Zaawansowane metody Uczenia maszynowego - Projekt nr 1

Jadwiga Słowik

Wykorzystane narzędzia

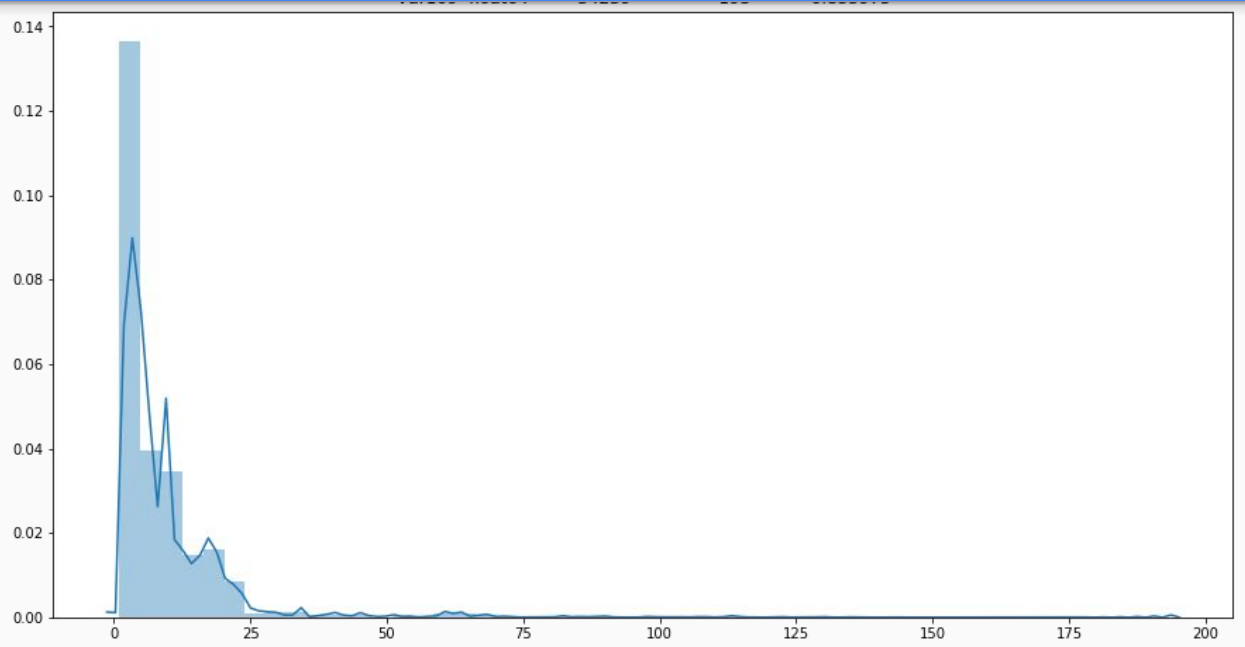
- Python
- Jupyter-notebook
- Scikit-learn
- Categorical_encoders

Analiza danych

- Obliczenie współczynnika korelacji dla wszystkich par kolumn

- Narysowanie wykresów rozkładu/histogramów dla każdej zmiennej

Przykładowy rozkład zmiennej (Var109)



Proces oczyszczania danych

- Usunięcie kolumn, które mają **więcej niż 30% braków danych**
- Obliczenie współczynnika korelacji dla wszystkich par kolumn, a następnie usunięcie zmiennych **skorelowanych >= 0,8**
- Zastąpienie braków danych:
  - Dla zmiennych **ilościowych**: **średnia**
  - Dla zmiennych **jakościowych**: **moda**
- Kodowanie zmiennych kategorycznych: **factor/mean encoding**

Testowane klasyfikatory

- Random forest
  - max_depth=30
- XGBoost
  - learning_rate=0.04
- Naive Bayes
- QDA

Random forest

	fit_time	score_time	test_f1	test_precision	test_recall	train_f1	train_precision	train_recall
0	16.881244	0.367998	0.704641	0.927778	0.568027	1.0	1.0	1.0
1	17.171040	0.370342	0.645435	0.853933	0.518771	1.0	1.0	1.0
2	16.915865	0.353736	0.716102	0.944134	0.576792	1.0	1.0	1.0
3	17.671090	0.357593	0.671053	0.938650	0.522184	1.0	1.0	1.0
4	17.412162	0.373471	0.646018	0.918239	0.498294	1.0	1.0	1.0
5	16.832811	0.360960	0.678038	0.903409	0.542662	1.0	1.0	1.0
6	16.952564	0.352811	0.686071	0.877660	0.563140	1.0	1.0	1.0
7	16.970798	0.445207	0.643478	0.886228	0.505119	1.0	1.0	1.0
8	16.941810	0.373797	0.675269	0.912791	0.535836	1.0	1.0	1.0
9	17.081994	0.355108	0.645161	0.872093	0.511945	1.0	1.0	1.0

XGBoost

	fit_time	score_time	test_f1	test_precision	test_recall	train_f1	train_precision	train_recall
0	27.246294	0.057673	0.744094	0.883178	0.642857	0.747777	0.873354	0.653773
1	27.341921	0.057446	0.690058	0.804545	0.604096	0.755411	0.880424	0.661486
2	27.002709	0.060656	0.752941	0.884793	0.655290	0.746482	0.870268	0.653525
3	26.821100	0.057945	0.718876	0.873171	0.610922	0.751027	0.874559	0.658074
4	26.835882	0.056246	0.701826	0.865000	0.590444	0.752921	0.877016	0.659591
5	26.679966	0.059034	0.741176	0.870968	0.645051	0.750000	0.875126	0.656179
6	27.010275	0.056951	0.698842	0.804444	0.617747	0.753028	0.876636	0.659970
7	26.865850	0.056101	0.705426	0.816143	0.621160	0.751782	0.873933	0.659591
8	26.394756	0.056070	0.724070	0.848624	0.631399	0.747728	0.870968	0.655042
9	26.762205	0.057797	0.678295	0.784753	0.597270	0.754439	0.879798	0.660349

Naive Bayes

	fit_time	score_time	test_f1	test_precision	test_recall	train_f1	train_precision	train_recall
0	0.034720	0.014985	0.178771	0.109339	0.489796	0.177292	0.107705	0.500948
1	0.032970	0.014978	0.176295	0.106782	0.505119	0.177799	0.108860	0.484837
2	0.034006	0.014729	0.214433	0.134251	0.532423	0.182451	0.113991	0.456785
3	0.033607	0.014551	0.171100	0.108379	0.406143	0.186104	0.116982	0.454890
4	0.034936	0.014762	0.181971	0.120805	0.368601	0.177501	0.118740	0.351403
5	0.033266	0.014964	0.168781	0.100658	0.522184	0.178147	0.106471	0.545110
6	0.031932	0.014374	0.182637	0.112520	0.484642	0.179972	0.110032	0.493935
7	0.031862	0.014946	0.173967	0.106175	0.481229	0.180583	0.110489	0.493935
8	0.032212	0.014344	0.172539	0.104952	0.484642	0.180374	0.110276	0.495072
9	0.032542	0.015239	0.189533	0.119536	0.457338	0.184208	0.116064	0.446171

QDA

	fit_time	score_time	test_f1	test_precision	test_recall	train_f1	train_precision	train_recall
0	0.036416	0.014106	0.178771	0.109339	0.489796	0.177292	0.107705	0.500948
1	0.032897	0.014850	0.176295	0.106782	0.505119	0.177799	0.108860	0.484837
2	0.032907	0.014835	0.214433	0.134251	0.532423	0.182451	0.113991	0.456785
3	0.032007	0.014848	0.171100	0.108379	0.406143	0.186104	0.116982	0.454890
4	0.032856	0.015543	0.181971	0.120805	0.368601	0.177501	0.118740	0.351403
5	0.034144	0.014816	0.168781	0.100658	0.522184	0.178147	0.106471	0.545110
6	0.033852	0.015308	0.182637	0.112520	0.484642	0.179972	0.110032	0.493935
7	0.033523	0.014828	0.173967	0.106175	0.481229	0.180583	0.110489	0.493935
8	0.031921	0.015304	0.172539	0.104952	0.484642	0.180374	0.110276	0.495072
9	0.033754	0.015100	0.189533	0.119536	0.457338	0.184208	0.116064	0.446171

Dziękuję za uwagę