ML System Design

Problem Formulation

Used car price evaluation

The developed models can be leveraged by anyone involved in selling or buying cars to facilitate informed pricing decisions. By embedding these models into second-hand car marketplaces, both sellers and buyers can benefit from enhanced pricing transparency and accuracy.

Benefits for Users:

1. Seller Benefits:
    - Optimized Pricing: Sellers can use the model predictions to set competitive and realistic prices based on market trends and vehicle attributes.
    - Improved Selling Experience: Enhanced pricing accuracy can attract more potential buyers and expedite the selling process.
2. Buyer Benefits:
    - Fair Pricing: Buyers can use model predictions to assess the fairness of listed prices and make informed purchasing decisions.
    - Enhanced User Experience: Access to accurate price estimates improves buyer confidence and satisfaction during the car-buying process.

Integration in Marketplaces:

> Embedding these models in second-hand car marketplaces enhances the overall user experience by providing reliable pricing guidance, thereby fostering trust and efficiency in transactions.

Can’t be used to evaluate antique cars, since metrics used won’t be relevant to their price or sometimes can negatively affect results.

Example - Kilometrage is not relevant for antique car price

Age positively affects price of antique car

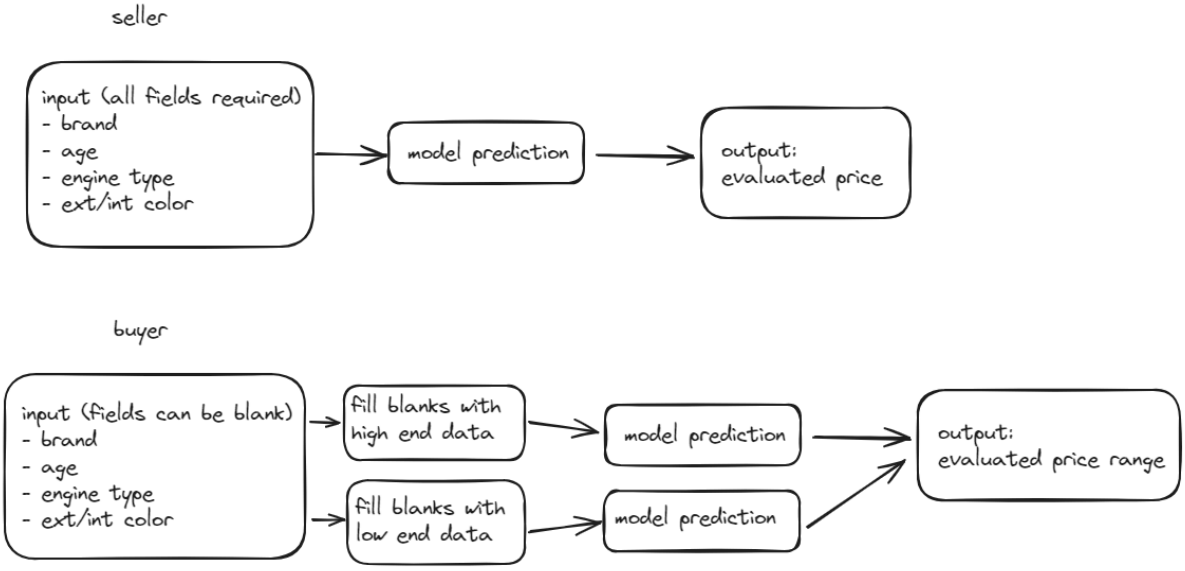
There are data sources available due to the fact that the used car market is big and records are saved well. If combining a couple of different data sources normalization needs to be done on metrics such as currency, mile/kilometer etc.

Model can evaluate used car prices well, since most of the deciding factors are straightforward.

Linear Regression will be used as baseline, since most of the features like kilometrage or age are linearly dependant on car price

Gradient boosting technique will be used as a comparison since Linear Regression struggles to handle non-linear relationships so features like brand and transmission type will be handled better

Architectural Components



For seller flow will be straightforward since they have to fill all the fields and model will evaluate price of the car

For buyers flow model will fill empty fields with low and high end data to determine range properly

Data Collection and Preparation

Will be using a dataset from kaggle, which represents the US market

For preprocessing following steps are done

Dropping Unnecessary Columns:
- The columns ext_col and int_col are dropped from the dataset as they are not relevant for the analysis.

Data Type Conversion:
- The model_year column is converted to an integer type for numerical analysis.

Milage Processing:
- The milage column, which originally contains values in the format "X mi", is cleaned to remove the "mi" suffix, leaving only the numerical value.

Engine Specifications:
- The engine column is split into two separate columns: HP (horsepower) and engine_volume for more granular analysis.

Transmission Type:
- The transmission column is categorized into two types: automatic and manual.

Price Processing:
- The price column, which originally contains values in the format "$X", is cleaned to remove the dollar sign, leaving only the numerical value.

Dropping Rare Brands:
- Brands with a count of fewer than 15 occurrences in the dataset are removed to focus on more common brands.

Handling Outliers:
- Outliers in each category are removed based on the range [mean - 2 * std, mean + 2 * std], where mean is the average value and std is the standard deviation of the category.

Categorical Data Conversion:
- Categorical data is converted into binary true/false (t/f) format for easier analysis and modeling.

.

Feature Engineering

In determining factors influencing car prices, emphasis is placed on selecting features with significant impact. For instance, variables like car color are excluded from consideration due to their negligible effect on pricing outcomes. This ensures that only influential factors are prioritized in the analysis and model development process.

Model Development and Offline Evaluation

Both Linear Regression and Gradient Boosting techniques will be employed to evaluate car prices. The performance of each model will be assessed rigorously to determine the superior performer for final implementation. This comparative analysis ensures the selection of the most effective model for accurate price prediction.

The dataset comprises 5,000 entries and encompasses 12 features. These features have been carefully selected as they are essential and sufficient for price prediction, mirroring the criteria commonly used by human experts in determining car prices.

For the evaluation of model performance, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) will be utilized. Using these metrics together ensures a comprehensive assessment of model performance, addressing both accuracy in prediction errors (MSE, RMSE) and overall model adequacy and fit (R2).

Following model training, a thorough examination will be conducted to identify areas where the model may struggle. This includes investigating whether certain features disproportionately influence results or if particular brands are consistently evaluated more accurately than others.

Areas of Focus:

1. Feature Influence Assessment:
    - Feature Contribution: Analysis will focus on understanding the impact of individual features on model predictions. Features that significantly skew results or contribute inconsistently will be identified for further scrutiny.
2. Brand Evaluation:
    - Brand Performance: The model's performance across different brands will be evaluated to ascertain if certain brands are predicted more accurately

than others. This analysis aims to uncover any biases or patterns that may affect overall predictive reliability.

Objective:

- The objective of this analysis is to refine the model by addressing any biases or inconsistencies observed during post-training assessment. Adjustments may include feature selection, fine-tuning model parameters, or applying techniques to mitigate brand-specific biases.