**Capstone Project Presentation**

**Title: Predicting Used Car Prices with Machine Learning: A Data-Driven Solution for Smarter Automotive Decisions**

**1. Background and Context**

Over the past few years, the used car market has grown tremendously, both in size and in the complexity of buyer-seller dynamics. With thousands of vehicle listings online, determining the fair price of a car has become a pressing challenge. Pricing isn’t straightforward—it depends on many interconnected factors such as the car's brand, model, mileage, engine size, transmission type, fuel type, number of previous owners, and more.

This capstone project draws from a rich dataset of 10,000 vehicle listings, each representing a unique combination of these attributes. The data serves as a foundation to explore patterns in car pricing and build an intelligent prediction model that could help buyers and sellers make more informed decisions.

**2. Problem Statement**

The used car market suffers from inconsistent pricing practices. Sellers often rely on guesswork or outdated valuation tools, leading to underpricing (causing revenue loss) or overpricing (which pushes potential buyers away). This lack of standardization in pricing introduces inefficiency, mistrust, and poor decision-making in the automotive resale industry.

How can we bring consistency, fairness, and data-backed intelligence to the way used cars are priced?

**3. Why This Problem Matters**

Accurate price estimation isn’t just a technical challenge—it’s a business need. Dealerships want to maximize profit without losing customers. Buyers want to avoid overpaying. Online marketplaces want to reduce negotiation time and improve user experience. If pricing can be predicted more precisely, all parties stand to benefit.

This project, therefore, doesn’t just stop at technical curiosity—it has real-world implications. It provides a foundation for intelligent pricing tools that can improve revenue, boost trust, and optimize the customer experience.

**4. Objectives of the Project**

* **Understand** the key variables that influence car prices.
* **Analyze** historical data to uncover pricing patterns.
* **Build** a machine learning model to predict car prices based on features like mileage, engine size, brand, and fuel type.
* **Evaluate** which features contribute most to price prediction.
* **Present** a solution that can help sellers set fair prices and buyers make confident purchases.

**5. Data Source and Description**

The data used for this project was collected from public online car marketplaces. It contains detailed information on **10,000 car listings**, and includes the following key attributes:

* **Brand and Model**
* **Year of manufacture**
* **Engine size (in liters)**
* **Transmission type** (manual, automatic, semi-auto)
* **Fuel type** (petrol, diesel, electric, hybrid)
* **Mileage**
* **Number of doors**
* **Number of previous owners**
* **Listed price** (which we are trying to predict)

Each entry represents a real car listing, offering us enough depth to identify trends and train a robust model.

**6. Data Challenges and Preprocessing**

As with most real-world datasets, the car price data required significant cleaning and preprocessing before modeling. Key challenges included:

* **Missing or inconsistent data**: Some records lacked engine size or had incomplete owner history.
* **Outliers and entry errors:** Unrealistically high or low prices and mileage values were detected, likely due to human error**.**

To prepare the data for modeling, we applied the following preprocessing steps:

* **Data cleaning**: Removed or corrected anomalies and outliers**.**
* **Categorical encoding:** Transformed categorical variables such as Brand, Fuel\_Type, and Transmission using appropriate encoding techniques.
* **Feature scaling**: Normalized continuous variables like Mileage and Engine\_Size to improve model performance**.**
* **Feature engineering:** Added new features such as**:**
  + **Car\_Age:** Calculated as the difference between the current year and the car’s manufacturing year.
  + **Mileage\_per\_Year:** Derived by dividing total mileage by car age, to better reflect usage intensity.

**7. Methodology**

**Step 1: Exploratory Data Analysis (EDA)**We began with a thorough EDA to uncover relationships between features and price. Visual tools included:

* Scatter plots – to assess how mileage impacts car price.
* Box plots – to compare price distributions across different brands.
* Correlation heatmaps – to identify strong relationships among numerical variables.

**Step 2: Model Development**The problem was framed as a supervised regression task, with Price as the target variable. We tested multiple models, including:

* Linear Regression – as a baseline.
* Random Forest Regressor – for its robustness to non-linearity and outliers.
* Gradient Boosting models – for optimized performance on structured data.

Model performance was evaluated using:

* Root Mean Square Error (RMSE) – to measure prediction error magnitude.
* R² Score – to assess the proportion of variance explained.

**Step 3: Model Interpretation**To enhance transparency, we examined feature importances. The most influential features were:

* Mileage
* Brand
* Engine Size
* Car Age

**8. Key Insights**

* **Mileage matters:** Lower-mileage cars consistently fetched higher prices, even when older.
* **Brand premium:** Well-known or luxury brands (e.g., BMW, Audi) commanded significantly higher resale values.
* **Transmission impact:** Automatic cars were generally priced higher than manual or semi-automatic ones.
* **Fuel innovation:** Electric and hybrid cars were less common in the dataset, but when present, had a distinct impact on pricing often reflecting newer models or perceived tech value.

**9. Limitations and Opportunities for Improvement**

* **Geographical data** was not included, yet location plays a huge role in pricing—future models can integrate city or country data.
* **Market trends** such as fuel prices and inflation were not factored in—these could enrich the model.
* Some **model assumptions** (like equal brand popularity) may not hold in all regions.

**10. Ethical Considerations**

We ensured the dataset does not contain any personal data—just car attributes. Since data was sourced from public listings, there are no privacy violations. However, for future deployment, we would need to ensure proper **GDPR** compliance and **consent mechanisms** for any live user data.

**11. Future Work and Practical Applications**

With the model developed, here’s how it could be used in the real world:

* Integrated into **dealership systems** to offer real-time car pricing.
* Embedded in **mobile apps** for private sellers to estimate fair value before listing.
* Used by **insurance companies** to set accurate premiums.
* Applied in **online marketplaces** to recommend price ranges based on features.

Next steps:

* Tune models further with cross-validation.
* Incorporate more external data.
* Develop a dashboard or lightweight API for deployment.

**12. Conclusion – So, What’s the Big Deal?**

In a world where data drives decisions, car pricing should no longer be left to guesswork. This project brings **structure, fairness, and intelligence** to how used cars are valued. With the right tools in place, sellers get the best deal, buyers feel confident, and platforms operate more efficiently.

By applying machine learning to real-world automotive data, we’re not just predicting car prices—we’re reshaping how people interact with the used car market.