**Predicting Used Car Prices with Machine Learning Algorithms**

Divyesh Rathod, Meet Shah, Vedant Thaker

*Arizona State University I*

# Abstract

In this study, we aim to predict the prices of used cars based on various features using different machine learning algorithms. We used a publicly available dataset of 20,063 used car records with 11 features. After processing the data, we trained and compared linear regression, lasso regression, ridge regression, and polynomial regression algorithms on a local computer, Google Colab, and AWS SageMaker instance. The results and performance of each model are discussed, along with their training times on different platforms.

# Introduction

The goal of this study is to develop a model that can predict the prices of used cars based on various features such as trim, mileage, year, color, and other factors. This is important because accurate price prediction can help buyers and sellers make informed decisions when purchasing or selling a used car. In this project, we used machine learning algorithms to train a model that can make such predictions with reasonable accuracy. Specifically, we experimented with linear regression, lasso regression, ridge regression, and polynomial regression algorithms, and compared their performance on different platforms.

# Data Description

The dataset used in this project consists of 20,063 used car records with 11 features: price, trim, isOneOwner, mileage, year, color, displacement, fuel, region, soundSystem, and wheelType. The target variable is the price, while the other 10 features represent various features of the cars. The dataset includes both numerical and categorical features. Table below provides information about the features present in the dataset.

[Table 1 about here.]

*Email addresses:* [drathod2@asu.edu](mailto:drathod2@asu.edu) (Divyesh Rathod), [mshah72@asu.edu](mailto:mshah72@asu.edu) (Meet Shah),

[vthaker1@asu.edu](mailto:vthaker1@asu.edu) (Vedant Thaker)

[Figure 1 about here.] [Figure 2 about here.] [Figure 3 about here.] [Figure 4 about here.] [Figure 5 about here.] [Figure 6 about here.] [Figure 7 about here.] [Figure 8 about here.] [Figure 9 about here.]

# Data Processing

* 1. The following steps were performed to process the data:Read the data.
  2. Read the data
  3. Replaced ’unsp’ with ’NaN’ (missing values in the ’color’, ’soundSystem’, and ’wheel- Type’ columns
  4. Dropped ’soundSystem’ and ’wheelType’ columns from the training data, as they had more than 40% of missing data.
  5. Filled in the missing values in the ’color’ column, which had nearly 4% missing data, using the IterativeImputer technique. To apply this method, we converted the cat- egorical values of the ’color’ column to numerical values using the OrdinalEncoder method.
  6. Selected the best features using the SelectKBest method. Based on this method, we chose the following five features: ’isOneOwner’, ’mileage’, ’year’, ’displacement’, and ’trim’.
  7. In addition to SelectKBest, we also used the PCA method for feature selection
  8. Split the data into train (80%) and test (20%) sets for model training and evaluation.

# Plan of Attack

We planned to train and compare 7 different regression algorithms to predict used car prices:

* 1. Decision Tree
  2. KNN Regression
  3. Lasso Regression
  4. Linear Regression
  5. Polynomial Regression
  6. Random Forest Regression
  7. Ridge Regression

These algorithms were chosen because they are commonly used for regression problems and can provide a good baseline for comparison. In addition, we trained the models on three different platforms – local computer, Google Colab, and AWS SageMaker – to compute the training times and compare their performance.

# Results

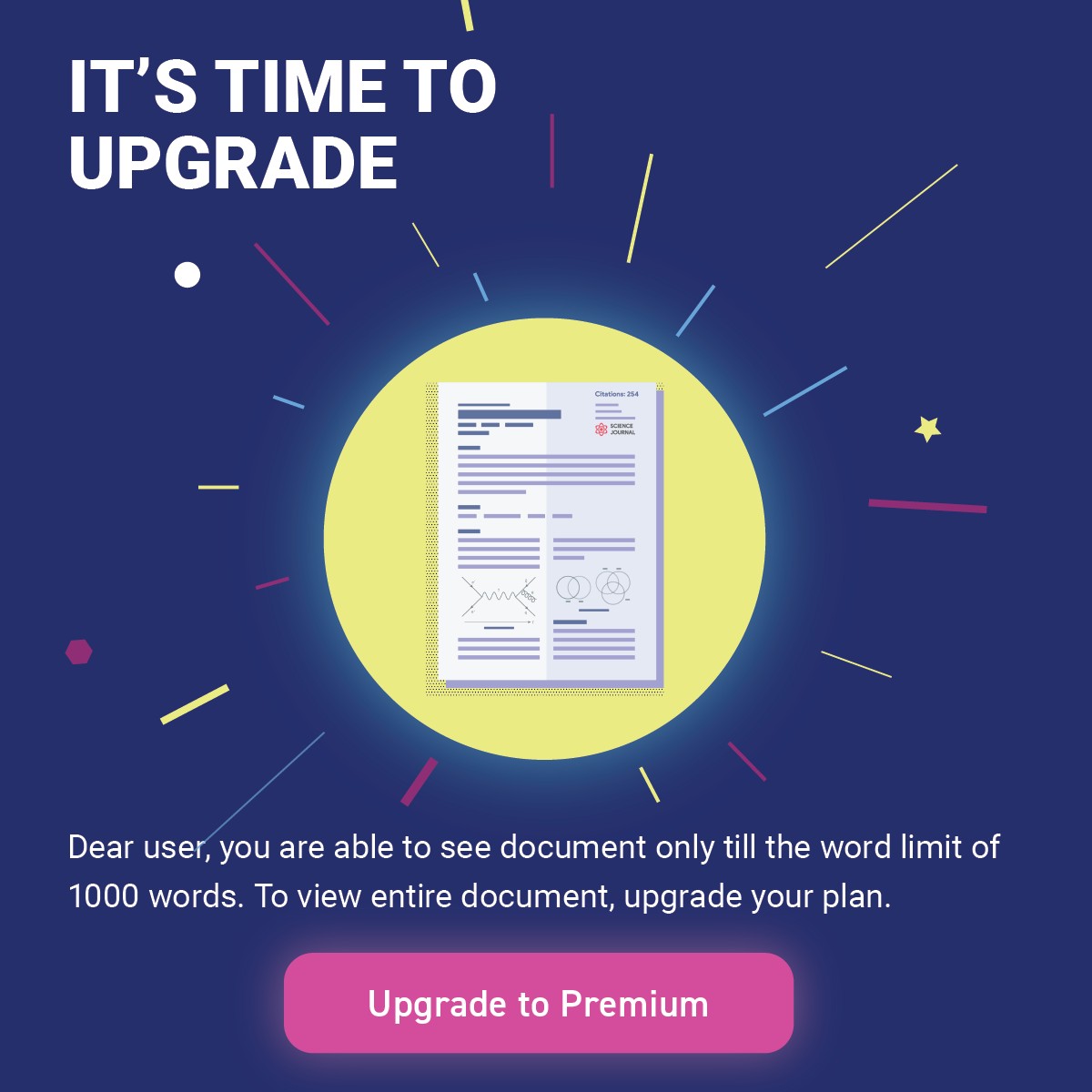
[Table 2 about here.]

[Figure 10 about here.]

We trained and evaluated various machine learning models to predict used car prices based on the selected features. Here is a detailed discussion of the results for each model:

* 1. **Decision Tree Model**: This model demonstrated high prediction accuracy with an R Square

[Edit this article on Typeset](https://typeset.io/edit/f38acd91-a904-49f9-9e46-f2bacd4c1257)



[www.typeset.io](https://typeset.io/) [Looking for a Discount?](https://www.typeset.io/orders/coupon/ZfAB8STU/?source=pricing-page-student-discount)

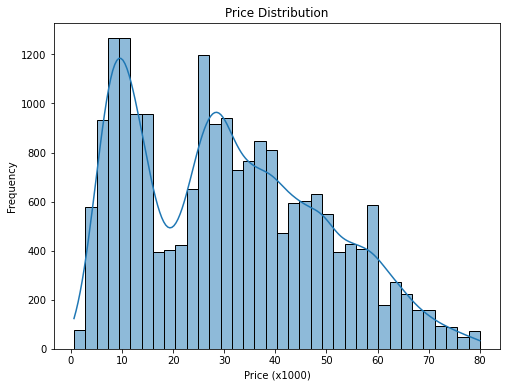


Figure 1: This visualization presents the distribution of car prices in the dataset. It is important to understand the overall distribution of the dependent variable (price) to identify any potential outliers or skewness in the data that might impact our model’s performance.

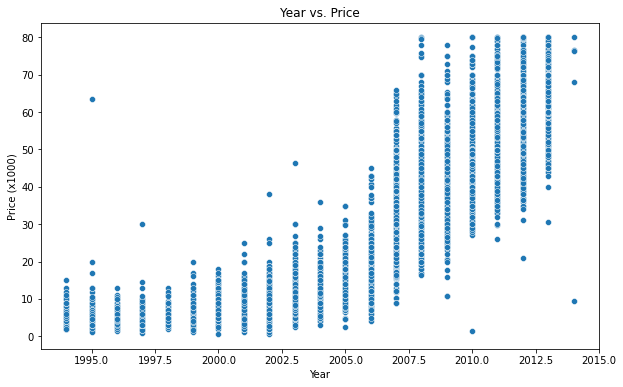


Figure 2: This scatter plot displays the relationship between the manufacturing year of the cars and their prices. It helps in identifying any trends or patterns between the age of the cars and their selling price, which might be a significant factor in predicting the car prices.

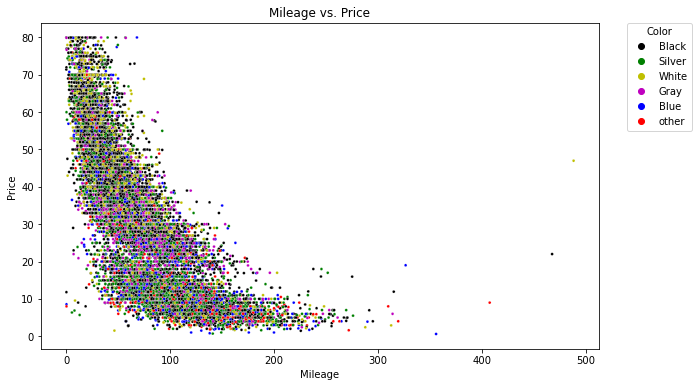


Figure 3: This scatter plot showcases the relationship between the mileage of the cars and their prices, with each data point colored according to the car’s color. It helps in understanding how the mileage affects the car price and whether the color of the car has any impact on the relationship between mileage and price.

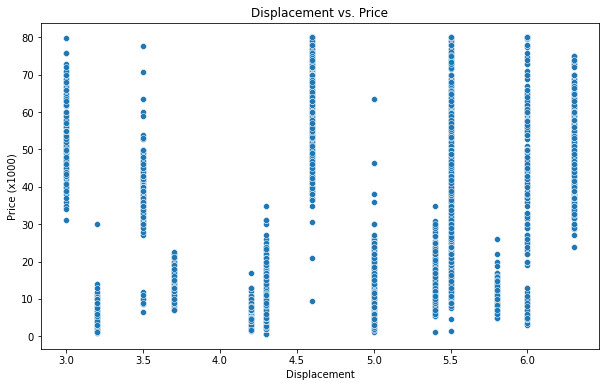


Figure 4: This scatter plot presents the relationship between the engine displacement of the cars and their prices. It helps in determining if there is a correlation between the engine size and the selling price of the cars, which might be a useful factor in predicting car prices.

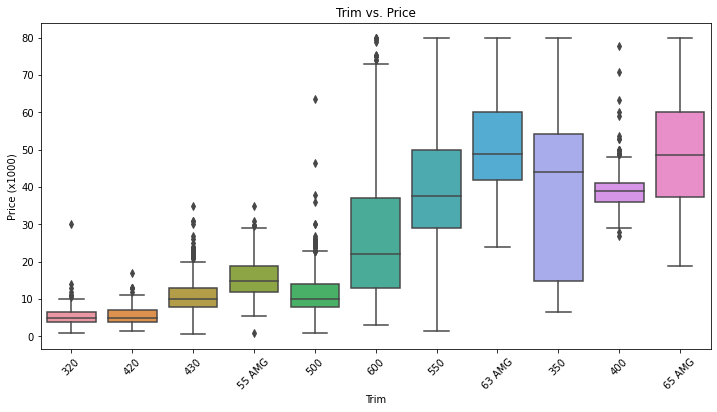


Figure 5: This box plot displays the distribution of car prices for different colors. It helps in determining if the color of the car has any significant impact on its price, which might be an important factor to consider when predicting car prices.

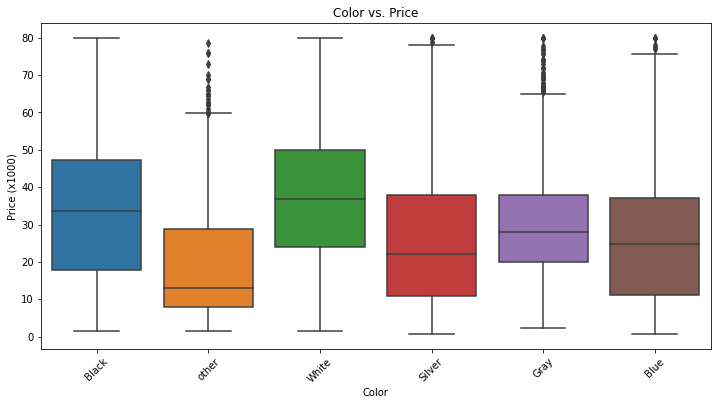


Figure 6: This box plot displays the distribution of car prices for different colors. It helps in determining if the color of the car has any significant impact on its price, which might be an important factor to consider when predicting car prices.

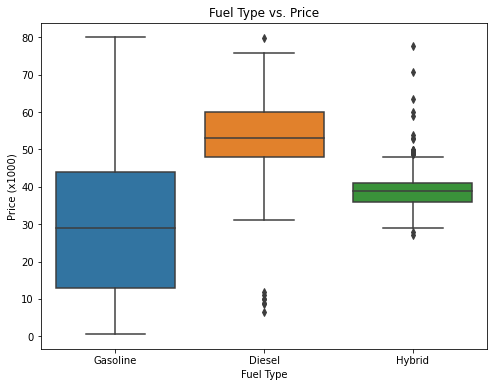


Figure 7: This box plot shows the distribution of car prices for different fuel types. It helps in understanding how the fuel type affects the car price and whether certain fuel types are more expensive than others, which might be a significant factor in predicting the car prices.

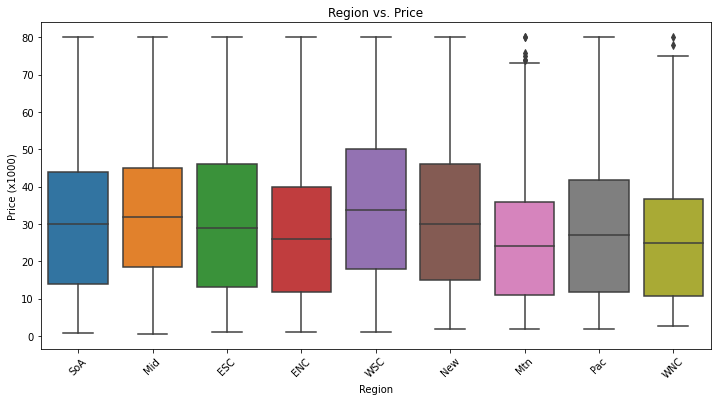


Figure 8: This box plot presents the distribution of car prices across different regions. It helps in identifying if there are any regional differences in car prices, which might be due to factors such as regional demand, taxes, or other economic factors that can influence the car’s value.

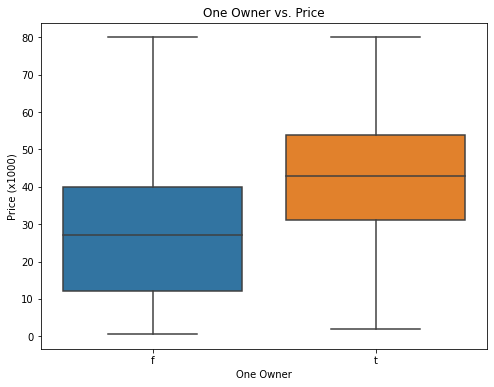


Figure 9: This box plot shows the distribution of car prices for cars with one owner versus multiple owners. It helps in understanding the impact of car ownership history on its price, as cars with a single owner might be perceived as better maintained or more reliable, which can affect their value.

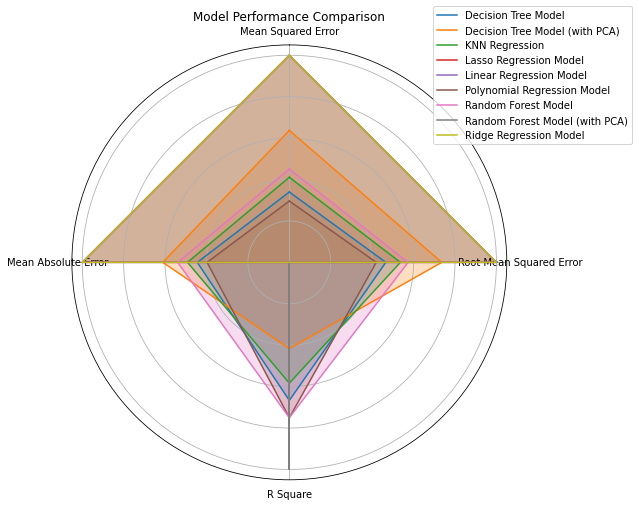


Figure 10: Graphical representation of model performance

# List of Tables

1. [Description of dataset](#_bookmark0) 16
2. [Model performance comparison](#_bookmark1) 17

Table 1: Description of dataset

|  |  |  |
| --- | --- | --- |
| **Feature** | **Category** | **Description** |
| **price** | Numerical | Price of the used car |
| **trim** | Categorical | Trim level of the car |
| **isOneOwner** | Categorical | Whether the car had only one owner |
| **mileage** | Numerical | Mileage of the car |
| **year** | Numerical | Year of manufacture |
| **color** | Categorical | Color of the car |
| **displacement** | Numerical | Engine displacement (in liters) |
| **fuel** | Categorical | Type of fuel used by the car |
| **region** | Categorical | Geographical region of the car |
| **soundSystem** | Categorical | Type of car sound system |
| **wheelType** | Categorical | Type of car wheel |

Table 2: Model performance comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Mean Squared Error** | **Root Mean Squared Er- ror** | **R Square** | **Mean Abso- lute Error** |
| **Decision Tree** | 18.91 | 4.35 | 0.94 | 3.1 |
| **Model** |  |  |  |  |
| **Decision Tree** | 31.11 | 5.58 | 0.91 | 3.7 |
| **Model (with** |  |  |  |  |
| **PCA)** |  |  |  |  |
| **KNN Regres-** | 21.83 | 4.67 | 0.93 | 3.27 |
| **sion** |  |  |  |  |
| **Lasso Regres-** | 45.97 | 6.78 | 0.86 | 5.1 |
| **sion Model** |  |  |  |  |
| **Linear Re-** | 45.98 | 6.78 | 0.86 | 5.1 |
| **gression** |  |  |  |  |
| **Model** |  |  |  |  |
| **Polynomial** | 17.13 | 4.14 | 0.95 | 2.93 |
| **Regression** |  |  |  |  |
| **Model** |  |  |  |  |
| **Random For-** | 23.45 | 4.84 | 0.95 | 3.44 |
| **est Model** |  |  |  |  |
| **Random For-** | 4.99 | 2.23 | 0.98 | 1.5 |
| **est Model** |  |  |  |  |

# (with PCA) Ridge Re- gression Model

45.97 6.78 0.86 5.1