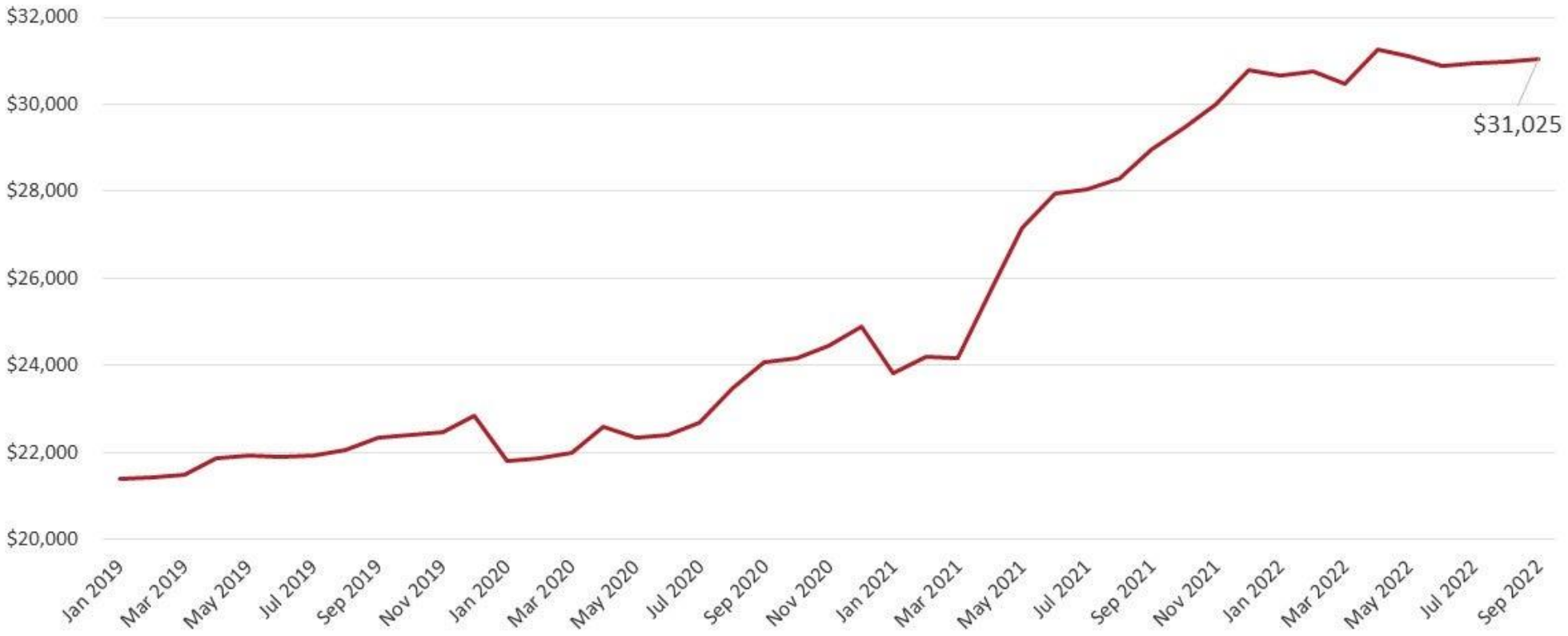
Used Car Price Prediction Modeling

INTRODUCTION

- Over the last 5 years, car prices have experienced major volatility due to inflation and a supply/demand imbalance
- Potential buyers need a tool that can give them an up-to-date fair prices for used cars based on a variety of factors, so that they are not forced to overpay
- Sellers/dealerships need to know what prices are fair so that they are able to operate profitably.

Used Vehicle Average Transaction Price

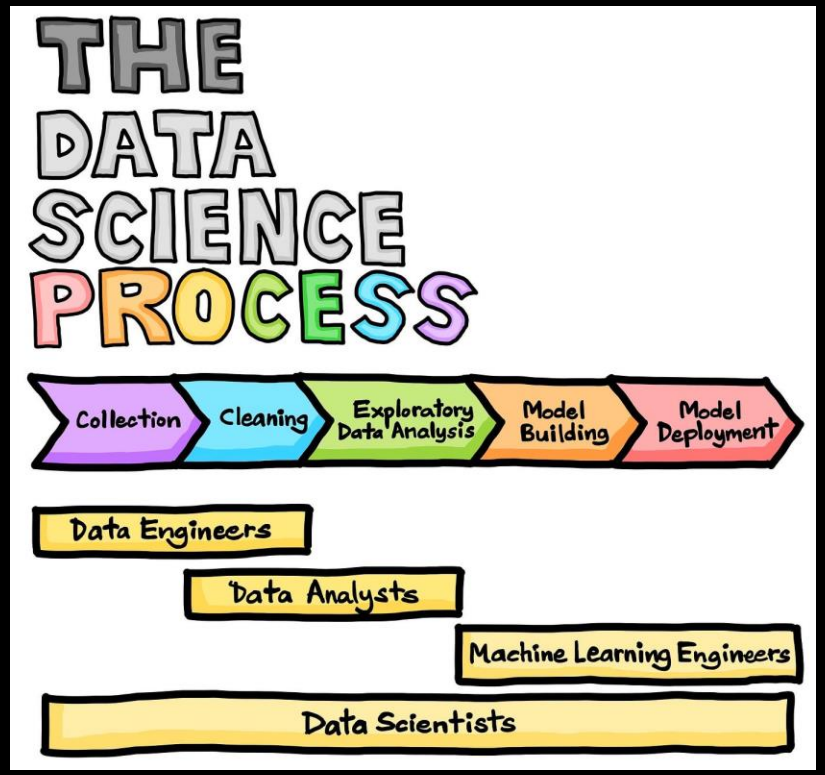


PROJECT GOALS

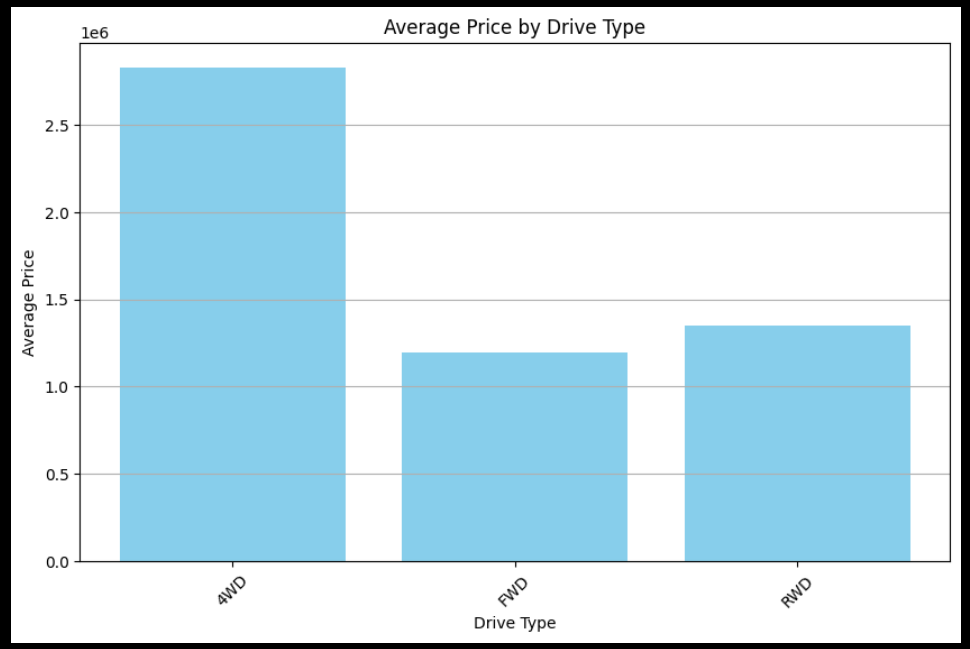
- Import and preprocess used car data, so that we can generate machine learning models
- Visualize and analyze the data before making ML models
- Create machine learning models that can predict the price of a used car
- Determine which model is the best performer and therefore most suitable in this context

PROJECT STRUCTURE

- Data collection
- Data exploration
- Data preparation
- Model training
- Results



DATA EXPLORATION



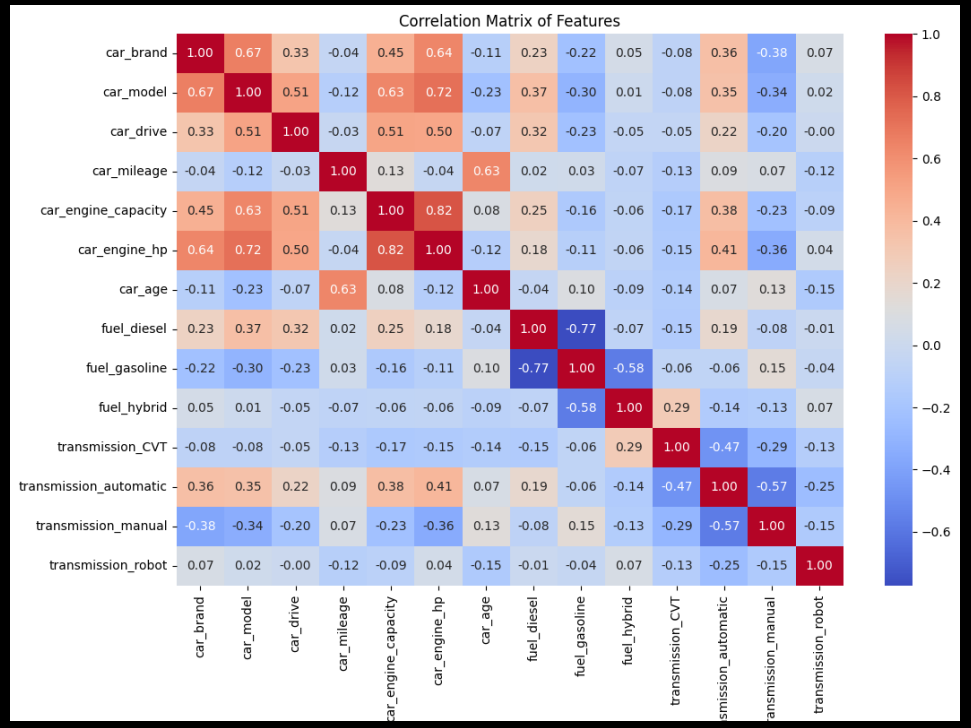
- Summary Statistics were used to understand key metrics
- Visualizing the data showed that most cars fall within the cheaper price range, because most people do not buy high end cars
- Looking at the graphs, we were able to see some potential factors that could result in a high price:
  - Low mileage
  - High horsepower
  - Drivetrain

DATA PREPARATION

- Nulls and duplicate columns were checked for
- Converted the price to U.S Dollars from Rubles
- Columns that we believe would not impact price were removed
- Most categorial columns were one-hot encoded such as
  - Transmission type
  - Fuel type
- Other categorial columns with enormous unique values were target encoded
  - Make
  - model

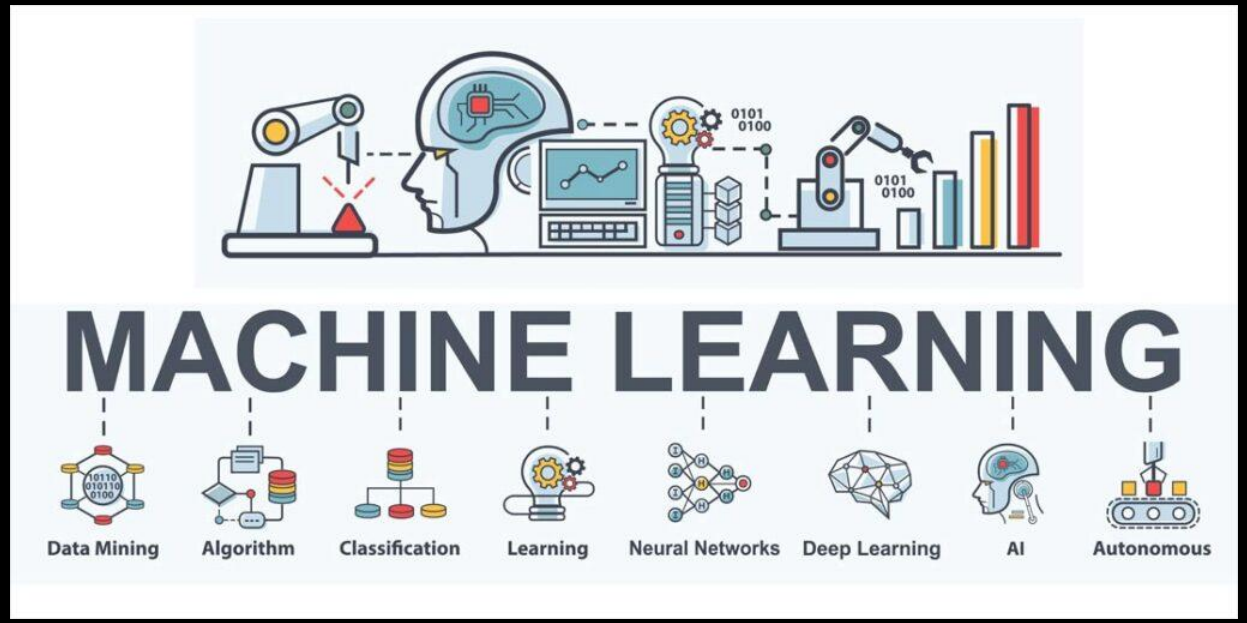
DATA PREPARATION CONTINUED

- Data was divided into testing and training
- Correlation matrix determined that the features seem to be relatively independent of each other
- Data was divided into testing and training
- Feature sets were standardized to prevent magnitude discrepancies from impacting results



ML ALGORITHMS USED

- Linear Regression
- Decision Tree Regression
- Random Forest Regression
- Gradient Boosting Regression



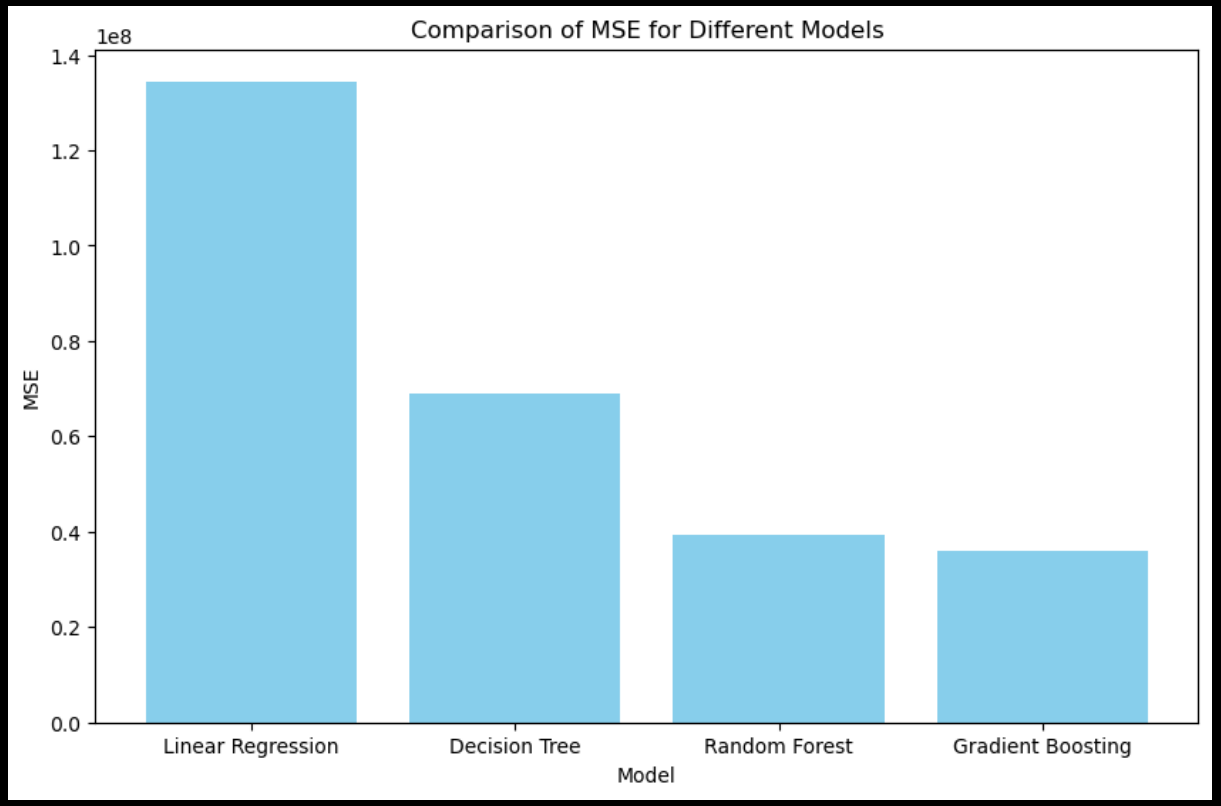
RESULTS

Linear Regression MSE: 134454153.62289524
Linear Regression MAPE: 0.6411544179601025

Decision Tree MSE: 68885951.91219285
Decision Tree MAPE: 0.2217123926240761 3

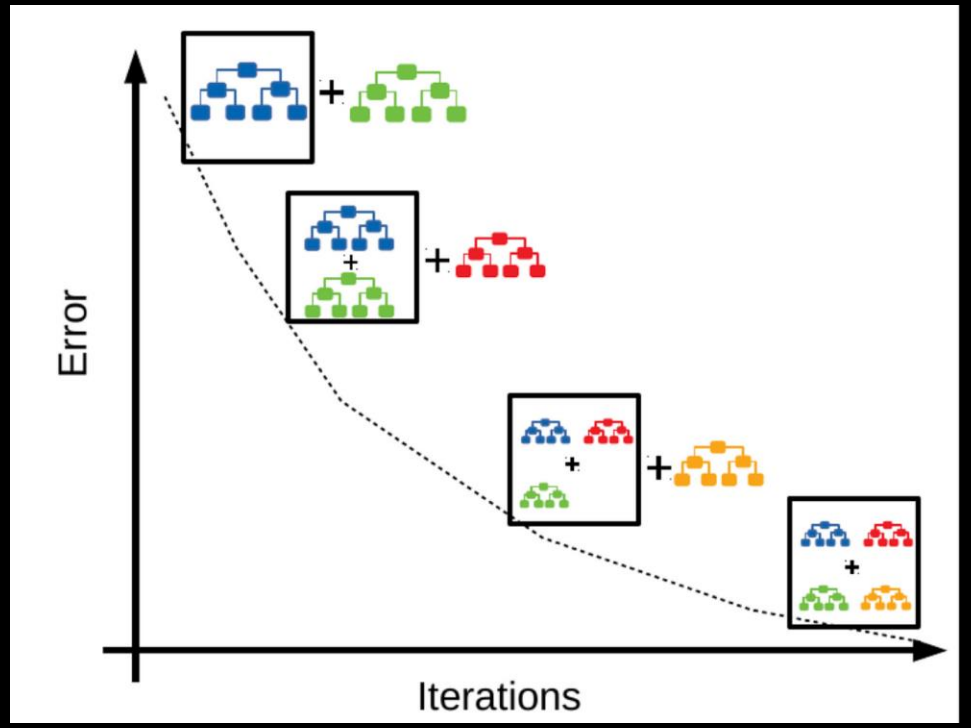
Random Forest MSE: 39245887.139780544
Random Forest MAPE: 0.1746835624727106 6

Gradient Boost MSE: 36020114.505509436
Gradient Boost MAPE: 0.1862963435326717 2

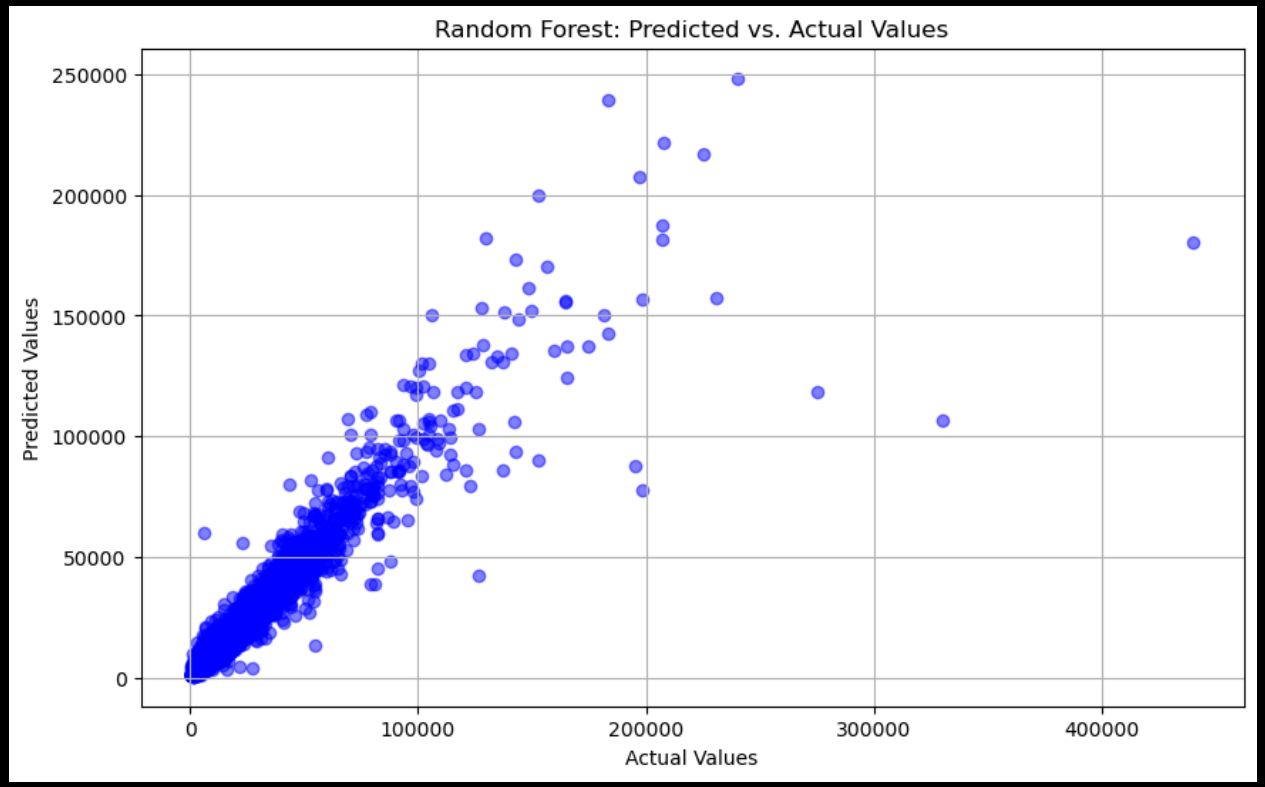


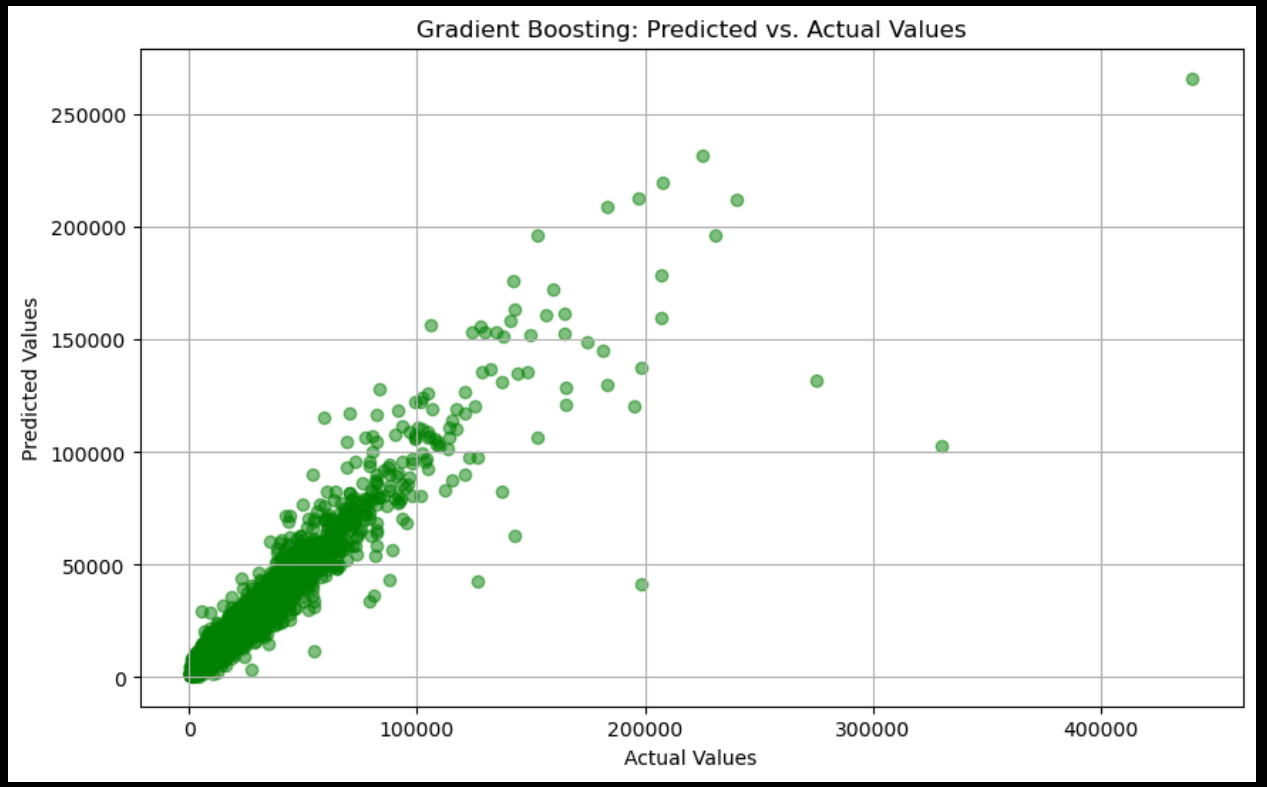
RESULTS

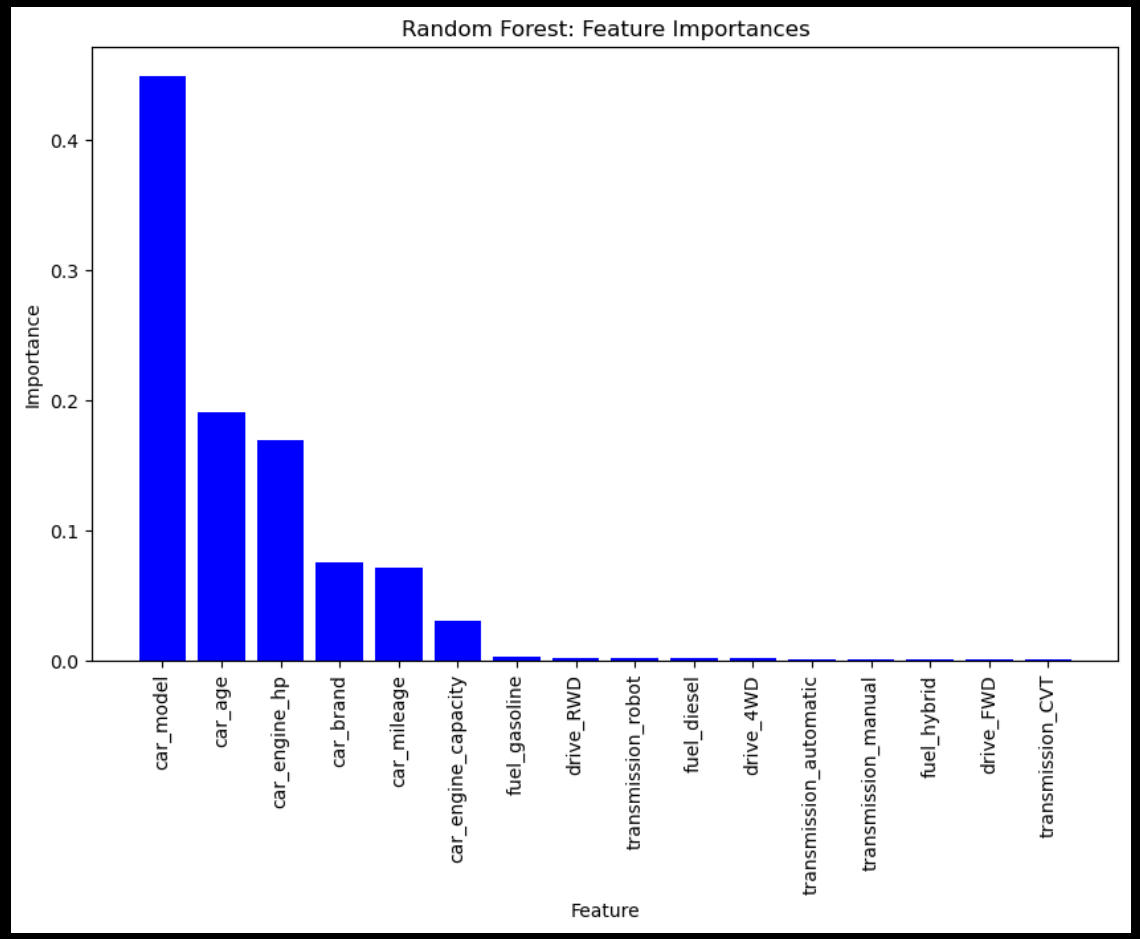
- Gradient Boosted Decision Tree and Random Forest substantially outperformed the other two models
- These models have the ability to adequately balance complexity with contingencies that prevent overfitting

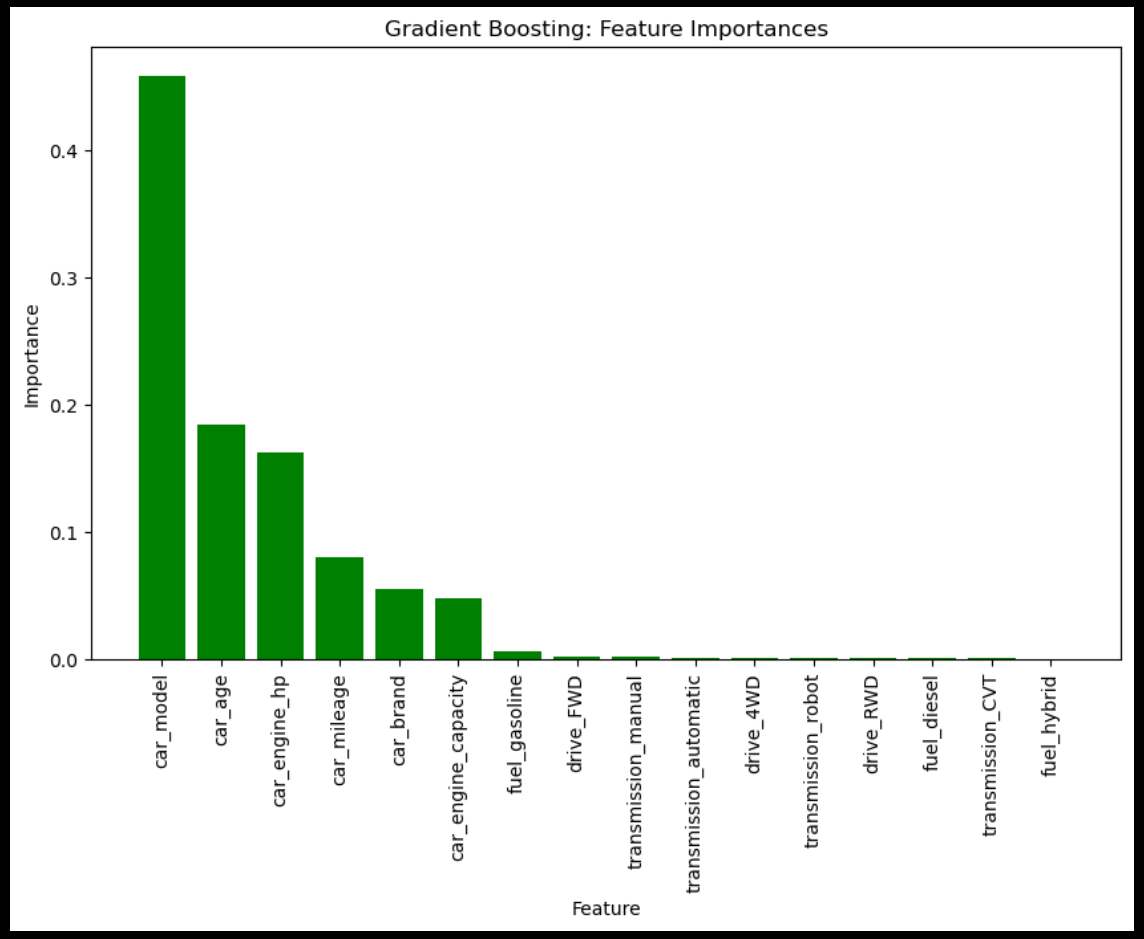


RESULTS CONTINUED






Random Forest: Feature Importances
Importance
0.0
0.1
0.2
0.3
0.4
car_model
car_age
car_engine_hp
car_brand
car_mileage
car_engine_capacity
fuel_gasoline
drive_RWD
transmission_robot
fuel_diesel
drive_4WD
transmission_automatic
transmission_manual
fuel_hybrid
drive_FWD
transmission_CVT
Feature


Gradient Boosting: Feature Importances
Importance
0.0
0.1
0.2
0.3
0.4
car_model
car_age
car_engine_hp
car_mileage
car_brand
car_engine_capacity
fuel_gasoline
drive_FWD
transmission_manual
transmission_automatic
drive_4WD
transmission_robot
drive_RWD
fuel_diesel
transmission_CVT
fuel_hybrid
Feature

CONCLUSION

- Millions of people across the world purchase and sell used cars every year
- This has contributed to an enormous market for these vehicles which involves a complex array of factors which influence their pricing
- We acquired a vast dataset of used cars and their attributes with the goal of constructing models to predict the price of these cars
- Data preprocessing, EDA, and modeling were executed
- GB Regression and RF Regression performed the best
- Models can be useful for the many individuals involved in the used car market

QUESTIONS