

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА**

**Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»
Кафедра математичного моделювання та статистики**

Освітньо-професійна програма

Галузь знань

Спеціальність

Спеціалізація

Економічна кібернетика

05 соціальні та поведінкові науки

051 економіка

Економічна кібернетика

Форма навчання: очна (денна)

КВАЛІФІКАЦІЙНА БАКАЛАВРСЬКА РОБОТА

на тему «Моделювання ціноутворення на вторинному авторинку»

здобувача Зварич Христини Олегівни

**Науковий керівник: кандидат економічних наук, доцент
Катуніна Ольга Сергіївна**

(підпис)

**Робота допущена до захисту перед екзаменаційною
комісією з атестації здобувачів вищої освіти (ЕК)**

**Завідувач кафедри: кандидат фізико-математичних наук,
професор Великоіваненко Г.І.**

(підпис)

Київ 2023

ЗМІСТ

ВСТУП.....	3
РОЗДІЛ 1 ТЕОРЕТИЧНІ ЗАСАДИ ДОСЛІДЖЕННЯ ТА МОДЕЛЮВАННЯ РОЗВИТКУ ВТОРИННОГО АВТОРИНКУ	5
1.1 Структура та особливості вторинного авторинку як об'єкту моделювання	5
1.2 Напрями дослідження розвитку вторинного авторинку у світі	9
1.3 Огляд інструментарію математичного моделювання ціноутворення та системи факторів формування цін на вторинному авторинку	13
РОЗДІЛ 2 МЕТОДОЛОГІЧНІ ПІДХОДИ ТА ІНСТРУМЕНТАРІЙ МОДЕЛЮВАННЯ ЦІНОУТВОРЕННЯ НА ВТОРИННОМУ АВТОРИНКУ	17
2.1 Інструментарій розвідувального аналізу та препроцесингу даних у дослідженнях процесів формування цін на автомобілі.....	17
2.2 Напрями використання методів машинного навчання для аналізу та прогнозування цін на вторинному авторинку	21
2.3 Моделювання процесів ціноутворення та системи факторів на базі бустінгових моделей	26
РОЗДІЛ 3 ПРАКТИЧНЕ ЗАСТОСУВАННЯ МАТЕМАТИЧНИХ МОДЕЛЕЙ ФОРМУВАННЯ ЦІН НА ВТОРИННОМУ АВТОРИНКУ	29
3.1 Аналіз результатів візуалізації та препроцесингу даних	29
3.2 Аналіз результатів моделювання факторів формування цін за допомогою інструментів машиного навчання	36
3.3 Аналіз результатів моделювання процесів ціноутворення за допомогою методів градієнтного бустингу	42
ВИСНОВКИ.....	48
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	50
ДОДАТКИ	54

ВСТУП

Актуальність теми. У сучасному світі вторинний авторинок набуває все більшої популярності. Останнім часом споживачі все частіше віддають перевагу покупці автомобіля, який був у використанні, замість нових, так як це є набагато доступнішим варіантом. Відповідно до цього, зростає кількість платформ, які допомагають продавати і відповідно купувати такі автомобілі.

Однак, одна з найбільших проблем, з якою стикаються покупці – це недостатня інформація та компетенція про вторинний авторинок, а також нестабільність цін. У більшості випадків продавці виставляють ціну на свій розсуд, що може вагомо відрізнятись від реальної ціни на ринку.

За допомогою аналізу факторів, які впливають на ціноутворення, продавці та покупці будуть мати більше інформації щодо цін на ринку, що допоможе споживачам приймати більш обґрунтовані рішення щодо покупки автомобіля, а продавцям збільшувати шанси продати їхній автомобіль.

Аналіз досліджень і публікацій. Технології аналізу автомобільного ринку в основному розглянуті в роботах зарубіжних вчених. Так, техніка множинної лінійної регресії для прогнозування ціни на автомобілі була розглянута в роботі (Kanwal & Sadaqat, 2017); застосування моделей множинного лінійного регресійного аналізу, k-найближчих сусідів (kNN), дерев рішень, наївного байєсівського класифікатора продемонстровано в роботі (Sameerchand Pudaruth, 2014).

Метою роботи є побудова математичних моделей ціноутворення на вторинному авторинку, які дозволять оцінювати ринок та прогнозувати ціни на основі впливу різних характеристик автомобілів. Для досягнення поставленої мети потрібно виконати такі завдання:

- дослідити функції вторинного авторинку;
- проаналізувати тенденції розвитку вторинного авторинку у світі;
- проаналізувати фактори, які можуть впливати на ціну автомобілей;

- дослідити основні техніки аналізу та математичний інструментарій побудови моделей;
- побудувати математичні моделі ціноутворення на основі методів машинного навчання;
- експериментально дослідити ефективність побудованих моделей, сформулювати висновки та рекомендації.

Об'єктом дослідження є функціонування вторинного авторинку у світі.

Предметом дослідження є математичні моделі ціноутворення на вторинному авторинку на базі машинного навчання.

Методи дослідження. Для дослідження теоретичних основ вторинного авторинку були використані методи аналізу та синтезу. Для аналізу сучасного стану ринку використано комплексний підхід збирання та обробки інформації, системний аналіз. Використано графічний метод для візуалізації залежностей та препроцесингу даних. Для прогнозування були використані такі методи як лінійна регресія, випадкові ліси, дерева рішень, а також градієнтний бустинг.

Інформаційною базою дослідження виступили роботи зарубіжних та вітчизняних науковців, інтернет-платформи вторинних авторинків.

Теоретична, методична та практична значущість отриманих результатів. Дослідження дозволяє зрозуміти особливості ціноутворення на вторинному авторинку. Теоретичні результати можуть допомогти в створенні теорії ціноутворення. Методологічні результати можуть стати основою для подальших досліджень та розробки нових або удосконалення старих методів аналізу. Практичні результати можуть бути корисними як для великого бізнесу так і для індивідуальних споживачів, для прийняття рішень про купівлю або продаж автомобілів на ринку.

РОЗДІЛ 1

ТЕОРЕТИЧНІ ЗАСАДИ ДОСЛІДЖЕННЯ ТА МОДЕЛЮВАННЯ РОЗВИТКУ ВТОРИННОГО АВТОРИНКУ

1.1 Структура та особливості вторинного авторинку як об'єкту моделювання

Автомобільний ринок – це ринок на якому здійснюється операція покупки-продажу автомобіля, що включає в себе як нові автомобілі так і ті, які були у використанні, а також всі пов'язані з цим послуги, такі як страхування, технічне обслуговування.

Вживаний автомобіль – це той, який спочатку був придбаний одним або кількома власниками. Вживані автомобілі продаються в різних місцях, включаючи компанії з прокату автомобілів, франшизи та незалежних автодилерів, аукціони та приватні розпродажі тощо.

Прийнято вважати, що товаром автомобільного ринку може бути тільки готовий автомобіль, але зустрічаються також мотоцикли, човни, або ж запчастини машини. Суб'єктами попиту вважають домашні господарства та комерційні і державні організації. Представниками пропозиції виступають виробники автомобілів, або ж автомобільні дилери. Варто зауважити, що тільки в рідкісних випадках промислові підприємства проводять операцію з кінцевим споживачем. Також важливим суб'єктом автомобільного ринку виступає держава. [1].

Важливість автомобільної промисловості полягає в [2]:

- Зв'язках з іншими секторами. втомобільна промисловість має важливий мультиплікаційний ефект в економіці. Це важливо для наступних галузей.
- Зайнятості. велика кількість людей по всьому світі працюють в автомобільному секторі. Виробництво забезпечує багато робочих місць

- Економіці. Оборот, який створений автомобільною промисловістю становить значний відсоток світового ВВП.

Основними функціями авторинку є купівля–продаж автомобілів, а також забезпечення інформацією та підтримка учасників. Розглянемо детальніше кожен з важливих функцій:

1) Купівля–продаж автомобілів. Однією з основних функцій авторинку є обмін автомобілями між учасниками ринку. Авторинок забезпечує можливість продажу автомобілів безпосередньо між власниками, а також шукає покупців для автодилерів та інших продавців. Власники автомобілів можуть звертатися до авторинку для знаходження покупців, а покупці можуть шукати необхідний автомобіль серед пропозицій, які є на ринку.

2) Інформаційна підтримка. Авторинок забезпечує учасників ринку інформацією про технічні характеристики автомобілів, рівень цін на автомобілі та ринкову ситуацію. Інформація, яка надається на ринку, допомагає учасникам приймати рішення щодо покупки або продажу автомобіля.

3) Підтримка учасників ринку. Авторинок забезпечує підтримку учасників ринку в процесі купівлі–продажу автомобілів. Авторинок може надавати послуги з оцінки автомобіля, консультації щодо вибору автомобіля, підготовки документів для продажу або покупки автомобіля. Додаткові послуги, які надає авторинок, можуть бути корисні для учасників ринку, оскільки допомагають зменшити ризики та забезпечити успішну операцію з купівлі–продажу автомобіля.

4) Регулювання цін на автомобілі. Авторинок дозволяє продавцям та покупцям знайти баланс між попитом та пропозицією, що може впливати на ціни. Якщо попит на автомобілі високий, а кількість доступних автомобілів низька, ціни можуть підвищитися. Навпаки, якщо кількість автомобілів перевищує попит, ціни можуть знизитися. Таким чином, авторинок регулює ціни, що дозволяє покупцям та продавцям знаходити взаємовигідні угоди.

Крім цього, варто згадати про те, що авторинок відіграє важливу роль в економіці кожної країни та світу загалом, забезпечуючи зайнятість великої частки населення та сприяючи розвитку підприємства.

Що ж стосується вторинного авторинку, то зменшення цін на вживані автомобілі дозволяє більшій частині населення мати можливість придбати транспортний засіб та забезпечує більшу мобільність населення планети.

Існують три основні чинники, що спонукають до сталого розвитку в автомобільній промисловості [3]:

1. Державне втручання.

Урядові установи відіграють впливову роль у зростанні імпульсу до стійких практик. Наприклад, такі директиви, такі як Європейська зелена угода та Паризька угода, змушують автомобільні компанії шукати більш стійкі рішення для досягнення суворих цілей щодо нейтрального викиду вуглецю.

2. Зміна очікувань.

Очікування клієнтів, інвесторів і навіть працівників змінюються, що має великий вплив на галузь. Технічно підковані та екологічно свідомі споживачі все більше вимагають інтуїтивно зрозумілих технологій в транспортних засобах, водночас змушуючи автовиробників зосереджуватися на екологічних методах і моделях альтернативного палива, таких як електромобілі.

3. Мегатренди.

Кілька мегатрендів в автомобільній промисловості є великими рушійними силами інноваційних змін. До них належать розвиток автономних транспортних засобів і автопарків, використання даних з підключених транспортних засобів, програми спільного використання автомобілів, альтернативні ініціативи транспортування на вимогу та, звичайно, зміна парадигми галузі в бік електромобілів.

Збільшення використання передових технологій і розширення виробництва електромобілів повністю змінили традиційне виробництво автомобілів. Перехід до виробництва електромобілів зробив деякі робочі місця зайвими, а також зробив процес традиційного конвеєрного виробництва застарілим.

Також варто згадати про глобалізацію, так як зараз дуже популярна практика об'єднання двох чи більше автовиробників, для того, щоб досягнути

спільної мети – вийти на закордонні ринки та стати більш конкурентноспроможними.

Ринок автомобілів – це олігополія, тобто на ринку присутні домінантні фірми. На рис. 1.1 можна побачити рейтинг найпопулярніших моделей машин, які продавались у 2022 році.

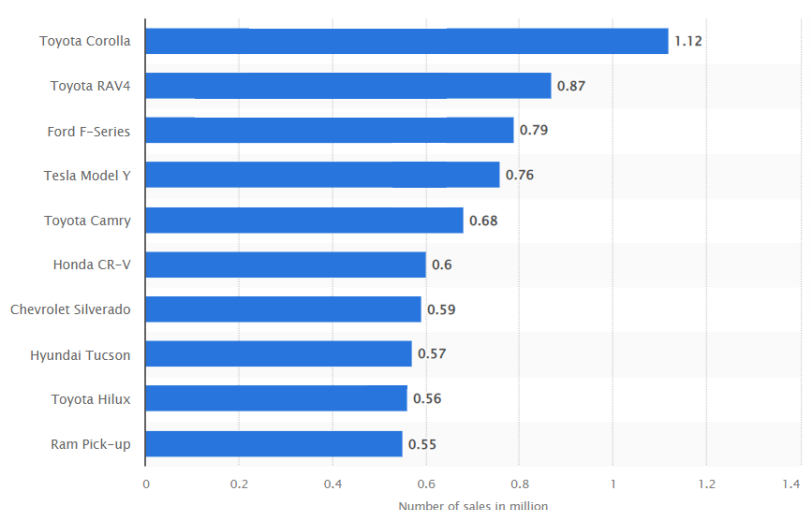


Рисунок 1.1 – Найпопулярніші машини у світі у 2022 році [4]

Ключові гравці ринку зосереджуються на розширенні клієнтської бази, щоб отримати конкурентну перевагу на ринку. Таким чином, постачальники вживають кілька стратегічних ініціатив, таких як співпраця, поглинання та злиття, а також партнерства. Наприклад, у 2020 році Volkswagen оголосив про серйозні інвестиції в ринок вживаних автомобілів шляхом співпраці власної мережі продажів вживаних автомобілів Das WeltAuto з різними платформами для вживаних автомобілів. Основні автовиробники також розширюють свою присутність у цьому просторі завдяки мережам продажу вживаних автомобілів, як-от True Value Maruti Suzuki, First Choice Wheels M&M Mahindra та U Trust Toyota. Серед відомих гравців світового ринку вживаних автомобілів є [5]:

- Alibaba.com;
- Asbury Automotive Group;
- AutoNation Inc.;
- CarMax Business Services, LLC;
- Cox Automotive;
- eBay Inc.;

- Group 1 Automotive Inc.;
- Hendrick Automotive Group;
- LITHIA Motor Inc.

Ринок вживаних автомобілів сегментований за типом палива, типом кузова, типом постачальника та каналом збуту. Детальніша інформація зображена у табл. 1.1.

Таблиця 1.1 – Сегменки ринку вживаних автомобілів [6]

Тип постачальника	Організовано
	Неорганізований
Тип палива	Бензин
	Дизель
	Електричний
	Інші
Статура	Хетчбек
	Седан
	Позашляховики та інше
Канал збуту	Онлайн
	Офлайн

1.2 Напрями дослідження розвитку вторинного авторинку у світі

Дослідження розвитку вторинного авторинку у світі є важливим напрямком в автомобільній галузі, оскільки цей ринок є значним чинником у формуванні світової економіки. Вторинний автомобільний ринок є місцем продажу вживаних автомобілів, і він досить розвинений у багатьох країнах світу.

До недавнього часу виробники та дилери автомобілів зосереджувалися переважно на бізнесі нових автомобілів, виключаючи вживані автомобілі, які часто розглядалися як побічний продукт. Проте конкуренція на ринку та загроза нових

входів спричинили значне зростання продажів вживаних автомобілів. Більше того, додана якість та надійність вживаних автомобілів змінила споживче ставлення та збільшила продажі вживаних легкових автомобілів. Інвестиції в управління вживаними автомобілями стали однією з вимог ринку, що характеризується зменшенням маржі, невинною конкуренцією та вимогливими споживачами.

Протягом останніх кількох десятиліть ринок комерційних вживаних автомобілів залишався переважно традиційним, зосередженим на місцевих дилерських центрах. 1990-ті та перше десятиліття 2000-х років принесли традиційні офлайн-реклами онлайн у формі веб-сайтів, таких як німецький *mobile.de* та французький *Leboncoin*. Лише останнім часом такі сайти почали пропонувати додаткові послуги, такі як пряма онлайн-покупка автомобіля та підписка. Наприкінці 2010-х з'явилися справжні цифрові роздрібні торговці: такі компанії, як *Autohero*, *Carvago* та *Driverama*, пропонують повний досвід цифрових продажів, включаючи віртуальну «випробувальну можливість», безкоштовну доставку, повну перевірку транспортних засобів, гарантії та безліч додаткових можливостей. додаткові послуги та гарантії.

Окрім платформ онлайн-продажів, існують також онлайн-платформи придбання, які купують вживані автомобілі як у споживачів, так і у комерційних продавців. Вони використовують бізнес-моделі з великою кількістю активів, як-от *Wirkaufendeinauto* (Німеччина), або моделі з невеликими активами, як-от *Motorway* (Великобританія). Після деяких проблем на початку ці онлайн-покупці тепер прозоро та надійно пропонують справедливі ринкові ціни зі значно меншими труднощами, згідно з незалежними ринковими перевітками.

Роль онлайн-продажів стала критичним фактором зростання на ринку. Інтернет-сайти на автомобільних ринках відіграли важливу роль у забезпеченні доступу споживачів одним дотиком. Поєднання таких подій викликало значне зростання попиту на старі автомобілі. Крім того, такі чинники, як доступність, доступність вживаних автомобілів, зростання потреби в особистій мобільності та поява різних онлайн-гравців для організації ринку призвели до зростання зростання ринку.

За даними Data Bridge Market Research, у 2022 році світовий ринок вживаних автомобілів оцінювався в 996 906,42 мільйона доларів США, а до 2030 року очікується, що він досягне 1700 106,13 мільйона доларів США при CAGR 6,90% протягом прогнозованого періоду [7]. За останні кілька років ринок став свідком значного зростання, оскільки цінова конкурентоспроможність серед нових гравців була одним із яскравих місць у промисловості вживаних автомобілів. Неможливість покупців придбати нові автомобілі стала однією з причин зростання обсягів продажу вживаних автомобілів, що доповнюється інвестиціями учасників галузі для створення дилерської мережі на ринку. Ці дилерські мережі допомогли учасникам ринку створити бренд і зробити варіанти вживаних автомобілів життєздатними [8].

Технологічні досягнення, такі як розвиток Інтернету та впровадження гібридних та електричних транспортних засобів, змінили позицію покупця на ринку. Крім того, споживачі тепер знають про транспортний засіб, його залишкову вартість, фінансові витрати на якість, доступність, застосовану ціну, а іноді й розмір прибутку, який продавець отримує під час укладання угоди. Ці знання змінили динаміку та зуміли перетворити розвідку клієнтів на їхню користь. Як результат, сьогодні споживачі більш схильні купувати вживані автомобілі.

Як бачимо на рис. 1.2 офлайн-продажі становили найбільшу частку обсягу понад 75,0% у 2021 році. Це пояснюється перевагою споживачів традиційного способу купівлі. Проте очікується, що сегмент каналів онлайн-продажів значно зросте найближчим часом. Розвиток онлайн-інструментів для продавців і покупців зробив ринок більш конкурентним. Крім того, онлайн-доступність інформації, включаючи ціни, відгуки та характеристики, призводить до значного зростання дилерів.



Рисунок 1.2 – Канали продажу використаних машин, у %

Джерело: розроблено автором на основі даних [5]

Ринок вживаних автомобілів загалом досить стабільний, навіть під час спаду: багато людей покладаються на свої автомобілі, щоб дістатися до роботи та побачити сім'ю, і вони радше пожертвують іншими покупками, ніж відмовляться від своїх автомобілів. Крім того, покупці нових автомобілів можуть перейти на вживані автомобілі, коли їх старі автомобілі не можна обслуговувати економічно, і вони більше не можуть дозволити собі нові автомобілі, тому на ринку вживаних автомобілів завжди буде оборот.

На рис. 1.3 можна побачити відношення ринку нових автомобілів та тих, що були у використанні. Як бачимо, в більшості країн перевагу має ринок використаних автомобілів, так як такі автомобілі є більш доступними для населення.

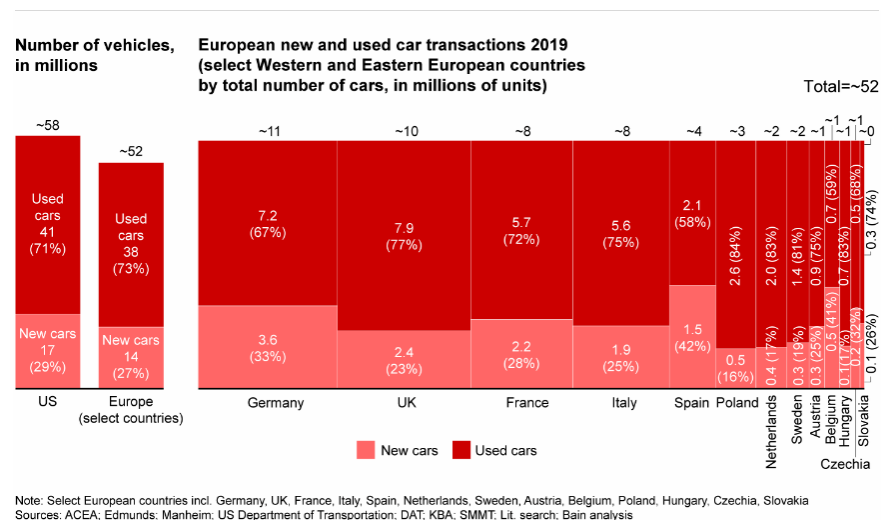


Рисунок 1.3 – Відношення обсягу авторинку нових та використаних автомобілів [9]

1.3 Огляд інструментарію математичного моделювання ціноутворення та системи факторів формування цін на вторинному авторинку

Використання машинного навчання є дуже привабливим підходом для роздрібних продавців. Замість використання, наприклад, агресивних загальних уцінок (що часто є поганою стратегією), вони можуть отримати вигоду від прогнозних моделей, які дозволяють їм визначити найкращу ціну. Сила полягає в тому, що розроблені алгоритми можуть вивчати шаблони з даних замість того, щоб бути явно запрограмованими. Моделі машинного навчання можуть постійно інтегрувати нову інформацію та виявляти нові тенденції чи нові вимоги.

Ось список часто використовуваних алгоритмів машинного навчання. Ці алгоритми можна застосувати майже до будь-якої проблеми з даними [10]:

- Лінійна регресія
- Логістична регресія
- Дерево рішень
- Випадковий ліс
- Алгоритми посилення градієнта
- GBM
- XGBoost
- CatBoost.

Для того, щоб моделювання ціни було економічно значущим потрібно визначити які фактори впливають на прогнозування найбільше. Висновок про те, чи вартий вживаний автомобіль вказаної ціни, коли ви бачите оголошення в Інтернеті, може бути складним [11]. Для цього потрібно розглянути які фактори можуть впливати на формування ціни автомобіля, що був у використанні.

Ціноутворення є ключовим фактором на вторинному авторинку, оскільки визначає конкурентоспроможність продукту та може вплинути на його популярність серед покупців. Ціна може бути рішучим фактором для покупців у виборі між двома або більше схожими автомобілями, що продаються на ринку.

У той же час, визначення оптимальної ціни на вторинному авторинку може бути складним завданням. Це залежить від багатьох факторів, таких як: рівень попиту, конкуренція, стан ринку, стан автомобіля, його технічні характеристики та історія використання. Отже, дослідження та розуміння ціноутворення на вторинному авторинку є важливим завданням для продавців та покупців, які бажають отримати максимальну вигоду від своїх угод.

Усі автомобілі швидко падають у ціні. Насправді вартість абсолютно нового автомобіля зазвичай падає приблизно на 10-15% щороку після того, як ви його придбали. Ця неминуча частина володіння автомобілем відома як амортизація [12].

Амортизація – це різниця у вартості автомобіля з моменту його придбання до моменту обміну або продажу. Зазвичай його виражають у відсотках, щоб показати, скільки втратила вартість автомобіля.

Деякі автомобілі добре тримають свою вартість, а інші дуже швидко знецінюються. Фактично, деякі нові автомобілі втрачають майже 60% своєї вартості протягом перших трьох років. Але розуміння факторів, які впливають на амортизацію, може допомогти вибрати автомобіль, який добре зберігатиме свою вартість. Розглянемо найбільш популярні фактори [13]:

1 . Ціна попередньої покупки

На жаль, вартість автомобіля майже втрачається, як тільки ви на ньому їдете.

Автосалон купує автомобілі за оптовою ціною, але для отримання прибутку ціна їх автомобілів зазвичай встановлюється вище, ніж ціна, яку вони отримують від виробника. Це означає, що як тільки покупець отримує ключі від свого автомобіля, ціна знову знижується до оптової, тож майже відразу авто коштує менше, ніж за нього заплатили.

2 . Репутація автомобіля

Виробник, модель і рік випуску автомобіля беруть участь у грі. Якщо виробник має репутацію надійних, міцних, витривалих автомобілів, це може допомогти витримати амортизацію.

Хороша репутація допомогла автомобілям таких брендів, як Mini, Ford і Volkswagen, добре зберігати свою вартість. Можна очікувати середньої амортизації їхніх автомобілів (із середнім пробігом 34 700 кілометрів) менше 50%. На іншому кінці спектру знаходиться Toyota, її автомобілі втрачають 74% своєї вартості після такої ж кількості миль.

3 . Скільки миль проїхано

Пробіг є одним із найважливіших факторів, що впливає на вартість автомобіля. Середній річний пробіг автомобіля становить від 12 000 до 15 000 миль. Все, що менше, зазвичай вважається «малим пробігом», а все, що перевищує «великий пробіг».

4 . Стан автомобіля

Перепродажна вартість автомобіля, природно, частково залежить від того, як за ним дбали. Річ іде про наявність пошкодження зовнішнього вигляду, салону чи двигуна. Присутність цього зменшує вартість автомобіля. Модифікації, такі як нова стереосистема або легкосплавні диски, також можуть мати негативний вплив. Більшість покупців віддають перевагу не модифікованим автомобілям, оскільки це значно полегшує їх страхування.

5 . Історія обслуговування автомобіля

Будь-який потенційний покупець захоче побачити сервісну історію автомобіля. Якщо автомобіль регулярно його обслуговується і присутні докази, це допоможе зменшити амортизацію.

6 . Скільки ще людей було за кермом

Як правило, чим менше власників, тим краще, коли йдеться про захист вартості автомобіля. Якщо він часто переходив з рук в руки, він, ймовірно, принесе менше.

На рис. 1.4 зображено найбільш важливі фактори, що впливають на вибір покупця під час покупки автомобіля. Відповідно тут буде доречно також згадати, що ті фактори, які впливають на вибір споживача також впливають і на ціну самого товару.

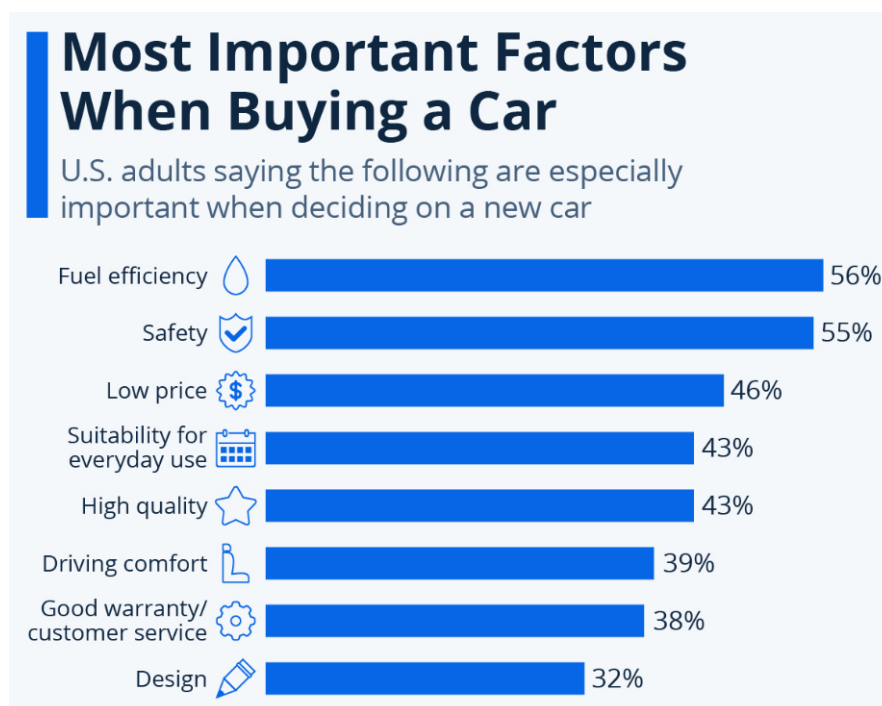


Рисунок 1.4 – Найважливіші фактори при покупці автомобіля [4]

Сьогодні на просторах інтернету існує дуже багато веб-додатків. Які пропонують свої послуги щодо розрахунку вартості автомобіля. Для продавців це корисно тим, що вони можуть правильно оцінити автомобіль та створити конкурентноспроможне середовище, а для покупців тим, що вони можуть перевірити.

Проте так звані «калькулятори» не можна вважати повністю достовірними, тому модель, яка буде враховувати різноманітні чинники та робити мінімальну похибку при розрахунку прогнозованої ціни допоможе правильно розраховувати ціну на той чи інший автомобіль.

РОЗДІЛ 2

МЕТОДОЛОГІЧНІ ПІДХОДИ ТА ІНСТРУМЕНТАРІЙ МОДЕЛЮВАННЯ ЦІНОУТВОРЕННЯ НА ВТОРИННОМУ АВТОРИНКУ

2.1 Інструментарій розвідувального аналізу та препроцесингу даних у дослідженнях процесів формування цін на автомобілі

У рамках життєвого циклу процесу Data Mining згідно з методологією CRISP DM [14] входить декілька етапів: спочатку – це розуміння даних (рис. 2.1), а потім вивчення даних, тобто так званий розвідувальний аналіз.

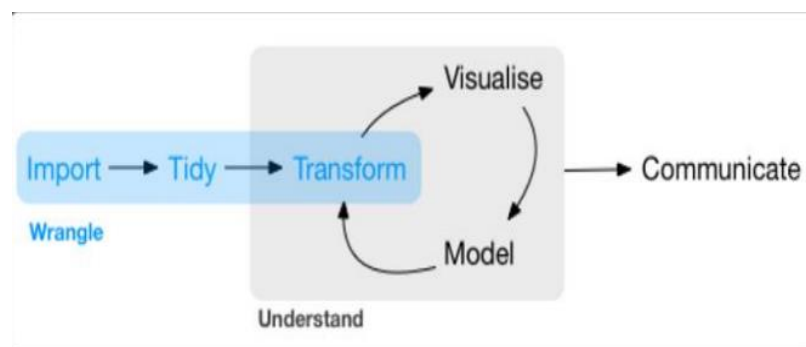


Рисунок 2.1 – Структура задачі розуміння даних на основі [14]

Поняття розвідувального аналізу даних можна узагальнити як процес дослідження основних властивостей даних.

Розвідувальний аналіз даних стосується практики використання штучного інтелекту та інструментів машинного навчання для аналізу та перетворення масивних наборів даних у інтелектуальну аналітичну інформацію, яку потім можна використовувати для подальшого аналізу.

Більш технічно процес такого перетворення можна описати як ETL (extract, transform, load) плюс представлення корисної інформації. На рис. 2.2 зображено структуру даного процесу.

The ETL Process Explained



Рисунок 2.2 – Пояснення процесу ETL [15]

Цілями розвідувального аналізу даних, які варто згадати, є:

- виявлення основних структур даних;
- виявлення аномалій;
- вибір найвагоміших змінних;
- перевірка даних.

R цілком здатний служити двигуном для процесів, пов'язаних з розвідувальним аналізом даних. R має природньо широкий спектр доступних статистичних інструментів (кілька сховищ із тисячами пакетів). Нижче наведено пакети, які безпосередньо стосуються етапів базового процесу розвідувального аналізу.

Для завантаження інформації:

- DBI – рідні драйвери баз даних для багатьох постачальників, найвища продуктивність.
- RODBC – підключення драйвера бази даних ODBC.
- RJDBC – підключення драйвера бази даних JDBC.
- data.table's fread – дуже швидкий читач файлів CSV.
- безліч інших пакетів для підтримки різних форматів даних (наприклад, xlsx, xml, json, sas, spss, stata).

Для трансформації:

- data.table – потужний інструмент перетворення даних, використовує синтаксис.

- dplyr – також потужний, але менш масштабований інструмент перетворення даних.

Для презентації:

- shiny – платформа веб-додатків для R.
- оренспі – HTTP API для R.
- httpuv – серверна бібліотека HTTP та WebSocket, також ядро пакету shiny.
- Rook – інтерфейс веб-сервера для R [16].

Представлення корисної інформації – це зовсім інше завдання, ніж процес ETL, його можна легко передати будь-якому інструменту інформаційної панелі розвідувального аналізу, просто заповнивши структуру даних, яку вимагає конкретний інструмент. Проте, використовуючи R, вам навіть не потрібно надсилати підготовлені дані до зовнішнього засобу представлення. Можна інформаційну панель веб-додатків безпосередньо з R.

Для більш глибокої обробки даних використовується також препроцесінг даних – попередня обробка даних таким чином, що з них виникають додаткові дані.

Препроцесінг даних – це процес очищення, перетворення та підготовки даних для подальшого аналізу та використання. Це включає в себе ряд операцій, таких як:

- обробка пропущених даних: включає в себе виявлення пропущених значень у даних та їх заповнення або видалення;
- кодування категоріальних даних: більшість алгоритмів машинного навчання працюють з числовими даними, тому необхідно перетворити категоріальні дані на числові;
- масштабування та нормалізація даних: це включає в себе масштабування даних до одного діапазону або нормалізацію даних до стандартного розподілу;
- видалення дублікатів: це включає в себе виявлення та видалення повторюваних рядків у даних;

- видалення викидів: це включає в себе виявлення та видалення даних, які відхиляються від нормального розподілу;
- вибір ознак: це включає в себе вибір найбільш інформативних ознак для використання у моделі.

Ці операції є необхідними для забезпечення якості та точності аналізу та моделювання даних. На рис. 2.3 зображено найважливіші кроки попередньої обробки даних

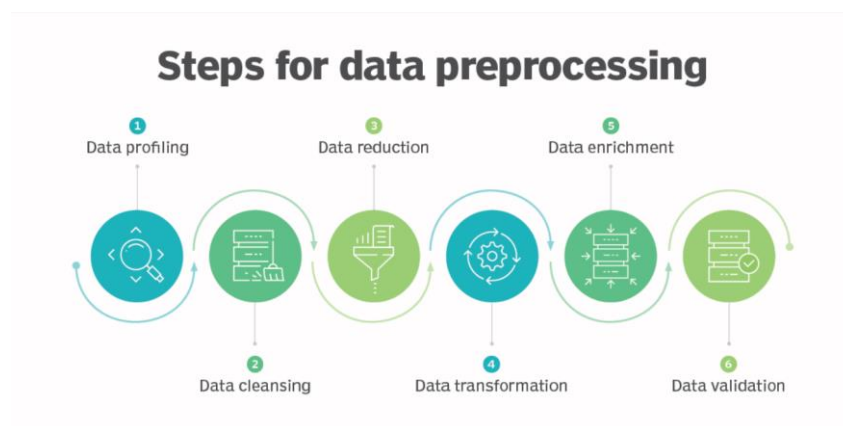


Рисунок 2.3 – Кроки попередньої обробки даних [17]

R, середовище аналізу даних і мова програмування з відкритим вихідним кодом, дозволяє користувачам виконувати ряд завдань, необхідних для ефективної обробки та аналізу великих даних. Однією з переваг R є те, що він полегшує процеси керування даними, такі як перетворення, підмножини та «очищення», і допомагає користувачам проводити пошуковий аналіз даних і готувати дані для статистичного тестування. R також містить численні готові до використання алгоритми машинного навчання та статистичного моделювання, які дозволяють користувачам створювати відтворювані дослідження та розробляти продукти даних.

Хоча існують альтернативні інструменти, наприклад Python, перевага R полягає в тому, що він єдина мова програмування з відкритим вихідним кодом, створена спеціально для статистичного аналізу. Він містить власні вбудовані статистичні алгоритми – величезна кількість математичних моделей і алгоритмів машинного навчання, доступних користувачам у пакетах базового R і сторонніх розробників, вражає й продовжує зростати.

Незважаючи на те, що обробку великих даних можна виконати за допомогою інших інструментів, саме коли ми переходимо до аналізу даних, R справді виділяється завдяки величезній кількості вбудованих статистичних формул і доступних алгоритмів сторонніх розробників. Це дозволяє користувачеві створювати вихідні дані в майже необмеженій кількості форматів, включаючи графіки, діаграми та інформаційні панелі [18].

2.2 Напрями використання методів машинного навчання для аналізу та прогнозування цін на вторинному авторинку

Для досягнення поставленої мети роботи будемо використовувати три різні моделі для прогнозування ціни.

1) Множинна лінійна регресія.

Завдяки своїй простоті, інтерпретації та швидкості лінійна регресія є одним із найбільш широко використовуваних алгоритмів штучного інтелекту в реальних задачах машинного навчання.

Множинна лінійна регресія – це техніка для розуміння зв'язку між однією залежною змінною та кількома незалежними змінними. Математично це можна відобразити на формулою (2.1):

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i + \varepsilon, \quad (2.1)$$

де y_i – залежна змінна;

x_i – пояснювальні змінні;

ε – термін «помилки» із середнім значенням 0 та деякою дисперсією σ^2 .

Навчання алгоритму лінійної регресії відбувається за допомогою методу звичайних найменших квадратів. Мета цього навчання в тому, що потрібно знайти вагові коефіцієнти факторів, що впливають на модель [10].

Щоб уникнути переобладнання моделі потрібно тренувати модель на навчальному наборі даних і створювати метрики оцінки на тестовому наборі.

2) Древа рішень

Древа рішень (DTs) – це непараметричний контрольований метод навчання, який використовується для класифікації та регресії. Мета полягає в тому, щоб створити модель, яка передбачає значення цільової змінної шляхом вивчення простих правил прийняття рішень, виведених із характеристик даних [19].

Древа рішень, як правило, є методом вибору для прогнозного моделювання, оскільки їх відносно легко зрозуміти, а також вони дуже ефективні. Основна мета дерева рішень – розділити сукупність даних на менші сегменти (рис. 2.4).

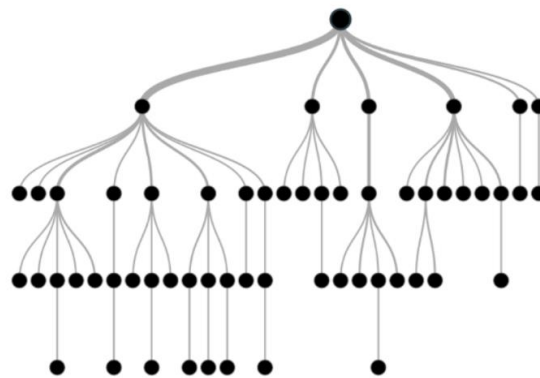


Рисунок 2.4 – Графічне відображення алгоритму дерева рішень [19]

Алгоритм CART широко використовується для створення дерев рішень за допомогою класифікації та регресії.

Найважливішим питанням під час дерева рішень виступає «який найкращий атрибут для кореневого вузла та підвузлів обрати?». Є два загальні варіанти:

- Метод найменших квадратів (2.2).

$$RSS = \sum_{i=1}^n (y_i - \widehat{y}_1)^2, \quad (2.2)$$

де y_i – фактичне (спостережуване) значення явища, що вивчається;

де $\widehat{y}_1$ – розрахункове значення моделі тренда;

n – число спостережень за явищем, що вивчається.

- Метод найменших абсолютних відхилень (2.3).

$$S = \sum_{i=1}^n |y_i - f(x_i)|, \quad (2.3)$$

де y_i – фактичне (спостережуване) значення явища, що вивчається;
 x_i – пояснювальні змінні.

У дереві рішень вузли розбиваються на підвузли на основі порогового значення атрибута. Алгоритм CART робить це шляхом пошуку найкращої однорідності для підвузлів за допомогою критерію індексу Джіні (2.4).

$$Gini = \sum_{i=1}^n p_i^2, \quad (2.4)$$

де p_i – імовірність вибору точки даних із класом i .

Кореневий вузол береться як навчальний набір і розбивається на два з урахуванням найкращого атрибута та порогового значення. Крім того, підмножини також розбиваються за тією ж логікою. Це триває до тих пір, поки в дереві не буде знайдено останню чисту підмножину або максимальну можливу кількість листків у цьому зростаючому дереві. Це також відоме як обрізка дерев [20].

Коли модель обмежується навчальним набором даних, то відбувається перенавчання моделі, що призводить до того, що модель не може узагальнити інші невідомі та невидимі набори даних. Для того, щоб уникнути перенавчання моделі варто використовувати метод обрізки. Цього можна досягнути:

- обмежуючи максимальну глибину дерева рішень;
- обмежуючи мінімальну кількість вибірок на простір рішень [21].

3) Рандомний ліс

Випадковий ліс – це контрольований алгоритм машинного навчання, який складається з алгоритмів дерева рішень.

Алгоритми випадкового лісу мають три основні гіперпараметри, які необхідно встановити перед навчанням. До них належать розмір вузла, кількість дерев і кількість відібраних функцій. Звідти класифікатор випадкового лісу можна використовувати для вирішення проблем регресії або класифікації.

Алгоритм випадкового лісу складається з набору дерев рішень, і кожне дерево в ансамблі складається із вибірки даних, отриманої з навчального набору із заміною, яка називається початковою вибіркою. З цієї навчальної вибірки одна третина відкладена як тестові дані, відомі як вихідна вибірка, до якої ми повернемося пізніше. Інший випадок випадковості потім вводиться через пакетування функцій, додаючи більше різноманітності до набору даних і зменшуючи кореляцію між деревами рішень. Залежно від типу проблеми визначення прогнозу буде різним. Для завдання регресії окремі дерева рішень будуть усереднені, а для завдання класифікації більшість голосів, тобто найпоширеніша категоріальна змінна, дасть прогнозований клас (рис. 2.5).

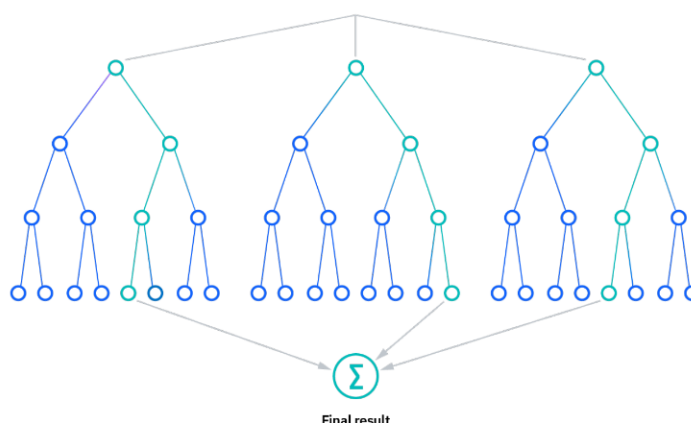


Рисунок 2.5 – Робота алгоритму випадкового лісу [22]

Наступні кроки пояснюють робочий алгоритм випадкового лісу [23]:

Крок 1: Вибрати випадкові зразки з заданих даних або навчального набору.

Крок 2. Цей алгоритм побудує дерево рішень для кожного навчального набору даних.

Крок 3: Голосування відбуватиметься шляхом усереднення дерева рішень.

Крок 4: Нарешті виберіть результат прогнозу, який отримав найбільшу кількість голосів, як кінцевий результат прогнозу.

Ключові переваги алгоритму [22]:

- Зменшення ризику переобладнання: дерева рішень мають ризик переобладнання, оскільки вони, як правило, щільно підходять для всіх зразків у навчальних даних. Однак, коли у випадковому лісі існує значна кількість дерев

рішень, класифікатор не буде переповнювати модель, оскільки усереднення некорельованих дерев зменшує загальну дисперсію та помилку передбачення.

- Забезпечує гнучкість: оскільки випадковий ліс може обробляти завдання як регресії, так і класифікації з високим ступенем точності, це популярний метод серед дослідників даних. Розміщення функцій також робить класифікатор випадкового лісу ефективним інструментом для оцінки відсутніх значень, оскільки він підтримує точність, коли частина даних відсутня.

- Легко визначити важливість функції: випадковий ліс дозволяє легко оцінити важливість змінної або внесок у модель. Є кілька способів оцінити важливість функції. Важливість Джіні та середнє зменшення домішок (MDI) зазвичай використовуються для вимірювання того, наскільки зменшується точність моделі, коли дана змінна виключається.

Для того, щоб перевірити наскільки добре працює модель існують різні показники. Розглянемо найбільш популярні з них [24].

- Середньоквадратична помилка (RMSE) . RMSE приймає значення MSE і застосовує до нього квадратний корінь. Він схожий на MSE, але набагато більш інтуїтивно зрозумілий для інтерпретації помилок. Це еквівалентно абсолютній похибці між нашою лінією лінійної регресії та будь-якою гіпотетичною точкою спостереження. На відміну від MSE та RSS (які використовують квадратичні значення), RMSE можна безпосередньо використовувати для інтерпретації «середньої помилки», яку робить наша модель прогнозування.

- R^2 або R-квадрат або R^2 оцінка. R-квадрат – це міра дисперсії залежної змінної, яку враховує наша лінійна функція. Цей показник є більш технічним, ніж два інших, тому він менш інтуїтивно зрозумілий для нестатистів. Як правило, значення R-квадрат, яке ближче до 1, є кращим, оскільки воно враховує більшу дисперсію.

2.3 Моделювання процесів ціноутворення та системи факторів на базі бустінгових моделей

Алгоритм градієнтного бустингу – це техніка, яка привертає увагу своєю швидкістю й точністю прогнозування, особливо з великими та складними даними.

Градiєнтне посилення – це тип посилення машинного навчання. Він покладається на інтуїцію, що найкраща наступна модель у поєднанні з попередніми моделями мінімізує загальну помилку передбачення. Ключова ідея полягає в тому, щоб встановити цільові результати для наступної моделі, щоб мінімізувати помилку. Як розраховуються цілі? Цільовий результат для кожного випадку в даних залежить від того, наскільки зміна прогнозу цього випадку впливає на загальну помилку прогнозу [25]:

- якщо невелика зміна прогнозу для випадку спричиняє значне зниження помилки, тоді наступним цільовим результатом випадку буде високе значення. Прогнози нової моделі, близькі до цілей, зменшать помилку.

- якщо невелика зміна прогнозу для випадку не спричиняє зміни помилки, тоді наступний цільовий результат випадку дорівнює нулю. Зміна цього прогнозу не зменшує помилку.

Підвищення градієнта назви виникає тому, що цільові результати для кожного випадку встановлюються на основі градієнта помилки щодо прогнозу. Кожна нова модель робить крок у напрямку, який мінімізує помилку передбачення, у просторі можливих передбачень для кожного випадку навчання.

Градiєнтний бустинг – це розширення бустингу, де процес адитивного генерування слабких моделей формалізовано як алгоритм градієнтного спуску над цільовою функцією. Підсилення градієнта встановлює цільові результати для наступної моделі з метою мінімізації помилок. Цільові результати для кожного випадку базуються на градієнті помилки (звідси й назва підвищення градієнта) щодо прогнозу [26].

Алгоритм побудови Gradient Boosting можна описуються наступними математичними кроками:

1. Створення базової моделі для прогнозування спостережень у навчальній виборці (2.5).

$$F_0(x) = \operatorname{argmin}_{\gamma} \sum_i^n L(y, \gamma), \quad (2.5)$$

де L – функція втрат;

argmin означає, що ми повинні знайти прогнозоване значення/гаму, для якої функція втрат є мінімальною.

2. Обчислення псевдо-залишків (2.6).

$$r_{im} = -a \left[\frac{\partial L(y_i, F(x_i))}{F(x_i)} \right]_{F(x)=F_{m-1}(x)}, \quad (2.6)$$

де a – швидкість навчання.

3. Знаходження вихідних значення для кожного листа дерева рішень (2.7).

$$\gamma_m = \operatorname{argmin}_{\gamma} \sum_i^n L(y_i, F_{m-1}(x) + \gamma h_m(x_i)), \quad (2.7)$$

де $h_m(x)$ – це наше дерево рішень, створене на основі цих залишків.

4. Оновлення прогнозів попередньої моделі (2.8).

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x), \quad (2.8)$$

де $F_{m-1}(x)$ – прогноз базової моделі (попередній прогноз).

GBDT ітеративно навчає ансамбль неглибоких дерев рішень, при цьому кожна ітерація використовує залишки помилок попередньої моделі, щоб відповідати наступній моделі. Остаточний прогноз є зваженою сумою всіх прогнозів дерева. Випадкове «розміщення» лісу мінімізує дисперсію та надмірне облаштування, тоді як «посилення» GBDT мінімізує зміщення та недообладнання.

XGBoost – це масштабована та високоточна реалізація посилення градієнта, яка розширює межі обчислювальної потужності для вдосконалених деревоподібних алгоритмів, створена в основному для підвищення продуктивності моделі машинного навчання та швидкості обчислень. За допомогою XGBoost дерева будуються паралельно, а не послідовно, як у GBDT. Він дотримується порівневої стратегії, скануючи значення градієнта та використовуючи ці часткові суми для оцінки якості розбиття на кожному можливому розбиванні в навчальному наборі [27].

Перелік переваг і атрибутів XGBoost великий і включає в себе наступне [28]:

- великий і постійно зростаючий список дослідників обробки даних у всьому світі, які активно сприяють розробці XGBoost з відкритим кодом;
- використання в широкому діапазоні програм, включаючи вирішення проблем регресії, класифікації, ранжування та визначених користувачем завдань прогнозування;
- бібліотека, яка є високопортативною та зараз працює на платформах OS X, Windows і Linux;
- хмарна інтеграція, яка підтримує AWS, Azure, кластери Yarn та інші екосистеми;
- активне виробниче використання в багатьох організаціях у різних вертикальних ринкових зонах;
- бібліотека, яка була створена з нуля, щоб бути ефективною, гнучкою та портативною.

РОЗДІЛ 3

ПРАКТИЧНЕ ЗАСТОСУВАННЯ МАТЕМАТИЧНИХ МОДЕЛЕЙ ФОРМУВАННЯ ЦІН НА ВТОРИННОМУ АВТОРИНКУ

3.1 Аналіз результатів візуалізації та препроцесингу даних

Інформація буде сильний та надійний фундамент для будь-якого підприємства, компанії чи навіть людини. Володіння інформацією дає змогу провести певний розвідувальний аналіз та зробити висновки про компанію, товар для покупця, а також знайти кращі шляхи вирішення будь-якої проблем та покращити прибутки для компаній [29].

Це стосується і рішення покупки автомобіля, адже зараз перед тим, як прийняти таке важливе рішення клієнт проводить досліджуваний аналіз характеристик того чи іншого автомобіля, щоб обрати кращий для себе.

Розвідувальний аналіз є одним з найважливішим у моделюванні, так як саме це впливає на якість побудованих в майбутньому моделей.

Для даної бакалавратської роботи використовується набір даних [30], який містить в собі інформацію з платформи продажів автомобілів, що уже були у використанні. У даному датасеті Міститься наступна інформація:

- Age_of_car – вік машини;
- Year – рік випуску;
- Location – місце знаходження;
- Mileage – стандартний пробіг, запропонований оригінальною компанією;
- Kilometers_Driven – пробіг в кілометрах;
- Fuel_Type – тип палива;
- Transmission – тип коробки передач;
- Owner_Type – скільки було нових власників;

- Engine – об’єм двигуна (числовий варіант колонки Engine);
- Power – максимальна потужність двигуна (числовий варіант колонки Power);
- Seats – кількість сидінь в машині;
- Name – модель автомобіля;
- Brand – марка автомобіля.

Деякі із цих колонок були штучно створені, наприклад колонка Age_of_car показує різницю між 2023 роком та роком випуску машини, для кращого аналізу впливу даного показника на ціну (рис. 3.1).

```
data <- data %>% mutate(Age_of_car = 2023-Year, .before = 0) %>% mutate(Age_of_car=as.numeric(Age_of_car))
```

Рисунок 3.1 – Створення колонки Age_of_car

Джерело: розроблено автором за даними [30]

Колонка Brand була створена для візуалізаційного аналізу інформації. Створення вона за допомогою кожного першого слова із колонки Name (рис. 3.2).

```
library(stringr)
data$brand <- word(data$Name, 1)
```

Рисунок 3.2 – Створення колонки Brand

Джерело: розроблено автором за даними [30]

Також були проведені певні маніпуляції з іншими даними. Для кращого аналізу та розуміння колонку Price було переведено до грошового формату дол. США.

Після даних перетворень даних, був проведений аналіз даних на наявність пропущених значень. Як виявилось були присутні пропущені значення в декількох колонках (рис.3.3).

##	Age_of_car	Name	Location	Year
##	0	0	0	0
##	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type
##	0	0	0	0
##	Mileage	Engine	Power	Seats
##	2	36	36	42
##	Price	brand		
##	0	0		

Рисунок 3.3 – Кількість пропущених значень в даних

Джерело: розроблено автором за даними [30]

Для якісного аналізу ми позбулися всіх пропущених значень. Тому після відсіювання непотрібної інформації під час розвідувального аналізу наш датасет містить 5975 спостережень та 14 змінних.

На даному етапі можна перейти до візуалізації даних.

Серед 5975 спостережень зустрічається 31 унікальна марка автомобіля. Їхня густота достатньо різна: присутні марки машин, які зустрічаються більше ніж сто разів, є такі, що зустрічаються тільки один раз. Тому для того, щоб результати були коректними буде доцільно звужити досліджуване коло марок машин та залишити тільки ті, які зустрічаються більше ніж 14 разів. Після даної маніпуляції залишилось 23 унікальних марки.

На рис. 3.4 побудований графік 10-ти найбільш популярних марок машин на вторинному авторинку серед досліджуваних спостережень.

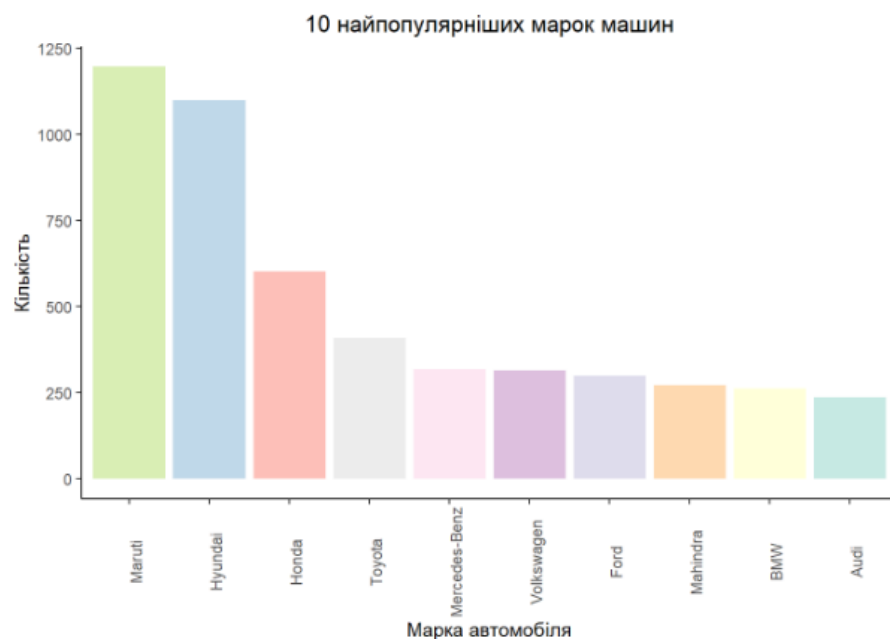


Рисунок 3.4 – Найбільш популярні марки автомобілів

Джерело: розроблено автором за даними [30]

Як бачимо найбільш популярними марками серед нашого дослідження є: Maruti, Hyundai, Honda. При тому що кількість оголошень з двома першими найбільш популярними марками машин перевищує 1000 оголошень.

Проведемо аналогічний аналіз із моделями автомобілів. Серед спостережень є 1838 унікальних моделей автомобілів. Велика кількість з яких

зустрічається на досліджуваному ринку не більше ніж 1 раз. Тому було б доцільно позбутись таких марок машин. Після зниження реєстру отримуємо 641 унікальних значень, які виставлені на продаж більше, ніж 3 рази, найбільш популярні з них можна побачити на рис. 3.5.

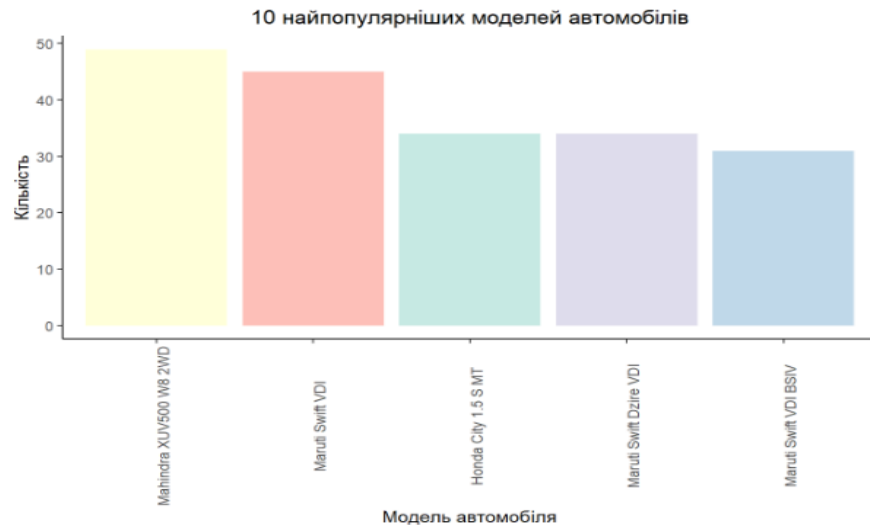


Рисунок 3.5 – Найбільш популярні моделі автомобілів

Джерело: розроблено автором за даними [30]

Так як раніше ми визначили найбільш популярні марки машин, на рис. 3.6 визначимо найбільш популярні моделі машин в розрізі трьох найбільш популярних марок автомобілів.

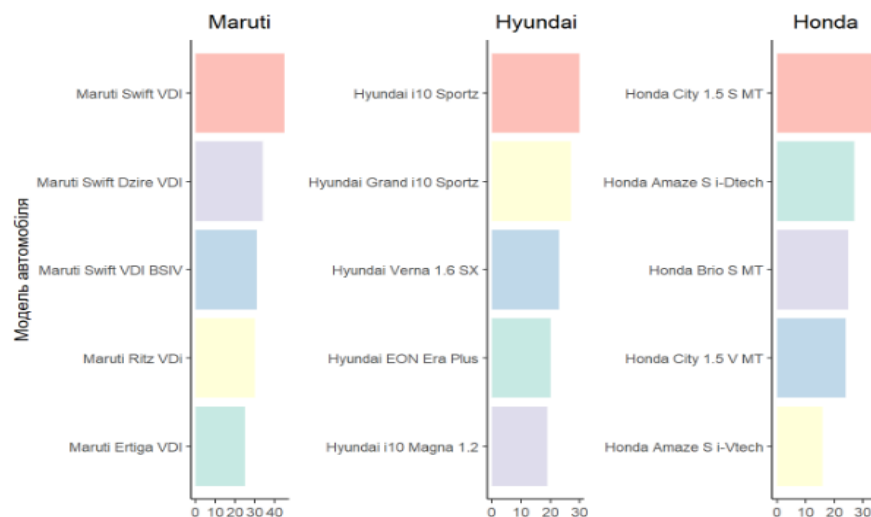


Рисунок 3.6 – Найпопулярніші моделі машин в розрізі найпопулярніших марок

Джерело: розроблено автором за даними [30]

Після попередніх видалень даних кінцевий набір даних тепер становить 4317 спостережень.

На рис. 3.7 подивимось на розподіл автомобілів за різними факторами, такими як: вік, ціна, потужність двигуна, об'єм двигуна автомобіля.

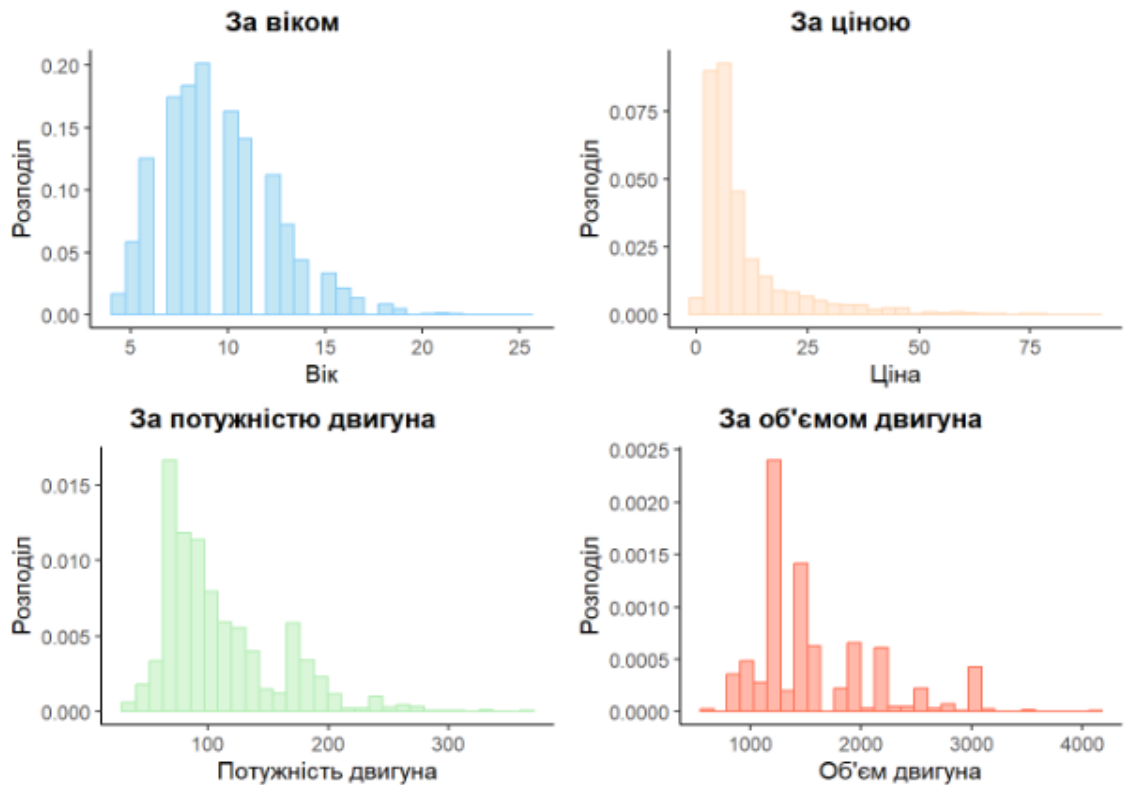


Рисунок 3.7 – Розподіл автомобілів за різними факторами

Джерело: розроблено автором за даними [30]

Згідно з результатами розподілу більшість автомобілів віку 7-13 років, але найбільша кількість машин мають 9 років. Що ж стосується ціни, то найбільше ціни варіюються в проміжку між 5-11 тис. дол. США, середня вартість автомобілів становить 10,87 тис. дол. США. У розподілах за потужністю та об'ємом двигуна не прослідковується чіткі межі густоти розсіювання. Середня потужність двигуна становить 110,60 кінських сил. Для об'єма двигуна середнє значення становить 1583 кубів.

На рис. 3.8 зображено розподіл машин за типом коробки передач у відсотках.

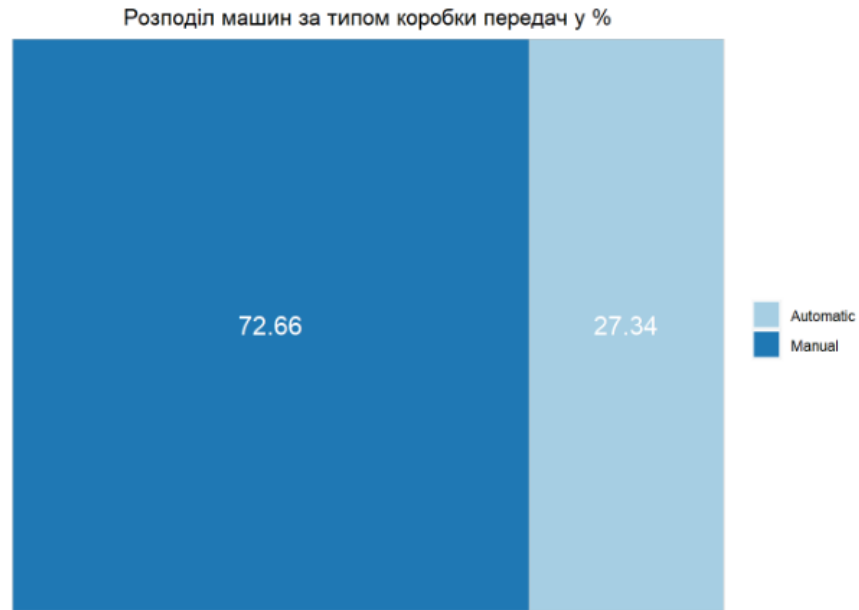


Рисунок 3.8 – Розподіл машин за типом коробки передач, у %

Джерело: розроблено автором за даними [30]

Як бачимо більшість, а саме 72,66% машин в оголошеннях мають ручну коробку передач, та тільки 27,34% – механічну.

Важливим етапом візуалізації даної роботи також буде відобразити взаємозалежність ціни на автомобілі від інших важливих характеристик.

На рис. 3.9 зображено взаємозалежність ціни від віку машини.

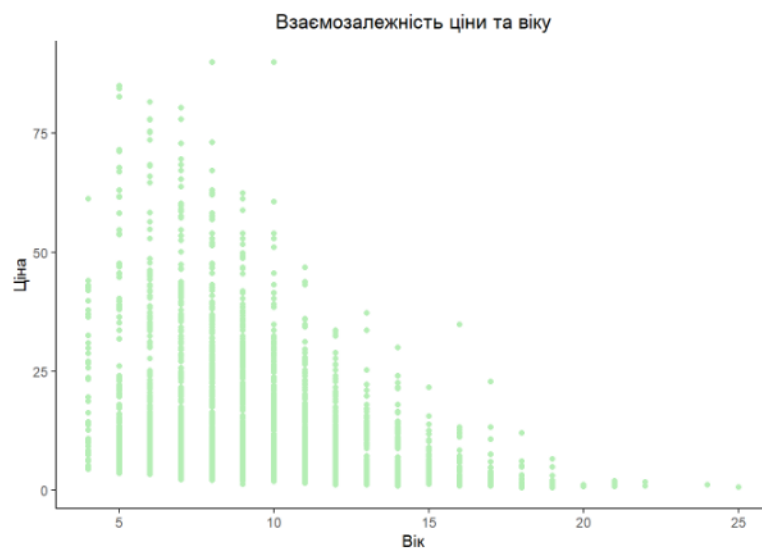


Рисунок 3.9 – Взаємозалежність ціни автомобіля та віку

Джерело: розроблено автором за даними [30]

Бачимо логічний взаємозв'язок: чим старша машина, тим менше вона коштує і навпаки.

Тепер розглянемо залежність ціни та потужності двигуна. Як бачимо на рис. 3.10 чим потужніший двигун тим більша ціна на автомобіль, тому що характеристика автомобіля буде кращою.

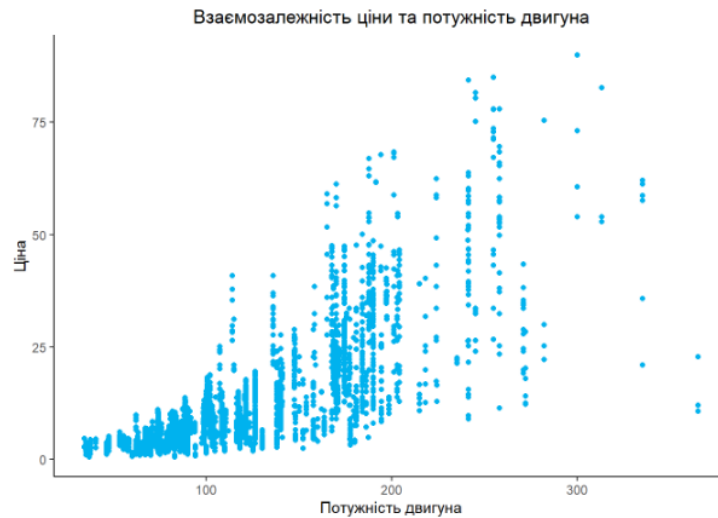


Рисунок 3.10 – Взаємозалежність ціни та потужності двигуна

Джерело: розроблено автором за даними [30]

Розглянемо також залежність ціни на автомобіль від фактора, який ми ще не розглядали. На рис. 3.11 зображено залежність ціни від кількості сидінь в ньому. Як можна побачити, важко прослідкувати якусь певну залежність між даними показниками. Впевнено можна сказати, що кількість сидінь 5 та 7 є найбільш стандартними для автомобілів, ціни на них розподілені майже рівномірно.

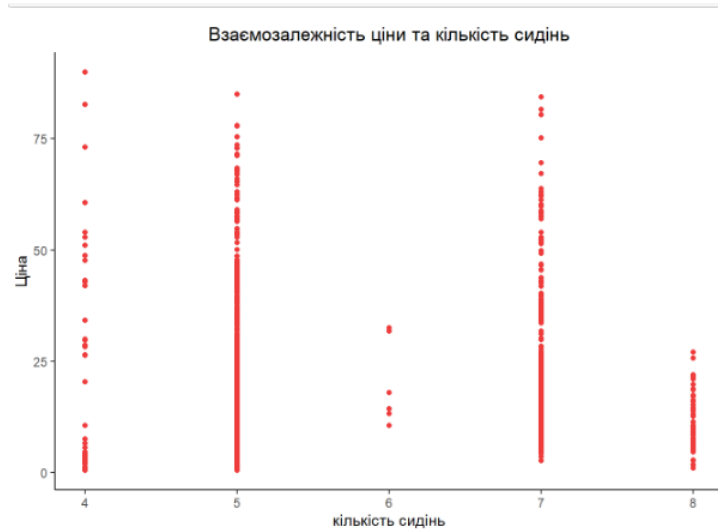


Рисунок 3.11 – Взаємозалежність ціни та кількість сидінь

Джерело: розроблено автором за даними [30]

На рис. 3.12 розглянуто розподіл ціни в залежності від типу палива автомобіля.



Рисунок 3.12 – Взаємозалежність ціни та типу палива

Джерело: розроблено автором за даними [30]

Як бачимо, для типу палива CNG ціни на автомобіль достатньо низькі, для Petrol – середні, а для Diesel – порівняно з іншими високі.

3.2 Аналіз результатів моделювання факторів формування цін за допомогою інструментів машинного навчання

Для того, щоб почати безпосередньо моделювання даних, важливим кроком повинно бути переведення даних, що містять текст у фактори (рис. 3.13). Завдяки переведення інформації до факторної унікальне значення зберігається тільки один раз, а дані зберігаються у вигляді вектора чисел [31].

```
data$Fuel_Type<- as.factor(data$Fuel_Type)
data$Transmission<- as.factor(data$Transmission)
data$Owner_Type<- as.factor(data$Owner_Type)
data$brand<- as.factor(data$brand)
```

Рисунок 3.13 – Переведення текстового типу даних на факторний

Джерело: розроблено автором за даними [30]

Дана маніпуляція дає нам змогу правильно включити дані, які містять в собі текстову інформацію в моделювання ціни на автомобілі.

Також наш датасет був розділений на тестову та навчальну виборки. Будуватися модель буде на навчальних даних, а ефективність буде перевірятися на тестових даних. В моєму випадку навчальний набір становить 80% всього датасету, коли тестовий – 20%.

Почнемо наше моделювання з побудови лінійної моделі, як було сказано раніше, на основі навчальних даних. Після навчання лінійної моделі використовуємо функцію `summary()` для отримання загального огляду даних, які ми аналізуємо (рис. 3.14).

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.230e+01  1.511e+00  8.143 5.34e-16 ***
Age_of_car   -1.122e+00  3.864e-02 -29.040 < 2e-16 ***
Kilometers_Driven -1.753e-05  3.106e-06 -5.643 1.81e-08 ***
Fuel_TypeDiesel  1.088e+00  9.737e-01  1.117 0.26388
Fuel_TypePetrol  4.482e-01  9.710e-01  0.462 0.64444
TransmissionManual -4.888e-01  3.092e-01 -1.581 0.11400
Owner_TypeFourth & Above 2.593e+00  2.639e+00  0.982 0.32597
Owner_TypeSecond -8.041e-01  2.613e-01 -3.077 0.00211 **
Owner_TypeThird  -7.967e-01  8.160e-01 -0.976 0.32898
Seats         1.313e+00  2.059e-01  6.380 2.01e-10 ***
brandBMW      -2.547e+00  6.277e-01 -4.057 5.07e-05 ***
brandChevrolet -1.236e+01  8.742e-01 -14.136 < 2e-16 ***
brandFiat     -1.336e+01  2.224e+00 -6.009 2.07e-09 ***
brandFord     -1.124e+01  7.385e-01 -15.215 < 2e-16 ***
brandHonda    -1.345e+01  6.277e-01 -21.424 < 2e-16 ***
brandHyundai   -1.240e+01  6.209e-01 -19.972 < 2e-16 ***
brandJaguar    6.639e-01  1.193e+00  0.557 0.57782
brandJeep     -8.847e+00  2.064e+00 -4.286 1.87e-05 ***
brandLand     8.068e+00  1.237e+00  6.523 7.91e-11 ***
brandMahindra -1.683e+01  8.037e-01 -20.941 < 2e-16 ***
brandMaruti   -1.142e+01  6.487e-01 -17.611 < 2e-16 ***
brandMercedes-Benz 1.387e+00  5.811e-01  2.387 0.01705 *
brandMini     6.327e+00  1.595e+00  3.968 7.41e-05 ***
brandMitsubishi -1.134e+01  1.951e+00 -5.813 6.69e-09 ***
brandNissan   -1.249e+01  9.756e-01 -12.805 < 2e-16 ***
brandPorsche  2.909e+01  2.719e+00  10.700 < 2e-16 ***
brandRenault  -1.276e+01  8.305e-01 -15.367 < 2e-16 ***
brandSkoda    -1.279e+01  8.257e-01 -15.494 < 2e-16 ***
brandTata     -1.371e+01  9.591e-01 -14.291 < 2e-16 ***
brandToyota   -1.115e+01  7.346e-01 -15.177 < 2e-16 ***
brandVolkswagen -1.268e+01  6.824e-01 -18.580 < 2e-16 ***
brandVolvo    -6.743e+00  3.075e+00 -2.193 0.02839 *
power_num     8.496e-02  6.079e-03  13.976 < 2e-16 ***
engine_num     2.512e-03  5.644e-04  4.450 8.85e-06 ***

```

Рисунок 3.14 – Результат функції `summary()` для лінійної моделі

Джерело: розроблено автором за даними [30]

Як бачимо на рис. 3.14 більшість наших даних є статистично значущими. Це означає, що вони достатньо сильно корелюють зі змінною, а саме ціною, та відповідно впливають на неї.


```
Linear Regression RMSE: 7.267369
Linear Regression R-squared: 0.625843
```

Рисунок 3.16 – Оцінки ефективності лінійної моделі

Джерело: розроблено автором за даними [30]

Наступною моделлю, яку ми розглянемо буде модель дерева рішень. Якщо розглядати вплив чинників на дану модель на рис. 3.17 то можна зробити висновок, що на дану модель мають найбільший вплив такі характеристики автомобіля як: потужність та об'єм двигуна, при чому вони впливають на модель візуально однаково, а також вік автомобіля та його марка.

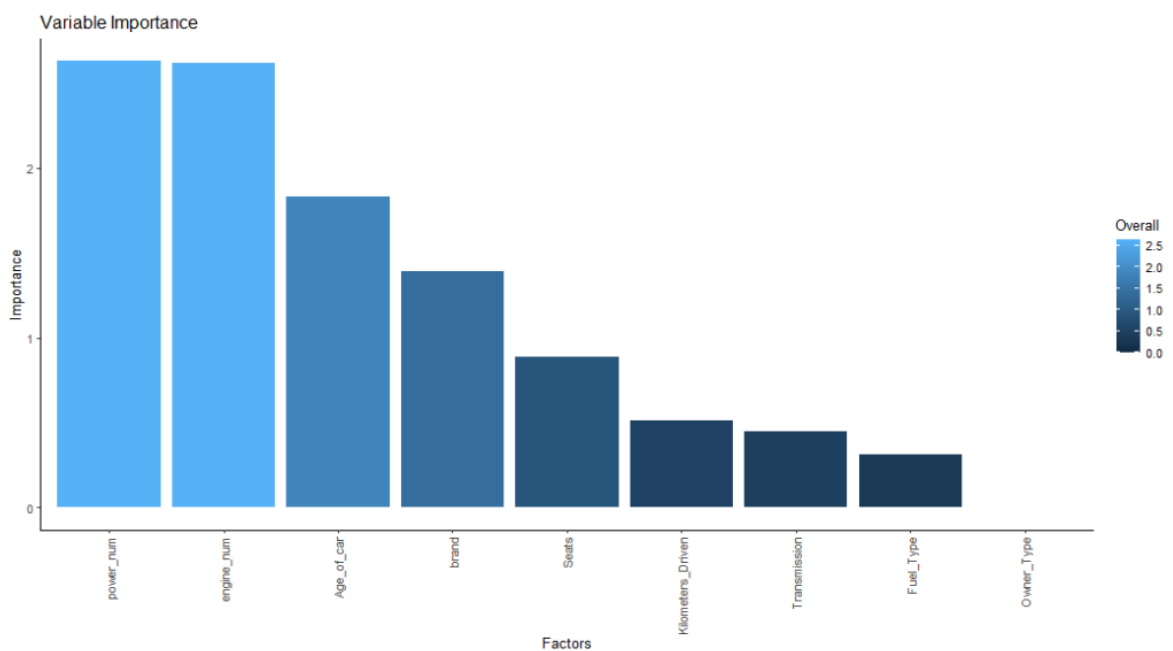


Рисунок 3.17 – Важливість змінних, що впливають на модель дерева рішень

Джерело: розроблено автором за даними [30]

Наступним кроком буде виконання прогнозування на тестовому наборі даних та оцінення ефективності моделі дерева рішень.

Як бачимо на рис. 3.18 R квадрат даної моделі становить 0,843 – модель може пояснити 84,3% варіації, інші 15,7% – невідомі, так як не враховані в наш аналіз, показник достатньо високий для того, щоб визнати модель економічно значущою. Що ж стосується показника RMSE, то прогнозована ціна відрізняється від експериментальної на приблизно 4,61 тис. дол. США.

Decision Tree RMSE: 4.607801
Decision Tree R-squared: 0.8436032

Рисунок 3.18 – Оцінки ефективності моделі дерева рішень

Джерело: розроблено автором за даними [30]

Далі побудуємо модель випадкового лісу. Так як ми не можемо використати функцію `summary()` для даної моделі, то одразу перейдемо до графічного зображення показників, які мають найбільший вплив на цю модель. Адже саме так можна зрозуміти, які показники можна опустити з моделі для того, щоб зробити її простішою та більш гнучкою для прогнозування у майбутньому.

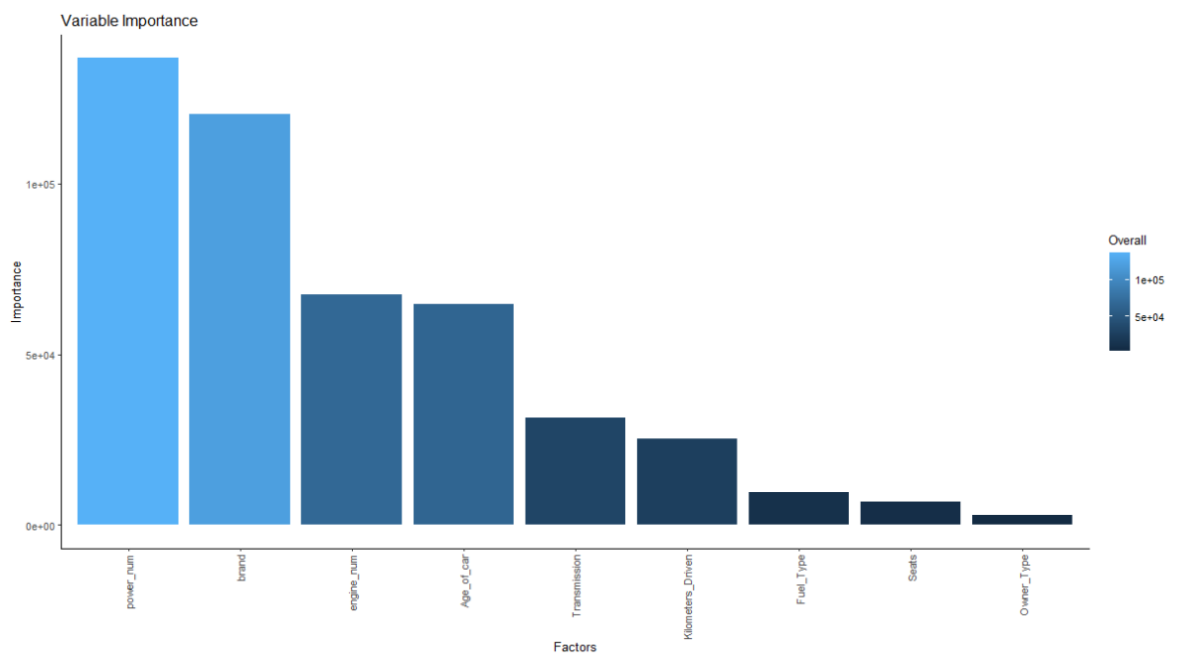


Рисунок 3.19 – Важливість змінних, що впливають на модель випадкових лісів

Джерело: розроблено автором за даними [30]

Для моделей випадкових лісів важливість визначається обчисленням відносного впливу кожної змінної: чи ця змінна була обрана для поділу під час процесу побудови дерева, і наскільки квадрат помилки (для всіх дерев) покращився (зменшився) у результаті обрахунків. [32].

Як бачимо на рис. 3.19 вплив будь-якого показника є дуже низьким, проте найбільший з них вплив на дану модель має числова характеристика автомобіля – потужність двигуна. Також марка автомобіля та об'єм двигуна входять у топ-3 найвпливовіших змінних.

Знову ж виконаємо прогнозування на тестовому наборі даних та оцінимо ефективність моделі випадкових лісів. На рис. 3.20 бачимо, що дана модель за своїми оцінками ефективності краща за попередню, так як R квадрат становить 0,935, тобто 93,5% варіації пояснюється даною моделлю, що вважається дуже високим показником для того, щоб використовувати модель випадкових лісів для подальшого аналізу та прогнозування. Показник RMSE вказує на те, що прогнозована ціна відрізняється від експериментальної на приблизно 2,98 тис. дол. США, що також вважається дуже високим показником. Варто згадати, що чим нижчий показник RMSE, тим краще є модель для прогнозування.

Random Forest RMSE: 2.985819
Random Forest R-squared: 0.935468

Рисунок 3.20 – Оцінки ефективності моделі випадкових лісів

Джерело: розроблено автором за даними [30]

Далі можна приступити до порівнювання результатів, які ми отримали під час моделювання та визначити, яка модель з розглянутих буде найкращою.

Model <chr>	RMSE <dbl>	R_squared <dbl>
Linear Regression	7.267369	0.6258430
Random Forest	2.985819	0.9354680
Decision Tree	4.607801	0.8436032

Рисунок 3.21 – Порівняння оцінок ефективності моделей

Джерело: розроблено автором за даними [30]

Для цього розглянемо оцінки ефективності трьох моделей на рис. 3.21, як бачимо, найкращий показник такої важливої оцінки як R квадрат у моделі випадкових лісів, а найгірший – у лінійній моделі. Що ж стосується іншого важливого показника RMSE, то він найкращий також у моделі випадкових лісів та найгірший – у лінійній.

Проаналізувавши результати та оцінивши їхні показники ефективності можна зробити висновок, що з розглянутих моделей найкращі показники у моделі випадкових лісів, на яку найбільший вплив чинять такі показники як потужність двигуна, об'єм двигуна, а також марка автомобіля.

3.3 Аналіз результатів моделювання процесів ціноутворення за допомогою методів градієнтного бустингу

Моделювання ціни на вторинному авторинку в рамках даного підрозділу буде здійснено за допомогою градієнтного бустингу.

Спочатку ми розділяємо наш датасет на дві вибірки: навчальну та тестову, з таким же співвідношенням, як ми це робили раніше, 80% та 20% відповідно. Також додатково переводимо їх у формат матриці для того, щоб правильно відображалась робота моделі. Що ж стосується маніпуляції з початковим датасетом, то повторюємо всі ті кроки, які були зроблені раніше, а саме: розрахунок віку машини, визначення бренду, конвертація валюти до доларів США, зменшення виборки з метою покращення аналізу та інше.

Першою побудуємо модель, яка будується на основі параметрів, які виступають за замовчуванням методу. Для кращого розуміння, параметри за замовчуванням включають в себе: швидкість навчання моделі 0,3. За замовчуванням методу градієнтного бустингу кількість дерев 6, а мінімальна кількість вузлів що може бути становить тільки 1.

В даному випадку у нас відбувається 10–кратна перехресна перевірка, в якій максимальна кількість ітерація становить 1000. Для ролі первинної моделі, що буде проводити навчання остаточної, було обрано лінійну.

Отримали результати зображені на рис. 3.22.

```
##      ntrees.train rmse.train ntrees.test rmse.test
## 1              924  0.8249258         46 0.8347485
```

Рисунок 3.22 – Результати моделі на основі параметрів за замовчуванням

Джерело: розроблено автором за даними [30]

Графічно ці ж результати виглядають наступним чином (рис.3.23):

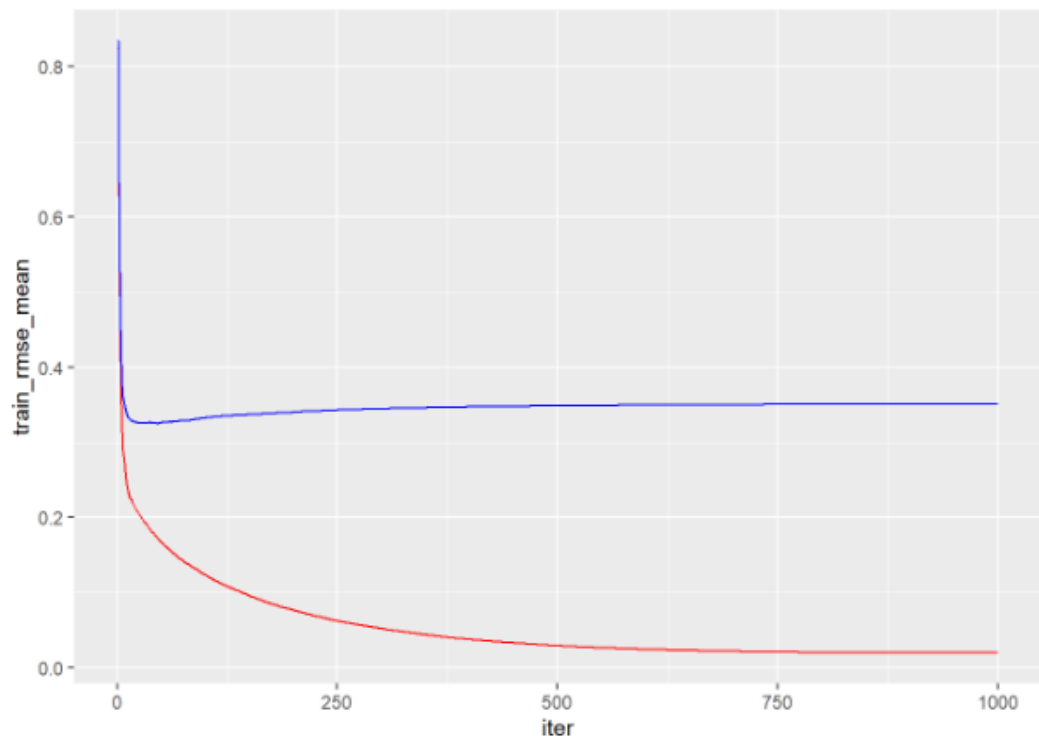


Рисунок 3.23 – Візуалізація результатів моделі на основі параметрів за замовчуванням

Джерело: розроблено автором за даними [30]

Щодо побудованої моделі можна зробити наступні висновки: було побудовано 924 дерева для навчальної вибірки, а також 46 для тестової. Значення показника RMSE становить 0,82 та 0,83 для тренувальної та навчальної виборок відповідно. Даний показник вказує на рівень помилок моделі на поточному кроці та слугує показником якості моделі. Як бачимо значення на тренувальній та навчальній виборках практично однакові.

Як можна побачити з рис. 3.23 для нашої навчальної виборки значення RMSE починає спадати приблизно після 40 проведеної ітерації. Що ж стосується тестової, то після приблизної тої ж ітерації значення RMSE набуває певної стабільності.

Наступним нашим кроком буде побудова другої моделі алгоритму. Цього разу параметри будуть не за замовчуванням, їх ми встановимо самостійно. Швидкість навчання становитиме 0,1, коефіцієнт регуляції для рідж-регресії – 0,5, коефіцієнт регуляції лассо-регресії дорівнюватиме 1. Максимальна кількість припустимих дерев дорівнюватиме 6, а мінімальна кількість дочірніх елементів –

2. Відсоток даних, які будуть навчатись для кожного дерева становитиме 80% загальних даних, частка стовпців для випадкових вибірок для кожного з дерев становитиме також 80% (рис. 3.24).

```
xgb_params <- list(
  eta = 0.1,
  lambda = 0.5,
  gamma = 1,
  max_depth = 6,
  min_child_weight = 2,
  subsample = 0.8,
  colsample_bytree = 0.8
)
```

Рисунок 3.24 – Задані параметри для навчання моделі

Джерело: розроблено автором за даними [30]

У цьому випадку будемо будувати модель з 5-ти кратною перехресною перевіркою. Кількість ітерацій збільшимо до максимального значення 5000. Лінійну модель знову ж використаємо в ролі початкової, яка розпочне навчання за алгоритмом градієнтного бустингу. Результати навчання можна побачити на рис. 3.25.

```
##      ntrees.train rmse.train ntrees.test rmse.test
## 1             173    1.01582       163    1.018516
```

Рисунок 3.25 – Результати моделі на основі заданих параметрів

Джерело: розроблено автором за даними [30]

Для навчальної виборки було побудовано 173 дерева, тоді коли для тестової – 163. Значення RMSE для обох виборок перевищує 1 та становить 1,016 для навчальної виборки, 1,019 – для тестової. Ці результати зображено графічно на рис. 3.26.

Як бачимо, коефіцієнт RMSE спадає на всьому досліджуваному інтервалі, при тому і для навчальної, і для тестової вибірки.

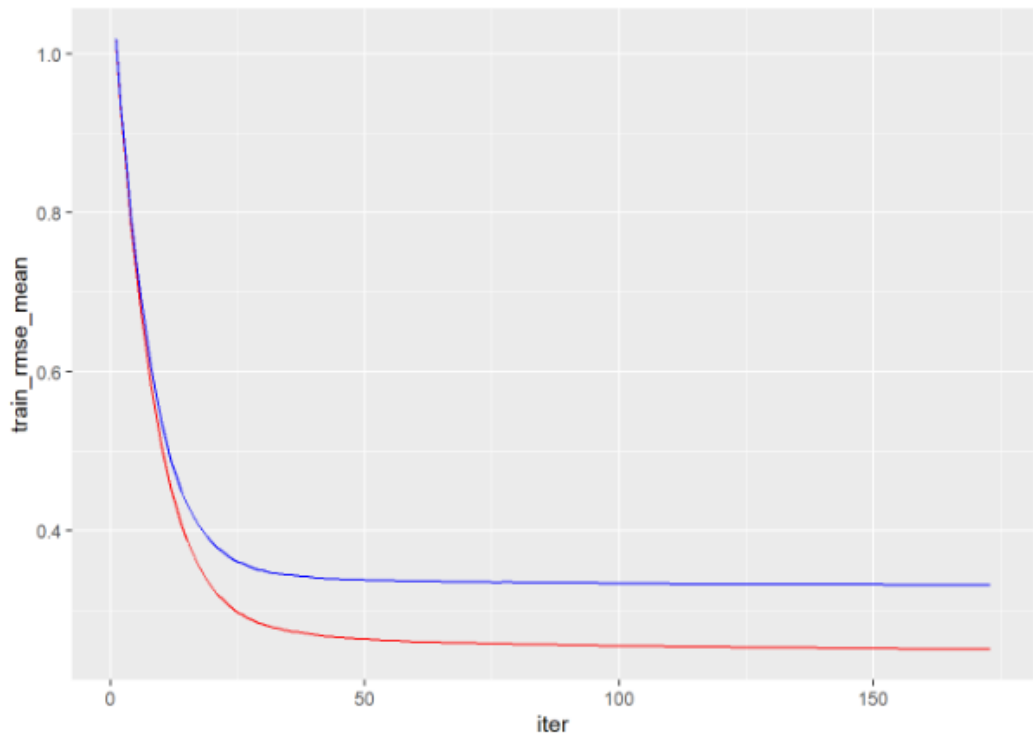


Рисунок 3.26 – Візуалізація результатів моделі на основі заданих параметрів

Джерело: розроблено автором за даними [30]

Побудуємо ще одну, останню модель за допомогою алгоритму градієнтного бустингу. Ця модель будується за допомогою налаштування гіпермаркетів. Для налаштування сітки гіпермаркетів були встановлені наступні показники: швидкість навчання моделі на рівні 0,01, 0,5, 0,1 та 0,3; коефіцієнт регуляції лассо-регресії на рівнях 0, 0, та 5; коефіцієнт регуляції для рідж-регресії – 0, 0,25 або 1; максимальна кількість дерев становить 1, 3, 5, 7; мінімальна кількість дочірних елементів – 1, 3, 5, 7; для кожної виборки кожного дерева буде навчатись 65% даних; частка стовпців для випадкових вибірок для кожного дерева становитиме 75%, 90% або 100%. Нашою метою є мінімізація показника RMSE ціни. В кінцевому результаті ми отримали 1728 значень настроюваної нами сітки гіпермаркетів. Виконавши пошук результатів, де результат навчання моделі був найкращий торимали результат зображений на рис. 3.27.

З отриманого результату можемо визначити, що найкращі моделі були отримані в результаті 1-4 ітерації, так як відсутні коефіцієнти лямбда та гама, але

так як швидкість навчання найкраща у 4-ої ітерації, то цю модель можна вважати найкращою.

	eta <dbl>	gamma <dbl>	lambda <dbl>	max_depth <dbl>	min_child_weight <dbl>	subsample <dbl>	colsample_bytree <dbl>	optimal_trees <dbl>	min_RMSE <dbl>
1	0.01	0.00	0	1	1	0.65	0.75	0	0
2	0.05	0.00	0	1	1	0.65	0.75	0	0
3	0.10	0.00	0	1	1	0.65	0.75	0	0
4	0.30	0.00	0	1	1	0.65	0.75	0	0
5	0.01	0.25	0	1	1	0.65	0.75	0	0
6	0.05	0.25	0	1	1	0.65	0.75	0	0
7	0.10	0.25	0	1	1	0.65	0.75	0	0
8	0.30	0.25	0	1	1	0.65	0.75	0	0
9	0.01	1.00	0	1	1	0.65	0.75	0	0
10	0.05	1.00	0	1	1	0.65	0.75	0	0

Рисунок 3.27 – Результат навчання за методом сітки гапермаркетів

Джерело: розроблено автором за даними [30]

На рис. 3.28 можна побачити, які параметри найбільше впливали на визначену нами найкращу модель.

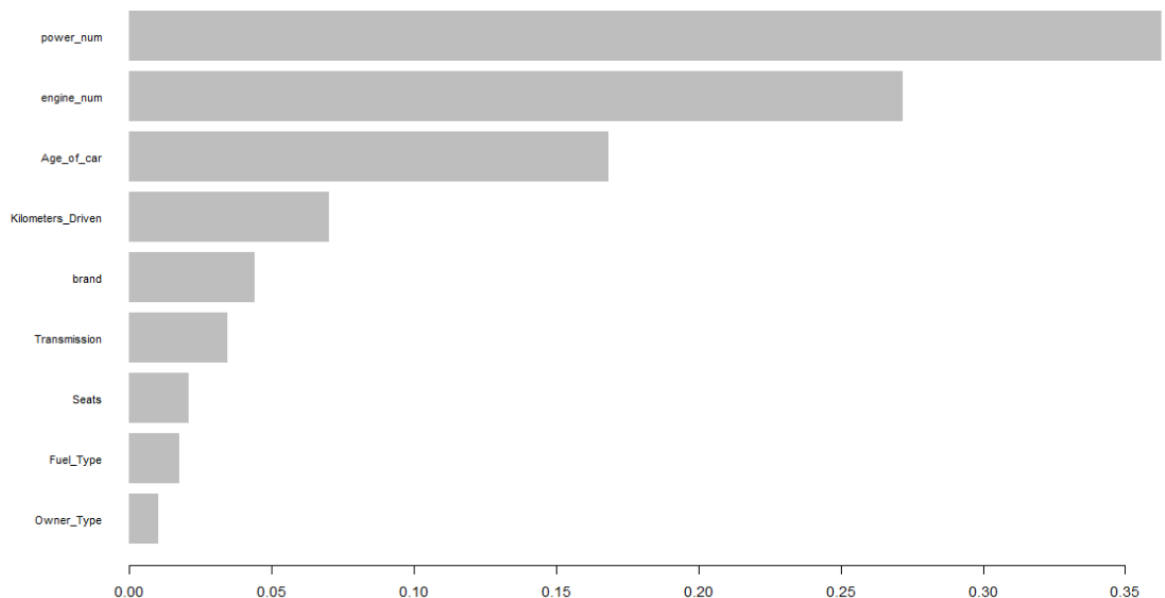


Рисунок 3.28 – Важливість змінних

Джерело: розроблено автором за даними [30]

В результаті проведеного аналізу за допомогою алгоритму градієнтного бустингу, ми отримали модель, похибка варіації у якої становить 0, значення індексу Джині – 0,75. Найбільший вплив на модель мають такі відомі нам характеристики як: потужність та об'єм двигуна, а також вік машини та кількість кілометрів, які проїхав автомобіль, також свій вплив на ціну мають такі інші характеристики як марка машини, тип коробки передач, кількість сидінь в машині, тип палива та скільки попередніх власників було.

Фінальним етап роботи буде порівняння всіх показників, які були отримані в результаті побудови різних моделей.

У табл. 3.1 спостерігаємо показники R-квадрата та RMSE для всіх досліджуваних моделей:

Таблиця 3.1 – Отримані результати оцінювання моделей

Model	RMSE	R_squared
Linear Regression	7,267369	0,625843
Random Forest	2,9851819	0,935468
Decision Tree	4,607801	0,8436032
XGBoost	1,018516	0,9452145

Джерело: розроблено автором за даними [30]

Модель, побудована за методом градієнтного бустингу має найкращі показники: найвищий R-квадрат – 94,9% та найменший RMSE показник – 1,02. Одні з найкращих результатів також були отримані методом навчання моделі випадкових лісів. R-квадрат становить 93,5%, а RMSE показник – 2,99.

Загалом результати та показники оцінки моделей на економічну значущість достатньо високі, на що безпосередньо мало вплив попереднє правильне очищення даних. Всі із розглянутих моделей можуть бути вдосконаленими та розгянутими для практичного використання, а саме прогнозування ціни на автомобіль, який був у використанні.

ВИСНОВКИ

З'ясовано, що автомобільний ринок – це ринок на якому здійснюється операція покупки–продажу автомобіля, що включає в себе як нові автомобілі так і ті, які були у використанні, а також всі пов'язані з цим послуги, такі як страхування, технічне обслуговування та інше.

Визначено, що основними функціями авторинку є: купля–продаж автомобілів, інформаційна підтримка, підтримка учасників ринку та регулююча функція на ціни на автомобіль.

Так як автомобільний ринок – олігополія, то за даними 2022 року найпопулярнішою маркою машини, що продавалась на ринку сатала Toyota Corolla. А найвідомішою онлайн–платформою продажу автомобілів, що були у використанні став Alibaba.com.

Автомобільний ринок є важливим напрямком автомобільної галузі так як цей ринок є значним чинником у формуванні світової економіки. За даними Data Bridge Market Research, у 2022 році світовий ринок вживаних автомобілів оцінювався в 996 906,42 мільйона доларів США, а до 2030 року очікується, що він досягне 1700 106,13 мільйона доларів США при CAGR 6,90% протягом прогнозованого періоду.

Ціна слугує дуже важливим чинником впливу на розвиток автомобільного ринку, так як саме за допомогою ціни формується попит на автомобілі. Отже, нами було проаналізовано які чинники найбільше впливають на ціноутворення на вторинному авторинку, серед них виявились такі показники як марка машини, вік машини, кількість проїханих кілометрів, об'єм та потужність двигуна, також з'ясовано, що має вплив також історія технічного обслуговування, зовнішній вигляд та навіть колір машини.

Досліджено основні техніки аналізу та математичний інструментарій побудови моделей множинної лінійної регресії, випадкових лісів, дерев рішень та градієнтного бустингу.

Виконано препроцесинг даних, створені додаткові змінні, так як «Вік автомобіля» на основі змінної «Рік випуску», а також змінна грошової одиниці у змінній «Ціна» для кращого розуміння аналізу. Також відбулось видалення не потрібних у нашому аналізі змінних, таких як «Регіон», змінено формат даних деяких змінних для коректної роботи з ними, здійснене розбиття масиву даних на навчальну та тестову виборки для подальшого аналізу.

Побудовано три моделі аналізу та прогнозування ціни автомобілів на вторинному авторинку:

- за алгоритмом множинної лінійної регресії;
- за алгоритмом дерев рішень;
- за алгоритмом випадкових лісів;

Оцінка побудованих моделей оцінювалась за допомогою показників R квадрат та RMSE. Також для кожної з моделей було графічно зображено чинники, які найбільше впливають на ціну в рамках однієї моделі.

Із проведеного аналізу та відповідно результатам, які ми отримали у дослідженні, можна стверджувати, що моделі побудовані методами дерева рішень та випадкових лісів мають високу оцінку якості, відповідно низьку похибку моделювання – 4,6 та 2,9 відповідно, що говорить про їхню статистичну значущість. За результатами моделювання з'ясовано, що на рівень ціни найбільше впливають такі показники як потужність та об'єм двигуна та марка автомобіля.

У результаті застосування алгоритму градієнтного бустингу було виявлено, що на ціну автомобіля у більшій мірі впливають такі показники як об'єм та потужність двигуна, а також вік машини. Перевірка моделі на адекватність підтвердила економічну значущість результатів та показала, що похибка моделі наближається до нуля.

Отже, побудовані моделі можуть ефективно використовуватись для прогнозування ціни на автомобілі за рахунок збільшення точності оцінювання їхньої вартості, а методи математичного моделювання є затребуваним інструментом збільшення прибутковості бізнесу в поточних умовах цифрової трансформації економіки.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Saberi. The role of the automobile industry in the economy of developed countries. 2020. vol 1 (3), 179–180.
2. Automotive industry. Internal Market, Industry, Entrepreneurship and SMEs. URL: https://single-market-economy.ec.europa.eu/sectors/automotive-industry_en (date of access : 14.05.2023).
3. Quang T. L., Khoa D. D. Trading. Behavior of institutional investors and CEO's market timing // *Asia Pacific Management Review*. 2023. vol. 1(28), 69–80.
4. Best-selling car models worldwide 2022. Statista. URL: <https://www.statista.com/statistics/239229/most-sold-car-models-worldwide/> (date of access : 14.05.2023).
5. Sameerchand Pudaruth. Predicting the Price of Used Cars using Machine Learning Techniques. *International Journal of Information & Computation Technology*. 2021. vol. 1 (4), 753–764.
6. Used car market size & share report, 2022–2030. Market Research Reports & Consulting. Grand View Research, Inc. URL: <https://www.grandviewresearch.com/industry-analysis/used-car-market> (date of access : 14.05.2023)
7. Used car market share, analysis & industry trends by 2030. Market Research Business Consulting and Strategy Planning Firm. Data Bridge Market Research Private Ltd. URL: <https://www.databridgemarketresearch.com/reports/global-used-car-market> (date of access : 14.05.2023).
8. Gegic, E., Isakovic, B., Keco, D., Masetic, Z. & Kevric, J. Car Price Prediction using Machine Learning Techniques. *Saraevo: TEM Journal*. 2020. vol. 1 (8), 113–118.
9. The outlook for the european used car market. Bain. URL: <https://www.bain.com/insights/the-outlook-for-the-european-used-car-market-brief> (date of access : 14.05.2023).

10. Kanwal, N., Sadaqat, J. Vehicle Price Prediction System using Machine Learning Techniques. *New York: International Journal of Computer Applications*. 2021. vol. 167, 27–31.
11. Fathalla, A., Salah, A., Li, K., Li, K., Francesco, P. Deep end-to-end learning for price prediction of second-hand items. *Knowl. Inf. Syst.* 2020. vol. 1(62), 4541–4568.
12. What affects the value of my car. Goodbye car. URL: <https://www.goodbyecar.uk/guides/what-affects-the-value-of-my-car> (date of access : 14.05.2023).
13. Factors that can affect used car trade-in value | capital one auto navigator. Capital One Auto Navigator. URL: <https://www.capitalone.com/cars/learn/managing-your-money-wisely/factors-that-can-affect-used-car-tradein-value/1224> (date of access : 14.05.2023).
14. Yi Lu Murphey, Liping Huang, Hao Xing Wang, Yinghao Huang. Vehicle Fault Diagnostics Using Text Mining, Vehicle Engineering Structure and Machine Learning. *International Journal of Intelligent Information Systems*. 2015. vol.4 (3), 58–70.
15. What is ETL?. Informatica. URL: <https://www.informatica.com/resources/articles/what-is-etl.html> (date of access : 14.05.2023).
16. R in business intelligence. r-bloggers. R-bloggers. URL: <https://www.r-bloggers.com/2015/01/r-in-business-intelligence/> (date of access : 14.05.2023).
17. Lawton G. Data preprocessing: definition, key steps and concepts. Data Management. URL: <https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing#:~:text=What%20is%20data%20preprocessing?,for%20the%20data%20mining%20process> (date of access : 14.05.2023).
18. Barton M., Lennox B. Model stacking to improve prediction and variable importance robustness for soft sensor development. *Digital Chemical Engineering*. URL: <https://doi.org> (date of access: 15.05.2023).

19. Decision trees. scikit-learn. URL: <https://scikit-learn.org/stable/modules/tree.html> (date of access : 14.05.2023).
20. Analytics Steps. Classification and Regression Tree (CART) Algorithm. URL: <https://www.analyticssteps.com/blogs/classification-and-regression-tree-cart-algorithm> (date of access : 14.05.2023).
21. Gupta P. Decision trees in machine learning. Medium. URL: <https://medium.com/towards-data-science/decision-trees-in-machine-learning-641b9c4e8052> (date of access : 14.05.2023).
22. IBM. Random forest. URL: <https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.>
23. Simplilearn. Random forest algorithm. URL: <https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm#:~:text=Step%201%3A%20Select%20random%20samples,as%20the%20final%20prediction%20result> (date of access : 14.05.2023).
24. Keboola. Linear Regression: Machine Learning. URL: <https://www.keboola.com/blog/linear-regression-machine-learning> (date of access : 14.05.2023).
25. Displayr. Gradient Boosting: The Coolest Kid on the Machine Learning Block. URL: <https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/> (date of access : 14.05.2023).
26. NVIDIA. XGBoost. URL: <https://www.nvidia.com/en-us/glossary/data-science/xgboost/#:~:text=The%20GPU%2Daccelerated%20XGBoost%20algorithm,dataset%20concurrently%20on%20the%20GPU> (date of access : 14.05.2023).
27. Machine Learning Mastery. A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning. URL: <https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/> (date of access : 14.05.2023).
28. XGBoost Documentation. URL: <https://xgboost.readthedocs.io/en/stable/> (date of access : 14.05.2023).

29. Kandilarov, I., Dobрева, E. "Random Forest" – въведение в един алгоритъм за класификация и регресия. Proceedings of the International Conference on Computer Systems and Technologies – CompSysTech'16, Ruse, Bulgaria. 2016. Pp. 59–64. URL: <https://core.ac.uk/download/pdf/147037361.pdf> (date of access : 14.05.2023).
30. Used Cars Price Prediction. Kaggle, CoLearningLounge, 2021. URL: <https://www.kaggle.com/datasets/colearninglounge/used-cars-price-prediction> (date of access : 14.05.2023).
31. Manish. Factors in R Programming. Finance Train. URL: <https://financetrain.com/factors-in-r-programming> (date of access: 04.12.2022)
32. Hyperparameter Tuning in Python: a Complete Guide. neptune.ai. URL: <https://neptune.ai/blog/hyperparameter-tuning-in-python-compleateguide> (date of access: 04.12.2022).

ДОДАТКИ

Додаток А. Початковий масив даних

Age_of_car	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	Price	brand
13	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5	2.100	Maruti
8	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5	15.000	Hyundai
12	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5	5.400	Honda
11	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7	7.200	Maruti
10	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5	21.288	Audi
10	Nissan Micra Diesel XV	Jaipur	2013	86999	Diesel	Manual	First	23.08 kmpl	1461 CC	63.1 bhp	5	4.200	Nissan
7	Toyota Innova Crysta 2.8 GX AT BS	Mumbai	2016	36000	Diesel	Automatic	First	11.36 kmpl	2755 CC	171.5 bhp	8	21.000	Toyota
10	Volkswagen Vento Diesel Comfortline	Pune	2013	64430	Diesel	Manual	First	20.54 kmpl	1598 CC	103.6 bhp	5	6.240	Volkswagen
11	Tata Indica Vista Quadrajet LS	Chennai	2012	65932	Diesel	Manual	Second	22.3 kmpl	1248 CC	74 bhp	5	2.340	Tata
5	Maruti Ciaz Zeta	Kochi	2018	25692	Petrol	Manual	First	21.56 kmpl	1462 CC	103.25 bhp	5	11.940	Maruti
8	Maruti Swift VDI BSIV	Jaipur	2015	64424	Diesel	Manual	First	25.2 kmpl	1248 CC	74 bhp	5	6.720	Maruti
9	Land Rover Range Rover 2.2L Pure	Delhi	2014	72000	Diesel	Automatic	First	12.7 kmpl	2179 CC	187.7 bhp	5	32.400	Land
11	Land Rover Freelander 2 TD4 SE	Pune	2012	85000	Diesel	Automatic	Second	0.0 kmpl	2179 CC	115 bhp	5	21.000	Land
9	Mitsubishi Pajero Sport 4X4	Delhi	2014	110000	Diesel	Manual	First	13.5 kmpl	2477 CC	175.56 bhp	7	18.000	Mitsubishi
7	Honda Amaze S i-Dtech	Kochi	2016	58950	Diesel	Manual	First	25.8 kmpl	1498 CC	98.6 bhp	5	6.480	Honda
6	Maruti Swift DDIS VDI	Jaipur	2017	25000	Diesel	Manual	First	28.4 kmpl	1248 CC	74 bhp	5	7.188	Maruti
9	Mercedes-Benz New C-Class C 220 CDI BE Avantgare	Bangalore	2014	78500	Diesel	Automatic	First	14.84 kmpl	2143 CC	167.62 bhp	5	33.600	Mercedes-Benz
9	BMW 3 Series 320d	Kochi	2014	32982	Diesel	Automatic	First	22.69 kmpl	1995 CC	190 bhp	5	22.260	BMW
13	Hyundai i20 1.2 Magna	Kolkata	2010	45807	Petrol	Manual	First	18.5 kmpl	1197 CC	80 bhp	5	2.244	Hyundai
13	Volkswagen Vento Petrol Highline AT	Kolkata	2010	33000	Petrol	Automatic	First	14.4 kmpl	1598 CC	103.6 bhp	5	3.420	Volkswagen
11	Nissan Micra Diesel XV	Hyderabad	2012	54000	Diesel	Manual	First	23.08 kmpl	1461 CC	63.1 bhp	5	5.100	Nissan
10	Maruti Alto K10 2010-2014 VXI	Hyderabad	2013	54000	Petrol	Manual	Second	20.92 kmpl	998 CC	67.1 bhp	5	3.300	Maruti
5	Honda WRV i-VTEC VX	Kochi	2018	37430	Petrol	Manual	First	17.5 kmpl	1199 CC	88.7 bhp	5	11.880	Honda
16	Toyota Innova 2.5 V Diesel 7-seater	Mumbai	2007	262000	Diesel	Manual	Fourth & Above	12.8 kmpl	2494 CC	102 bhp	7	4.800	Toyota
10	Renault Duster 110PS Diesel RxZ Pack	Bangalore	2013	62000	Diesel	Manual	First	19.01 kmpl	1461 CC	108.45 bhp	5	9.300	Renault
11	Toyota Corolla Altis G	Delhi	2012	87615	Petrol	Manual	First	14.53 kmpl	1798 CC	138.1 bhp	5	8.100	Toyota
10	Mahindra Saangyong Xeston RX7	Bangalore	2013	72000	Diesel	Automatic	Second	11.18 kmpl	2696 CC	184 bhp	7	15.000	Mahindra
13	Audi A6 2.7 TDI	Kolkata	2010	35000	Diesel	Automatic	First	12.4 kmpl	2698 CC	179.5 bhp	5	13.800	Audi
9	Volkswagen Vento 1.6 Highline	Coimbatore	2014	55431	Petrol	Manual	Second	16.09 kmpl	1598 CC	103.5 bhp	5	8.376	Volkswagen
14	Maruti Wagon R LXI BSIII	Jaipur	2009	83816	Petrol	Manual	Second	14.0 kmpl	1061 CC	64 bhp	5	2.220	Maruti
5	Maruti Vitara Brezza ZDI Plus	Kochi	2018	50075	Diesel	Manual	First	24.3 kmpl	1248 CC	88.5 bhp	5	10.356	Maruti
10	Mercedes-Benz M-Class ML 350 CDI	Pune	2013	85000	Diesel	Automatic	First	11.74 kmpl	2987 CC	254.8 bhp	5	33.600	Mercedes-Benz

Додаток Б. Модифікований масив даних

Age_of_car	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	Price	brand	power_num	engine_num
2	8 Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5	15,000	Hyundai	126.20	1582
3	12 Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5	5,400	Honda	88.70	1199
4	11 Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7	7,200	Maruti	88.76	1248
5	10 Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5	21,288	Audi	140.80	1968
6	10 Nissan Micra Diesel XV	Jaipur	2013	96999	Diesel	Manual	First	23.08 kmpl	1461 CC	63.1 bhp	5	4,200	Nissan	63.10	1461
7	7 Toyota Innova Crysta 2.8 GX AT BS	Mumbai	2016	36000	Diesel	Automatic	First	11.36 kmpl	2755 CC	171.5 bhp	8	21,000	Toyota	171.50	2755
8	10 Volkswagen Vento Diesel Comfortline	Pune	2013	64430	Diesel	Manual	First	20.54 kmpl	1598 CC	103.6 bhp	5	6,240	Volkswagen	103.60	1598
9	11 Tata Indica Vista Quadrajet LS	Chennai	2012	65932	Diesel	Manual	Second	22.3 kmpl	1248 CC	74 bhp	5	2,340	Tata	74.00	1248
10	5 Maruti Ciaz Zeta	Kochi	2018	25692	Petrol	Manual	First	21.56 kmpl	1462 CC	103.25 bhp	5	11,940	Maruti	103.25	1462
11	8 Maruti Swift VDI BSIV	Jaipur	2015	64424	Diesel	Manual	First	25.2 kmpl	1248 CC	74 bhp	5	6,720	Maruti	74.00	1248
12	9 Land Rover Range Rover 3.2L Pure	Delhi	2014	72000	Diesel	Automatic	First	12.7 kmpl	2179 CC	187.7 bhp	5	32,400	Land	187.70	2179
13	11 Land Rover Freelander 2 TD4 SE	Pune	2012	85000	Diesel	Automatic	Second	0.0 kmpl	2179 CC	115 bhp	5	21,000	Land	115.00	2179
14	9 Mitsubishi Pajero Sport 4X4	Delhi	2014	110000	Diesel	Manual	First	13.5 kmpl	2477 CC	175.56 bhp	7	18,000	Mitsubishi	175.56	2477
15	7 Honda Amaze S i-Dtech	Kochi	2016	58950	Diesel	Manual	First	25.8 kmpl	1498 CC	98.6 bhp	5	6,480	Honda	98.60	1498
16	6 Maruti Swift DDiS VDI	Jaipur	2017	25000	Diesel	Manual	First	26.4 kmpl	1248 CC	74 bhp	5	7,188	Maruti	74.00	1248
17	9 Mercedes-Benz New C-Class C 220 CDI BE Avantgare	Bangalore	2014	78500	Diesel	Automatic	First	14.84 kmpl	2143 CC	167.62 bhp	5	33,600	Mercedes-Benz	167.62	2143
18	9 BMW 3 Series 320d	Kochi	2014	32982	Diesel	Automatic	First	22.69 kmpl	1995 CC	190 bhp	5	22,260	BMW	190.00	1995
19	13 Hyundai i20 1.2 Magna	Kolkata	2010	45907	Petrol	Manual	First	18.5 kmpl	1197 CC	80 bhp	5	2,244	Hyundai	80.00	1197
20	13 Volkswagen Vento Petrol Highline AT	Kolkata	2010	33000	Petrol	Automatic	First	14.4 kmpl	1598 CC	103.6 bhp	5	3,420	Volkswagen	103.60	1598
21	11 Nissan Micra Diesel XV	Hyderabad	2012	54000	Diesel	Manual	First	23.08 kmpl	1461 CC	63.1 bhp	5	5,100	Nissan	63.10	1461
22	10 Maruti Alto K10 2010-2014 VXI	Hyderabad	2013	54000	Petrol	Manual	Second	20.92 kmpl	998 CC	67.1 bhp	5	3,300	Maruti	67.10	998
23	5 Honda WRV i-VTEC VX	Kochi	2018	37430	Petrol	Manual	First	17.5 kmpl	1199 CC	88.7 bhp	5	11,880	Honda	88.70	1199
24	16 Toyota Innova 2.5 V Diesel 7-seater	Mumbai	2007	262000	Diesel	Manual	Fourth & Above	12.8 kmpl	2494 CC	102 bhp	7	4,800	Toyota	102.00	2494
25	10 Renault Duster 110PS Diesel RxZ Pack	Bangalore	2013	62000	Diesel	Manual	First	19.01 kmpl	1461 CC	108.45 bhp	5	9,300	Renault	108.45	1461
26	11 Toyota Corolla Altis G	Delhi	2012	87615	Petrol	Manual	First	14.53 kmpl	1798 CC	138.1 bhp	5	8,100	Toyota	138.10	1798
27	10 Mahindra Ssangyong Reixton RX7	Bangalore	2013	72000	Diesel	Automatic	Second	11.18 kmpl	2696 CC	184 bhp	7	15,000	Mahindra	184.00	2696
28	13 Audi A6 2.7 TDI	Kolkata	2010	35000	Diesel	Automatic	First	12.4 kmpl	2698 CC	179.5 bhp	5	13,800	Audi	179.50	2698
29	9 Volkswagen Vento 1.6 Highline	Coimbatore	2014	55431	Petrol	Manual	Second	16.09 kmpl	1598 CC	103.5 bhp	5	8,376	Volkswagen	103.50	1598
30	14 Maruti Wagon R LXI BSIII	Jaipur	2009	83816	Petrol	Manual	Second	14.0 kmpl	1061 CC	64 bhp	5	2,220	Maruti	64.00	1061
31	5 Maruti Vitara Brezza ZDi Plus	Kochi	2018	50075	Diesel	Manual	First	24.3 kmpl	1248 CC	88.5 bhp	5	10,356	Maruti	88.50	1248
32	10 Mercedes-Benz M-Class ML 350 CDI	Pune	2013	85000	Diesel	Automatic	First	11.74 kmpl	2987 CC	254.8 bhp	5	33,600	Mercedes-Benz	254.80	2987
33	10 Volkswagen Polo Diesel Trendline 1.2L	Chennai	2013	62000	Diesel	Manual	First	22.07 kmpl	1199 CC	73.9 bhp	5	6,000	Volkswagen	73.90	1199
34	15 Maruti Alto LXI	Delhi	2008	90000	Petrol	Manual	Second	19.7 kmpl	796 CC	46.3 bhp	5	1,500	Maruti	46.30	796
35	9 Hyundai i20 Magna Optional 1.2	Kochi	2014	35487	Petrol	Manual	First	18.5 kmpl	1197 CC	82.9 bhp	5	5,688	Hyundai	82.90	1197
36	9 Hyundai Elantra 2.0 SX Option AT	Bangalore	2014	46374	Petrol	Automatic	First	14.62 kmpl	1999 CC	149.92 bhp	5	10,620	Hyundai	149.92	1999
37	9 Hyundai Xcent 1.1 CRDi S	Jaipur	2014	75000	Diesel	Manual	First	24.4 kmpl	1120 CC	71 bhp	5	4,620	Hyundai	71.00	1120
38	8 Audi A4 2.0 TDI 177 Bhp Premium Plus	Kochi	2015	13648	Diesel	Automatic	First	17.11 kmpl	1968 CC	174.33 bhp	5	25,716	Audi	174.33	1968
39	9 Maruti Swift LDI	Delhi	2014	52000	Diesel	Manual	First	22.9 kmpl	1248 CC	74 bhp	5	5,100	Maruti	74.00	1248
40	9 Maruti Swift Lxi BSIV	Kolkata	2014	53000	Diesel	Manual	First	17.8 kmpl	1248 CC	75 bhp	5	3,960	Maruti	75.00	1248

Showing 1 to 40 of 4375 entries. 16 total columns

```

```{r}
figure1 <- ggarrange(annotate_figure(data %>% ggplot(aes(x=Age_of_car)) +
 geom_histogram(aes(y=..density..), colour="lightskyblue", fill="skyblue", alpha=0.5)+
 theme_classic()+
 labs(x="Вік", y="Розподіл", title = NULL), top = text_grob("За віком",
 color ="black", face = "bold", size = 12)) + rremove("xlab"),annotate_figure(data %>%
 ggplot(aes(x=Price))+
 geom_histogram(aes(y=..density..), colour="peachpuff", fill="peachpuff", alpha=0.5)+
 theme_classic()+
 labs(x="Ціна", y="Розподіл", title = NULL), top = text_grob("За ціною",
 color ="black", face = "bold", size = 12)) + rremove("ylab")+ rremove("xlab"),annotate_figure(data %>%
 ggplot(aes(x=power_num)) +
 geom_histogram(aes(y=..density..), colour="darkseagreen2", fill="darkseagreen2", alpha=0.5)+
 theme_classic()+
 labs(x="Потужність двигуна", y="Розподіл", title = NULL), top = text_grob("За потужністю двигуна",
 color ="black", face = "bold", size = 12)) + rremove("ylab")+ rremove("xlab"), annotate_figure(data %>%
 ggplot(aes(x=engine_num)) +
 geom_histogram(aes(y=..density..), colour="coral1", fill="coral1", alpha=0.5)+
 theme_classic()+
 labs(x="Об'єм двигуна", y="Розподіл", title = NULL), top = text_grob("За об'ємом двигуна",
 color ="black", face = "bold", size = 12))+ rremove("ylab")+ rremove("xlab"))

figure1
```

```

Додаток Д. Гістограма важливості факторів

```

```{r}
ggplot(var_imp1, aes(x = reorder(rownames(var_imp1), -Overall), y = Overall, fill = Overall)) +
 geom_bar(stat = "identity") +
 theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
 xlab("Factors") +
 ylab("Importance") +
 ggtitle("Variable Importance")
```

```