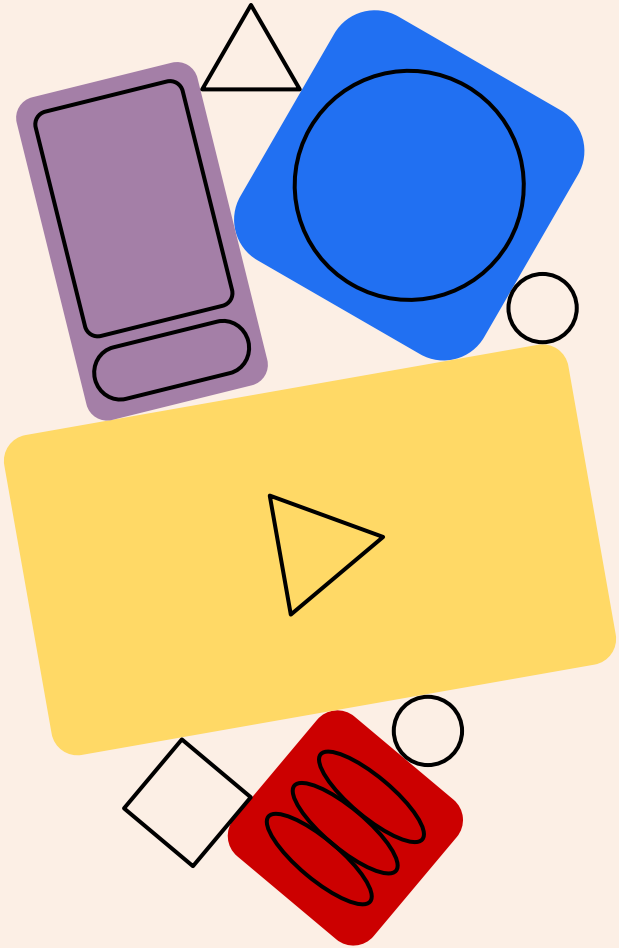
CIS 5190

Feng Jiang, Jiayi Chen, Amber Yan
Group 113

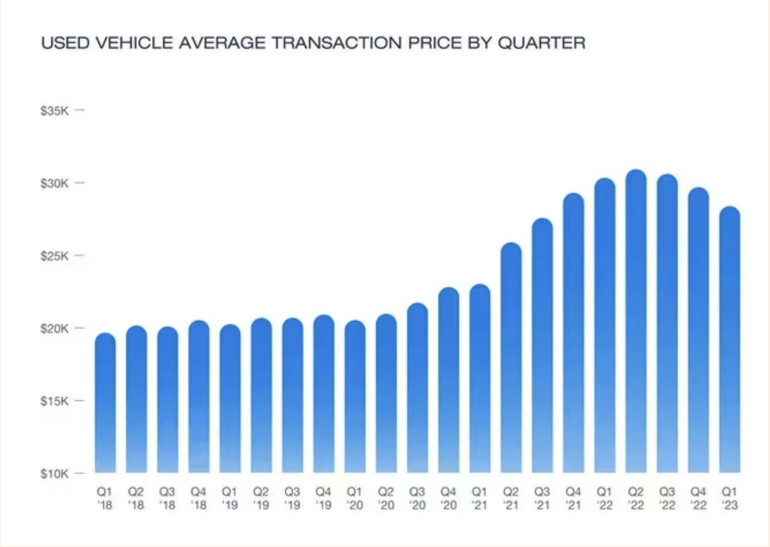
Used Car Prices Prediction



9/12/2024

Introduction

- Objective:
  - Develop a model to predict used car prices.
  - Explore key factors influencing car prices.
- Value Proposition:
  - **For Consumers:** Better buying/selling decisions.
  - **For Dealerships:** Optimized inventory and pricing.
  - **For Analysts:** Scalable data exploration methodology.



Dataset Overview

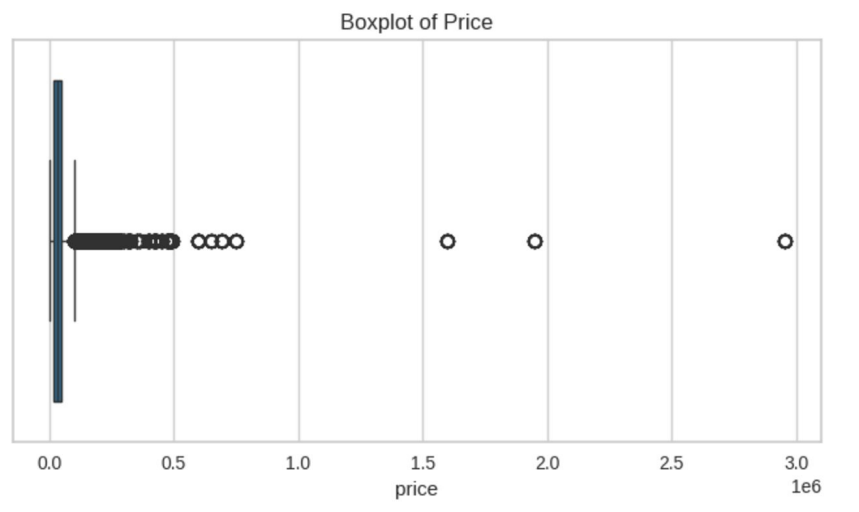
- Source: Kaggle
- Number of Records: 188533
- Number of Columns:13 columns
    - 11 numerical and categorical features and 1 target column (price)

	id	brand	model	model_year	milage	fuel_type	engine	transmission	ext_col	int_col	accident	clean_title	price
0	0	MINI	Cooper S Base	2007	213000	Gasoline	172.0HP 1.6L 4 Cylinder Engine Gasoline Fuel	A/T	Yellow	Gray	None reported	Yes	4200
1	1	Lincoln	LS V8	2002	143250	Gasoline	252.0HP 3.9L 8 Cylinder Engine Gasoline Fuel	A/T	Silver	Beige	At least 1 accident or damage reported	Yes	4999
2	2	Chevrolet	Silverado 2500 LT	2002	136731	E85 Flex Fuel	320.0HP 5.3L 8 Cylinder Engine Flex Fuel Capab...	A/T	Blue	Gray	None reported	Yes	13900
3	3	Genesis	G90 5.0 Ultimate	2017	19500	Gasoline	420.0HP 5.0L 8 Cylinder Engine Gasoline Fuel	Transmission w/Dual Shift Mode	Black	Black	None reported	Yes	45000
4	4	Mercedes-Benz	Metris Base	2021	7388	Gasoline	208.0HP 2.0L 4 Cylinder Engine Gasoline Fuel	7-Speed A/T	Black	Beige	None reported	Yes	97500

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 188533 entries, 0 to 188532
Data columns (total 13 columns):
 #   Column        Non-Null Count   Dtype 
---  ------        --------------   ----- 
 0   id            188533 non-null  int64 
 1   brand         188533 non-null  object
 2   model         188533 non-null  object
 3   model_year    188533 non-null  int64 
 4   milage        188533 non-null  int64 
 5   fuel_type     183450 non-null  object
 6   engine        188533 non-null  object
 7   transmission  188533 non-null  object
 8   ext_col       188533 non-null  object
 9   int_col       188533 non-null  object
 10  accident      186081 non-null  object
 11  clean_title   167114 non-null  object
 12  price         188533 non-null  int64 
dtypes: int64(4), object(9)
memory usage: 18.7+ MB
```

Exploratory Data Analysis

Remove Outliers in Price using Interquartile Range (IQR) method

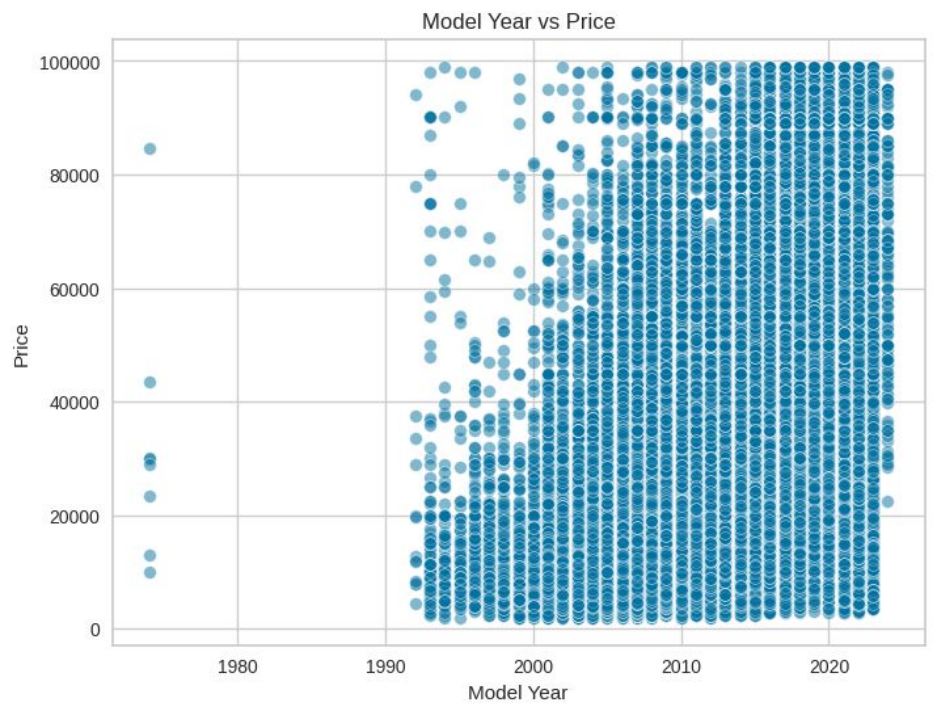


Exploratory Data Analysis

- Negative Correlation between Mileage and price

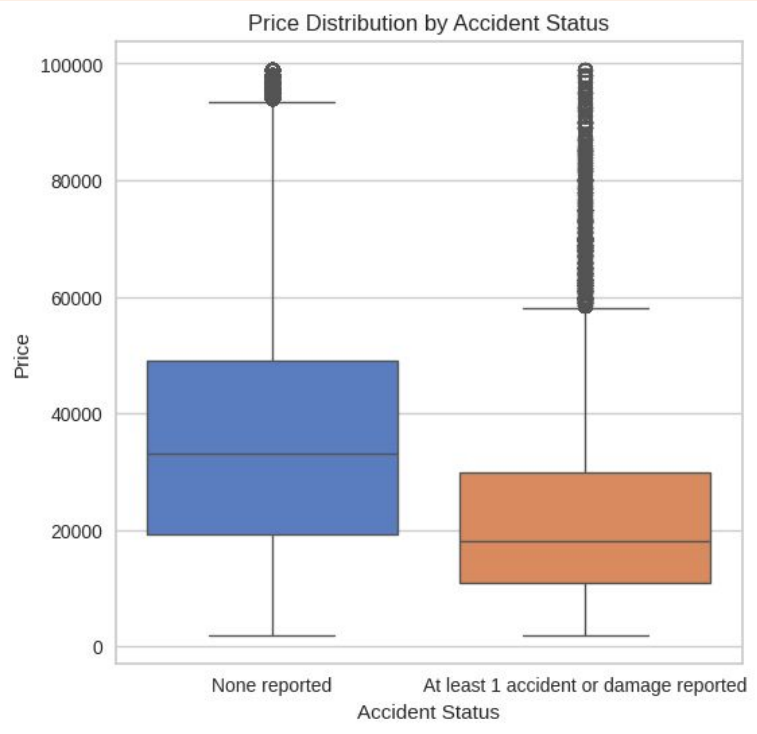
Exploratory Data Analysis

- Newer models tend to have higher price



Exploratory Data Analysis

- Vehicles with accidents reported have significantly lower prices.



Data Preprocessing & Feature Engineering

Extract useful numerical features from ‘engine’ column

- Tools used: Regex
- Features extracted
  - horse_power
  - engine_size
  - cylinders
- Example Features extraction from the first row:
  - horsepower: 172.0
  - engine_size: 1.6
  - cylinders: 4

	engine
0	172.0HP 1.6L 4 Cylinder Engine Gasoline Fuel
1	252.0HP 3.9L 8 Cylinder Engine Gasoline Fuel
2	320.0HP 5.3L 8 Cylinder Engine Flex Fuel Capab...
3	420.0HP 5.0L 8 Cylinder Engine Gasoline Fuel
4	208.0HP 2.0L 4 Cylinder Engine Gasoline Fuel

Data Preprocessing & Feature Engineering

Mapping ext_col and int_col into standard colors

- **Problem**: Features ext_col (Exterior color) and int_col (Interior color) have high cardinality
  - ext_col: 318 unique values
  - int_col: 156 unique values
- **Solution**: Mapping Colors to Standard Categories
  - Group unique colors into standardized categories like Black, White, Gray, etc.
  - Example:
    - "Carbon Black", "Jet Black", "Obsidian" → **Black**
    - "Snowflake White Pearl", "Ivory", "Crystal White" → **White**
- Reduced to a manageable set of categories for easier encoding and modeling.

Data Preprocessing & Feature Engineering

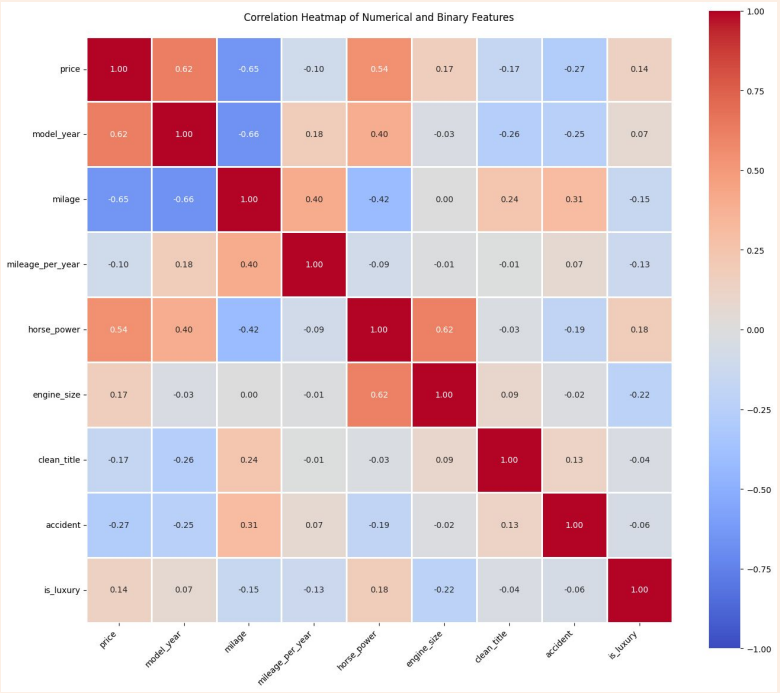
- Create feature mileage_per_year
  - mileage_per_year = mileage / age (age = current_year - model_year)
  - Capture the relationship between the car's age, usage, and price
- Unify categories for transmission (Automatic, Manual, CVT, Other)
  - Example: "8-Speed A/T" → Automatic
- Create binary feature is_luxury to identify luxury vehicles
  - Example: Mercedes-Benz → 1, Chevrolet → 0
- Handle missing values
  - Fuel Type: Missing values filled with mode
  - Accident History: Missing values filled with 0
  - Horsepower & Engine Size: Median imputation grouped by brand
  - Results: Improved consistency across the dataset

Data Preprocessing & Feature Engineering

- Baseline models like Linear Regression and Random Forest cannot process categorical values directly
- Target Encoding:
  - Applied to high-cardinality features like brand by replacing each category with the mean price for that brand
- One-Hot Encoding:
  - Used for nominal variables (fuel_type, int_col, ext_col) to create binary columns

Correlation Analysis of Features

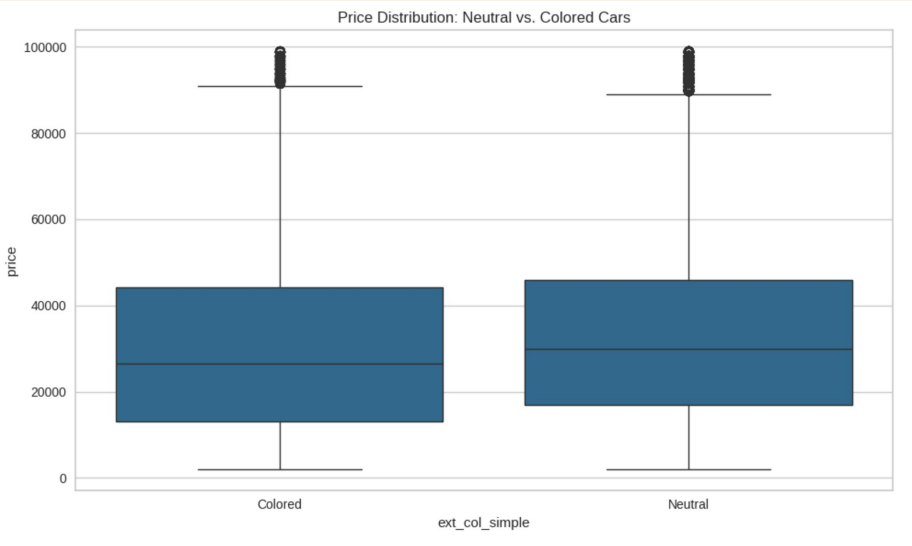
- Price:
  - Positively correlated with model_year (0.62) and horse_power (0.54).
  - Negatively correlated with mileage (-0.65) and accident (-0.27).
- Mileage:
  - Strong negative correlation with model_year (-0.66), reflecting older cars accumulate higher mileage.
- Horsepower & Engine Size:
  - Positively correlated (0.62), aligning with expectations for vehicle specifications.



Hypothesis Testing

Hypothesis 1: Price Premium on Color Hypothesis

- **H0:** No price difference between cars with neutral colors (Black, White, Silver, Gray) and vibrant colors.
- **H1:** Neutral-colored cars are priced higher.



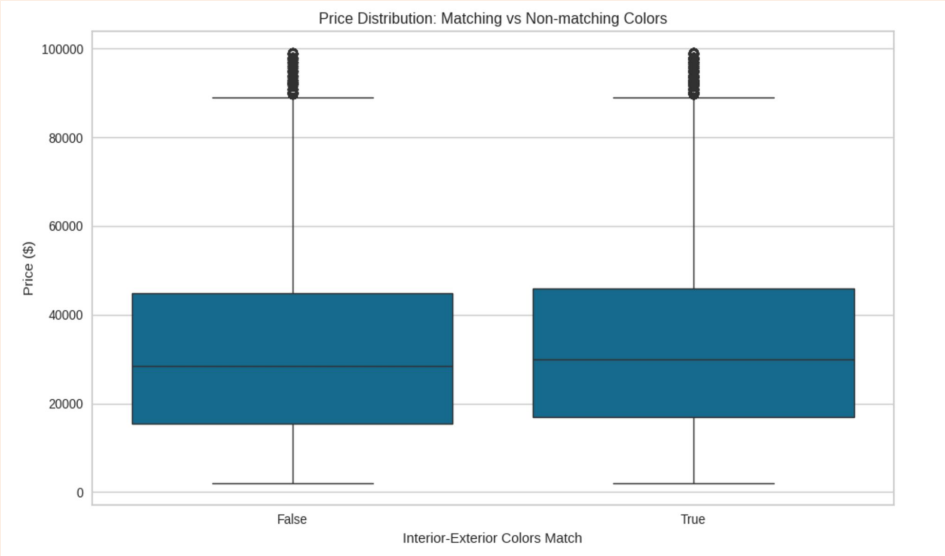
Result and Analysis:

- Reject the null hypothesis
- P-value = 8.50e-89 < 0.05
- Mean Prices:
  - Neutral: $33,895.82
  - Colored: $31,314.07

Hypothesis Testing

Hypothesis 2: Exterior & Interior Matching Colors

- **H0:** No difference in prices between cars with matching and non-matching colors
- **H1:** Cars with matching colors have significantly different prices



Result and Analysis:

- Reject the null hypothesis
- P-value = 3.97e-18 < 0.05
- Mean Prices:
  - Matching: $34,120.50
  - Non-Matching: $32,890.75

Model Performance and Analysis

Model	Train RMSE	Test RMSE	Train R^2	Test R^2
Linear Regression	14349	14297	0.55	0.55
Random Forest	13222	13295	0.62	0.61
XGBoost (Unencoded)	12536	12873	0.66	0.63
XGBoost (Encoded)	12688	12898	0.65	0.63
LightGBM (Unencoded)	12723	12875	0.65	0.63
LightGBM (Encoded)	12802	12919	0.64	0.63
PyCaret (Tuned LGBM)	12400	12880	0.67	0.63

Linear Regression Lags Significantly:

- **Struggles with non-linear relationships and complex data patterns.**
- **Performance is significantly lower than advanced models.**

Tree-Based Models Excel:

- **Handle non-linear relationships effectively.**
- **Boosting models (XGBoost, LightGBM) deliver better generalization than Random Forest.**

Impact of Encoding:

- **Encoding features offers minimal improvement for LightGBM and XGBoost.**
- **Models like LightGBM perform better with native categorical data handling.**

PyCaret's Performance:

- **PyCaret successfully identifies LightGBM as the best model.**
- **Hyperparameter tuning marginally improves train metrics but shows limited impact on test performance.**

Implications and Insights

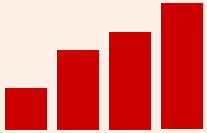
Key Market Insights:

- Colors significantly impact used car prices.
- Mileage, model year, and accident history are critical determinants of price.

Practical Applications:

- For Consumers: Accurate valuation for fair pricing
- For Dealerships: Optimize inventory with high-demand features.
- For Industry: Develop AI-powered price prediction tools.

Challenges and Limitations



1. Missing and Imbalanced Data

Challenge: Addressing missing values and imbalances in key features during preprocessing.

Limitation: Imputation strategies (e.g., median, mode) sometimes affected model generalization, particularly in features with skewed distributions.

2. Handling Categorical Features

Challenge: Deciding whether to manually encode categorical features or rely on models like LightGBM that natively handle categorical data.

Future Agenda

☑️ **Advanced Modeling Techniques**

Deep Learning Models:

- such as neural networks, can capture complex relationships between features that traditional models might miss.

☑️ **Handling Temporal Features**

Significance of Temporal Features:

- Temporal features, such as the year of manufacture or economic trends, influence car prices significantly.

☑️ **Encoding Techniques for Categorical Data:**

- Replace one-hot encoding with Target Encoding or Frequency Encoding to handle high-cardinality categories like brand names.

☑️ **Scalability:**

- Deploy the model in a user-friendly application. E.g. Carmax Used Car Evaluation
- Implement dynamic model updates with new data.

Future Work and Enhancements

Thank you for listening