УДК 004.8:338+629.33

В. В. Гіжевський
А. Р. Яцолт

РОЗРОБКА ІНТЕЛЕКТУАЛЬНОЇ ТЕХНОЛОГІЇ АНАЛІЗУ ТА ПЕРЕДБАЧЕННЯ ЦІН НА ВЖИВАНІ АВТОМОБІЛІ

Вінницький національний технічний університет

Анотація

У тезах вказано основну проблематику, описано здійснену роботу в розвідувальному аналізі даних, налаштування, видалення та наповнення даних за допомогою стандартних модулів та методів на мові програмування Python. Розроблено ефективну модель машинного навчання для задачі передбачення цін на вживані автомобілі.

Ключові слова: *Python, EDA, передбачення цін на авто, робота з даними, аналіз даних, Pandas, Numpy, RandomForestRegressor, GridSearchCV, RMSE.*

Abstract

The thesis outlines the main issues, describes the work done in intelligence analysis, configuration, deletion, and filling of data using standard modules and methods in the Python programming language. An effective model of machine learning for the problem of predicting the prices of used cars has been developed.

Keywords*: Python, EDA, prediction of car prices, data processing, data analysis Pandas, Numpy, RandomForestRegressor, GridSearchCV, RMSE.*

Вступ

На сьогоднішній день, транспортна галузь вважається однією з основ економіки. Автомобілі в розвинених країнах називають «Промисловістю промисловостей». За словами професіоналів, автомобільна промисловість як в США, так і в Україні відчуває значне зростання. Більше того, значно зростають оберти використання автомобілів місцевим населенням, які припадають в Україні.

Сьогодні майже кожен хоче мати власний автомобіль, але через такі фактори, як доступність та/або економічність умов, багато хто воліє обирати вживані автомобілі. Точне прогнозування цін на вживані автомобілі вимагає експертних знань через характер їх залежності від різноманітних факторів та особливостей. Ціни на вживані автомобілі є непостійними на ринку, а тому, як покупцям, так і продавцям потрібна інтелектуальна технологія, яка дозволить їм прогнозувати правильні ціні ефективно. У цій інтелектуальній системі найскладнішою проблемою є збір набору даних, який містить усі важливі елементи, такі як рік виробництва автомобіля, його тип газу, його стан, пройдені милі, потужність, двері, кількість фарбувань автомобіля, відгуки клієнтів, вага автомобіля тощо. Зрозуміло, що на ціну товару впливає багато факторів і саме потрібний набір даних надає веб-платформа Kaggle розроблена компанією Google.

Метою роботи є застосування основних методів аналізу даних та розроблення ефективної моделі для вирішення задачі передбачення цін на вживані автомобілі.

Основна частина

Для дослідження було обрано набір даних Used Cars Dataset викладений Austin Reese у вільний доступ на веб-платформі Kaggle[1]. Датасет містить майже всю відповідну інформацію по продажу автомобілів, яку надає компанія Craigslist. А саме, такі стовпці, як ціна, стан, виробник, широта/довгота та 21 інших категорій.

Першою задачею було проведення розвідувального аналізу даних (EDA)[2], для ознайомлення із набором даних, очищення даних від аномальних та не справжніх значень, визначено більшість закономірностей між ознаками набору даних.

Для вирішення задачі, обробки та аналізу даних, було обрано мову програмування Python та модулі(програмні пакети), такі як, Numpy, Pandas, MatPlotLib, Seaborn, SkLearn та інші[2].

Датасет начислює цілих 426880 елементів та 26 ознак, з яких 773 елементи мали відсутні 9 та більше ознак, через що, їх було видалено. Дані із відсутньою кількість елементів менше 9 було залишено, та їх нараховується 426107 одиниць елементів.

Значення набору даних було відфільтровано за його основними ознаками в кількості 14 одиниць, після чого, дані були готові до якісного та швидкого тренування моделі прогнозування цін на вживані автомобілі.

Ознаки, за якими було проведено машинне навчання моделі:

- (‘year’) – рік випуску автомобіля;
- (‘manufacturer’) – виробник автомобіля;
- (‘model’) – модель автомобіля;
- (‘condition’) – стан автомобіля;
- (‘cylinders’) – кількість циліндрів в автомобілі;
- (‘fuel’) – вид палива автомобіля (gas, diesel, hybrid, electric, other);
- (‘odometer’) – пробіг автомобіля;
- (‘title_status’) – титульний статус автомобіля (clean, rebuild, lien, salvage, missing, parts only);
- (‘transmission’) – трансмісія автомобіля (manual, automatic, other);
- (‘drive’) – привід автомобіля (rwd, 4wd, fwd);
- (‘size’) – система класифікації, довжина автомобіля (full-size, mid-size, compact, sub-compact);
- (‘type’) – тип кузова автомобіля (pickup, truck, couple, SUV, hatchback, mini-van, sedam, offroad, bus, van, convertible, wagon, other);
- (‘paint_color’) – колір автомобіля (white, blue, red, black, silver, grey, brown, yellow, orange, green, custom та purple).

Як виглядають готові дані для прогнозування, але поки без видаленої ознаки (‘price’), зображено на рисунку 1.

	price	year	manufacturer	model	condition	cylinders	fuel	odometer	title_status	transmission	drive	size	type	paint_color
27	33590	2014.0	gmc	sierra 1500 crew cab slt	good	8 cylinders	gas	57923.0	clean	other	Unkown	full-size	pickup	white
28	22590	2010.0	chevrolet	silverado 1500	good	8 cylinders	gas	71229.0	clean	other	Unkown	full-size	pickup	blue
29	39590	2020.0	chevrolet	silverado 1500 crew	good	8 cylinders	gas	19160.0	clean	other	Unkown	full-size	pickup	red
30	30990	2017.0	toyota	tundra double cab sr	good	8 cylinders	gas	41124.0	clean	other	Unkown	full-size	pickup	red
31	15000	2013.0	ford	f-150 xlt	excellent	6 cylinders	gas	128000.0	clean	automatic	rwd	full-size	truck	black

Рисунок 1 – Приклад готового датасету вживаних автомобілів для прогнозування

Після очищення та фільтрування набору даних, останній був повторно зменшений у розмірі, за рахунок видалення аномальних (‘price’ > 150 тисяч доларів) та не справжніх (‘price’ = 1234567, 1111111, 9999999 доларів, price < 1500 доларів) значень цін та не тільки. Саме за допомогою квантелів, було визначено, які дані являються максимальними, які мінімальними, середніми та викидними.

Результати дослідження

Для задачі створення моделі передбачення цін на вживані автомобілі було обрано одну із найкращих моделей машинного навчання, яка в більшості випадків, дає чудові результати, це модель Random Forest Regressor. Але для того, щоб вручну не перебирати гіперпараметри, та скоротити час розробки моделі, було використано Scikit-Learn GridSearchCV[3].

Окрім цього, було використано не менш важливий алгоритм, для визначення коефіцієнта детермінації R^2. Даний показник визначається наступним чином: Якщо $\hat{y}_i$ – це передбачене значення -го зразка, а y_i – відповідне істинне значення, то частка правильних передбачень над $n_{samples}$ визначається як[4]:

$$accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i) \quad (1)$$

Знаючи значення даного коефіцієнта, ми із повною впевненістю можемо сказати, наскільки чітко наша модель передбачила дані.

Також, для оцінки ефективності моделі, використовують значення RMSE[5], що розшифровується, як середньоквадратична похибка, або ж root-mean-square-error.

Результат роботи, RMSE та точність передбачення даних моделлю Random Forest Regressor із застосуванням GridSearchCV наведено на рисунку 2.

```
y_pred = clf_GSCV.best_estimator_.predict(X_test_normed)
rmse2 = np.sqrt(MSE(y_test, y_pred))
print("RMSE = {:.2f}".format(rmse2))
accuracy = clf_GSCV.score(X_test_normed,y_test)
print(accuracy*100,'%')

RMSE = 4537.24
90.32195993536037 %
```

Рисунок 2 – Результат роботи моделі передбачення цін на вживані автомобілі

Як можна побачити, результат роботи моделі показав себе на відмінно, що підтверджує ефективність моделі Random Forest Regressor для розв'язання задач машинного навчання.

Код інтелектуальної технології аналізу та передбачення цін на вживані автомобілі розміщений та доступний на веб-платформі Kaggle[6].

Висновки

Використовуючи дані Used Cars Dataset, можливості мови програмування Python та набір модулів для аналізу та попередньої обробки даних, було здійснено аналіз датасету, проведено розвідувальний аналіз даних(EDA), проведено Feature Engineering, що дало змогу обробити дані, очистити їх, та профільтрувати і вже після, провести повноцінне тренування моделі із автоматично підставленими гіперпараметрами за допомогою GridSearchCV, що в свою чергу дало результат RMSE = 4537,24 та неймовірні ~90.3% точності прогнозування.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Used Cars Dataset. Kaggle. 2021. URL: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data
2. Python Data Science Handbook: Essential Tools for Working with Data. 2016. URL: https://g.co/kgs/JxxZP5
3. sklearn.model_selection.GridSearchCV. URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
4. Metrics and scoring: quantifying the quality of predictions. URL: https://scikit-learn.org/stable/modules/model_evaluation.html
5. Root-mean-square deviation. URL: https://en.wikipedia.org/wiki/Root-mean-square_deviation
6. Car prices prediction. Python. Used Cars Dataset, Kaggle. URL: https://www.kaggle.com/code/hizhevskyivladyslav/car-prices-prediction/notebook

Гіжевський Владислав Віталійович – студент групи СА-18б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, g-mail: vladgiz2000@gmail.com

Ящолт Андрій Русланович – к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, e-mail: yasholt@gmail.com

Hizhevskyi Vladislav Vitaliyovych – student of the CA-18b, Faculty of Intelligent Information Technology and Automation, Vinnytsia National Technical University, Vinnytsia, g-mail: vladgiz2000@gmail.com

Yashcholt Andrey Ruslanovich - Ph.D., Associate Professor of Systems Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: yasholt@gmail.com