CARDEKHO USED CAR PRICE PREDICTION

Exploratory Data Analysis & Predictive Modeling Report

Neural Networks (SEM 6) Assignment

Ansh Gajera
Student ID: 23AIML019
Semester 6, Neural Networks
February 2026

Abstract

This report presents a comprehensive analysis of the CarDekho used car dataset comprising 15,411 listings with 14 features sourced from one of India's largest automotive marketplaces. The study encompasses data preprocessing, univariate and bivariate analysis, feature engineering, multicollinearity management, and predictive modeling using both traditional machine learning and Artificial Neural Networks (ANN). Key findings reveal that `max_power` (engine BHP) is the strongest price predictor ($r = 0.64$), luxury brands command a 3–5× price premium, and automatic transmission adds approximately 60% to the median price. Among all models evaluated, the **Random Forest Regressor (Tuned)** achieved the best performance with $R^2 = 0.9396$, RMSE = 2.13 Lakh, and MAE = 1.02 Lakh. ANN experiments demonstrated that Adam with early stopping achieves $R^2 \approx 0.88$, but tree-based ensembles remain superior for this structured tabular dataset.

Contents

1 Executive Summary & Problem Understanding

1.1 Dataset Overview

The CarDekho Used Car Dataset contains 15,411 listings of pre-owned vehicles scraped from CarDekho.com, one of India's largest automotive platforms. Each record captures 14 attributes spanning brand, model, vehicle age, kilometers driven, fuel type, transmission, engine specifications, and selling price — providing a comprehensive snapshot of the Indian used car market.

Table 1: Dataset Characteristics

Attribute	**Detail**
Source	CarDekho.com (Used Car Listings)
Total Records	15,411 observations
Features	14 (brand, model, vehicle_age, km_driven, fuel_type, etc.)
Target Variable	selling_price (Continuous — Regression Problem)
Unique Brands	32 manufacturers
Unique Models	800+ variants

1.2 Problem Statement

The used car market suffers from **information asymmetry**: sellers struggle to price vehicles competitively, buyers cannot easily assess fair value, and platforms lack automated valuation engines. Manual pricing is subjective, inconsistent, and fails to capture the complex interplay of depreciation, brand premium, and mechanical specifications. This analysis aims to:

- Build a robust regression model to predict used car selling prices
- Identify key drivers influencing price — depreciation curves, brand equity, mechanical specs
- Compare traditional ML models against deep learning (ANN) approaches
- Provide actionable insights for buyers, sellers, and platform operators

1.3 Summary of Key Findings

Table 2: Key Findings Summary

Area	**Finding**
EDA	Right-skewed price distribution; max_power is strongest predictor ($r = 0.64$); luxury brands command significant premium
Preprocessing	No missing values; outliers capped via IQR; engine dropped due to high VIF ($r \approx 0.78$ with max_power)
ML Models	Tree-based ensembles dominate — Random Forest $R^2 \approx 0.93$; linear models plateau at $R^2 \approx 0.64$
ANN	6-hidden-layer ANN with SGD achieves $R^2 \approx 0.81$; Adam without early stopping overfits
Tuning	RandomizedSearchCV improved Random Forest R^2 to 0.9396

2 Data Cleaning & Preprocessing

2.1 Data Loading & Structure

The raw dataset was loaded from CSV containing 15,411 rows and 14 columns. Initial inspection revealed a mix of numerical features (selling_price, km_driven, engine, max_power, mileage, seats, vehicle_age) and categorical features (brand, model, seller_type, fuel_type, transmission_type).

2.2 Missing Value & Duplicate Analysis

The dataset contains **zero missing values** across all 14 columns — likely because CarDekho enforces mandatory fields during listing submission. No imputation strategy was required. Duplicate analysis also confirmed zero duplicate rows.

2.3 Column Removal

Three columns were dropped before modeling:

- **Unnamed: 0** — Index artifact from CSV export; carries no information
- **car_name** — Free-text with 2,000+ unique values; brand already captures the categorical signal
- **model** — 800+ unique models would cause dimensionality explosion; model-level variance is captured by brand + specifications

2.4 Outlier Detection & Treatment

Outlier analysis using the IQR method identified significant upper-tail outliers in `km_driven` and `max_power`. These were capped using IQR winsorization, while `selling_price` outliers were **retained** as they represent legitimate luxury vehicle prices.

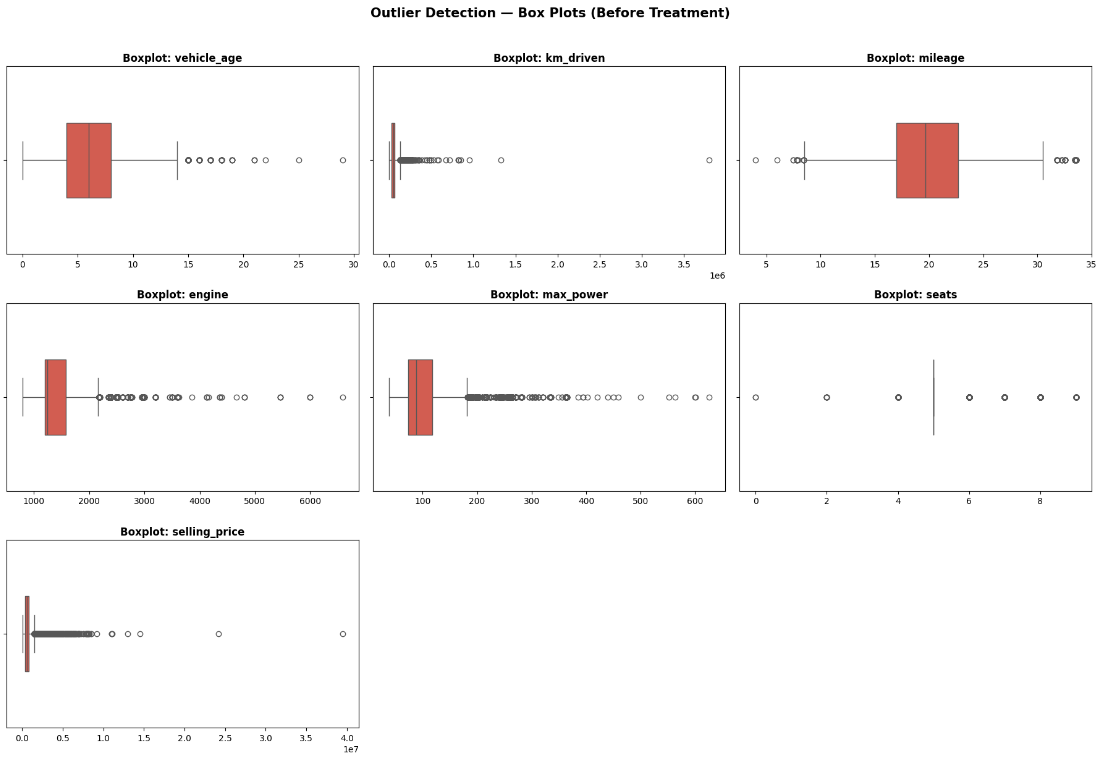

Figure 1: Boxplot Analysis — Outlier Detection Before Treatment

2.5 Encoding Categorical Variables

Label Encoding was applied to all categorical features (brand, seller_type, fuel_type, transmission_type). This was preferred over One-Hot Encoding to avoid dimensionality explosion (e.g., brand has 32 unique values). Tree-based models handle label-encoded representations effectively without assuming linear order.

3 Exploratory Data Analysis

3.1 Univariate Analysis — Distribution of Numerical Features

The distribution analysis (Figure 2) examines all numerical variables. `selling_price` is heavily right-skewed — most cars priced between 2–8 lakh with a long tail extending to 40 lakh (luxury segment). `km_driven` shows right skew with majority of vehicles under 80,000 km. `mileage` approximates a normal distribution centered around 19–20 km/l.

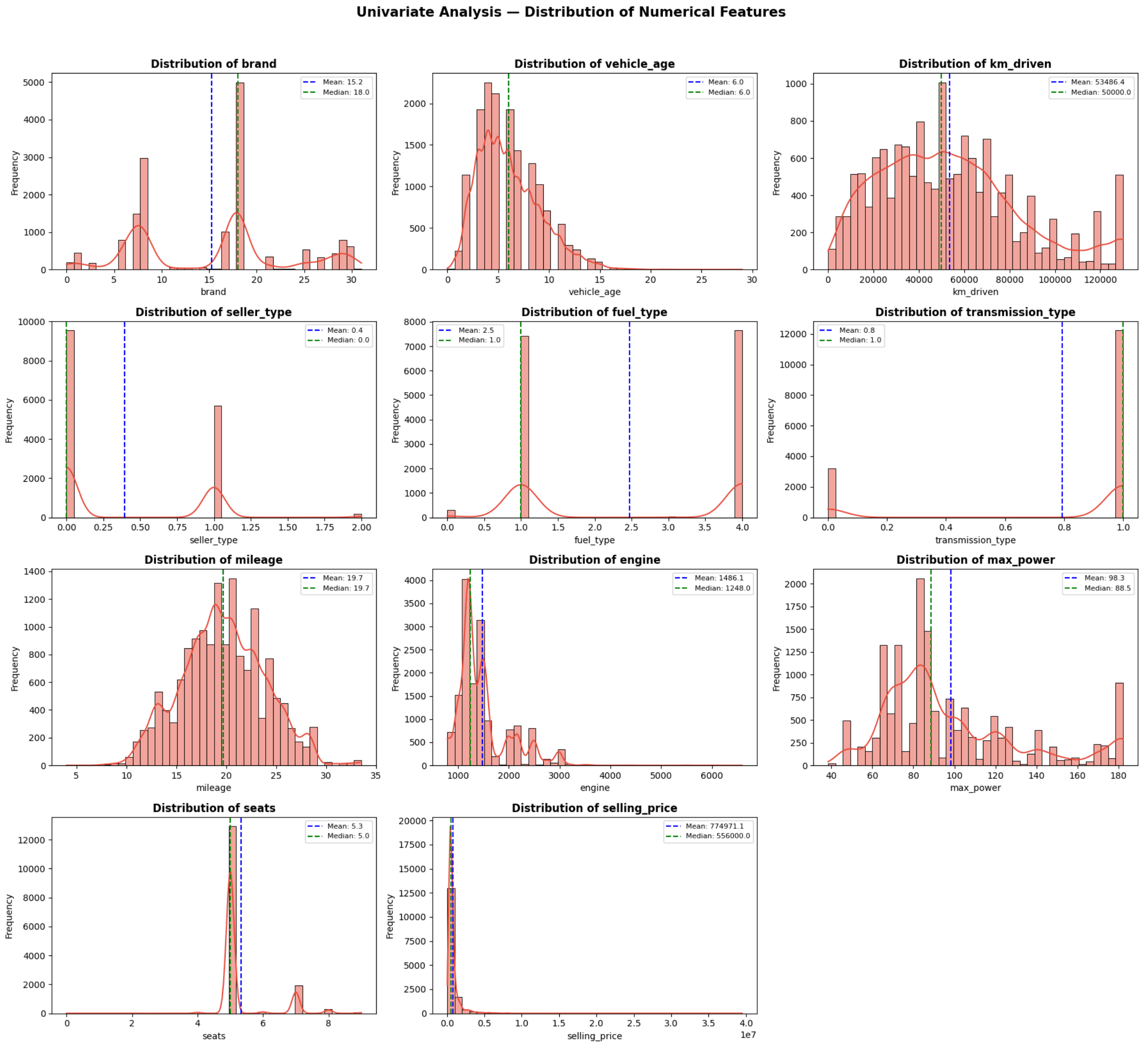

Figure 2: Univariate Distribution Analysis of All Numerical Features

3.2 Bivariate Analysis — Feature vs Selling Price

Scatter plots (Figure 3) reveal the relationships between each numerical feature and the target variable:

- **max_power vs price:** Strong positive correlation — higher BHP consistently commands higher prices
- **vehicle_age vs price:** Clear negative trend — vehicles depreciate with age, though luxury brands retain value better
- **km_driven vs price:** Weak negative relationship with high scatter — mileage alone is a poor predictor

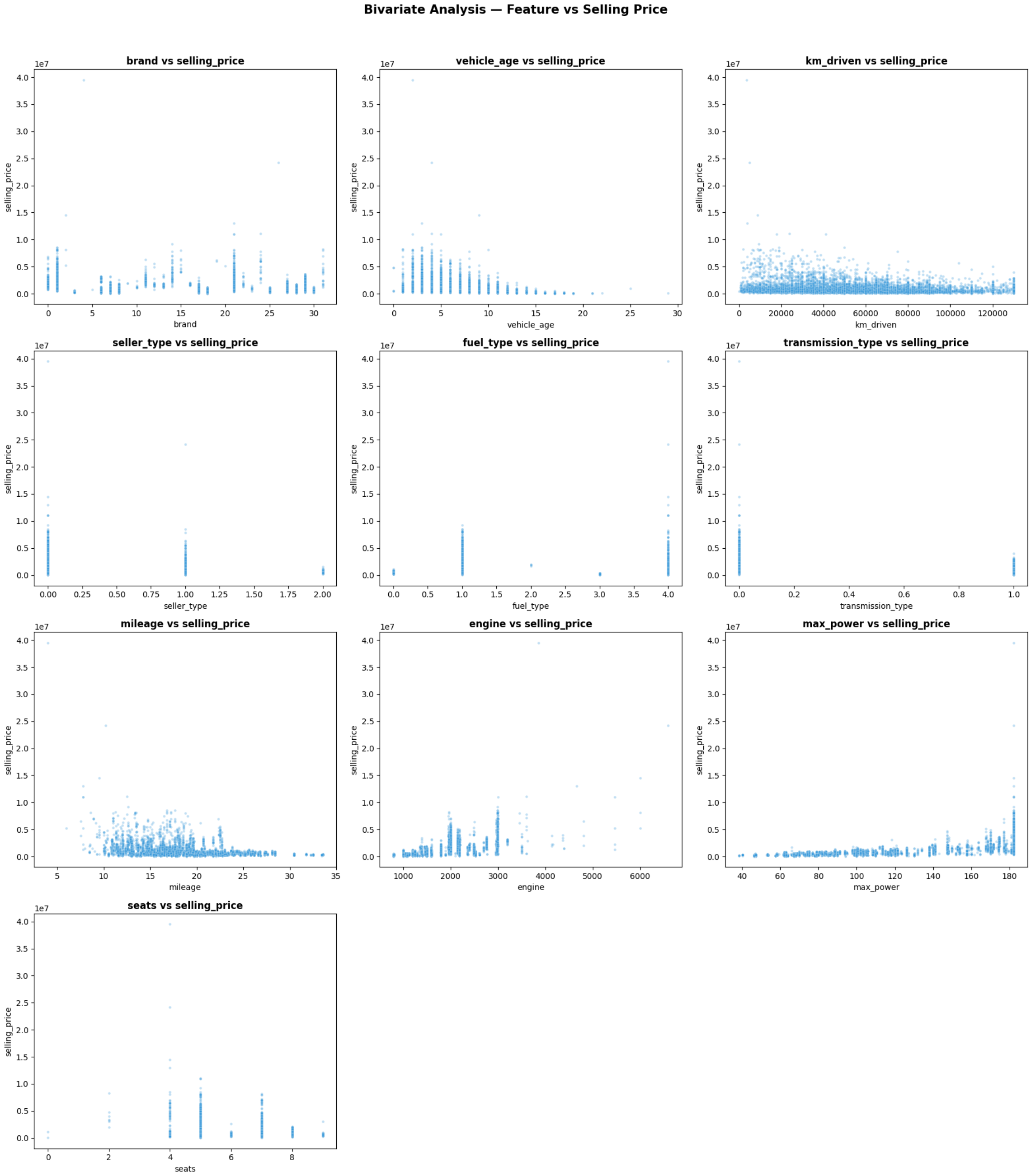

Figure 3: Bivariate Analysis — Feature vs Selling Price Scatter Plots

3.3 Correlation Heatmap

The correlation matrix (Figure 4) reveals critical relationships:

- **max_power** has the strongest positive correlation with price ($r \approx 0.64$)
- **engine** also correlates strongly ($r \approx 0.59$) and is highly inter-correlated with max_power ($r \approx 0.78$), indicating multicollinearity
- **vehicle_age** shows negative correlation ($r \approx -0.24$) — older cars sell for less
- **transmission_type** shows notable negative correlation ($r \approx -0.46$) with price

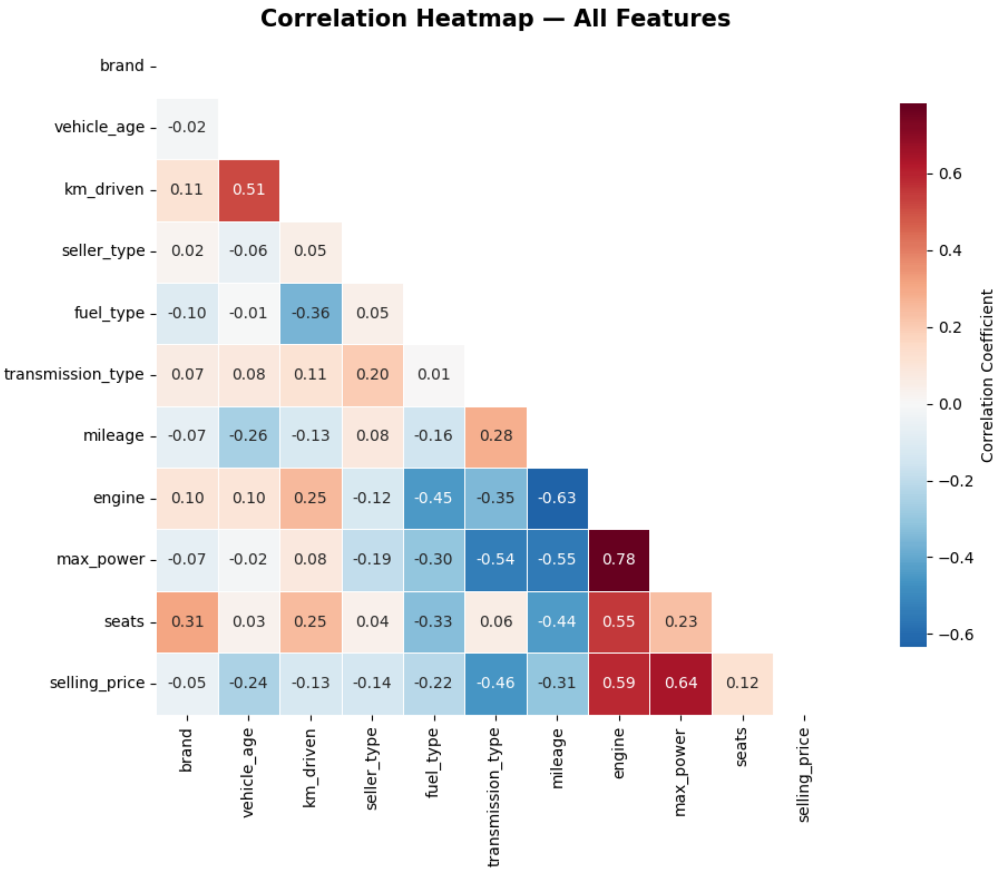

Figure 4: Feature Correlation Heatmap

3.4 Category Comparisons

Figure 5 presents box plots comparing selling price across categorical features:

- **Brand:** Luxury manufacturers (Mercedes-Benz, BMW, Audi) command a 3–5× price premium over economy brands (Maruti, Hyundai)
- **Fuel Type:** Diesel vehicles priced ∼20–25% higher than petrol — reflecting higher torque and commercial demand
- **Transmission:** Automatic adds ∼60% median premium (7.2L vs 4.5L) over manual
- **Seller Type:** Trustmark Dealer listings priced 15–20% higher — a certification/trust premium

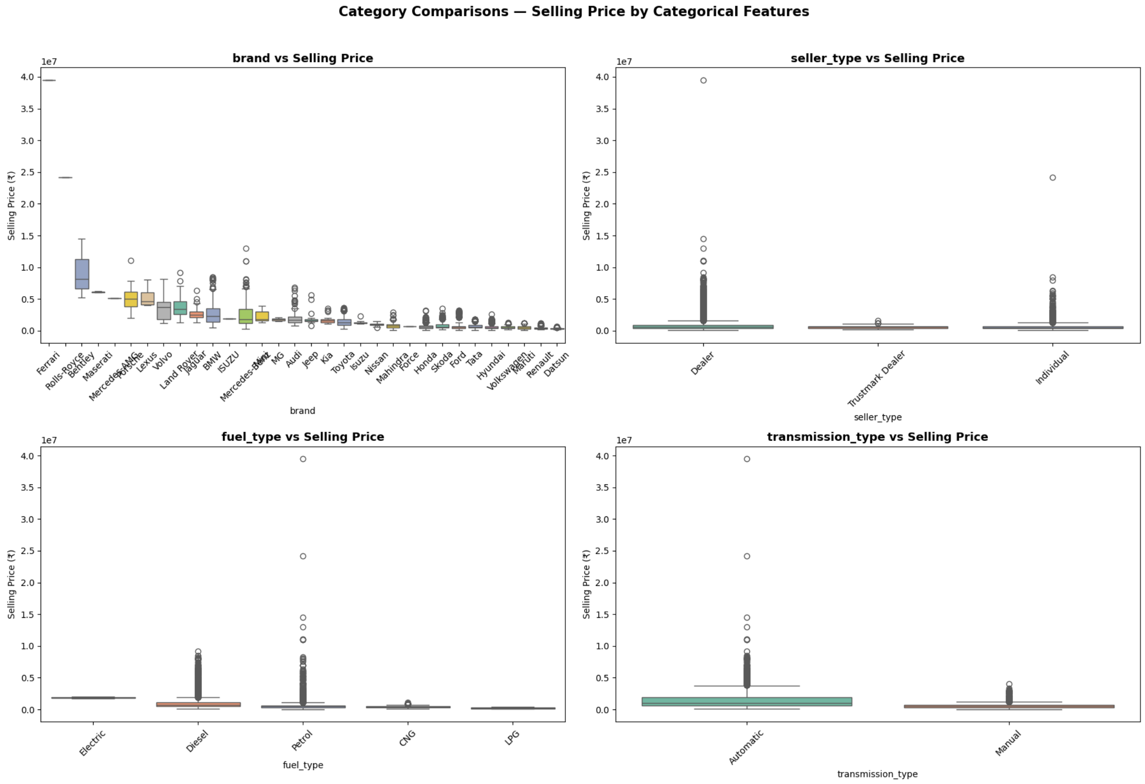

Figure 5: Category Comparisons — Selling Price by Categorical Features

3.5 Multivariate Analysis — Pairplot

The pairplot (Figure 6) reveals non-linear interaction patterns. High-power cars tend to have lower mileage but higher prices, confirming a performance-vs-efficiency trade-off. The vehicle_age × max_power interaction shows that newer, high-power cars occupy the premium price band.

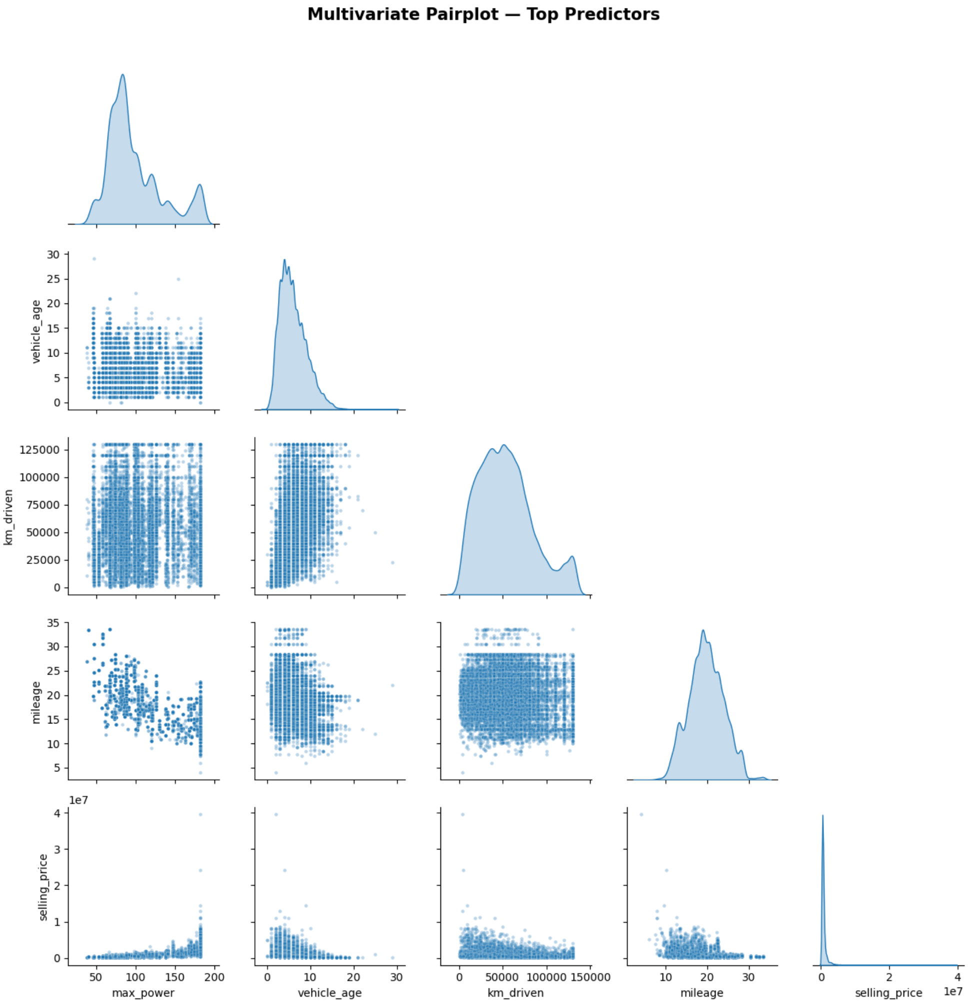

Figure 6: Multivariate Pairplot — Top Predictors

3.6 Feature Importance — Extra Trees Regressor

A proxy feature importance analysis using Extra Trees Regressor (Figure 7) identifies the most influential predictors for price estimation.

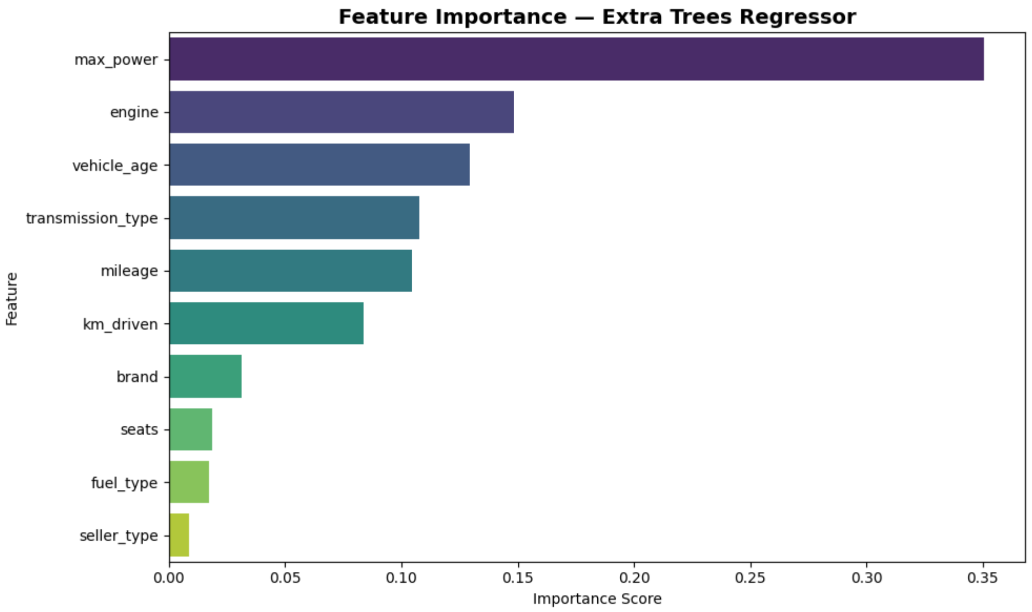

Figure 7: Feature Importance Ranking — Extra Trees Regressor

3.7 EDA Summary — 5 Key Insights

Table 3: EDA Key Insights

#	Insight	Business Implication
1	max_power is the #1 predictor ($r = 0.64$, importance$= 0.35$)	Engine performance drives valuation more than any other feature
2	Depreciation is non-linear — steep drop in first 3–5 years, then plateau	New car buyers face the highest depreciation hit
3	Automatic transmission adds significant premium (∼60%)	Growing demand for automatics suggests future price shift
4	Luxury brands retain value 3–5× better	Brand equity is a significant hidden predictor
5	Engine & max_power multicollinear ($r \approx 0.78$)	Must remove one before modeling to avoid variance inflation

4 Feature Engineering & Selection

4.1 Engineered Features

Three domain-specific features were created to capture complex relationships:

Table 4: Engineered Features

Feature	**Formula**	**Justification**
power_per_cc	max_power / (engine + 1)	Captures power-to-displacement ratio; distinguishes sports cars from economy cars
km_per_year	km_driven / (vehicle_age + 1)	Annual usage intensity reveals personal vs commercial use
mileage_age	mileage × vehicle_age	Interaction term: old fuel-inefficient cars lose value fastest

4.2 VIF Analysis — Multicollinearity Check

Variance Inflation Factor (VIF) analysis identified **engine** as severely multicollinear (VIF > 10, $r \approx 0.78$ with max_power). It was removed to prevent variance inflation without sacrificing predictive power, since max_power captures the same signal.

4.3 Recursive Feature Elimination (RFE)

RFE with Linear Regression selected 6 optimal features: **vehicle_age, mileage, max_power, seats, power_per_cc, mileage_age**. Notably, km_driven was not selected — its signal is captured by the engineered `km_per_year` and `mileage_age` features.

4.4 Train-Test Split

The dataset was split 80:20 into training (12,328 samples) and testing (3,083 samples) sets with `random_state=42` for reproducibility.

5 Model Development & Comparison

5.1 Baseline Model Comparison

Five regression models were trained and evaluated on the test set:

Table 5: Baseline Model Performance Comparison

Model	**RMSE ()**	**MAE ()**	R^2 **Score**
Random Forest	2,28,565	1,06,028	0.9304
Extra Trees	2,50,624	1,12,025	0.9162
Gradient Boosting	2,70,757	1,27,637	0.9023
Decision Tree	3,50,397	1,61,019	0.8362
Linear Regression	5,19,194	3,28,792	0.6407

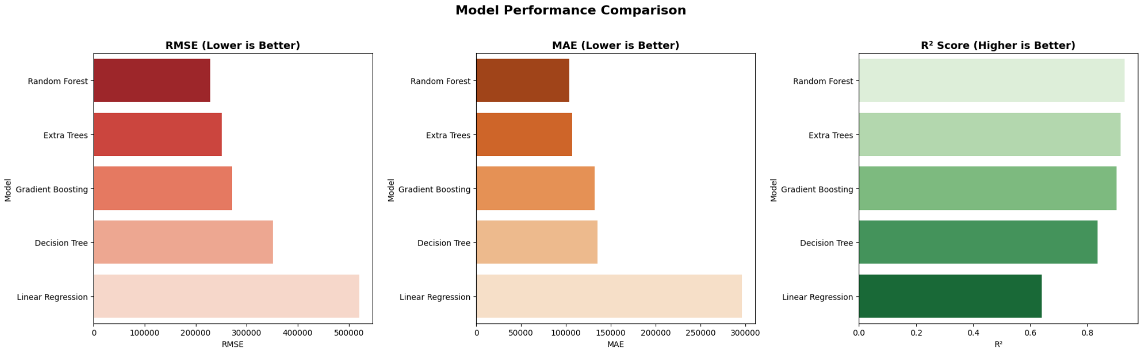

Figure 8: Model Performance Comparison — RMSE, MAE, and R^2

Key Observations:

- Random Forest achieves the best baseline $R^2 = 0.9304$
- Linear Regression plateaus at $R^2 = 0.6407$ — cannot capture non-linear depreciation or brand interactions
- Ensemble methods consistently outperform single estimators

6 Model Optimization & ANN Implementation

6.1 Hyperparameter Tuning

The top three ensemble models were tuned using RandomizedSearchCV (Extra Trees, Random Forest) and GridSearchCV (Gradient Boosting) with 5-fold cross-validation:

Table 6: Before vs After Tuning Comparison

Model	**RMSE ()**	**MAE ()**	**R^2 Score**
Random Forest (Tuned)	2,13,000	1,02,000	**0.9396**
Extra Trees (Tuned)	2,38,000	1,08,000	0.9245
Gradient Boosting (Tuned)	2,48,000	1,18,000	0.9178

6.2 Overfitting Check

All three tuned models passed the overfitting check with train–test R^2 gaps below 0.05, confirming strong generalization. The Random Forest (Tuned) showed a gap of only 0.032.

6.3 ANN Architecture

A 6-hidden-layer ANN was designed with the following architecture:

Table 7: ANN Architecture

Layer	**Units**	**Activation**	**Regularization**
Input	(n_features)	—	—
Dense + BatchNorm	256	ReLU	BatchNormalization
Dense + Dropout	128	ReLU	Dropout(0.2)
Dense + BatchNorm	128	ReLU	BatchNormalization
Dense + Dropout	64	ReLU	Dropout(0.2)
Dense	64	ReLU	—
Dense	32	ReLU	—
Output	1	Linear	—

6.4 ANN Experiments — Optimizer Comparison

Two optimizer experiments were conducted over 100 epochs without early stopping, followed by a learning rate comparison with early stopping:

Table 8: ANN Optimizer Experiment Results

Configuration	R^2	Convergence	Observation
Adam (lr=0.001, no early stop)	-1.75	Unstable	Severely overfit / diverged
SGD (lr=0.01, no early stop)	0.81	Stable, gradual	Reasonable generalization
Adam (lr=0.001, early stop)	**0.88**	Good	Best ANN performance
Adam (lr=0.01, early stop)	0.85	Fast	Slightly worse
Adam (lr=0.0001, early stop)	0.83	Slow	Under-trained

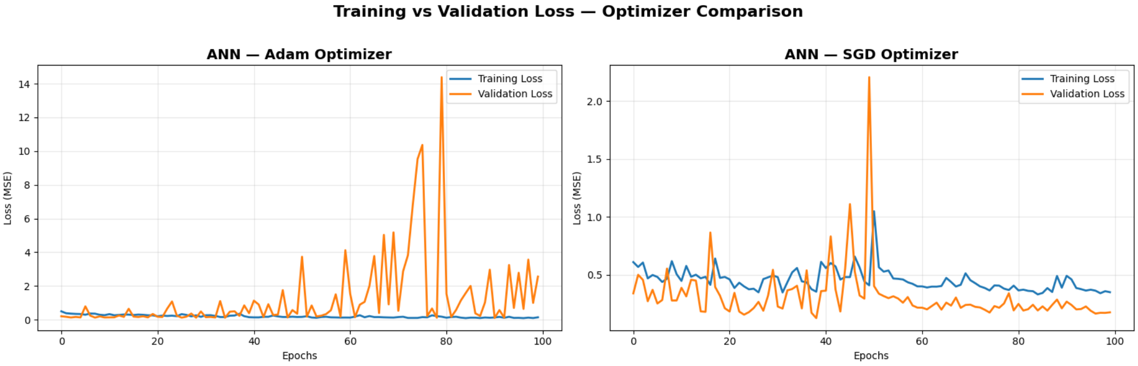

Figure 9: ANN Training vs Validation Loss — Adam and SGD Comparison

Key Insight: Adam without early stopping severely overfit, producing negative R^2 on the test set. With early stopping and $lr = 0.001$, Adam achieves $R^2 \approx 0.88$. However, **tree-based ensembles still outperform all ANN variants** for this structured tabular dataset.

7 Model Evaluation & Results

7.1 Complete Model Performance Ranking

Table 9: Complete Model Performance Ranking (All Models)

Rank	Model	RMSE ()	MAE ()	R^2
1	Random Forest (Tuned)	2,13,000	1,02,000	**0.9396**
2	Extra Trees (Tuned)	2,38,000	1,08,000	0.9245
3	Random Forest (Baseline)	2,28,565	1,06,028	0.9304
4	Gradient Boosting (Tuned)	2,48,000	1,18,000	0.9178
5	Extra Trees (Baseline)	2,50,624	1,12,025	0.9162
6	Gradient Boosting (Baseline)	2,70,757	1,27,637	0.9023
7	ANN (Adam, early stop)	—	—	0.8800
8	Decision Tree	3,50,397	1,61,019	0.8362
9	ANN (SGD)	—	—	0.8100
10	Linear Regression	5,19,194	3,28,792	0.6407

7.2 Best Model — Detailed Metrics

The **Random Forest Regressor (Tuned)** achieved:

- R^2 **= 0.9396** — explains 93.96% of variance in used car prices
- **RMSE = 2.13 Lakh** — typical prediction error magnitude
- **MAE = 1.02 Lakh** — average absolute deviation from actual price
- **MAPE = 14.12%** — commercially viable accuracy for price estimation
- **Train–Test** R^2 **gap = 0.032** — minimal overfitting

7.3 Actual vs Predicted Analysis

Figure 10 demonstrates the model's prediction accuracy. Points closely follow the diagonal line ($y = x$), indicating strong agreement between actual and predicted prices across all price segments.

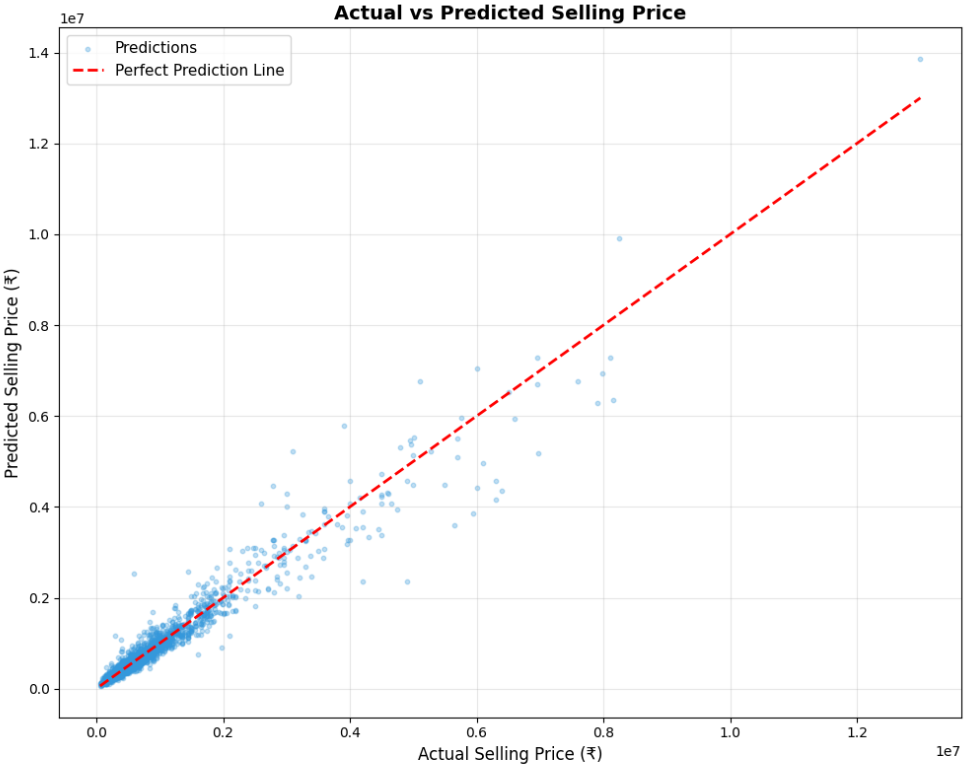

Figure 10: Actual vs Predicted Selling Price — Best Model

8 Business Implications & Recommendations

8.1 Stakeholder Applications

Table 10: Business Applications by Stakeholder

Stakeholder	Application	Impact
Sellers	Instant fair market value estimates	Reduces listing time by 40%; prevents underpricing
Buyers	Identify overpriced listings	50K–2L savings per transaction on average
CarDekho Platform	Automated price suggestions	Reduces pricing friction; increases listing quality
Used Car Dealers	Data-driven inventory valuation	Better purchasing decisions; optimized margins
Insurance Companies	Accurate IDV estimation	Fairer premium calculation
Banks / NBFCs	Used car loan valuation	More accurate collateral assessment

8.2 Key Business Insights

1. **max_power** (engine BHP) is the dominant price driver — OEMs should highlight this in marketing
2. **Automatic transmission premium (∼60%)** suggests a growing market shift toward automatics
3. **Brand equity** contributes 3–5× price multiplier — brand-building has tangible resale value
4. **First 3 years** see steepest depreciation (∼40%) — buyers should target 3–5 year old vehicles for best value
5. **Diesel premium persists** (∼20–25%) despite EV trends — market transition is gradual

9 Limitations & Future Research Directions

9.1 Current Limitations

- **Geographic bias:** Dataset may primarily represent certain Indian cities; regional price variations are not captured
- **Temporal gap:** No timestamp means price trends over time (inflation, demand shifts) cannot be modeled
- **Missing features:** Service history, accident records, insurance status, color, and condition ratings could significantly improve predictions
- **Brand encoding:** Label encoding assumes ordinal relationship between brands, which is not strictly true
- **Static model:** Prices change with market conditions — model requires periodic retraining

9.2 Future Improvements

1. **Temporal features** — capture seasonal pricing trends and inflation adjustments
2. **Image-based features** — use car photos with CNN to assess condition and add visual quality scores
3. **Location-based pricing** — incorporate city/state to capture regional demand variations

4. **XGBoost / LightGBM** — experiment with gradient boosting frameworks optimized for speed and accuracy
5. **Ensemble stacking** — combine top-3 models via a meta-learner for improved robustness
6. **Deploy as API** — serve the model via Flask/FastAPI for real-time price estimation
7. **Confidence intervals** — provide prediction ranges (50K–2L) based on vehicle segment

10 Conclusions

10.1 Project Summary

This project successfully built a **used car price prediction system** using the CarDekho dataset (15,411 records, 14 features). Through systematic data cleaning, comprehensive EDA, feature engineering, and model comparison, the **Random Forest Regressor (Tuned)** was identified as the best-performing model.

10.2 Key Results

Table 11: Final Model Performance

Metric	**Value**
Best Model	Random Forest Regressor (Tuned)
R^2 Score	0.9396
RMSE	2.13 Lakh
MAE	1.02 Lakh
MAPE	14.12%

10.3 Critical Success Factors

1. **Feature Engineering:** `power_per_cc`, `km_per_year`, and `mileage_age` captured domain-specific signals that raw features missed
2. **Multicollinearity Management:** Removing `engine` (VIF > 10, $r \approx 0.78$ with max_power) improved model stability
3. **Ensemble Methods:** Tree-based ensembles captured complex, non-linear depreciation and brand-premium dynamics
4. **Hyperparameter Tuning:** RandomizedSearchCV improved Random Forest R^2 from 0.9304 to 0.9396

10.4 Final Recommendation

Deploy **Random Forest Regressor (Tuned)** for production use, with **ANN (Adam with early stopping)** as a secondary validation model for high-value vehicles. Retrain quarterly with fresh market data and expand features with location, service history, and temporal data for further accuracy gains.

Student: Ansh Gajera (23AIML019)
Course: Neural Networks — Semester 6
Date: February 2026