**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**імені ТАРАСА ШЕВЧЕНКА**

****

**ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**

**Кафедра прикладних інформаційних систем**

**Звіт до лабораторної роботи 3**

**з курсу**

**«Data Science»**

*студентки 4 курсу*

*групи ПП-41*

*спеціальності 122 «Комп’ютерні науки»*

*ОП «Прикладне програмування»*

Ветряченко Т.С.

*Викладач:*

Білий Р.О.

**Київ – 2023**

**Методи аналізу та вибору значущих ознак в моделі автомобільного ринку**

Метою цієї лабораторної роботи є набуття практичних навичок аналізу та вибору значущих ознак для моделі автомобільного ринку. Для досягнення цієї мети ми використовували кореляційний аналіз, таблиці сопряження, аналіз багатомірних залежностей, дихотомії, дисперсійний аналіз ANOVA, критерій Хі-квадрат та інші методи.

**Основні етапи**

* Очищення та обробка даних

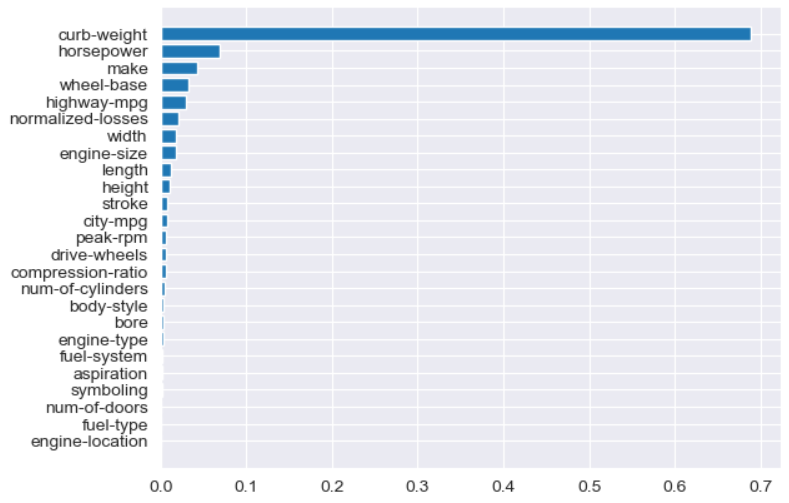
Початковим етапом було проведено очищення та підготовку даних. Виявлені відсутні значення, позначені символом "?", у колонці "normalized-losses", були замінені на значення None для подальшого аналізу. Далі була розроблена функція find\_mean, яка обчислює середні значення для конкретних категорій, таких як "symboling" і "curb-weight". Це дозволило зберегти більше даних для подальшого аналізу, уникнувши їх видалення.

* Експлоративний аналіз даних (EDA)

Під час EDA була зосереджена увага на вивченні різних ознак та їх потенційного впливу на ціну автомобілів. Були проаналізовані параметри, такі як вага автомобіля, тип палива, кількість дверей, стиль кузова, тип двигуна та інші. Виявлено, що деякі ознаки мають значущий вплив на ціну, такі як тип палива та вага автомобіля.

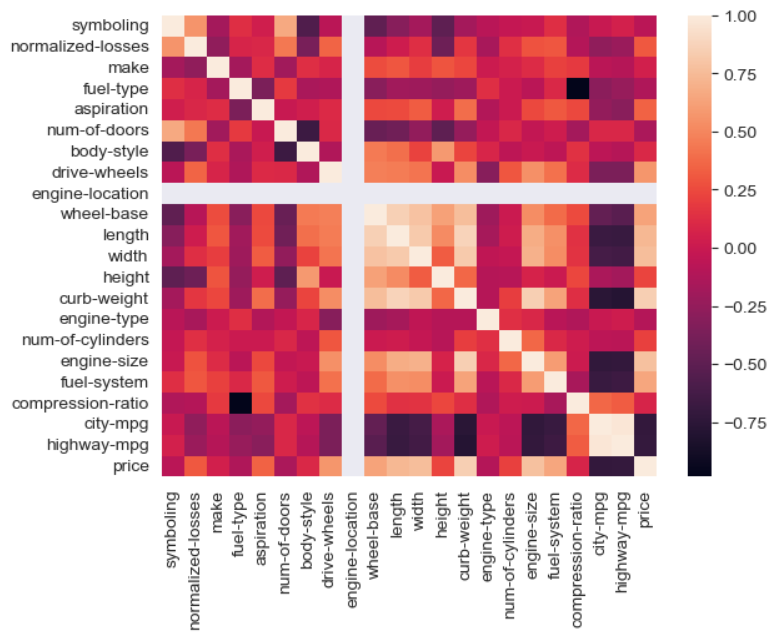
* Однофакторний аналіз ознак

Для визначення найважливіших ознак, впливаючих на ціну автомобіля, використовувалися методи VarianceThreshold і SelectKBest. VarianceThreshold дозволив виявити ознаки з низькою варіативністю, менш значущі для моделювання. SelectKBest використовувався для вибору топ-ознак, які найбільше корелюють з ціною автомобіля. Отримані результати підтвердили важливість таких параметрів, як тип двигуна, стиль кузова та тип паливної системи.



* Аналіз взаємозалежності ознак

Проведено аналіз кореляцій між різними ознаками за допомогою теплових карт (heatmaps) у Seaborn. Це дозволило виявити сильні взаємозв'язки між ознаками, наприклад, висока позитивна кореляція між вагою автомобіля та його потужністю.



Використання RandomForestRegressor для визначення важливості ознак

Модель RandomForestRegressor використовувалась для оцінки важливості різних ознак. Цей метод надав кількісну оцінку впливу кожної ознаки на ціну автомобіля. Результати підтвердили, що ознаки, визначені як значущі в однофакторному аналізі, також мали високу важливість у цьому аналізі.

Висновки: аналіз дав глибоке розуміння ключових факторів, що впливають на ціну автомобіля. Встановлено, що тип двигуна, стиль кузова, вага автомобіля та тип паливної системи мають вирішальне значення. Результати можуть бути корисні для виробників автомобілів, дилерів і споживачів для розуміння формування вартості автомобілів та розробки точних моделей прогнозування цін на авто, що може вплинути на ринок автомобілів.