Machine Learning

Capstone Briefing

Automobile Resale Value in India

Joshua W Julian
MIT-PE Applied Data Science
Jan ‘24 Cohort

- Overview
- BLUF
- Approach
- Key Findings & Insights
- Recommendations

Overview

- By volume, used car sales outpace new car sales
  - 3.6m new vs 4.0m used
- Slowdown in new car sales
- Used car sales pricing is uncertain

Proposal

Cars4U focuses on a small market with a specific subset of car features that maximizes resale price and keeps the startup as lean and agile as possible

Car Features

High Power
Low Age
Low KM Driven
Automatic Transmission
Diesel

Locations

Hyderabad
Delhi

Monetization Strategy

Percent Commission based on sale price, focusing on coordinating the higher reselling vehicles

Data Analysis

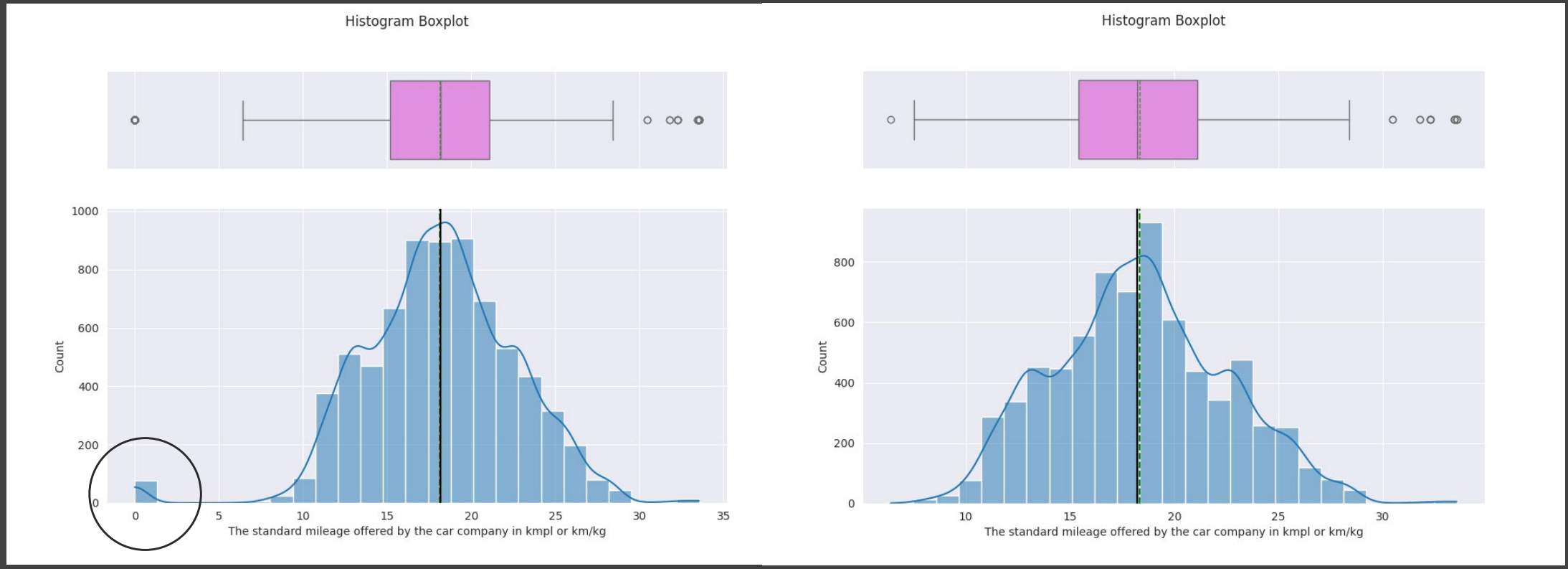
Working with the current historic data set we have from 2020.

Serial Number
Name (Make & Model)
Location of Sale
Year
KM Driven
Fuel Type (Petrol, Diesel, Electric, LPG, CNG)
Transmission Type

Owner #
Mileage
Engine Size
Engine Power
Seats
New Price
Target = Sale Price

Data Cleaning

Example of Bad Input



Before

After

Problem: 0 for mileage
78 of 80 was for Petrol or Diesel
Solution: Median Mileage

Cleaned Data

Simplify the data set while focusing on more important features

~~Serial Number~~

~~Name (~~Make ~~& Model)~~

Location of Sale

~~Year~~ Age (2020)

KM Driven

Fuel Type (Petrol, Diesel, ~~Electric,~~ ~~LPG, CNG~~, Other)

Transmission Type

Owner #

Mileage

Engine Size

Engine Power

Seats

New Price

Target = Log(Sale Price)

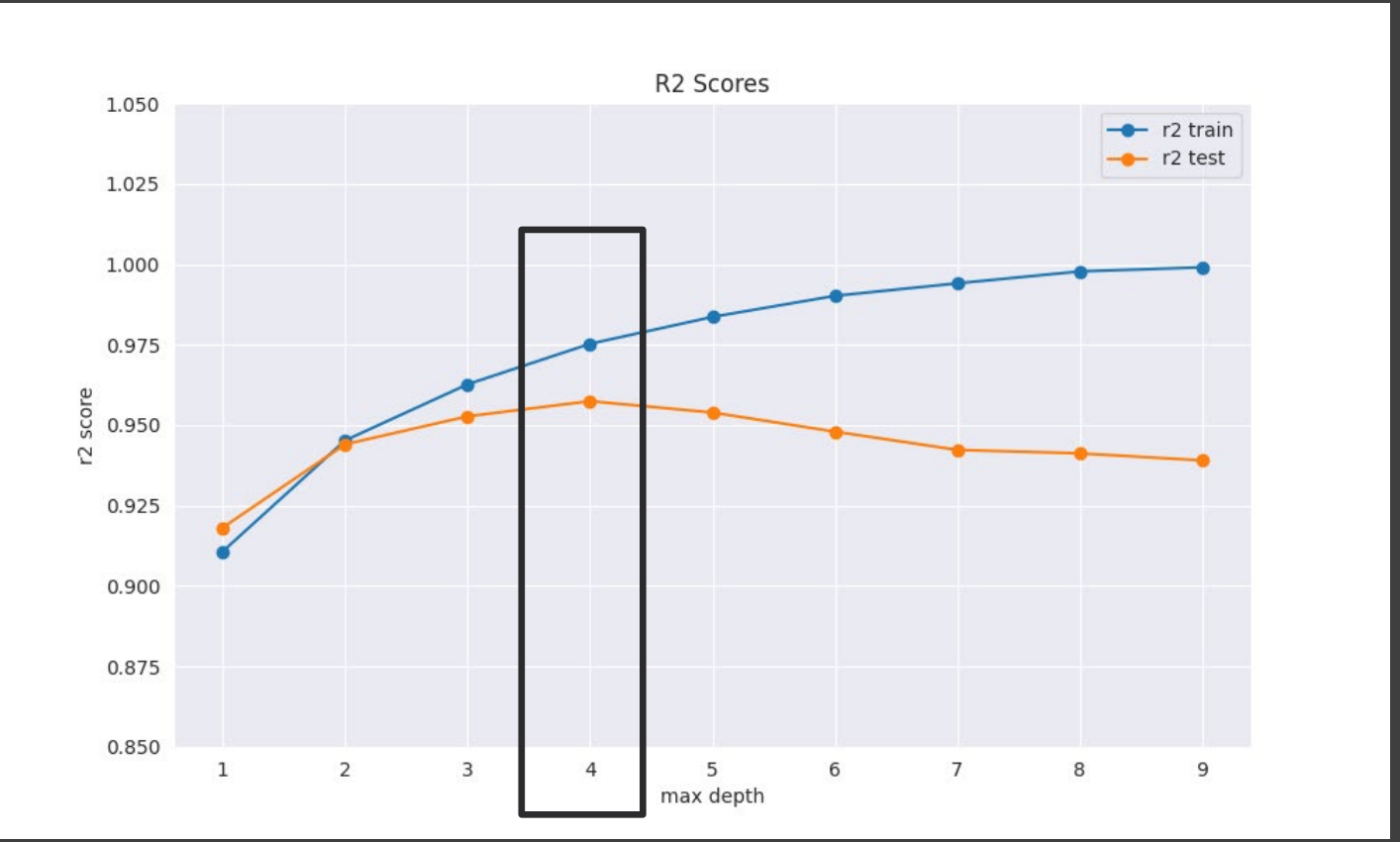
Model Comparison

	MAE		RMSE		MAPE		R^2	
Model	Train	Test	Train	Test	Train	Test	Train	Test
Linear Regression	.19	.18	.25	.52	.25	.21	.91	.91
Ridge Regression	.19	.19	.25	.25	.10	.26	.91	.91
Decision Tree	.21	.23	.28	.32	.30	.31	.89	.85
Gradient Boost	.15	.16	.199	.21	2.65	.21	.94	.94
XGB (depth=4)	.10	.13	.13	.18	.14	.13	.98	.96

MAE/RMSE/MAP – Lower is better | R^2 Higher is better

Hyper parameters where adjusted and tuned to find best fits/compromises for all models.

Model Refinement



Rationale - Highest test data R2 score with smallest split between the train data

Key Findings and Insights

Model Chosen: XG Boost (depth = 4)

Car Features

High Power

Low Age

Low KM Driven

Automatic Transmission

Diesel

Locations

Hyderabad & Delhi

Recommendations

Implementation

- Act as a coordinator for used car sales focusing on specific areas that have the highest resale
- Utilize model to help with pricing + % to help cover the commission
- Market as an effective marketplace for sale of higher quality (higher resale) vehicles

Risk

- Smaller market
- Limited Initial Exposure
- Feature Set might be too condensed

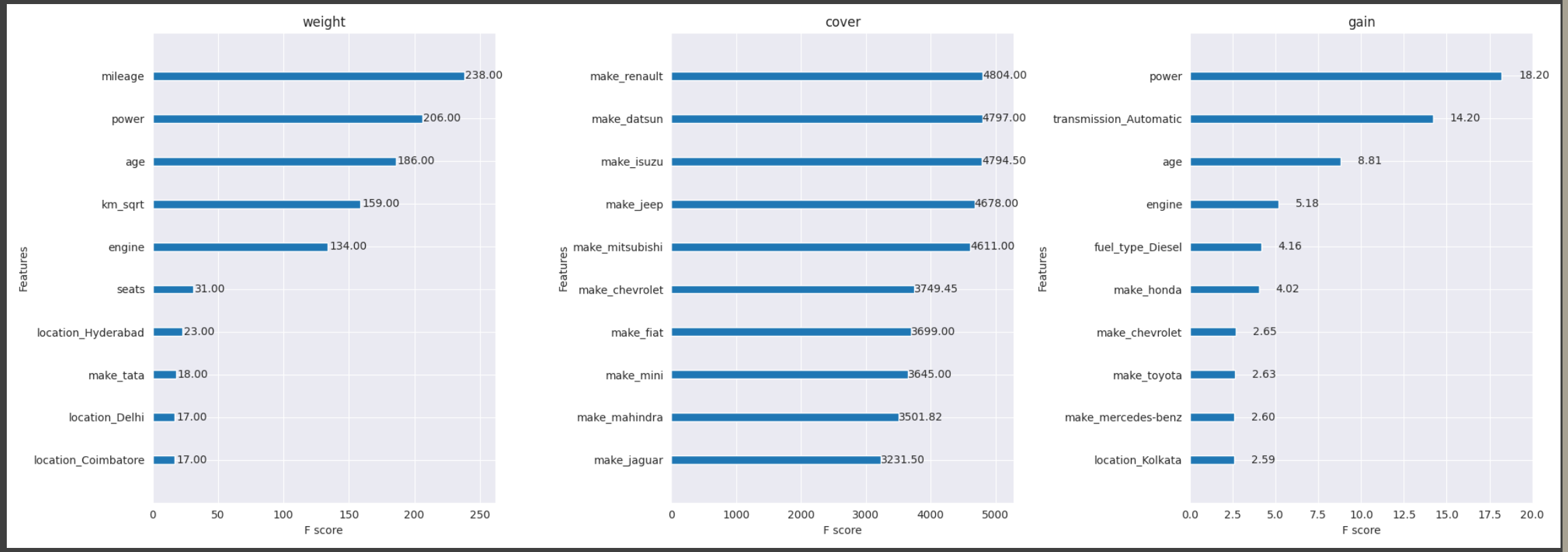
Further Analysis

- Constant refinement of features
- Market may move away from diesel or current metro areas we focus on
- Expand to other metro areas

Questions and Comments

BACKUP SLIDES

XGB - Feature Importance



Error Formulas

$$MAE(y, \hat{y}) = \frac{1}{N}\sum_{i=0}^{N-1} |y_i - \hat{y}_i|$$

$$RMSE(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}}$$

$$MAPE(y, \hat{y}) = \frac{1}{N}\sum_{i=0}^{N-1} \frac{|y_i - \hat{y}_i|}{|y_i|}$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{N}(y_i - \bar{y})^2}$$