ІНФОРМАЦІЙНА СИСТЕМА РЕКОМЕНДУВАННЯ ЦІНИ ВЖИВАНОГО АВТО

Керівник МКР д.т.н., професор
Мокін В. Б.
Розробив студент гр. ІСТ-18м
Лосенко А.В..

м. Вінниця 2019 р.

Мета роботи: систематизувати сучасні методи розвідувального аналізу даних на Python, запропонувати технологію аналізу та передбачення ціни на вживані авто та перевірити її за даними зі США та України.

Задачі:

- провести характеристику аналогічних систем прогнозування ціни;
- здійснити аналіз обраних масивів даних методами розвідувального аналізу даних;
- застосувати методи машинного навчання в розробці системи прогнозування ціни вживаного автомобіля;
- проаналізувати отримані результати роботи застосованих методів, та обрати серед них найбільш ефективні для використання в подальшому.

Пошук та порівняння аналогічних систем

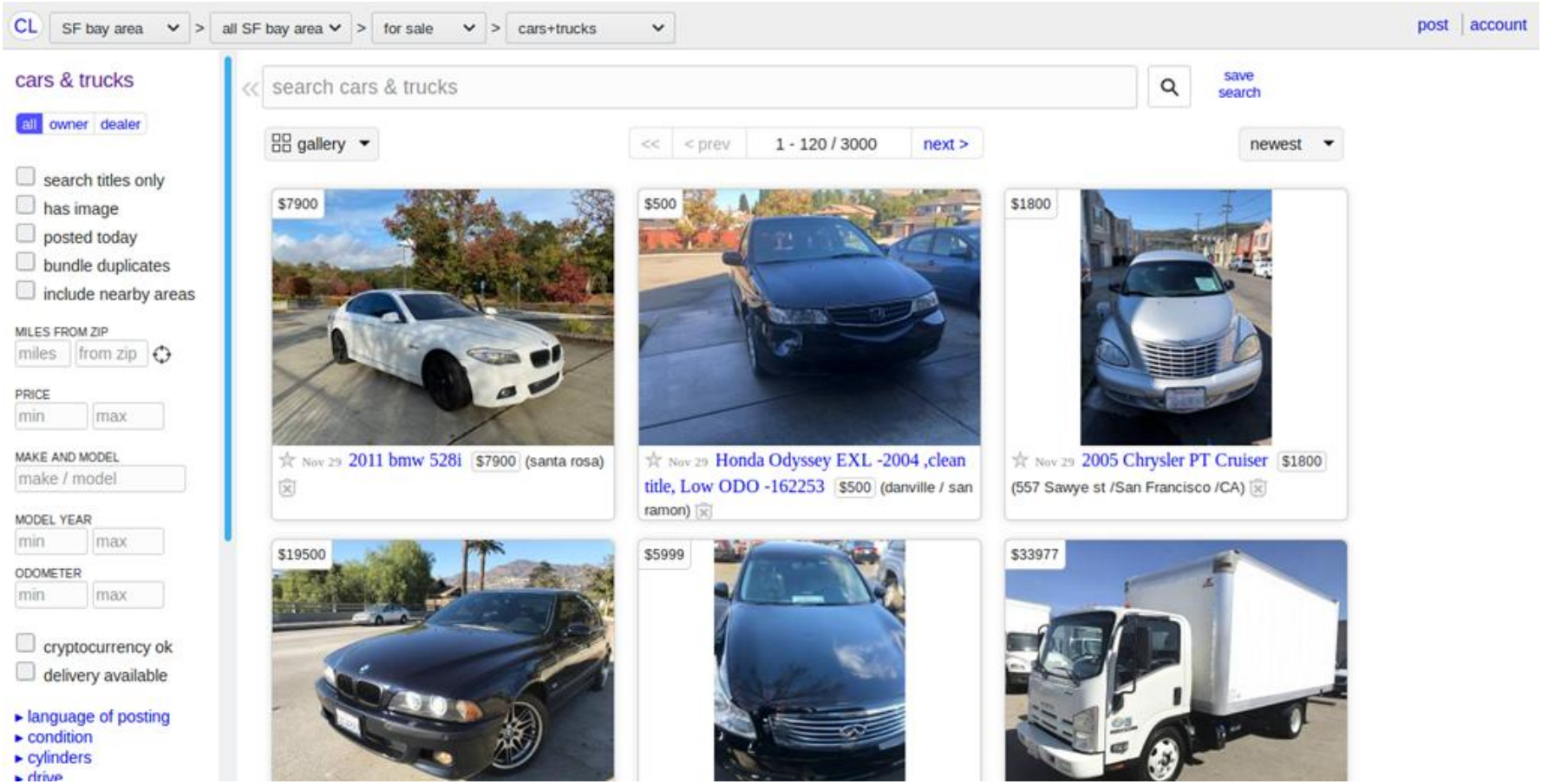
Серед чисельних онлайн-сервісів, які займаються продажем авто на ринку України єдиним сервісом, що надає послуги, наближені до прогнозування ціни автомобіля, є веб-сайт auto.ria.com. Даний веб-сайт надає доступ до власного API, що містить функцію надання середньої ціни. Слід зауважити, що дана функція реалізована лише за допомогою агрегованого запиту до бази даних.

Приклад даних з США

Набір даних, являє собою датасет, розміщений у вільному доступі на платформі Kaggle, що містить дані продажів вживаних автомобілів на сайті craigslist.com. Датасет включає в себе записи про 525839 автомобілі.

	price	year	manufacturer	make	condition	fuel	odometer	transmission	drive	type
0	9000	2,009.00	chevrolet	suburban lt2	good	gas	217,743.00	automatic	rwd	SUV
3	6000	2,002.00	gmc	sierra 1500	good	gas	195,000.00	automatic	4wd	pickup
4	37000	2,012.00	chevrolet	3500	excellent	diesel	178,000.00	automatic	4wd	pickup
12	9700	2,010.00	cadillac	srx luxury collection	good	gas	140,000.00	automatic	fwd	SUV
13	2500	2,001.00	chevrolet	silverado 1500	fair	gas	220,000.00	automatic	rwd	pickup

Інтерфейс сайту craigslist.com

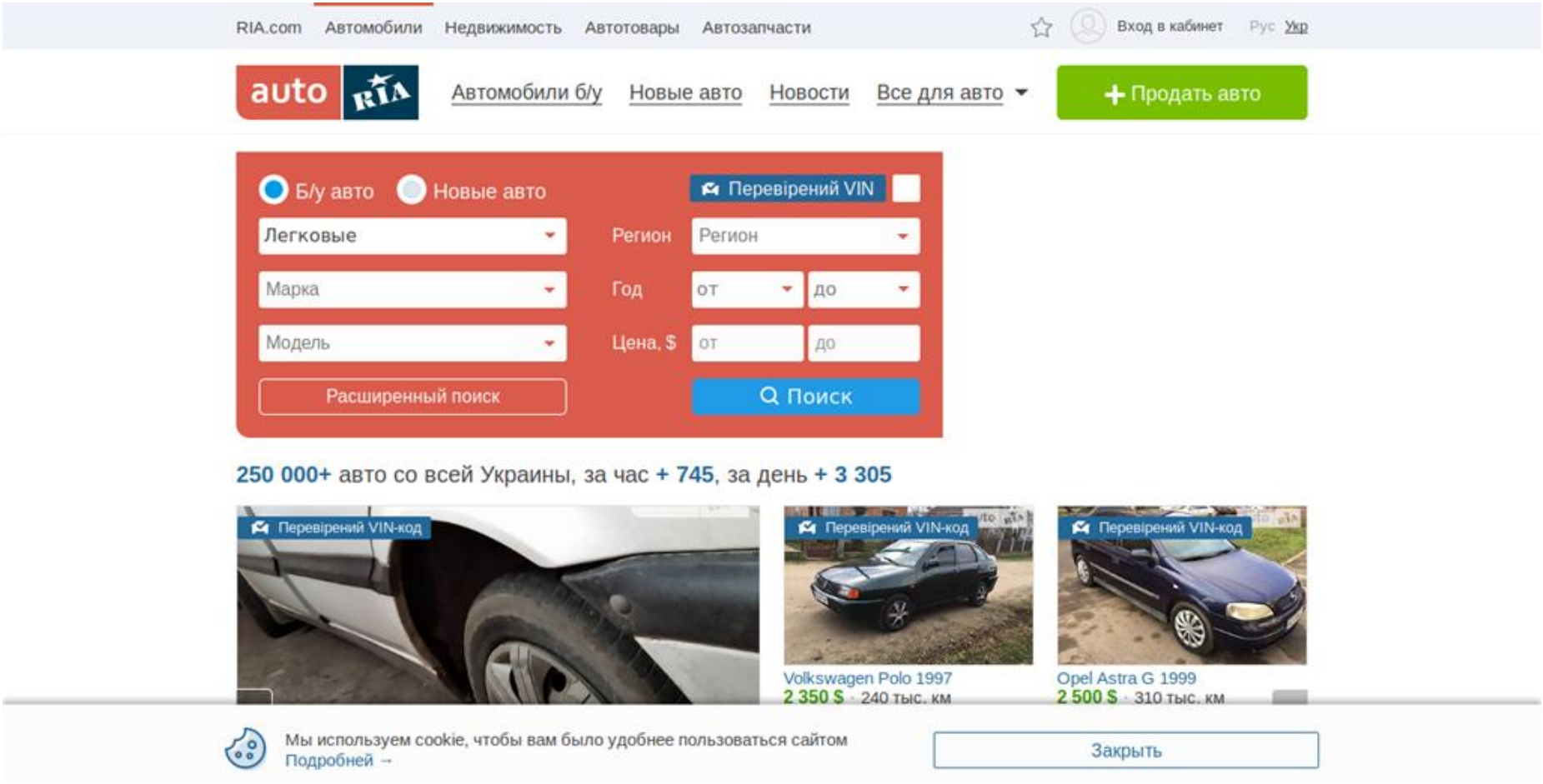


Приклад даних з України

Також для дослідження були використані дані веб-системи медіа-корпорації «RIA» по Україні (по 5432 автомобілях), які були оброблені в межах договору про науково-технічне співробітництво між цією медіа-корпорацією та ВНТУ.

	manufacturer	make	year	price	odometer	transmission	fuel	type	drive
1	ford	focus	2018	12400	6000	automatic	gas	hatchback	fwd
3	lexus	ls 460	2010	21500	164000	automatic	gas	sedan	4wd
4	ford	focus	2013	8600	70000	manual	gas	hatchback	fwd
5	ford	focus	2012	9800	242000	automatic	diesel	wagon	fwd
6	toyota	rav4	2014	22200	94000	automatic	diesel	SUV	4wd

Інтерфейс сайту auto.ria.com



Систематизація сучасних методів EDA

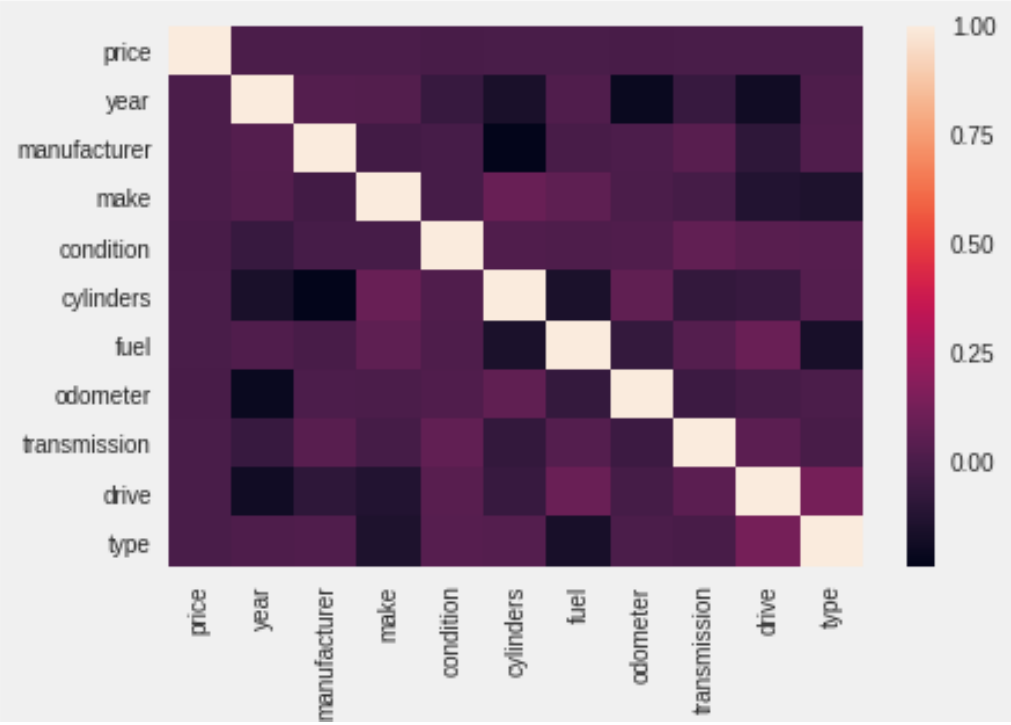
- Базова статистика Describe: по кожній ознаці кількість значень, мінімальне, максимальне, середнє і середньоквадратичне значення та значення квантилів, які можна задавати які завгодно списком (за замовчуванням: 25%, 50%, 75%);
- Бібліотеки Matplotlib та Plotly дозволяють побудувати багато графіків для відображення різних особливостей датасету в цілому та окремих його ознак зокрема, переважно двовимірних, хоча за допомогою Mpl_toolkits.mplot3d («MPL» – це скорочення від «MatPlotLib») дозволяє будувати й тривимірні графіки.
- Бібліотека Seaborn дозволяє будувати різні графіки для аналізу статистичних особливостей даних, наприклад графік для вивчення особливостей взаємозв’язку двох показників, коли по одній осі відкладається одна гістограма, по іншій – інша, а між ними – двовимірна функція їх взаємного розподілу.
- Бібліотека Sklearn дозволяє здійснювати інтелектуальний аналіз та очищення і доповнення даних, наприклад, масштабування і стандартизацію даних, їх імпутинг (інтерполяцію за різними методами по сусідніх даних); побудову різних моделей штучного інтелекту та, за ними – діаграм важливості ознак (як правило, на основі дерев рішень та регресійних моделей), що потім дозволяє з них відібрати найважливіші; кластеризацію та класифікацію даних за різними критеріями тощо.

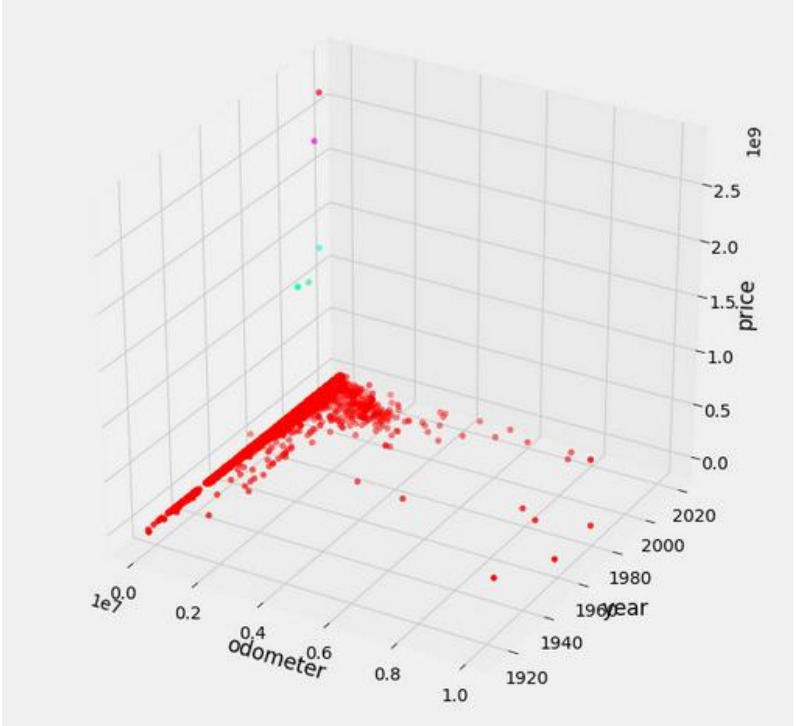
Метод ProfileReport бібліотеки pandas-profiling, який автоматично виконує більшість типових операцій аналізу даних, з яких починається вивчення датасету:

- наводить загальну статистику: кількість ознак (стовпців) загалом і по кожній ознаці зокрема; кількість спостережень (кількість рядків); кількість та відсоток пропущених даних; кількість дублікатів; обсяг пам'яті, яку займає датасет і кожен його рядок в середньому; типи даних ознаки;
- наводить статистику по кожній ознаці окремо (у структурованій гіпертекстовій форми): кількість унікальних значень, пропущених, середнє, мінімальне, максимальне, сума, кількість нульових, квантилі (5%, 25%, 50%, 75%, 95%), середньоквадратичне відхилення, дисперсія, коефіцієнт ексцесу, коефіцієнт асиметрії, графік гістограми, 10 найбільш частих значень, 5 найбільших і 5 найменших значень тощо;
- будує і візуалізує кореляційні матриці з використанням як відомих методів Пірсона, Спірмена і Кендала, так і метода Крамера для категоріальних ознак та найновішого метода φ_K, запропонованого у 2018 р. у роботі [9], для аналізу кореляції значень різного типу (числових, категоріальних та ін.), між якими можуть бути як лінійні, так і нелінійні залежності;
- будує і наводить статистику по пропущених даних у датасеті у вигляді гістограми, матриці, теплової карти та дендрограми;
- наводить 10 перших і 10 останніх рядків датасету.

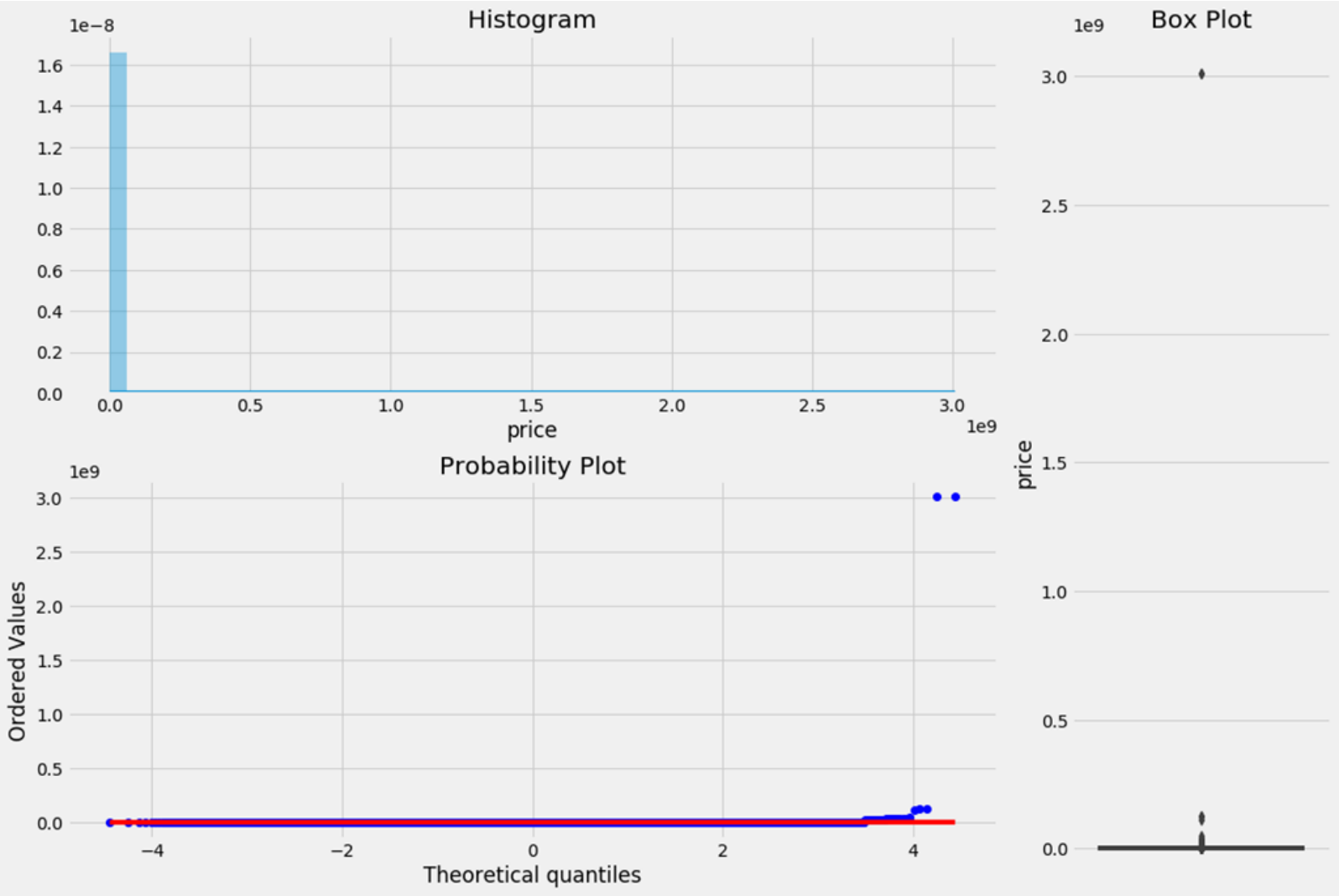
- Бібліотека Scipy.stats містить багато статистичних функцій, у т.ч. закони розподілу та їх аналіз за χ^2-критерієм і критерієм Стьюдента, кореляційний, регресійний, дисперсійний і факторний аналіз, перетворення Бокса-Кокса для перетворення заданого закону розподілу на нормальний та ін.
- Кластеризація даних та виявлення їх прихованих закономірностей тощо за допомогою інтерактивного сервіса https://projector.tensorflow.org.
- Базові бібліотеки Python, наприклад Pandas та NumPy, для роботи з основними типами даних у подібних задачах, теж мають ряд методів для виявлення пропущених, помилкових даних, аналізу їх типів, статистичних даних та їх виправлення за певними алгоритмами.
- Інші бібліотеки від різних розробників, у т.ч. MS (наприклад, lightgbm для побудови дерев рішень методом бустингу та аналізу важливості ознак у них), теж мають багато потужних можливостей, які часто перевищують можливості наведених вище технологій і методів.
- У разі, якщо точність передбачення вийшла низькою, тоді можна збільшити кількість ознак, за рахунок їх генераторів за допомогою бібліотек Featuretools (додає статистичні показники до кожної ознаки з відповідним агрегування по середньому, дисперсії, мінімуму та ін. та/або AutoML (застосовує різні математичні функції до значень ознак: піднесення у степінь, логарифм, тригонометричні функції та ін.), а потім будується діаграма важливості і мало важливі ознаки відкидаються.

Аналіз даних до фільтрування за аномальними ознаками





Розподілення значень ціни автомобіля до застосування фільтрів

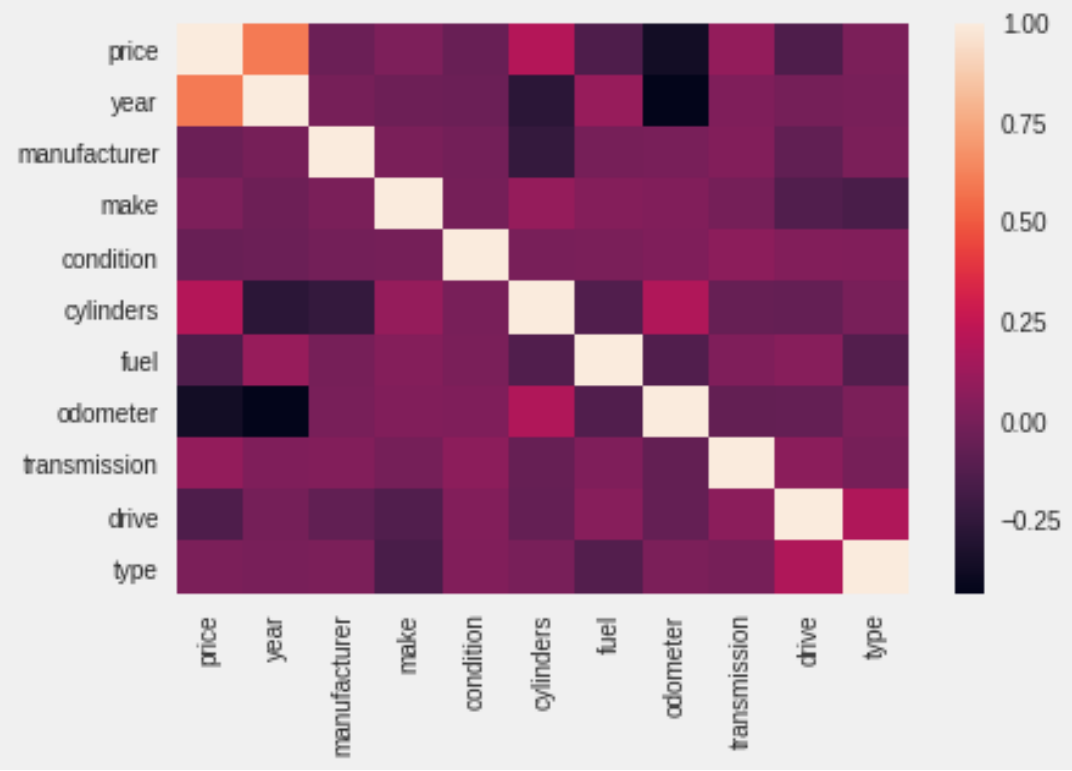


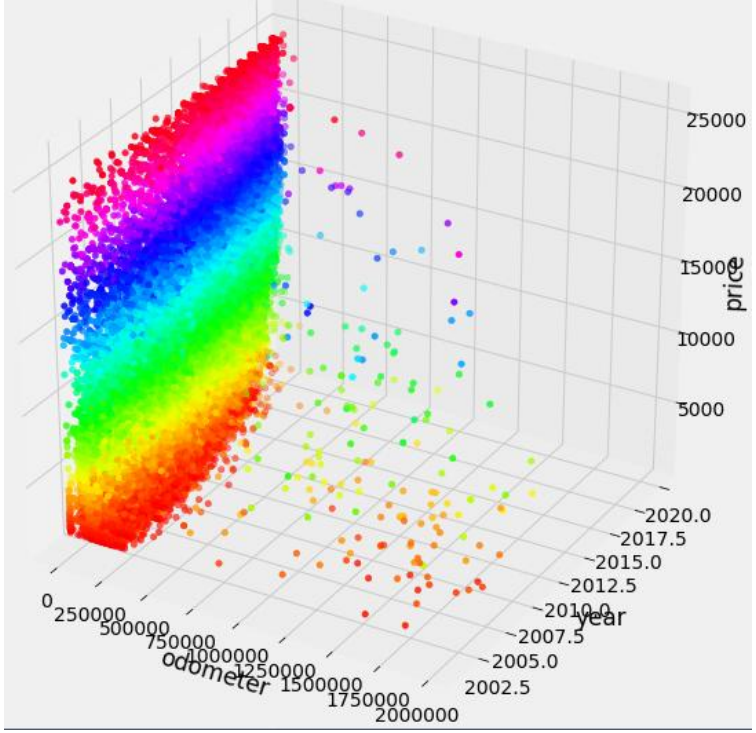
Фільтрування аномальних ознак

	price	year	manufacturer	condition	cylinders	fuel	odometer	transmission
count	1.449030e+05	144903.0	144903.000000	144903.000000	144903.000000	144903.000000	1.449030e+05	144903.000000
mean	6.198665e+04	NaN	18.252176	1.078190	4.658013	1.886310	1.134756e+05	0.117375
std	1.120434e+07	NaN	10.939406	1.163766	1.280567	0.535248	1.235779e+05	0.386070
min	0.000000e+00	1900.0	0.000000	0.000000	0.000000	0.000000	0.000000e+00	0.000000
10%	1.750000e+03	2001.0	7.000000	0.000000	3.000000	2.000000	3.000000e+04	0.000000
50%	8.495000e+03	2010.0	14.000000	0.000000	5.000000	2.000000	1.070000e+05	0.000000
90%	2.520000e+04	2016.0	37.000000	3.000000	6.000000	2.000000	1.900000e+05	0.000000
max	3.009549e+09	2020.0	39.000000	5.000000	7.000000	4.000000	1.000000e+07	2.000000

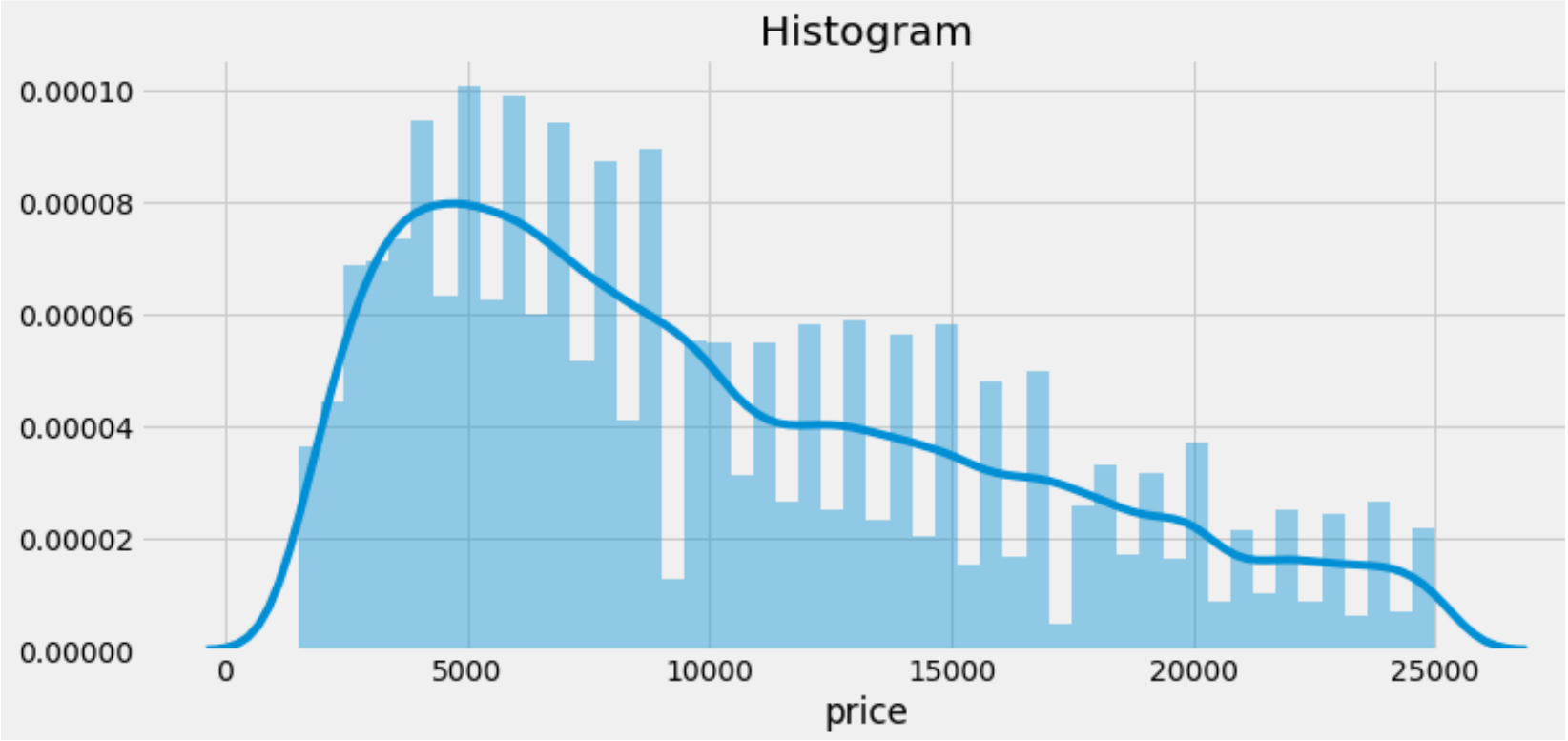
```
#Filter: price (upper (90%) and lower (10%)), year (lower - 10%) and odometer (upper - 90%)
train = train[((train['price'] >= 1500)
               & (train['price'] < 25000)
               & (train['year'] >= 2001)
               & (train['odometer'] < 2000000))]
```

Аналіз даних після фільтрування за аномальними ознаками

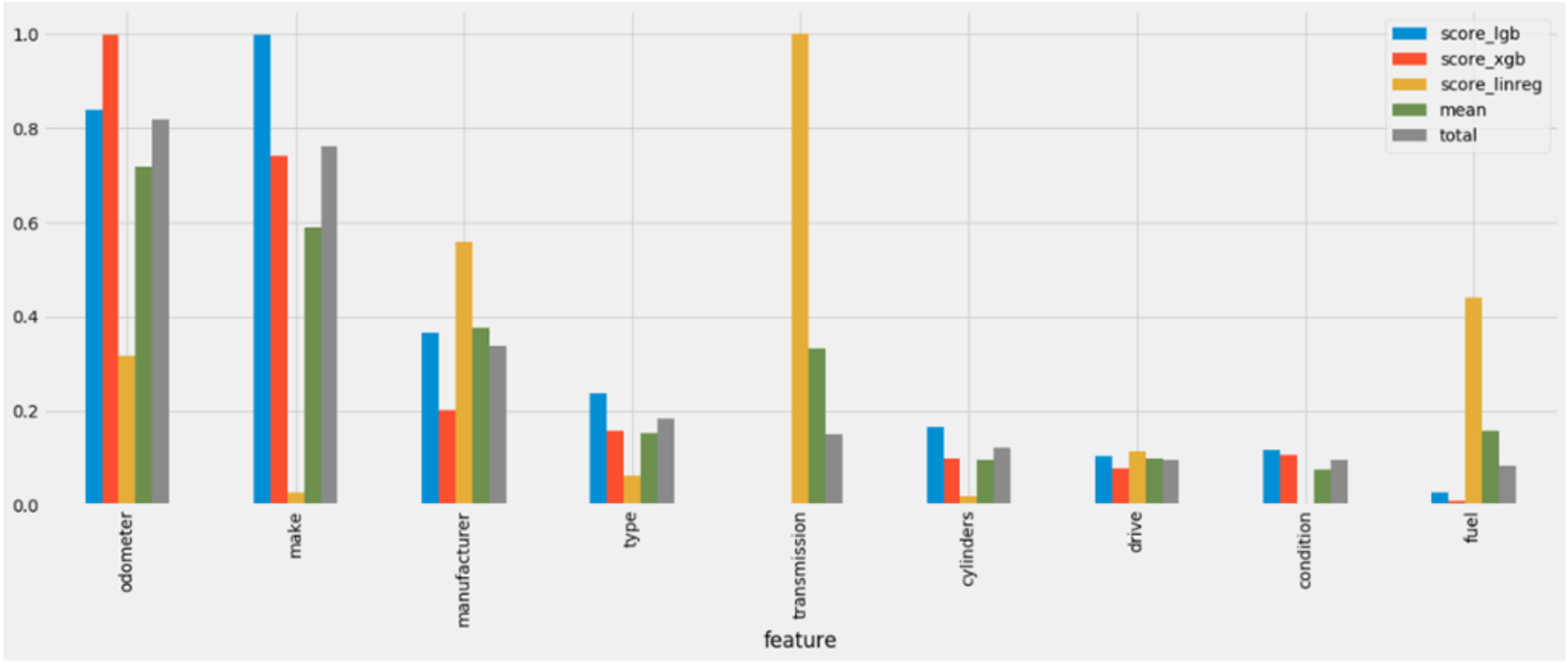




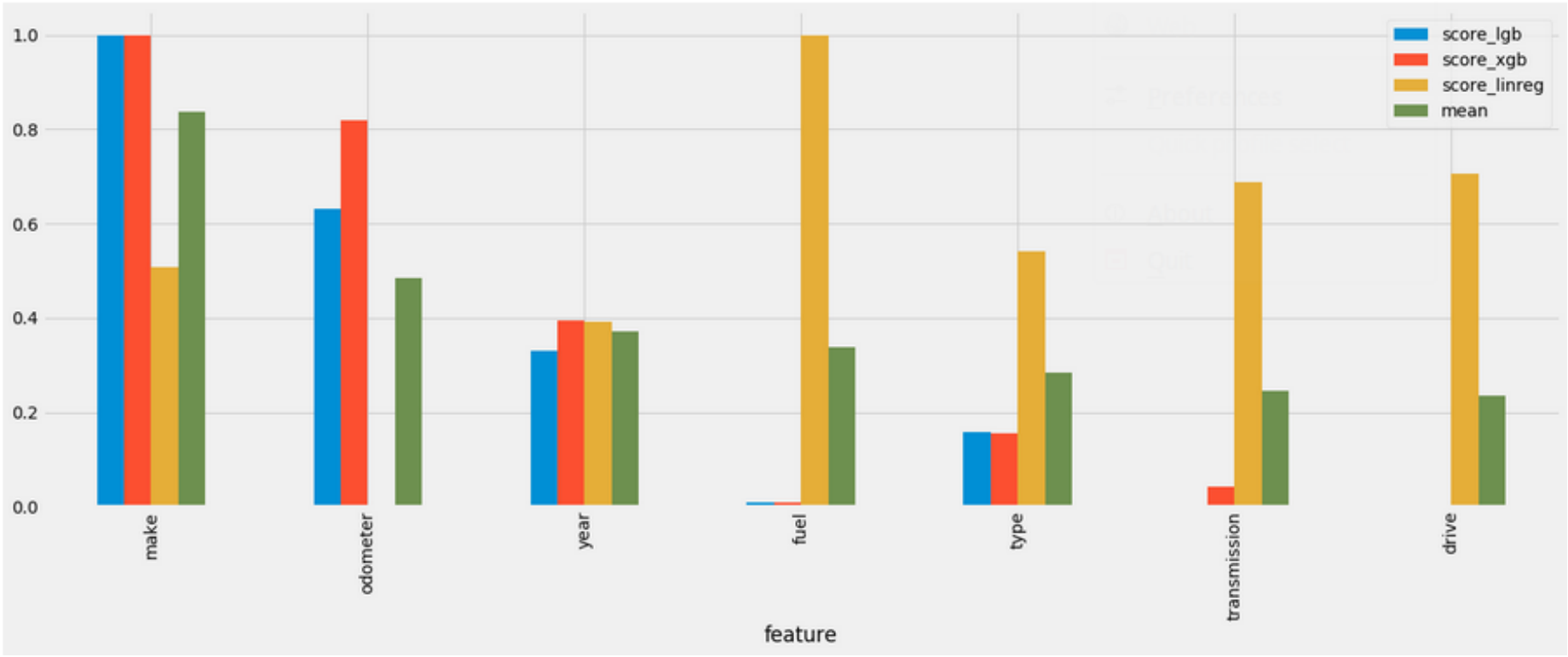
Розподілення значень ціни автомобіля до застосування фільтрів



Графік важливості ознак датасету з США



Графік важливість ознак датасету з України



Вибір методів машинного навчання для вирішення поставленої задачі

- Linear Regression
- Support Vector Machines
- Linear SVR
- MLPRegressor
- Stochastic Gradient Descent
- Decision Tree Regressor
- Random Forest with GridSearchCV
- XGB
- LGBM
- GradientBoostingRegressor with HyperOpt
- RidgeRegressor
- BaggingRegressor
- ExtraTreesRegressor
- AdaBoost Regressor
- VotingRegressor

Порівняння точності моделей з використанням американського датасету

	Model	r2_train	r2_test	d_train	d_test	rmse_train	rmse_test
8	LGBM	91.06	86.12	12.53	14.56	179,382.56	208,609.17
12	ExtraTreesRegressor	99.95	84.19	0.13	13.73	12,790.79	227,775.52
6	Random Forest	97.17	84.11	5.82	14.51	97,206.65	225,093.18
11	BaggingRegressor	97.19	84.10	5.80	14.50	96,970.11	224,957.57
7	XGB	88.33	83.54	14.57	15.84	204,923.27	223,658.23
5	Decision Tree Regressor	99.95	77.11	0.13	16.98	12,789.23	288,694.89
3	MLPRegressor	67.37	68.04	21.65	21.74	297,952.31	297,286.15
9	GradientBoostingRegressor	62.21	62.57	21.81	21.97	296,603.86	297,399.42
14	VotingRegressor	28.82	30.39	29.35	29.36	389,901.74	387,985.65
0	Linear Regression	26.95	28.58	29.45	29.45	389,782.46	387,856.11
10	RidgeRegressor	26.92	28.56	29.45	29.45	389,782.47	387,856.51
4	Stochastic Gradient Decent	22.25	23.99	29.58	29.56	390,531.29	388,697.83
2	Linear SVR	11.13	12.93	29.68	29.74	409,624.83	407,699.79
13	AdaBoostRegressor	-102.47	-102.86	34.83	35.04	414,316.82	416,350.42
1	Support Vector Machines	-1,017.94	-1,020.26	38.45	38.60	508,854.61	510,915.06

Порівняння точності моделей з використанням українського датасету

	Model	r2_train	r2_test	d_train	d_test	rmse_train	rmse_test
4	LGBM	92.51	85.55	31.92	36.05	38,031.21	42,672.70
5	GradientBoostingRegressor	94.14	85.45	6.45	35.26	10,965.31	46,303.93
6	BaggingRegressor	87.26	83.61	14.22	35.21	20,641.95	46,322.15
3	XGB	91.98	83.12	29.38	35.23	35,710.24	42,575.48
2	Random Forest	92.35	81.38	14.33	34.85	20,998.43	46,093.82
1	Decision Tree Regressor	84.16	78.43	1.40	38.18	8,257.61	57,887.24
0	MLPRegressor	60.75	57.71	53.76	53.12	53,822.12	53,586.36
7	ExtraTreesRegressor	27.75	25.63	1.40	35.70	8,257.72	49,729.96

ВИСНОВКИ

Розглянуто проблему прогнозування ціни на вживаний автомобіль, шляхом розвідувального аналізу, з наступним використанням методів машинного навчання. За рахунок розвідувального аналізу були досліджені обрані датасети, здійснене очищення та фільтрування даних з метою покращення якості їх вмісту.

Результати прогнозів моделей були порівняні за кількома критеріями, після чого була обрана модель з найкращим показником точності. Серед обраного набору моделей такою стала модель LGBM, з точністю 86%.

Наукова новизна Дістала подальший розвиток інтелектуальна технологія аналізу та передбачення ціни на вживані авто, за рахунок удосконалення параметрів фільтрів, вибраних під час розвідувального аналізу даних, та підходу щодо вибору оптимальної моделі із багатьох, отриманих у т.ч. з оптимізацією гіперпараметрів, що дозволило підвищити точність передбачення ціни автомобілів за їх параметрами.

Практичне значення одержаних результатів полягає у реалізованому програмному модулі розвідувального аналізу даних та визначення ключових ознак, а також реалізованого програмного модулю рекомендування ціни вживаного авто.

Результати теоретичних та експериментальних досліджень МКР **опубліковані** у вигляді наукової статті у науковому журналі “Вісник ВПІ”, що входить до міжнародної наукометричної бази даних.

Дякую за увагу