The Price is Right:
Predicting Car Values with Machine Learning

- ❏ RYAN
- ❏ JOSÉ
- ❏ LOUIS
- ❏ GIUSEPPE

Introduction

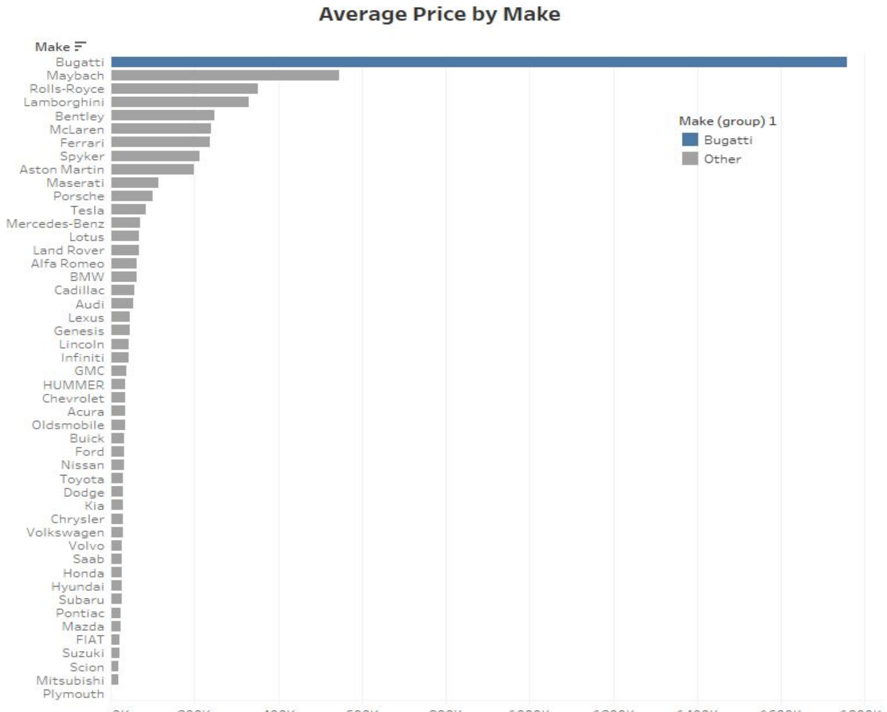
As a collaborative team of four car resellers established in 1990, we utilize data collection to gain a deeper understanding of average car sales prices within the market.

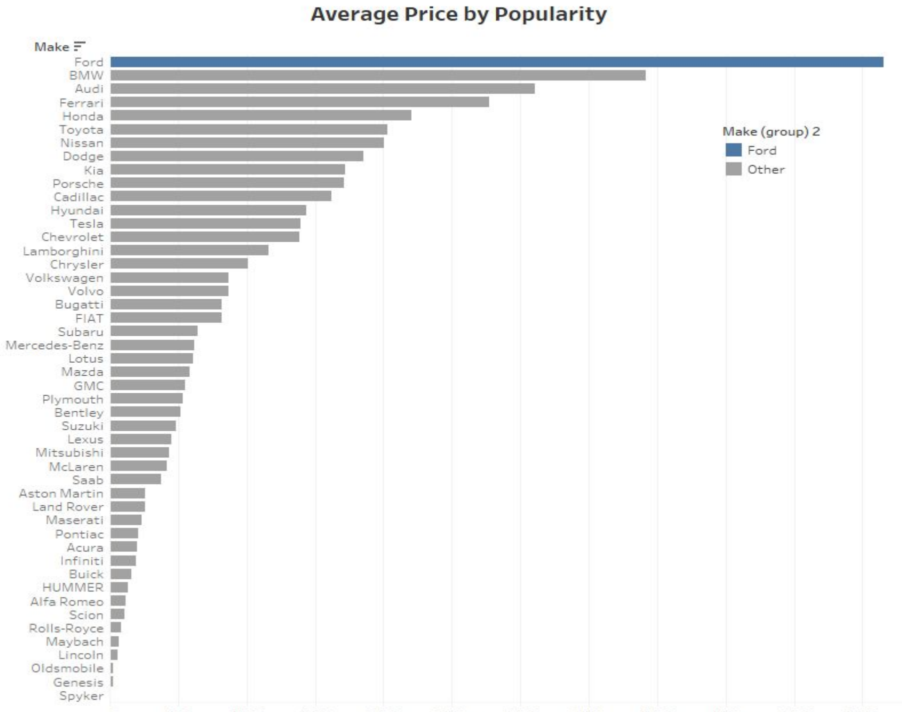
Project Scope :

Our aim in collecting this data is to build accurate car price predictions using advanced Machine Learning models and methodologies.

Data overview & Insights

An examination of these charts suggests that Bugatti cars comes with the highest average price point, while Ford emerges as the most prevalent brand within the dataset.





Data selection

This project utilizes a dataset consisting of various car features (e.g., brand, model, engine specs, and car specs) and corresponding pricing information. The goal is to develop a machine learning model that can accurately predict car prices based on these features.

- **8,172** cars in our dataset

Data cleaning

Null Values

- Dropped row with missing value in ‘Market Category’;
- Replaced missing values with mean and mode;

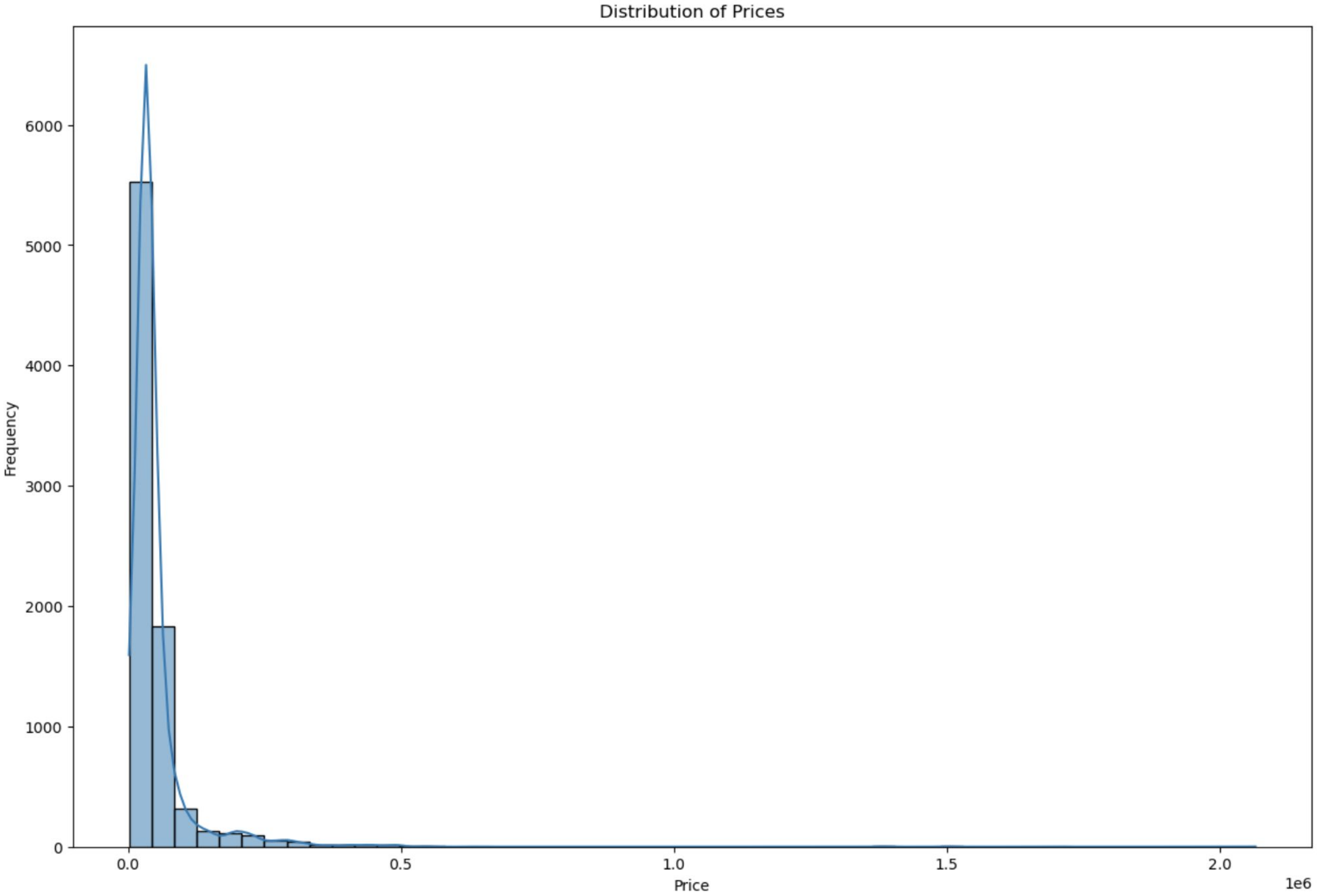
Column Names

- Renaming columns for consistency;

Feature Engineering

This stage focused on transforming non-numerical variables.

- Ordinal variables
- Rank categorical features
- Normalization (X_train & X_test)
- Standardization (y_train & y_test)



Distribution of Prices
Frequency
Price
1e6
0
1000
2000
3000
4000
5000
6000
0.0
0.5
1.0
1.5
2.0

Feature Engineering

This stage focused on transforming non-numerical variables.

- Ordinal variables
- Rank categorical features
- Normalization (X_train & X_test)
- Standardization (y_train & y_test)

Model Building and Evaluation

Performance Comparison of Different Models

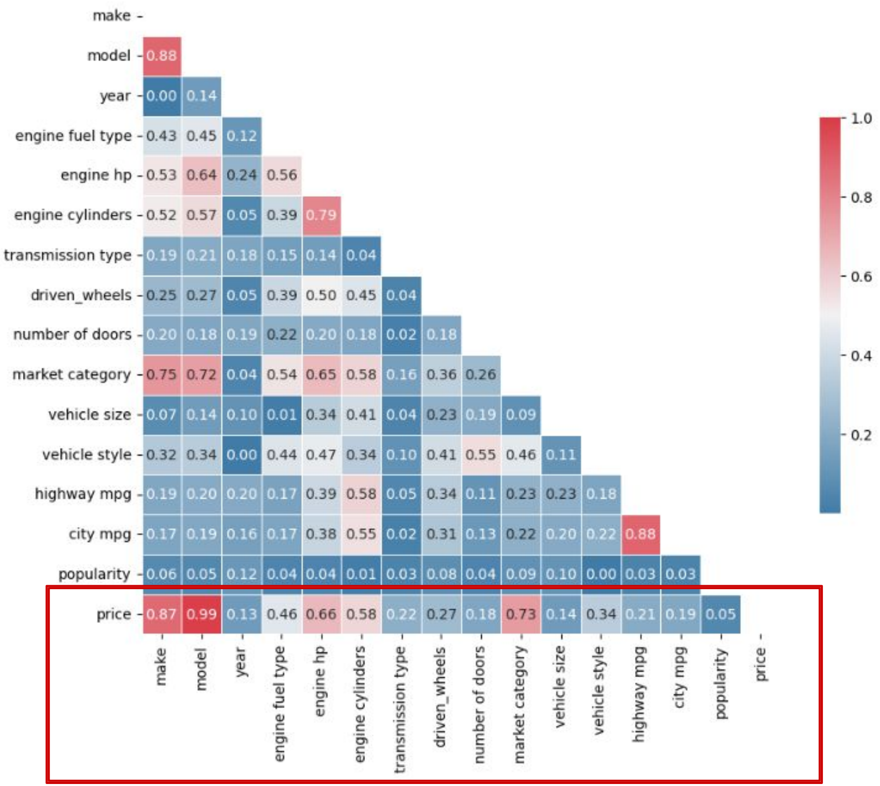
	R2 score	MAE	RMSE
Random Forest	0.9895	51884.22	85320.23
Gradient Boosting	0.9859	52061.92	86794.86
AdaBoost	0.9858	52213.92	86345.13
Bagging & Pasting	0.9855	51255.59	80080.91
Decision Tree	0.9741	19659.09	224732.61
Linear Regression	0.7189	57183.68	272882.6

High outliers, like luxury car prices, can mislead error measures (MAE, RMSE) by inflating the overall error for the typical data.

Random Forest and Gradient Boosting have established themselves as best models to achieve a balance between accuracy and robustness in predictive modeling.

In order to achieve optimal prediction performance, we will be employing the Random Forest approach.

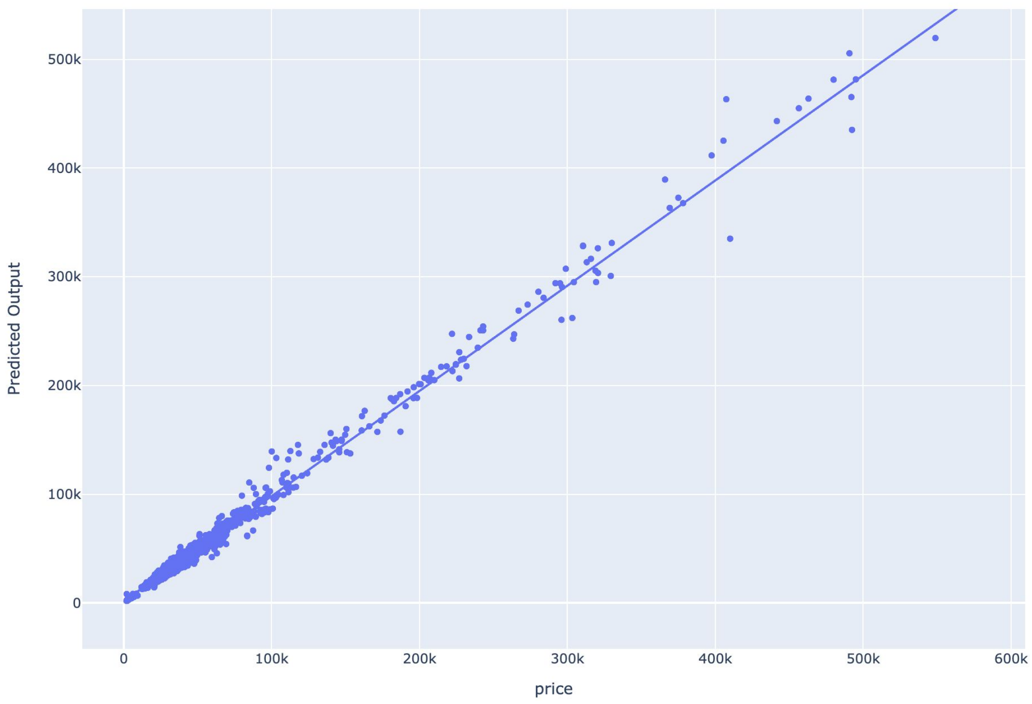
Key Findings and Insights



In our analysis, several key features have emerged as significant predictors of car prices. This features stand out on our model's predictive performance:

- Make and Model
- Market Category
- Engine Horsepower
- Engine Cylinders

Comparison of Predicted Price vs Actual Price



The alignment of points along the diagonal line demonstrates that our Random Forest model accurately predicts prices across a range of observations.

Ex Machina vs Iron Hackers

AUTOMATIC

2 DOORS

2 SERIES COUPE FROM 2017

Vehicle Size: Compact
Market: Luxurious
Drive type: Rear wheel drive
248 of Horsepower
4 engine cylinders

Predicted Price	*28630.92*
Actual Price	*33150.00*

Real-World application and impact

- Manufacturers can identify the optimal price points for new models based on various features and market trends.
- Dealerships can set competitive prices for new and used cars.
- Predict the best-selling models and adjust inventory to meet demand.
- Compare prices of similar models to make informed purchasing decisions.
- R&D departments can focus on features that significantly increase car value.

Challenges and Learnings

- Converting categorical to numerical values
- Normalization Vs Standardization
- Figuring out collaboration on github (branches)


THANK YOU