EDA INSIGHTS FOR IMPORT CAR DEALERSHIP SUCCESS

MYRAJOY KIGANANE

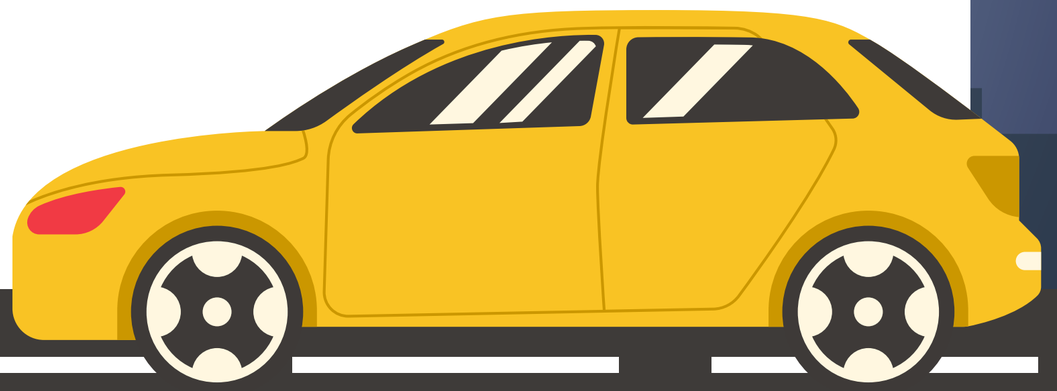
Presented By Barcelle Group

INTRODUCTION

- A new international car dealership is entering the market and seeks guidance on selecting imported vehicles with competitive pricing and low insurance risks.
- To support this goal, the team is analyzing the 1985 Auto Imports dataset to uncover how vehicle features impact key outcomes.
- The focus is on understanding the influence of features on market price, insurance risk ratings, and normalized annual loss.
- The aim is to provide actionable insights to shape the dealership’s purchasing and marketing strategies for a successful launch.

KEY BUSINESS QUESTIONS

1. Which car brands command higher market prices and what characteristics drive this?
2. What car body types are associated with higher insurance risk?
3. What impact do engine size and horsepower have on the price of a car?
4. How does fuel type influence the normalized annual losses and insurance risk of vehicles?
5. How does drive-wheel configuration (FWD, RWD, 4WD) affect the average price and insurance rating of cars?



DATA OVERVIEW

1

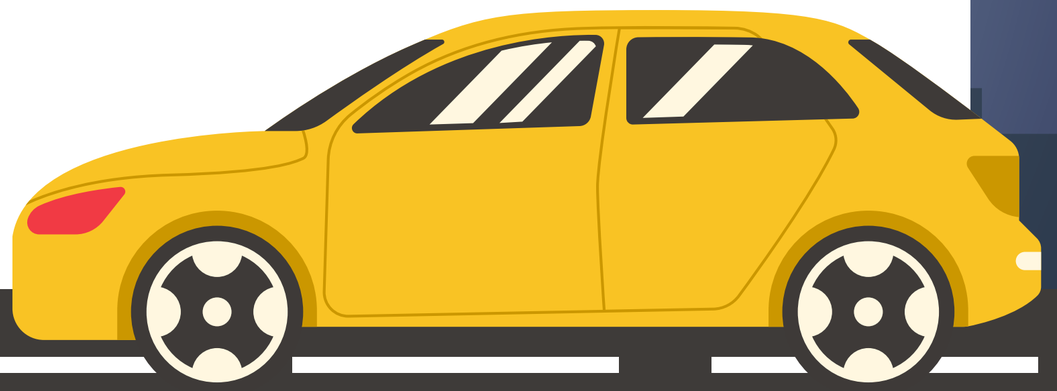
The dataset, sourced from Kaggle, contains detailed 1985 specifications and insurance data for imported cars. (https://www.kaggle.com/datasets/sumaya23abdul/automobile-database?resource=download)

2

The features used to predict price include make, body-style, fuel-type, engine-size, horsepower, symboling, and normalized-losses

3

The target variable is price, representing the market value of each vehicle.

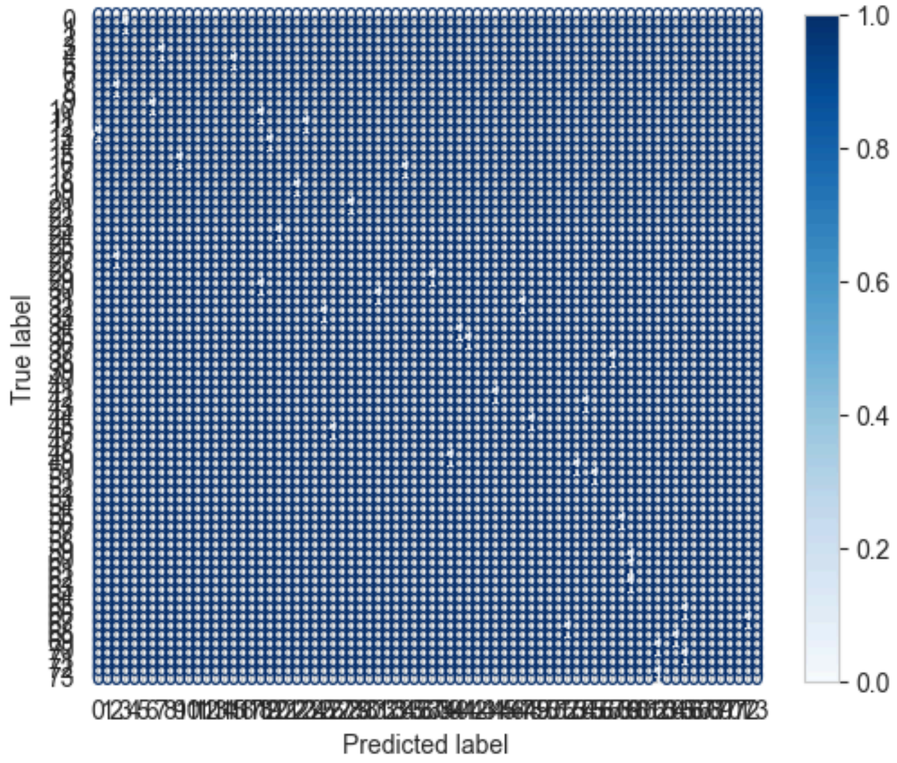


METHODOLOGY

1. Data Collection: Extracted relevant car data from Automobile Dataset
2. Data Preprocessing: Cleaned and split data using test_train_split
3. Data modelling : Uncovered prediction in the mean of the baseline model using linear regression and decision tree regressor
4. Data Evaluatuon: Used classification metrices, confusion martix and ROC

VISUALIZATIONS

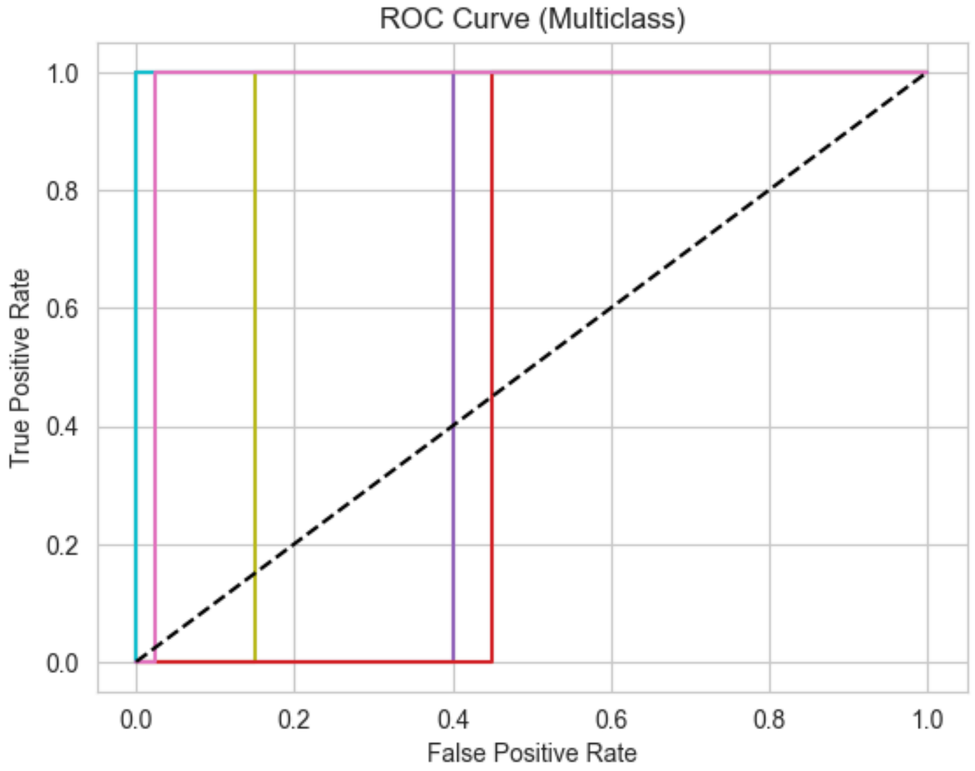
CONFUSION MATRIX



Based on the dispersed distribution of predictions across the confusion matrix and the absence of strong diagonal concentrations, it's evident that the model is struggling with accurate classification, indicating a need for further analysis and potential improvements to its performance.

VISUALIZATIONS

ROC



The ROC curve above illustrates the performance of a multiclass classification model, plotting the true positive rate (TPR) against the false positive rate (FPR). The colored lines represent the various classes' trade-offs between sensitivity and specificity, while the dashed diagonal line indicates random performance.

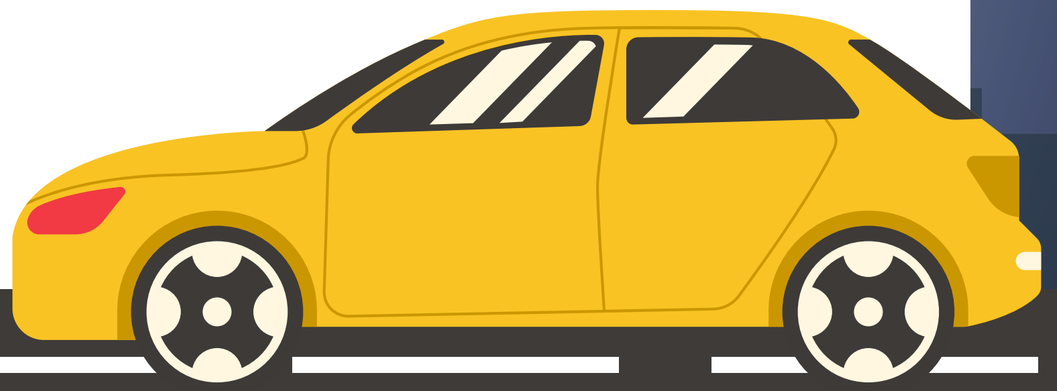
KEY FINDINGS

1. Ridge Regression performed best ($R^2 = 0.94$), while Linear Regression overfit the training data and the Decision Tree risked overfitting

2. The classification model failed with all metrics at 0.02 and AUC not computable due to model or data issues.

3. Ridge Regression balanced accuracy and generalization best, while OLS overfit and Decision Tree needed tuning.

4. OLS showed severe multicollinearity, making its coefficients unstable and hard to interpret

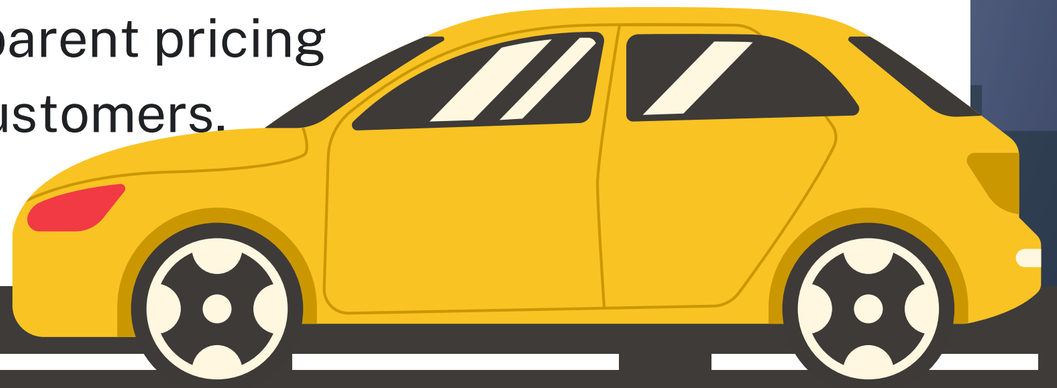


RECOMMENDATIONS

Investing in the development and marketing of vehicles with efficient, powerful engines is crucial, as larger engines and higher horsepower have a strong positive correlation with higher perceived value and pricing.

Cars with Rear-Wheel Drive (RWD) and 4WD configurations tend to have higher average prices and appeal to performance-conscious buyers, making it valuable to evaluate the cost-benefit of producing more AWD or RWD vehicles in premium segments.

Certain body types may lead to higher insurance symboling values, so partnering with insurance companies to offer transparent pricing based on body-type risk categories would benefit customers.



CONCLUSION

Data-driven decision-making is essential for optimizing vehicle design, pricing, and marketing strategies.

- Future steps: Conduct deeper analysis of engine performance, drive configurations, and insurance-related factors.
- Implement insights to enhance product value, attract target buyers, and improve overall profitability.

THANK
YOU