# Used Car Price Prediction

## CISC 520 Final Project

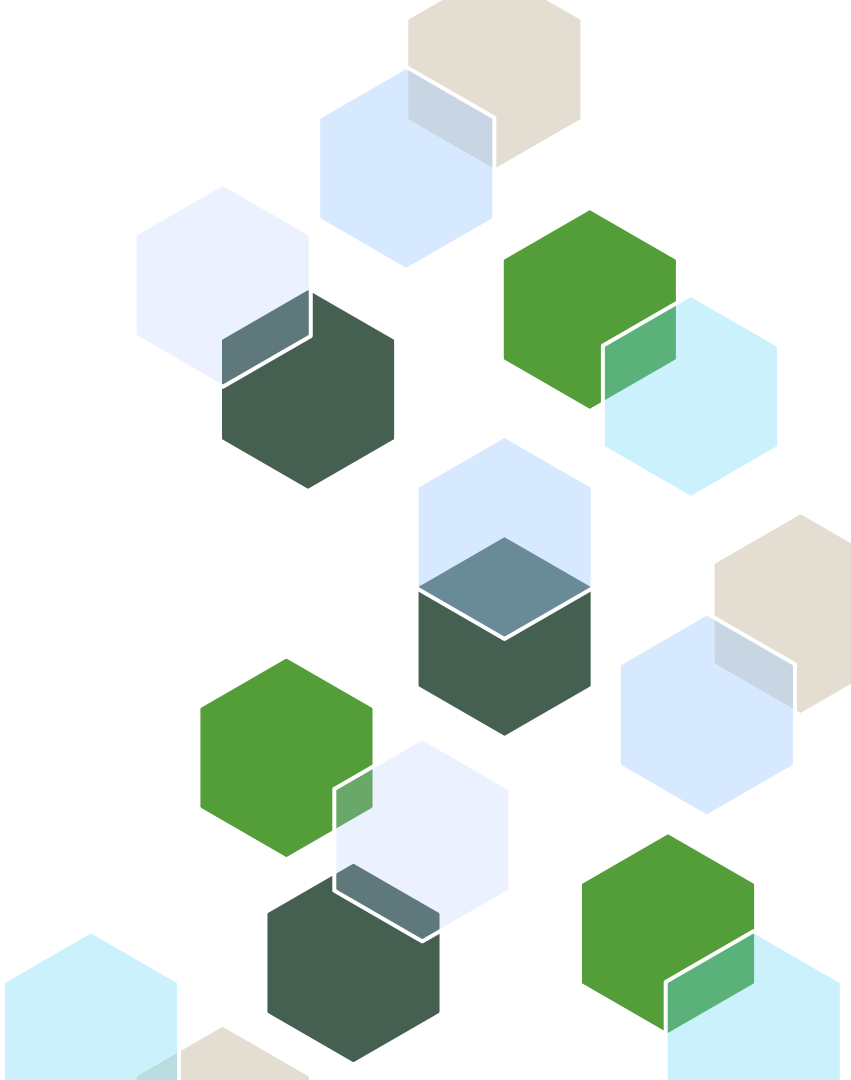
Siyuan Feng
Xinyuan He
Yingyu Cao

HARRISBURG UNIVERSITY
OF SCIENCE AND TECHNOLOGY

# Content

# 01

# Introduction



# Introduction

The used car market has experienced significant growth in recent years, necessitating accurate pricing mechanisms to facilitate fair transactions between buyers and sellers. Traditional pricing methods, often reliant on manual assessments, fail to capture the complex interactions among various factors influencing car prices. This project aims to develop a predictive model for second-hand car prices using advanced machine learning techniques.

Our approach involves comprehensive data visualization, preprocessing, feature engineering, and model evaluation to ensure the robustness of the predictions. The detailed steps include utilizing permutation importance and auto feature selection with Random Forest to identify the most significant features; implementing imputation, encoding, and grid search on Random Forest and Gradient Boosting models to optimize performance. The final model, a Gradient Boosting Regressor, achieved a test accuracy of 79%.

# 02

# Problem Statement

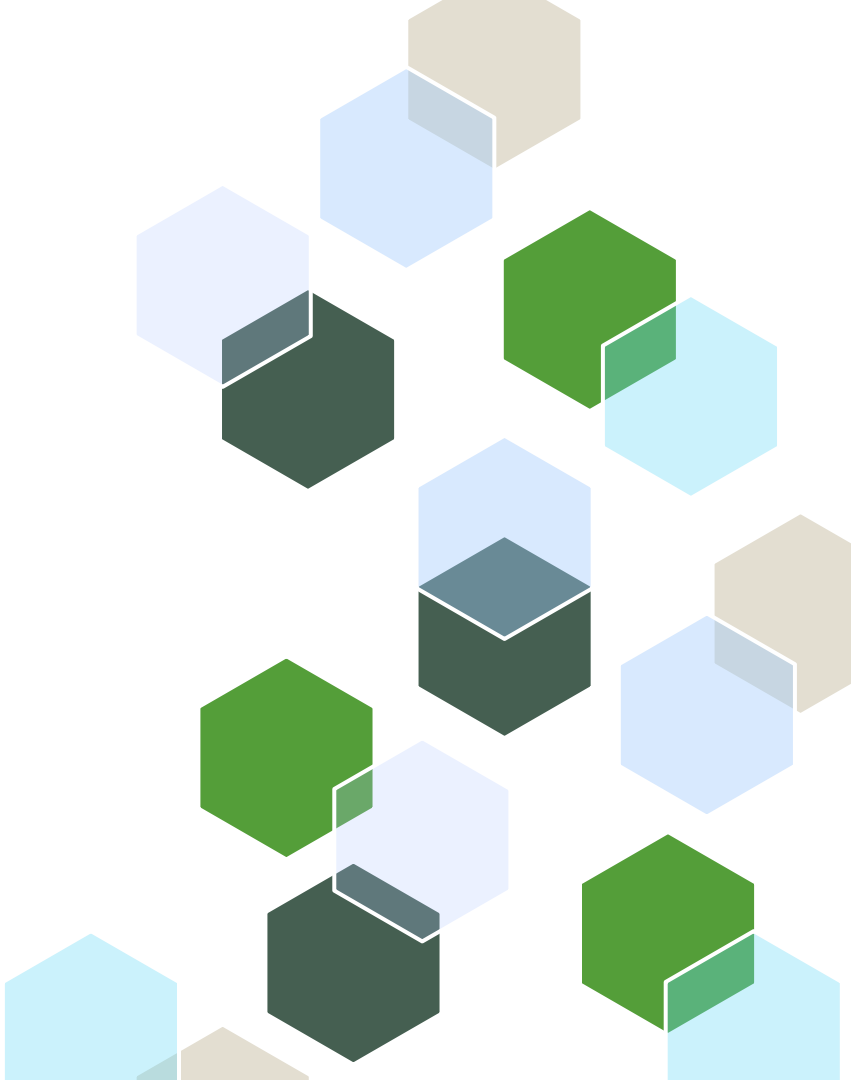
# Problem Statement

Determining the fair price of a used car is a multifaceted problem involving numerous variables. Traditional methods of pricing, while useful, are limited in their ability to capture non-linear relationships and interactions between these variables. This project seeks to address this limitation by developing a machine learning-based predictive model for second-hand car prices using Craigslist data.

By leveraging historical data and advanced algorithms, we aim to create a tool that provides accurate and data-driven price estimates. This model will assist stakeholders in the used car market by enhancing transparency and efficiency, ultimately leading to fairer transactions.

# 03

# Methodology



# Methodology

## Linear Regression

Base model. Easy to implement.

## Random Forest

Non-linear relationship.

## Gradient Boosting

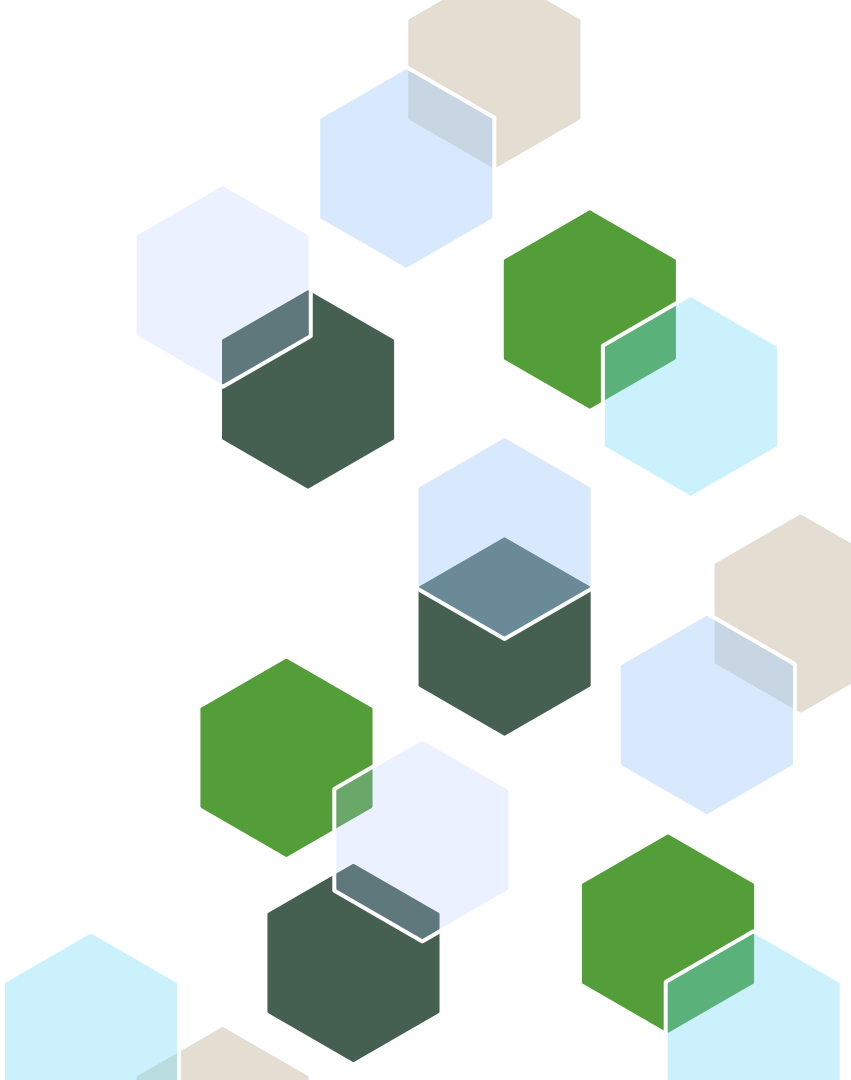
Train on previous errors. Better performance.

## Decision Tree

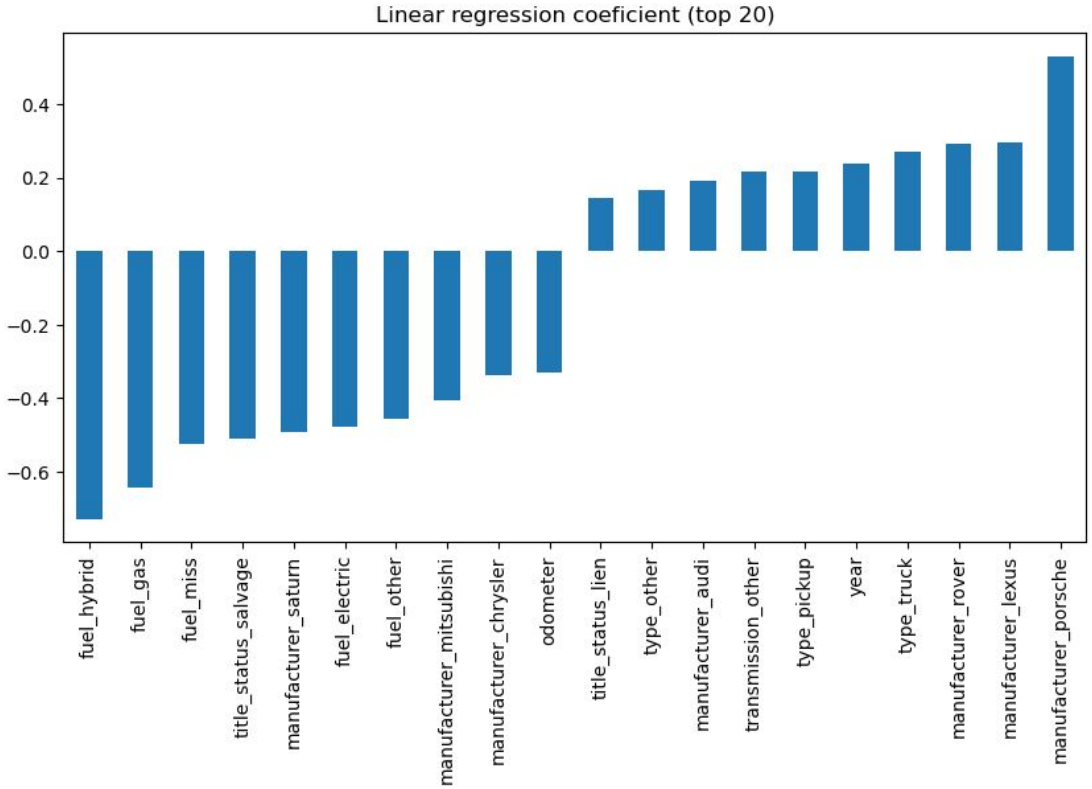
Explainable.

# 04

# Modeling



# Linear Regression



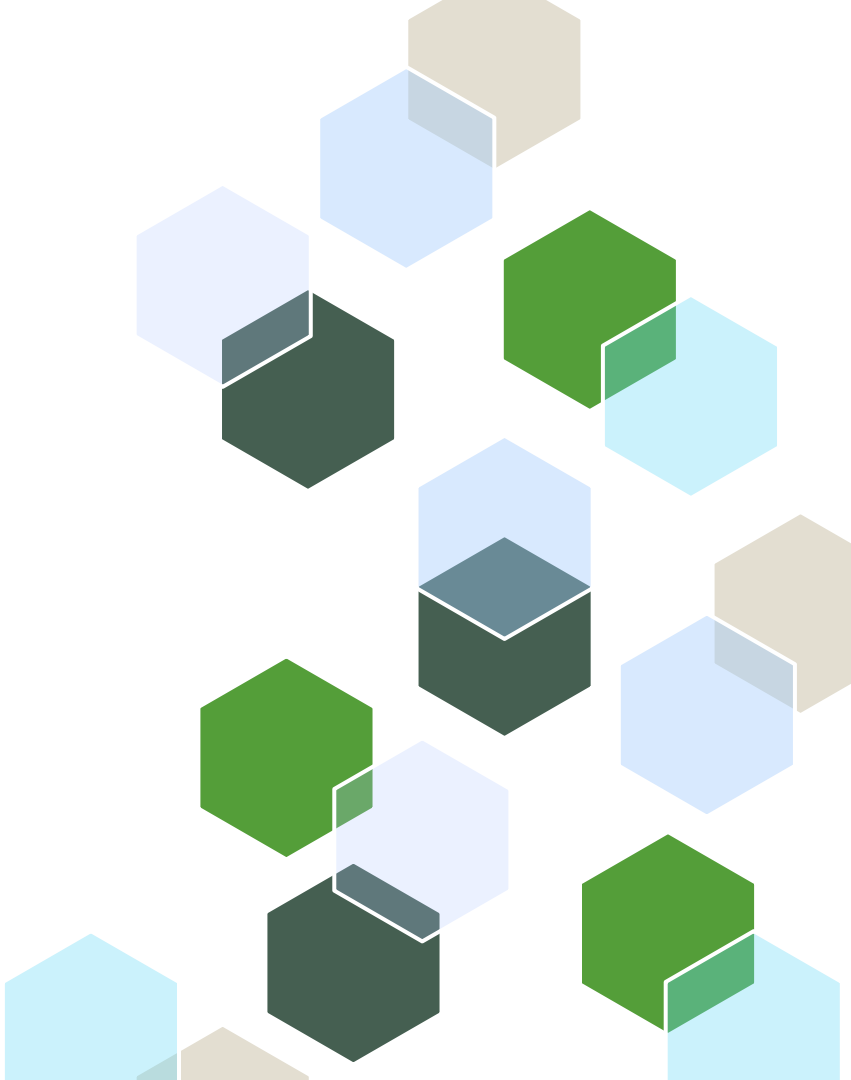
# Grid Search

- Evaluated on R-square

| Model | Hyper-Parameters | Values |
|---|---|---|
| Random Forest | Number of Estimators | [100, 150, 200] |
| | Max Depth | [7, 9, 11] |
| Gradient Boosting | Loss Function | ['squared_loss', 'absolute_error', 'huber'] |
| | Learning Rate | [0.001, 0.01, 0.1, 0.5] |
| | Max Depth | [7, 9, 11] |
| Decision Tree | Max Features | [3, 4, 5, ..., 21] |
| | Max Depth | [1, 2, ..., 6] |

# 05

# Results



# Experiment Set-up

## Cleaning

**Dropped the outliers**; imputed ***mean*** for missing values.

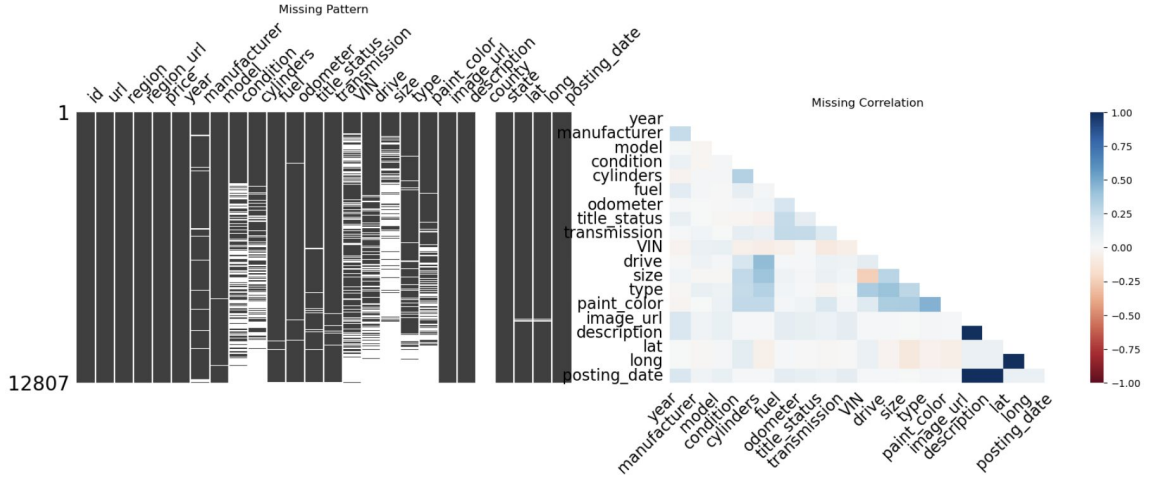
## Transformation

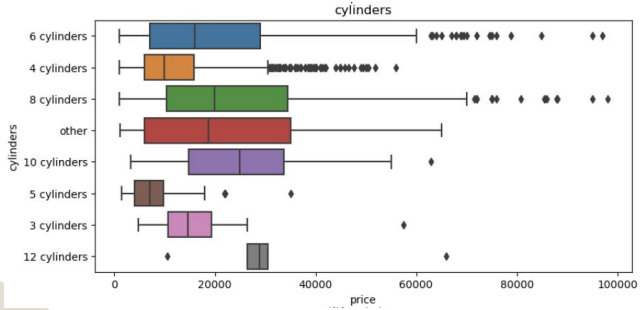
***standardized*** numerical variables;
***one-hot encoded*** categorical variables;
Log-transformed the target variable

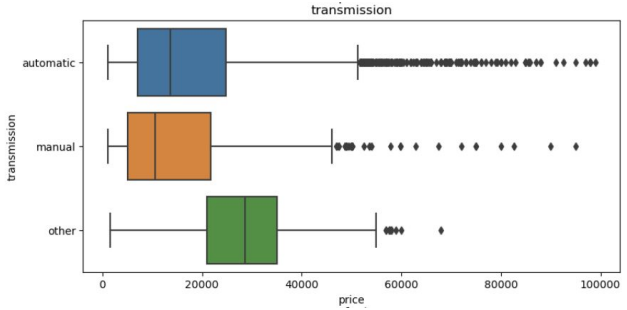
## Feature Selection

removed unnecessary features and only kept 12 features: year, manufacturer, condition, cylinders, fuel, odometer, transmission, drive, size, type, paint color, state, and title_status.

# Experiment Set-up



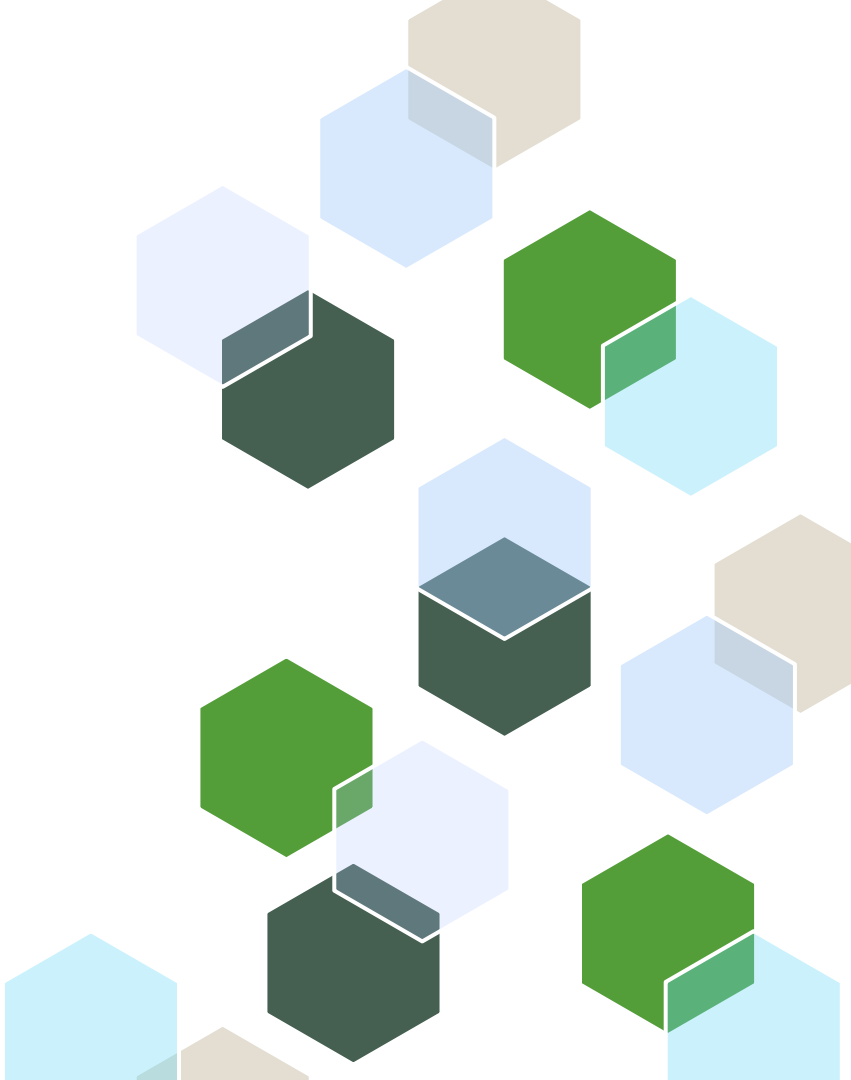



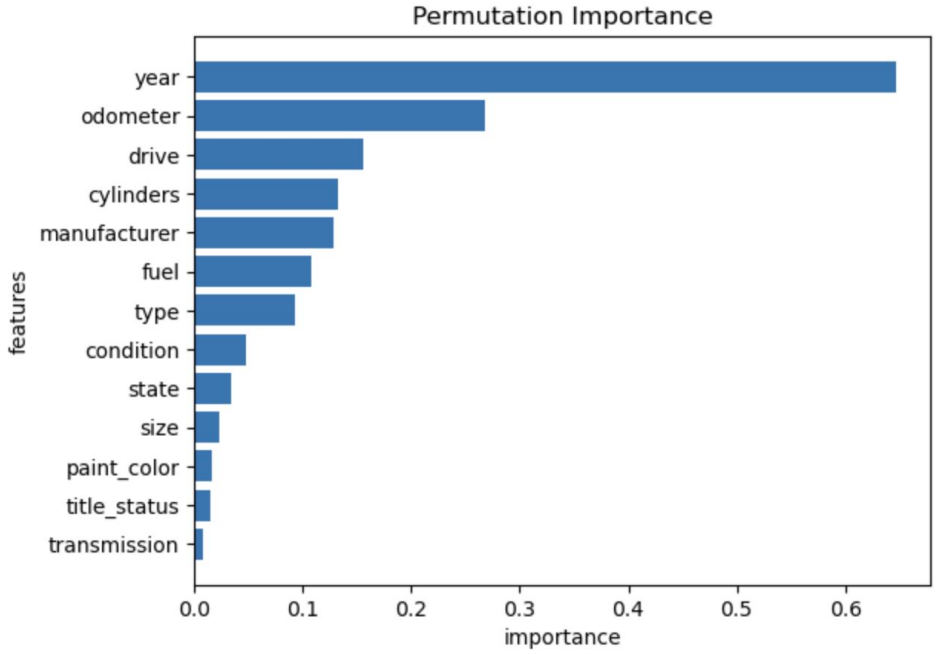
# Model Performance

| Model | Hyper-Parameters | R-square |
|---|---|---|
| Linear Regression | - | 0.57 |
| Decision Tree | Max_depth: 5 | 0.58 |
| Random Forest | Number of Estimators :100<br>Max_depth: 9 | 0.74 |
| Gradient Boosting | Learning rate: 0.1,<br>Loss function: huber,<br>max_depth: 9 | 0.79 |

# 06
# Discussion

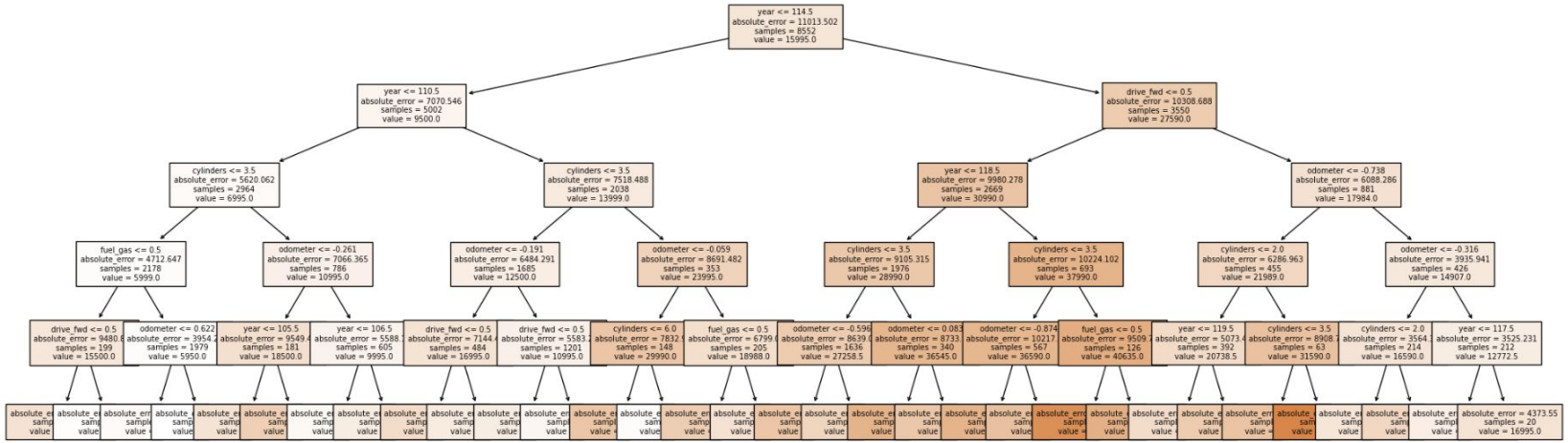


# Trade-off between Accuracy and Interpretability



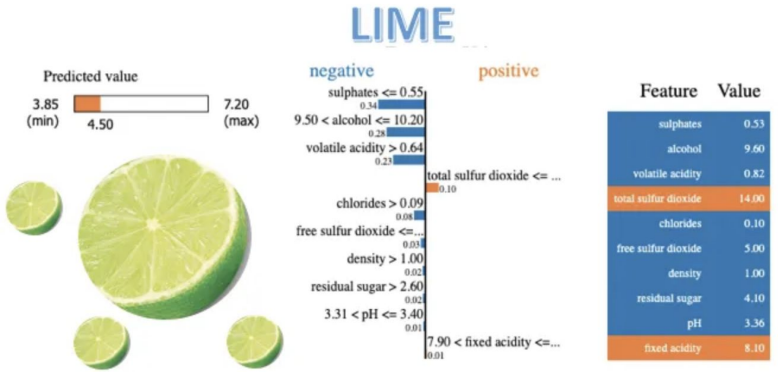

Feature Importance from Gradient-Boosting

# Trade-off between Accuracy and Interpretability

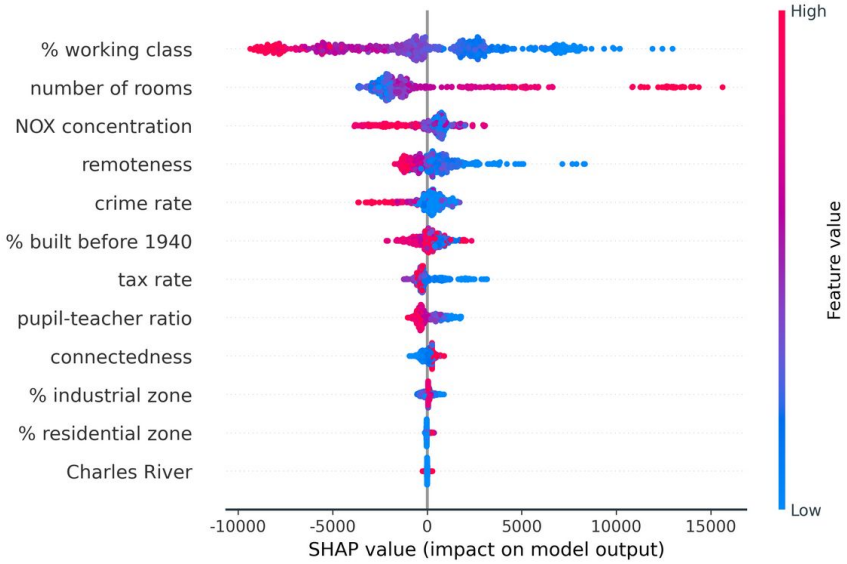



Decision-making process from Decision Tree

# Trade-off between Accuracy and Interpretability




LIME
(Local Interpretable Model-agnostic Explanations)




SHAP
(SHapley Additive Explanations)

# References


- Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189–1232.
- Tin Kam Ho. 1995. Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition, volume 1, pages 278–282. IEEE.
- Hung Truong Thanh Nguyen, Hung Quoc Cao, Khang Vo Thanh Nguyen, and Nguyen Dinh Khoi Pham. 2021. Evaluation of explainable artificial intelligence: Shap, lime, and cam. In Proceedings of the FPT AI Conference, pages 1–6.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011.
- Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12(85):2825–2830.
- Austin Reese. 2021. Used cars dataset. https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data. Online; accessed 15 June 2024.

# Thanks!