PREDICT USED CAR PRICE

Capstone project

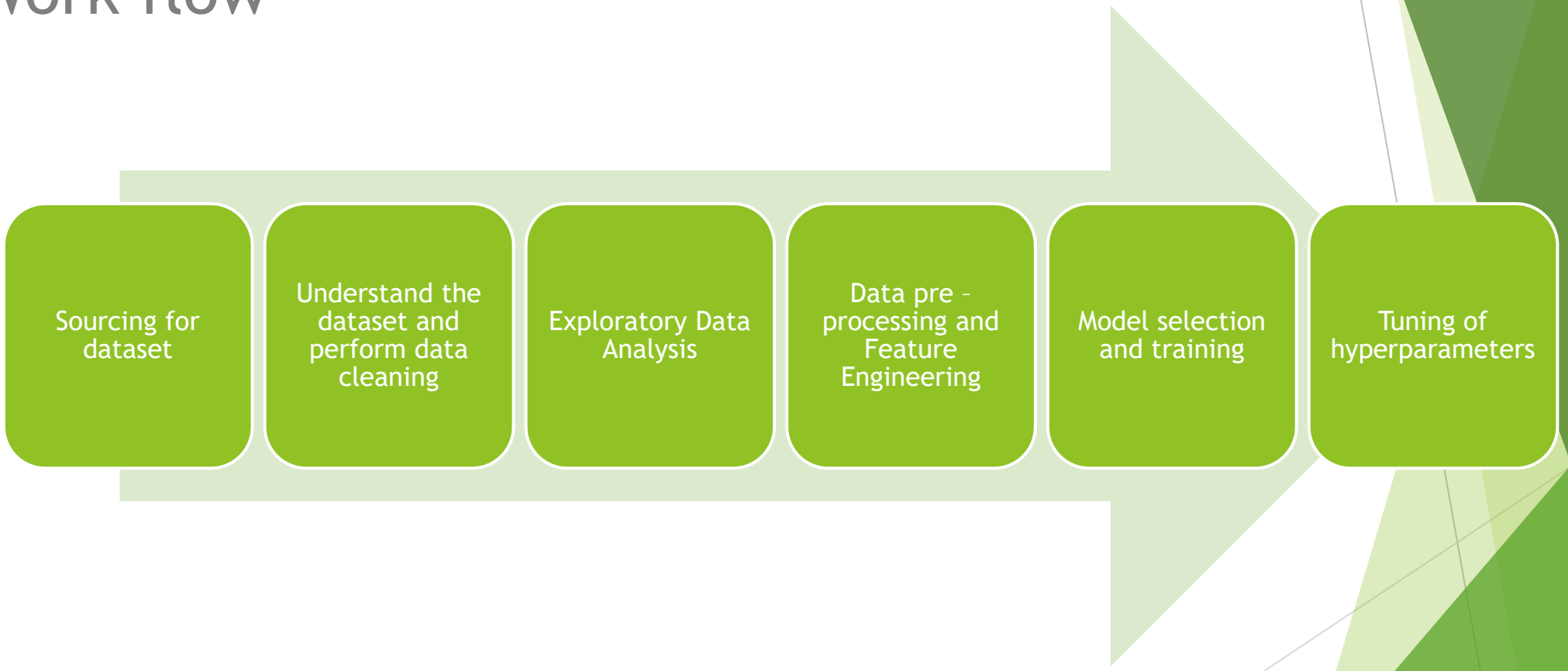
PROBLEM STATEMENT

Due to the reduction of COE quota, we had seen an increase in the COE premium. As a result, more buyers are turning to the used car market. But the prices for a particular make and model can have a lot of difference in the used car market due to the a few reasons like:

- How old is the car
- Mileage
- Number of owners etc

For this project, we would like to train a model to predict the price of the car so as to let buyers have a guide on how much is the car that they are looking at and work within their budget.

Work flow


Sourcing for dataset
Understand the dataset and perform data cleaning
Exploratory Data Analysis
Data pre – processing and Feature Engineering
Model selection and training
Tuning of hyperparameters

DATASET

- Using the data that scrape from find & buy used cars for sale in singapore – stcars

	0	1	2	3	4
make_model	MERCEDES-BENZ C-CLASS C200K (COE TILL 10/2028)	NISSAN X-TRAIL 2.0A PREMIUM 7-SEATER SUNROOF	NISSAN ELGRAND 2.5A HIGHWAY STAR	MERCEDES-BENZ CLS-CLASS CLS450 MILD HYBRID AMG...	TOYOTA HARRIER 2.4A G (COE TILL 09/2029)
Price	$63,800	$81,800	$88,000	$367,988	$58,500
Registration Date	19-Dec-2008	21-Mar-2017	03-Aug-2016	30-Oct-2020	23-Sep-2009
COE Remaining	7yrs 5mths 13days	5yrs 10mths 2days	5yrs 2mths 15days	9yrs 5mths 11days	8yrs 4mths 4days
Manufactured	2008	2016	2016	2020	2008
Mileage	165,000 km	74,000 km	82,225 km	652 km	174,000 km
No. of Owners	4	1	1	1	5
Transmission	Auto	Auto	Auto	Auto	Auto
Engine Capacity	1,796 cc	1,997 cc	2,488 cc	2,999 cc	2,362 cc
fuel_type	NaN	NaN	NaN	Petrol-Electric	NaN
COE	$32,279	$53,300	$57,501	$32,801	$37,941
OMV	$43,106	$23,955	$34,974	$83,552	$27,778
Paper Value	$24,081	$50,286	$60,687	$122,808	$31,694
Depreciation	$8,550 / year	$11,810 / year	$12,950 / year	$32,440 / year	$7,000 / year
Type	Luxury Sedan	SUV	MPV	Luxury Sedan	SUV

DATASET

- Consist of 11,251 rows and 15 features:

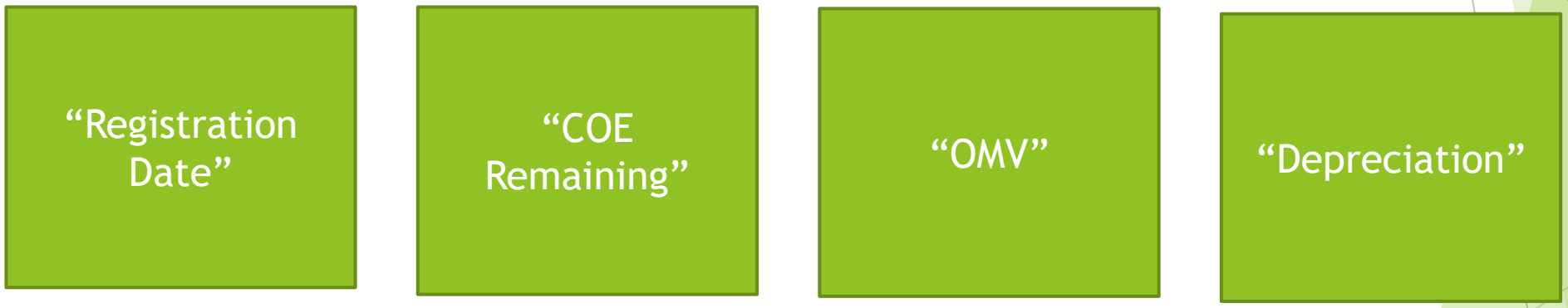
    1. make_model – make and model of the car
    2. Price – price of the car (our label)
    3. Registration Date – first registration date of the car
    4. COE Remaining – remaining COE of the car
    5. Manufactured – year the car is manufactured
    6. Mileage – mileage of the car in km
    7. No. of Owners – car is owned by how many owner before

DATASET

8. Transmission – auto or manual gear
9. Engine Capacity – engine capacity of the car in c.c
10. fuel_type – petrol, diesel, petrol – electric
11. COE – COE premium paid for the car when first registered
12. OMV – open market value of the car
13. Paper Value – the amount you get if the car is deregistered
14. Depreciation – how much the car depreciate per year
15. Type – type of the car (luxury sedan, MPV, SUV etc)

DATASET

- Features that could be important:



Questions to ask ourself

- Exploratory Data Analysis:
  I. How each features will affect the price
  II. The mean price of each car make
  III. The age distribution of the cars advertised
  IV. Is “green” car getting more popular

Questions to ask ourself

- Machine Learning (Supervised Regression Model) :

  I. Are we going to create or remove any features that may affect the model performance

  II. How are we going to deal with the outliers and missing values

  III. What to do if the feature distribution is skewed

  IV. Which encoding method to use for categorical features

  V. How many baseline models are we going to train and perform hyperparameter tuning

CHALLENGES

- Doing web scrapping to get the dataset
- Cleaning the dataset as it is a raw data
- Choosing encoding method for “make_model” as there are many variables inside
- Model selection and hyperparameters tuning

QUESTION?