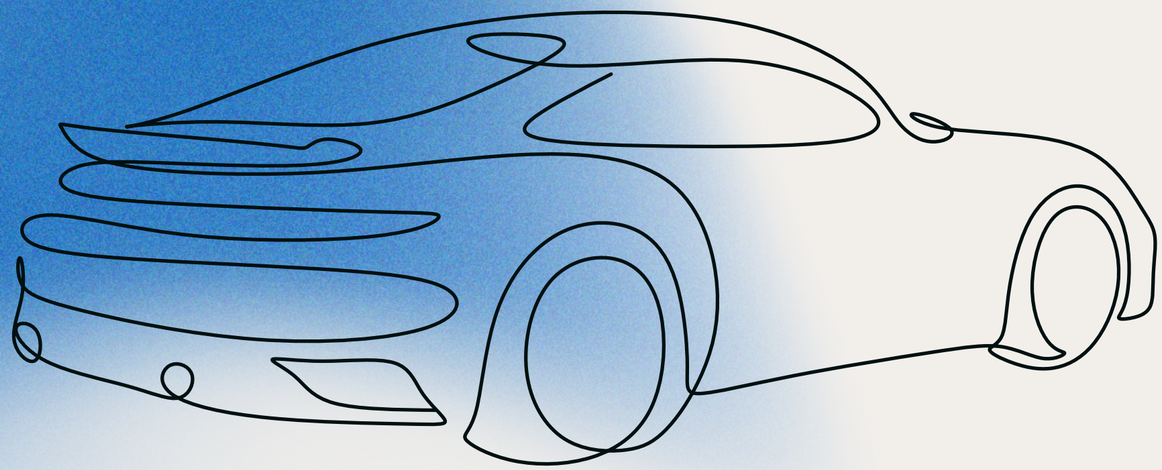
Multimodal Car Analytics



PRESENTED BY: F20DL GROUP 12

Our Topic

- To use regression models on tabular data of car specifications and price, and predict the resale price of used cars.
- To use classiciation model on image data, to classify car images from various angles and determine its details such as, maker, model, color, and body type of the car.

Dataset: DVM-CAR

- 1.45M cleaned car images (background removed, front/side views)
- These images cover 899 car models in the UK market over past two decades.
- We mainly used the ad_table and price_table for the features of car specifications, like engine size, fuel type, model, mileage, etc, and trained the model on these to accurately predict the resale price of used cars.

Dataset Citation (Source of Dataset):
Jingming Huang, Bowei Chen, Lan Luo, Shigang Yue, and Iadh Ounis. (2022). "DVM-CAR: A large-scale automotive dataset for visual marketing research and applications". In Proceedings of IEEE International Conference on Big Data, pp.4130–4137. [PDF_link]

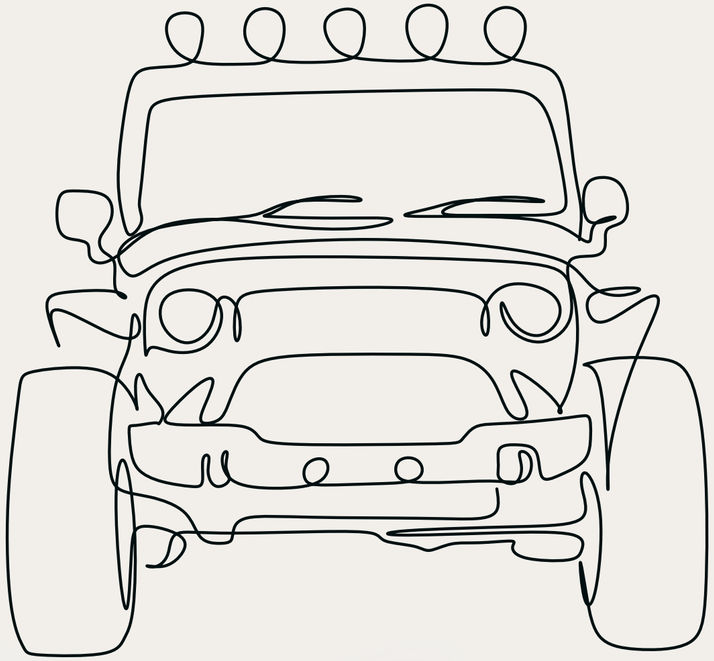
Tabular Data Pre-Processing

Data Cleaning & Feature Creation

- **Tables Used:** Price_table & Adv_table.
- **Initial Cleaning:** Removed whitespaces from column names.
- **Numeric Conversion:** Cleaned and converted Target Price and Runned_Miles by removing non-numeric characters ($).
- **Engine Size:** Cleaned Engin_size (e.g., '2.0L') to create a numeric Engine_size_L.
- **Data Merge:** Joined tables on Maker, Genmodel_ID, and Year to incorporate Entry_price (MSRP) as a feature.

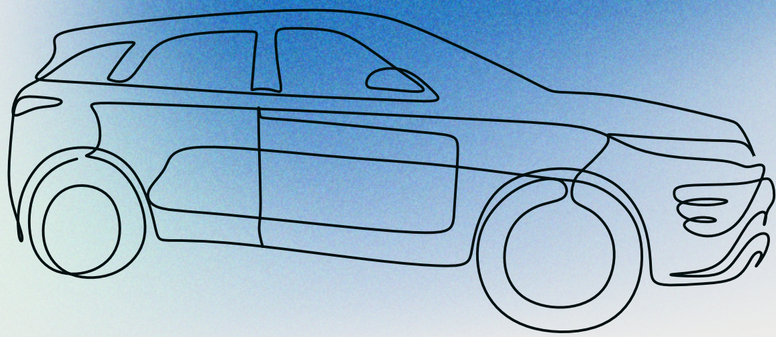
Missing Values & Feature Selection

- **Null Handling:** Removed rows missing core values (Miles, Engine, Seat/Door count).
- **Imputation:** Filled remaining nulls in Entry_price with the median.
- **Feature Sets:** Selected both numerical and categorical features (Maker, Bodytype, Fuel_type, etc.).



Final Preparation & Splitting

- **Encoding:** Applied One-Hot Encoding to all categorical features.
- **Train/Test Split:** Used a custom 80/20 split performed per Maker to ensure proportional representation.



Linear Regression

Baseline Price Prediction Model

Why Linear Regression was used?

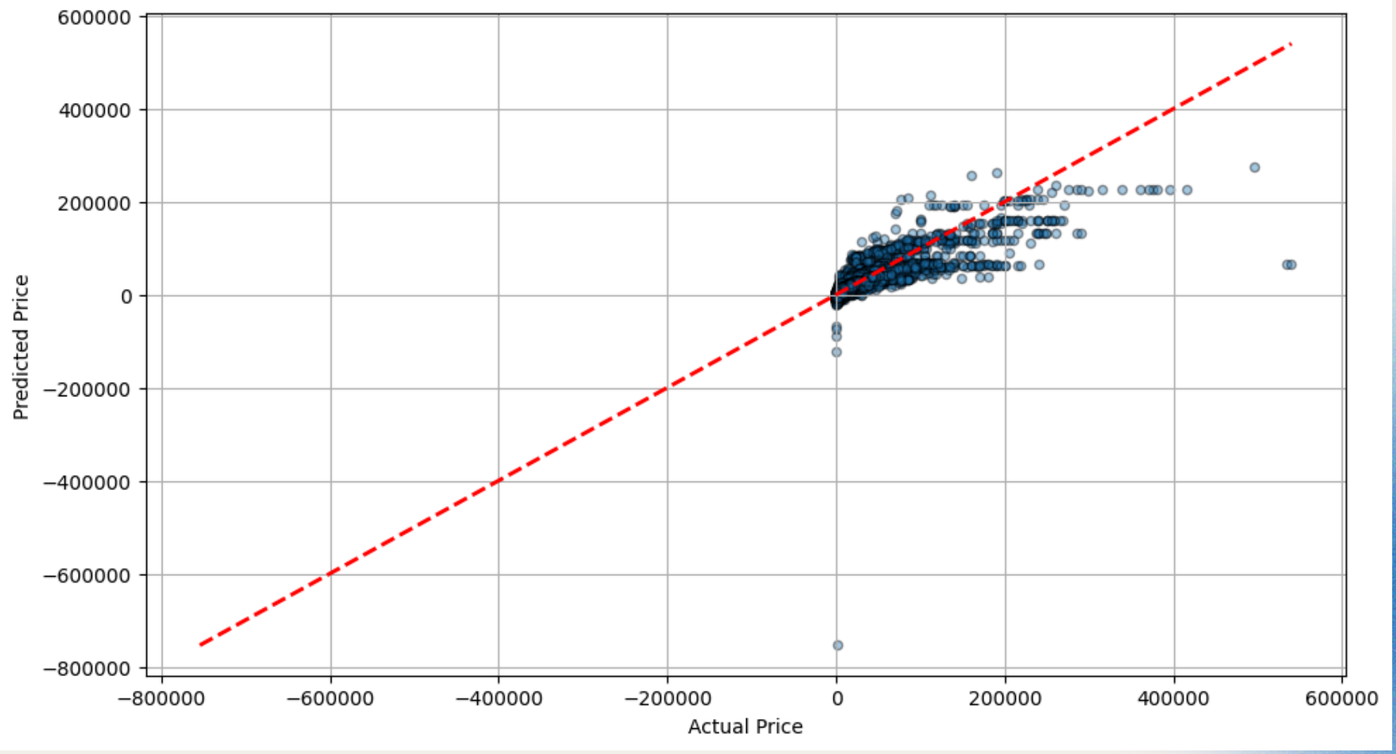
- Establishes a baseline to compare with complex models
- Shows how far linear assumptions can go on car pricing

Performance

- R^2 : **0.718**
- MAE : **£4,330**
- RMSE : **£10,050**
- Errors are high for **premium brands** and **non-linear patterns.**

Takeaway

Car prices follow non-linear depreciation and brand value differs greatly, so Linear regression underperforms so we proceed to trying other models such as XGboost and Random Forest.



Car Resale Price Prediction – Actual vs Predicted

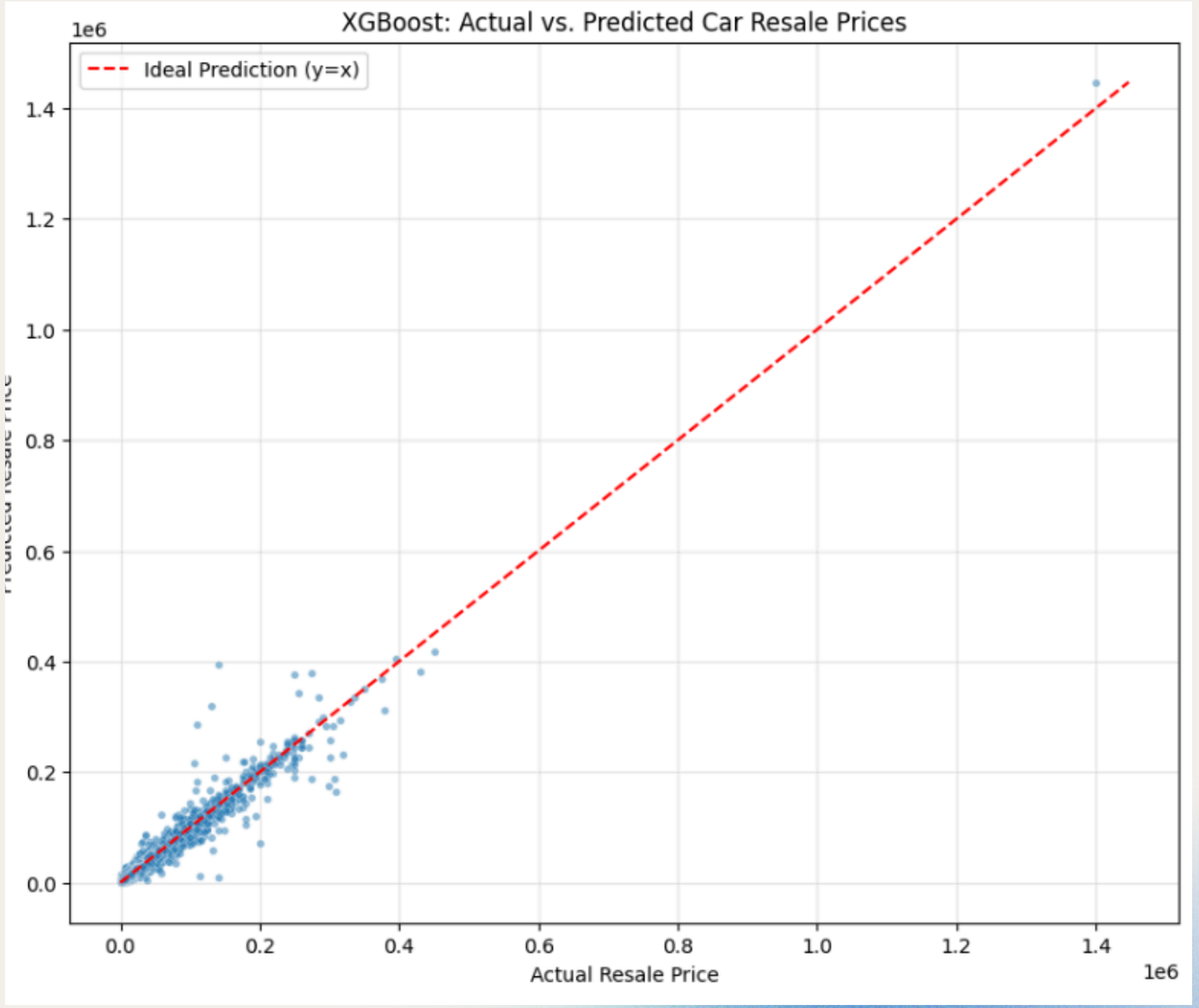
eXtreme Gradient Boosting (XGBoost)

Hyperparameter Tuning

Parameter	Value	Description
objective	'reg:squarederror'	Sets the task to Regression (minimizing MSE).
n_estimators	100	The number of decision trees built in the ensemble.
max_depth	10	High complexity setting to capture detailed, non-linear relationships.
learning_rate	0.1	To control the step size and prevent overfitting.
random_state	42	Ensures reproducibility of the model results across runs.

Results

R-Squared (R^2)	0.9546
MSE	2030.93
RMSE	4247.05



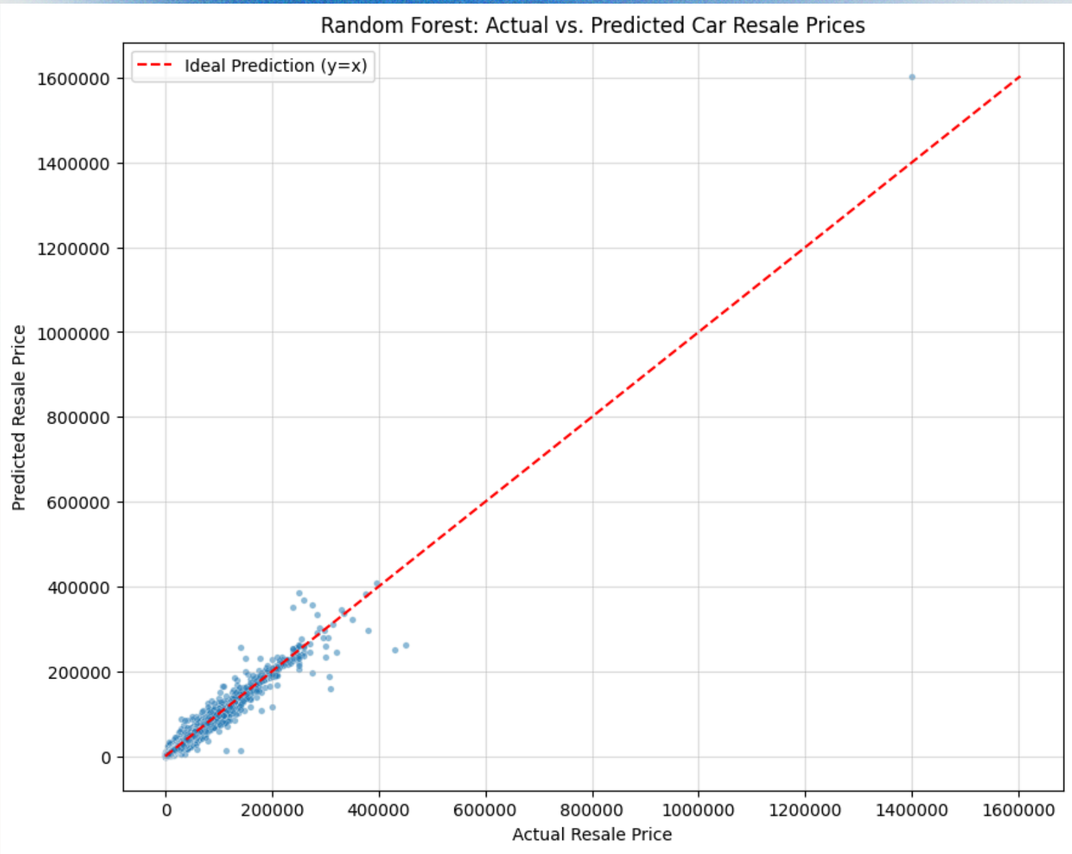
Random Forest

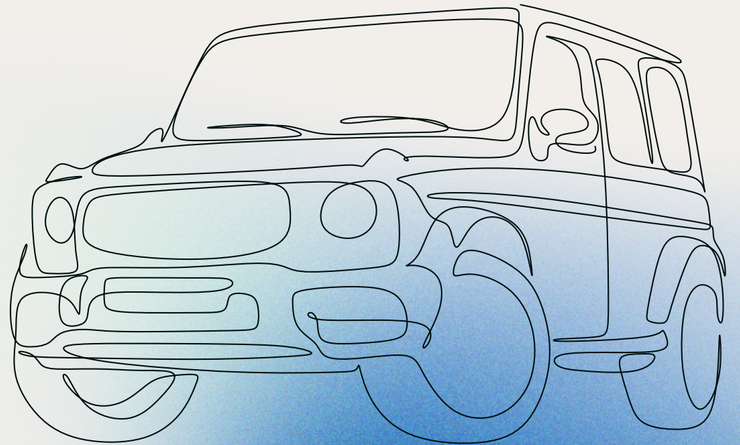
- Builds multiple decision trees and uses average of their predictions.
- Works very well with non-linear relationships and mixed features, making it very suitable for analysing complex car price patterns.

Experimenting

- Experimented with different values for the parameters (number of trees, max_depth, and min_split_size)
- Increasing the number of trees increased the accuracy but also the computation time.
- After testing, figured the most optimal vlaues.

R-Squared (R^2)	0.965
MAE	1,474.49
RMSE	3,683.92





MULTIMODAL CAR ANALYTICS

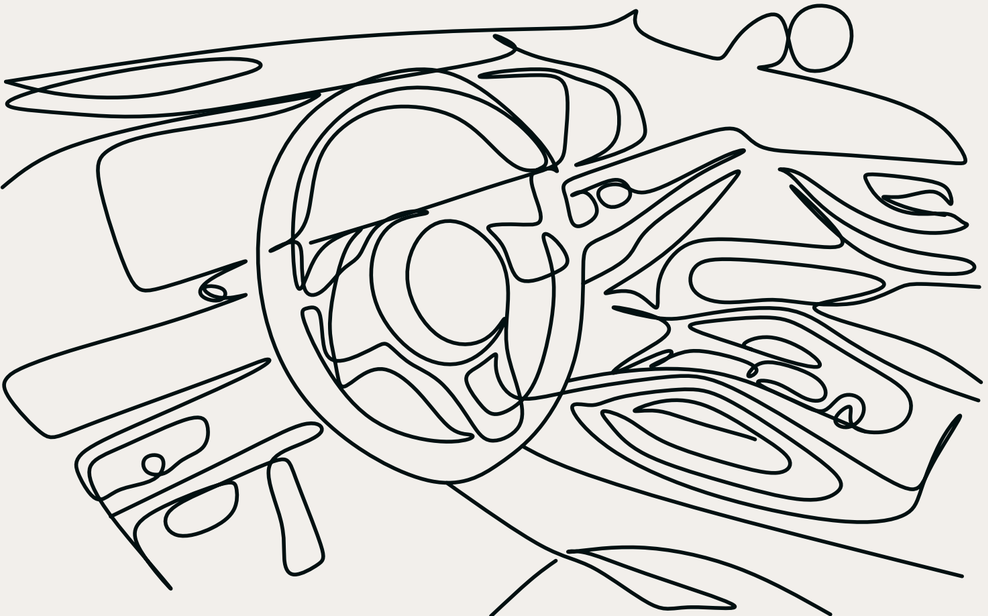
CNN

RQ2 (Classification)

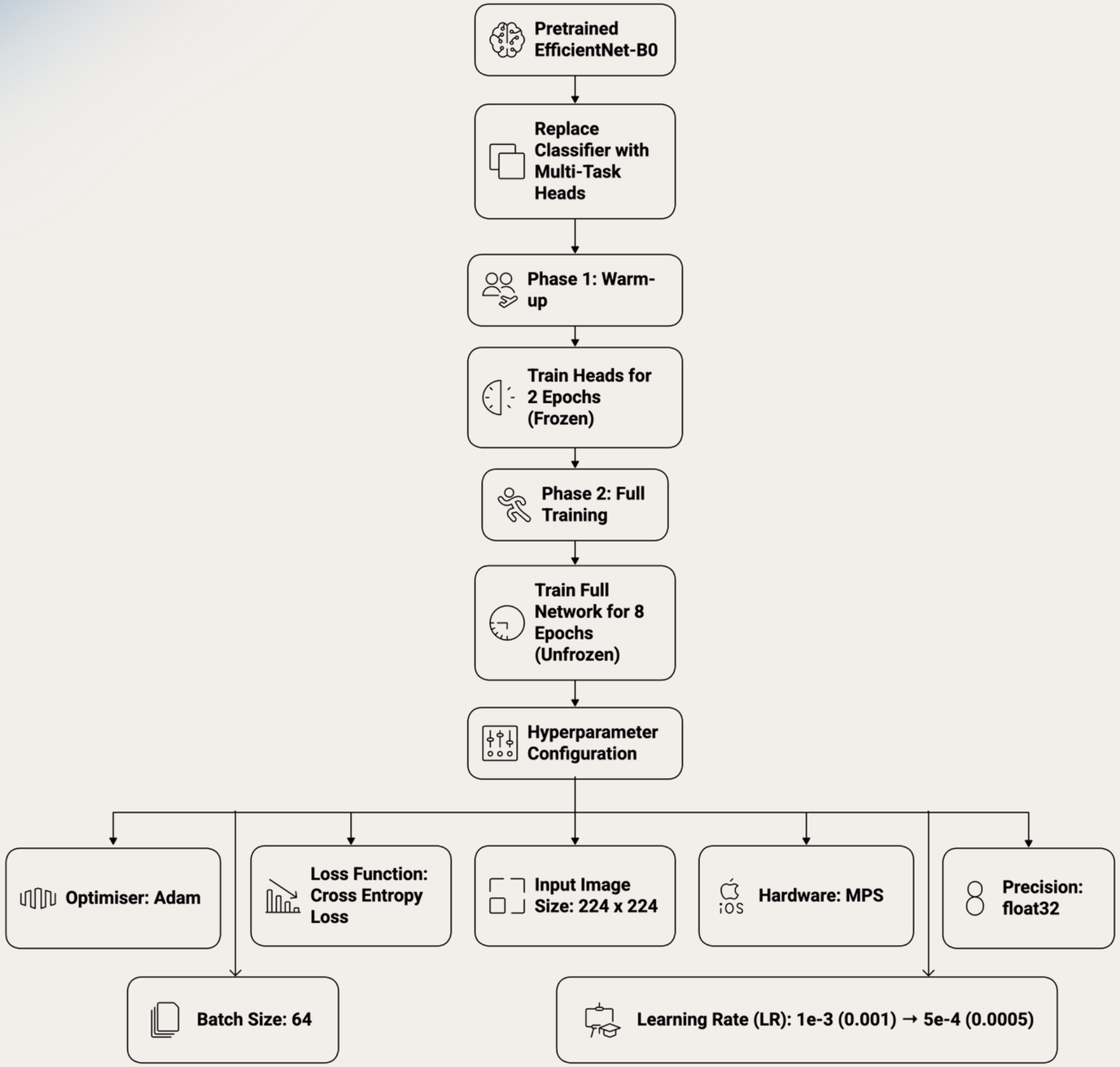
Can body type and manufacturer be accurately inferred from listing images of varying angles?

Success indicators

RQ2: Accuracy & F1 scores
precision + recall
(↑F1 = ↑recognition)



DualHeadCarNet Training Process



CNN Results

1. The Evaluation Workflow (evaluate.py)

- Checkpoint Loading: The system reconstructs the DualHeadCarNet using the saved .pt file and restores class mappings (Indices to Labels).
- Data Processing: Runs on a dedicated Test Subset (unseen during training) using a batch size of 64.
- Logits vs. Probabilities:
- The model outputs Logits for all 4 heads.
- For Metrics: Uses Logits to determine the winning class index directly.
- For User Display: Uses Softmax to convert Logits into confidence percentages (e.g., "98% Audi").

2. Experimental Results (Test Subset)

- **Phase 1 (Frozen Backbone):**
  - Excellent performance on Maker/Body (~97%) due to strong pre-trained ImageNet features.
  - Poor performance on Colour (~26%) and Model Variant (~88%).
- **Phase 2 (Unfrozen Fine-Tuning):**
  - Colour Accuracy: Jumped from 26% to 86%
  - Model Variant: Improved to 93%.
  - Conclusion: The two-phase strategy successfully adapted the backbone without catastrophic forgetting of basic shapes.

Metric	Accuracy (Frozen)	Accuracy (Unfrozen)	Improvement	Final F1 Score
Loss	2.6811	0.7975	-70.2%	-
Maker Head	97.44%	97.44%	Stable	0.7329
Body Head	97.02%	97.21%	+0.19%	0.8045
Model Head	88.34%	93.84%	+5.5%	0.5574
Colour Head	26.26%	86.39%	+60.13%	0.7189

APPENDIX

Literature Review:

Huang, J., Chen, B., Luo, L., Yue, S., & Ounis, I. (2022). "DVM-CAR: A large-scale automotive dataset for visual marketing research and applications." IEEE Big Data 2022.

- The original dataset paper describes the DVM-CAR dataset itself, with ~1.45 million images and six relational tables (Basic, Sales, Price, Trim, Ad, and Image). They show example applications such as using CNNs (e.g., fine-tuned VGG/ResNet) to infer "design modernity scores" from images and correlate them with sales trends or model withdrawals.
- We build directly on their dataset, but whereas they only showcase illustrative uses, we will implement a rigorous supervised pipeline aimed at classification (body/manufacturer) and regression (price & depreciation) on identical adverts, with error breakdowns, leakage control, and a dual-task focus they don't carry out.

Du, Siyi; Zheng, Shaoming; Wang, Yinsong; Bai, Wenjia; O'Regan, Declan P.; Qin, Chen (2024). "TIP: Tabular-Image Pre-training for Multimodal Classification with Incomplete Data." ECCV 2024.

- They use 176,414 image–tabular pairs sampled from DVM to train a car model classification task with 283 classes. Their method (TIP) combines masked tabular reconstruction, image–tabular contrastive matching, and an interaction module, using CNN + tabular encoder architectures.
- Unlike TIP, which is a multimodal classification work, we separate tasks: image classification (RQ2) and tabular regression (RQ1). That decomposition simplifies attribution of performance and lets us study the independent predictive power of each modality. We will also tackle the regression of price and depreciation, which TIP does not.

Yaman Abu Ghareebaih (2025). "Car Price Prediction Using Machine Learning: Analyzing the DVM-CAR Dataset." Master's Thesis, East Tennessee State University.

- This thesis explicitly uses DVM-CAR to build and compare regression models that predict used car prices from features like mileage, engine power, registration year, etc. He performs cleaning and encoding and applies linear regression and random forest models, evaluating with MSE/MSEP metrics.
- His work aligns with our RQ1 regression focus, but he does not integrate image classification or depreciation prediction. We enhance his methodology by also modelling % depreciation, plus combining classification (RQ2) and regression (RQ1) in one unified project, with leakage control and error stratification.

Roy, K., Krämer, L., et al. (2025). "Multi-Modal Contrastive Pre-training for Enhanced Tabular Data Analysis (MT-CMTM)." Preprint / arXiv.

- They evaluate MT-CMTM on DVM. Their method fuses contrastive learning and masked tabular modelling with image signals, using a 1D-ResNet + attention tabular backbone plus contrastive alignment. They report a 2.38% absolute accuracy gain on DVM classification tasks over baseline modelling.
- Their work is advanced in multimodal fusion for classification, but they don't target regression of price or depreciation. Our plan differs by assessing classification and regression as separate tasks initially, before optional fusion—thus giving clearer insights and more direct alignment with price prediction goals.

Du, S., Luo, X., O'Regan, D. P., & Qin, C. (2025). "STiL: Semi-supervised Tabular-Image Learning for Comprehensive Task-Relevant Information Exploration in Multimodal Classification." CVPR 2025.

- STiL is a semi-supervised multimodal classification framework that is evaluated partly on DVM. It learns from both labelled and unlabelled image-tabular pairs using a disentangled contrastive consistency module, consensus-guided pseudo-labelling, and prototype-guided label smoothing. It uses CNN encoders for images and transformer/tabular encoders for tabular data.
- We don't plan to use semi-supervised or fusion models at first. Instead, we treat classification (RQ2) and regression (RQ1) as separate supervised tasks, which gives clearer attribution of performance to each modality. We also introduce percentage depreciation prediction, strict advert-level data splits, and error breakdowns by make/price decile, which STiL's classification focus does not cover.