

Predicting Prices of Used Cars for Potential Buyer in Singapore



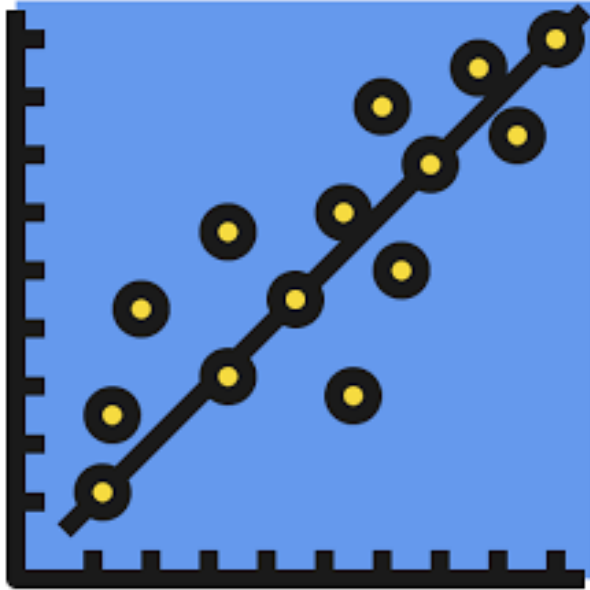
Prepared by Anita

Objective



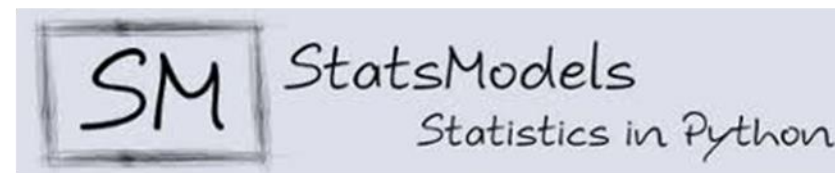
- To help potential buyer in Singapore to predict the prices of used cars as accurate as possible to be used as a benchmark before buying a used car.

Methodology





- Linear Regression Model
- Ridge Regression Model
- Lasso Regression Model
- Polynomial Regression Model

Tools Used




Data Collection



- Webscrapping from SgCarMart website using BeautifulSoup.


SGCARMART.COM Login | Sign up  Connect 






[New Cars](#) **Used Cars** [Rental Cars](#) [Sell My Car](#) [Directory](#) [Products](#) [Insurance](#) [Articles](#) [Forum](#) [Resources](#)




EUROPEAN ENGINEERING, NOW AFFORDABLE IN SMART WAYS  **REGISTER NOW** **ŠKODA**



FOR SALE BY OWNER **Post an Advertisement**
Sell it yourself! Advertise it at just **\$58 until it's SOLD!**
[Post an Ad](#) [Advertiser Login](#) [Ways of Selling](#)

Oct 2015 Toyota Vellfire 2.5A Z A-Edition At \$93,800
 1.98% Financing Rate. 1 Owner. Immaculate Condition. Mileage 91000km. Low Depre. [Roadlink Auto](#) [StarAd](#) 

PRIME MOTOR & LEASING
(CHENG YONG CREDIT ENTERPRISES PTE LTD GROUP OF COMPANIES)
With over 200 units of vehicles to choose from, ranging from all European to Japanese Makes. 

Browse by Category   Save Search  Print « Back (1 2 3 4 5 ... 281) Next » Sort by Date Posted  100 results/page 

28,013 vehicles **Search** Car Make / Model  Advanced Search  Search  View All

Make	Model	Price	Depreciation	Reg Date	Eng Cap	Mileage	Veh Type	Status
	<input type="checkbox"/> Mazda 2 HB 1.5A R (New 5-yr COE)	\$31,800	\$6,360 /yr	28-Sep-2010	1,498 cc	144,850 km	Hatchback	Available
Well Maintained Unit, Accident Free. Come Down To View It For Yourself! High In-house Loans And High Trade-ins Are Available. Conta...								
☆ Posted: 15-Sep-2020 Tags: 2010 Mazda 2, Mazda 2, Mazda, 2								
	<input type="checkbox"/> Honda Freed Hybrid 1.5A G 7-Seater Honda Sensing	\$84,800	\$8,880 /yr	28-Jan-2019	1,496 cc	13,000 km	MPV	Available
Fuel Type: Petrol-Electric 1 Owner! \$0 Down Full Loan Available! Low Mileage! Very Good Condition! Seldom Used! S Pass, EP Or LTV Pass Welcome To Loan! In...								
88 Motor Trading ☆ Posted: 15-Sep-2020 Tags: 2019 Honda Freed, Honda Freed, Honda, Freed								

Data Collection (cont...)

- Webscrapping from SgCarMart website using BeautifulSoup.

The screenshot displays the SgCarMart website interface. At the top, there's a navigation bar with links for New Cars, Used Cars (highlighted), Rental Cars, Sell My Car, Directory, Products, Insurance, Articles, Forum, and Resources. A search bar is also present. Below the navigation bar is a banner for 'CASUAL LUXURY AT SEA' featuring 'THE RITZ-CARLTON YACHT COLLECTION'. The main content area shows a search result for a 'Mazda 2 HB 1.5A R (New 5-yr COE)'. The car's details are listed in a table format, including Price, Depreciation, Mileage, Road Tax, Dereg Value, COE, Engine Cap, Curb Weight, Reg Date, Manufactured, Transmission, OMV, ARF, Power, and No. of Owners. To the right of the table, there are several images of the car, including a large interior view and smaller exterior views, along with a 'Location Map' button.

Mazda 2 HB 1.5A R (New 5-yr COE)			
Price	\$31,800		
Depreciation	\$6,360 /yr	Reg Date	28-Sep-2010 (5yrs COE left)
Mileage	144,850 km (14.5k /yr)	Manufactured	2009
Road Tax	\$684 /yr	Transmission	Auto
Dereg Value	N.A.	OMV	\$17,160
COE	N.A.	ARF	\$17,160
Engine Cap	1,498 cc	Power	77.0 kW (103 bhp)
Curb Weight	1,041 kg	No. of Owners	4

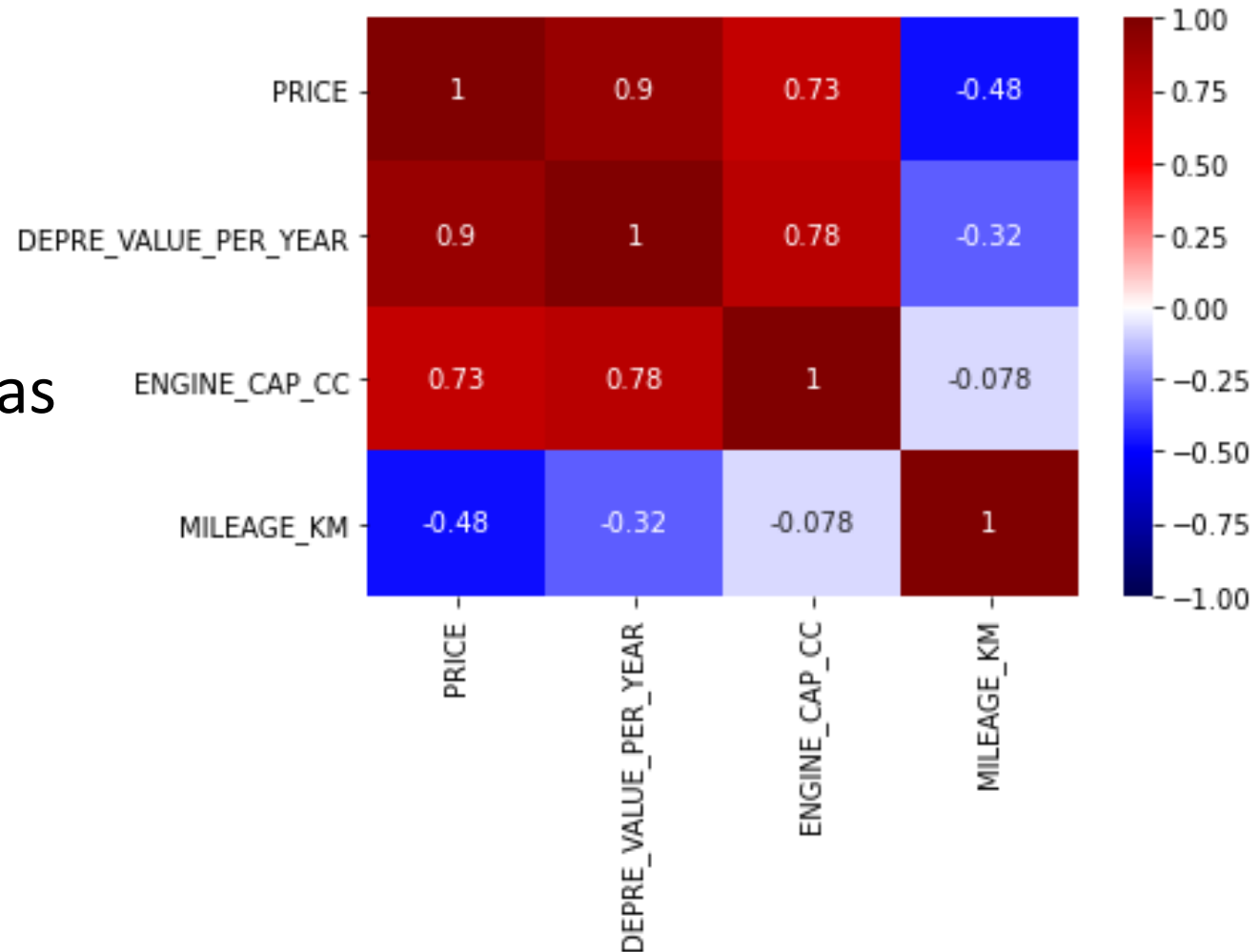
Data Cleaning



- 322 rows
- 6 columns:
 - Price (target)
 - Make (categorical feature)
 - Depreciation value per year (numerical feature)
 - Engine Cap cc (numerical feature)
 - Mileage km (numerical feature)
 - Vehicle Type (categorical feature)

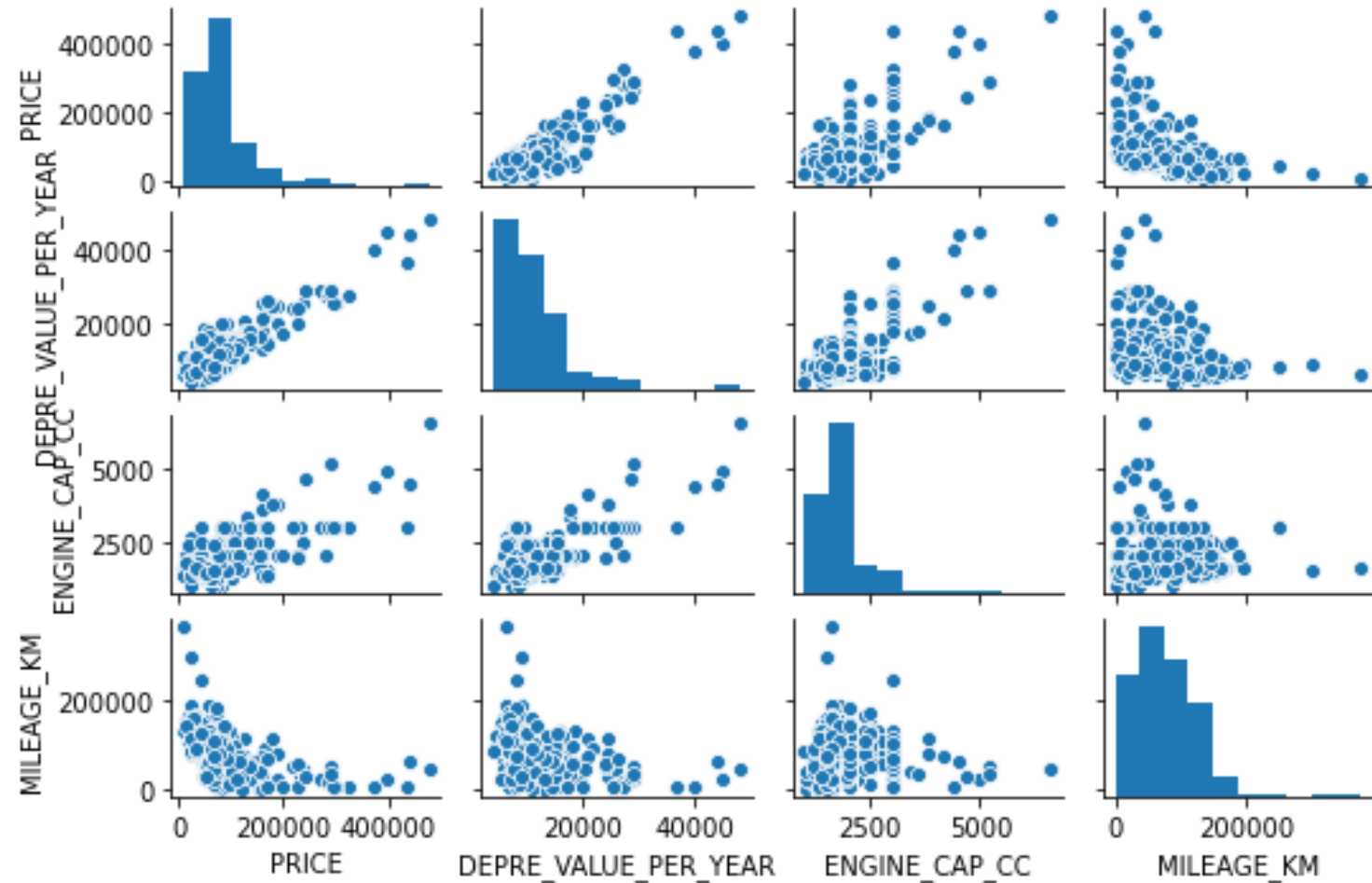
Data Analysis

- To view the correlation between feature to feature and feature to target.
- No feature is removed as there is no high correlation between features.



Data Analysis (cont...)

- To view the distribution plot of features and target.



Data Analysis (cont...)

- Fit in statsmodels
- Based from the p-value, all features are significant.
- Adj. R-squared = 0.859

OLS Regression Results

Dep. Variable:	PRICE	R-squared:	0.861
Model:	OLS	Adj. R-squared:	0.859
Method:	Least Squares	F-statistic:	654.4
Date:	Tue, 15 Sep 2020	Prob (F-statistic):	1.14e-135
Time:	16:48:58	Log-Likelihood:	-3723.9
No. Observations:	322	AIC:	7456.
Df Residuals:	318	BIC:	7471.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	-7642.9409	4923.592	-1.552	0.122	-1.73e+04	2043.990
DEPRE_VALUE_PER_YEAR	7.4692	0.396	18.854	0.000	6.690	8.249
ENGINE_CAP_CC	17.0354	3.367	5.060	0.000	10.412	23.659
MILEAGE_KM	-0.3456	0.033	-10.516	0.000	-0.410	-0.281

Omnibus:	23.665	Durbin-Watson:	2.065
Prob(Omnibus):	0.000	Jarque-Bera (JB):	86.289
Skew:	0.035	Prob(JB):	1.83e-19
Kurtosis:	5.535	Cond. No.	3.11e+05

Data Analysis (cont...)

- Log transform the mileage
- Fit in statsmodels
- Based from the p-value, all features are significant.
- Adj. R-squared = 0.875

OLS Regression Results

Dep. Variable:	PRICE	R-squared:	0.876
Model:	OLS	Adj. R-squared:	0.875
Method:	Least Squares	F-statistic:	751.3
Date:	Tue, 15 Sep 2020	Prob (F-statistic):	6.09e-144
Time:	16:49:02	Log-Likelihood:	-3704.6
No. Observations:	322	AIC:	7417.
Df Residuals:	318	BIC:	7432.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	1.636e+05	1.59e+04	10.262	0.000	1.32e+05	1.95e+05
DEPRE_VALUE_PER_YEAR	7.3177	0.370	19.792	0.000	6.590	8.045
ENGINE_CAP_CC	18.9153	3.179	5.951	0.000	12.661	25.169
np.log(MILEAGE_KM)	-1.821e+04	1417.194	-12.852	0.000	-2.1e+04	-1.54e+04

Omnibus:	46.742	Durbin-Watson:	2.079
Prob(Omnibus):	0.000	Jarque-Bera (JB):	116.250
Skew:	-0.696	Prob(JB):	5.71e-26
Kurtosis:	5.593	Cond. No.	1.61e+05

Data Analysis (cont...)

- Create dummy variables into dataset.



Data Analysis (cont...)

- After fit in statsmodels again, based from the p-value, only following features are significant:
 - Depreciation value
 - Engine cap
 - Log mileage
 - Make Ferrari
 - Make Mini
 - Make Rolls Royce
 - Vehicle type SUV
- Adj. R-squared = 0.882

OLS Regression Results

Dep. Variable:	PRICE	R-squared:	0.885			
Model:	OLS	Adj. R-squared:	0.882			
Method:	Least Squares	F-statistic:	345.3			
Date:	Tue, 15 Sep 2020	Prob (F-statistic):	2.26e-143			
Time:	16:50:49	Log-Likelihood:	-3692.8			
No. Observations:	322	AIC:	7402.			
Df Residuals:	314	BIC:	7432.			
Df Model:	7					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	1.676e+05	1.59e+04	10.523	0.000	1.36e+05	1.99e+05
DEPRE_VALUE_PER_YEAR	7.0185	0.372	18.864	0.000	6.286	7.751
ENGINE_CAP_CC	17.8814	3.157	5.664	0.000	11.670	24.093
LOG_MILEAGE_KM	-1.822e+04	1395.843	-13.054	0.000	-2.1e+04	-1.55e+04
MAKE_Ferrari	8.341e+04	2.48e+04	3.364	0.001	3.46e+04	1.32e+05
MAKE_MINI	-5.858e+04	2.36e+04	-2.484	0.014	-1.05e+05	-1.22e+04
MAKE_Rolls_Royce	5.241e+04	2.55e+04	2.055	0.041	2240.757	1.03e+05
VEHICLE_TYPE_SUV	6301.6449	3284.534	1.919	0.056	-160.833	1.28e+04
Omnibus:	49.299	Durbin-Watson:	2.055			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	157.598			
Skew:	-0.647	Prob(JB):	6.00e-35			
Kurtosis:	6.173	Cond. No.	2.77e+05			

Cross Validation

Linear Regression Cross Val Score: [0.93333532 0.52448046 0.87522598 0.8712238 0.88953408]
Mean cv r^2 : 0.819 +- 0.149

Ridge Cross Val Score: [0.92969153 0.54775207 0.86333013 0.86641585 0.89195441]
Mean cv r^2 : 0.82 +- 0.138

Lasso Cross Val Score: [0.93296689 0.52822619 0.87501454 0.87075869 0.88946554]
Mean cv r^2 : 0.819 +- 0.147

Degree 3 Poly Regression Cross Val Score: [-0.22285902 0.12113914 0.86517433 0.91012124 0.90049184]
Mean cv r^2 : 0.515 +- 0.475

- It seems like Ridge Regression provides the highest R^2 as compared to others. Therefore, will choose to use Ridge regression.

Results

Prediction

Ridge Regression RMSE - train: 24160.46597176425

Ridge Regression R2 Score - train: 0.8874145159737489

Ridge Regression RMSE - test: 18951.41407514264

Ridge Regression R2 Score - test: 0.8436070187572133

- From the train dataset, 89% of data variation explained by model and root mean squared error is \$24,160.
- From the test dataset, 84% of data variation explained by model and root mean squared error is \$18,951.

Results

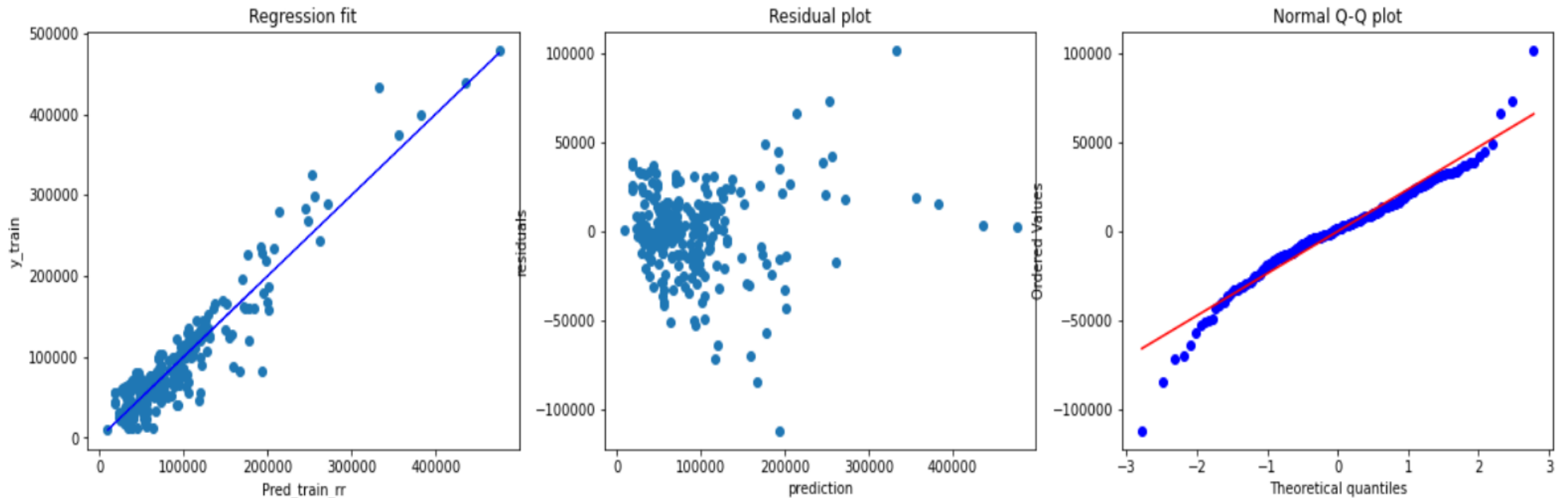
Coefficients

```
[('DEPRE_VALUE_PER_YEAR', 44577.66498575446),  
 ('ENGINE_CAP_CC', 13419.433108206378),  
 ('LOG_MILEAGE_KM', -19028.67403481724),  
 ('MAKE_Ferrari', 5508.863705839609),  
 ('MAKE_MINI', -3433.150015257121),  
 ('MAKE_Rolls_Royce', 3655.35391418426),  
 ('VEHICLE_TYPE_SUV', 3408.6844210868144)]
```

- Depreciation value, engine cap, make Ferrari, make Rolls-Royce and vehicle type SUV have positive impact on the price of used cars.
- Log mileage and make Mini have negative impact on the price of used cars.

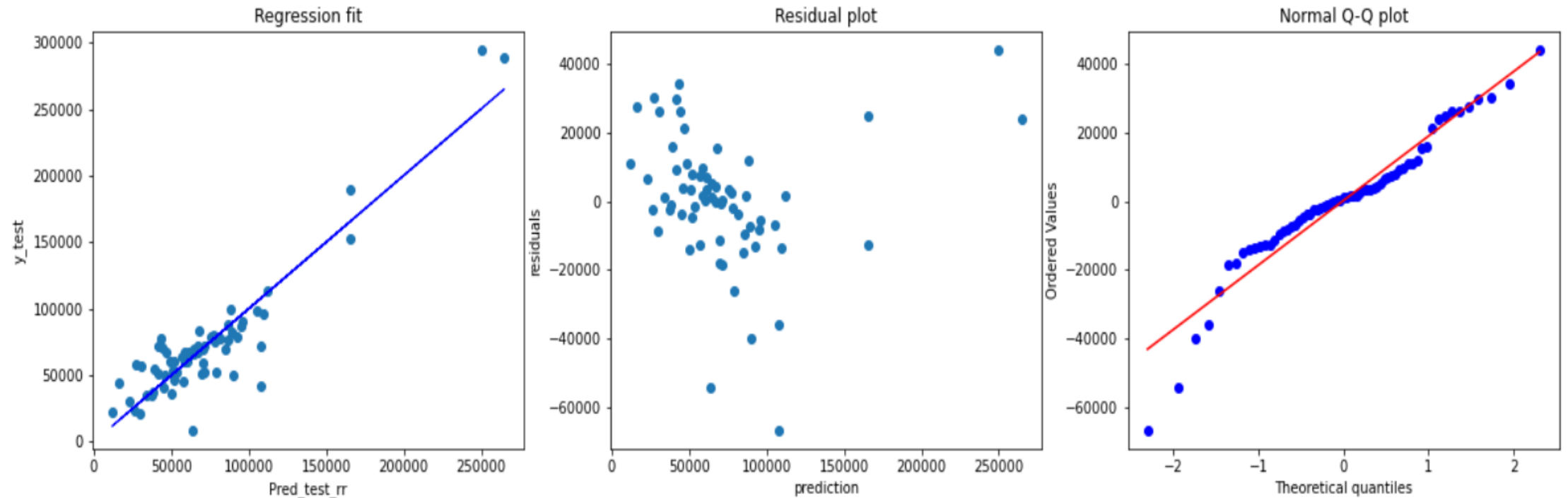
Linear Regression Assumption

- Train Dataset



Linear Regression Assumption

- Test Dataset



Conclusions



- This model still cannot do prediction accurate enough, with RMSE \$18,951 from test dataset.
- This model can only explain 84% of data variation from test dataset.
- Other features that could affect the price of used cars are not captured.
- Observation data collected are not big enough.
- Outliers in dataset was not investigated and removed from dataset.



**THANK
YOU**