Data Analytics Immersive Programme DAI01 Capstone Project

Prepared by Er Ning

Contents

1. Introduction

From the World Bank International Comparison Programme 2017, Singapore was ranked the most expensive country in the world to own a car. The average cost of a mid-range sedan in Singapore is around $125,300, more than twice the cost in the United States. This is attributed by Certificate of Entitlement (COE) and various additional fees and taxes imposed by the government to keep the car population in Singapore in check.

With the exception of certain ultra-luxurious car brands, owning a car in Singapore is widely seen as a depreciating asset. Car owners are not expected to profit from reselling their cars, because the value of a car decreases over time due to the decreasing validity of the COE once the car is registered. In general, there are two ways for car owners in Singapore to sell their used cars:

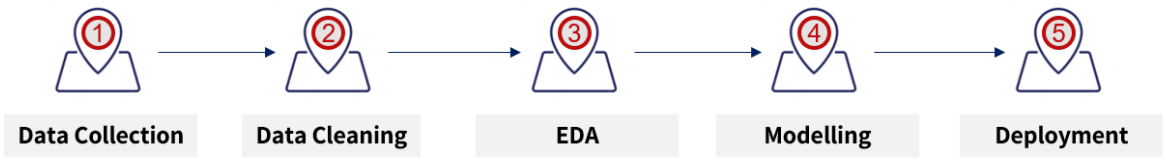
1) Selling to a car dealer. The resale car price is determined by car dealers and could vary greatly among different car dealers. The process of selling car to a car dealer is fuss-free but car owners may have to approach different car dealers to compare their quotes in order to get the best price.
2) Listing the car on an online platform and dealing with interested buyer directly. Car owners can compare the selling price with similar listings on the website or obtain a quote from the online platform. The exact mechanism of car valuation by the online platforms is unknown as car owners are still subject to price variation.

A machine learning-powered car valuation is a novel methodology that can bring tremendous benefit to the used car market in Singapore, which is forecasted to grow at a Compounding Annual Growth Rate of above 4% annually and reaching a market value of SGD 68 billions by 2027. It helps online platforms simplify their car valuation process by enabling instantaneous car price estimation. This is a crucial step in gaining market share among the technology-savvy car owners, whose population is growing rapidly as the society advances.

This project aimed to develop a machine learning pipeline for used car price prediction in the Singapore's context. Real used car data was collected from a market leader in the online space and a robust predictive model was successfully developed, with an accuracy score of 97%.

2. Data Workflow

Overview of data workflow for this project:



a. Data Collection

A total of 11,212 used car listings were scraped from SGCarMart. These car listings were posted from the period between 10 January and 25 March 2023, and they were still available for sale as of 27 March 2023. Python package BeautifulSoup was used and the scraping process was detailed in the Jupyter Notebook "Scraping from SGCarMart" (Annex A). Data fields scaped from the website are:

1	Listing's URL	10	Engine Capacity
2	Posted On	11	Curb Weight
3	Title of Listing	12	Manufactured
4	Price	13	Transmission
5	Seller	14	OMV
6	Mileage	15	ARF
7	Road Tax	16	Power
8	Deregistration Value	17	No. of Owner
9	COE	18	Registration Date

b. Data Cleaning

Summary of pre-cleaned data:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11211 entries, 0 to 11210
Data columns (total 18 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   url                11211 non-null  object
 1   Posted On          11211 non-null  object
 2   Car                11211 non-null  object
 3   Price              11185 non-null  object
 4   Seller             11211 non-null  object
 5   Mileage            9485 non-null   object
 6   Road Tax           11107 non-null  object
 7   Deg Value          10834 non-null  object
 8   COE                10945 non-null  object
 9   Engine Cap         11157 non-null  object
 10  Curb Weight        11084 non-null  object
 11  Manufactured       11211 non-null  int64
 12  Transmission       11211 non-null  object
 13  OMV                11200 non-null  object
 14  ARF                11195 non-null  object
 15  Power              11126 non-null  object
 16  No. of Owner       11211 non-null  object
 17  Registration Date  11209 non-null  object
dtypes: int64(1), object(17)
```

The following steps were performed to clean and process the raw data:

	Description
1	Removed unwanted characters and substrings.
2	Extracted substrings and placed them in separate columns.
3	Converted data to appropriate data types.
4	Inspected and imputed null values. Dropped rows that could not be imputed.
5	Converted string values to numeric values.
6	Created new columns (feature engineering).
7	Dropped unnecessary columns and rearranged the data frame.

The detailed steps of data cleaning process can be read from the Jupyter Notebook “Data Cleaning and Feature Engineering” (Annex A).

c. Feature Engineering

The following new columns were created using information from the scraped data:

- **Total COE left:**
  Calculated using registration date and COE validity period.
- **Car age:**
  Calculated using year of manufactured.
- **Brand category:**
  Cars were categorized into “Luxury” and “Regular” brand categories based on the car brand information on their official website.
- **Car category:**
  Cars were categorized into four groups-

Group 1	Luxury Brand with COE Category A
Group 2	Luxury Brand with COE Category B
Group 3	Regular Brand with COE Category A
Group 4	Regular Brand with COE Category B

- **Price and deregistration value ratio:**
  Calculated using listing price and deregistration value. This column was used to inspect and remove outliers but not used as a feature because it was a result of the intended predicted outcome- price.
- **OMV category:**
  Calculated using 25th, 50th and 75th percentile of the OMV values. Cars were grouped into 4 categories:

Group 1	OMV less than or equal to 25th percentile
Group 2	OMV less than or equal to 50th percentile
Group 3	OMV less than or equal to 75th percentile
Group 4	OMV more than 75th percentile

Summary of cleaned and processed data:

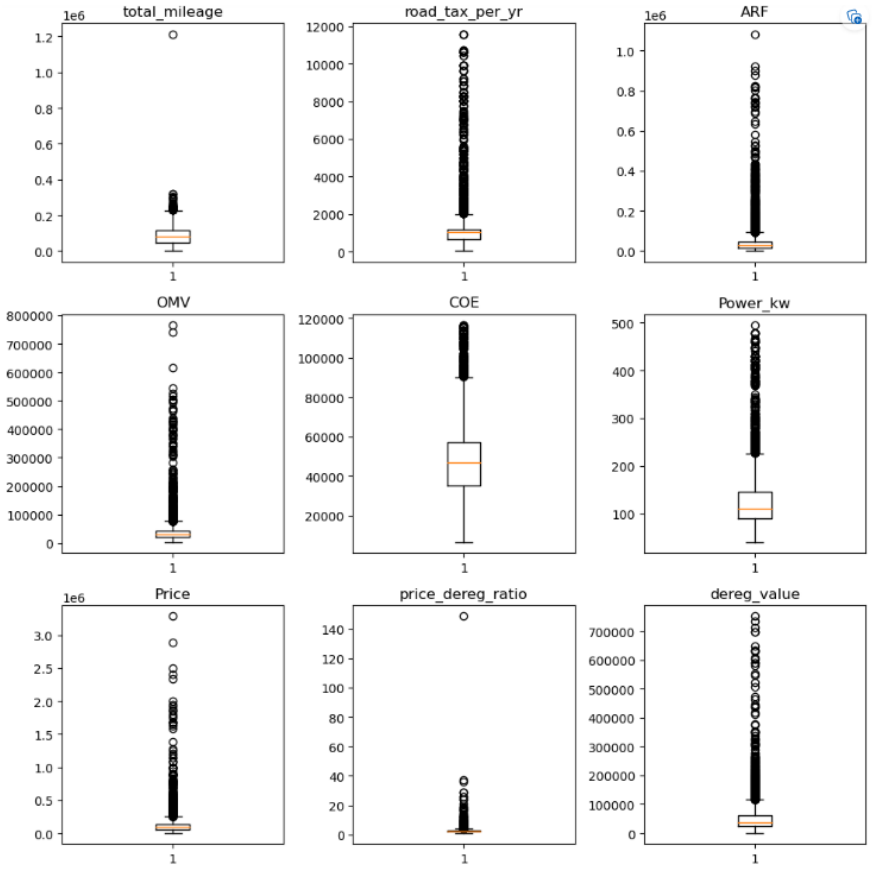
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10482 entries, 0 to 10481
Data columns (total 28 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   url                10482 non-null  object
 1   Posted On          10482 non-null  datetime64[ns]
 2   car_make           10482 non-null  object
 3   Seller             10482 non-null  object
 4   seller_type        10482 non-null  int64
 5   Price              10482 non-null  float64
 6   dereg_value        10482 non-null  float64
 7   COE                10482 non-null  float64
 8   Manufactured       10482 non-null  int64
 9   car_age            10482 non-null  int64
 10  road_tax_per_yr    10482 non-null  float64
 11  OMV                10482 non-null  float64
 12  ARF                10482 non-null  float64
 13  total_mileage      10482 non-null  float64
 14  Total_COE_left     10482 non-null  int64
 15  engine_cap_cc      10436 non-null  float64
 16  curb_weight_kg     10443 non-null  float64
 17  Transmission       10482 non-null  object
 18  transmission_type  10482 non-null  int64
 19  Power_kw           10482 non-null  float64
 20  owner_number       10482 non-null  int32
 21  registration_date  10482 non-null  datetime64[ns]
 22  car_brand          10482 non-null  object
 23  brand_category     10482 non-null  int64
 24  price_dereg_ratio  10482 non-null  float64
 25  OMV_cat            10482 non-null  int64
 26  COE_cat            10482 non-null  int64
 27  car_category       10482 non-null  int64
dtypes: datetime64[ns](2), float64(11), int32(1), int64(9), object(5)
```

d. Data Dictionary

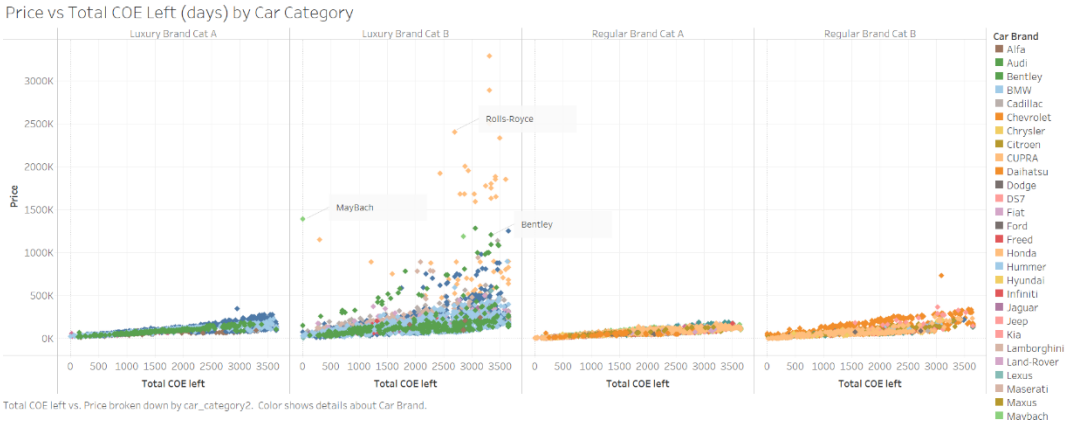
Column	Scraped or created?	Description	Data Type
URL	Scraped	URL of listing	Object
Posted On	Scraped	Date of listing posted on the website	Datetime
Car_make	Scraped	Car model	Object
Seller	Scraped	Seller of the listing: Name of car dealer or Ownership Transfer (directly listed by car owner)	Object
Seller Type	Created	0: Directly listed by car owner 1: Listed by car dealer	Integer
Price	Scraped	Listing price	float
Dereg_value	Scraped	Deregistration value or commonly known as scrap value. This is the amount a car owner will get back from the Land Transport Authority (LTA) upon deregistration the vehicle for use in Singapore. It is the sum of the COE rebate and the Partial Additional Registration Fee (PARF) rebate.	float
COE	Scraped	Price of COE premium paid by the car owner when the car was purchased	float
Manufactured	Scraped	Year of manufacture	Integer
Car_age	Created	Age of car from the year of manufacture	Integer
Road_tax_per_yr	Scraped	The amount of road tax paid by the car owner per year	float
OMV	Scraped	Open Market Value. It is the value of the car determined by Singapore Customs when the car is being imported into Singapore. It can be regarded as the raw cost of a brand-new car, before being subjected to other additional fees and taxes.	Float
ARF	Scraped	Additional Registration Fee. It is the tax payable by the car owner at the time of vehicle registration. It is a percentage of the vehicle's OMV. The prevailing ARF calculation: • First $20,000 of OMV: 100% of OMV • Next $20,000 of OMV: 140% of OMV • Next $20,000 of OMV: 190% of OMV • Next $20,000 of OMV: 250% of OMV • Above $80,000 of OMV: 320% of OMV This calculation was used for imputing the null ARF values in the dataset.	Float
Total_mileage	Scraped	Total mileage (in kilometers) of the vehicle	Float
Engine_cap_cc	Scraped	The engine capacity of the vehicle (in cc). Electric vehicles do not use this measurement.	Float

		This applies to fuel or gas-powered vehicles only.	
Curb_weight_kg	Scrapped	The weight of the car (in kilograms)	Float
Transmission	Scrapped	Auto or manual	Object
Transmission_type	Created	0: Manual 1: Auto	Integer
Power_kw	Scrapped	Power of engine (in kilowatts). Used by all types of vehicles as the measurement of engine power, including electric vehicles.	Float
Owner_number	Scrapped	Number of Owner(s)	Integer
Registration_date	Scrapped	Date of COE registration	Datetime
Car_brand	Scrapped	Car brand	Integer
Brand_category	Created	Luxury or regular brand.	Integer
Price_dereg_ratio	Created	A ratio of price and deregistration value. It is an indication of whether the car is over or under-priced.	Float
OMV_cat	Created	1: OMV less than or equal to 25th percentile 2: OMV less than or equal to 50th percentile 3: OMV less than or equal to 75th percentile 4: OMV more than 75th percentile	Integer
COE_cat	Created	Category A or Category B. It is dependent on the engine capacity or power: 1: Category A- cars ≤ 1,600cc or 110kW 2: Category B- cars above 1,600cc or 110kW	Integer
Car_category	Created	4: Luxury Brand with COE Category B 3: Regular Brand with COE Category B 2: Luxury Brand with COE Category A 1: Regular Brand with COE Category A	Integer

e. Exploratory Data Analysis



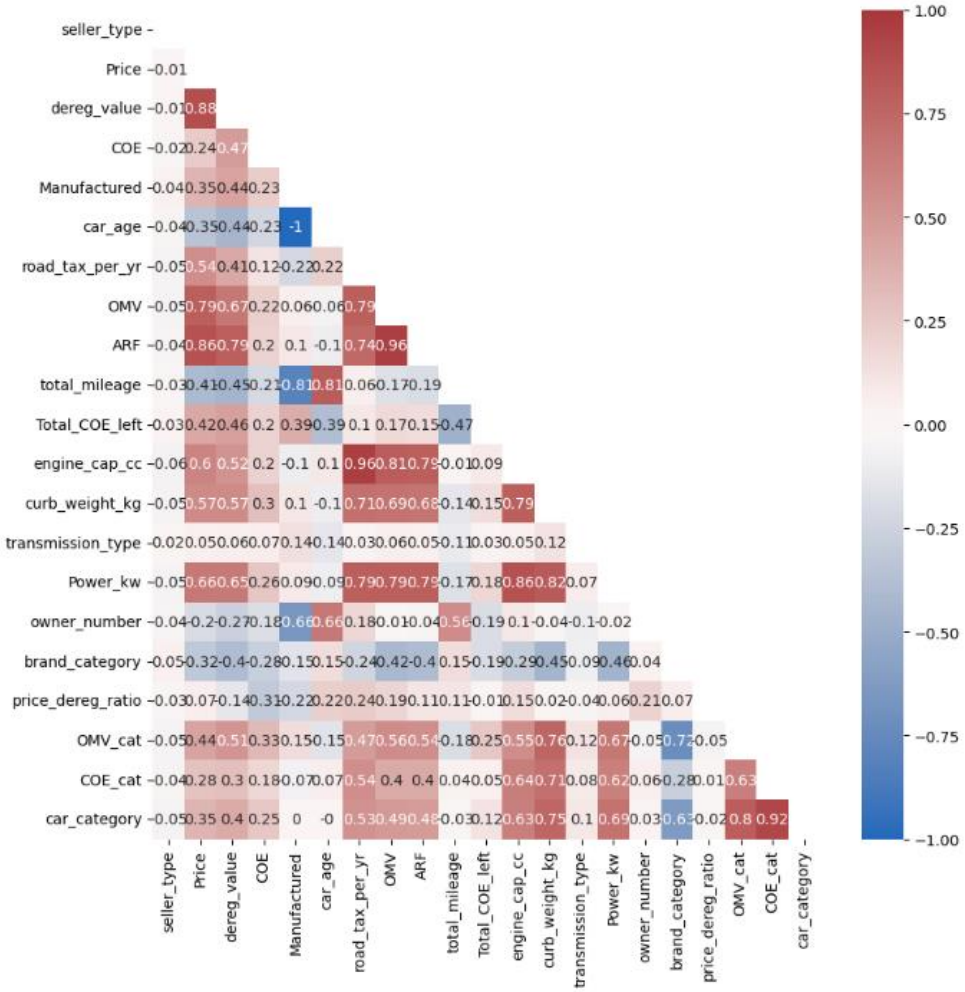
The box plot above illustrated the presence of outliers in the respective columns. As a result, outliers in total_mileage, price and price_dereg_ratio were dropped.



Price variation within Luxury Brand Category B is much larger than the other three categories, particularly among cars of certain brands, such as Rolls-Royce, Maybach and Bentley. Cars of these three brands were being dropped as they belong to the ultra-luxurious car category which has a different set of valuation metrics, such as rarity, collectability and craftmanship. Additionally, huge price variation was observed among cars that were priced above $500,000. It is presumed that cars of this price range are

being valued differently from the mass-market cars, which are the target audience of this project. Therefore, cars above $500,000 were being excluded from the subsequent predictive modelling.

After trimming outliers and further defining the scope of this project, the final dataset consists of 10,331 listings, approximately 8% reduction from the original dataset.



The above correlation matrix highlights features that have strong correlation coefficients with the intended outcome “Price”.

```
Price                1.00
dereg_value          0.88
ARF                  0.86
OMV                  0.79
Power_kw             0.66
engine_cap_cc        0.60
curb_weight_kg       0.57
road_tax_per_yr      0.54
OMV_cat              0.44
Total_COE_left       0.42
car_category         0.35
Manufactured         0.35
COE_cat              0.28
COE                  0.24
price_dereg_ratio    0.07
transmission_type    0.05
seller_type         -0.01
owner_number        -0.20
brand_category      -0.32
car_age             -0.35
total_mileage       -0.41
Name: Price, dtype: float64
```

Using the correlation coefficients as reference, potential features for modelling were selected: dereg_value, ARF, OMV, Power_kW, road_tax_per_yr, Total_COE_left, car_category, car_age and total_mileage. Engine_cap_cc and curb_weight_kg was not selected due to the high number of null values in these two columns.

f. Predictive Modelling

Prediction of used car price is a regression problem. The following regression models were attempted and fine-tuned to select the best-performing model that has the highest R^2 score and the lowest Root Mean Square Error (RMSE) for the test data set:

1. Linear Regression
2. K-Nearest Neighbour Regression
3. Random Forest Regression
4. Lasso Regression
5. Ridge Regression
6. Support Vector Machine Regression
7. XGBoost Regression

Train-test split of the dataset was performed with a ratio of 0.7 for training and 0.3 for testing. Various sets of scaler, transformer and normalizer were incorporated to pre-process the data and bring them into a pre-defined range, so that they could be properly fitted into the model. QuantileTransformer with output_distribution="uniform" performed the best for this dataset.

	scaler	R2_test	RMSE_test	R2_train	RMSE_train
0	QuantileTransformer(output_distribution='normal')	0.971396	11735.489586	0.999986	262.514659
1	QuantileTransformer()	0.970772	11862.758485	0.999986	262.514659
2	RobustScaler(quantile_range=(25, 75))	0.964005	13164.559732	0.999986	262.514659
3	PowerTransformer()	0.970629	11891.781933	0.999986	262.514659
4	StandardScaler()	0.968313	12351.661529	0.999986	262.514659

After numerous iterations of modelling, the most optimal set of features is as follows: 'dereg_value', 'OMV', 'ARF', 'road_tax_per_yr', 'total_mileage', 'Total_COE_left', 'Power_kw', 'car_age' and 'car_category'.

The following table summarizes the performance of each model:

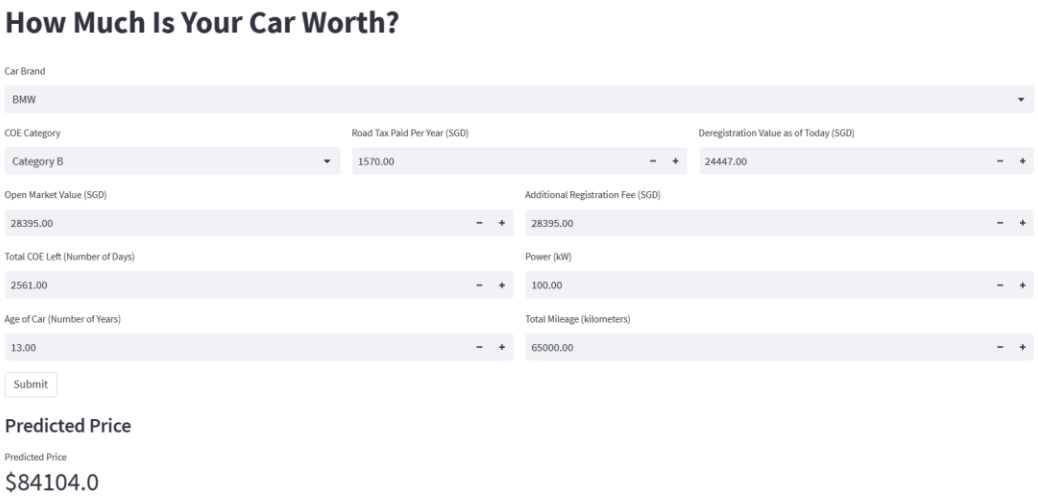
	model	R2_test	RMSE_test	R2_train	Parameter
0	Linear Regression	0.893462	22648.358246	0.89398	nil
1	K-Nearest Neighbor Regression	0.971506	11712.746305	0.999986	k=7
2	Random Forest Regression	0.965071	13044.580655	0.995733	n_estimators=600
3	Support Vector Machine Regression	0.468227	51961.413864	nil	kernel= linear
4	Lasso Regression	0.804028	30717.191796	nil	alpha= 0.2
5	Ridge Regression	0.804028	30717.179558	nil	alpha= 0.2
6	XGBoost Regression	0.973528	11289.671815	0.999266	max_depth= 9.0, n_estimators= 100.0

XGBoost is the best-performing model with an accuracy score of 0.97 on the test dataset. A RMSE of 11289.67 is approximately 10.5% of the average car price. The accuracy score of train dataset is 0.99, 2% above the test dataset, indicating that the model did not overfit the training dataset.

Detailed steps on the predictive modelling can be found in the Jupyter Notebook "Predictive Modelling" (Annex A).

g. Deployment to Streamlit

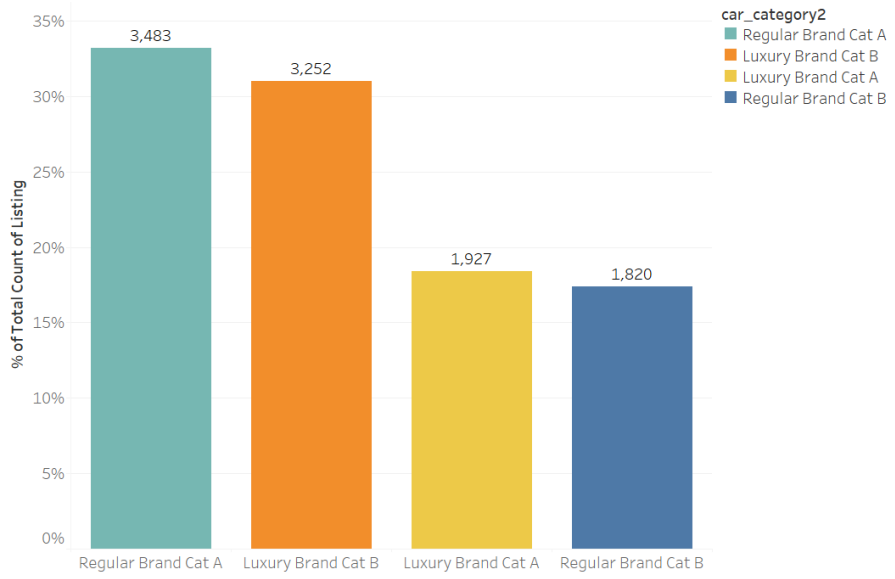
The final predictive model was deployed to a Streamlit prototype. Users are able to feed the required car info into the prototype and obtain an estimated price for the car.



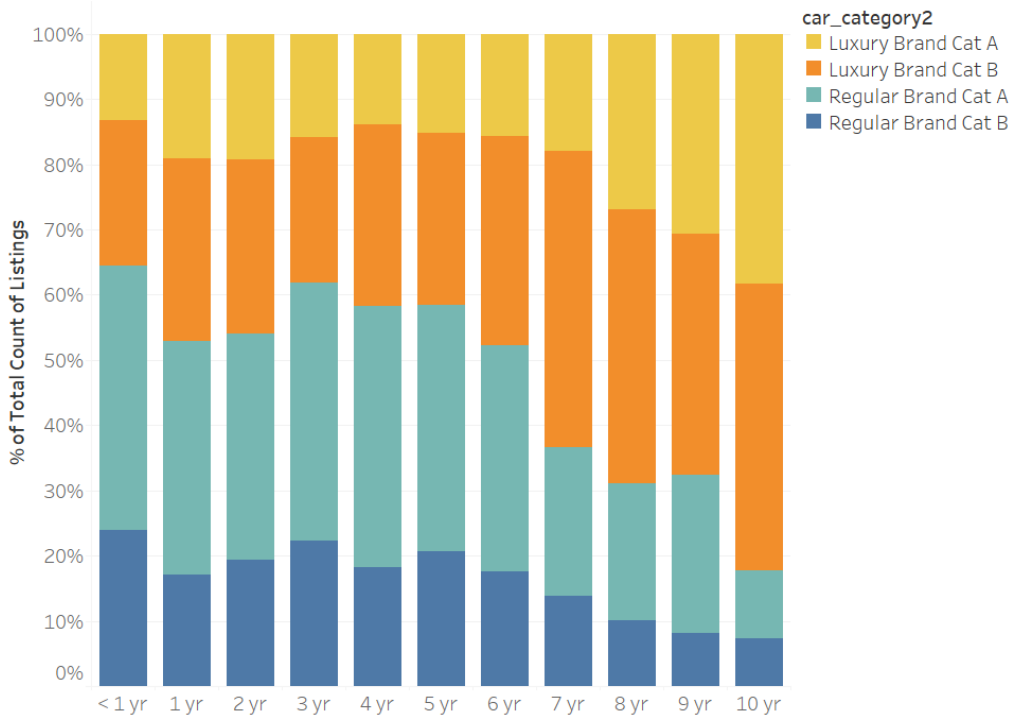
Codes for the Streamlit prototype can be found in Python file "car_price_prediction.py" (Annex A).

3. Findings and Insights

Count of Listings by Car Category



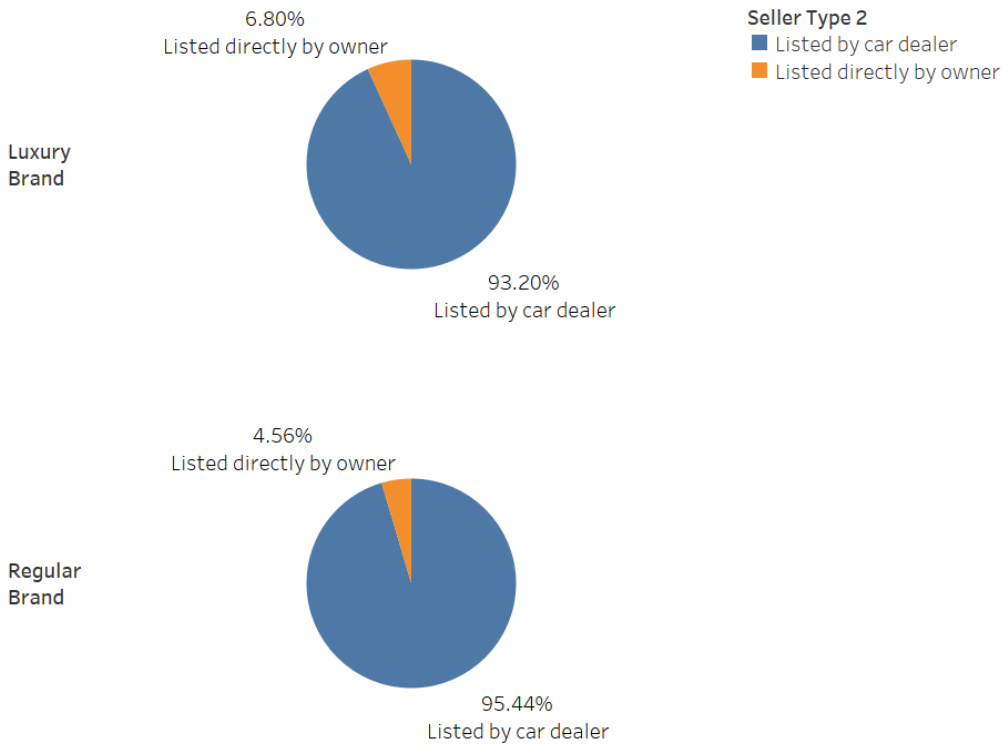
Percentage of Car Category Listings by Total COE Left



Regular Brand Cat A has the highest number of listings on the website, followed by Luxury Brand Cat B, Luxury Brand Cat A and Regular Brand Cat B. When further segregating them according to the total COE left, the percentage of Regular Brand (both Cat A and Cat B) decreased with the increase of total COE left, while the percentage of Luxury Brand (both Cat A and Cat B) increased. This shows that more luxury car listings are available than regular car listings if potential buyers

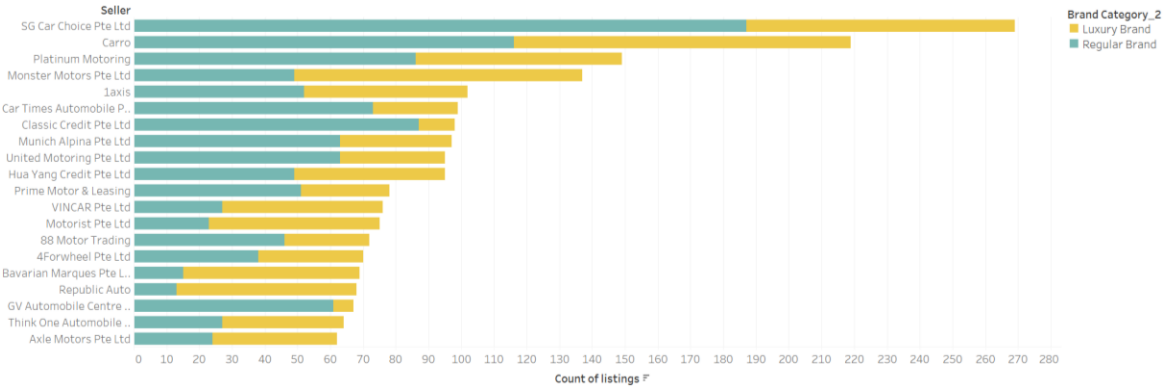
are looking for cars with longer period of valid COE. On the other hand, more regular car listings can be found if potential buyers are looking for cars with shorter period of valid COE, possibly due to cheaper cost.

Seller Type by Brand Category

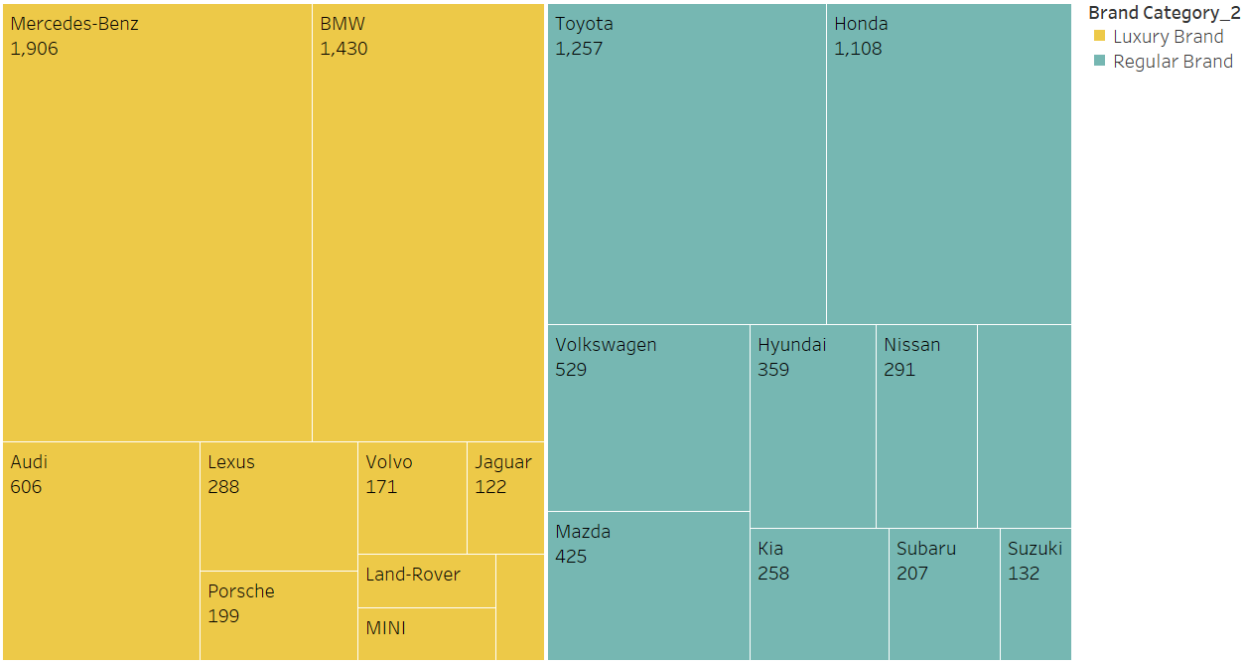


Most listings (more than 93%) are listed by car dealers. Direct listing by car owners accounted for less than 7% of the total listings. Similar trend is seen in both luxury and regular category. Luxury car brand has slightly higher percentage of listings by owner than regular car brand (6.8% and 4.56%).

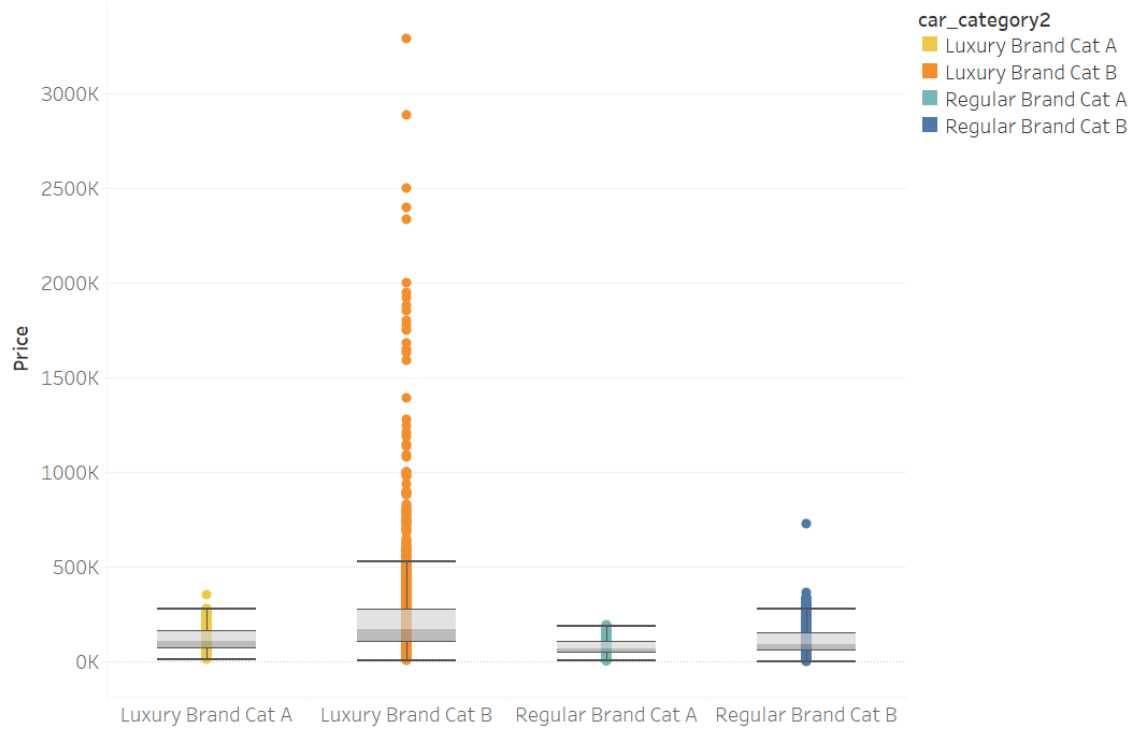
Top 20 Car Dealers on SGCarMart by Number of Listings



Top 20 Car Brands on SGCarMart by Number of Listings



Car Price Range by Car Category



Luxury Brand Category B cars have the highest variation in price range as compared to the other three category. Regular Brand Category A has the smallest variation in price range.

4. Recommendation and Conclusion

a. Business Recommendations

A predictive modelling with high accuracy score was successfully developed, indicating that the use of machine learning in car valuation is highly attractive for the automotive industry. It has shown promising potential in shortening the time needed for potential sellers to receive a quote for their car. This could also bring price transparency to consumers, thereby fuelling the growth of the market. Predictive modelling using machine learning could help online market players to gain a bigger piece of the pie from the traditional offline players, who still occupy the largest market share in Singapore.

The analysis on seller type has shown that car dealers are still the vast majority even in the online space. It is possibly because of the challenges and hassle faced by car owners when selling the car themselves and engaging buyers directly. Limited insights can be obtained from this dataset due to the lack of actual sales data, which is important for understanding market trends and buyer behaviour. More research and analysis can be done by the company to gain a deeper understanding of the seller and buyer dynamics in the online used car market. This may help them identify potential business opportunities and attract more car owners to list their used car on the platform. Both the diversity and volume of car listings are crucial drivers for an online car marketplace to stand out in an increasingly competitive market, with more players joining the scene.

b. Key assumptions

- The dataset scraped from SGCarMart, who is one of the leading online car marketplace, is representative of the used car price in Singapore.
- The car information listed on the website is validated and true. Falsified information or misdeclaration would undermine the credibility and accuracy of this model.
- The listed price is close to the actual sale price. Actual sales data could not be obtained from the website, and therefore listing price is presumed to be approximating sale price.

c. Conclusion

- Even though the predictive model has obtained the desired accuracy score, the RMSE score can be improved if the model can be further fine-tuned with data on the actual sales price. The RMSE score of 10.5% of the average listing price could be an indication of high price variation observed on the market currently.
- Overall, with an accuracy score of 97%, it is likely that this predictive model could be ready for implementation by the online car marketplace if the RMSE score can be reduced to less than 5% compared to the actual sales price.

5. Annex

Most parts in this project were done using Jupyter Notebook:

1. Scraping from SGCarMart.ipynb
2. Data Cleaning & Feature Engineering.ipynb
3. EDA.ipynb
4. Predictive Modelling.ipynb

Deployment to Streamlit:

1. Car_price_prediction.py