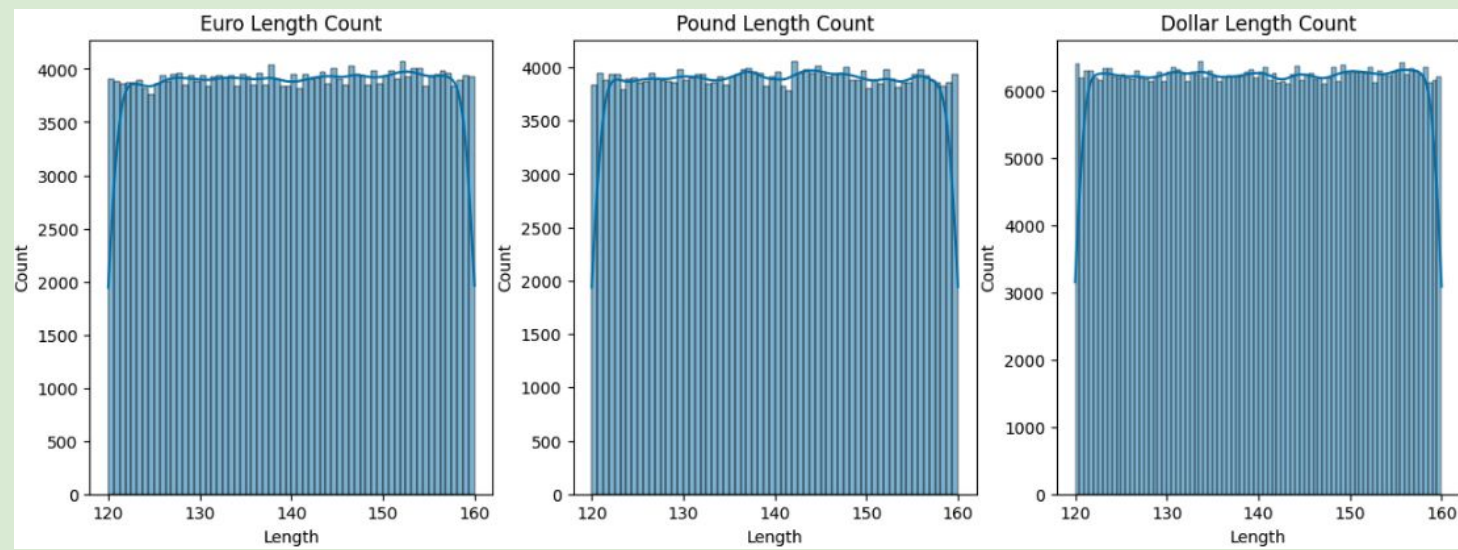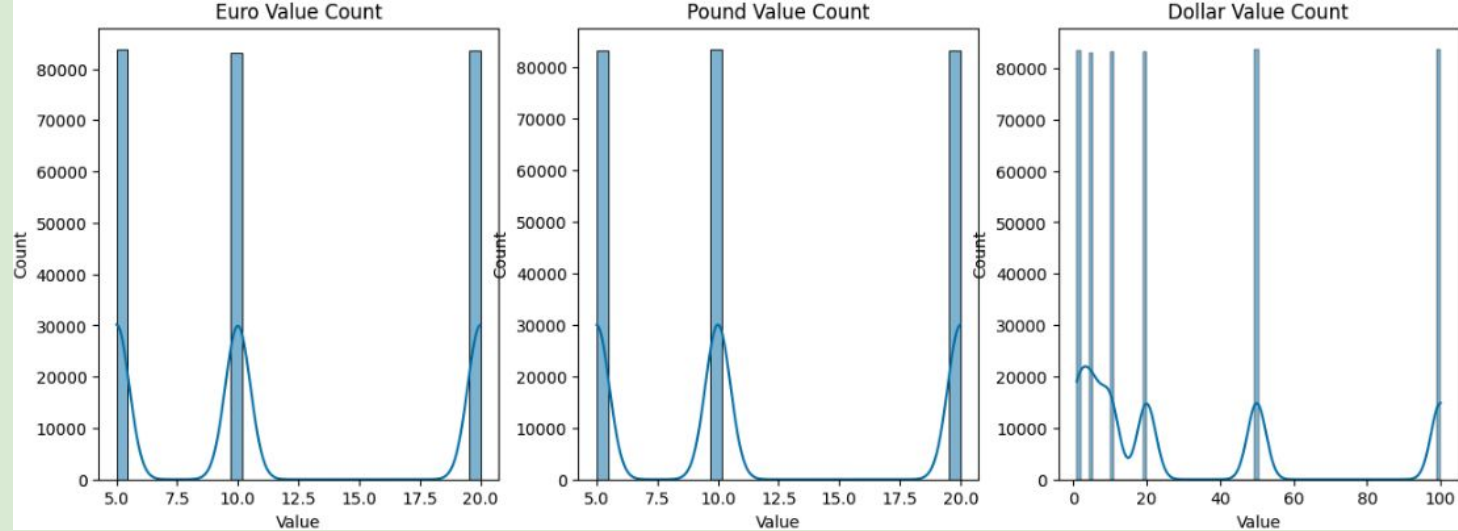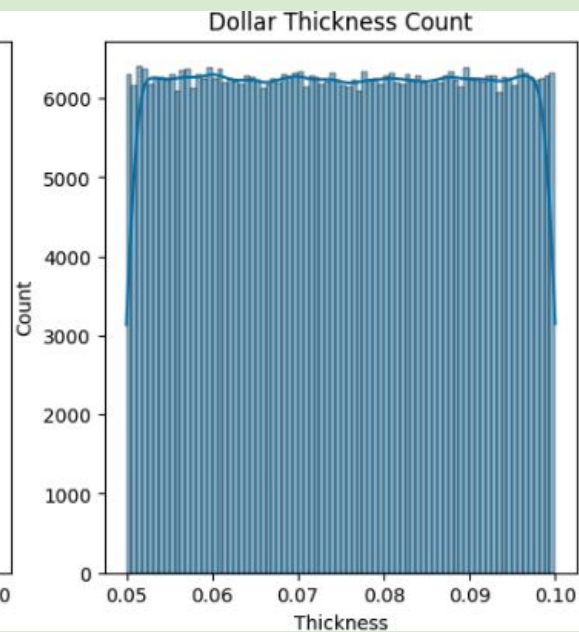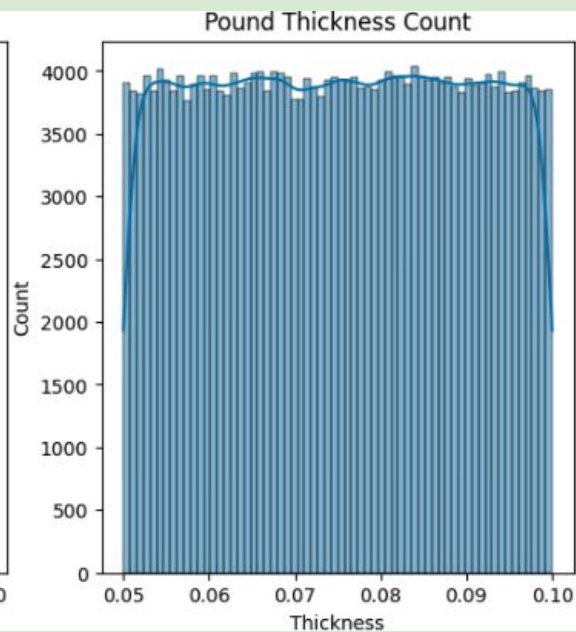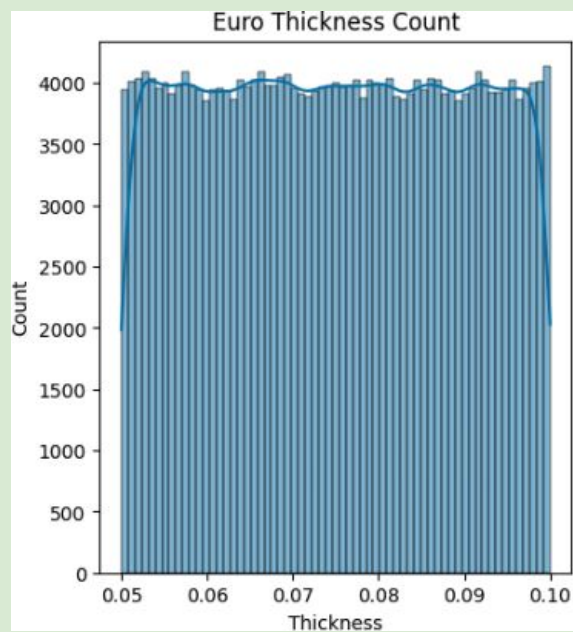
Univariate Analysis (Categorical)

Bivariate Analysis (Continuous)

	Value	SerialNumber	Weight	Length	Width	Thickness
Value	1.000000	0.001795	-0.001224	-0.000296	0.000373	-0.000434
SerialNumber	0.001795	1.000000	0.000688	-0.000604	-0.000613	0.000410
Weight	-0.001224	0.000688	1.000000	-0.000037	0.000589	0.000269
Length	-0.000296	-0.000604	-0.000037	1.000000	-0.000152	-0.001139
Width	0.000373	-0.000613	0.000589	-0.000152	1.000000	0.000497
Thickness	-0.000434	0.000410	0.000269	-0.001139	0.000497	1.000000

Counterfeit	Value	SerialNumber	Weight	Length	Width	Thickness
0	21.398738	5.503707e+07	1.650006	140.037239	70.000000	0.074988
1	21.336208	5.500808e+07	1.649525	140.003805	70.007899	0.075002

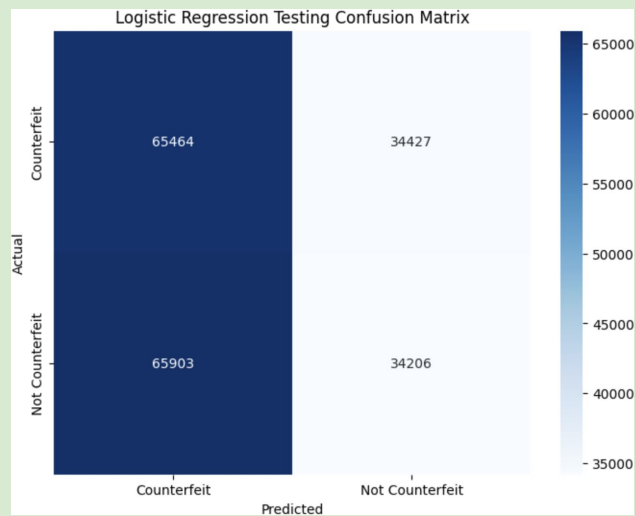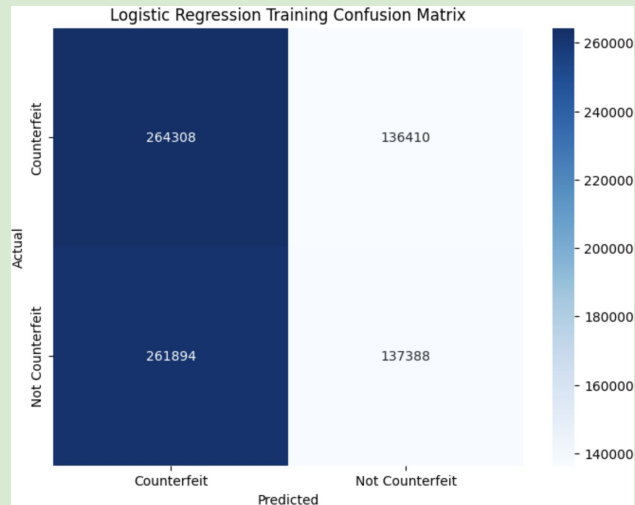# Bivariate Analysis (Categorical)
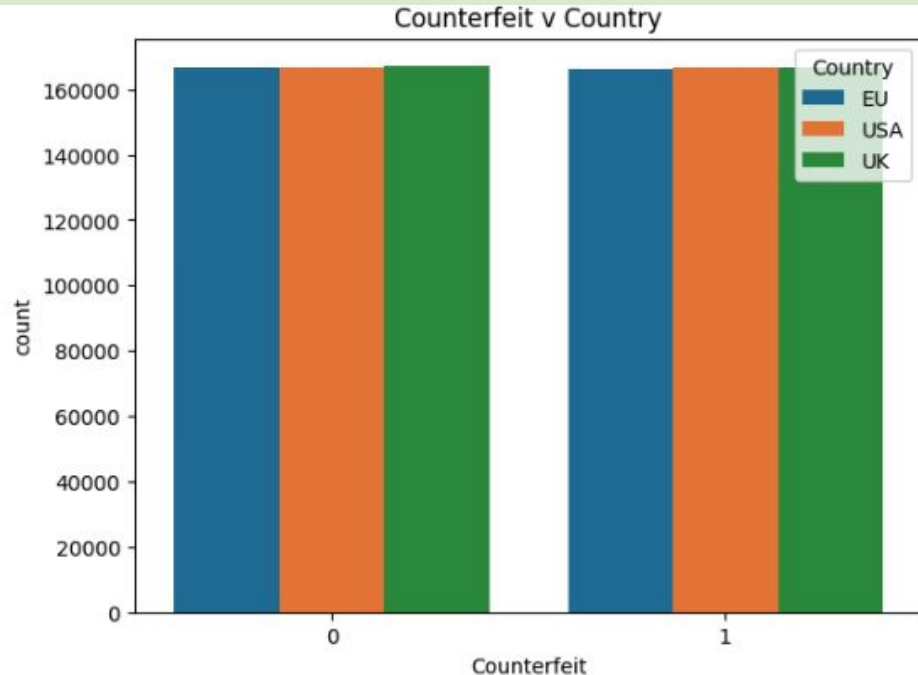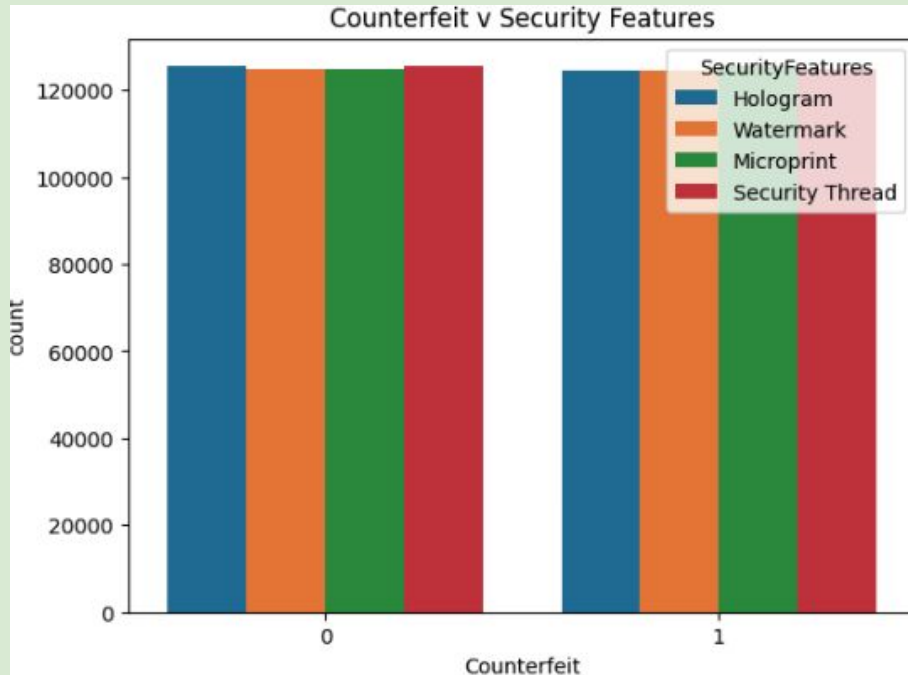
Model Development and Testing

Logistic Regression

- Solver : Stochastic Average Gradient A
- Max Iterations: 1000
- Training Score: ~50.21%
- Testing Score: ~49.84%

Logistic Regression Training Confusion Matrix

Logistic Regression Testing Confusion Matrix

SGD Classifier

- Early Stopping
- Training Score: ~50.07%
- Testing Score: ~50.00%

SGD Classifier Training Confusion Matrix

SGD Classifier Testing Confusion Matrix