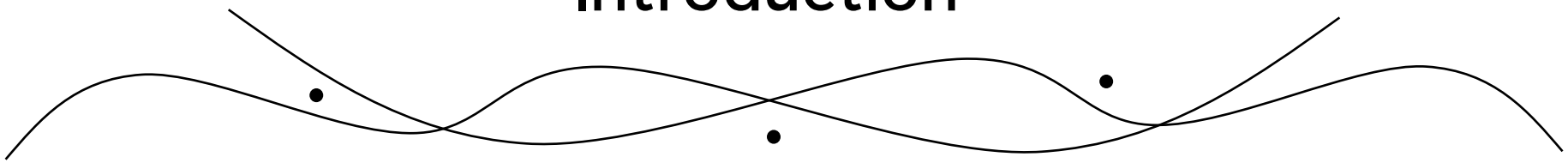
Predicting the Price of Used Cars

Weipeng Zhang, Aravindh Gowtham Bommisetty, Sai Vineeth Kaza | DS 5220

Part 01

Introduction



Introduction

Goal : Analyze the dataset and build a used cars' price predictor.

Online pricing services can offer better price estimates of a used car given some characteristics.

Dealers can better understand what features makes a car desirable and offer better services.

Individuals can make use of the model to better know the used cars market.

WHY? Any business value?

A Peek Into the Data

Dataset was originally built by using web crawlers on *carguru.com*

3M records

66 variables

27 numerical features

11 boolean features

24 categorical features

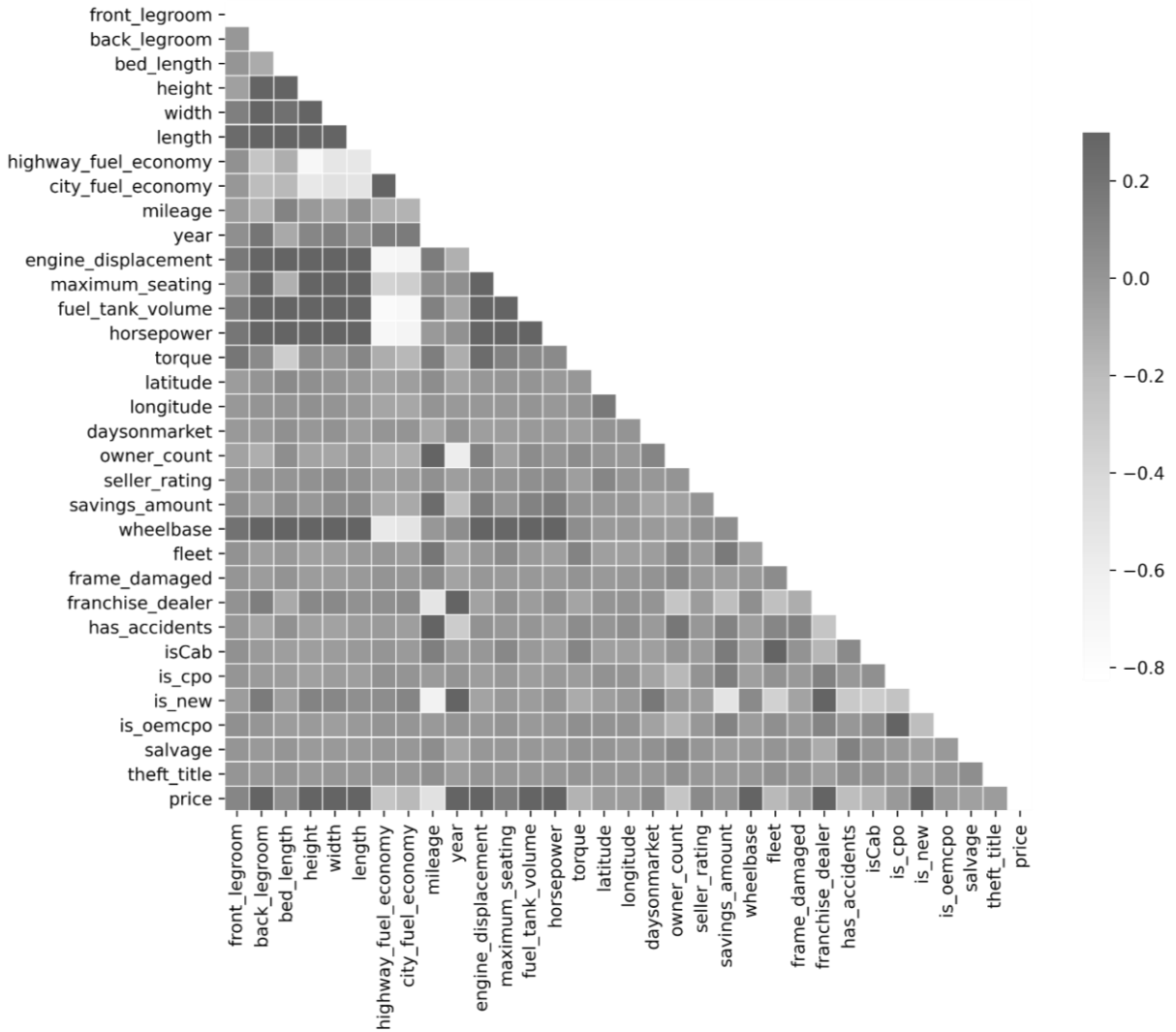
Information of cars and dealers.

	make_name	model_name	mileage	fuel_type	year	price
0	Dodge	Grand Caravan	85500	Gasoline	2001	5550.0
1	Dodge	Grand Caravan	18128	Flex Fuel Vehicle	2015	44995.0
2	Dodge	Grand Caravan	19532	Flex Fuel Vehicle	2015	44995.0
3	Dodge	Grand Caravan	197625	Gasoline	2008	7990.0
4	Dodge	Grand Caravan	29709	Flex Fuel Vehicle	2015	41595.0
5	Dodge	Grand Caravan	90074	Flex Fuel Vehicle	2012	9995.0
6	Dodge	Grand Caravan	251000	Gasoline	2008	10995.0
7	Dodge	Grand Caravan	130860	Flex Fuel Vehicle	2015	11995.0

Part 02

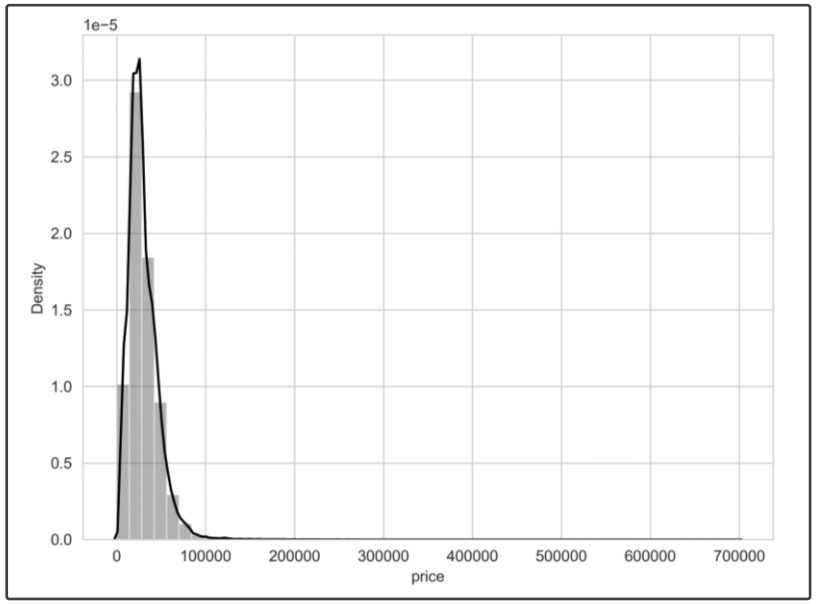
Exploratory Analysis

Exploratory Analysis



The strongest correlation is between ***price*** and ***power*** (0.61) followed by ***mileage*** (-0.48) and ***year*** (0.41).

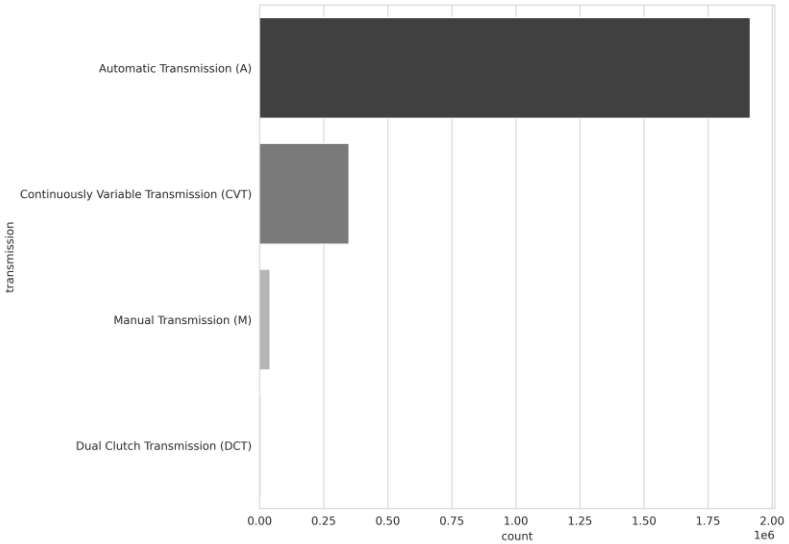
The target variable ***price*** is right skewed with exotic cars costing over 3m.

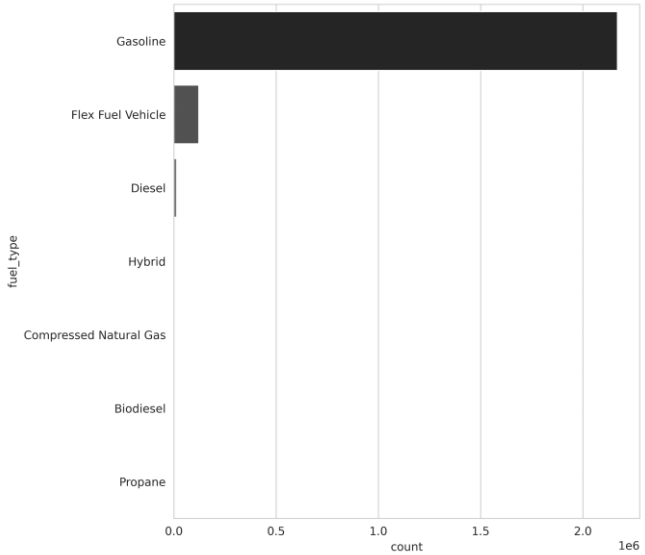


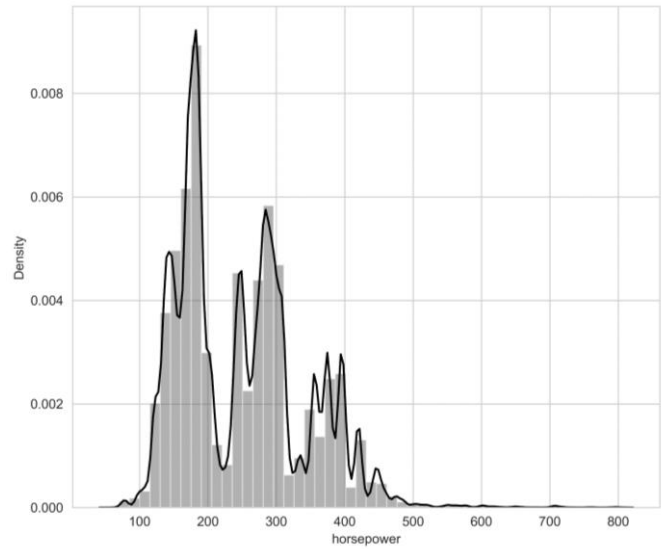
Exploratory Analysis

As the data is collected from US, most of the vehicles have automatic transmission and gasoline as fuel.

The ***horsepower*** ranges from 80 to 1001 and highest value corresponds to *Bugatti Veyron*.







Part 03

Feature Extraction and Preprocessing

Data Preprocessing

NA Analysis

16 variables have NA percentage as high as ***45%***

9 were dropped

7 were retained which will be imputed

NA Imputation

Continuous variables were imputed with mean.

Categorical variables were imputed with mode.

Deleting non-imputable records.

Special cases like electric cars were dealt separately.

Nonsense Variables

20 variables were dropped as they were not useful for the final model

2 variables were dropped because of duplicate information

Feature Engineering

Groupby Features

mean milage of each model in each year
number of cars of each model in each year
mean milage of each type of fuel
mean milage of each type of engine
...

Target Encoding

mean price of each model
mean price of each brand
mean price of each type of engine
mean price of each body type
...

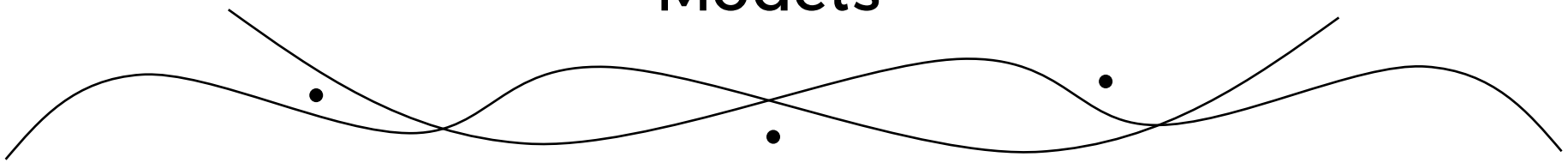
Other Features

mileage per year
estimated fuel spent in city
estimated fuel spent on highway
...

43 **new features generated**

Part 04

Models



Models

Slow

Random Forest Regressor

Support Vector Regressor

K Neighbors Regressor

CatBoost

>30 min

Fast

Decision Tree Regressor

Linear Regressor

Ridge Regressor

Lasso Regressor

<10 min

Fast with GPU

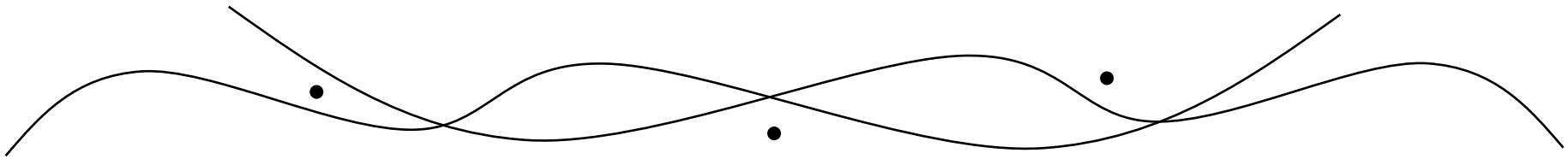
LightGBM

XGBoost

≈15 min

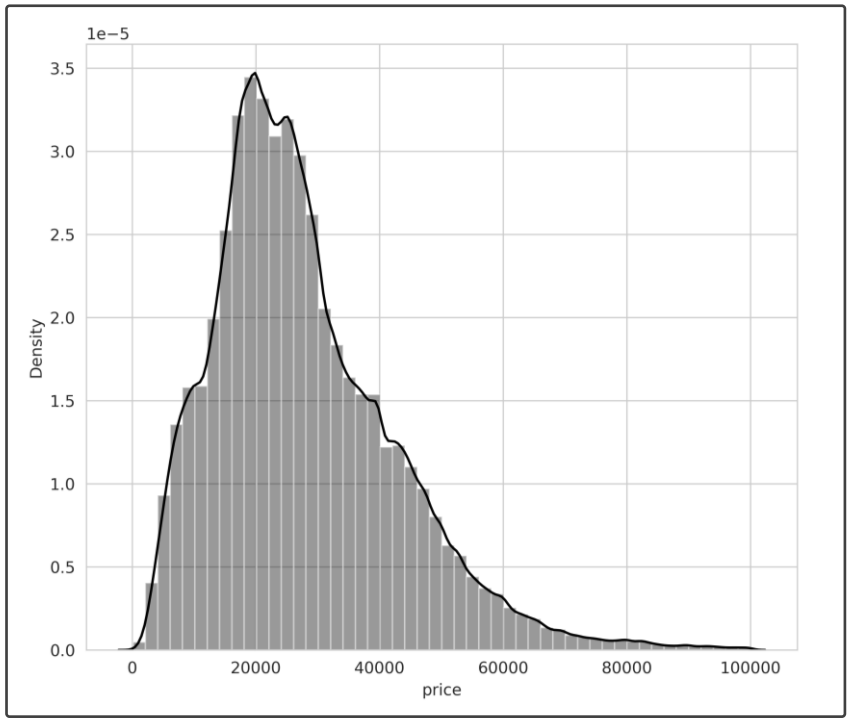
Part 05

Experiments And Evaluation

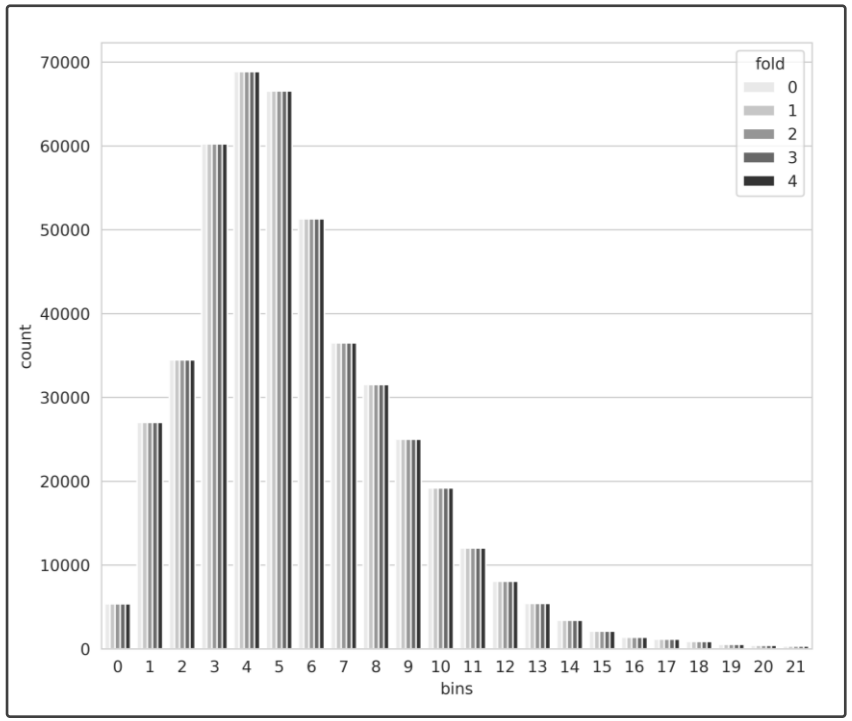


Data Split

Target Distribution



Stratified Splits



1 **Target Binning**

2 **Create Splits**

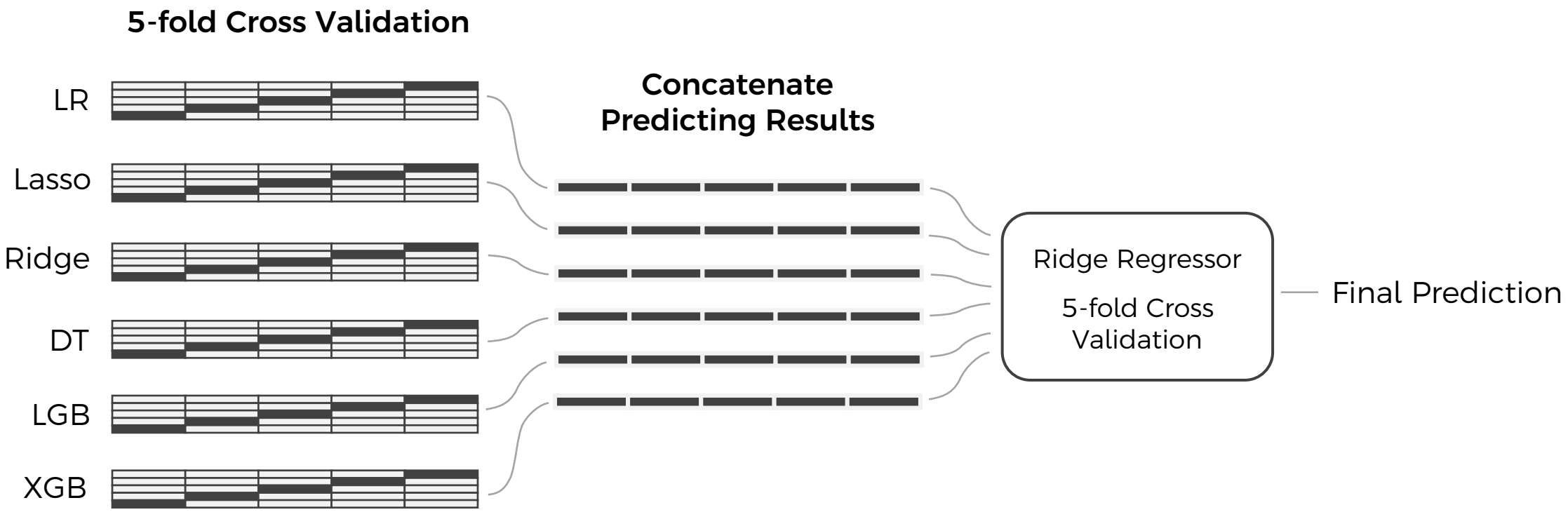
3 **Train-Test Split**
Train-Valid Split

Model Evaluation

Root Mean Square Error (RMSE)

Model Name	Baseline	Data Cleaning	Feature Engineering	Bayesian Parameter Estimation
Linear Regressor	14121	7039	4196	---
Ridge Regressor	14121	7039	4196	4196
Lasso Regressor	14121	7039	4197	4197
Decision Tree Regressor	7490	3242	3183	3051
LightGBM	7728	2942	3134	3007
XGBoost	7825	2938	2870	2852

Model Ensembling



Part 06

Summary

Summary

34 Features of Basic Car Information

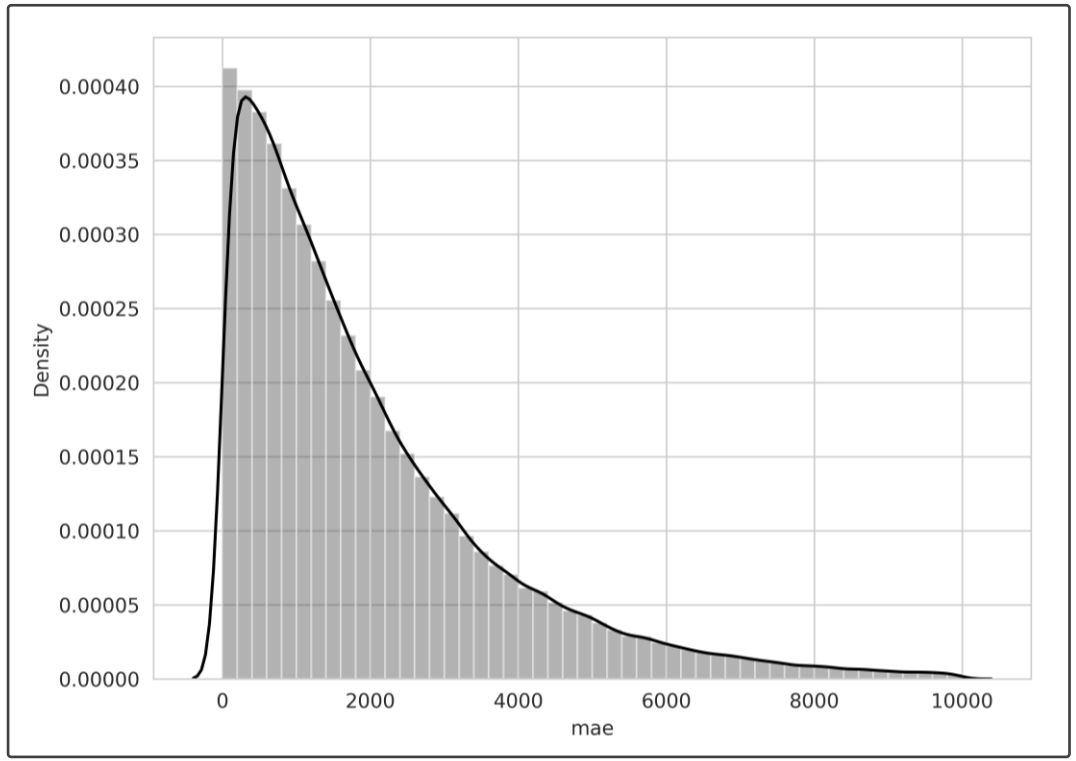
These features can be easily fetched thus letting our model have high applicability.

Ensembled 6 Different Models

Improved the precision and robustness of our predicting result

RMSE = 2851

Predicting **90%** of the variability in used cars with an average error of **4500$**



Error Distribution

Top important features mile_per_year, mileage, make_name, height, city_fuel_economy

THANK YOU

US Used Car Price Prediction

DS 5220 Final Project

Weipeng Zhang, Aravindh Gowtham Bommisetty, Sai Vineeth Kaza