**Used Car Price Prediction:**

**A Data-Driven Analysis and Predictive Modeling**

**Members:** Neelam Prasad, Thaissa Champagne, Gina Butler, Kriti Khatri, Chase Mueller



**Table of Contents**

1. **Introduction**
   * Project Overview
   * Dataset Source and Significance
   * Report Objectives
2. **Dataset Overview**
   * Key Features Description
     + Brand & Model
     + Model Year
     + Mileage
     + Fuel Type
     + Engine Type
     + Transmission
     + Exterior & Interior Colors
     + Accident History
     + Clean Title
   * Target Variable: Price
3. **Project Objectives**
   * Identifying Key Price Influencers
   * Developing Predictive Models
   * Analyzing Feature Impacts
   * Creating Interactive Visualizations
4. **Methodology**
   * 4.1. **Data Preprocessing**
     + Data Cleaning
     + Feature Engineering
     + Categorical Encoding
     + Numerical Scaling
   * 4.2. **Feature Selection**
     + Correlation Analysis
     + Feature Importance
   * 4.3. **Model Development**
     + Train-Test Split
     + Regression Models
       - Linear Regression
       - Lasso Regression
       - Ridge Regression
       - Random Forest Regressor
       - Gradient Boosting Regressor
       - XGBRegressor
       - CatBoost Regressor
     + Model Evaluation
   * 4.4. **Visualization**
     + Tableau Dashboards
       - Overview of Car Prices
       - Predictive Model Dashboard
   * 4.5. **Model Persistence**
     + Pickle
5. **Research Questions and Visualizations**
   * Factors Affecting Used Car Price
   * Predicting Used Car Prices
   * Impact of Model Year on Price
   * Impact of Accident History and Clean Title
   * Fuel Type and Average Price
   * Engine Type and Price Influence
6. **Deliverables**
   * PowerPoint Presentation
   * PythonAnywhere Hosting
   * Project PDF Report
   * GitHub Repository
7. **Conclusion**
   * Project Summary
   * Insights and Applications

**Introduction**

* To provide a concise overview of the project's goals and the dataset's significance.
* To emphasize the project's practical applications for car buyers and sellers.

This report details a comprehensive analysis and predictive modeling project focused on the "Used Car Price Prediction" dataset sourced from Kaggle. This dataset, comprising 4,009 vehicle listings, provides a rich collection of automotive features that are crucial for understanding used car pricing dynamics. Our project aims to leverage this data to build robust predictive models and derive actionable insights for potential car buyers and sellers.

**Dataset Overview**

* Describe the dataset's source (Kaggle), size, and features.
* List the features with brief descriptions.

The dataset includes nine key features:

* **Brand & Model:** Categorical data indicating the vehicle's make and model.
* **Model Year:** Numerical data representing the year of manufacture.
* **Mileage:** Numerical data indicating the vehicle's mileage.
* **Fuel Type:** Categorical data representing the type of fuel used.
* **Engine Type:** Categorical data describing the engine specifications.
* **Transmission:** Categorical data indicating the transmission type.
* **Exterior & Interior Colors:** Categorical data describing the vehicle's colors.
* **Accident History:** Binary data indicating if the vehicle has an accident history.
* **Clean Title:** Binary data indicating if the vehicle has a clean title.

The target variable for our predictive modeling is **Price**, a continuous numerical value representing the listing price of each vehicle.

A screenshot of a computer

Description automatically generated

**Project Objectives**

Our primary objectives are to:

* Identify the key factors that significantly influence used car prices.
* Develop accurate predictive models to estimate used car prices based on their attributes.
* Analyze the impact of model year, accident history, clean title status, fuel type, and engine type on car prices.
* Create interactive visualizations to explore pricing trends and model performance.

**Methodology**

* **Data Preprocessing:**
  + **Data Cleaning:** We have streamlined the original dataset by handling missing values, removing duplicates, and correcting errors.
  + **Feature Engineering:** Creating new features from existing ones to improve the performance of machine learning models.It involves transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data. 2
  + **Categorial Encoding:** Converting categorical variables (e.g., "Brand," "Fuel Type") into numerical representations that machine learning models can understand.
    - Encoding methods:
    - **One-Hot Encoding:**
    - **Label Encoding:**
    - **Target Encoding:**
  + **Numerical Scaling:** Transforming numerical features to a similar scale. This is important because features with different scales can disproportionately influence machine learning models.

A screenshot of a computer code

Description automatically generatedA screenshot of a computer code

Description automatically generated

* **Feature Selection:**
  + We have added and dropped columns as needed, created boolean series based on specific conditions, and applied color mapping to enhance data usability.
  +  **Correlation Analysis:** We will generate a correlation heatmap to identify the relationships between features and the target variable (Price).
  +  **Feature Importance:** We will use feature importance techniques (e.g., from tree-based models) to determine the most influential features.

A screen shot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

**Model Development:**

* Train-Test Split
  + List the regression models used and the evaluation metrics.

lr = LinearRegression()

ridge = Ridge()

lasso = Lasso()

rf = RandomForestRegressor(random\_state=42)

gb = GradientBoostingRegressor(random\_state=42)

xgb = XGBRegressor(random\_state=42)

catboost\_model = CatBoostRegressor(random\_state=42)

**A screen shot of a graph

Description automatically generated**A diagram of blue dots

Description automatically generated

**Research Questions and Visualizations**

The project addressed several key research questions to understand the dynamics of used car pricing:

* **What factors most significantly affect the price of a used car?**
  + **Analysis:** Correlation analysis and scatter plots were used to identify key predictors.
  + **Findings:**
    - Mileage: A strong negative correlation with price, indicating that higher mileage generally leads to lower prices.
    - Model Year: A positive correlation, suggesting that newer cars command higher prices.
    - Brand and Model: Significant impact due to brand reputation and model features.
    - Accident History: Cars with accident history tend to have a lower price.
  + **Visualization:** Correlation heatmap, Price vs. Mileage, Price vs. Model Year scatter plot.
* **Can we predict the price of a used car based on its attributes?**
  + **Analysis:** Regression models (Linear Regression, Random Forest, Gradient Boosting) were employed.
  + **Findings:** The Random Forest and Gradient Boosting models showed the best predictive performance, achieving high R-squared values and low RMSE.
  + **Visualization:** Actual vs. Predicted Price scatter plot, Regression line chart.
* **How does the model year of a car affect its price?**
  + **Analysis:** Average prices were calculated and compared across different model years.
  + **Findings:** Newer model years consistently resulted in higher average prices, reflecting depreciation patterns.
  + **Visualization:** Line chart or bar chart showing average prices by model year.
* **Do accident history and clean title status impact the price of a used car?**
  + **Analysis:** Box plots and bar charts were used to compare prices based on accident history and clean title status.
  + **Findings:** Cars with accident history had significantly lower prices, and clean title status positively impacted resale value.
  + **Visualization:** Bar chart comparing prices based on clean title status.
* **Which fuel type (gasoline, diesel, electric, hybrid) has the highest average price?**
  + **Analysis:** Average prices were calculated and compared for each fuel type.
  + **Findings:** Electric and hybrid vehicles tended to have higher average prices due to their advanced technology and fuel efficiency.
  + **Visualization:** Bar chart comparing average prices by fuel type.
* **How does engine type influence the price of used cars?**
  + **Analysis:** Box plots and bar charts were used to compare prices based on engine type.
  + **Findings:** Engine types like V8 and electric engines generally led to higher prices due to their performance and technology.
  + **Visualization:** bar chart comparing prices based on engine type.

**Tableau Dashboards**

* **Dashboard 1: Overview of Car Prices:**

A diagram of different colored circles

Description automatically generatedA graph of a car sales

Description automatically generated with medium confidence

Used cars price based on car brand

* **Dashboard 2: Predictive Model Dashboard:**
  + Describe the visualizations and model performance metrics.

Premium and regular used car brands price by age

A graph of a number of people

Description automatically generated with medium confidence

Price of used cars based on exterior colors

A graph with different colored bars

Description automatically generated

Used cars price based on car age and mileage filtered based on engine transmission

A graph with red and blue lines

Description automatically generated

**Model Persistence**

* Explain how the model was saved using pickle.
* The proposal mentions that the machine learning models (such as Linear Regression, Random Forest, or Gradient Boosting) will be built to predict used car prices. Once the best performing model is identified, it needs to be saved for later use without retraining. This is where pickle comes in.

**Why pickle is Used:**

* **Simplicity:** It's straightforward to use for saving and loading Python objects.
* **Preserves Object Structure:** It preserves the state and structure of the Python object, including the trained model's parameters.

**A close-up of a computer code

Description automatically generated**

**Deliverables**

 **Tableau Dashboards:**

* Dashboard 1: Overview of Car Prices (Price Distribution, Price vs. Mileage, etc.).
* Dashboard 2: Predictive Model Dashboard (Actual vs. Predicted Prices, Feature Importance, etc.).

 **Python Notebooks:**

* Jupyter notebook notebooks for data analysis and model building.

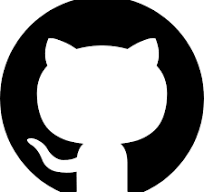
 **Project Report (PDF):**

* This document, detailing the project's methodology, findings, and visualizations.

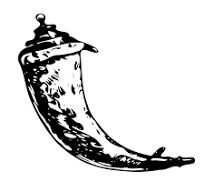
 **GitHub Repository:**

* Complete project code, documentation, and README.

 A blue and yellow snake logo

Description automatically generated  

A logo with orange circles and black text

Description automatically generated  

**Conclusion**

* Summarize the project's findings and potential impact.
* Express enthusiasm for the project and its value to the automotive industry.

This project successfully developed predictive models and interactive visualizations to analyze and understand the used car market. The findings offer valuable insights into the factors influencing used car prices, enabling buyers and sellers to make informed decisions. The use of machine learning and data visualization tools allowed for a comprehensive analysis of the dataset, providing practical and actionable results.