Name: Nishchal Sharma

Github Repo: https://github.com/nishchalacharya/Car-Price-Predict-Model

Portfolio: https://nishchalportfolio.vercel.app/

Contact: 9867932976

PROJECT: Car Price Prediction Using Machine Learning

OBJECTIVE:

Predict the selling price of used cars using machine learning and exploratory data analysis. Used large language models (LLMs) like ChatGPT to support data understanding, feature creation/selection, modeling, and result interpretation.

DATA OVERVIEW:

- The raw car dataset contains 8128 rows and 12 columns.
- It contains many missing values in various columns so, they are filled as per their nature like skewness.
- The name column contains more than 2000 unique values, can't visualize so, only looked after 10 most famous name of car from existing datasets.
- Performed various EDA(uni-variate like hist plot , box-plot,count-plot,pie-chart for analysis) and multi variate tools like scatterplot,box_plot,heatmap to see correlation between different features.
- Removed outliers to help model learn more useful dataset only.

FEATURE ENGINEERING:

- Extract age of car ,power_per_cc and remove unnecessary columns like age.
- Extracted brand name only from name column and removed name column from datataset.
- Ordinal Encoding for 'owner' column and one-hot encoding is done for other other categorical dataset for model training.

DATA PREPARATION/SPLITTING :

- Splitting data into train and test datasets with proportion(80% training and 20% test datasets) with random_state value 42 .

MODEL TRAINING :

Since, it's a regression problem ,so we choose model on basis of that.

- Used Random Forest Regression and XGBoost Regression (ensemble models) for better prediction.
- Fine tuned on them using GridSearchCV to find best parameters for their respective models.
- Performed model training by without removing outliers and removing outliers in various column to check overall model performance.

MODEL EVALUATION :

- Used various evaluation metrics like root mean square error(rmse), mean absolute eror(mae), R_squared error (r^2).
- Also used Dummy Regressor with strategy 'mean' to check performance of model.

RESULTS:

Let's look after each value for all these models.

Model	RMSE	MAE	R2
Random Forest(Tuned)	65,871.70	46,050.00	0.9168
XGBoost(Tuned)(used)	63,106.68	45,442.35	0.9236 (high)
Dummy Regressor	228,365.65	191,795.67	-0.00037

- After evaluating those parameters, XG Boost model is chosen as right one for predicting.
- Also, used various tools like scatterplot to evaluate the target vs actual value.
- The above parameters shows quite good training of data (pretty good fit) .

TOOLS USED:

- Python, Numpy, Pandas, Matplotlib, Seaborn
- Scikit-Learn, XGBoost, LLM(Chatgpt),Vscode, Git/Github
- FastAPI for api- end points development.

USE OF LLMs (ChatGPT) for Contribution:

Preprocessing

- Used chatgpt to replace missing values for different columns like by their median, mean on different conditions like looking skewness of dataset and their nature.
- Chatgpt also asssist me to understand different unique values associated with various columns ,their meaning and visualizing them.

Data Visualization and Analsysis/EDA

- For EDA, mostly in columns like name, which have more than 2000 unique values ,the plotting and comparing it was extremely difficult to cover all unique data and find their relationship with other columns and fine insights ,so Chatgpt assist me for only find insights for 10 most common name and use them to visualize.
- It also helps me choosing right relationship between columns and which graph will help me finding proper insights from those datasets.

Feature Engineering

- LLM assist me in generating new useful features for training like car age,power_per cc and also remove some columns like 'year' .
- LLM also assist me in creating new column like brand and removing name columns,so analysis can be done more properly. The more unique values of name columns have no significance,so created new column with their brand representation.

Model Selection and Training

- Take help for finding good model as per my datasize and its behaviours (like present of outliers).
- I asked chatgpt for various parameters in GridSearchCV for fine tuning purpose and finding best parameters for model for good accuracy.

Model Evaluation

- I provided mean, median values for target columns and predict rmse,mse, r_sq to find best model,know how well it is performing.
- Asked assist from LLMs if I should only use most important columns to train dataset(by removing more) and see,but it doesnt work quite good.
- Assist in plotting datasets for finding actual vs predicted target values in scatterplot to see how good it is performing.
- I provided R_square values of Dummy Regressor to take idea of what it is representing and other values too.
- It helps me guaranting that my model is predicting right and don't predict as like just guessing So can be concluded that it has learnt properly.