https://doi.org/10.31649/1997-9266-2019-147-6-62-72

УДК 004.8:338+629.33

В. Б. Мокін[1]
А. В. Лосенко[1]
М. В. Дратований[1]

ІНТЕЛЕКТУАЛЬНА ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПЕРЕДБАЧЕННЯ ЦІН НА ВЖИВАНІ АВТОМОБІЛІ

[1]Вінницький національний технічний університет

Для вигідного продажу вживаного автомобіля слід керуватись не лише власною оцінкою або оцінкою сторонніх експертів, але й використовувати всі інші придатні для цього ресурси. Такими ресурсами можуть слугувати системи передбачення ціни, які за допомогою загальних ознак того чи іншого автомобіля (як-от виробник автомобіля, модель автомобіля, пробіг, вид палива, тип кузова тощо) здатні прогнозувати можливу ціну автомобіля. Такі системи можуть допомогти під час прийняття рішень не лише пересічним продавцям вживаних авто, а й агентствам, які займаються замовленням та масовим перевезенням вживаних авто з-за кордону. Для вибору ключових ознак та ідентифікації за ними оптимальної структури і параметрів моделей необхідно вибрати релевантні датасети, провести їх розвідувальний аналіз та відбір ознак, побудувати моделі машинного навчання, з яких вибрати оптимальну за певними критеріями. Для побудови інформаційної системи та перевірки працездатності запропонованої інтелектуальної технології вибрано два зіставні датасети по вживаних автомобілях США та України. Здійснено систематизацію методів та бібліотек на Python для проведення розвідувального аналізу даних і сформульовано загальні рекомендації щодо їх застосування для поставленої задачі. Запропоновано загальні принципи інтелектуальної технології, яка апробована на відібраних датасетах. Зокрема, проведено розвідувальний аналіз даних по США та обґрунтовано правило для фільтрування аномальних, а можливо й помилкових, даних. Вибрано множину можливих моделей, здійснено їх тренування та вибрано оптимальну серед них за R^2-критерієм. Здійснено передбачення вартості авто, з точністю 86,1 %. Аналогічна задача розв'язана і для даних по Україні. Досягнуто точність 85,6 %. Це довело працездатність запропонованої технології та дозволило отримати корисні для використання на практиці результати.

Ключові слова: інтелектуальна технологія, розвідувальний аналіз даних, передбачення ціни, вживаний автомобіль, моделі машинного навчання.

Вступ

Розв'язання будь-якої задачі аналізу та передбачення даних з використанням інтелектуальних методів машинного навчання, зазвичай здійснюється у такі етапи [1]:

– очищення даних (виявлення і вилучення помилкових та аномальних даних);
– розвідувальний аналіз даних (англ.: EDA — Exploratory Data Analysis);
– видобування ознак (англ.: FE — Feature Engineering);
– ідентифікація та вибір оптимальної моделі передбачення даних.

Перед багатьма людьми періодично виникає задача придбання власного автомобіля. Причому, як в Україні, так і за кордоном, останнім часом, поширеним є придбання вживаних автомобілів [2]. І покупці, і продавці таких автомобілів намагаються встановити оптимальну для себе ціну за них. Популярним є створення й використання веб-сайтів, де можна підібрати оптимальні для себе параметри та вибрати такий автомобіль. Накопичена вже чимала статистика таких даних, яка робить можливим застосування інтелектуальних методів їх аналізу та передбачення ціни автомобіля за його характеристиками [3]. А це, у свою чергу, дозволяє виявити деякі закономірності щодо формування цієї ціни та краще вибирати маркетингову політику щодо них.

Інтелектуальні технології розв'язання такої задачі досліджувались в усьому світі, оскільки ця проблема є інтернаціональною [4], однак, більшість дослідників недостатньо уваги приділяють етапу розвідувального аналізу даних, що суттєво погіршує точність передбачення та робить отримані результати малокорисними на практиці.

Оскільки найпоширенішою мовою програмування для інтелектуального аналізу і передбачення даних в наш час є мова Python, то автори обмежились аналізом саме можливостей бібліотек на Python, хоча, варто зазначити, що бібліотеки мови *R*, пакетів програм Matlab та ін. є теж потужними, але, по-перше, це — предмет окремих статей, а по-друге, частка дослідників, які досі використовують ці рішення у противагу рішенням на Python, у світі з кожним роком стає дедалі меншою [5].

Мета дослідження — систематизувати сучасні методи розвідувального аналізу даних на Python, запропонувати технологію аналізу та передбачення ціни на вживані автомобілі та перевірити її, використовуючи дані з США та України.

Систематизація технологій аналізу даних на Python у задачах передбачення за табличними даними

Як зазначено вище, розв'язанню поставленої задачі на Python мають передувати етапи очищення даних та розвідувальний аналіз даних. Проведено систематизацію сучасних підходів до автоматизації оброблення даних на цих етапах. Зазвичай, для аналізу табличних даних (у форматі Pandas.DataFrame) використовуються підходи, методи та технології, розглянуті в [6]—[8]:

1. Базова статистика Describe: за кожною ознакою кількість значень, мінімальне, максимальне, середнє і середньоквадратичне значення та значення квантилів, що можна задавати будь-які списком (за замовчуванням: 25 %, 50 %, 75 %);

2. 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 за кожною ознакою 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 найчастіших 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 за пропущеними даними у датасеті у вигляді гістограми, матриці, теплової карти та дендрограми;

– наводить 10 перших і 10 останніх рядків датасету.

3. 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.

4. 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.

5. Бібліотека Sklearn дозволяє здійснювати інтелектуальний аналіз та очищення і доповнення даних, наприклад, масштабування і стандартизацію даних, їх імпутинг (інтерполяцію різними методами за сусідніми даними); 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.

6. 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.

7. Ще більше можливостей дають інтерактивні динамічні технології: інтерактивні графіки за допомогою методів бібліотеки Bokeh.

8. Кластеризація даних та виявлення їх прихованих закономірностей тощо за допомогою інтерактивного сервіса https://projector.tensorflow.org.

9. Існують бібліотеки для роботи з просторовими даними Google Maps (kml-формат та ін.), ArcGIS (shp-формат та ін.) та ін., наприклад, для виведення даних просто на карту і проведення їх просторового аналізу та виявлення відповідних закономірностей.

10. Базові бібліотеки Python, наприклад Pandas та NumPy, для роботи з основними типами даних у подібних задачах, теж мають низку методів для виявлення пропущених, помилкових даних, аналізу їх типів, статистичних даних та їх виправлення за певними алгоритмами.

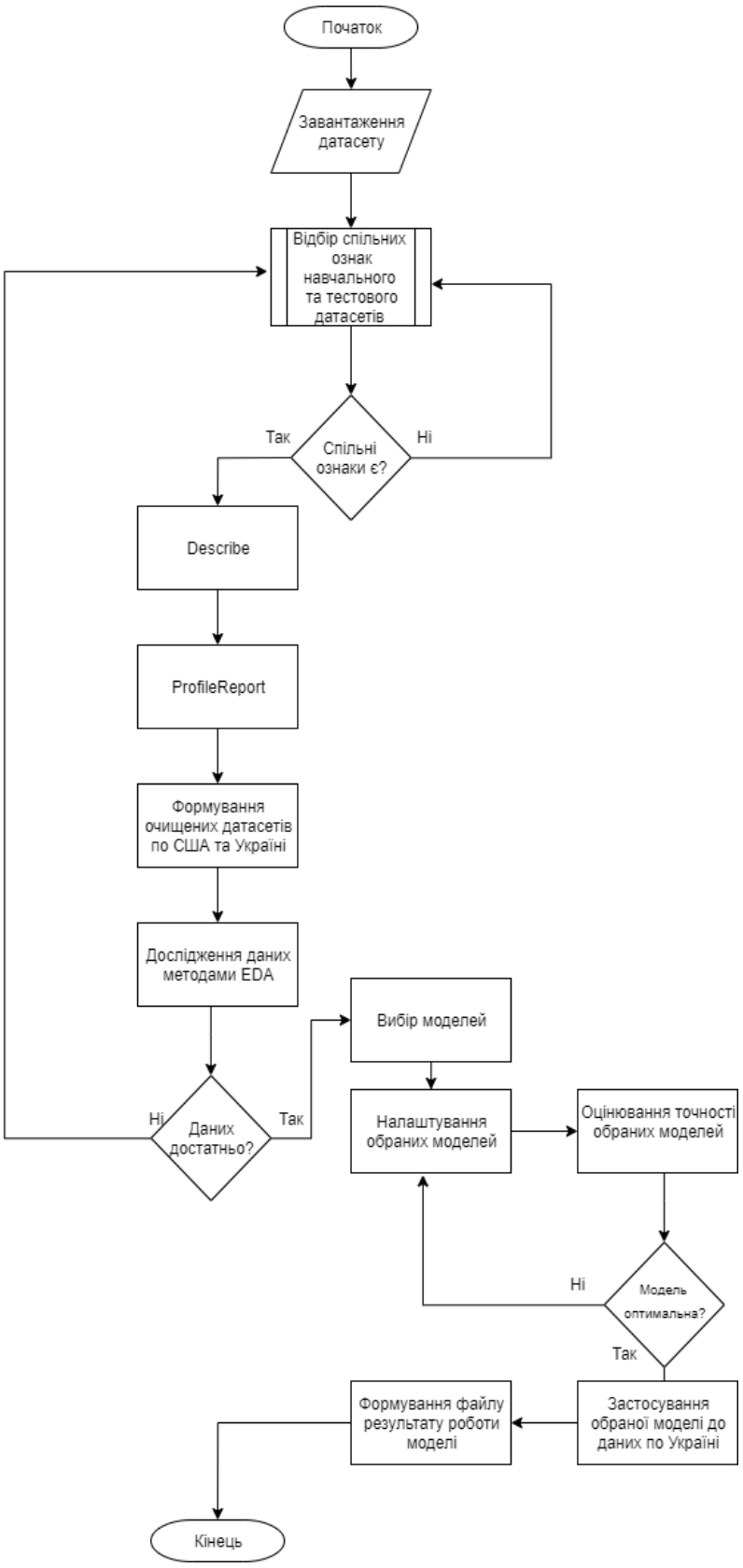


Рис. 1. Блок-схема алгоритму запропонованої інтелектуальної технології

11. 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 вищенаведених технологій і методів.

12. У разі, якщо точність передбачення низька, тоді можна збільшити кількість ознак, за рахунок їх генераторів за допомогою бібліотек Featuretools (додає статистичні показники до кожної ознаки з відповідним агрегуванням по середньому, дисперсії, мінімуму та ін. та/або AutoML (застосовує різні математичні функції до значень ознак: піднесення у степінь, логарифм, тригонометричні функції тощо), а потім будується діаграма важливості і маловажливі ознаки відкидаються.

В цій задачі можна використовувати усі ці технології, але, враховуючи поставлену мету та досвід у сфері розв'язання подібних задач, автори пропонують використовувати таку технологію аналізу та передбачення ціни на вживані авто на основі Python:

1. Вибрати датасети, видобути основні ознаки (FE), спільні для обох датасетів, та попередньо очистити дані від помилкових і відсутніх;

2. Провести розвідувальний аналіз даних та остаточний вибір ознак і даних, за якими варто здійснювати їх моделювання і передбачення: побудувати базову статистику та оцінити різні значення квантилів: 5%, 10%, 25%, 50%, 75%, 90%, 95% — відібрати такі параметри фільтру, за яких основна кількість даних залишиться в основній вибірці, але аномальні зна-

чення (які суттєво відрізняються від інших), будуть видалені; здійснити EDA з використанням pandas-profiling, в якому, в першу чергу, проаналізувати кількість пропущених даних, найбільші і найменші значення, гістограму значень та кореляційну матрицю – на основі цього, розробити остаточні правила для фільтрування помилкових, пропущених та аномальних даних; побудувати дво- і тривимірні графіки, які дозволять уточнити по яких саме ознаках варто робити фільтрування аномальних даних;

3. Вибір і налаштування (навчання) моделей та їх застосування для передбачення даних: натренувати моделі та вибрати серед них найкращу; якщо її точність буде нижчою 70 %, тоді спробувати розширити кількість ознак з використанням бібліотек Featuretools та AutoML з подальшим відкиданням мало важливих за діаграмою важливості, побудованою за методом LGBM.

Блок-схема алгоритму запропонованої інтелектуальної технології аналізу та передбачення ціни на вживані авто показана на рис. 1.

Вибір датасетів, видобування основних ознак та попереднє очищення даних

Для проведення дослідження пропонуємо використовувати такі дані:

– по США (по 525839 автомобілях) із датасету на базі платформи Kaggle з ліцензією «CC0: Public Domain», тобто без обмежень на копіювання і використання (https://www.kaggle.com/austinreese/craigslist–carstrucks–data);

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

Дані по США містять такі ознаки (рис. 2) [10]:

– модель автомобіля (“make”);
– виробник автомобіля (“manufacturer”);
– стан автомобіля (“condition”): "good", "excellent", "like new", "salvage", "fair", "new";
– рік випуску (“year”);
– вид палива (“fuel”): "gas", "diesel", "hybrid", "electric", "other";
– пробіг автомобіля, км (“odometer”): 25000, 120000, 30000...;
– трансмісія (“transmission”): "automatic", "manual", "other";
– привід автомобіля (“drive”): "fwd", "rwd", "4wd";
– тип кузова (“type”): "coupe", "sedan", "wagon", "hatchback", "pickup", "SUV", "truck", "minivan", "other", "van", "convertible", "bus", "offroad".

	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

Рис. 2. Приклад параметрів автомобілів у датасеті щодо США [10]

Дані по Україні медіа-корпорації «RIA» містять такі ознаки (http://auto.ria.com):

– виробник автомобіля (“manufacturer”);
– модель автомобіля (“make”);
– пробіг автомобіля, км (“odometer”);
– рік випуску (“year”);
– трансмісія (“transmission”);
– привід (“drive”);
– вид палива (“fuel”);
– тип кузова (“type”).

Порівняння даних по США та Україні показує, що спільними для них є такі ознаки: виробник автомобіля, модель автомобіля, пробіг, рік випуску, трансмісія, вид палива, тип кузова, привід, тобто варто зосередитись на передбаченні ціни авто саме за цим набором ознак.

Здійснимо очищення даних за запропонованим вище алгоритмом. Послідовне використання методів Describe та методу ProfileReport бібліотеки pandas-profiling під назвою до даних по США дозволило з’ясувати таке: є 48708 авто з ціною $0, тому їх видалено. Крім того, 315307 даних містять відсутні значеннях у певних ознаках, що ускладнює їх передбачення. Звичайно, можна було спробувати замінити пропущені значення по певних ознаках (показниках) на середні по певних класах, як це часто роблять, але тоді це будуть вже не первинні дані, а тому прийнято рішення

цього не робити. Після очищення даних отримано датасет з 161824 авто з повними наборами даних і ненульовою ціною. Стосовно українських даних застосовано аналогічний фільтр та отримано дані щодо 2347 авто. Спробуємо застосувати запропоновану технологію аналізу та передбачення ціни на вживані авто і до даних по США, і до даних по Україні та порівняємо результати.

Розвідувальний аналіз даних та остаточний вибір ознак і даних про вживані авто

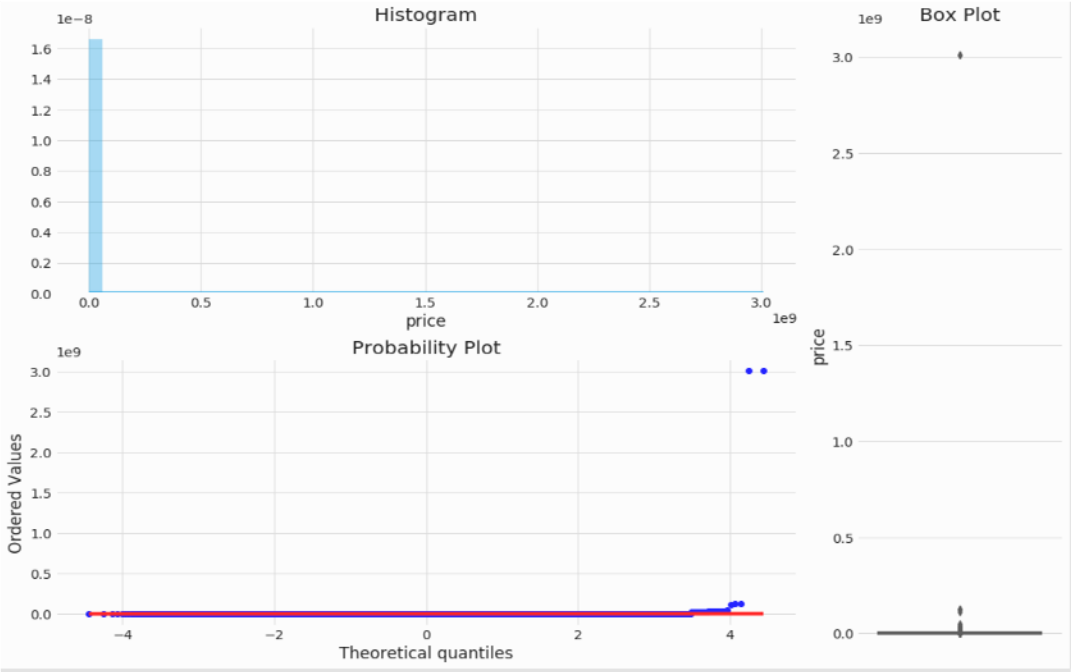
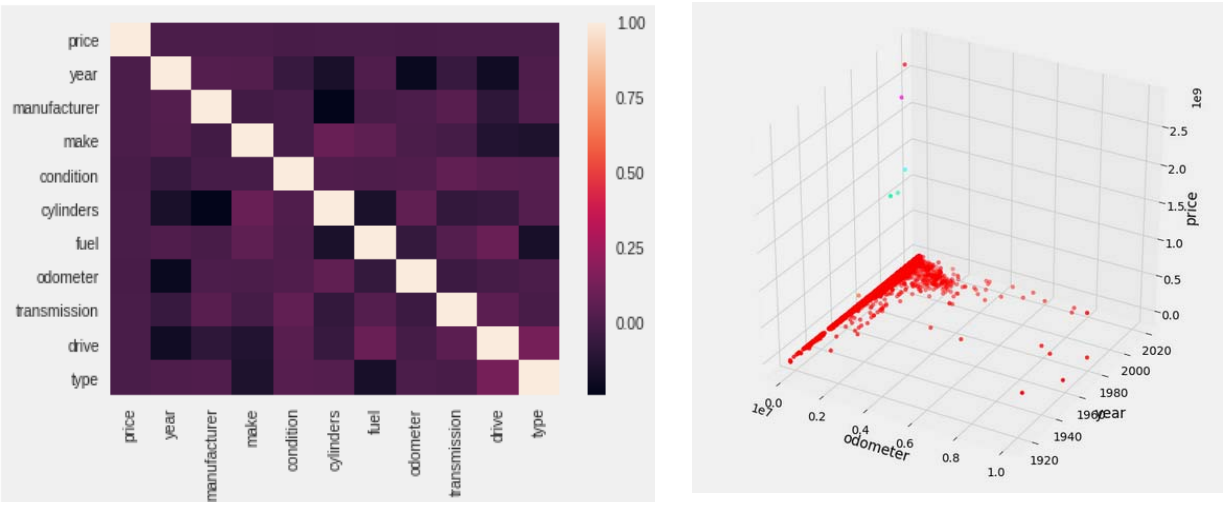


Рис. 3. Перевірка на аномальні дані методами Matplotlib, Pandas та Seaborn

Після очищення даних, як пропонувалось вище, проведемо розвідувальний аналіз даних. Для визначення аномальних даних побудуємо гістограми ознак датасету, для прикладу візьмемо ціну автомобіля (рис. 3).

Оскільки невідомо чи потрібно враховувати аномалії, слід побудувати кореляційну матрицю (рис. 4а), з якої видно, що дані погано корелюють, а на рис. 4б видно одну з можливих причин цього — помилкові або аномальні значення і по пробігу (10 млн км), і по вартості (2,5 млрд дол. — явна помилка), і по року випуску (1920 р.), отже потрібно побудувати правило для фільтрування таких даних, оскільки, наприклад, ціна авто за 2015 рік з порівняно невеликим пробігом явно формується не так, як ціна на авто 1920 року чи з пробігом в 10 млн км.



а) б)

Рис. 4. Розвідувальний аналіз американського набору даних по вживаних автомобілях:
а) кореляційна матриця ознак; б) графік пробігу, року випуску і ціни автомобілів

Відповідно до наведених вище рекомендацій застосовано метод Describe для різних значень квантилів. Найвдалішим вважаємо варіант (10 %, 50 %, 90 %), показаний на рис. 5.

	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

Рис. 5. Значення квантилів для ключових ознак американського набору даних

Добре видно, як суттєво відрізняється мінімальний рік випуску від року з квантилем 10 %, максимальний пробіг від пробігу з квантилем 90 %, мінімальна ціна від ціни з квантилем 10% та максимальна ціна від ціни з квантилем 90 %. Це дозволило визначити правило для фільтрування аномальних значень, подане у вигляді коду на Python, на рис. 6.

```
#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))]
```

Рис. 6. Приклад коду на Python застосування фільтрів за верхньою та нижньою межею значень квантилів 10 % і 90 % по ряду ознак для американського набору даних

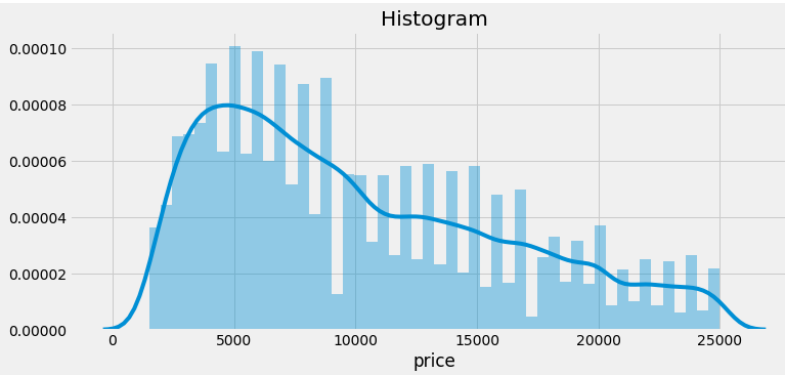
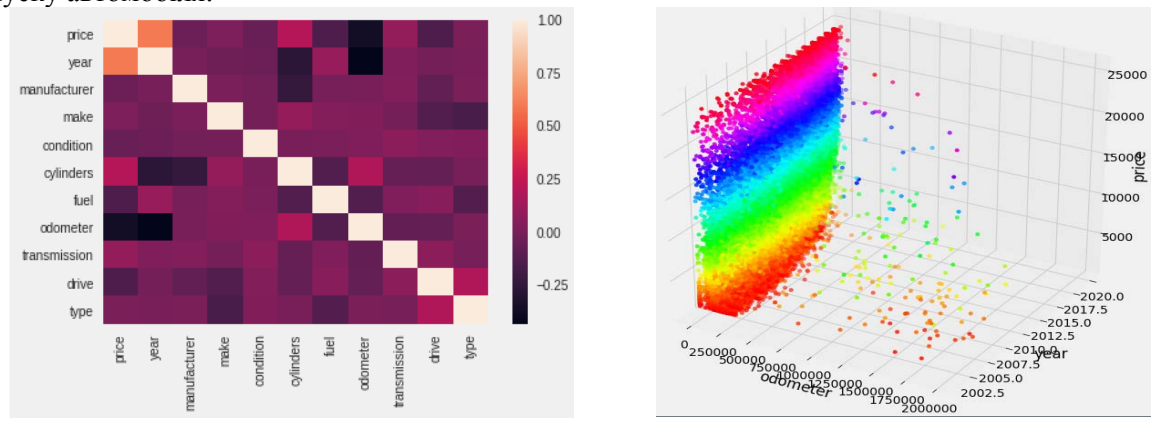


Рис. 7. Розподілення даних за ціною автомобіля на основі американського набору даних

За правилом з рис. 6 здійснено фільтрування набору даних по США, яке ще зменшило датасет до 115186 авто, а по Україні — до 2128 авто.

В американському наборі даних міститься 4290 унікальних значень ціни автомобіля, гістограма для якої подана на графіку на рис. 7.

На рис. 8 подано кореляційну матрицю відфільтрованого датасету по США та графік залежності значень пробігу, ціни та року випуску автомобіля:



а) б)

Рис. 8. Розвідувальний аналіз американського набору даних по вживаних авто після фільтрування за правилом з рис. 6: а) кореляційна матриця ознак; б) графік пробігу, року випуску і ціни автомобілів

За допомогою моделей лінійної регресії, LGBM та XGBoost побудовані діаграми важливості ознак для американського (рис. 9) українського набору даних (рис. 10), з якої видно, що варто будувати модель з використанням таких найважливіших ознак, як: пробіг, виробник автомобіля, модель автомобіля, рік випуску, тип кузова, кількість циліндрів, привід, стан, вид палива.

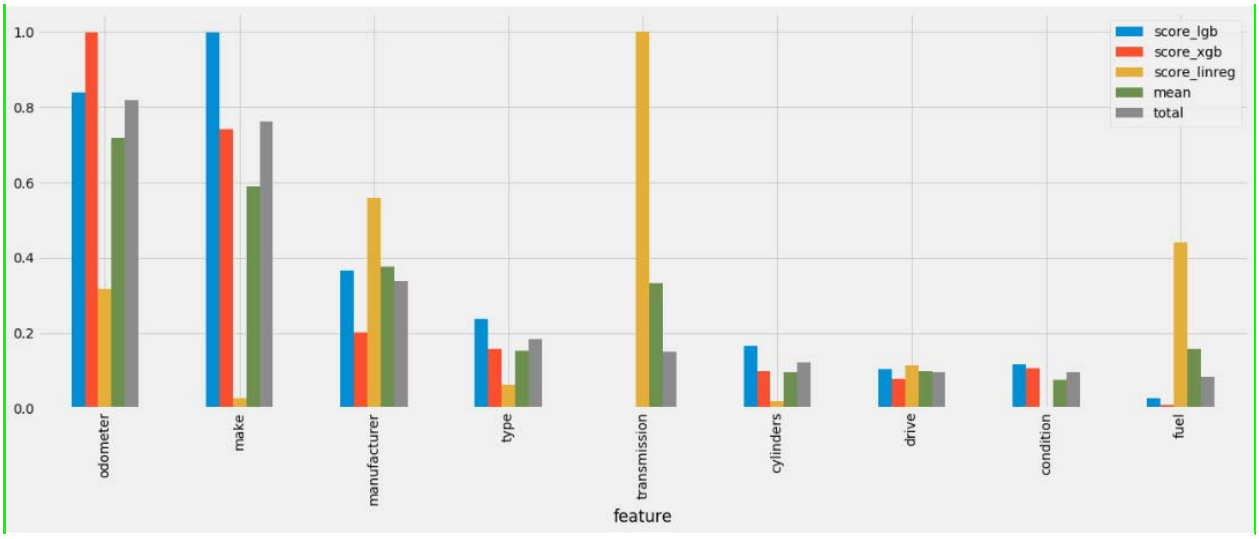


Рис. 9. Діаграма важливості ознак американського набору даних

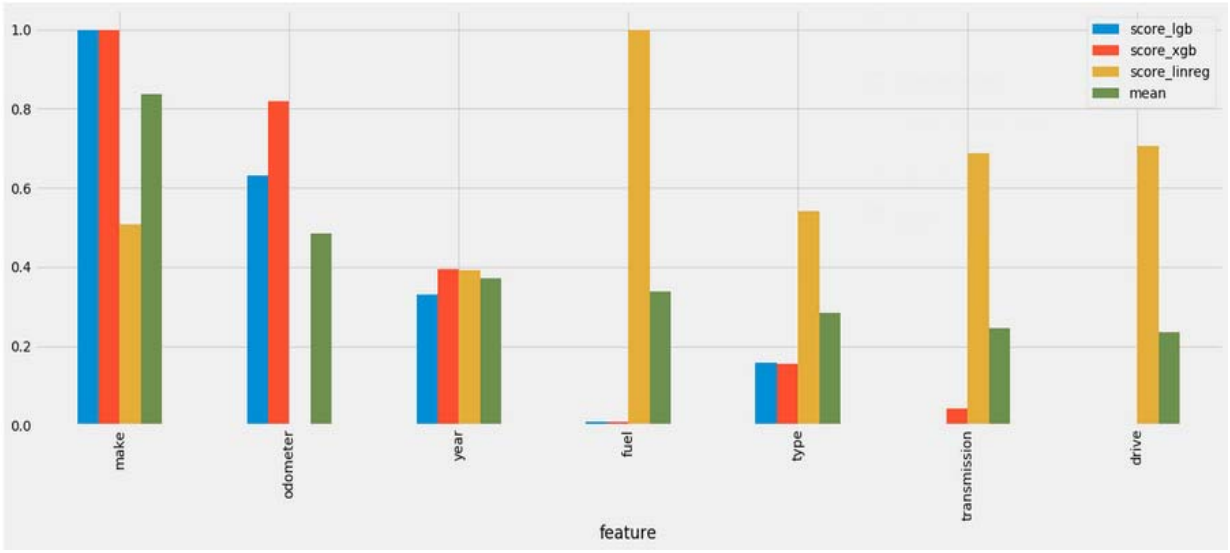


Рис. 10. Діаграма важливості ознак українського набору даних

Вибір і налаштування (навчання) моделей та їх застосування для передбачення даних

Як бачимо, для визначення ціни автомобіля потрібно дослідити залежність її величини від ознак того чи іншого автомобіля. Для розв'язання такої задачі зазвичай використовують регресійні методи дослідження даних. Враховуючи велику кількість класів, для розв'язання поставленої задачі пропонується застосувати моделі-регресори на основі дерев рішень (Decision Tree Regressor, Extra Trees Regressor, Random Forest) або регресійних моделей (Linear Regression, Ridge Regressor, Bagging Regressor, AdaBoost Regressor, XGB, LGBM, Voting Regressor) [11], методу опорних векторів тощо, а не моделі-класифікатори на їхній основі.

На попередньому етапі відфільтровано дані, відібрано оптимальні ознаки, визначено типи моделей, які варто використовувати для передбачення даних. Для розв'язання задачі передбачення використано авторський досвід, зокрема, програми-кернели із профіля Мокіна В. Б. на веб-платформі Kaggle (https://www.kaggle.com/vbmokin). Їх застосування дозволило ранжувати всі дані по США за точністю R^2-критерію (Sklearn.r2_test) (рис. 11).

	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

Рис. 11. Ранжування за R^2-критерієм результатів передбачення моделей за величиною помилки R^2-критерію, натренованих на американських даних

Точність моделей оцінюється за трьома критеріями: RMSE (root-mean-square-error або середньоквадратична похибка), R^2 (R-squared або коефіцієнт детермінації) і за відносною похибкою δ на основі вбудованої функції MAE (абсолютне значення середньої похибки).

Пояснимо яким чином використовуються ці методи оцінювання точності. RMSE обчислюється у вигляді квадратного кореня з середнього значення квадратів різниці результатів натренованої моделі, та даних, що містяться у тренувальному наборі даних [12]

$$RMSE = \sqrt{\frac{\sum_{t=1} T\left(\hat{y}_t - y_t\right)^2}{T}}, \tag{1}$$

де T — загальна кількість результатів (рядків даних); $\hat{y}$ — результат обчислення моделі; y — відповідна величина з тренувального набору даних.

Коефіцієнт детермінації є статистичним показником, що використовується в статистичних моделях як показник залежності варіації залежної змінної від варіації незалежних змінних, що вказує наскільки отримані спостереження підтверджують модель [13]

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{t=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{t=1}^{n}\left(y_i - \overline{y}_i\right)^2}, \tag{2}$$

де y — істинне значення з тренувального набору даних; $\hat{y}$ — передбачене значення відповідного набору ознак; $\overline{y}$ — середнє арифметичне істинних значень з тренувального масиву даних.

Відносна похибка обчислюється за відомою формулою

$$\delta = \frac{\sum_{t=1}^{n}\left|y_i - \hat{y}_i\right|}{\sum_{t=1}^{n}\left|y_i\right|} \cdot 100, \%. \tag{3}$$

Аналіз показав, що найкращою моделлю за R^2-критерієм є модель LGBM. Її застосування до даних США дозволило отримати точність передбачення 86,12 %.

Результат аналогічного застосування запропонованої технології до українського набору даних показано на рис. 12. Найкращою моделлю за R^2-критерієм є знову LGBM. Її точність передбачення склала 85,55 %.

	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

Рис. 12. Ранжування за R^2-критерієм результатів передбачення моделей, натренованих на українських даних

Висновки

Дослідження масивів даних, що містять інформацію про продажі вживаних авто в США та Україні, за запропонованою інтелектуальною технологією на основі бібліотек Python показало, що, по-перше, для точного передбачення ціни необхідно провести докладний попередній розвідувальний аналіз і відфільтрувати помилкові та аномальні дані, а також відкинути ознаки, які у них відрізняються, щоб результати можна було порівнювати. Після цього можна переходити до тренування моделей та порівняння їхньої точності для вибору оптимальної. Обґрунтовано, що для розв'язання поставленої задачі найкраще обрати моделі-регресори, оскільки завдання передбачення великої кількості класів ціни полягає в аналізі залежності деякої величини (у нашому випадку ціни автомобіля) відносно інших ознак (виробник, модель, обсяг пробігу, трансмісія тощо). Вибрані 15 моделей і натреновані на американських та українських даних. Результат показав зіставні результати: оптимальною моделлю в обох випадках є модель LGBM з бібліотеки Lightgbm, з точністю 85…86°%. Таким чином, отримала подальший розвиток інтелектуальна технологія аналізу та передбачення ціни на вживані авто, за рахунок удосконалення параметрів фільтрів, вибраних під час розвідувального аналізу даних за запропонованим алгоритмом, та підходу щодо вибору оптимальної моделі з багатьох, отриманих зокрема й з оптимізацією гіперпараметрів, що дозволило підвищити точність передбачення ціни автомобілів за їх параметрами.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] A. Bezerra, I. Silva, L. A. Guedes, D. Silva, G. Leitão, and K. Saito, "Extracting Value from Industrial Alarms and Events: A Data-Driven Approach Based on Exploratory Data Analysis," *Sensors*, 2019, no 19, issue 12, pp. 11-32.

[2] Stefan Lessmann, and Stefan Voß, "Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy," *International Journal of Forecasting*, 2017, no 33, issue 4, pp. 864-877.

[3] Kanwal Noor, and Sadaqat Jan, "Vehicle Price Prediction System using Machine Learning Techniques," *International Journal of Computer Applications*, 2017, no 167, issue 9, pp. 27-31.

[4] Sun, Ning & Bai, Hongxi & Geng, Yuxia & Shi, Huizhu, "Price evaluation model in second-hand car system based on BP neural network theory," *IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, 2017, pp. 431-436.

[5] *Python leads the 11 top Data Science, Machine Learning platforms: Trends and Analysis.* [Electronic resource]. Available: https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html .

[6] *Comprehensive Data Exploration with Python* [Electronic resource]. Available: https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python .

[7] *Module pandas_profiling.* [Electronic resource]. Available: https://pandas-profiling.github.io/pandas-profiling/docs/

[8] *Matplotlib API Overview*. [Electronic resource]. Available: https://matplotlib.org/api/index.html .

[9] *A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics.* [Electronic resource]. Available: https://arxiv.org/abs/1811.11440 .

[10] *Used Cars Dataset, Vehicles listings from Craigslist*. [Electronic resource]. Available: https://www.kaggle.com/austinreese/craigslist-carstrucks-data .

[11] *Supervised Learning API Overview*. [Electronic resource]. Available: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning .

[12] T. Houska, P. Kraft, A. Chamorro-Chavez, and L. Breuer, *SPOTting Model Parameters Using a Ready-Made Python Package*. [Electronic resource]. Available: https://doi.org/10.1371/journal.pone.0145180 .

[13] *Metrics and scoring: quantifying the quality of predictions*. [Electronic resource]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score .

Рекомендована кафедрою системного аналізу, комп'ютерного моніторингу та комп'ютерної графіки ВНТУ

Стаття надійшла до редакції 9.12.2019

Мокін Віталій Борисович — д-р техн. наук, професор, завідувач кафедри системного аналізу, комп'ютерного моніторингу та комп'ютерної графіки, e-mail: vbmokin@gmail.com ;

Лосенко Арсен Володимирович — студент факультета компьютерных систем и автоматики, e-mail: arsenlosenko@gmail.com ;

Дратований Михайло Володимирович — аспірант кафедри системного аналізу, комп'ютерного моніторингу та інженерної графіки

V. B. Mokin[1]
A. V. Losenko[1]
M. V. Dratovanyi[1]

Intellectual Technology of Analysis and Price Forecasting of Used Cars

[1]Vinnytsia National Technical University

For the profitable sale of a used car, people should not only be guided by their own or third-party experts' evaluation, but also use all other suitable resources. Such resources can serve as price prediction systems that, using the common features of a car (such as a car manufacturer, car model, mileage, fuel type, body type, etc.), are able to predict the possible price of a car. Such systems can help in decision-making not only to ordinary car dealers, but also to agencies involved in the ordering and bulk transportation of used cars from abroad. To select the key features and identify the optimal structure and parameters of the models, relevant datasets should be selected, the intelligence analysis and selection of features will be conducted, after which building of a number of machine learning models has begun, from which the optimal model was chosen by certain criteria. In order to build an information system and test the functionality of the proposed intellectual technology, two comparable datasets for used cars of the USA and Ukraine were selected. Python methods and libraries have been systematized for intelligence analysis and general recommendations for their application for the task have been formulated. The general principles of intellectual technology, which is tested on the selected datasets, are offered. In particular, a exploratory data analysis of US data was conducted and a rule for filtering anomalous, and possibly erroneous, data was substantiated. Many possible models were selected, their training was carried out and the optimal one was selected according to the R-squared criterion. The cost of the car has been predicted to an accuracy of 86.1%. A similar problem is solved for data on Ukraine. An accuracy of 85.6% was achieved. This has proven the workability of the proposed technology and has yielded useful results in practice.

Keywords: intellectual technology, data mining, price prediction, used car, machine learning models.

Mokin Vitalii B. — Dr. Sc. (Eng.), Professor, Head of the Chair of System Analysis, Computer Monitoring and Computer Graphics, e-mail: vbmokin@gmail.com

Losenko Arsen V. — Student of the Department of Computer Systems and Automation, e-mail: arsenlosenko@gmail.com ;

Dratovanyi Mykhailo V. — Post-Graduate Student of the Chair of System Analysis, Computer Monitoring and Engineering Graphics

В. Б. Мокин[1]
А. В. Лосенко[1]
М. В. Дратований[1]

Интеллектуальные технологии анализа и предвидения цен на подержанные автомобили

[1]Винницький национальный технический университет

Для выгодной продажи подержанного автомобиля следует руководствоваться не только собственной оценкой или оценкой сторонних экспертов, а также использовать все другие подходящие для этого ресурсы. Такими ресурсами могут служить системы предвидения цен, которые с помощью общих признаков того или иного автомобиля (например производитель автомобиля, модель автомобиля, пробег, вид топлива, тип кузова и другие) способны прогнозировать возможную цену автомобиля. Такие системы могут помочь при принятии решений не только рядовым продавцам подержанных автомобилей, но и агентствам, которые занимаются заказами и массовым перевозкам подержанных авто из-за рубежа. Для выбора ключевых признаков и идентификации по ним оптимальной структуры и параметров моделей необходимо выбрать релевантные датасеты, провести их разведывательный анализ и отбор признаков, построить ряд моделей машинного обучения, из которых выбрать оптимальную по определенным критериям. Для построения информационной системы и проверки работоспособности предложенной интеллектуальной технологии были выбраны два сопоставимые датасеты по подержанным автомобилям США и Украины. Проведена систематизация методов и библиотек на Python для проведения разведовательного анализа данных и сформулированы общие рекомендации по их применению для поставленной задачи. Предложены общие принципы интеллектуальной технологии, апробированой на отобранных датасетах. В частности, проведен разведывательный анализ данных по США и обоснованно правило для фильтрации аномальных, а возможно и ложных, данных. Выбрано множество возможных моделей, осуществлены их тренировки и выбрана оптимальная среди них по R^2-критерию. Осуществлены предсказания стоимости автомобиля, с точностью 86,1 %. Аналогичная задача решена и для данных по Украине. Достигнута точность 85,6 %. Это доказало работоспособность предлагаемой технологии и позволило получить полезные для использования на практике результаты.

Ключевые слова: интеллектуальная технология, разведывательный анализ данных, предсказания цены, подержанный автомобиль, модели машинного обучения.

Мокин Виталий Борисович — д-р техн. наук, профессор, заведующий кафедрой системного анализа, компьютерного мониторинга и компьютерной графики, e-mail: vbmokin@gmail.com ;

Лосенко Арсен Владимирович — студент факультета компьютерных систем и автоматики**,** e-mail: arsenlosenko@gmail.com ;

Дратований Михаил Владимирович — аспирант кафедры системного анализа, компьютерного моринге и инженерной графики