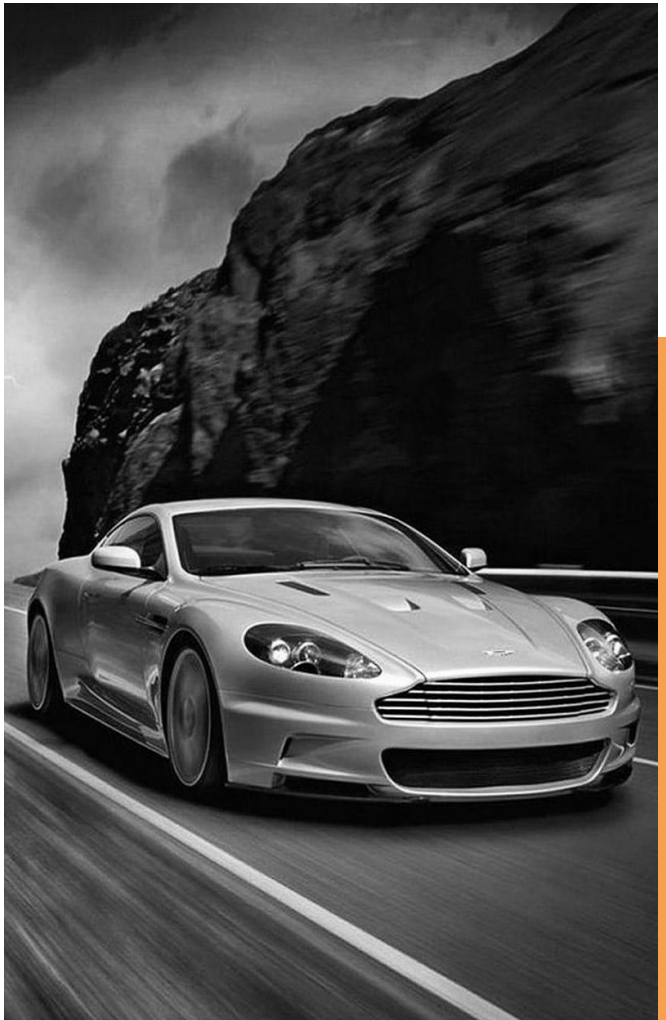
USED CAR DATA ANALYTICS

Capstone Project Presentation

Yoni Kazovsky | Ohm Patel

EXECUTIVE SUMMARY

- **Challenge**: Accurately predict car prices in a large and diverse dataset in order to provide consumers and business with valuable insights into the used car market
- **Approach**: We will use feature engineering and advanced ML algorithms to build predictive models and incorporate them into a user interface allowing consumers and businesses to get meaningful insights with a few clicks of a button
- **Outcome**: Tailored ML models with high performance & an Interactive Shiny app for predictions and dataset navigation.



Introduction

THE DATASET

Data Source

Dataset sourced from Kaggle (~10GB, highly varied data)

Crawler on Cargurus inventory (2020)

Overview

3 million rows, diverse numerical (hp), categorical (color), and textual (description) features.
Predictors: mileage, make, model, year, description
Target: **price**

Initial Hurdles

Opening the dataset

What software to utilize?

Computational power

DATA CLEANING

DROP NA VALUES

Remove all columns with more than 90% missing values, followed by further row dropping based on significance of variable, with others being imputed

Standardizing Values

Simplifying formatting by grouping synonyms (gas, petrol, gasoline), turning everything into lowercase, removing whitespace

Capping Values

We capped price at 150K, mileage at 300K, and age at 30 years, allowing us to handle outliers and trim data into a more manageable size

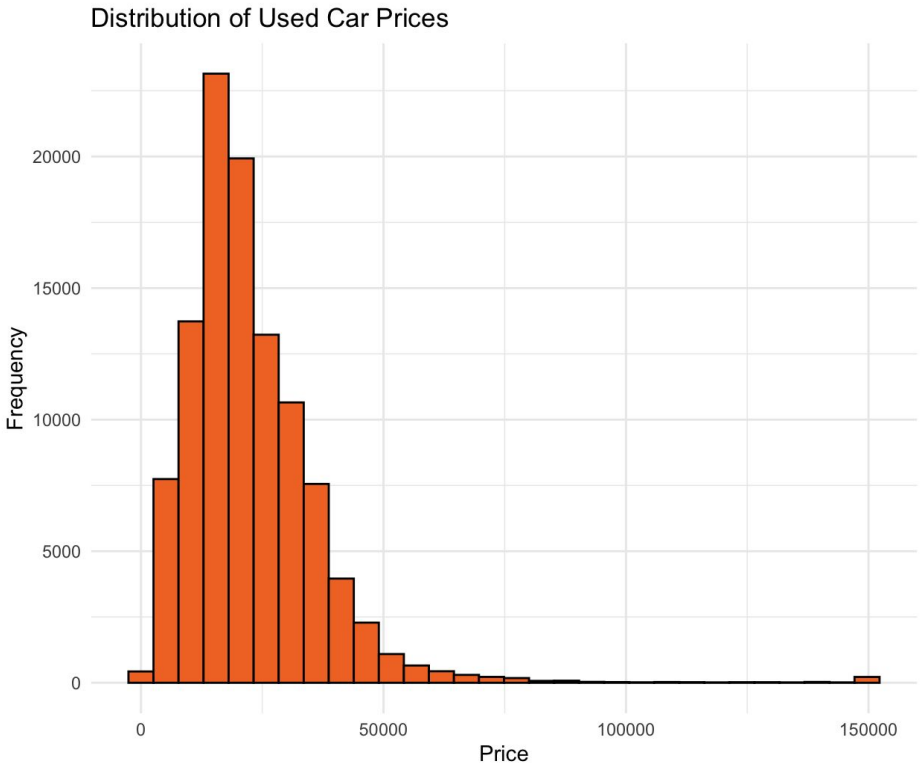
Random Sample

The sheer size of the dataset introduced challenges in relation to the computational power we had access too. Random sampling would allow us to extract significant amounts of information, while being computationally possible for us to build models

EXPLORATORATION (EDA)

The next step was to perform exploratory data analytics to gauge a better understanding of data trends and behaviour.

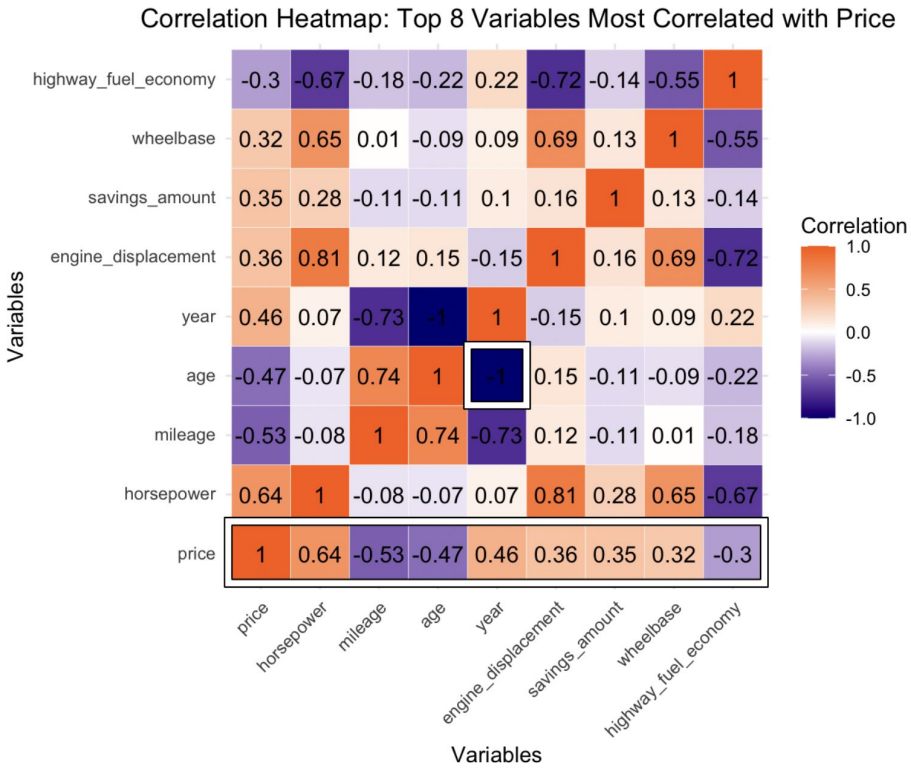
PRICE DISTRIBUTION



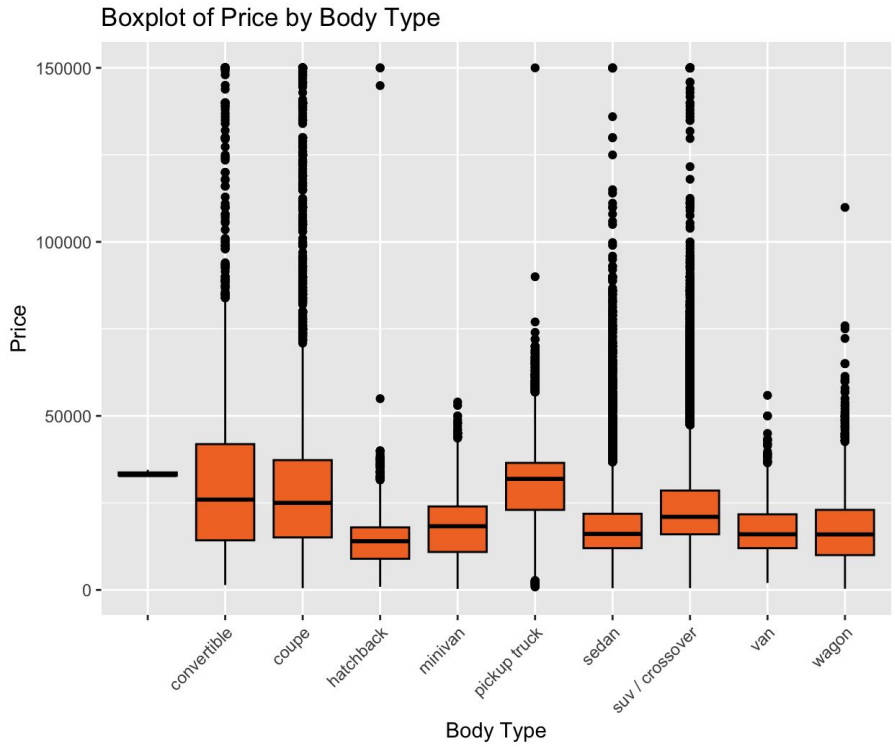
As we can see a vast majority of our used-cars fall between the price ranges of $5,000-$50,000. As a result of this we might see differing trends among vehicles in different prices ranges such as those above $50,000.

CORRELATION MATRIX

. Strongly correlated variables that provide lots of predictive power. The strongest are horsepower, mileage, and age.
. Note that age and year have a correlation of 1 because they are essentially inverses. It is important that our model only incorporates one of them to avoid multicollinearity.



BODY-TYPE TRENDS



As we can see in the box plot. Vehicles such as hatchbacks, minivans, and sedans tend to cost less with a narrower distribution of prices as opposed to coupes and convertibles which have a higher average price alongside a much wider distribution of prices.

FEATURE ENGINEERING

Age

Turning manufacturing date into a simpler variable, “age” by subtracting 2024 from the date of manufacture

Body Types

We had to split up the data into more manageable pieces, and we decided to segment by body type. Pickup Truck, Sedan, and Small/Mid/Large SUV.

Attempted: Color

Region

Instead of relying on longitude, latitude, and dealership zip code, we chose to broaden this variable into four regions.
NE, NW, SE, SW

ML MODELS

Model Type

We decided on RF models due to its ability to capture non-linear relationships. Ideally, we would try XGBoost and GBM, however computationally we believe this was a happy medium of performance/efficiency

Of Models

After numerous attempts to make a single model, we eventually took advantage of the body type segmentation and made separate models. This allowed us to process more data and have better accuracy (specific traits)

Feature Importance

After running the models, out suspicion was verified that there were critical variables that heavily affected price. Age and mileage were the two most influential factors, along with wheelbase, horsepower, and fuel economy

MODEL PERFORMANCE

Model Type	R2	RMSE	Focus On Model
Pickup Truck	.897	5391.82	Reliability, Workload Capability
Sedan	.863	5998.47	Comfort, Urban Usability
Small SUV	.896	3141.84	Efficiency, Compact family utility
Mid Size SUV	.882	5791.95	Balanced Utility and Space
Large SUV	.899	7571.67	Luxury, Towing Capacity, Maximum space

SHINY APP METHODOLOGY

Our goal is to build a shiny app with a used-car price predictor, and dataset traversal features, which can be used by both consumers and business to better gauge the used-car market.

KEY CONSIDERATIONS

Useful Model

Random Forest

Dataset Traversal

Car finder tab

UI/UX

Use of UI/UX principles

Visually Pleasing

Ensuring consistent fonts and visuals

APP DEMO

RESULTS AND INSIGHTS

Price Drivers

Dealerships should focus on key price influencers, such as mileage, age, and highlight unique features of the car, not restating information.

This can also come into play when it comes to inventory management, dealerships should focus on vehicles with high-demand attributes.

Body Type Insights

Understanding how consumers differ with different car types should prompt marketing to have different campaigns based on what consumers prioritize.

This can also relate to a dealerships financing/warranty deals. Offering tailored financing or packages depending on the consumer, and intended use of vehicle could attract more people.

Sentiment Scores

Scoring keywords from descriptions show that buyers value certifications (pre owned, certified, etc.), premium features (leather seats, heated seats), and warranty guarantees. Many descriptions that we saw were restating information already on the listing, we believe that this space would be better used highlighting what consumers ACTUALLY value and what's UNIQUE about the car/dealership.

SUMMARY/NEXT STEPS

Summary

Throughout the process, our team navigated through challenges relating to the size of datasets, and ended up with accurate models that could reliably predict prices of cars. We also made a consumer sided interface where anyone could interact with it (with no previous knowledge) and gain understanding of the car market.

Model Refinement

In the future, we would like to explore using cloud computing to be able to process more data, and we would also like to explore different models, such as XGBoost and GBM. Ideally, we could also integrate external sources (API, etc.) to continuously get recent data about the industry and factor that into our models.

Further Research

Given the rise of EV's and alternative fuel vehicles, we are curious to see if the primary fuel source of the car affects how consumers behave, and what factors affect the price (and by how much). We would also like to dive more into regions and how certain locations might affect consumer behavior, or how weather conditions could affect resale value (rust, etc.)

REFERENCES

- **Automotive Market Trends**
  https://www.statista.com/topics/1986/automotive-industry/
- **Understanding Used Car Depreciation**
  https://www.edmunds.com/car-buying/depreciation.html
- **Used Car Pricing and Analysis**
  https://www.kbb.com/car-prices/
- **Automotive Data and Research**
  https://www.cars.com/research/
- **US Used Car Market Insights**
  https://www.autotrader.com/cars-for-sale