ЛАБОРАТОРНАЯ РАБОТА #2

Линейная классификация

Ивченко Матвей Сергеевич
Пермяков Герман Алексеевич
Крушинин Никита Игоревич
М8О-307Б-23

Цели

- Построить, оптимизировать и оценить классификаторы.
- Понять интерпретацию признаков (веса, permutation importance).
- Освоить расширение и отбор признаков (feature tuning).
- Изучить влияние регуляризации, калибровки и гиперпараметров

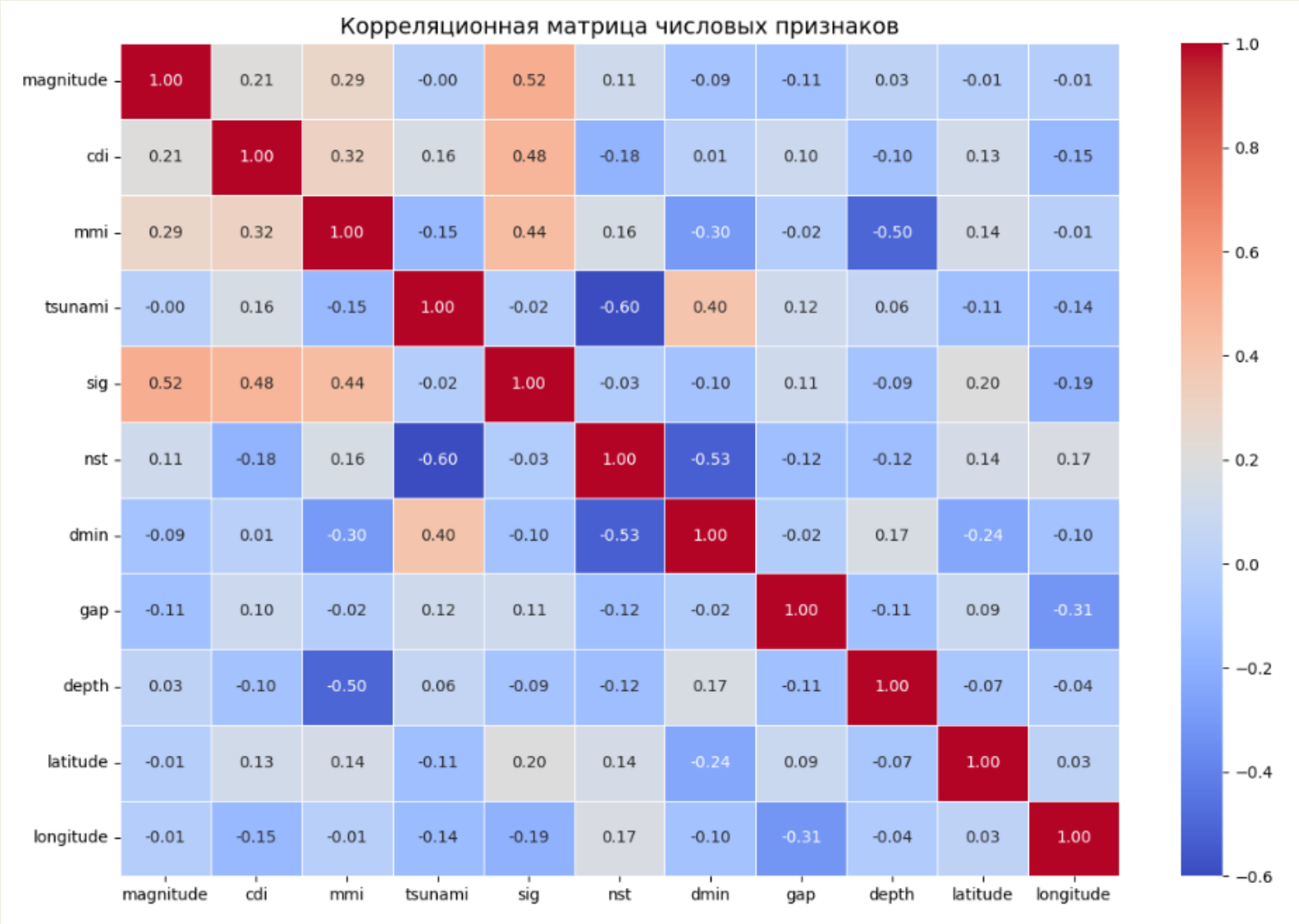
Условия

- Numpy
- Pandas
- scikit-learn
- XGBoost

EDA

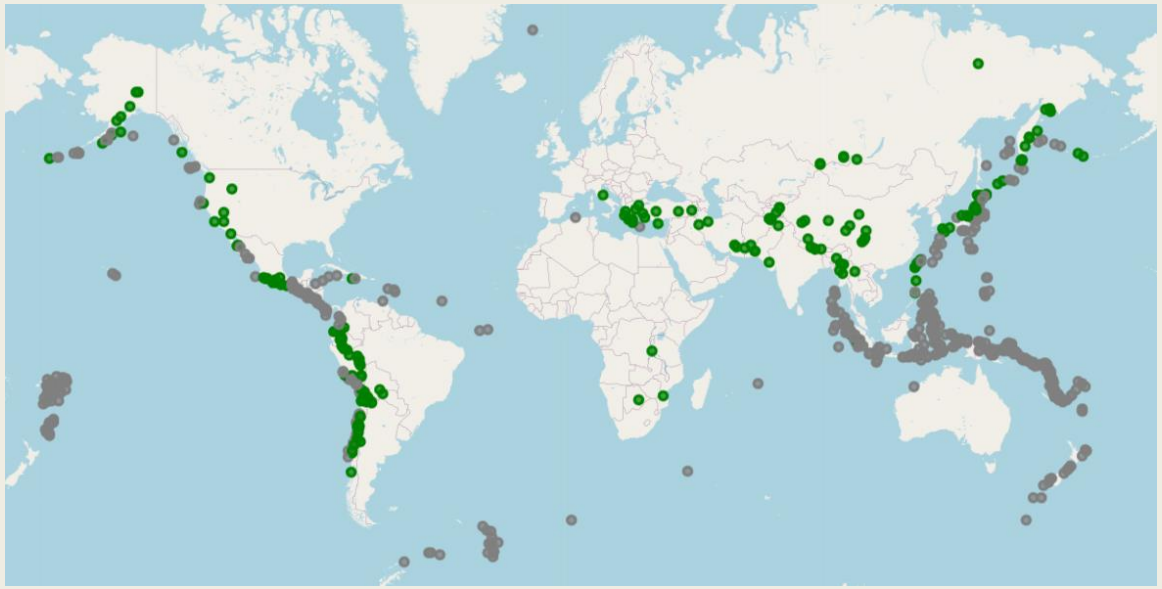
- В датасете в основном нет пропусков, кроме continent, country и alert (таргет)
- Есть категориальные и вещественные, а также геоданные (в общем, тоже категориальные)
- Распределение классов неравномерное (наибольший класс – 325 примеров, наименьший – 12)
- Вещественные признаки имеют преимущественно нормальное и логнормальное распределение

EDA

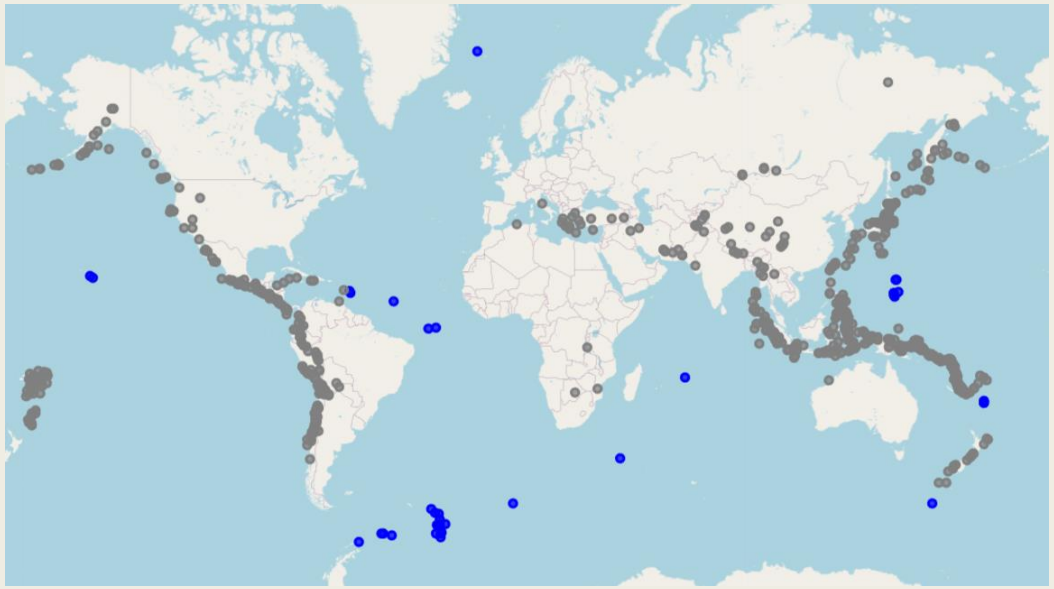

Корреляционная матрица числовых признаков
magnitude 1.00 0.21 0.29 -0.00 0.52 0.11 -0.09 -0.11 0.03 -0.01 -0.01
cdi 0.21 1.00 0.32 0.16 0.48 -0.18 0.01 0.10 -0.10 0.13 -0.15
mmi 0.29 0.32 1.00 -0.15 0.44 0.16 -0.30 -0.02 -0.50 0.14 -0.01
tsunami -0.00 0.16 -0.15 1.00 -0.02 -0.60 0.40 0.12 0.06 -0.11 -0.14
sig 0.52 0.48 0.44 -0.02 1.00 -0.03 -0.10 0.11 -0.09 0.20 -0.19
nst 0.11 -0.18 0.16 -0.60 -0.03 1.00 -0.53 -0.12 -0.12 0.14 0.17
dmin -0.09 0.01 -0.30 0.40 -0.10 -0.53 1.00 -0.02 0.17 -0.24 -0.10
gap -0.11 0.10 -0.02 0.12 0.11 -0.12 -0.02 1.00 -0.11 0.09 -0.31
depth 0.03 -0.10 -0.50 0.06 -0.09 -0.12 0.17 -0.11 1.00 -0.07 -0.04
latitude -0.01 0.13 0.14 -0.11 0.20 0.14 -0.24 0.09 -0.07 1.00 0.03
longitude -0.01 -0.15 -0.01 -0.14 -0.19 0.17 -0.10 -0.31 -0.04 0.03 1.00
magnitude cdi mmi tsunami sig nst dmin gap depth latitude longitude
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4 −0.6

Feature Engineering

- Признаки longitude и latitude были дискретизированы и закодированы модой в каждом дискретном участке.
- Пропущенные значения в continent и country были заполнены с помощью RadiusNearestNeighbor



До заполнения. Серые точки - пропуски



После заполнения. Серые точки – континент, синие - океан

Pseudolabeling

- Так как таргет содержит пропуски, попробуем применить трюк под названием псевдолейблинг (хотя он скорее для нейронок, но всё же)
- Суть в том, чтобы использовать модель, обученную на размеченных данных, для разметки тех объектов, на которых она очень уверена, а потом обучить модель на доразмеченных данных, повторяя итерации, пока качество на отложенной выборке растёт.
- Удалось увеличить precision и recall почти на 0.1

```
--- Iteration 1 ---
Добавлено псевдометок: 97
Accuracy: 0.840
Weighted Precision: 0.721
Weighted Recall: 0.812
--- Iteration 2 ---
Добавлено псевдометок: 12
Accuracy: 0.850
Weighted Precision: 0.808
Weighted Recall: 0.851
--- Iteration 3 ---
Добавлено псевдометок: 4
Accuracy: 0.845
Weighted Precision: 0.795
Weighted Recall: 0.902
--- Iteration 4 ---
Добавлено псевдометок: 2
Accuracy: 0.855
Weighted Precision: 0.798
Weighted Recall: 0.890
--- Iteration 5 ---
Добавлено псевдометок: 2
Accuracy: 0.843
Weighted Precision: 0.806
Weighted Recall: 0.896
```

Обучение моделей

model	accuracy	f1-score	roc_auc
Best CatBoost	0.902	0.767	0.971
CatBoost	0.896	0.758	0.970
XGBoost	0.906	0.757	0.972
Logistic Regression	0.896	0.730	0.973
RandomForest	0.896	0.713	0.972
Boosting	0.895	0.704	0.961
DesicionTree	0.866	0.695	0.810
SVM	0.866	0.511	0.958
KNN	0.851	0.477	0.887