Car Price Prediction

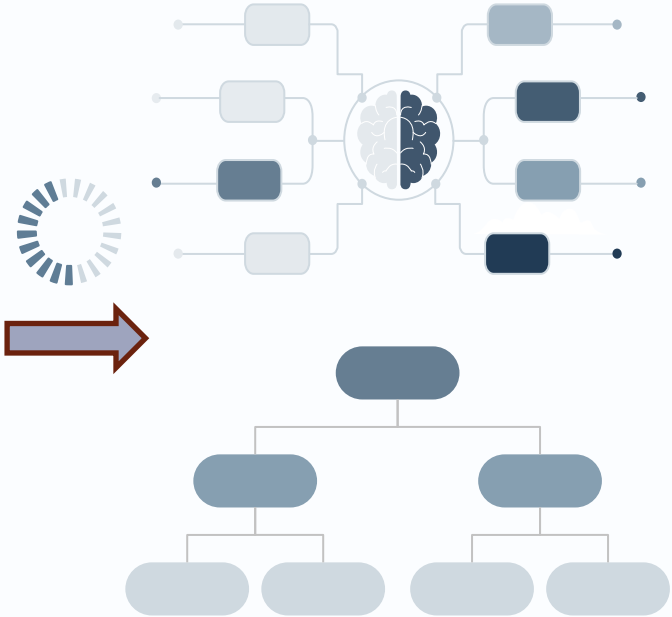
A data-driven approach to predicting car prices using various machine learning models

Presented by Chik Hung, Tang (Ricky)

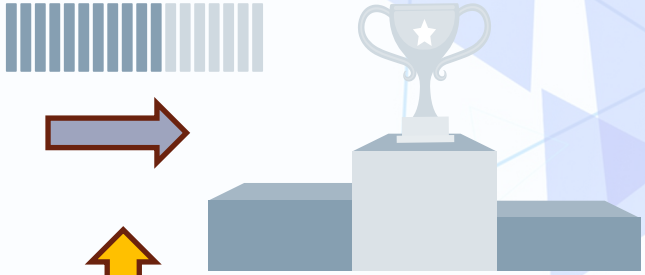
Table of contents

How machine learning works?

Input data



Model selection



Fine tune variables
and hyperparameters

Select the best model

Data Preparation

Duplicate values → Unique values → Null / missing values → Typos → Eliminate reductant column

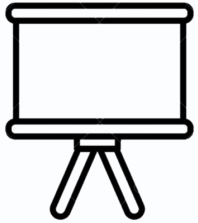
Raw Data

Clean Data

Executive Summary

Introduction

Geely Auto seeks to capture a larger market share in the US and Europe with appropriate pricing strategies



Accuracy

Our project centers on refining machine learning models to predict car prices with a minimum adjusted R-squared of 0.8 to ensure reliable and accurate predictions

Strategies

Perform data cleaning to eliminate missing and duplicate values, and analyze the results using various regression models

Assumptions

Data Authenticity

- Data source is reliable

Consistent Variables

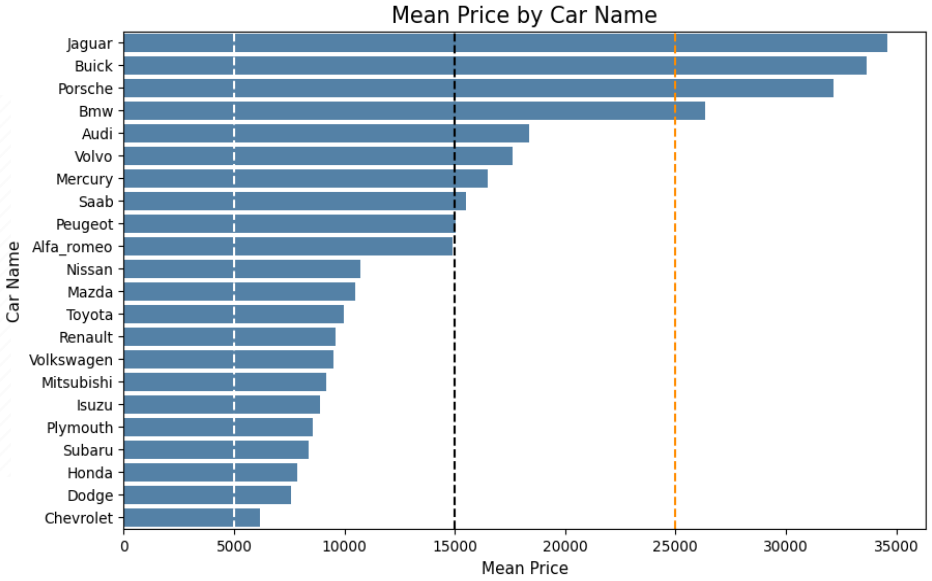
- Vehicle features and price do not across various periods or durations
- Avoid time confusion

Population Representatives

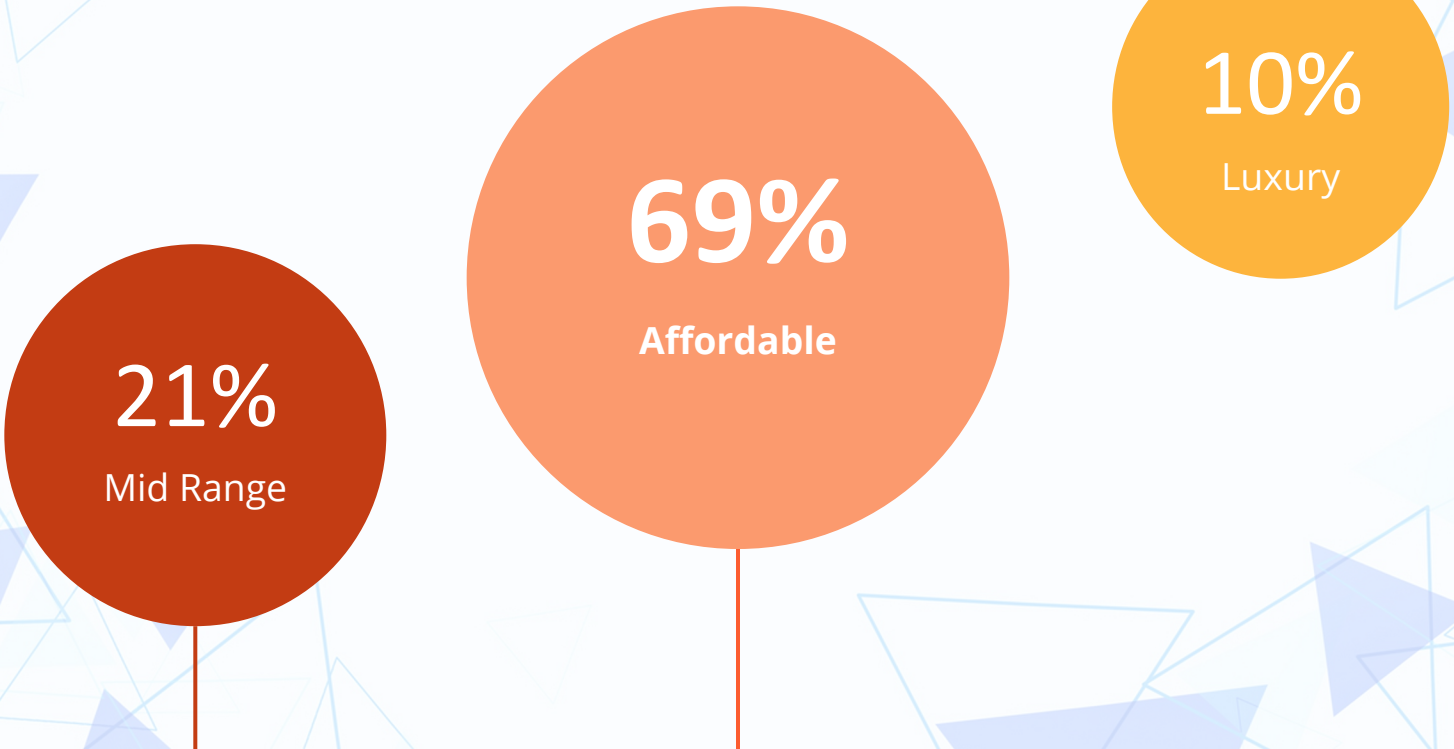
- Random samples
- A diverse range of observations from the larger population

Feature Engineering

Derived Variable	→	CarClass
$5,000 - $15,000		Affordable
$15,000 - $25,000		Mid Range
$25,000 or above		Luxury



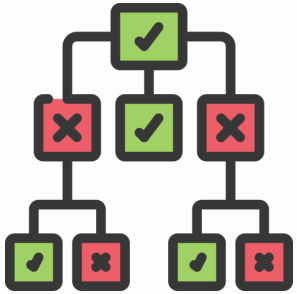
Car Class Percentages



Modelling

Linear Regression

Establish a linear relationship between car features and prices for prediction purposes



Decision Trees

A tree-like structure that splits data based on feature conditions to perform prediction tasks

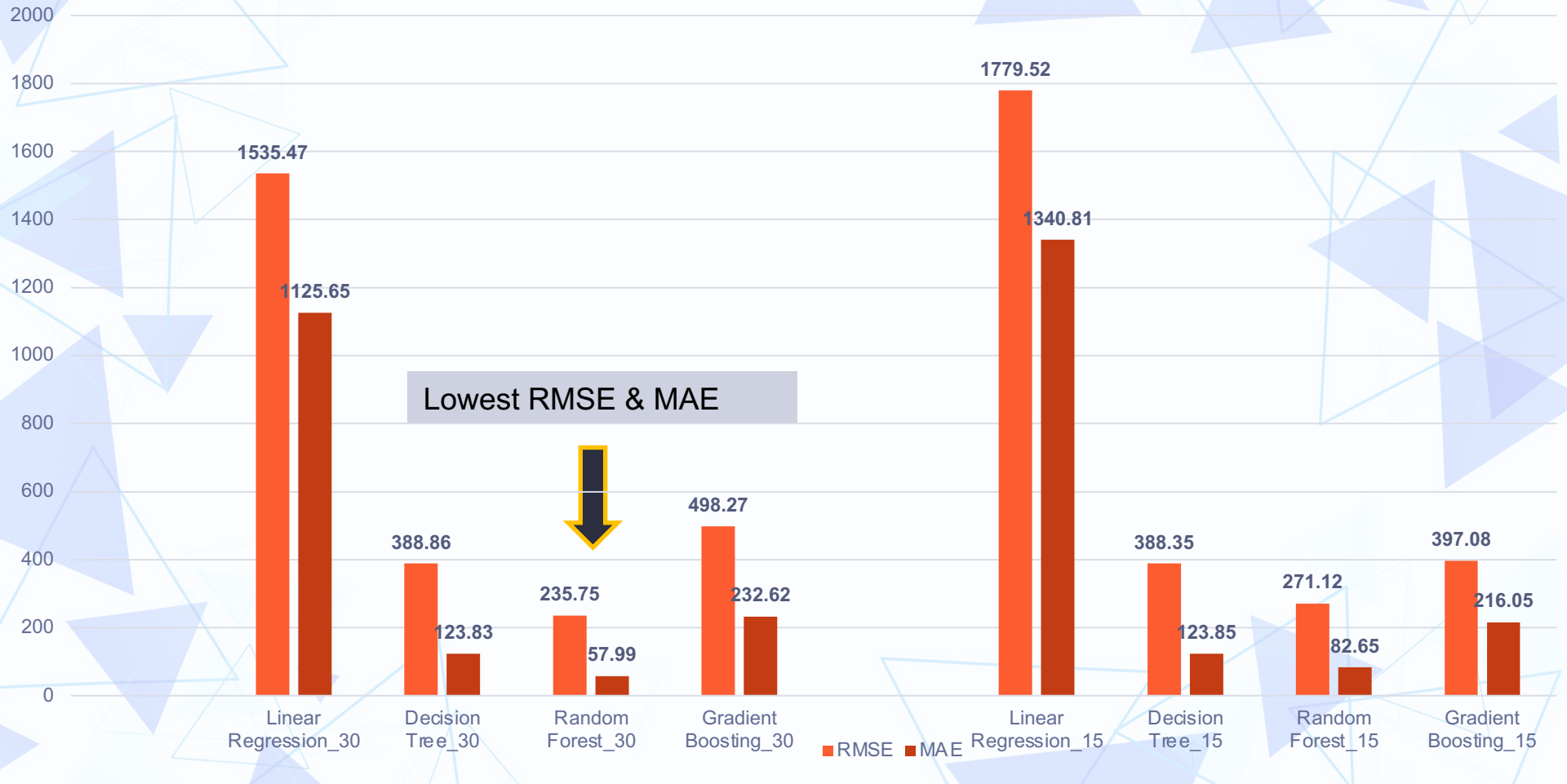
Random Forest

An advanced multiple decision trees approach to improve price prediction accuracy and reduce overfitting

Gradient Boosting

Combine weak features together to refine prediction accuracy through minimize errors

Model Performance



Model Findings

Model Fit Statistics

Root Mean Squared Error (RMSE)

The squared differences between the predicted and actual car prices from the random forest model have an average error of $241.64 dollars

Mean Absolute Error (MAE)

The predicted car prices are off by $57.34 compared to the actual car prices on average

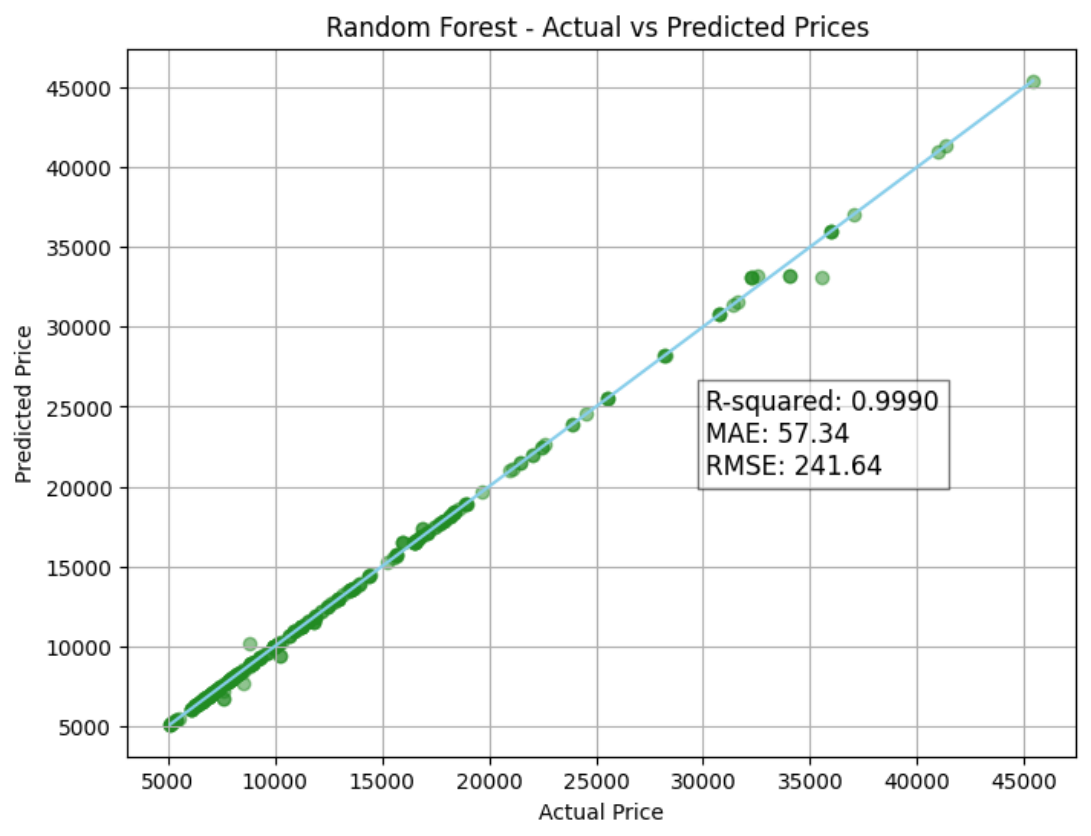
Adjusted R-Squared

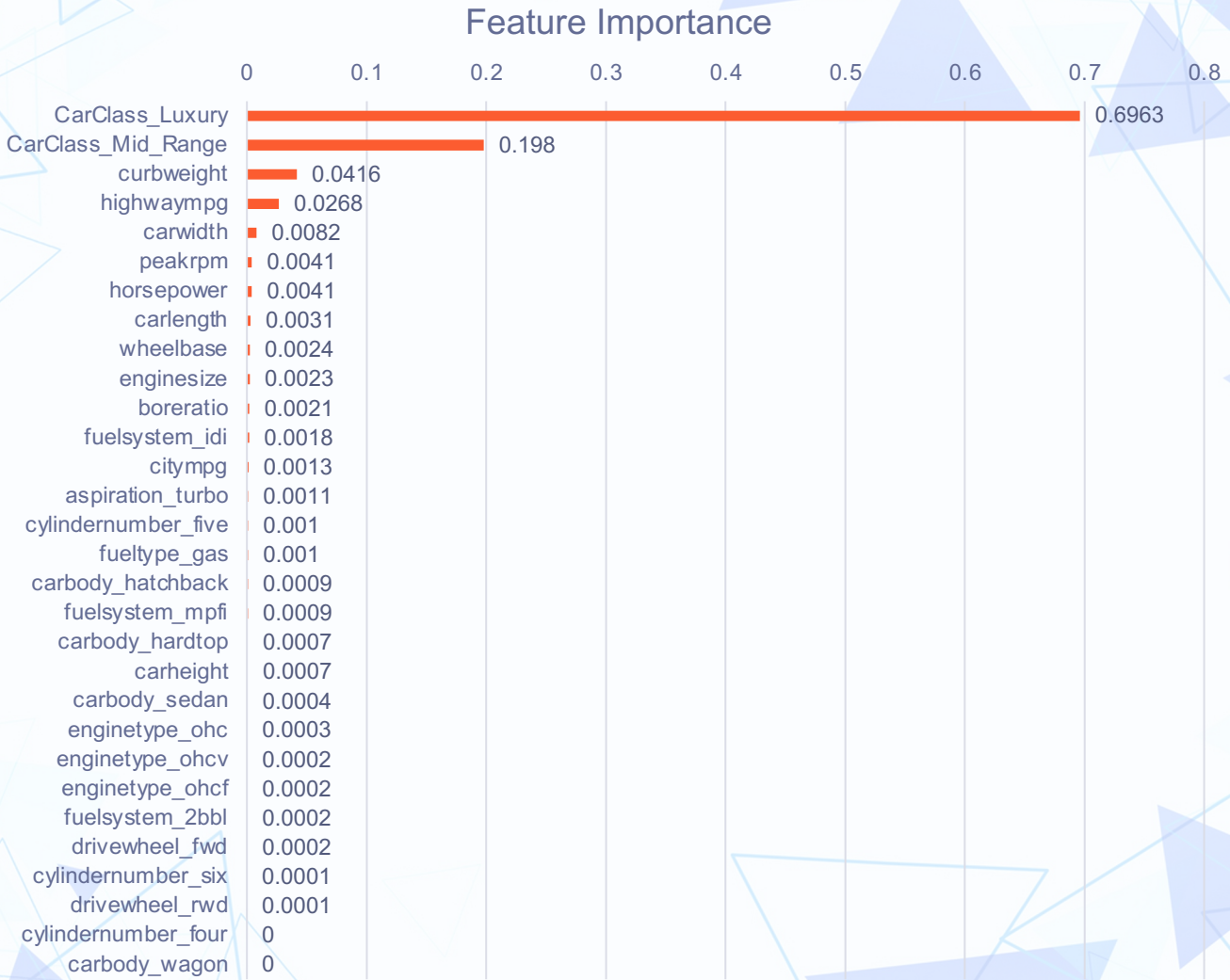
99.8% of the data points in car prices can be explained by the variables used in the final random forest model

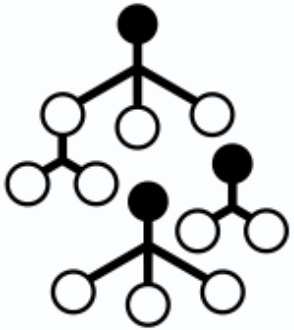
RMSE vs MAE?

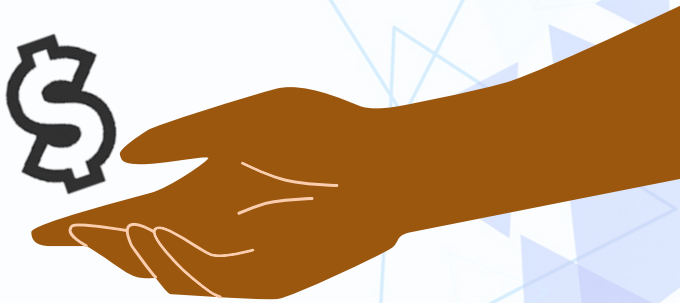
- RMSE gives more weight and is more sensitive to outliers
- MAE treats all errors equally (less sensitive to outliers)
- RMSE would be a better indicator if the business goal is to emphasize the impact of larger errors on the model performance

Actual vs Predicted Prices




Feature Importance
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
CarClass_Luxury
0.6963
CarClass_Mid_Range
0.198
curbweight
0.0416
highwaympg
0.0268
carwidth
0.0082
peakrpm
0.0041
horsepower
0.0041
carlength
0.0031
wheelbase
0.0024
enginesize
0.0023
boreratio
0.0021
fuelsystem_idi
0.0018
citympg
0.0013
aspiration_turbo
0.0011
cylindernumber_five
0.001
fueltype_gas
0.001
carbody_hatchback
0.0009
fuelsystem_mpfi
0.0009
carbody_hardtop
0.0007
carheight
0.0007
carbody_sedan
0.0004
enginetype_ohc
0.0003
enginetype_ohcv
0.0002
enginetype_ohcf
0.0002
fuelsystem_2bbl
0.0002
drivewheel_fwd
0.0002
cylindernumber_six
0.0001
drivewheel_rwd
0.0001
cylindernumber_four
0
carbody_wagon
0





Random Forest with 30 Features

Business Recommendations

Focus on Luxury and Mid-Range classes

These two features have the highest importance, indicating they significantly impact the car prices in the model performance

Increase car weight

Heavier cars tend to have a positive correlation with higher prices

Fuel efficiency

Fuel efficiency also implies city and highway miles per gallon (MPG), peakrpm, and horsepower have moderate importance in predicting accurate car prices.

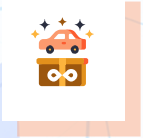
Actions

- Geely Auto should promote and target foreign automotive with their luxury and mid-range vehicles

- Consider exporting those cars to the US and European markets, emphasizing heavier weight and better fuel efficiency for both city and highway driving

- Avoid entering the foreign market with vehicles that have *four* cylinders and belong to the *wagon* car type

Conclusion



Model Optimization

Geely Auto can position its luxury and mid-range vehicles for maximum potential profit in the US and European markets

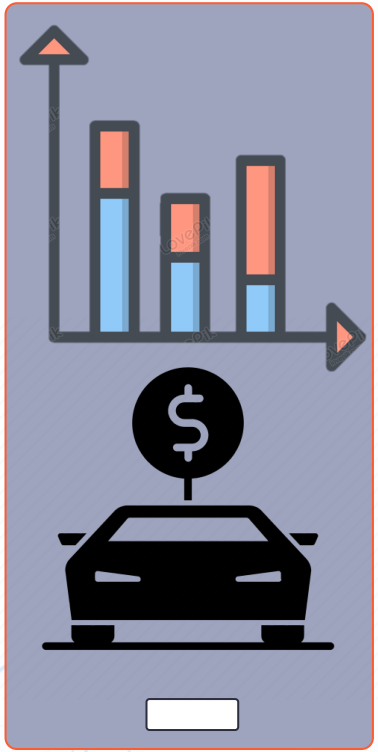
Accuracy Monitoring

Frequent monitoring to ensure the best model continue to perform with acceptable RMSE, MAE, and adjusted R-squared

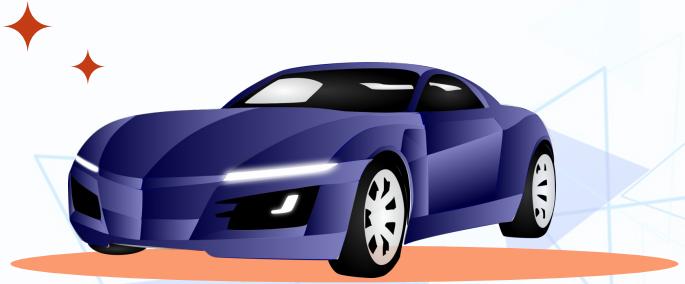
Develop Pricing Strategies

Make use of selected features with high importance to make informed pricing and generate business insights in the foreign market

Thank you!



Q&A Session



Appendix

Potential Questions:

Q: How can your model’s insights impact Geely's pricing and marketing strategies?
A: Geely can strategically position its vehicles in foreign markets to maximize profits and competitiveness based on the key variables.

Q: How did you come up with the recommendation to promote luxury and mid-range vehicles in the US and European markets?
A: My insights are based on my feature importance analysis, where luxury and mid-range classes were identified as the top two significant variables. This demonstrates their strong impact on car prices, and focusing on these classes enables our learning model to predict prices efficiently.

Q: How did you handle the outliers in your dataset?
A: I kept them just to represent the real-world variability and preserve this information as accurately as possible. Removing them might enhance the model’s performance, but it could potentially introduce data bias in the analysis.

Q: How do you make sure the model remains up-to-date with fluctuating market conditions nowadays?
A: Model monitoring and drift detection are implemented to regularly measure the model's performance against real-time data and ready to take appropriate actions whenever necessary. This ensures that the variables within the model does not go beyond the acceptable range of minimum and maximum over time.