CAR PRICE PREDICTION

ALX1_AIS3_S1e IBM Data Scientists

TEAM MEMBERS

Mariam Naeim Nassim
Menna Mahmoud Abd EL-Rahman
Yomna Wael Muhammad

DEPI Final Project

Predicting Used Cars Prices Using Machine Learning

• Problem Statement:

The used car market is highly dynamic, with prices fluctuating based on various factors like brand, model, mileage, engine type, and accident history. Currently, buyers and sellers face difficulties in determining the fair market price for a used vehicle due to the sheer complexity and number of influencing factors. The lack of accurate price prediction models leads to inefficient pricing strategies, underpricing, or overpricing of cars, which ultimately hampers market transparency and trustThis project aims to build a predictive model that can accurately estimate the price of a used car based on its features. Such a model can be leveraged by car dealerships, individual sellers, and buyers to ensure fair transactions, improve customer satisfaction, and increase market efficiency.

• Objectives:

The primary objectives of this project are:

- To develop a machine learning model that predicts the price of a used car based on its features such as brand, model year, mileage, engine type, and accident history.
- To identify the key factors that have the greatest impact on used car prices.
- To improve the accuracy of price estimation over traditional pricing models or manual pricing strategies.

• Data Description:

The dataset used for this project contains information on used cars and their respective prices, gathered from an online car marketplace. It includes two sets: a training dataset with 188,533 records and a test dataset with 125,690 records. Both datasets contain 13 columns, which represent various features that can influence the price of a used car.

The following columns are included in the dataset:

- **id**: Unique identifier for each car.
- **brand**: The car's brand (e.g., Toyota, BMW).
- **model**: The car's specific model (e.g., Corolla, X5).
- **model_year**: The year the car model was manufactured.
- **mileage**: The total distance the car has traveled (in kilometers).
- **fuel_type**: Type of fuel used by the car (e.g., gasoline, diesel, electric).
  Missing values: 5,083 in the training data and 3,383 in the test data.
- **engine**: Engine size in liters.
- **transmission**: Type of transmission (e.g., automatic, manual).
- **ext_col**: Exterior color of the car.
- **int_col**: Interior color of the car.
- **accident**: Whether the car has a recorded accident history (Yes/No).
  Missing values: 2,452 in the training data and 1,632 in the test data.

- **clean_title**: Whether the car has a clean title (Yes/No).
  Missing values: 21,419 in the training data and 14,239 in the test data.
- **price**: The price of the car (present only in the training data).

The dataset contains missing values primarily in the **fuel_type**, **accident**, and **clean_title** columns, which will need to be addressed during the data preprocessing phase.

• Problem Definition:

- Build a model that accurately predicts the price of a car based on various features (e.g., brand, model, model_year, engine, Fuel_Type…).
- Identify the stakeholders and set clear evaluation metrics (e.g., Mean Squared Error, R-squared, Accuracy , F1Score) to measure the model's success.

• Data Preprocessing:

-Data Cleaning:

Handle missing values, outliers, and noisy data. Replace or impute missing values for continuous data (e.g., average mileage) or drop if unimportant.

-Feature Engineering:

Categorical Encoding: Convert categorical features (e.g., brand, model, fuel type) into numerical representations using techniques like One-Hot Encoding or Label Encoding.

-Feature Scaling:

Normalize or standardize numerical features (e.g., mileage, engine size) for better model performance.

-Feature Selection:

Select the most important features using techniques such as correlation analysis, variance thresholding, or more advanced methods like Random Forest Feature Importance or Linear Regression

-Exploratory Data Analysis (EDA):

Visualization: Use histograms, scatter plots, and box plots to visualize relationships between features and car prices.

Correlation Analysis: Identify which features are most correlated with the car price using correlation matrices heatmaps.

-Insights: For example, newer cars, lower mileage, and certain brands may be more valuable. EDA helps understand these relationships.

-**Model Selection:**

Choose a mix of regression models that are suitable for prediction problems:

Linear Regression: A simple baseline model that assumes a linear relationship between features and price.

Decision Trees: Captures non-linear relationships but may lead to overfitting.

Random Forest : Ensemble models that handle both linear and non-linear relationships effectively.

XGBoost: A powerful model known for winning many machine learning competitions.

- ## Tools and techniques:

- **Data Collection Tools**

  kaggle

- **Data Processing and Manipulation**

  Python: Widely used for handling and manipulating data.

  Pandas: Ideal for data wrangling, including loading, cleaning, and transforming car price datasets.

  NumPy: Used for numerical computations and handling large datasets efficiently.

- **Data Visualization**

  Matplotlib and **Seaborn**: Useful for visualizing relationships between features (e.g., car age vs. price, mileage vs. price).

  Plotly: For interactive visualizations that can be shared with stakeholders.

- **Machine Learning Libraries**

  Scikit-Learn: Essential for building models, performing feature scaling, and training algorithms like **Linear Regression**, **Decision Trees**, and **Random Forests**.

- ## TimeLine:

October 1 - October 12: Data Preprocessing

Handle missing values in key columns and normalize categorical variables.

October 12: Proposal Submission

Finalize and submit the project proposal outlining objectives and methodologies.

October 13 - October 16: Model Development and Evaluation

Develop and train machine learning models using the prepared dataset.

Evaluate model performance and fine-tune parameters.

October 17: Final Report Submission

Compile findings, insights, and model performance into a comprehensive report.

October 25: Final Presentation

Prepare and deliver a presentation summarizing the project outcomes and insights.

Challenges and Risks

- **Insufficient Data:**
  - Missing values in critical features may affect model accuracy.
- **Computational Resources:**
  - Limited access to high-performance computing resources could slow down model training.
- **Model Overfitting:**
  - Risk of the model performing well on training data but poorly on unseen data.

- **Mitigation Strategies:**
  - **For Insufficient Data:**
    Implement data imputation techniques and consider augmenting the dataset if necessary.
  - **For Computational Resources:**
    Utilize cloud computing services for scalable resources.
  - **For Model Overfitting:**
    Use cross-validation and regularization techniques to ensure the model generalizes well.

Conclusion

The project on predicting used car prices is crucial in addressing the complexities of the automotive market. By developing a machine learning model, we aim to provide accurate price estimations that enhance market transparency and trust. This initiative will ultimately contribute to more efficient pricing strategies, benefiting both buyers and sellers in the used car market.