Capstone Project: Used Cars Price Prediction

By: Marcus Chan

Executive Summary:

The model chosen to accurately predict second hand cars in India is the Random Forest Regressor. Data we used to make this decision was collected through 2010 – 2019 We found that this model responded the best to the data set and performed better than other test models with low RMSE scores. Through tests we determined that imputing all variables where possible resulted in the overall best RMSE scores for test and training data. Regarding the model the Random Forest Regressor creates multiple decision trees, making understanding the process of the decision like a black box operation. However, we are able to determine the most significant variables of the decision allowing a general walk through of the evaluation of the vehicle if demanded. It is recommended that the implementation of this model comes through in several stages. Regularly check model results at the head office and send out the optimization data. Then fit the data over at the dealerships, and implement it to onsite machines, and websites to allow consumers to enjoy the fastest service predicting car price.

Problem Summary:

As India's GDP grows and improves the quality of their citizen's lives it created a demand for vehicles and transportation. Thus, in the Indian market a was a huge demand in cars was created. Then as the market saturated with new cars, it has slowed down in recent years. This is due to the increased interest in second-hand cars causing used cars sales to grow to fill the demand gap. The price for new cars were dictated in large part to the OGM, however reselling the car has no such standard. The issue for emerging corporations such as Cars4U is listing a price that is acceptable for consumers to sell a considerable portion of their stock. The key objective for this project is the determine a pricing model that will fulfil these requirements. The records of already sold second-hand cars allow us to model pricing after successful negotiations and streamline the process so that selling a car can be done more efficiently and with some level of expectation. The analysis and discussion of these model will help understand the kinds of factors that determine vehicle pricing for the current retain environment. The data that we use for this project come from 2010 – 2019 a period of time that hasn't had any major global shocks that would destabilize the markets.

Solution Design:

Several different regression models were tested to determine their efficacy on data collected as early as 2019. In addition to the different regression models, we also tested the results against different preprocessing methods and determine which methods offer the best performance given the best kind of model. We determined that the most successful model is a Random Forest Regressor that has been hyper tuned. As observed below in Figure 1 the Random Forest Regressor has a RMSE value of 2.09 for training data and 4.408 for test data.

]:

	Model	Train_r2	Test_r2	Train_RMSE	Test_RMSE
0	Linear Regression	0.739520	0.754674	5.608260	5.203630
1	Decision Tree	0.831597	0.787488	4.509368	4.843134
2	Ridge	0.735656	0.751703	5.649706	5.235044
3	Lasso	0.730818	0.750121	5.701171	5.251696
4	ElasticNet	0.732668	0.753339	5.681550	5.217768
5	Decision Tree Tuned	0.863602	0.764046	4.058312	5.103273
6	Random Forest Tuned	0.963723	0.823909	2.092937	4.408632
7	KNN Regressor	0.999899	0.519604	0.110492	7.281729

Figure 1: Chart of RMSE of Regression Models

Based on the readings of the other models used the Random Forest model is the more accurate in its predictions given a test and training set by a significant margin. As discussed later linear regressors not KNN and Decision Tree models returned extremely similar results. This is due to the conditions surrounding the dataset used for training that pigeonholed the linear regressors to certain results such that all of their predictions are graphed on a scatter plot against the true values in the same way. Examples of these can be found in the Appendix.

Analysis and Key Insights:

Some of the data collected from the data set have an inherent skewness, and so a derived variable is employed to make their histogram graph more normal. The effect of these variables added to the dataset is to add multicollinearity to the dataset. This would reduce the validity of our results so the original variables are removed. To remove any other multicollinearity in the dataset VIF values are employed and kept within a score of 5. Both of the dataset that we employ must be processed in this method first.

Due to the method that we use to preprocess the data several of the models that we tested the Random Forest Model against are equivalent to a linear regression model. The Lasso Model is better against unscaled data and Ridge Data is better against multicollinearity. However due to our preprocessing these 2 models are essentially the same as linear regression.

While testing the multiple datasets we keep in mind that they have a common base and the preprocessing is largely the same as well. Both data sets filled missing data based on data found from rows with a matching brand and model name. Where they differ is one data set is imputed completely where filling all NA values, the other is also imputed but only after dropping rows from a large section of NA values. We used the data set with rows cut to determine the model efficacy since so much data was imputed on the other data set, we could not be confident about the validity of the results.

However we found that the imputed data performed better over all with a RMSE of 2.75 and 3.32 for training and testing sets respectively. There is roughly an 8 % increase in performance and offsets the signs of overfitting seen in the model fitted with the data dropped. This means when possible that the data set should be scaled and imputed when possible, to optimize the accuracy of the predictions.

Another insight is that once the model parameters been properly optimized that the model git with a data set approximately 7000 entries large takes about 1 second to fit. If the model is fitted and parameters hyper tuned, the prediction is almost instant, despite the model containing 120 instances of decision trees. As for the hyper tuning that process takes about a minute assuming a non exhaustive list of parameters.

This may indicate that there could be a 3-stage implementation method for this model with different level of the corporation handling each stage depending on the scale of the company. The consumer should obviously be working with just the prediction at either a set location of a place where they can easily access. This could be through an implementation of a website for example. As for the fitting and determining the hyper tuning that could be done at a parking lot where a 1 second delay isn't too bad. Then finally for the training this would have to be done elsewhere as a one-minute delay is too long to wait on a service such as quoting a price. However, the model could be exported to their distribution centers and dealerships.

Limitations and Recommendations for Further Analysis:

As the nature of the model chosen it is bit like a black box, since the Random Forest takes into consideration all iterations of decision trees and aggregates the results as a vote to determine the prediction. Then on top of that each decision tree has an increasing complexity based on the number of significant variables. It makes is difficult to limit the decision tree when a large number of variables are significant, causing a sprawling mess of a tree as shown below.

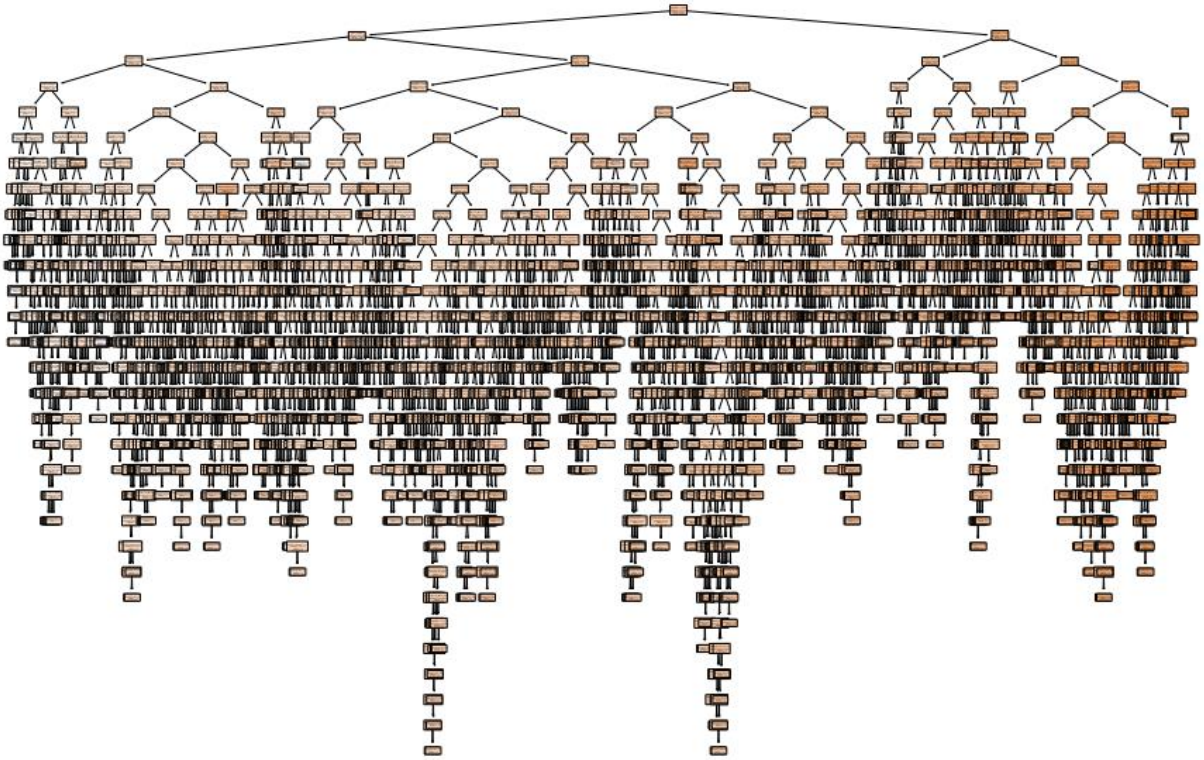

Figure 2: 1 Decision Tree in Random Forest

It is impossible to easily distinguish factors that determine pricing from Figure 2. However, we are able to extract a list of variables that dictate the order of significance that goes into the pricing. This should dissuade the consumer from looking too deep into the pricing model decisions. From Figure 3 we can see that the key values that influence the decision trees of the random forest like in Figure 2 the power, the year and the Mileage are the most important factors.

	Imp		
log_power	0.698079	Location_Jaipur	0.000874
Year	0.201473	Brand_Ford	0.000862
Mileage	0.033002	Brand_Tata	0.000737
new_price_log	0.023373	Brand_Mahindra	0.000726
Seats	0.010519	Brand_Renault	0.000664
Fuel_Type_Petrol	0.002056	Brand_Nissan	0.000595
Location_Pune	0.001680	Brand_Audi	0.000569
Location_Kochi	0.001642	Owner_Type_Third	0.000479
Brand_Hyundai	0.001638	Brand_Skoda	0.000463
Owner_Type_Second	0.001606	Brand_Land	0.000393
Location_Kolkata	0.001599	Brand_Chevrolet	0.000387
Transmission_Manual	0.001588	Brand_Jaguar	0.000320
Location_Mumbai	0.001588	Brand_Fiat	0.000256
Brand_Toyota	0.001583	Brand_Isuzu	0.000231
Location_Hyderabad	0.001469	Brand_Mini	0.000086
Location_Delhi	0.001386	Brand_Mitsubishi	0.000056
Location_Chennai	0.001300	Brand_Datsun	0.000049
Location_Coimbatore	0.001282	Brand_Jeep	0.000042
Brand_Honda	0.001233	Brand_Volvo	0.000041
Location_Bangalore	0.001147	Brand_Porsche	0.000016
Brand_Volkswagen	0.001073	Owner_Type_Fourth & Above	0.000011
Brand_Mercedes-Benz	0.000910	Fuel_Type_LPG	0.000008
Brand_BMW	0.000906	Fuel_Type_Electric	0.000003
		Brand_Smart	0.000000

Figure 3: Significance List of Random Forest

Like all models the Random Forest is limited by the quality of data that is processed into the model. With the current models we were fortunate that we were able to fill in a largely missing column from other rows such that the imputation had positive results. However, if too much of the data set is missing it could mislead the imputation, causing to drop either the rows or the column entirely an overall better solution. This limits the influence of the imputation but also reduces the number of entries that we can draw data from for our model.

At that point it, an extremal library to generate synthetic data should be utilized to provide a better coverage of values. It is not ideal but the best solution given the quality of data provided to the system.

Recommendations for Implementation:

One way to implement this across the company depends on the scale of the operation. There are certain assumptions that the corporation has multiple locations to accept buying back the second-hand cars and an office to collaborate the redistribute of them to resell. The hyper-tuning takes the longest amount of time to operate so we could operate the hyper tuning out of the office keeping track of the records. This allows the model easier access to the data while also not impacting the rest of the operations of the business.

Then that office could export the model weights to the offices around the country where the model could be fit with recent data. That should provide a any consumer that wants a prediction a price the is suitable for a local level. A prediction could be carried out almost instantly once the parameters of their vehicle are fit on the local level.

Finally, a model should also be available to the consumer via a website so they can measure their vehicles before they get to the car resale dealership.

Then each will have to be done again on a periodic timescale since 1 model might not hold true forever especially if what is valued by people changes. If the Random Forest Regression Model is hyper tuned every month it should cover any kind of changes well enough to capture the general trend. Then the model could be refit with new data every 2-3 days, the time taken isn't very long and could be handled onsite. Finally both of these updates should also be reflected on the website as quickly as possible so consumers can reference the most up to date models.

Appendix

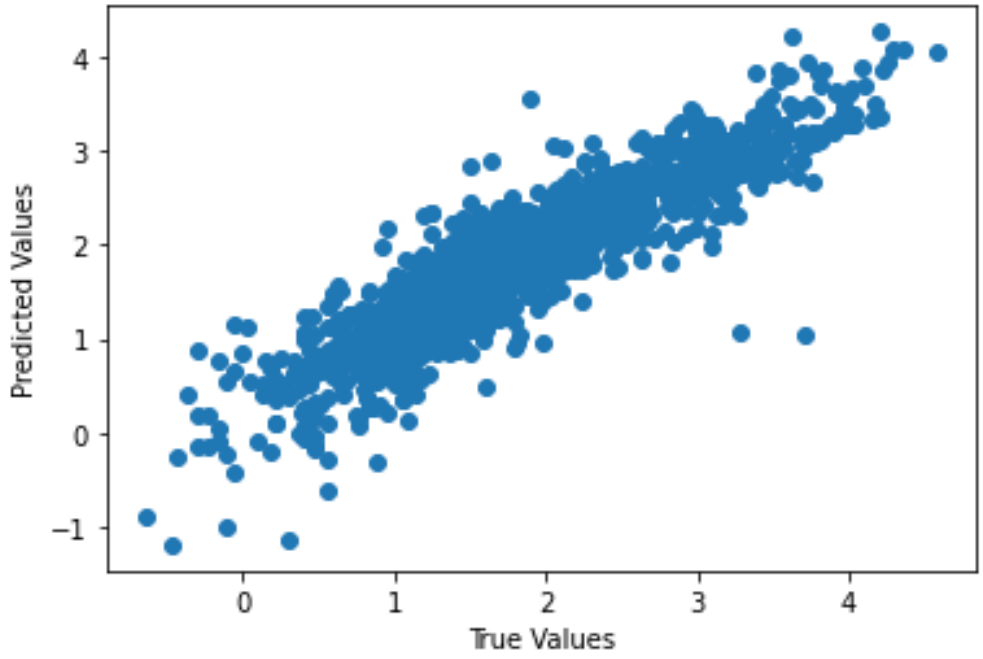

Figure 4 Scatter Plot of Linear Regression Prediction vs. True Values

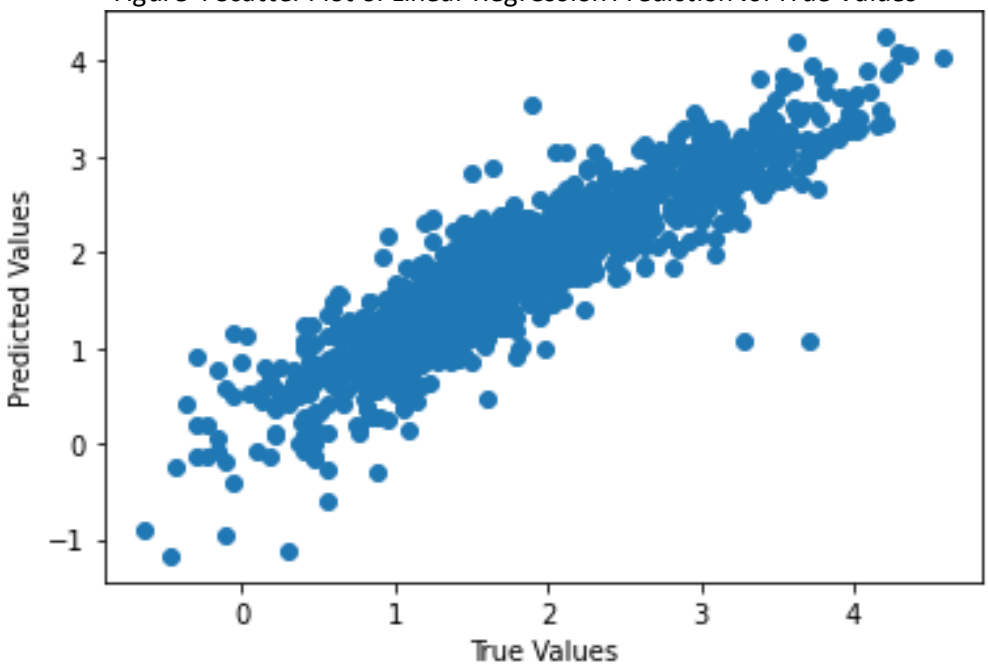

Figure 5 Scatter Plot of Ridge Regression Prediction vs. True Values

VIF Scores:

Year	1.206202
Mileage	1.716580
Seats	1.156330
new_price_log	1.102872
log_power	1.538614
Location_Bangalore	2.498506
Location_Chennai	2.897784
Location_Coimbatore	3.549097
Location_Delhi	3.099061
Location_Hyderabad	3.700127
Location_Jaipur	2.633354
Location_Kochi	3.503737
Location_Kolkata	3.130878
Location_Mumbai	3.987681
Location_Pune	3.373652
Fuel_Type_Electric	1.008857
Fuel_Type_LPG	1.006785
Fuel_Type_Petrol	1.364012
Transmission_Manual	2.038112
Owner_Type_Fourth & Above	1.005935
Owner_Type_Second	1.080204
Owner_Type_Third	1.050748
Brand_Audi	1.580677
Brand_BMW	1.602831
Brand_Chevrolet	1.081275
Brand_Datsun	1.013028
Brand_Fiat	1.018877
Brand_Ford	1.199712
Brand_Honda	1.388874
Brand_Hyundai	1.574994
Brand_Isuzu	1.005744
Brand_Jaguar	1.097582
Brand_Jeep	1.018022
Brand_Land	1.137018
Brand_Mahindra	1.225315
Brand_Mercedes-Benz	1.617275
Brand_Mini	1.057610
Brand_Mitsubishi	1.013976
Brand_Nissan	1.062628
Brand_Porsche	1.019186
Brand_Renault	1.098270
Brand_Skoda	1.166429
Brand_Smart	1.013406
Brand_Tata	1.111333
Brand_Toyota	1.324883
Brand_Volkswagen	1.210310
Brand_Volvo	1.037909

dtype: float64

Figure 6 VIF Scores for Given Data Set

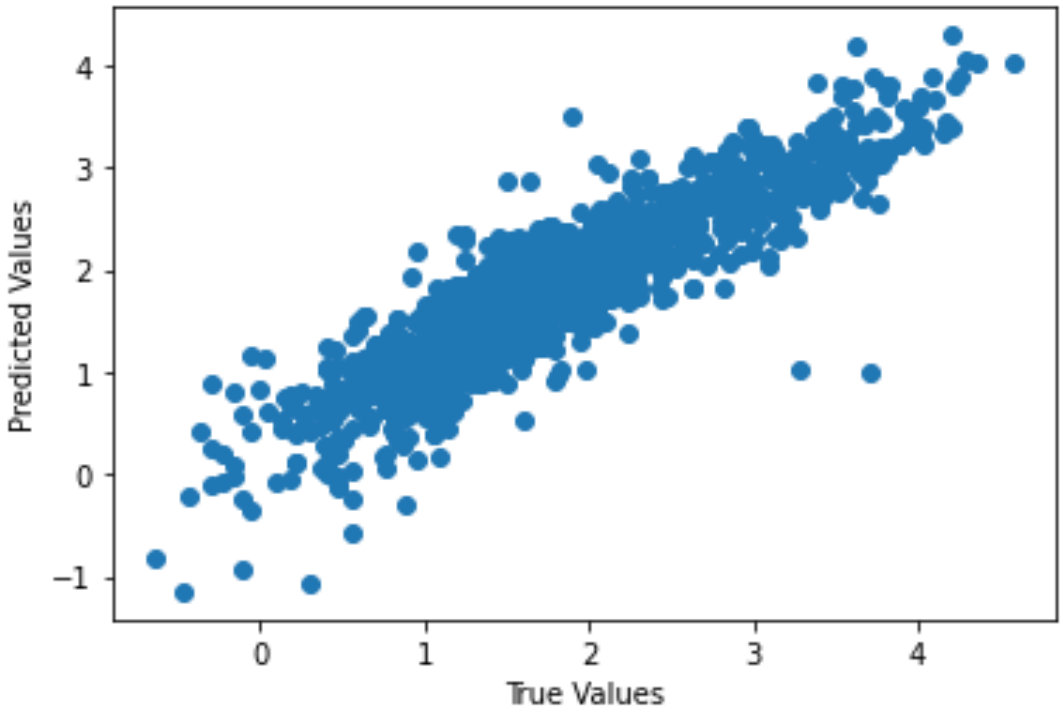

Figure 5 Scatter Plot of Lasso Regression Prediction vs. True Values