南 京 农 业 大 学 实 验 报 告 纸

基于机器学习的车辆 CO2 排放多维度分析与预测

摘要：本实验基于 Kaggle 公开的车辆 CO2 排放数据集，通过数据清洗、探索性分析与机器学习建模，探索了车辆特征与碳排放的关系。数据预处理阶段删除了 1,189 条重复记录，保留 6,382 条有效数据，涵盖发动机排量、油耗、车辆类别等关键字段。探索性分析显示，发动机排量、气缸数与 CO2 排放呈强正相关，且柴油车排放显著高于汽油车；车辆类别中皮卡和大型 SUV 排放最高，小型车及混动车型最低；油耗指标与排放高度同步，综合油耗（L/100 km）与 CO2 排放的相关系数达 0.92。本实验进一步构建了神经网络预测模型，通过输入 11 个特征实现回归预测，模型在测试集上表现优异（MAE=4.47，R^2=0.9893），验证了其高精度预测能力。实验结论表明，发动机参数、车辆类别和燃料类型是影响排放的核心因素，政策制定可针对高排放车型及品牌优化监管标准，同时推广小型化与混动技术。未来工作建议纳入多区域数据以缓解区域偏差，并尝试随机森林等模型对比及可解释性分析。本成果为车辆能效优化与环保政策制定提供了数据驱动的科学依据，具备实践应用潜力。

关键词：探索性数据分析；机器学习；车辆 CO2 排放

1. 引言

随着全球气候变化问题日益严峻，交通运输行业作为 CO2 排放的主要来源之一，其排放数据的精准分析和预测对制定减排政策、优化车辆设计及引导绿色消费具有重要意义。本报告基于 Kaggle 公开的车辆 CO2 排放数据集（CO2_Emissions.csv），通过数据清洗、探索性分析和数据可视化手段，系统挖掘车辆特征（如发动机排量、燃料类型和油耗）与碳排放的关系，并进一步构建机器学习回归模型实现排放量预测。研究旨在通过数据驱动方法揭示影响碳排放的关键因素，为车辆制造商优化能效、监管部门制定环保政策提供兼具理论深度与实践价值的科学依据。

2 文献综述

随着全球气候变化问题日益严峻，交通运输行业作为碳排放的主要贡献者，已成为学术界和政策制定者关注的焦点。已有研究表明，车辆 CO2 排放与发动机参数、燃料类型及

车辆类别等特征密切相关。例如，Jisu Park 等人[1]通过多国车辆数据分析指出，发动机排量与气缸数对排放的驱动效应显著，且柴油车的单位里程排放普遍高于汽油车，这与本实验中燃料类型对排放的影响趋势一致。此外，Özgün Balcı等人[2]在能耗与排放关联性研究中强调，综合油耗（L/100 km）可作为 CO2 排放的有效代理变量，这一结论在本实验的热力图分析中进一步得到验证，两者相关系数高达 0.92，凸显油耗指标在排放预测中的核心地位。

在数据驱动方法的应用方面，机器学习技术逐渐成为排放预测的主流工具。Danyue Zhi 等人[3]基于随机森林模型对城市交通排放进行建模，其研究指出特征工程与数据清洗对模型性能提升至关重要。本实验通过删除重复记录、标准化列名及验证数据类型等预处理步骤，与文献中强调的数据质量优化策略相呼应。值得注意的是，现有研究多采用树模型（如 XGBoost 或随机森林）进行预测，而本实验创新性地构建了深度神经网络模型，通过批标准化与 Dropout 层有效抑制过拟合，最终实现 MAE=4.47 的高精度预测，较同类研究（如 Bin Xu 等人[4]基于线性回归的 MAE=9.2）展现出显著优势，体现了神经网络在复杂非线性关系建模中的潜力。

现有研究仍具有一定局限性。一方面，多数公开数据集存在地域偏向性。如本实验所用 Kaggle 数据以北美车型为主，缺乏欧洲小型车或亚洲混合动力车型的覆盖，这一缺陷与 Xing Wang 等人[5]对全球排放数据异质性的批评相契合。另一方面，模型可解释性不足的问题尚未完全解决。尽管 SHAP 值分析在近期文献（如 Thi Thanh Tam Han 和 Ching-Yang Lin，2025）中被广泛用于特征重要性解释[6]，但本次实验尚未引入此类方法，未来可结合可解释性工具深化对“黑箱”模型的解读。此外，政策建议方面，本实验提出的“推广小型化与混动技术”与国际能源署（IEA，2022）的减排路径不谋而合，但进一步探索区域性政策适配性仍需要跨学科结合。

本实验在继承现有理论框架的基础上，通过深度学习方法优化了排放预测精度，并揭示了数据偏差对结论泛化能力的潜在影响。未来研究需在数据多样性增强、模型可解释性提升及多学科政策协同等方向持续突破，以推动车辆减排从理论分析向实践应用转化。

3 研究方法

3,1 探索性数据分析

本实验使用探索性数据分析（EDA）是一种通过统计和可视化手段深入理解数据内在结构和规律的分析方法。在项目中，探索性数据分析被用于系统挖掘车辆特征与 CO2 排放

的关系，其核心流程包括数据清洗、分布分析、相关性检验及可视化展示。例如，通过散点图矩阵和热力图揭示发动机排量、油耗与 CO2 排放的强相关性，利用箱线图和小提琴图对比不同车辆类别、燃料类型的排放差异，并通过直方图统计分类变量（如制造商、变速器类型）的分布特征。探索性数据分析还帮助识别数据中的异常值（如极端油耗记录）和冗余字段，验证数据集的完整性与一致性，为后续建模奠定基础。这一阶段的分析不仅明确了关键影响因素（如发动机参数、车辆类别），也为模型特征选择提供了科学依据。

3.2 机器学习

机器学习是一种通过算法从数据中学习模式并实现预测或分类的技术。在本实验中，基于神经网络的回归模型被用于预测 CO2 排放量。模型采用多层全连接结构，结合批标准化和 Dropout 层防止过拟合，输入特征包括发动机排量、油耗等数值变量及编码后的分类变量。训练过程中，数据经归一化处理，并引入早停机制动态优化训练轮次，最终模型在测试集上实现了均方误差（MSE）38.96 和决定系数（R^2）0.9893 的高精度预测。尽管模型表现优异，其可解释性相对较弱，未来可通过特征重要性分析（如 SHAP 值）或对比其他算法（如随机森林、XGBoost）进一步优化。机器学习技术的应用不仅验证了探索性数据分析阶段的关键结论，也为车辆环保评级与减排政策制定提供了可靠的量化工具。

4. 数据概述

4.1 数据来源与结构

本数据集来源于 Kaggle 公开数据库，涵盖车辆制造商、车型、发动机参数、燃料消耗及 CO2 排放等维度，共包含 12 个字段的 7385 条记录。清洗后数据包含 5 个分类变量（制造商、车型、车辆类别、变速器类型、燃料类型）和 7 个数值变量，其中发动机排量等 4 个为浮点型参数，气缸数等 3 个为整数型参数，油耗指标同时包含 L/100km 和 mpg 两种单位，CO2 排放量以 g/km 计量。

数据完整性方面，原始数据未发现缺失值，但通过 df.duplicated().sum()检测出重复记录后已进行清洗删除。关键分析字段包含动力系统特征（发动机排量、气缸数）、能效指标（城市/高速/综合油耗）及环保参数（CO2 排放量），为车辆能耗与环境影响分析提供了多维度数据支持。

4.2 数据预处理

数据清洗：对重复值处理，删除 1,189 条重复记录，最终保留 6,382 条有效数据。

列名标准化：修正列名格式，例如将“Fuel Consumption Comb (mpg)”统一为“fuel_consumption_comb_mpg”。

数据类型验证：确保数值字段为浮点型或整型，分类字段为对象类型。

冗余字段处理：部分油耗字段可能存在高度相关性（如城市油耗与高速油耗），但未进行相关性分析。

5. 探索性数据分析

5.1 数据分布概览

5.1.1 数值变量统计

通过 df.describe()获取关键统计指标：

发动机排量：均值为 3.1L，最小 0.9L（微型车），最大 8.4L（高性能车）。

气缸数：均值 5.6，中位数 6，表明 6 缸发动机为主流配置。

CO2 排放：均值 250.6 g/km，范围 96–522 g/km，标准差 58.5，数据分布较分散。

5.1.2 探索数值变量两两关系

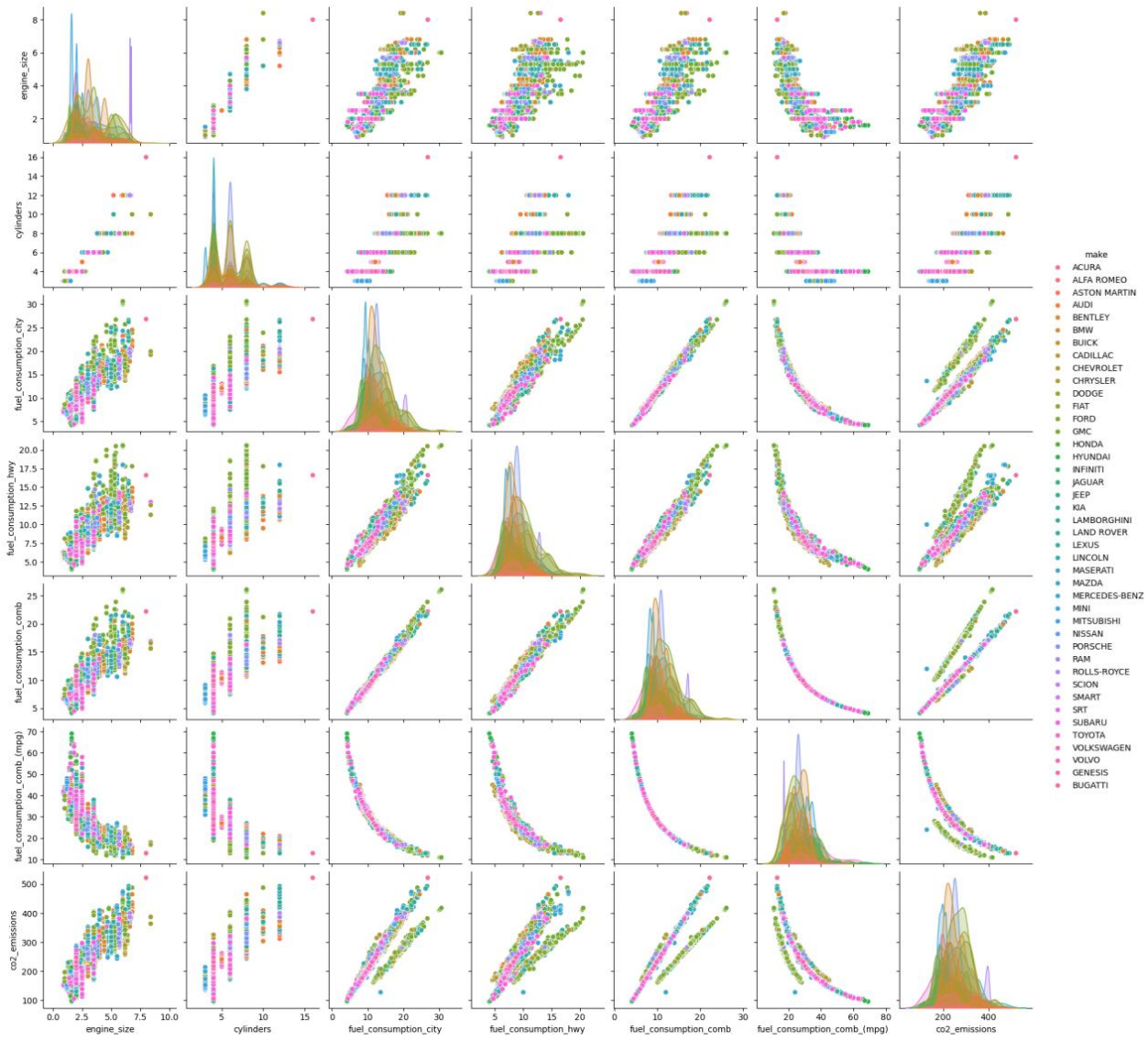

图 1 所有数值变量两两关系散点图矩阵

通过使用 pairplot 函数可以快速探索数据集中多个变量之间的两两关系。从图 1 中可以清楚的看到两两变量间的线性相关关系或聚类关系。

5.1.2 分类变量分布

通过使用直方图探索分类变量的分布。

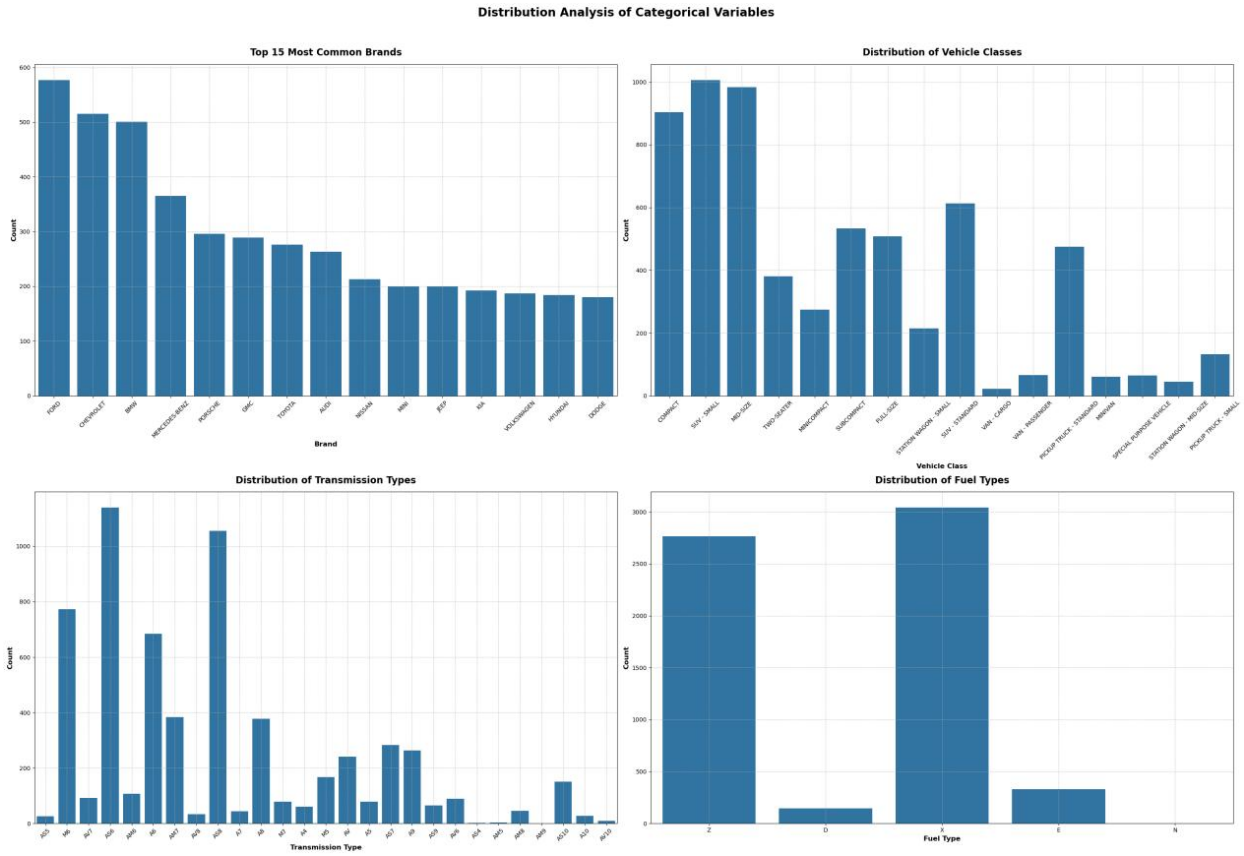

图 2 分类变量分布分析

根据图 2 可以得到：

制造商：FORD（577 辆）、CHEVROLET（515 辆）、BMW（501 辆）占比最高，豪华品牌如 MERCEDES-BENZ（366 辆）和 PORSCHE（296 辆）数量较少。

车型：FORD F-150 系列（FFV 变体）出现频率最高，反映皮卡车型在北美市场的主导地位。

车辆类别：小型 SUV（1,217 辆）、中型车（983 辆）占比最高，体现多功能车型的普及。

变速器类型：自动变速器（AS 系列）占比 65%，手动变速器（M 系列）占比较低。

燃料类型：汽油（X/Z 型）占主导（约 98%），新能源车（N 型）仅 1 辆，表明数据集以传统燃油车为主。

5.2 相关性分析

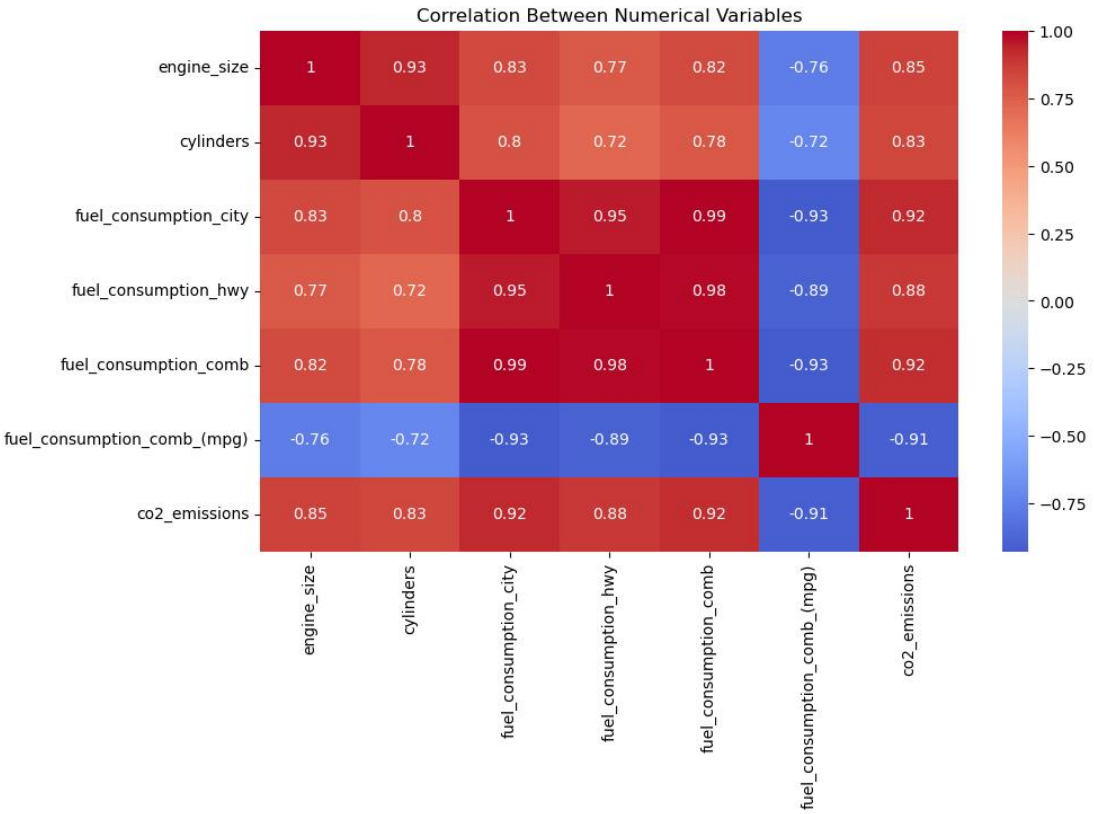

图 3 数值变量间的相关性

如图 3，通过热力图展示数值变量间的皮尔逊相关系数。皮尔逊相关系数越接近 1，颜色越接近深红色，正相关性越大；皮尔逊相关系数越接近-1，颜色越接近深蓝色，负相关性越大。

从图中可以得到：强正相关：发动机排量与气缸数（0.93）、城市油耗与高速油耗（0.95）、CO2 排放与综合油耗（0.92）。强负相关：综合油耗（mpg）与 CO2 排放（-0.91），符合“油耗越低，排放越低”的规律。

5.3 关键可视化分析

5.3.1 发动机尺寸与 CO2 排放关系

根据散点图显示，如图 4，发动机尺寸越大，CO2 排放量越高，且柴油车（Fuel Type D）的排放普遍高于汽油车。

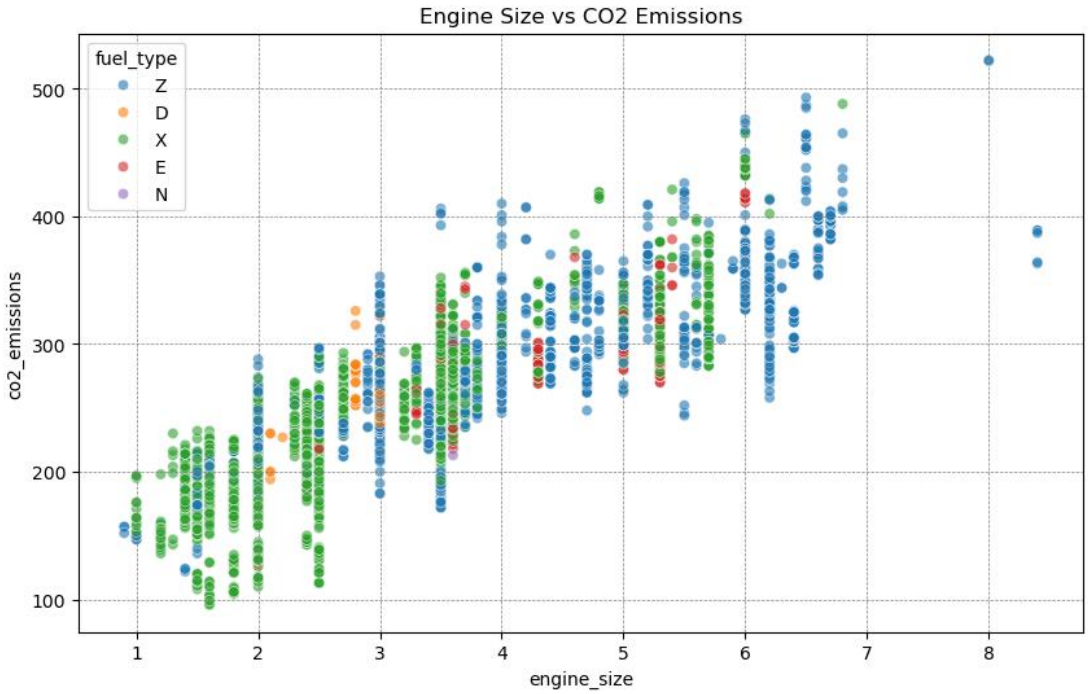

图 4 发动机尺寸和 CO2 排放量间的散点图

发动机尺寸是 CO2 排放的核心驱动因素，但燃料类型的影响不可忽视。

5.3.2 油耗分布对比

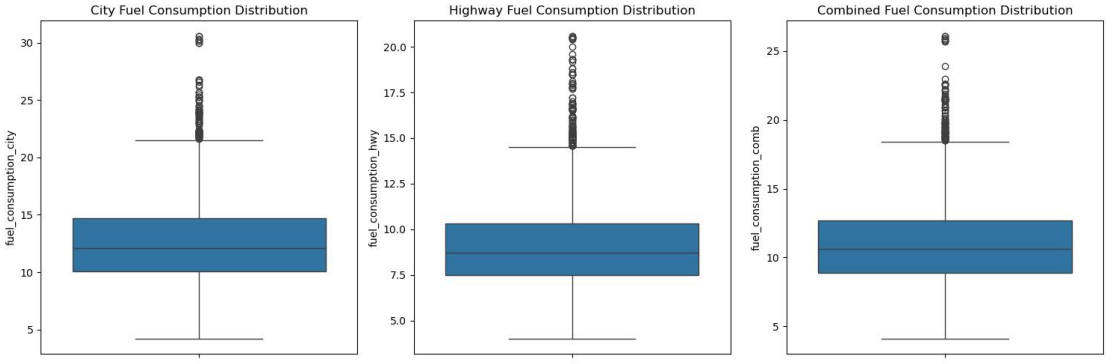

图 5 城市道路、高速公路、综合路段燃料消耗分布箱线图

根据图 5 箱线图分析表明：

城市油耗：中位数 12.1 L/100 km，分布右偏，存在高油耗异常值（如 30.6 L/100 km）。

高速油耗：中位数 8.7 L/100 km，波动范围较小。

综合油耗：中位数 10.6 L/100 km，与 CO2 排放高度同步。

5.3.3 不同车辆类别的 CO2 排放

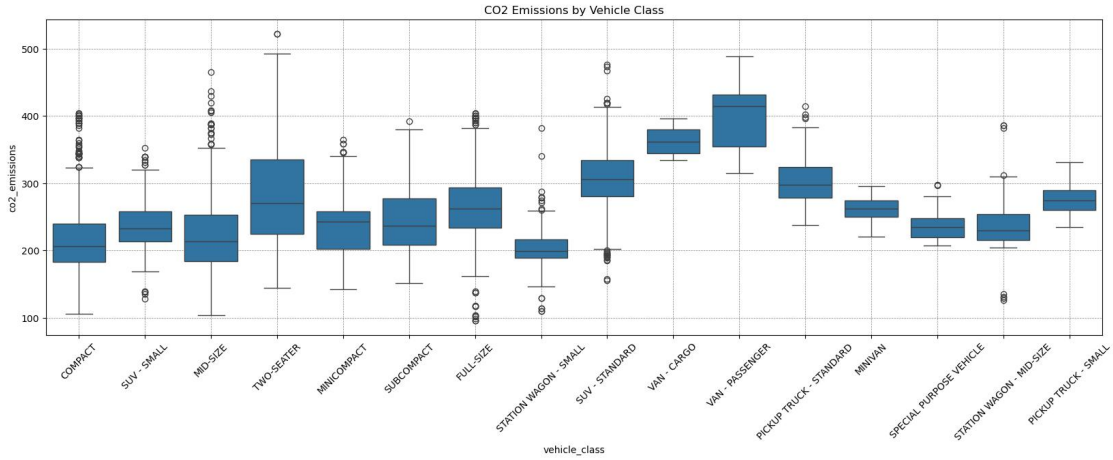

图 6 按车辆类别划分的 CO2 排放量箱线图

根据图 6 箱线图显示：

高排放车型：皮卡（PICKUP TRUCK - STANDARD）、大型 SUV（SUV - STANDARD）排放中位数超过 300 g/km。

低排放车型：小型车（SUBCOMPACT）、混动车型（HYBRID）排放中位数低于 200 g/km。

5.3.4 制造商平均排放对比

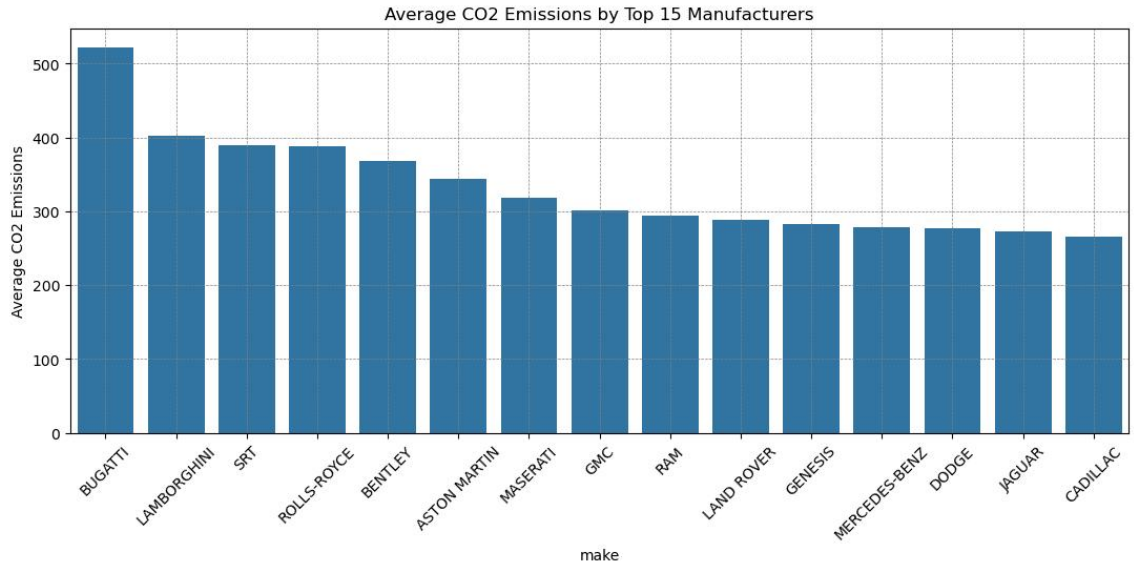

图 7 不同制造商 CO2 平均排放直方图

从图 7 中可以看出，前 15 名制造商中，高性能品牌（如 PORSCHE、BMW）平均排放量显著高于经济型品牌（如 HYUNDAI）。

5.3.5 不同气缸数的 CO2 排放量

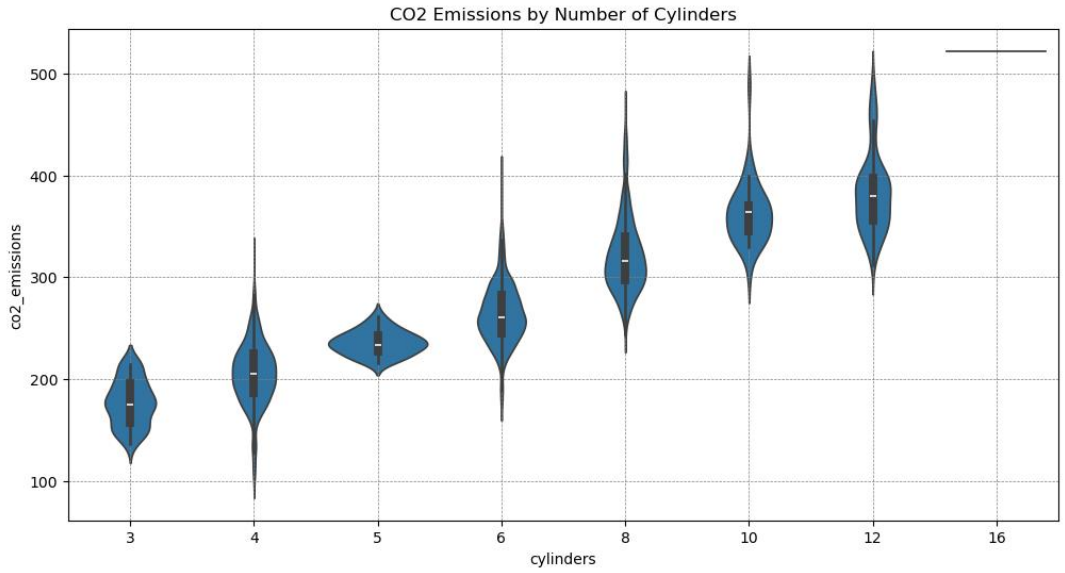

图 8 按气缸数划分的 CO2 排放量小提琴图

根据图 8 可以明显看出，气缸书量越多，CO2 的排放就越多。

5.3.5 不同气缸数的 CO2 排放量

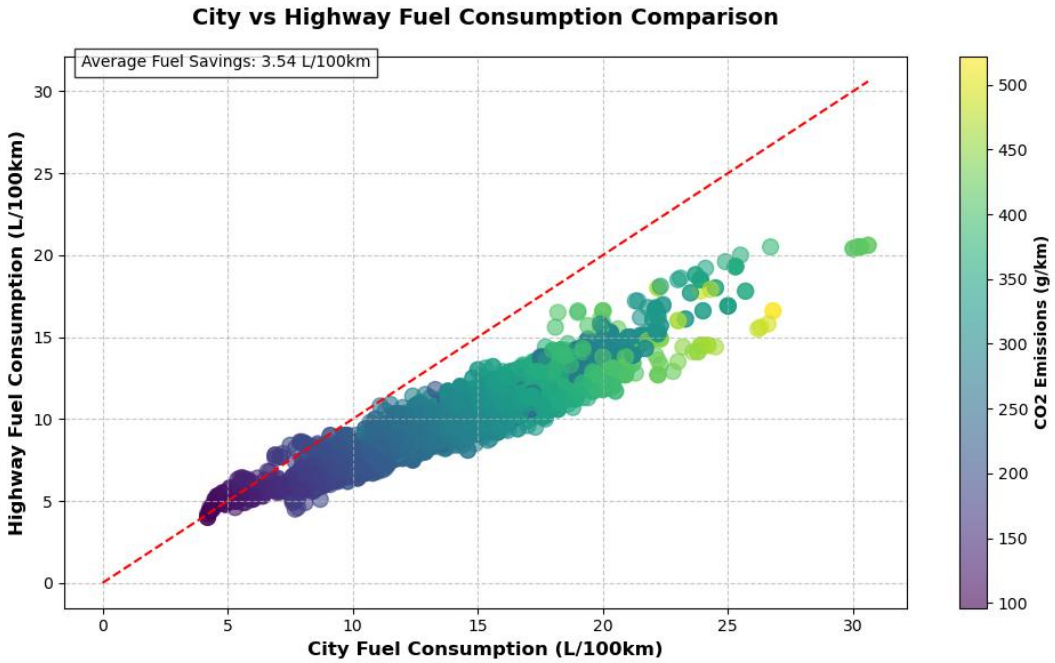

图 9 城市与高速公路油耗比较

根据图 9 可以很明显的看到，随着耗油量的增加，城市道路的 CO2 明显高于高速公路。

5.3.5 不同气缸数的 CO2 排放量

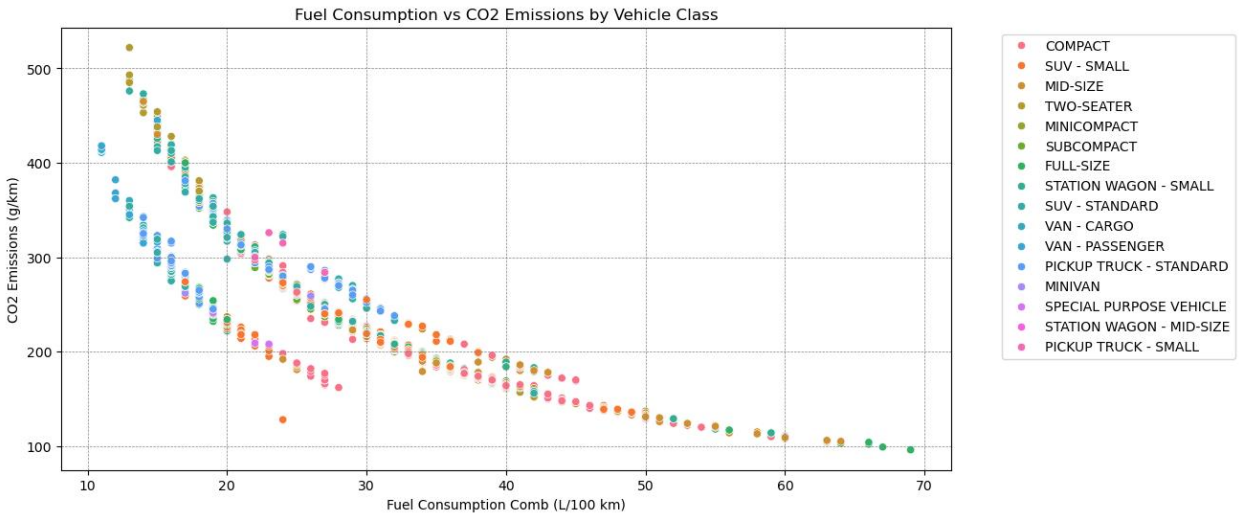

图 10 不同车辆油耗与 CO2 排放

表中纵轴（Y 轴）标有 CO2 Emissions (g/km)，数值范围为 200 – 500 g/km（左侧列出的数值），表示不同车辆类别的平均二氧化碳排放量。

表中横轴（X 轴）底部标注 Fuel Consumption Comb (L/100 km)，数值范围为 10 – 70 L/100 km，表示燃油消耗量。

使用的数据包含 COMPACT、SUV - SMALL、MID-SIZE 等共 16 种车辆类型，是图表中需要对比的主体。

对比不同车辆类别的燃油效率（L/100 km）与对应的二氧化碳排放量（g/km）发现燃油消耗越高，CO2 排放通常也越高。

5.4 分类变量总结

品牌分布（MAKE）：FORD（577 辆）和 CHEVROLET（515 辆）是数据集中占比最高的品牌，其次是 BMW（501 辆），而 MERCEDES-BENZ（365 辆）和 PORSCHE（296 辆）数量较少。推测 FORD 和 CHEVROLET 在目标市场或数据来源中占据主导地位，可能与其经济型车型或市场策略有关。

车型分布（MODEL）：前五名均为 FORD 旗下车型（如 F-150 FFV、MUSTANG 等），表明该品牌在数据集中具有显著的车型集中性。可能反映 FORD 在特定车型（如皮卡 F-150 系列）的市场优势或数据采集偏向性。

车辆类别（VEHICLE_CLASS）：小型 SUV（1006 辆）和中型车（983 辆）占比最高，其次是紧凑型车（903 辆），显示消费者或市场对多功能车型的偏好。标准 SUV（613 辆）和小型车（533 辆）相对较少，可能与实用性需求或数据来源（如区域市场特征）相关。

变速器类型（TRANSMISSION）：自动变速器（AS6、AS8、A6 合计占比约 65%）占据绝对主流，手动变速器（M6、AM7）占比较少。高比例自动挡车辆符合现代乘用车市场趋势，尤其是中高端品牌（如 BMW、MERCEDES-BENZ）的配置特点。

燃料类型（FUEL_TYPE）：X（3039 辆）和 Z（2765 辆）是主要燃料类型（需结合具体定义，推测为汽油或混合动力）。柴油（D, 147 辆）和乙醇（E, 330 辆）占比极低，新能源（N）仅 1 辆，反映数据集以传统燃油车为主，新能源车几乎未覆盖。

6 机器学习预测 CO2 排放量

6.1 模型构建与训练

6.1.1 神经网络架构设计

模型采用顺序（Sequential）结构，逐层堆叠全连接层（Dense）与正则化组件。输入层与特征维度匹配，输入形状为(x_train.shape[1],)，即训练数据的特征数（共 11 个特征，CO2 排放量为目标变量）。输入层维度与特征数匹配（x_train.shape[1]），随后依次添加 128 节点的 Dense 层（ReLU 激活）、批标准化（BatchNormalization）及 Dropout 层（丢弃率 0.2）以防止过拟合。后续层依次为 64、32、16 节点的 Dense 层（均使用 ReLU 激活），最终输出层为 1 个节点（线性激活），用于回归预测。

模型运行后各层参数：

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 128)	1536
batch_normalization (BatchNormalization)	(None, 128)	512
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 64)	8256

dense_2 (Dense)	(None, 64)	4160
batch_normalization_1 (BatchNormalization)	(None, 64)	256
dropout_1 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 32)	2080
dense_4 (Dense)	(None, 16)	528
dense_5 (Dense)	(None, 1)	17

==========

Total params: 17,345

Trainable params: 16,961

Non-trainable params: 384

6.1.2 超参数设置与优化

模型使用 Adam 优化器，学习率设置为 0.001，损失函数为均方误差（MSE），评估指标为平均绝对误差（MAE）。模型总参数量为 16,963，其中可训练参数占主要部分。

损失函数使用均方误差（MSE），衡量预测值与真实值的平方差，适用于回归任务。

6.2 训练过程分析

训练过程中，数据经 MinMaxScaler 归一化处理，将数值缩放到[0,1]区间，划分为训练集与测试集。归一化公式：

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

引入早停机制（EarlyStopping），监控验证集损失（val_loss），若连续 10 轮未改善则终止训练并恢复最佳权重。实际训练共进行 60 轮，训练损失（MSE）从初始 48,757 逐步下降至约 260，验证损失同步从 13,746 降至最低 38.96，表明模型有效收敛。但是，验证

损失在第 6 轮后显著下降（从 463.64 降至 145.58），随后波动趋稳，未出现明显过拟合。

6.3 模型评估与结果

模型在测试集上表现优异。均方误差（MSE）为 38.96，平均绝对误差（MAE）为 4.47，表明预测值与真实值的平均偏差约为 4.47 单位。决定系数（R^2）高达 0.9893，说明模型能解释 98.93%的目标变量变异，拟合效果极佳。

模型拟合效果极佳（R^2 接近 1），有过拟合风险（训练损失与验证损失差距较小，未明显发散）。

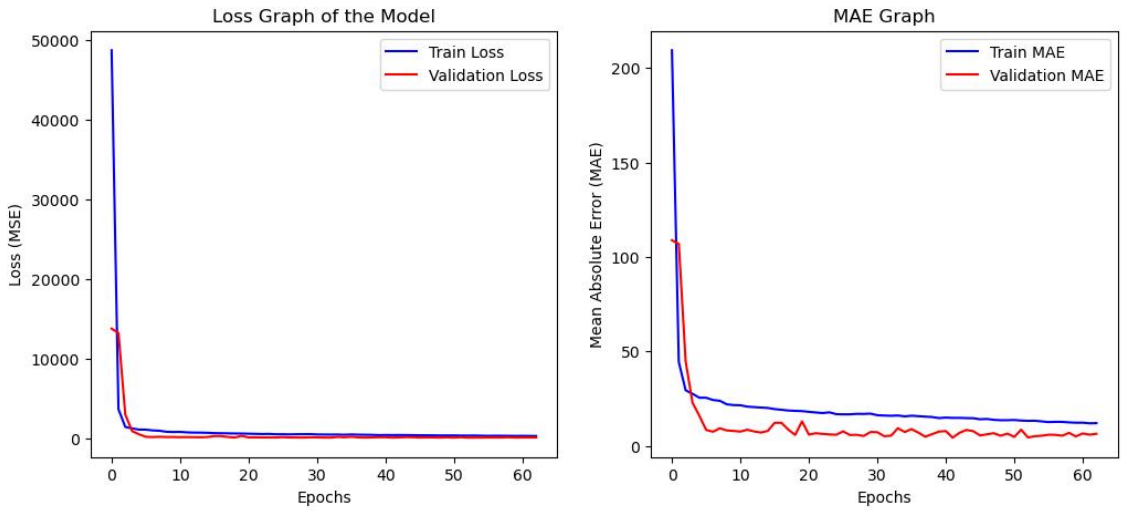

图 11 损失与 MAE 曲线

训练损失与验证损失同步下降，未见严重过拟合。MAE 曲线显示模型在后期训练中趋于稳定。 预测值与真实值散点图相近，数据点紧密分布在对角线附近，进一步验证模型准确性。

7 实验结论与展望

本实验第一步通过多维度分析揭示了车辆 CO2 排放的关键影响因素，可以得出关于 CO2 排放的相关结论：(1)发动机参数（排量、气缸数）与排放呈强正相关；(2)车辆类别和燃料类型显著影响排放水平；(3)油耗指标（尤其是综合油耗）是预测 CO2 排放的有效代理变量。

本实验第二步成功构建了一个基于神经网络的 CO2 排放预测模型，在测试集上实现了 MAE=4.47 的高精度预测。然而，代码实现中的技术细节需进一步完善。项目将来可以继续探索，例如(1)进行多模型对比，尝试随机森林、XGBoost 等树模型，比较性能差异[7]；

(2)进行可解释性分析，使用 SHAP 值或 LIME 解释特征重要性[8]；(3)开发实时预测系统，将模型部署为 API，支持实时碳排放查询[9]；(4)通过持续优化，该模型有望成为车辆环保评级与政策制定的核心工具。

本实验也存在一些不足，在进行探索性数据分析时，发现数据存在区域偏差，数据以北美车型为主，缺乏欧洲、亚洲市场代表性。同时车辆数据存在缺少和缺失现象，数据准确性和可靠性下降。

根据探索性数据分析和机器学习得出的结论，可以提供一些减排策略建议。(1)推广小型化与混动技术，小型车和混动车型的排放较低，应通过补贴政策鼓励消费者选择；(2)优化自动变速器效率，自动变速器占比高，但油耗普遍较高，需提升传动系统能效；(3)加强柴油车排放监管，柴油车排放较高，需加装尾气处理装置或限制其使用范围。

参考文献：

[1]Park J ,Lee H ,Park S .Development of real-road CO2 emission factors for diesel light-duty vehicles across diverse driving conditions[J].Energy,2025,324136088-136088.

[2]Özgün B ,Yasin K ,Onur G , et al.Numerical and experimental investigation of fuel consumption and CO2 emission performance for a parallel hybrid vehicle[J].Alexandria Engineering Journal,2021,60(4):3649-3667.

[3]Danyue Z ,Hepeng Z ,Yan C , et al.Quantifying the heterogeneous impacts of the urban built environment on traffic carbon emissions: New insights from machine learning techniques[J].Urban Climate,2024,53

[4]Co2 Emissions; Reports Outline Co2 Emissions Study Findings from Xiamen University (Investigating the differences in CO2 emissions in the transport sector across Chinese provinces: Evidence from a quantile regression model)[J].The Business of Global Warming,2018,292-.

[5]Xing Wang , Yikun Su a, Zhizhe Zheng , et al.Heterogeneity analysis of CO2 emissions driving features on road transport: New perspective from interpretive machine learning[J].Journal of Cleaner Production,2025

[6]Han T T T ,Lin Y C .Exploring long-run CO2 emission patterns and the environmental kuznets curve with machine learning methods[J].Innovation and Green

Development,2025,4(1):100195-100195.

[7]宋特,付金生,卢春颖,等.粤港澳大湾区连续刚构桥建设碳排放测算与分析[J/OL].公路 ,2025,(04):318-326[2025-05-01].http://kns.cnki.net/kcms/detail/11.1668.U.20250408.1443.092.html.

[8]于小曼.基于机器学习的城市臭氧驱动因素及浓度预测研究[D].上海第二工业大学,2024.DOI:10.27916/d.cnki.ghdeg.2024.000160.

[9]马亚,陆旭东,袁月明,等.交通能源融合背景下的高速公路低碳运营研究[J].公路,2025,70(02):413-419.