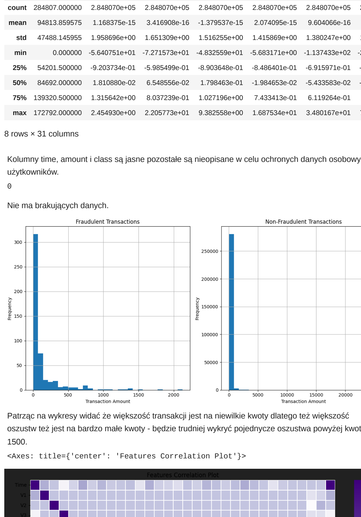
Wczytanie danych i analiza

Out[ ]:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...

5 rows × 31 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 284807 entries, 0 to 284806
Data columns (total 31 columns):
 #   Column  Non-Null Count   Dtype  
---  ------  --------------   -----  
 0   Time    284807 non-null  float64
 1   V1      284807 non-null  float64
 2   V2      284807 non-null  float64
 3   V3      284807 non-null  float64
 4   V4      284807 non-null  float64
 5   V5      284807 non-null  float64
 6   V6      284807 non-null  float64
 7   V7      284807 non-null  float64
 8   V8      284807 non-null  float64
 9   V9      284807 non-null  float64
 10  V10     284807 non-null  float64
 11  V11     284807 non-null  float64
 12  V12     284807 non-null  float64
 13  V13     284807 non-null  float64
 14  V14     284807 non-null  float64
 15  V15     284807 non-null  float64
 16  V16     284807 non-null  float64
 17  V17     284807 non-null  float64
 18  V18     284807 non-null  float64
 19  V19     284807 non-null  float64
 20  V20     284807 non-null  float64
 21  V21     284807 non-null  float64
 22  V22     284807 non-null  float64
 23  V23     284807 non-null  float64
 24  V24     284807 non-null  float64
 25  V25     284807 non-null  float64
 26  V26     284807 non-null  float64
 27  V27     284807 non-null  float64
 28  V28     284807 non-null  float64
 29  Amount  284807 non-null  float64
 30  Class   284807 non-null  int64  
dtypes: float64(30), int64(1)
memory usage: 67.4 MB
```



Widać ze zbiór jest bardzo niezbalansowany. Klasa mniejszościowa nie stanowi nawet 0.2% całego zbioru.

Out[ ]:

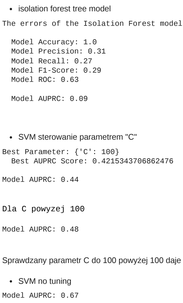
	Time	V1	V2	V3	V4	V5
count	284807.000000	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070
mean	94813.859575	1.168375e-15	3.416908e-16	-1.379537e-15	2.074095e-15	9.604066e-16	1.487311
std	47488.145955	1.958696e+00	1.651309e+00	1.516255e+00	1.415869e+00	1.380247e+00	1.332271
min	0.000000	-5.640751e+01	-7.271573e+01	-4.832559e+01	-5.683171e+00	-1.137433e+02	-2.616051
25%	54201.500000	-9.203734e-01	-5.985499e-01	-8.903648e-01	-8.486401e-01	-6.915971e-01	-7.682956
50%	84692.000000	1.810880e-02	6.548556e-02	1.798463e-01	-1.984653e-02	-5.433583e-02	-2.741871
75%	139320.500000	1.315642e+00	8.037239e-01	1.027196e+00	7.433413e-01	6.119264e-01	3.985649
max	172792.000000	2.454930e+00	2.205773e+01	9.382558e+00	1.687534e+01	3.480167e+01	7.330163

8 rows × 31 columns

Kolumny time, amount i class są jasne pozostałe są nieopisane w celu ochronych danych osobowych użytkowników.

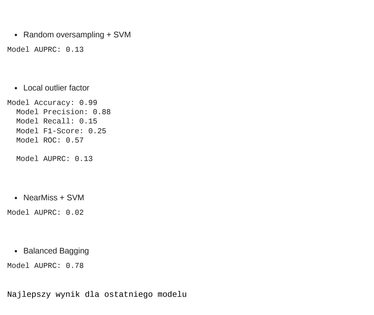
Out[ ]: 0

Nie ma brakujących danych.



Patrząc na wykresy widać że większość transakcji jest na niewielkie kwoty dlatego też większość oszustw też jest na bardzo małe kwoty - będzie trudniej wykryć pojedncze oszustwa powyżej kwoty 1500.

Out[ ]: <Axes: title={'center': 'Features Correlation Plot'}>



Obróbka danych

```
Rozmiar zbioru treningowego: (170884, 29)
  Rozmiar zbioru walidacyjnego: (56961, 29)
  Rozmiar zbioru testowego: (56962, 29)
```

3. Modele

- isolation forest tree model

```
The errors of the Isolation Forest model is  152

  Model Accuracy: 1.0
  Model Precision: 0.31
  Model Recall: 0.27
  Model F1-Score: 0.29
  Model ROC: 0.63

  Model AUPRC: 0.09
```

- SVM sterowanie parametrem "C"

```
Best Parameter: {'C': 100}
  Best AUPRC Score: 0.4215343706862476

Model AUPRC: 0.44
```

Dla C powyżej 100

```
Model AUPRC: 0.48
```

Sprawdzany parametr C do 100 powyżej 100 daje jeszcze odrobinę lepsze wyniki

- SVM no tuning

```
Model AUPRC: 0.67
```

- SMOTE + SVM

```
Model AUPRC: 0.08
```

- Random undersampling + SVM

```
Model AUPRC: 0.08
```

- Random oversampling + SVM

```
Model AUPRC: 0.13
```

- Local outlier factor

```
Model Accuracy: 0.99
  Model Precision: 0.88
  Model Recall: 0.15
  Model F1-Score: 0.25
  Model ROC: 0.57

  Model AUPRC: 0.13
```

- NearMiss + SVM

```
Model AUPRC: 0.02
```

- Balanced Bagging

```
Model AUPRC: 0.78
```

Najlepszy wynik dla ostatniego modelu