Final Report: Predicting Used Car Selling Prices

Overview

This project aims to build a predictive model to estimate the selling prices of used cars using the provided dataset (car.csv). The dataset contains 8,128 entries with features such as car age, kilometers driven, fuel type, seller type, transmission, owner history, mileage, engine capacity, max power, and seats.

Data Understanding and Cleaning

- The target variable is selling_price, which ranges from approximately 30,000 to 10,000,000 NPR.
- The dataset features a wide price range and some columns with missing or inconsistent values.
- Features like fuel, seller_type, transmission, and owner were encoded numerically using label encoding.
- Continuous variables were checked for skewness, and the target variable was log-transformed (log1p) to reduce the effect of extreme values.
- Missing values were imputed using median or mean based on skewness analysis.
- Data was then scaled using StandardScaler for model training.

Exploratory Data Analysis (EDA)

- Visualizations such as scatter plots between features and selling price, box plots to detect outliers, and correlation matrices were used to understand relationships.
- High variance and wide distribution of the selling price suggested that log transformation would improve model performance.

Model Building

- Several regression models were trained and evaluated:
  - Linear Regression
  - Decision Tree Regressor
  - Random Forest Regressor
  - Neural Networks with varying layer sizes (30, 60, 90 neurons)
- Models were evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2).
- Log transformation of the target was applied before training; predictions were inverse-transformed before evaluation.

Results

Model	MSE (NPR^2)	MAE (NPR)	R^2 Score
Linear Regression	1.04×10^{11}	149,868	0.7674
Decision Tree	3.50×10^{10}	92,468	0.9440
Random Forest	2.71×10^{10}	78,439	0.9543
Neural Network (30)	6.19×10^{10}	120,721	0.886
Neural Network (60)	5.94×10^{10}	121,167	0.919
Neural Network (90)	4.94×10^{10}	113,368	0.909

- The Random Forest model performed best overall, with the lowest MSE and MAE and highest R^2.
- Neural networks showed competitive results but slightly higher errors compared to tree-based models.
- Linear Regression had the poorest performance due to the non-linear relationships in data.

Use of Large Language Models (LLMs)

- LLMs like ChatGPT were used for:
  - Understanding feature meanings and suggesting data cleaning methods.
  - Brainstorming feature engineering ideas (e.g., creating age from year, transforming skewed features).
  - Guidance on model selection, evaluation metrics, and code snippets for preprocessing.
  - Assistance in writing and formatting code and reports.