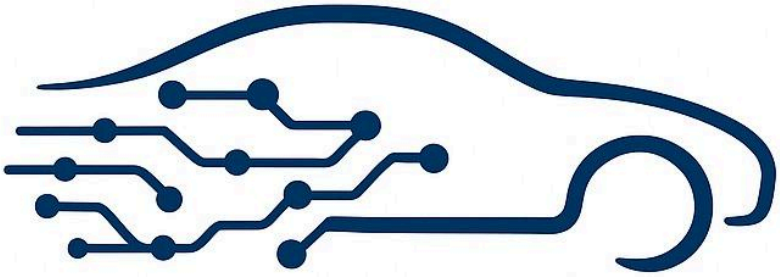
ie UNIVERSITY
SCHOOL OF SCIENCE & TECHNOLOGY



AutoWorth

Before you sell, Know it well

Mariana Saca
Isaura Llort
Jim Vincent
Louay Mohaisen
Shambhavi Chhabra
Teresita Alessandri

Table of contents

01. Executive Summary

The automotive resale market suffers from price uncertainty and inconsistent valuation methods. Private sellers frequently undervalue their cars due to limited market knowledge, while buyers risk overpaying. Dealers and online resale platforms struggle to align expectations between both sides without a transparent, data-driven mechanism.

AutoWorth proposes an intelligent ML system that predicts a car's **fair resale value** using features such as make, model, year, mileage, fuel type, engine size, tax, and mpg. It will act as a **plug-in valuation engine** for online marketplaces like Cars24, AutoScout24, or OLX Autos. Sellers can instantly check an objective price before listing; buyers can validate whether an asking price is reasonable; and platforms can use the same API to automate pricing recommendations at scale.

This addresses a real customer pain point: lack of trust and time lost in negotiation. **AutoTrader UK**[1] (2024), reports that private listings[2] take an *average of 27 days* to sell, with price negotiations being the major source of delay. We project that **predictive valuation can reduce sale duration by 10–15%**, based on benchmarked marketplace behavior. The model will estimate a "fair price" using supply-side features such as vehicle attributes and historical market listings, providing a transparent and data-driven reference for buyers and sellers. By capturing pricing patterns across brands, fuel types, and model years, AutoWorth can help align expectations between both sides, improving buyer confidence and platform efficiency.

The model will be developed and deployed following the **MLOps Iterative Loop** (Scope → Data → Model → Deploy → Monitor) to ensure reliability, scalability, maintainability, and adaptability.

Our objective is to achieve a Mean Absolute Percentage Error (MAPE) below 12% on the test set, corresponding to a predicted price accuracy within ±£1,200 for most vehicles. In a simulated marketplace scenario, this performance level is expected to reduce average days-to-sell by at least 10%.

Expected impact

1. Reduction in average negotiation and sale duration.
2. Higher conversion and return-user rate on resale platforms
3. Foundation for future AI features such as depreciation forecasting and fairness audits.

[1] The largest digital automotive marketplace in the UK
[2] Refer to vehicles advertised directly by individual sellers rather than dealerships or professional traders.

02. Business Case & Objectives

2.1 Business Rationale

Pricing in the used-car market is a complex equilibrium between consumer perception, vehicle depreciation, and information asymmetry.
Traditional valuation approaches such as dealer heuristics or static price guides cannot fully capture temporal dynamics like seasonality, fuel-price changes, or shifts in consumer preference.
An ML-based valuation engine offers a data-driven complement to these conventional methods by learning multi-factor relationships directly from historical listings.

From a business-systems perspective, integrating predictive valuation into an online marketplace can:

1. Reduce the need for manual pricing oversight,
2. Support dynamic pricing strategies, and
3. Create a consistent, explainable framework for both private sellers and dealers.

The UK used-car sector, valued at roughly **£118 billion (2023)** and involving more than seven million annual transactions (Statista 2024), provides ample scope for incremental efficiency gains. Academic and consulting literature (e.g., McKinsey 2022) suggests that applying predictive analytics to pricing decisions can improve conversion rates and negotiation outcomes by *approximately* 10–20 % in similar retail contexts.

This project therefore explores whether an interpretable regression model could capture the structural relationships between vehicle attributes and observed sale prices, rather than attempting to guarantee any specific commercial impact.

The model's quantitative targets align directly with business KPIs: a test-set MAE ≤ £1,000 or MAPE ≤ 12% ensures price transparency within a tolerable negotiation margin; latency ≤150 ms per API call maintains smooth platform UX; and fairness checks across fuel types and brands will monitor residual bias below ±5% of mean price. These thresholds operationalize our success criteria and link ML performance directly to user trust and conversion rates.

2.2 Stakeholders and Value Alignment

1. **End Users (buyers / sellers):** benefit from greater pricing transparency and reduced uncertainty.
2. **Dealerships and Platforms:** obtain scalable, auditable price recommendations that support internal valuation processes.
3. **Data Science Team:** gains a controlled environment for testing lifecycle automation (tracking, deployment, retraining) using the MLOps framework taught in class.

The project also provides a practical demonstration of the MLOps maturity framework (Level 1), emphasizing automation, reproducibility, and model monitoring.

03. Scope, Constraints & Risks

3.1 In Scope

The project develops a full ML pipeline for used car price prediction, covering data cleaning, feature engineering, and regression modeling (Random Forest, XGBoost). Experiments are tracked in MLflow for reproducibility. The best model is containerized with Docker, deployed via FastAPI, and monitored through a dashboard that visualizes performance and drift demonstrating version control, retraining, and the full MLOps loop.

3.2 Out of Scope

Full-scale **production integration** with commercial platforms or third-party APIs lies beyond this project's scope. Similarly, no **frontend or mobile application** will be developed, as the focus is strictly on backend automation and lifecycle management. The work is designed to remain within academic boundaries, emphasizing **methodology over commercial implementation**.

3.3 Constraints

The project will be conducted over **four weeks**, following the course timeline. Computational resources are limited to **standard personal computing resources**, constraining model complexity. The dataset contains only **UK car listings**, which limits generalization to other markets. These boundaries are acceptable, as the objective is to demonstrate reproducible model development and MLOps principles rather than global deployment.

Risks and Mitigation

- **Brand Imbalance:** Certain manufacturers (e.g., Audi, BMW) dominate the dataset, potentially biasing results.
    - **Mitigation**: Apply **stratified sampling** and **class weighting** during training to ensure balanced representation.
- **Data Drift:** Market prices and patterns may shift over time, reducing model reliability.
    - **Mitigation**: Implement **scheduled retraining** and **continuous performance monitoring** to maintain model accuracy.
- **Overfitting:** The model may learn patterns specific to training data, lowering generalization.
    - **Mitigation**: Use **cross-validation**, **regularization**, and **test set evaluation** to preserve robustness.
- **Latency Risk:** Containerized deployment may cause slower API responses under load.
    - **Mitigation**: Optimize **Docker configurations**, resource allocation, and **batch inference size** to maintain responsiveness.

A practical trade-off will be accepted between model complexity and response speed: we may tolerate up to +2 percentage-points higher MAE to ensure p95 latency stays below 150 ms per prediction.

04. Data Plan

4.1 Data Sources

The project uses the **"100,000 UK Used Car Dataset" (Desai, 2019)** from Kaggle, containing public listings across major UK marketplaces. Data spans brands like Audi, BMW, Ford, Hyundai, Mercedes-Benz, Toyota, Vauxhall, Skoda, and Volkswagen, with features such as model, year, price, mileage, fuel type, tax, mpg, and engine size. All brand files are merged into one dataset with a *brand* column, forming a clean, labeled foundation for supervised regression.

4.2 Data Quality and Pre-processing

Although largely clean and numerical, the dataset requires preprocessing for completeness, consistency, and representativeness:

1. **Missing values:** Median imputation per brand for *tax* and *mpg*.
2. **Outliers:** Capped via the IQR method to reduce distortion.
3. **Encoding:** Continuous variables standardized; categorical ones one-hot encoded for interpretability.
4. **Feature engineering:** Derived variables such as *car_age = current_year – year* and *mileage_per_year = mileage / car_age* capture depreciation and usage intensity.

All steps follow a logged **ELT pipeline** for reproducibility. Though diverse, the dataset overrepresents premium brands; future iterations may add enrichment data to improve representativeness.

4.3 Labeling Strategy

The target (*price*) is provided. Anomalies (zero or unrealistic values) will be removed, and duplicate listings averaged to ensure consistent, reliable labels.

4.4 Sampling and Imbalance Considerations

Premium brands dominate the data. To reduce bias, use **stratified sampling** and, if needed, **weighted regression**. Data will be split **70/15/15** (train/validation/test) with fixed random states to prevent leakage.

4.5 Avoid Common Data Pitfalls

1. **Label multiplicity:** Average duplicates.
2. **Data leakage:** Exclude or verify correlated variables.
3. **Bias detection:** Analyze residuals by brand, fuel type, and transmission.
4. **Reproducibility:** Version datasets and pipelines with **MLflow**.

These steps ensure transparent, balanced, and reproducible data management consistent with **MLOps Level 1** standards.

05. Modelling Approach

The goal of this phase is to develop a reliable, interpretable model capable of predicting used car prices based on structured vehicle data. The project focuses on building a single predictive pipeline that balances performance, interpretability, and reproducibility within an MLOps framework.

5.1 Model Selection

Since the dataset is tabular and primarily numerical, the modeling process will use supervised regression algorithms suited for structured data. Experiments will begin with **Linear Regression** to capture overall trends and continue with **tree-based ensemble models** such as **Random Forest Regressor** and **XGBoost**, which handle non-linearities and feature interactions effectively. These models are well-established for tabular prediction problems and offer strong generalization with relatively low computational requirements. All model development will follow the iterative experimentation process introduced in the course, with **MLflow** used to track parameters, metrics, and results for reproducibility.

A simple rule-based baseline (e.g., predicting median price per brand and year) will serve as a reference. Any ML model achieving at least 20% lower MAE than this baseline will be considered “useful” under our defined business criteria.

5.2 Feature Engineering

Feature transformations will capture domain-relevant aspects of car depreciation and usage. Derived features will include **vehicle age**, **mileage per year**, and **engine efficiency metrics** using **mpg and engineSize** Categorical variables such as brand, transmission, and fuel type will be encoded using **one-hot encoding**, ensuring compatibility with regression and ensemble models.
All preprocessing steps will be integrated into a reproducible pipeline to maintain consistency across training and deployment.

5.3 Evaluation

Model performance will be measured using standard regression metrics: **Mean Absolute Error (MAE)** for interpretability in currency units, **Root Mean Squared Error (RMSE)** to penalize large deviations, and **R^2** to capture the proportion of explained variance. A **70/15/15 train–validation–test split** will be used to assess generalization, complemented by **k-fold cross-validation** for stability. The focus will be on obtaining a model that performs consistently rather than on maximizing a single metric.

5.4 Interpretability

To ensure transparency, the project will analyze **feature importance** and use **SHAP values** to explain individual predictions. Residual plots will be reviewed to detect any systematic bias, such as consistent over- or underestimation for specific brands or car segments. These methods ensure that the final model is both accurate and explainable, aligning with responsible AI and MLOps best practices.

06. System Preview & Project Plan

The **AutoWorth** will operate as an integrated MLOps pipeline moving from data collection to model deployment and monitoring. The system follows the iterative lifecycle introduced in class :
Scope → Data → Model → Deploy → Monitor,
ensuring traceability and continuous improvement.

Data from the Kaggle dataset will first be ingested and preprocessed using a reproducible transformation pipeline. The cleaned and feature-engineered data will feed into a **regression model** trained through MLflow-tracked experiments. The best-performing model will then be **containerized using Docker** and deployed as a **FastAPI endpoint**, allowing real-time predictions through HTTP requests.

A lightweight **Streamlit dashboard** will serve two purposes:

1. Allow users to input car details and view predicted prices, and
2. Monitor performance trends, latency, and potential data drift.

All predictions and metrics will be logged for retraining and evaluation, closing the loop between deployment and monitoring.

6.1 System Architecture

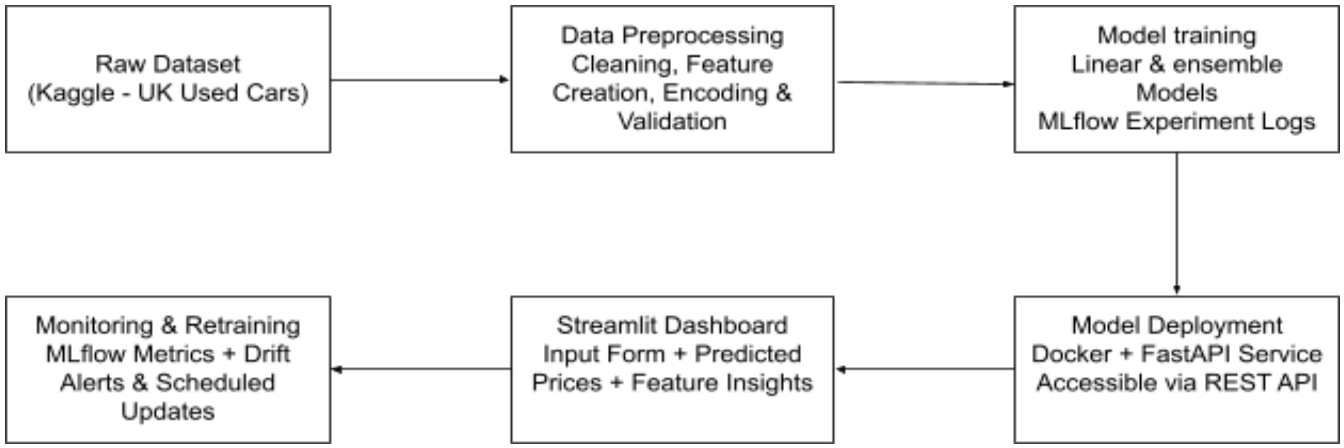


Figure 1. AutoWorth MLOps pipeline architecture

Each component will be modular and version-controlled, enabling easy updates and collaborative development. The pipeline will be deployed locally or in a lightweight cloud environment suitable for academic demonstration.

Monitoring KPIs will include model latency (p95 ≤ 150 ms), prediction error (rolling MAE ≤ £1,000), and drift detection alerts triggered when input feature distributions shift by >10% from training baseline.

6.2 Timeline

Week 1	Data cleaning, exploratory analysis and preprocessing design
Week 2	Feature Engineering and initial model training (MLflow tracking setup)
Week 3	Model tuning, containerization and API integration
Week 4	Dashboard development, monitoring setup and final evaluation

6.3 Team Roles

The project will be executed collaboratively by a six-member team, with each member responsible for complementary areas that together cover the entire MLOps workflow.

- **Shambhavi:** Focuses on **model development and experimentation**, including algorithm selection, parameter tuning, and MLflow tracking.
- **Teresita:** Oversees **system deployment and automation**, handling Docker containerization, API integration, and monitoring scripts.
- **Isaura:** Works on **data preparation**, ensuring dataset quality through cleaning, transformation, and feature encoding.
- **Jim:** Contributes to **feature engineering and model optimization**, evaluating correlations and creating derived metrics such as mileage per year and fuel efficiency.
- **Louay:** Responsible for **model evaluation and performance tracking**, implementing drift detection and visualizing key metrics through the monitoring dashboard.
- **Mariana:** Coordinates **documentation and visualization**, maintaining project records, producing dashboards and charts, and consolidating final materials for presentation.

All members will participate in each stage of the pipeline, from data preprocessing to monitoring to ensure shared understanding and consistent results.

07. References

AutoTrader UK. (2024). *Retail price index & market report.* Retrieved from https://www.autotrader.co.uk/market-insights/retail-price-index

CarsGuide. (2023). *Selling a used car privately vs through a dealer.* Retrieved from https://www.carsguide.com.au/car-advice/sell-your-car-privately-vs-dealer

Desai, A. (2019). *100,000 UK used car dataset (Ford and Mercedes).* Kaggle. Retrieved from https://www.kaggle.com/datasets/adityadesai13/used-car-dataset-ford-and-mercedes

Google Cloud. (2023). *Application deployment and testing strategies.* Retrieved from https://cloud.google.com/architecture

McKinsey & Company. (2022). *Automotive retail: The future of car buying.* Retrieved from https://www.mckinsey.com/industries/automotive-and-assembly

Muruaga, X. (2025). *IE MLOps lecture slides – Sessions 1–7.* IE University.

Statista. (2024). *Value of the used car market in the United Kingdom (2015–2023).* Retrieved from https://www.statista.com/statistics/308753