Group Members:

Yate Zhang, Yuchen Bi, Chris Wang, Jiancong Zhu

Executive Summary

The automotive resale market presents significant challenges in accurately estimating the value of used vehicles. Variability in pricing arises due to factors such as mileage, brand perception, fuel efficiency, and model year. This project aims to address these challenges by developing a machine learning model capable of predicting the prices of second-hand vehicles based on various attributes. The dataset sourced from Kaggle consists of Craigslist listings of used cars, providing a diverse range of vehicles with details such as make, model, year of production, odometer reading, fuel type, and transmission. Our objective is to explore and refine this dataset, engineer meaningful features, and implement predictive models to generate reliable price estimates.

Following the CRISP-DM framework, the project begins with exploratory data analysis to identify patterns in vehicle pricing, understand depreciation trends, and detect potential data inconsistencies. The next phase involves feature engineering, where key attributes such as car age and mileage-to-price ratios are constructed to enhance model performance. Several machine learning models, including linear regression, decision trees, random forests, and XGBoost, will be evaluated to determine their efficacy in price prediction. Model performance will be assessed using root mean squared error (RMSE) and R-squared metrics. Finally, insights from our findings will be translated into actionable recommendations for car buyers and sellers, optimizing pricing strategies and improving transparency in the used car market.

Problem Statement / Research Objectives

Accurate price prediction for used cars is crucial for both buyers and sellers in the second-hand vehicle market. Buyers seek fair market prices to avoid overpaying, while sellers aim to set competitive yet profitable prices. However, pricing used cars is inherently complex due to the multitude of factors influencing value. Traditional valuation methods, such as dealership estimates or online appraisal tools, often fail to capture the full spectrum of variables affecting a vehicle's price. This project aims to bridge that gap by leveraging machine learning techniques to create a robust predictive model for used car pricing.

The primary objective of this study is to analyze the key determinants of vehicle pricing and develop an automated model that accurately predicts the market value of used cars based on historical Craigslist listings. Through exploratory data analysis, we will identify correlations between attributes such as mileage, brand reputation, model year, and fuel type with the listed price. Feature engineering techniques will be applied to construct meaningful input variables that improve model interpretability. Various machine learning algorithms will be implemented and compared to select the most effective model for price prediction. The ultimate goal is to provide an accurate and scalable pricing model that can assist individual buyers, sellers, and automotive dealerships in making data-driven pricing decisions.

Literature Review

The use of machine learning in price prediction has been extensively studied across various domains, including real estate, consumer electronics, and automobiles. Prior research indicates that regression-based models and tree-based algorithms are particularly effective for handling structured numerical data such as vehicle specifications and pricing. Studies have explored the impact of traditional features, including mileage and model year, alongside non-traditional factors such as consumer demand trends and geographic location.

One relevant study examines the role of feature selection in car price prediction, demonstrating that incorporating engineered features, such as car age and fuel efficiency, significantly improves model performance. The study compares linear regression, support vector machines, and ensemble methods, concluding that tree-based models like random forests and gradient boosting outperform simpler algorithms in handling non-linear relationships between vehicle attributes and price.

Another key area of research focuses on handling data inconsistencies and missing values in automobile pricing datasets. Studies highlight that outliers in odometer readings and price listings can skew model accuracy, necessitating preprocessing techniques such as median imputation and log transformations. Research on large-scale automotive datasets further suggests that model interpretability is crucial for industry adoption. Decision trees and SHAP (SHapley Additive exPlanations) values have been explored as tools for understanding how each feature contributes to price predictions.

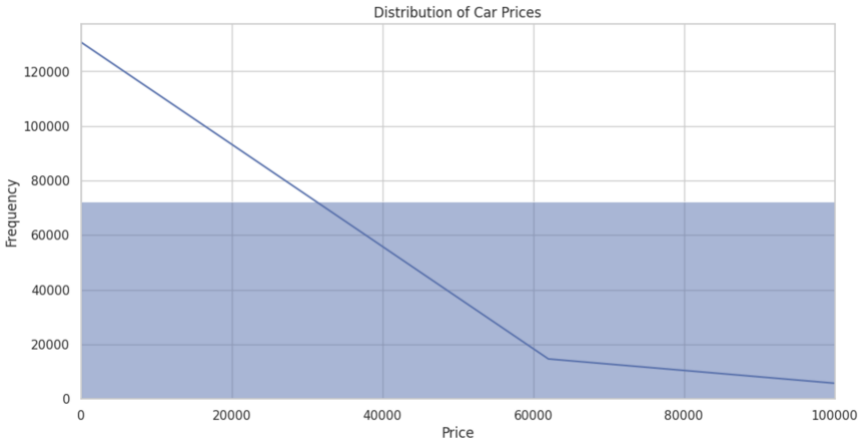
In the context of online listings, studies have analyzed the impact of textual descriptions and images on car prices, leveraging natural language processing (NLP) and computer vision. While these techniques provide additional insights, structured tabular data remains the most reliable basis for price prediction in large datasets like Craigslist vehicle listings.

Building on this body of research, our project applies machine learning techniques to the Craigslist Cars & Trucks dataset, ensuring that best practices in data preprocessing, feature engineering, and model evaluation are implemented. By comparing multiple modeling approaches and validating feature importance, our study aims to refine existing methodologies for used car price prediction and contribute to the growing body of machine learning applications in the automotive industry.

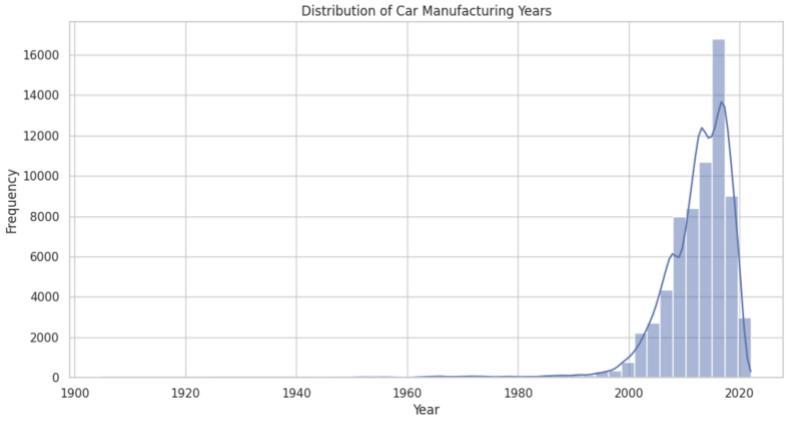
How We Deal with EDA

For the EDA, the first step involved loading the used car dataset on Kaggle and examining its structure. The dataset contained 426,880 rows and 26 columns, covering various attributes of used cars such as price, year, manufacturer, model, condition, fuel type, odometer readings, and transmission type. The initial inspection of the dataset using df.info() and df.describe() provided an overview of data types, numerical distributions, and missing values. The dataset included both numerical and categorical variables, requiring different strategies for handling missing data and feature engineering.

A significant part of the EDA focused on identifying and handling missing values. A missing values report showed that several columns contained a high percentage of missing values, including size, cylinders, and VIN, making them unsuitable for further analysis. Therefore, columns with more than 50% missing values were dropped to prevent excessive imputation errors. For numerical columns such as year and odometer, missing values were replaced with their respective median values to maintain consistency without being overly influenced by outliers. Categorical variables like manufacturer, model, fuel, and transmission were imputed using the most frequent category (mode). Additionally, paint_color, which had a moderate number of missing values, was filled with "unknown" to retain as much information as possible.

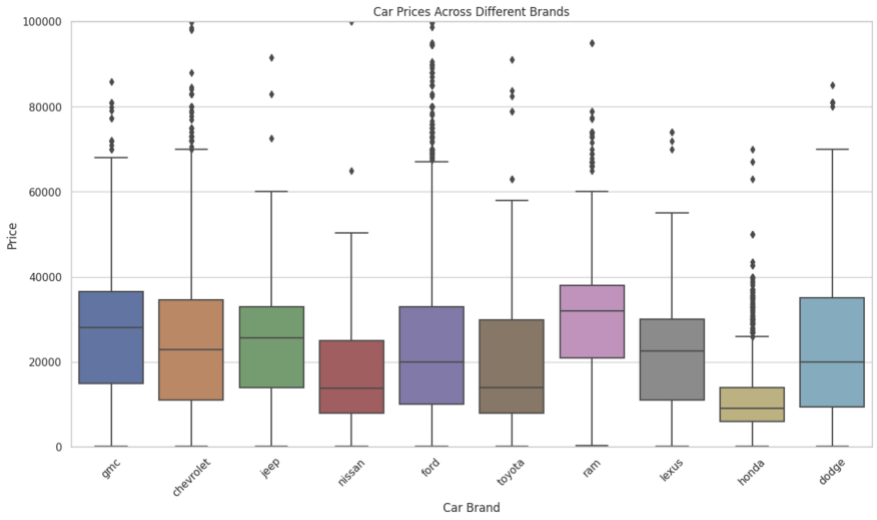


Following data cleaning, we examined the distribution of car prices using a histogram that was showed above. The results showed that car prices were highly skewed, with many listings having extremely low or unrealistically high values. To improve the reliability of the analysis, listings with prices below $100 and above $100,000 were removed to filter out unrealistic values and extreme outliers. A similar histogram below was generated for car manufacturing years, revealing that most listed vehicles were produced between 1995 and 2020, with a noticeable concentration in more recent years. This pattern aligns with market trends, where newer vehicles are more frequently listed for resale.



To explore relationships between key variables, scatter plots were generated for price vs. manufacturing year and price vs. odometer readings. As expected, newer vehicles tended to have higher prices, while older vehicles showed a downward trend in value due to depreciation.

Similarly, cars with higher mileage generally had lower prices, confirming the expected negative correlation between mileage and resale value. However, a few anomalies were present, where some older vehicles maintained high prices, likely due to classic or luxury car categories.



The price distribution was also analyzed across different manufacturers using a boxplot that is shown above. The visualization highlighted that brands such as Tesla, Porsche, and Mercedes-Benz had higher median prices compared to brands like Ford, Honda, and Toyota, which are more common in the used car market. This variation suggests that luxury brands hold their value better or are priced at a premium due to perceived quality and demand. Furthermore, categorical variables such as fuel type and transmission type were explored using count plots.

The majority of the listings were gasoline-powered vehicles, while electric and hybrid cars made up a smaller proportion of the dataset. Similarly, automatic transmission cars were more common than manual ones, reflecting market preferences.

Data Preprocessing & Feature Engineering

```python
from sklearn.preprocessing import LabelEncoder, StandardScaler

# Feature Engineering - Creating new meaningful features

# Create a new feature: Car Age (from year)
df["car_age"] = 2025 - df["year"]

# Create a mileage per year feature (assuming a car should have some miles every year)
df["mileage_per_year"] = df["odometer"] / df["car_age"]
df["mileage_per_year"] = df["mileage_per_year"].replace([np.inf, -np.inf], np.nan)
df["mileage_per_year"] = df["mileage_per_year"].fillna(df["mileage_per_year"].median())

# Encoding categorical variables using Label Encoding
encoder = LabelEncoder()
categorical_cols = ['manufacturer', 'model', 'fuel', 'transmission', 'paint_color', 'drive', 'type']
for col in categorical_cols:
    df[col] = encoder.fit_transform(df[col])

# Standardize numerical features to improve model performance
scaler = StandardScaler()
df[['odometer', 'mileage_per_year']] = scaler.fit_transform(df[['odometer', 'mileage_per_year']])

# Save the cleaned and preprocessed dataset
df.to_csv("/kaggle/working/cleaned_vehicles.csv", index=False)

# Display final dataset information
print("\nFinal Dataset Information:")
print(df.info())
print("\nData preprocessing and feature engineering completed successfully!")
print("The cleaned dataset has been saved as 'cleaned_vehicles.csv'.")
```

For the data preprocessing and feature engineering phase, due to the raw data had many missing values, irrelevant columns, and categorical variables that needed transformation, we systematically cleaned and refined the dataset to make it suitable for machine learning models.

The first step was to remove unnecessary columns that did not contribute to price prediction. Columns like id, url, region, region_url, VIN, image_url, and description were dropped because they contained non-informative or unique values that would not generalize well for a predictive model. Additionally, the county column had no values at all, making it redundant and necessary to remove. After eliminating these irrelevant fields, we addressed missing values by identifying columns where more than 50% of the data was missing. Any column exceeding

this threshold was dropped since excessive imputation could introduce bias or inconsistencies in the dataset.

For numerical variables, we replaced missing values in the year and odometer columns with their respective median values. This approach was chosen because the median is robust to extreme outliers, preventing skewed results that could arise from using the mean. For categorical variables such as manufacturer, model, fuel, transmission, title_status, and drive, we imputed missing values with the most frequently occurring category. This ensured that the dataset remained as complete as possible without introducing artificial patterns. Additionally, the paint_color column, which had a moderate number of missing values, was filled with "unknown" to retain it in the dataset while preserving its categorical nature.

To improve the reliability of the price prediction model, we applied outlier removal techniques. The price distribution showed extreme values, including listings with prices set to zero or values exceeding $100,000. Such listings could distort the model's predictions, so we filtered the dataset to retain only reasonable price values between $100 and $100,000. Similarly, we removed duplicate records to eliminate redundant data points that could bias the model.

Once the dataset was clean, we proceeded with feature engineering to create new variables that could enhance predictive performance. One of the most critical transformations was computing the car's age by subtracting the year of manufacture from 2024. This car_age variable provided a more intuitive measure of depreciation, which is a key factor in determining the resale price of a vehicle. Additionally, we created a new feature, mileage_per_year, by dividing the odometer reading by the car's age. This feature was designed to capture the average annual usage of a vehicle, which can influence its market value. Since some vehicles had missing

or zero values for mileage, we replaced infinite values resulting from division by zero with the median mileage per year to maintain consistency in the dataset.

After feature creation, we encoded categorical variables to make them compatible with machine learning algorithms. we used LabelEncoder to transform categorical columns like manufacturer, model, fuel, transmission, paint_color, drive, and type into numerical values. This approach allowed the model to interpret categorical data while preserving the relative distinctions between different brands and fuel types. Finally, we standardized numerical features such as odometer and mileage_per_year using StandardScaler. Standardization ensured that these features operated on a comparable scale, preventing models like linear regression and neural networks from being dominated by variables with large numerical ranges.

Once all preprocessing steps were completed, we saved the cleaned dataset as cleaned_vehicles.csv in the Kaggle working directory. This final dataset was structured, free of missing values, and contained meaningful engineered features, making it ready for machine learning model training and evaluation. The preprocessing and feature engineering process transformed the raw Craigslist dataset into a refined and optimized dataset that could be leveraged for accurate price predictions**.**