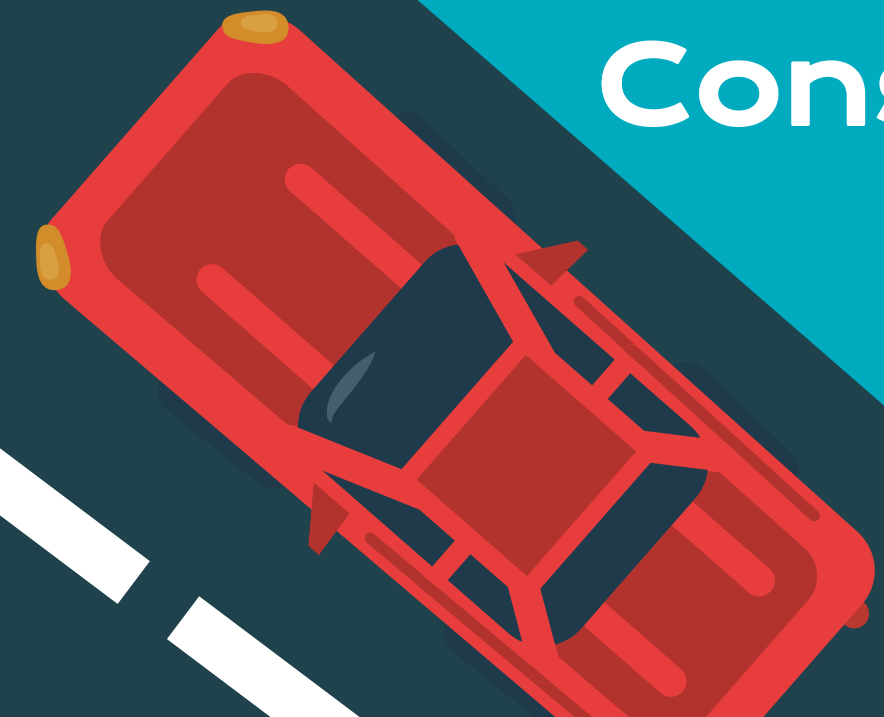


Car-eful Consideration

An ML approach to pricing cars

Team 10
Aishwarya Sanjay Maloo
Chua Wee Yuan (Marcus)
Lau Ho Yin Amanda Faith
Prachi Rajendra Ashani
Uthara Venkatachari

Have you bought a resale car?

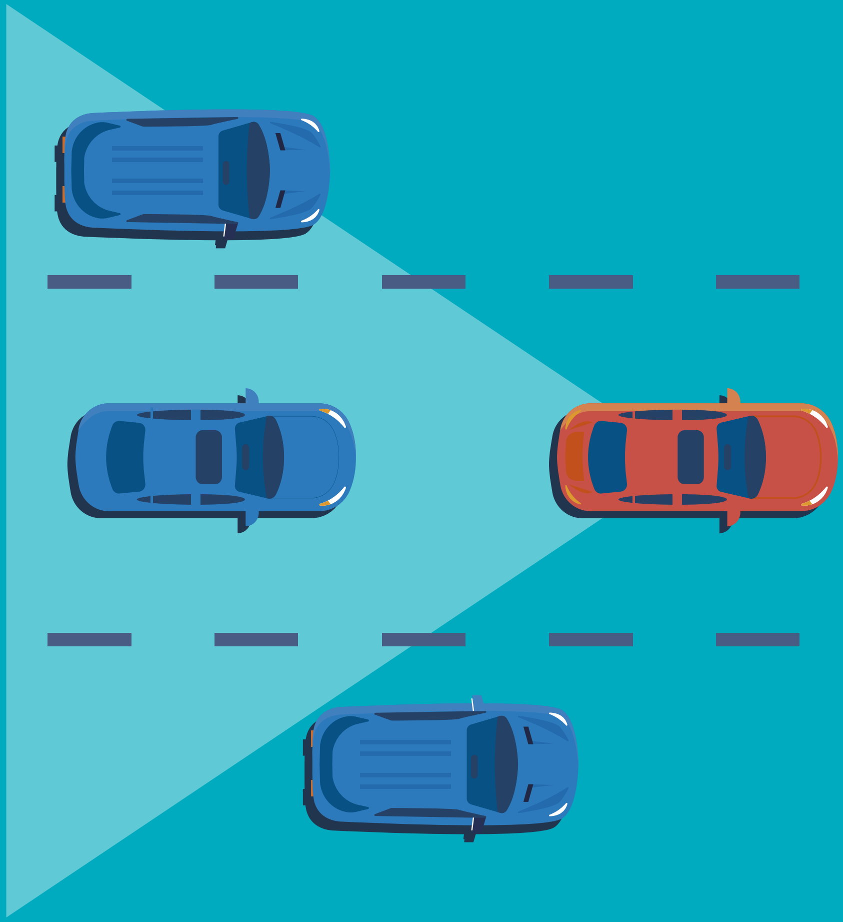
Used car pricing is often ambiguous – it can depend on multiple factors.

However, how these affect the price of the car is often unclear to most potential buyers.

Our Objective

We seek to build an **online platform** to bring **clarity** to buyers in their **decision-making process** and put them in a **better position** to **negotiate** a better price

Let us show you



DEMO

01

DATA

EDA & Processing

02

COMPUTER VISION

Feature Selection
Model Build & Analysis

03

PRICE PREDICTION

Feature Selection
Model Builds & Analysis

04

CONCLUSION

Learnings

About the data

Car Images Dataset

- **Purpose:** Used for computer vision model
- **Source:** Stanford
- **Sample Size:** 16,185
- **Feature Size:** N/A

USA Used Cars Dataset

- **Purpose:** Used for price prediction model
- **Source:** Kaggle
- **Sample Size:** 3,000,040
- **Feature Size:** 166 features

CV Multiclass Model data processing

Car Images

Resize

Resized Image to higher resolution 224 x 244 pixels or 229 x 229 pixels common for CNNs

Normalization

Normalized mean and standard deviation to reduce impact of variations in light, color, and contrast

Augmentation

Generated ‘new’ images by applying various transformations (rotation, translation, scaling, flipping, and color jittering)

Data processing for prediction models

Data with features

ELIMINATION

Dropped data with **NAs** and **columns** based on correlation results

NEW VARIABLES

Created new variables such as **age** of the car and **state** from where they are listed

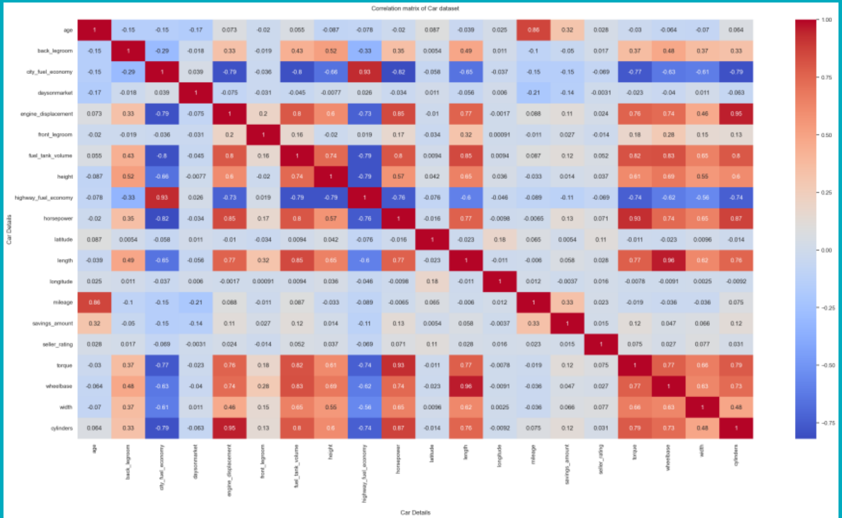
CREATED DUMMIES

Used **one-hot encoding** for categorical variables

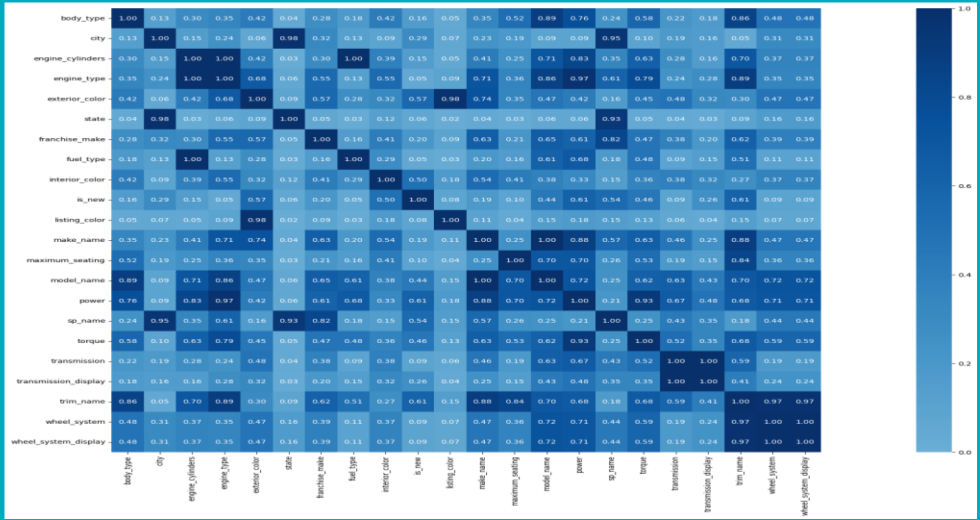
EDA: CORRELATION ANALYSIS (1/4)

We check correlation between variables to avoid inputting highly correlated variables into the model

CorrPlot for continuous variables

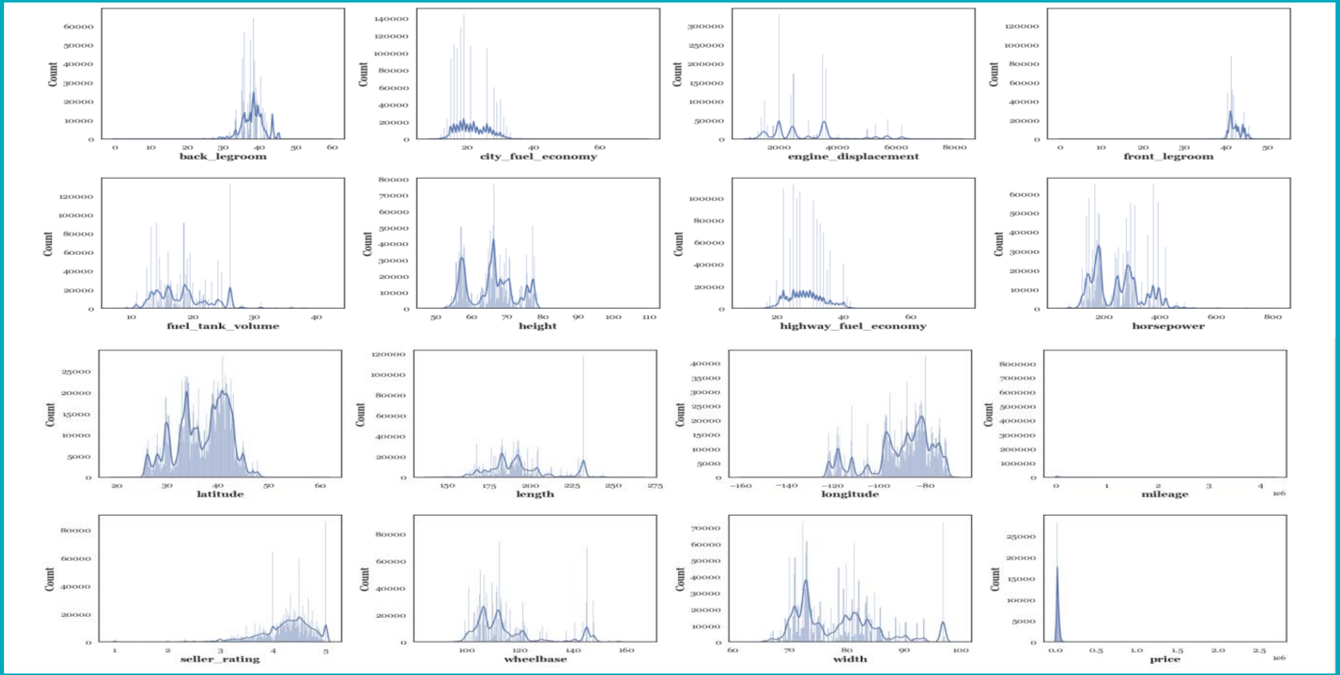


Cramer's V for categorical variables



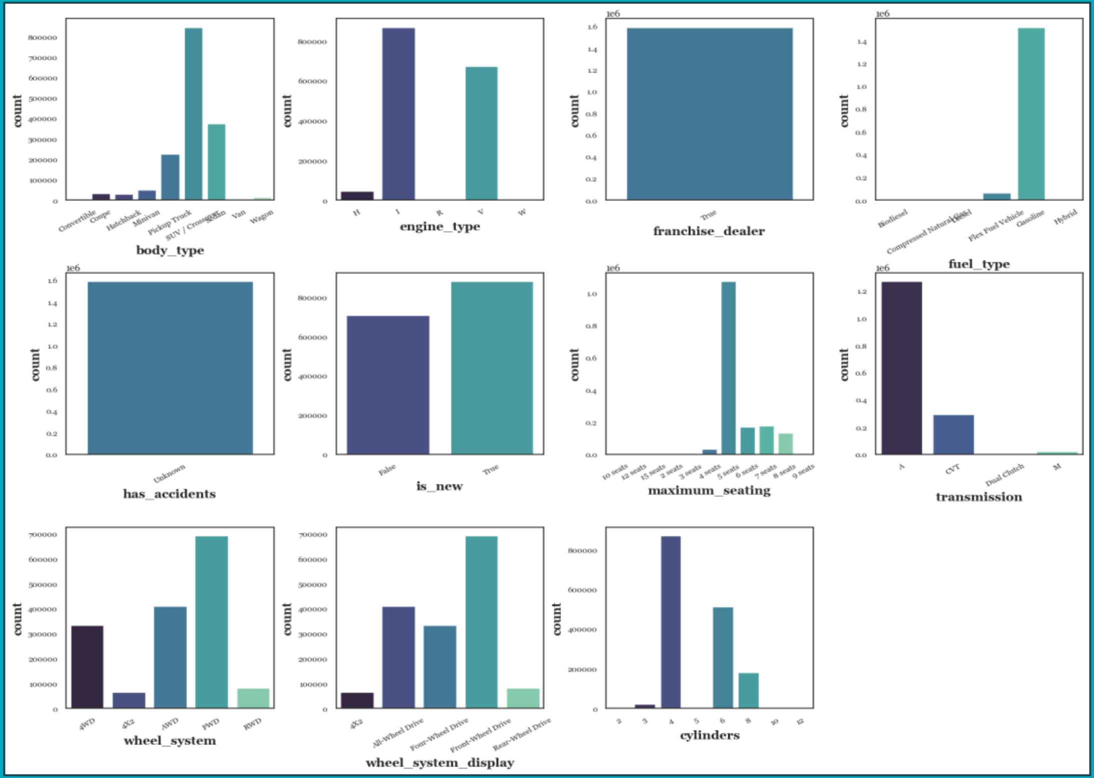
EDA: CHECKING THE DATA DISTRIBUTION (2/4)

Since the input variables are not normally distributed, data normalization will be required prior to developing the machine learning models



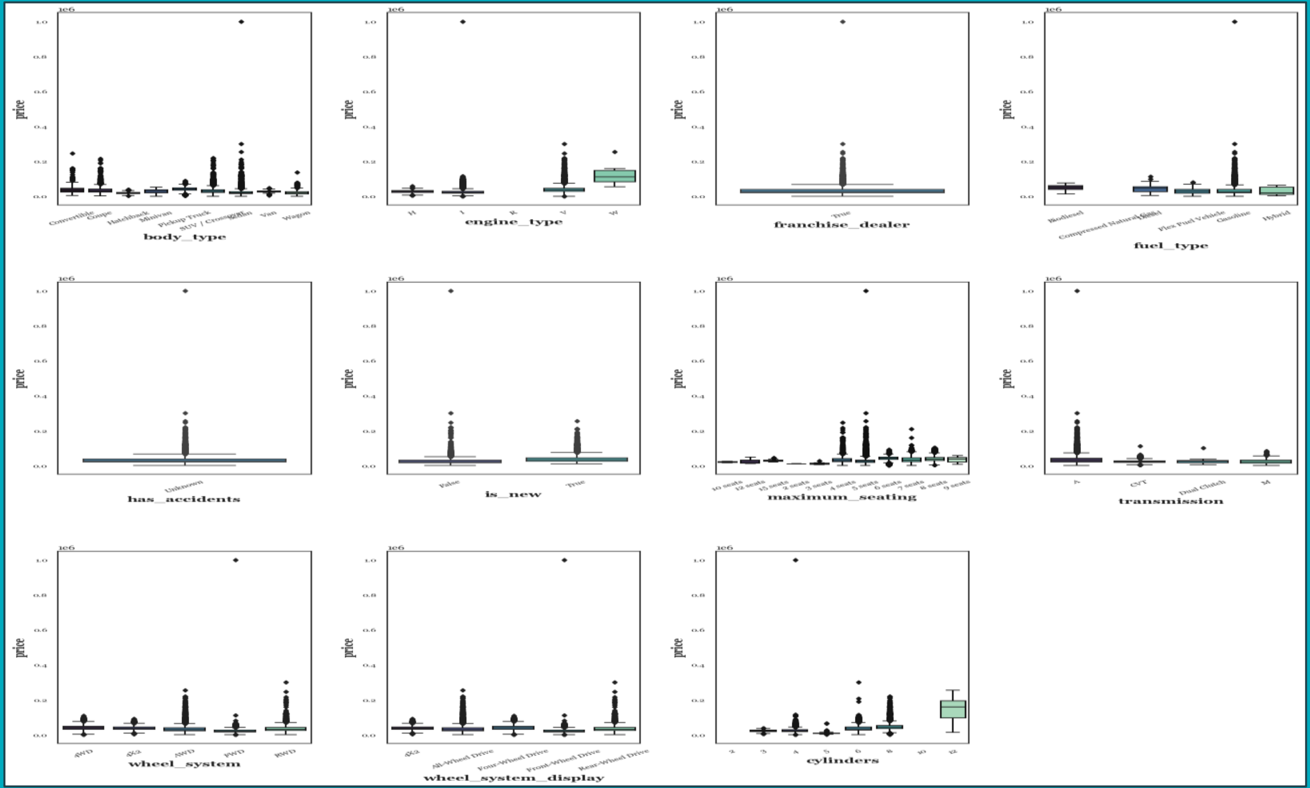
EDA: CHECKING THE DATA DISTRIBUTION (3/4)

Since the input variables are not normally distributed, data normalization will be required prior to developing the machine learning models



EDA: CHECKING FOR OUTLIERS (4/4)

We check for the presence of outliers in categorical variables, data normalization will be required prior to developing the machine learning models



01

DATA

EDA & Processing

02

COMPUTER VISION

Feature Selection
Model Build & Analysis

03

PRICE PREDICTION

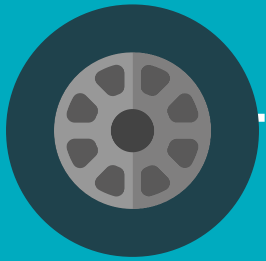
Feature Selection
Model Builds & Analysis

04

CONCLUSION

Learnings

Process for CV multi-class model selection



01

Training

Trained ~100 (20/80) epochs with various torchvision transfer learning backbones: Resnet50, Resnet152, Effcientnetv2_L, Inceptionv3

02

Evaluation

Saved Model after every epoch, evaluated and selected the best model on precision/ recall/ accuracy

03

Fine Tuning

Further trialed the model by varying hyper parameters like learning rate, addition of FC layer & Input resolution

Trials best model summary

Final Model

	Resnet50	Resnet152 w FC	Efficientnet v2_L w FC	Inceptionv3 w FC
Precision	77%	90%	93%	85%
Recall	77%	89%	92%	84%
Accuracy	79%	89%	92%	84%

Final Selected Model

Parameters

Model Parameters

EfficientnetV2	FC Layer
Depth: 23 layers	Depth: 2 layers
Width: 4.0	Activation F: Relu
Activation F: Swish	598k parameters
Pooling: Adaptive Average Pooling	
307 mil parameters	

Training Parameters

Optimizer	Adam
Learning R	0.001,0.0001
LR Scheduler	Cosine Annealing
Dropout	0.3
Batch size	32

Classification Report

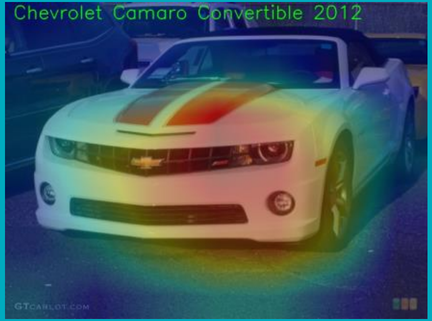
	precision	recall	f1-score	support
AM General Hummer SUV 2000	0.98	0.98	0.98	44
Acura Integra Type R 2001	0.98	0.93	0.95	44
Acura RL Sedan 2012	0.82	0.88	0.85	32
Acura TL Sedan 2012	0.79	0.98	0.88	43
Acura TL Type-S 2008	1.00	1.00	1.00	42
Acura TSX Sedan 2012	1.00	0.78	0.87	40
Acura ZDX Hatchback 2012	1.00	0.87	0.93	39
Aston Martin V8 Vantage Convertible 2012	0.90	0.78	0.83	45
Aston Martin V8 Vantage Coupe 2012	0.82	0.78	0.80	41
Aston Martin Virage Convertible 2012	1.00	0.82	0.90	33
Aston Martin Virage Coupe 2012	0.77	0.97	0.86	38
Audi 100 Sedan 1994	0.71	0.90	0.79	40
Audi 100 Wagon 1994	1.00	0.83	0.91	42
Audi A5 Coupe 2012	0.71	0.90	0.80	41
Audi R8 Coupe 2012	0.98	0.95	0.96	43
Audi RS 4 Convertible 2008	0.89	0.92	0.90	36
Audi S4 Sedan 2007	0.96	0.96	0.96	45
Audi S4 Sedan 2012	0.97	0.74	0.84	39
Audi S5 Convertible 2012	0.84	0.76	0.80	42

...

	precision	recall	f1-score	support
Tesla Model S Sedan 2012	1.00	0.97	0.99	38
Toyota 4Runner SUV 2012	0.98	1.00	0.99	40
Toyota Camry Sedan 2012	0.95	0.95	0.95	43
Toyota Corolla Sedan 2012	0.97	0.91	0.94	43
Toyota Sequoia SUV 2012	0.97	0.89	0.93	38
Volkswagen Beetle Hatchback 2012	1.00	1.00	1.00	42
Volkswagen Golf Hatchback 1991	1.00	0.98	0.99	46
Volkswagen Golf Hatchback 2012	0.95	0.93	0.94	43
Volvo 240 Sedan 1993	0.96	0.96	0.96	45
Volvo C30 Hatchback 2012	1.00	0.95	0.97	41
Volvo XC90 SUV 2007	0.95	0.98	0.97	43
smart fortwo Convertible 2012	0.95	1.00	0.98	40
accuracy			0.92	8041
macro avg	0.93	0.92	0.92	8041
weighted avg	0.93	0.92	0.92	8041

Visualizing class activation maps

Front View



Diagonal View

Side View

Back View

Class Activation Mapping Code Repurposed from https://github.com/zhoubolei/CAM/blob/master/pytorch_CAM.py

Limitations and Future Considerations

Limitations

Insufficient computing power

Realistically, number of brand, model and make is very high

Resolution of input images need to be high

Future Work

Add more data and add more classes, split models up by classes

Try more transfer-learning backbone

Optimizing NN hyperparameters further (batch-size, epochs, optimizers,)

01

DATA

EDA & Processing

02

COMPUTER VISION

Feature Selection
Model Build & Analysis

03

PRICE PREDICTION

Feature Selection
Model Builds & Analysis

04

CONCLUSION

Learnings

Our model evaluation metrics

RMSE $ To give a dollar value indication of pricing error

R^2 % To show variance explained

MAPE % To show pricing error whilst considering for large price values

Bias-Variance To check for model fit

K-fold Cross-Validation For final model validation

In any price negotiation…

MAPE 12.0%

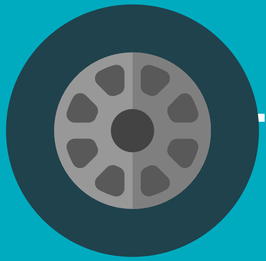
Negotiation band

Our predicted price

Seller's min price

Buyer's max price

Overview of our process for price prediction



01 — Scaling

Variations of Robust and Standard scaling used

02 — Feature Selection

Based on Decision Tree model built, feature importance and model metrics informed this process

03 — Model Builds & Evaluation

Decision Tree, XGBoost, MLR and Neural Network

Models were built on scaled data after each feature was omitted in feature selection

Train-Test-Val Split: 60-20-20

Scaling

By Sci-kit Learn

Robust

Addresses outliers

Standard

Centers data around the mean

Feature selection – variables omitted

Based on Decision Tree model

01
SAVINGS AMOUNT
Definition unclear

02
TRANSMISSION TYPE
Low feature importance

03
FUEL TYPE
Low feature importance

04
MAX. SEATING
Low feature importance for some values

05
ENGINE TYPE
Low feature importance

06
BODY TYPE
Low feature importance

07
LISTING COLOUR
Low feature importance

08
STATE
Low feature importance

Feature Selection Justification

Models

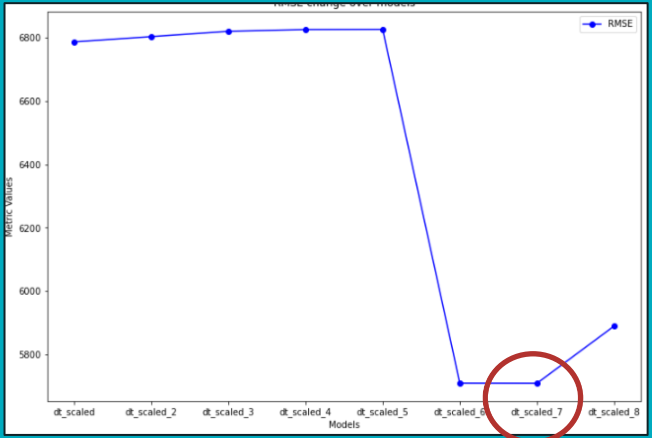
- 01 SAVINGS AMOUNT
- 02 TRANSMISSION TYPE
- 03 FUEL TYPE
- 04 MAX. SEATING
- 05 ENGINE TYPE
- 06 BODY TYPE
- 07 LISTING COLOUR
- 08 STATE

Model 7 Analysis

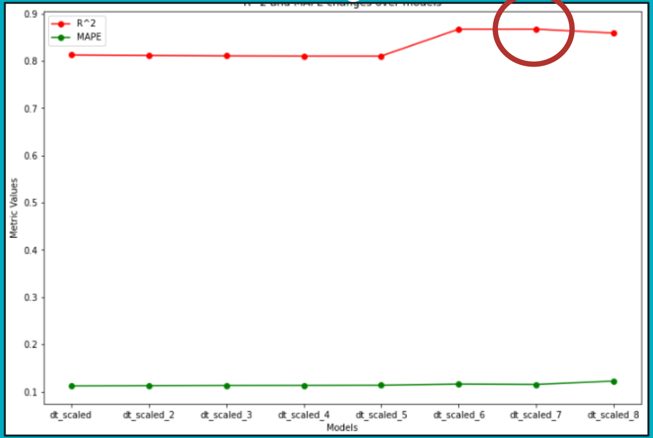
- Among the **Lowest RMSE**
- **R^2** goes up to **~0.86**
- **Abs. Mean Error** drops to **~150**
- **Variance** increases quite significantly to **~2.1**

- **Bias-Variance tradeoff – prioritize lower bias since test data shows model is less overfit**

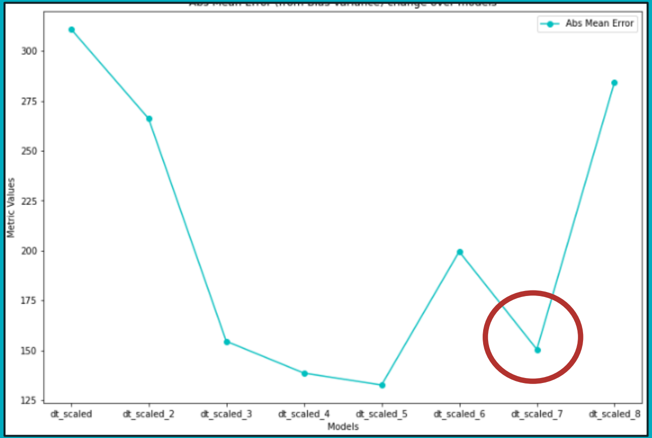
RMSE Change over Models



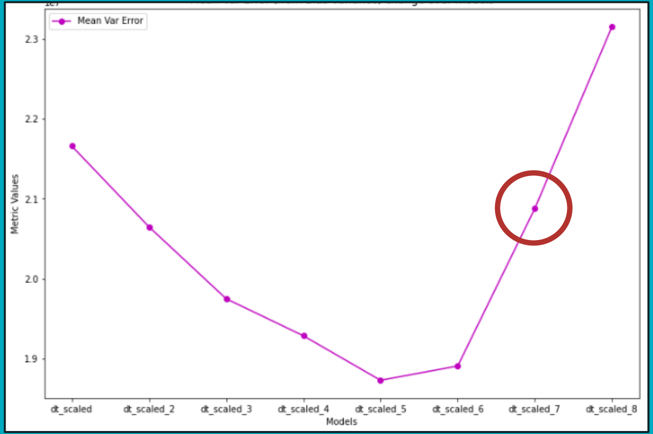
R^2 and MAPE Change over Models



Absolute Mean Error (Bias)



Mean Variance Error (Variance)



Final Model Variables

Is New
1st hand car

Make Name
Brand of Car

Max. Seating
Seats in the car: 4,5,6,7,8

Wheel system
4x2, All Wheel, Forward, Rear drive

State
US state

Horsepower
Vroom, vroom

Days on market
Duration car has been for sale

Mileage
Distance travelled

Seller rating
Proxy for trustworthiness

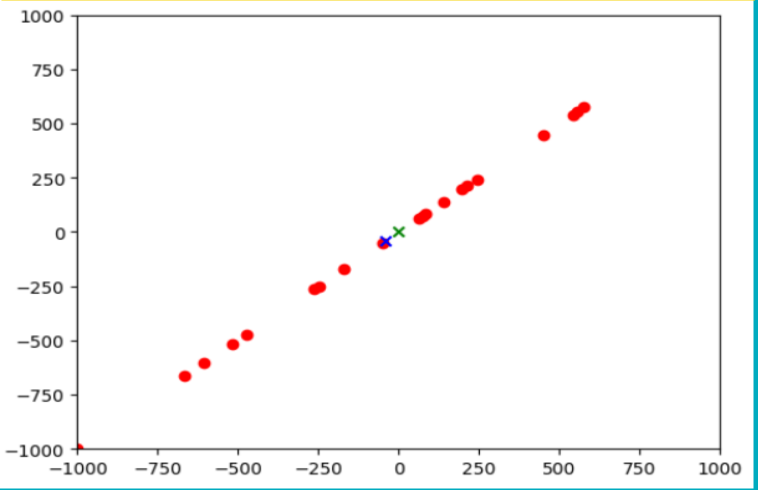
Final Decision Tree – Results on Test data

RMSE 5255

R^2 88.4%

MAPE 12.0%

Bias Variance, Mean Error: 43.1



K-fold Cross-Validation (10 splits)

Average R^2 84.5%

Average MAPE 10.0%

XGBoost model tried – results were not comparable

Preprocessing for multilinear regression

01 — Duplicate

Duplicated original dataset used

02 — Feature Selection

Based on Decision Tree model built, dropped columns that were not important

03 — Scaling

Only implemented Robust Scaling

04 — Train-Test Split

70-30 split without validation set

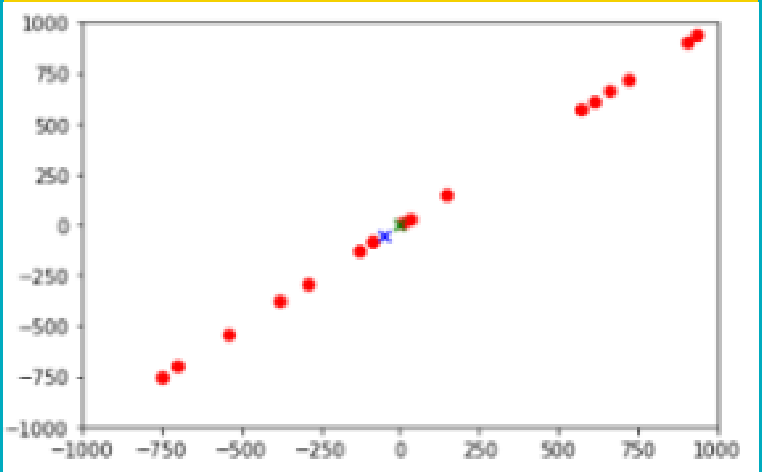
Final Multilinear Regression – Results on Test data

RMSE 6954

R^2 80%

MAPE 18.0%

Bias Variance, Mean Error: 53

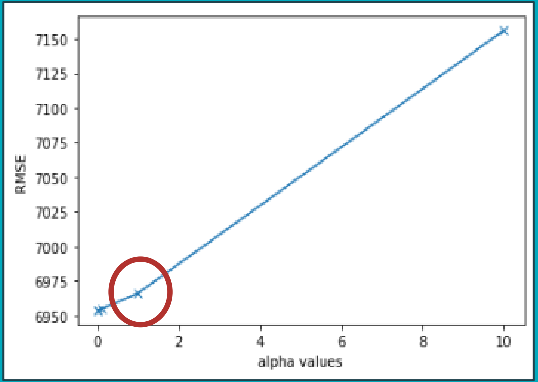


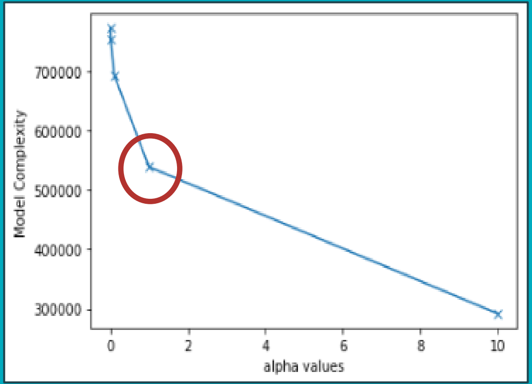
Optimal Alpha = 1

Regularization

Lasso Metrics by Alpha

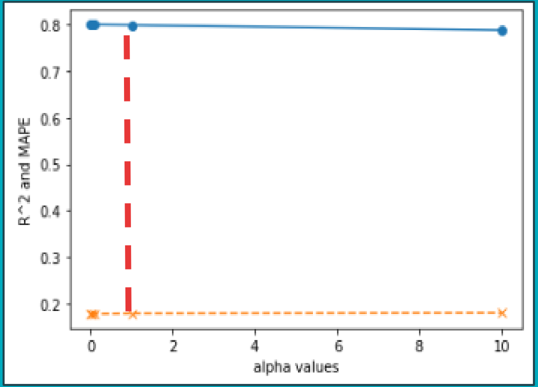
RMSE

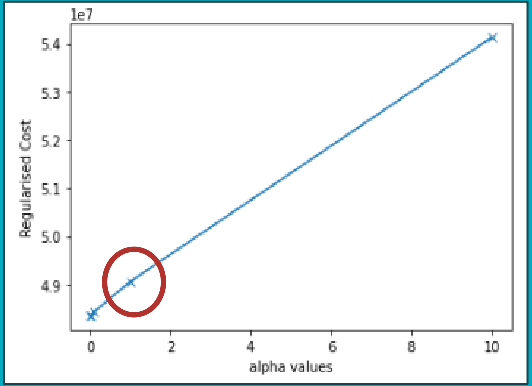




Model Complexity

R2 and MAPE





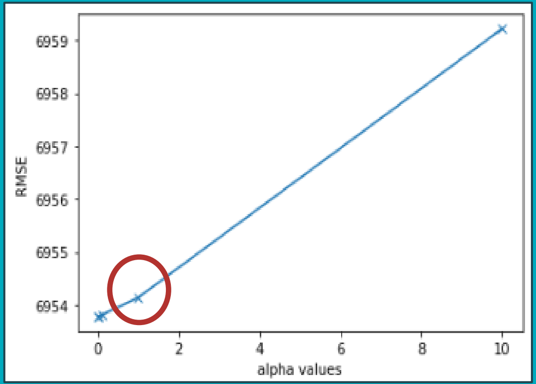
Regularized Cost

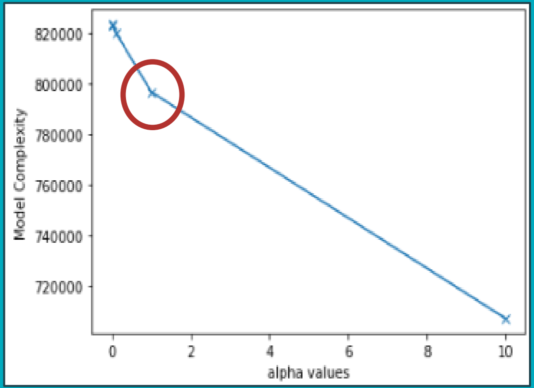
Regularization

Optimal Alpha = 1

Ridge Metrics by Alpha

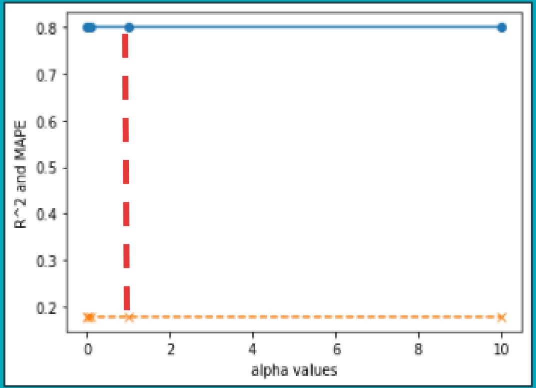
RMSE

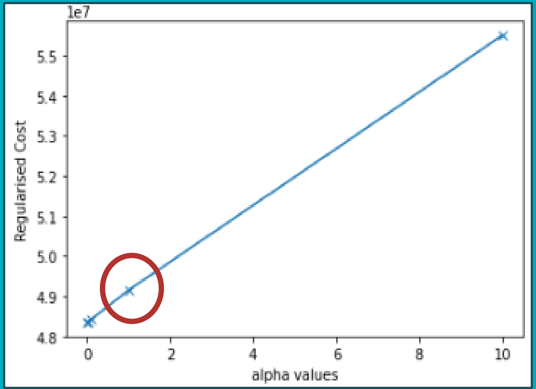




Model Complexity

R2 and MAPE





Regularized Cost

Preprocessing for neural network

01 Duplicate

Duplicated original dataset used

02 Feature Selection

Based on Decision Tree model built, eliminated columns that were not important

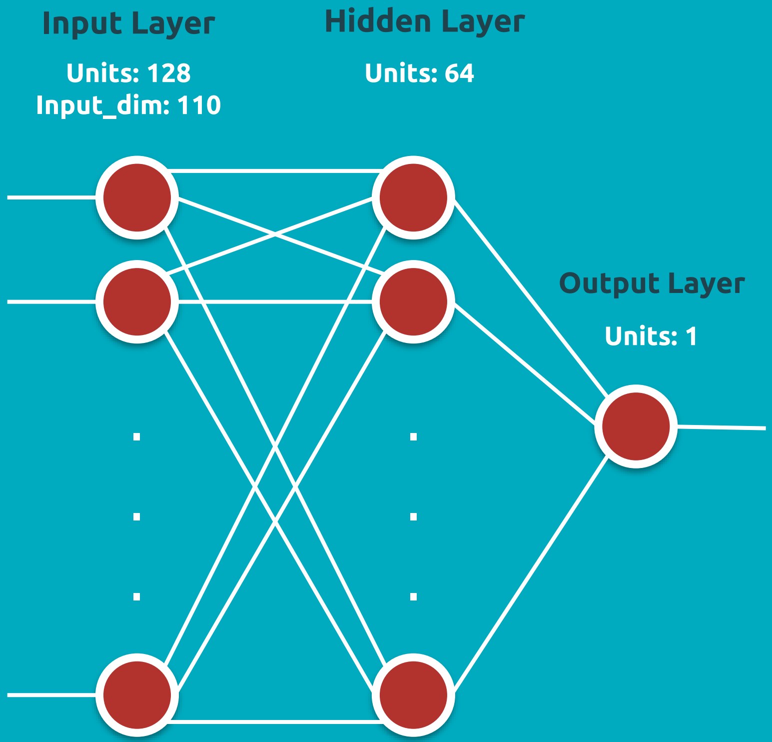
03 Scaling

Implemented both Robust and Standard Scaler

04 Train-Test-Val Split

60-20-20 split of train, test, and validation set

Model of the neural network



Parameters

- **Sample size:** 750k
- **Package used:** Tensorflow, Keras
- **Learning rates:** 0.001, 0.01, 0.025, 0.05, 0.1
- **Activation function:** Leaky ReLU
- **Epochs:** 100
- **Batch_size:** 1024
- **Evaluation metrics:** MAPE, R^2, RMSE

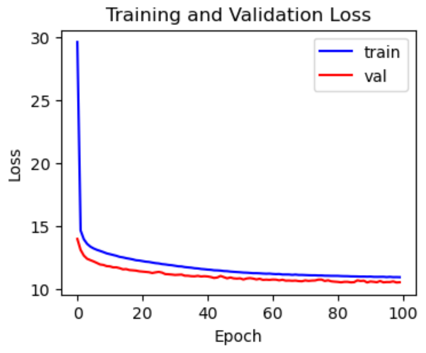
Final Neural Network – Results on Test data

RMSE **4943**

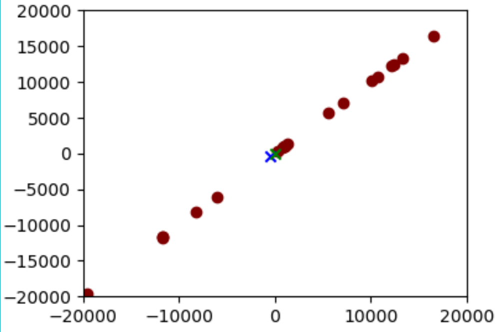
R^2 **89.2%**

MAPE **10.5%**

Learning Rate: 0.01



Bias Variance, Mean Error: -49.2



K-fold Cross-Validation (10 splits)

Average R^2 **85.7%**

Average MAPE **10.4%**

01

DATA

EDA & Processing

02

COMPUTER VISION

Feature Selection
Model Build & Analysis

03

PRICE PREDICTION

Feature Selection
Model Builds & Analysis

04

CONCLUSION

Learnings

Overall results

Final Model

	Decision Tree	MLR	Neural Network
RMSE	5255	5255	4943
R^2	88.4%	88.4%	89.2%
MAPE	12.0%	12.0%	10.4%

Limitations and Future Considerations

Limitations

Insufficient computing power

Transferability of model limited for SG, etc

Additional variables that are difficult to quantify (e.g. condition of car, market conditions)

Future Work

Performing Polynomial Regression

Regularization for Decision Tree

Optimizing NN (pruning, batch-size, epochs, optimizers)

Be Car-eful now

Thank you

APPENDIX

XGBoost with GridSearchCV and K-fold

Parameters

TUNED

n_estimators 200, 300, 400, 500

learning_rate 0.001, 0.01, 0.1, 0.2

FIXED

lambda 2

objective squared error

eval_metric MAPE

OTHERS

K-fold 10 splits, random shuffle

GridSearchCV MAPE

RMSE 10516

R^2 53.7%

MAPE 17.8%