Capstone Project on Used Car Price

Executive Summary

Used car market is rapidly growing in India now. The price of used car keeps rising dramatically. There are significant amount of used car buyers and sellers that it is hard to obtain specific car data and the fair price because of lack of information.

We are a start-up car dealer named Cars4U. We research how we can make effective business with pricing model.

Intended goal is making a pricing model. Predict the price in terms of the value of the used car using this prediction model during the market goes too fast and rapidly growing. Using prediction model of used car pricing helps us to create more new strategy to make profitable business which provides both buyers and sellers are satisfied with the valuation.

This project proposes the Gradient Boost model for the prediction of used car pricing in order to solve the current business issue with the hardness to price appropriately and quickly while the used car market in India is rapidly expanding. The suggested model is able to predict the price with high speed, high accuracy, and optimal interpretability and it is flexibly handle the data for outliers. That can solve not only predicting the price itself, but also it can reduce the time to evaluate car, time to decide if consumers purchase the used car, the cost regarding pricing such as labor fee to evaluate and process the price of used car.

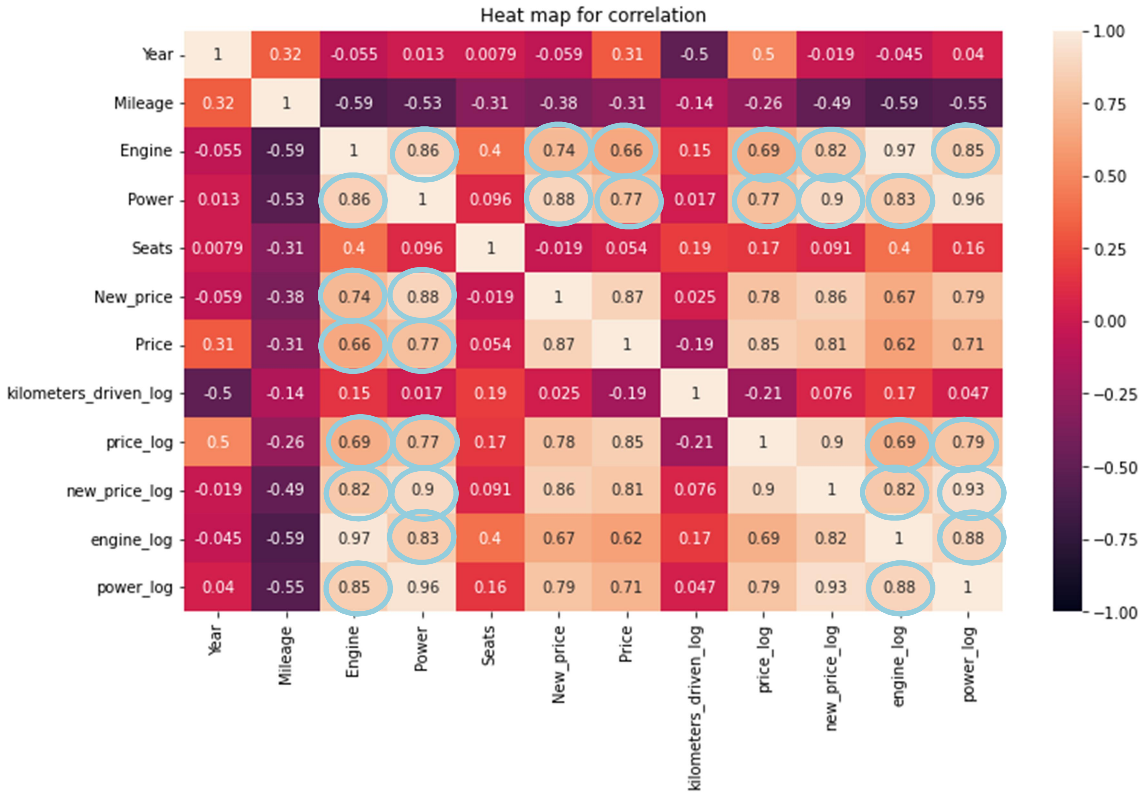
It can utilize the marketing strategies and inventory management for used car as well. However, the model is still not perfect which needs to improve the remaining error. We may think about ensemble methods to improve it. In order to be more accurate model, it needs more data. The model needs to up-to-date, maintain, improve to adapt the market and consumers' preference and to be competitive in the market to overcome to the competitors.

Problem Summary

It is important to solve this problem because there are a lot of used car buyers and sellers uncertainty about the price while the market is growing too fast. The key objective of this project is to make a pricing model with machine learning which effectively predicts the price of used cars that helps us to create more new strategy to make profitable business which provides both buyers and sellers are satisfied with the valuation.

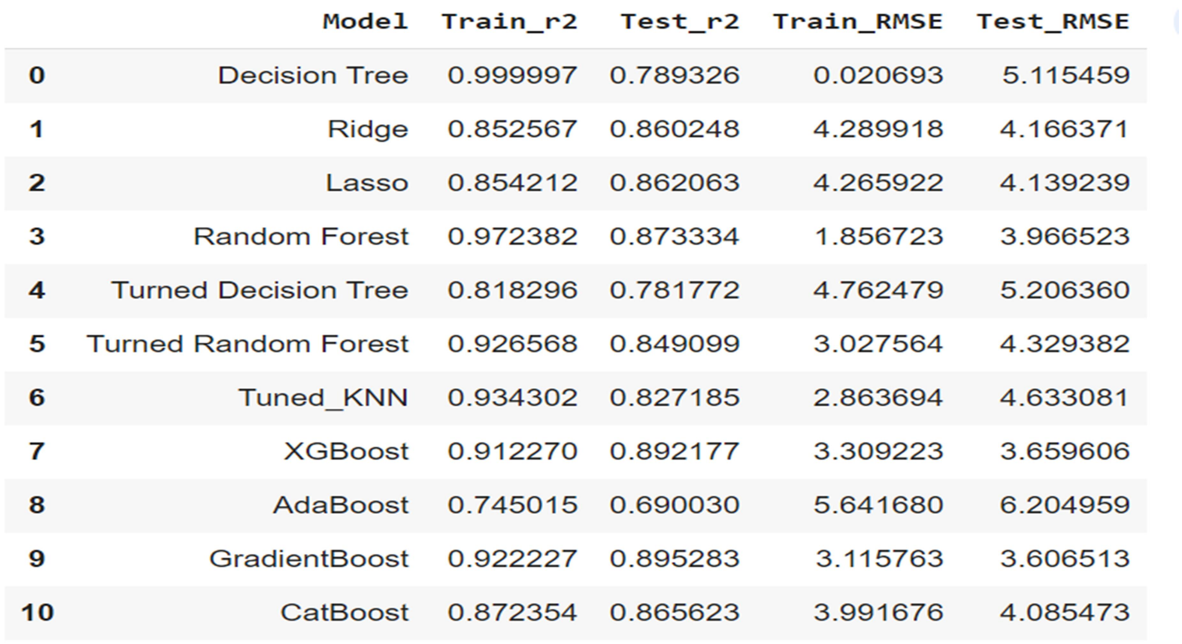
Solution Design and Summary

We explore the data as the first step. We found some variables are strongly correlated. The blue circles mark high correlation as below in Figure1 and also Figure2 in Appendix. Engine and Power are high correlated each other. The two are strongly correlated Price, New_price, price_log and New_price_log as well.



There are a lot of machine learning regression methods and models were examined in terms of identify which method and model are the best fit with the used car pricing prediction. The methods and models that were explored are total 14 models with Linear Regression, Statistics model, Ridge regression, Lasso regression, KNN, Tuned KNN, Decision Tree, Random Forest, Tune Decision Tree, Tuned Random Forest, XGBoost, AdaBoost for Decision Tree, Gradient Boost, and CatBoost. Please take a look at the Table1: It is comparison of the results of models (Some of models are not included.)

Table1



	Model	Train_r2	Test_r2	Train_RMSE	Test_RMSE
0	Decision Tree	0.999997	0.789326	0.020693	5.115459
1	Ridge	0.852567	0.860248	4.289918	4.166371
2	Lasso	0.854212	0.862063	4.265922	4.139239
3	Random Forest	0.972382	0.873334	1.856723	3.966523
4	Turned Decision Tree	0.818296	0.781772	4.762479	5.206360
5	Turned Random Forest	0.926568	0.849099	3.027564	4.329382
6	Tuned_KNN	0.934302	0.827185	2.863694	4.633081
7	XGBoost	0.912270	0.892177	3.309223	3.659606
8	AdaBoost	0.745015	0.690030	5.641680	6.204959
9	GradientBoost	0.922227	0.895283	3.115763	3.606513
10	CatBoost	0.872354	0.865623	3.991676	4.085473

Gradient Boost provides the best performance and accuracy in terms of adaptiveness of the dataset and the level of effectivity in the models what I explored. (XGBoost looks like very similar result as Gradient Boost. I am going to talk about why I don't choose XGBoost later on.)

Please take a look at Table2 that shows the results of R-square and RMSE on train data and test data of Gradient Boost. The R-square on train data and test data are a good balance with the gap is approximately 0.03 as train data of R-square with 0.92 and test data of R-square with 0.89. It may a little overfitting but the difference of 0.03 is very small relatively. Less than 0.15 can be a good balance. The score of R-square is higher is better. The score indicates how much it can explain the data. This combination of R-square is the highest score in the models. The RSME on train data and test data are relatively low in the models. The score of RSME is lower is better. The score indicates how much it is accurate.

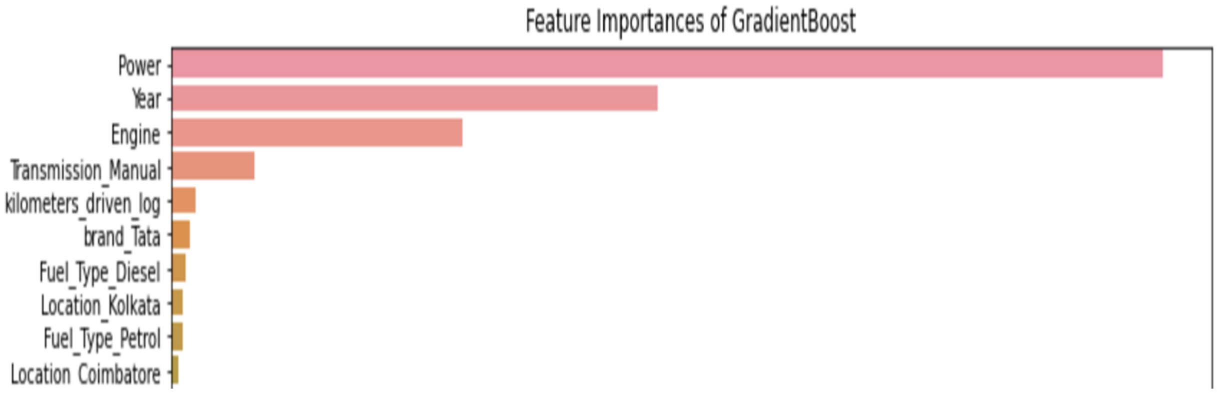
Table2

Model	GradientBoost
Train_R2	0.92
Test_R2	0.89
Train_RMSE	3.12
Test_RMSE	3.60

This result is the best in the models that are examined. Gradient Boost is able to perform better than others. This is because it is boosting method which minimizes loss from the previous learning. And then 2nd learning improves its learning. And then 3rd learning continues the process. This characteristic helps to build the better model. I conducted hyperprarmeter tuning as well.

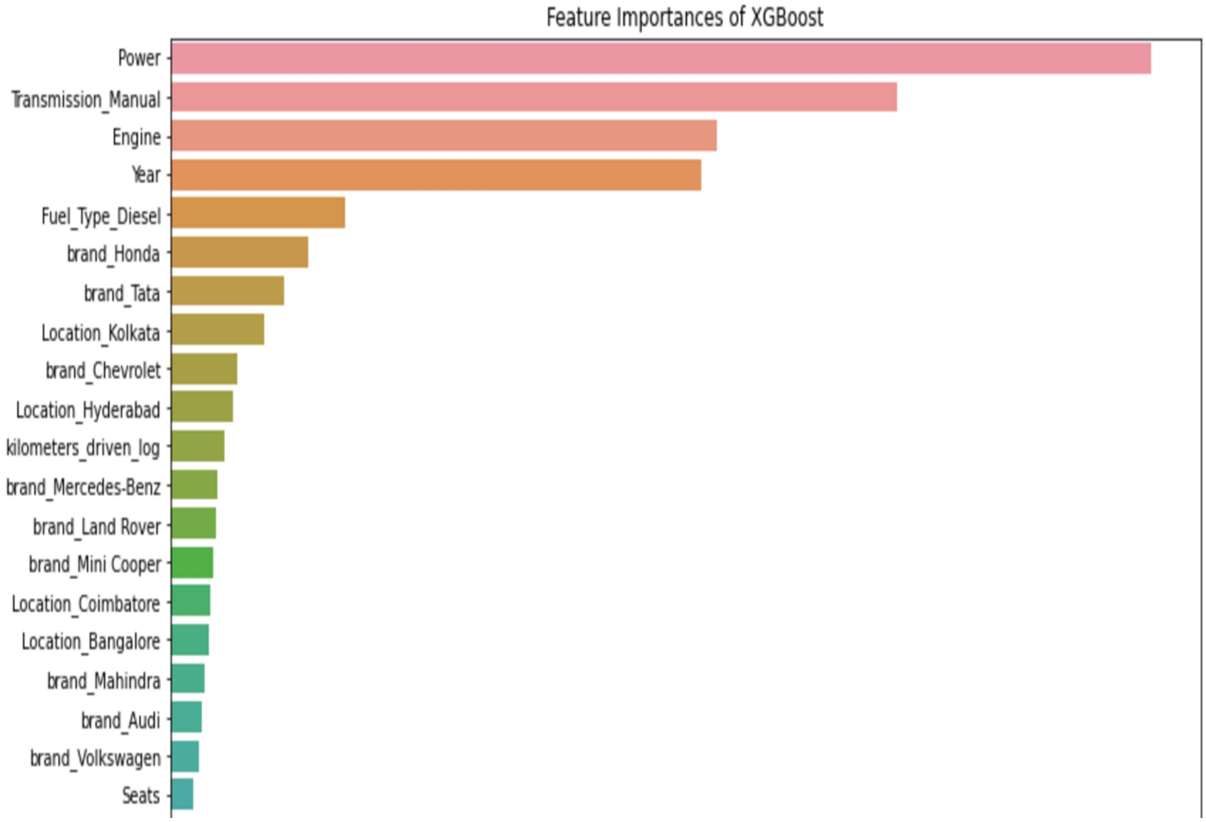
Examining Feature Importance is also significant point to evaluate which models is the best for the prediction. Most of models show the Feature Importance is 1st: Power, 2nd: Year, and 3rd: Engine. Please see Figure2.

Figure2



However, the model of XGBoost shows the Feature Importance as 1st: Power, 2nd: Transmission_Manual, 3rd: Engine, and 4th: Year. Please see Figure3, and also please see Figure5 in Appendix. This is because XGBoost is very sensitive to outliers so the order of Feature Importance is changed. This is because outliers made impact the coefficient then some of features became more significant and others became less significant. The result of R-square and RMSE of XGBoost was excellent as Gradient Boost which can be the best model as well as Gradient Boost. Unfortunately, however, the order of Feature Importance of XGBoost was different from the Feature of Importance of other models' figure such as decision tree and random forest because of sensitive to outliers; it was not selected as the best model.

Figure3



Key Points

Gradient Boost is the best model which can predict price with high accuracy and have superior explanatory ability. It has an R-squared of 0.89 on The test data, which means that our model can explain 89% variation in our data. Also the RMSE on test data is 3.60 which mean we can predict very closely to the original values. This is a very good model and we can use this model in production.

We discussed the importance of the performance and accuracy in terms of adaptiveness of the dataset and the level of effectivity in Solution design.

Since the dataset of used car is changing every moment because of the market growing too fast, the pricing model of Gradient Boost is able to perform faster than other models such as random forest and tuned random forest.

The dataset has a lot of outliers because used cars have a lot of varieties even if checking just one factor such as brand, model, price, year, and engine. In General idea, outliers can be dropped when it is cleaning so the accuracy and explanation ability will rise after cleaning data with dropping outliers. If we do cleaning data, the model of XGBoost can be used for the price prediction. But the removing outlier will not be covered anymore. In business, the dataset is fluctuating every time and it will be more cost and time to checking and evaluating every time if the outlier should be dropped or not while market expanding too fast. And then, even if one of value is an outlier, someday or sometimes it may become not outlier because of the fast-changing used car market. Also, increasing process will cause increasing mistakes. If we focus on real world business, we should not increase the manual process such as evaluating outliers after running a model. Cumulating data is important instead of dropping in this particular market and the dataset. Instead of removing, using the model with robust to outliers, which is hard to get impact from outliers, will be benefit to the pricing prediction of used car accurately.

- **Why is this a ‘valid’ solution that is likely to solve the problem? The reason for the proposed solution design. How it would affect the problem/business?**

This is because not only Gradient Boost can perform high accuracy and optimal interpretability but also it is robust to outliers. Other models do not have both of them. Random forest and decision tree are robust to outliers but the overall performance is relatively lower than Gradient Boost. XGBoost can perform well as same as Gradient Boost but XGBoost is highly sensitive to outliers. Even if we can take care of outliers in XGBoost, there will be more time, cost, process, and possibility to mistake that are not good in term of efficiency.

Gradient model can provide the business effectiveness of pricing which can adopt the fast-changing market and can be flexible to outliers which are sometimes changing in short-term.

Recommendations for implementation

- **What are some key recommendations to implement the solution?**

Our used car prediction model has to improve the remaining error which will cause inaccuracy of the result. To improve the model, it can be ensemble with other methods. It must be 0% error to use in business.

- **What are the key actionable for stakeholders?**

Bring more “First-Owner” used cars, “Automatic Transmission” used cars, “Engine with High Power” used cars, and data to analyze about used car in India will be more profitable. Please see Figure 8, 9, 10 in Appendix.
Car dealers can convince their clients using the fair price created by the pricing model.
Potential buyers can make purchasing decisions.
Marketing people can think about marketing strategies.

- **What is the expected benefit and/or costs?**

If we are able to predict the price accurately, then we can put the price of used car appropriately, sales will be increasing significantly. And our customer can do appropriate purchasing decision based on their preference, then they can save money. But if we can’t predict the price accurately, then it will cause loss of money of both parties. We put the price of used car too low and then it will be lowering sales. Customers may purchase car with too much cost or too low cost. Please take a look at the Figure7 below. This shows how much the actual price and predicted price are different in our model. The distance between the dot of actual (blue color) and the dot of predicted (red color) is the loss or the gain. The list of the difference between actual price and predicted price is below as Table3. The current accuracy is 3.6 (RSME) which is relatively good score in our models. For the validation, please see Figure11 in Appendix. The difference between Actual price and Predicted price is pretty close but we can improve it by using more data and advanced method such as ensemble method. Or we can use K-means method to clustering the predicted dots and identify the characteristics why the dots are getting error. The characteristics may too old to put the price or it has premier that is too expensive that cannot put price etc..

Figure7: Actual vs Predicted in Gradient Boost

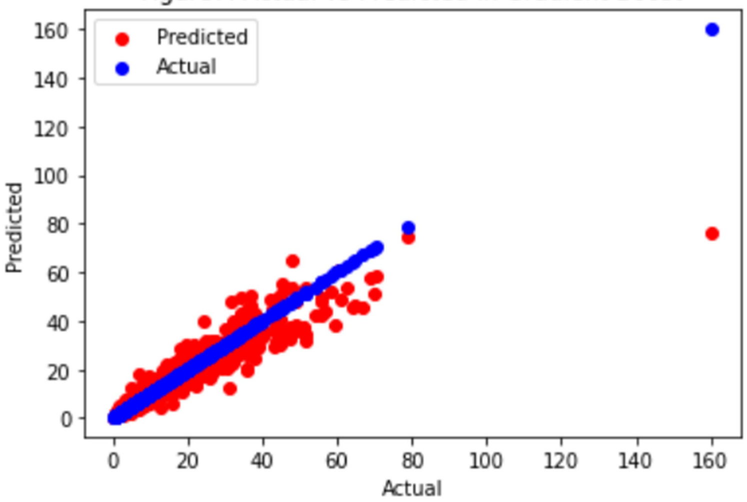
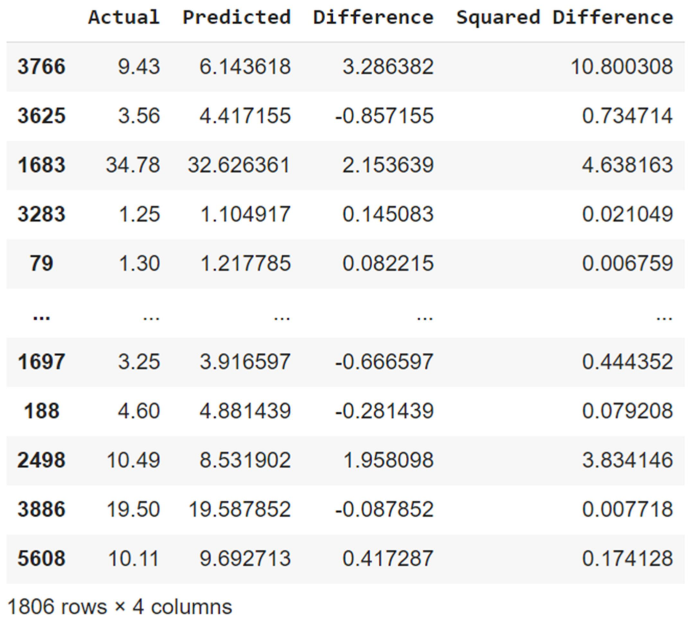


Table3



	Actual	Predicted	Difference	Squared Difference
3766	9.43	6.143618	3.286382	10.800308
3625	3.56	4.417155	-0.857155	0.734714
1683	34.78	32.626361	2.153639	4.638163
3283	1.25	1.104917	0.145083	0.021049
79	1.30	1.217785	0.082215	0.006759
...	...	...	...	...
1697	3.25	3.916597	-0.666597	0.444352
188	4.60	4.881439	-0.281439	0.079208
2498	10.49	8.531902	1.958098	3.834146
3886	19.50	19.587852	-0.087852	0.007718
5608	10.11	9.692713	0.417287	0.174128

1806 rows × 4 columns

There is another expected benefit and/or cost. To deploy the model with machine learning, the cost will be significant while the benefit may also significant. Generally running and maintain machine learning cost at least $60,000 for the first 5 years. But this is not including the decreasing the performance when using for a long term. To maintain the performance in the future, it needs to approximately $95,000 for the 5 years. If the sales and revenue (and subtract other running cost) are expected lower than that, it may be not a good choice. We need to think about cost and benefit (profit).

- **List the benefits of the solution**

By clarifying fair price by the pricing model:
-Customer can buy used car with fair price.
-We (dealer) can convince our customer with clarified analysis and information.
-Both parties can satisfy by the pricing model.
-The pricing model reduces cost and time to evaluate and process regarding the used car price manually.
-Our marketing people can make new strategies using this solution. From the data, they would know which used cars are profitable and popular.
-Inventory management can be more predictable and be focused on profitable and popular used cars. Right now, the popular cars' brand is Maruti and Hyundai.
-Auto repair store. Since the data of model and brand show which used car is popular and majority, we can prepare tools and equipment to fix them when our customer have car problems. This is another way to making money. Repeatedly coming customers for fixing car are larger amount than the customer who buys a used car. The mechanic person who has skill to fix car can be also a resource to make money as well.

- **What are the key risks and challenges?**

-Accuracy of data is one of key risks. Right now is correct. But in the future, it might bias, out of date, or a lot of missing data.
-The pricing may not right in the future some point. The performance and the process may not accurate and appropriate. The result will be overfitting, under fitting.
-The market may change. Consumers change their preference, and other factor may impact the accuracy of the pricing model. To adapt the changes, the pricing model needs to be updated regularly.
-Recently, using AI is not very special. A lot of competitors can use to make their business profitable. If competitors use similar kind of pricing model with machine learning, we need to improve ours and keep it up to date with the latest developments.
-It may lack of ethical mindset or security.

> -To chasing profitable, some business might become unethical. The pricing model might be misused by them and the model might be changed without ethics.

-We need to always keep up to date our cyber security. There is a lot of risk recently. If hacker changes some of process of the pricing model, that will be a significant issue.

- **What are the potential risks or challenges of the proposed solution design**

The model tends to be overfitting. Using hyperparameter, it can be fixed such as low learning rate and applying penalties. It may costly and take time if the data is large.

- **What further analysis needs to be done or what other associated problems need to be solved?**

Further precise valuation of used car, it may be able to take a picture or video to see the condition (scratches, accidents, or repairs) to evaluate the picture of used car. That can be applied to deep learning. The conditions of used car impact the car's price significantly. It's great if we have this.

Overall, while the used car prediction model will be profitable and efficient, it is important to be mindful of the risks and challenges associated with it, and to continuously strive to improve the model's accuracy and adapt to changes in the market.

Reference:

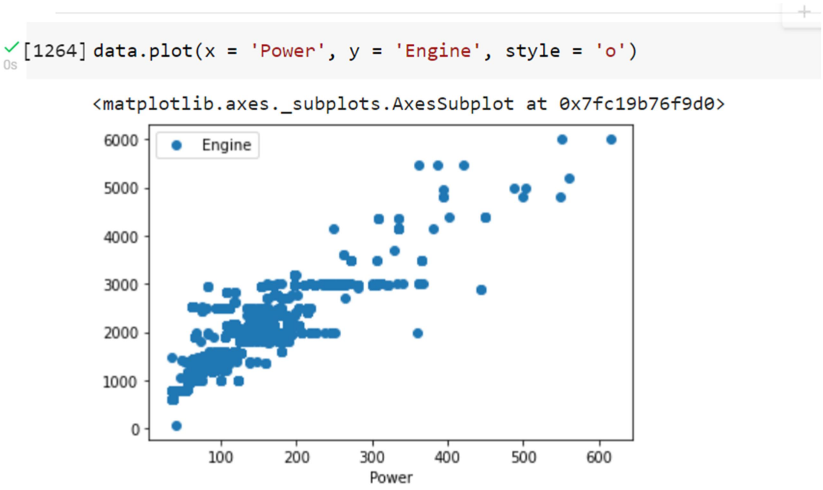
Analyticsindiamag. Retrieved from https://analyticsindiamag.com/top-xgboost-interview-questions-for-data-scientists/

Boardinfinity. Retrieved from https://discuss.boardinfinity.com/t/gradient-boosting-advantages-and-disadvantages/12577/

Phdata. Retrieved from https://www.phdata.io/blog/what-is-the-cost-to-deploy-and-maintain-a-machine-learning-model/

Appendix

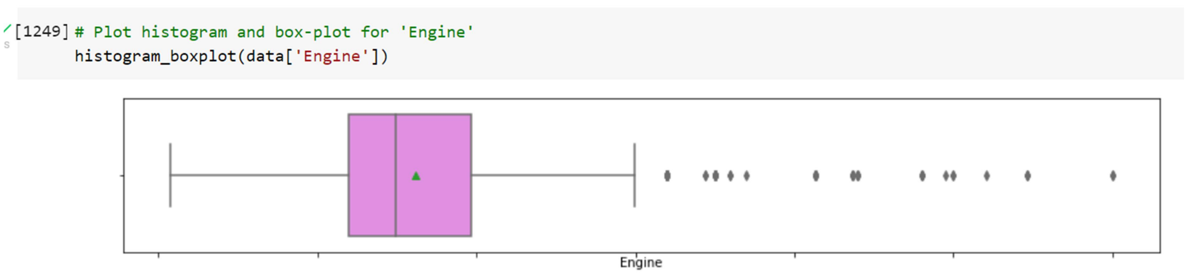
Figure2



```
✓ [1264] data.plot(x = 'Power', y = 'Engine', style = 'o')
0s

<matplotlib.axes._subplots.AxesSubplot at 0x7fc19b76f9d0>
```

Figure5



```
[1249] # Plot histogram and box-plot for 'Engine'
       histogram_boxplot(data['Engine'])
```

Figure 8

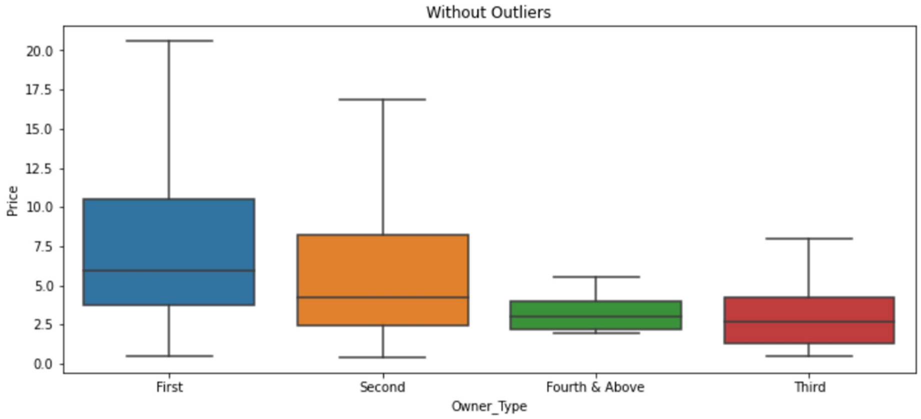


Figure 9

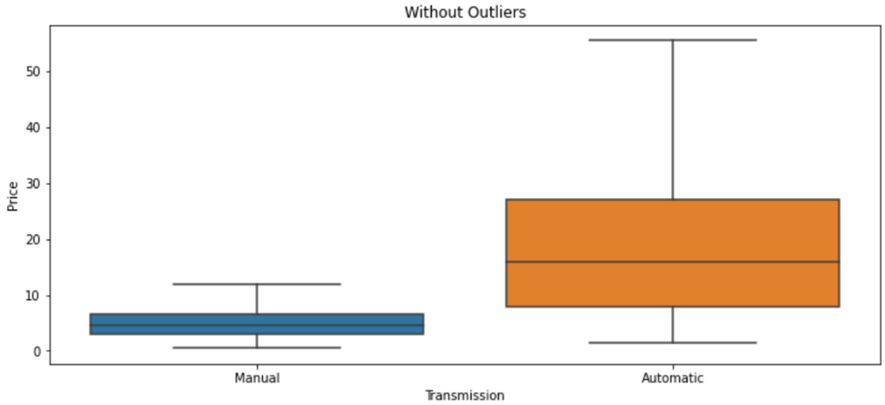
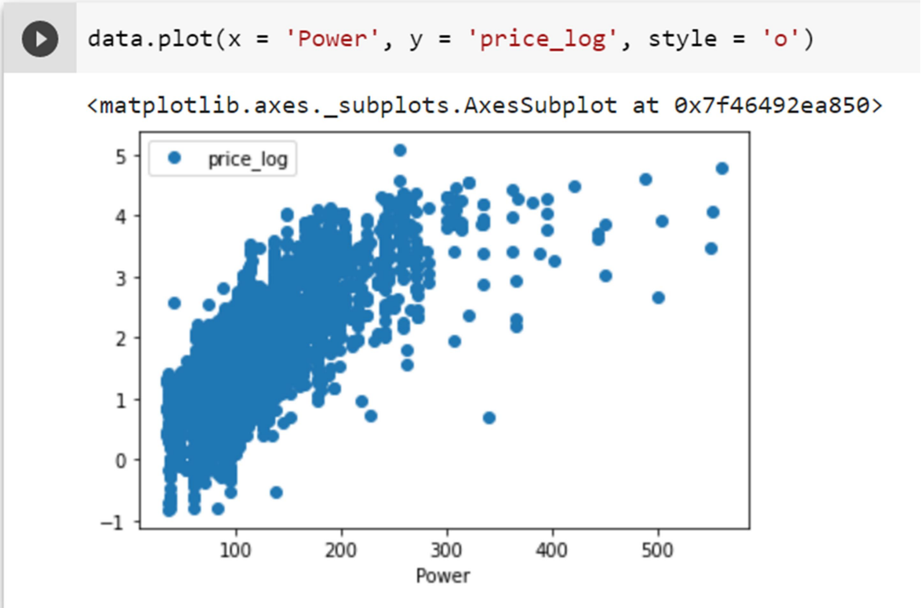


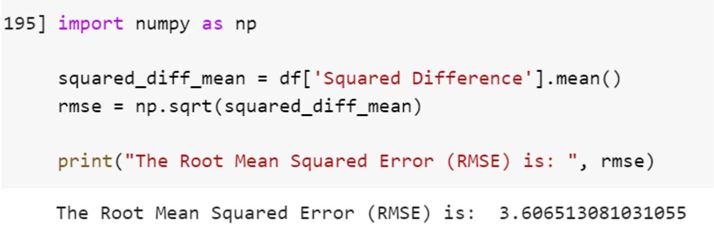
Figure 10



```
data.plot(x = 'Power', y = 'price_log', style = 'o')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f46492ea850>
```

Figure11



```
195] import numpy as np

     squared_diff_mean = df['Squared Difference'].mean()
     rmse = np.sqrt(squared_diff_mean)

     print("The Root Mean Squared Error (RMSE) is: ", rmse)
```

```
     The Root Mean Squared Error (RMSE) is:  3.606513081031055
```