PriceSense : Mastering Car Pricing Strategies

Group 35
Aavash Neupane, 24887438
Bharti Bulchandani, 24880888
Ram Puri, 24619900
Rika Gurung,24872475
Rusan Vaidya, 24886400

Statistical Thinking for Data Science
TD School
University of Technology Sydney

UTS CRICOS 00099F

Executive Summary

Car Pricing can be a complex problem as there are numerous factors that can influence its value in the market.

PriceSense takes factors like Manufactured Year, kilometers the car has driven and various features of it that would have an impact on price of the car to predict more accurate and efficient evaluation.

PriseSense identified that "Gearbox" significantly influences car prices (F-statistic: 544.70, p-value: 1.30e-118) with accurate predictions (RMSE: 0.2215, R-squared: 90.19%).

This project uses data to help buyers make informed decisions about car costs, promoting transparency and sustainability in the automotive industry.

This project supports auto industry analysts and businesses with cost analysis and marketing insights for informed decision-making.

R2 Score: 90.19%

Gearbox Impact

Table of Content

Introduction

Introduction

Problem Statement

The car market is complex, with various factors affecting price of car. The issue emerges from requirement for stakeholders to have a solid apparatus for evaluating showcase esteem of cars, encouraging better buying and offering choices. To tackle this problem, a machine-learning model is created that can precisely anticipate costs of cars. Reliable pricing data can lead to more superficial, efficient car sales.

Rationale

The method of reasoning for tending to the issue of foreseeing car costs is established in its potential to form a more straightforward, effective, and reasonable car commercial center. This endeavor benefits personal customers and the industry by guaranteeing better-informed choices, progressed showcase elements, and a positive financial effect.

Project aims and objectives

This research aims to predict the price of cars and develop effective strategies for pricing in the automobile industry using valuable insights, enabling stakeholders to make informed pricing decisions.

Following questions were raised to support the project's aim:

1. Do year, kilometers and car features impact price variations?
2. Does the “price” of the vehicle depend on all the factors available equally?
3. Does “price” prediction depend on all factors equally?

Following objectives were established to support these questions:

1. Identify relevant datasets.
2. Cleaning data.
3. Visualize data to understand data and market trends better.
4. Identify different factors that affect the price of cars.

Methodology

Methodology

Methods Overview

For this experiment, the car price dataset from Kaggle consisting of 17025 rows and 16 features is used to target car prices for Australia, with the response variable being “Price”.

Methods details

For this experiment, following steps are executed:

- Data acquisition
- Data manipulation
- Exploratory Data Analysis
- Formal Analyses

Data Exploration & Manipulation:

In this project, python libraries like Pandas, Numpy, Matplotlib, Dataprep and Seaborn were used. Further, the data were explored and cleaned. The response variable being left-skewed, a logarithmic transformation is performed. Further, the categorical columns were converted using one-hot encoding to include for training the model.

The predictor variables such as “ID” and “Name” were removed as including them would result in overfitting while model training. (see Appendix for Data acquisition, exploration and manipulation)

Modeling:

RandomForestRegressor models were chosen for training since the target variable was continuous. (see Appendix for modeling)

Result

Results

Key findings

This research showed that the car feature “Gearbox” has a surprisingly major influence on car price with a remarkable F-statistic of 544.70 and a very low p-value of 1.30e-118. Additionally, the use of RandomForestRegressor was able to make accurate predictions.

In-depth results

Do the “year”, “kilometers” and car features impact the “price” of the car?

The correlation coefficient result indicates that “Price” has a strong link with the variables "Year," which is positively correlated(0.6760) with “Price”, and “Kilometers”, which is negatively correlated(-0.5997) with “Price” and car feature such as “CC” has a small p-value of 5.23e-305 and a positive correlation of 0.2803, suggesting that these factors affect the car's pricing.

Does “price” prediction depend on all factors equally?

The RMSE score for RandomForestRegressor was 0.2292 for the validation set against 0.08590 for the training (baseline) model. (see Appendix for modeling)

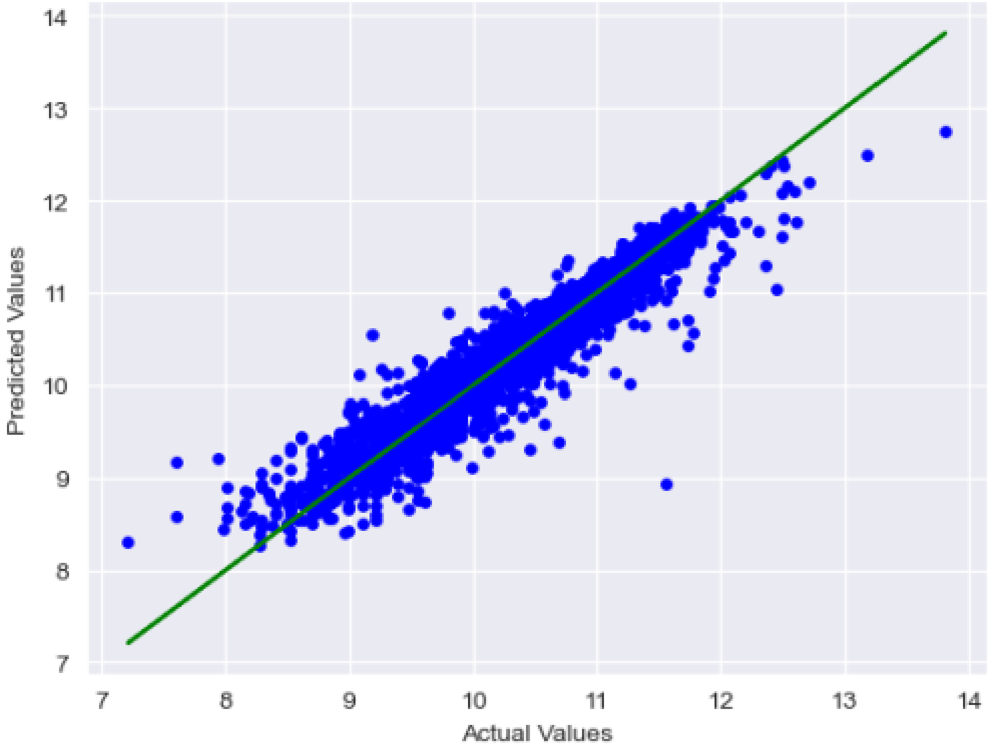


Figure 1: Validation set predictions vs actual values plot

Conclusion

Conclusion

Take home message

This project anticipates car costs to make data-driven decisions and provides buyers with more straightforwardness, empowering them to form better choices while promoting a sustainable car industry.

Discussion

This research aims to predict the price of cars accurately. Factors like “Year”, “Kilometers” that it was driven, and surprisingly, the type of “Gearbox” have a heavy influence on the pricing of the car. These findings contribute to a deeper understanding of pricing strategies that benefit buyers and sellers.

Project limitations and caveats

- The data is used from Australia until 2022.
- Removing non-essential variables that streamline the examination, but it may lead to a less comprehensive understanding of the circumstance.
- The dataset contains fewer records to train the model.

Stakeholder Analysis and Project Outcomes

Data analysts and investigators within the car industry can utilize the venture's experiences to get cost elements better and advertise trends. Car producers and dealerships can use the demonstration to estimate their vehicles competitively and understand advertising requests. The extent may incorporate a proposal framework that proposes reasonable advertising costs to buyers and vendors, supporting them in making good choices within the car market.

REFERENCES

OpenAI. (2023). *ChatGPT* (September 25 Version) [Large language model]. https://chat.openai.com

sklearn.linear_model.Ridge. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html#sklearn.linear_model.Ridge

sklearn.linear_model.LinearRegression. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression

sklearn.ensemble.RandomForestRegressor. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

APPENDICES

APPENDIX A

DATA

Data Acquisition, Exploration and Manipulation:

For this experiment, the dataset used is from Kaggle. The dataset contains 17025 rows and 16 features that will be used for training and testing the model performance. The dataset targets only the Australian market for this project. It contains predictor variables like ID, name of the car, model, vehicle variant, vehicle series, year of manufacture, kilometers driven, type of car, car gearbox, if its used or new, fuel type needed, engine capacity, its color and the seating capacity.

The response variable “Price” is left skewed and hence logarithm transformation was performed to help with the skewing.
Similarly, the predictor variables “Kilometers” and “CC” were left skewed and log transformed for model training.

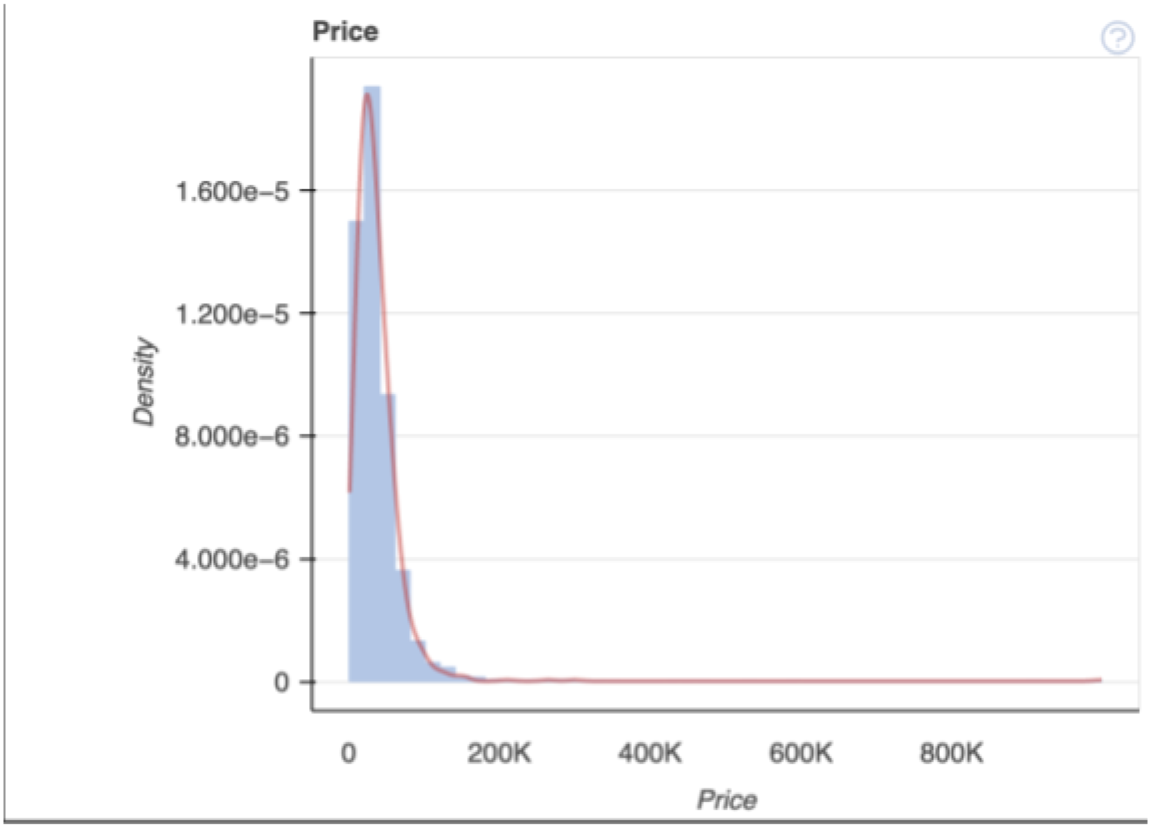


Figure 2: Response variable "Price"

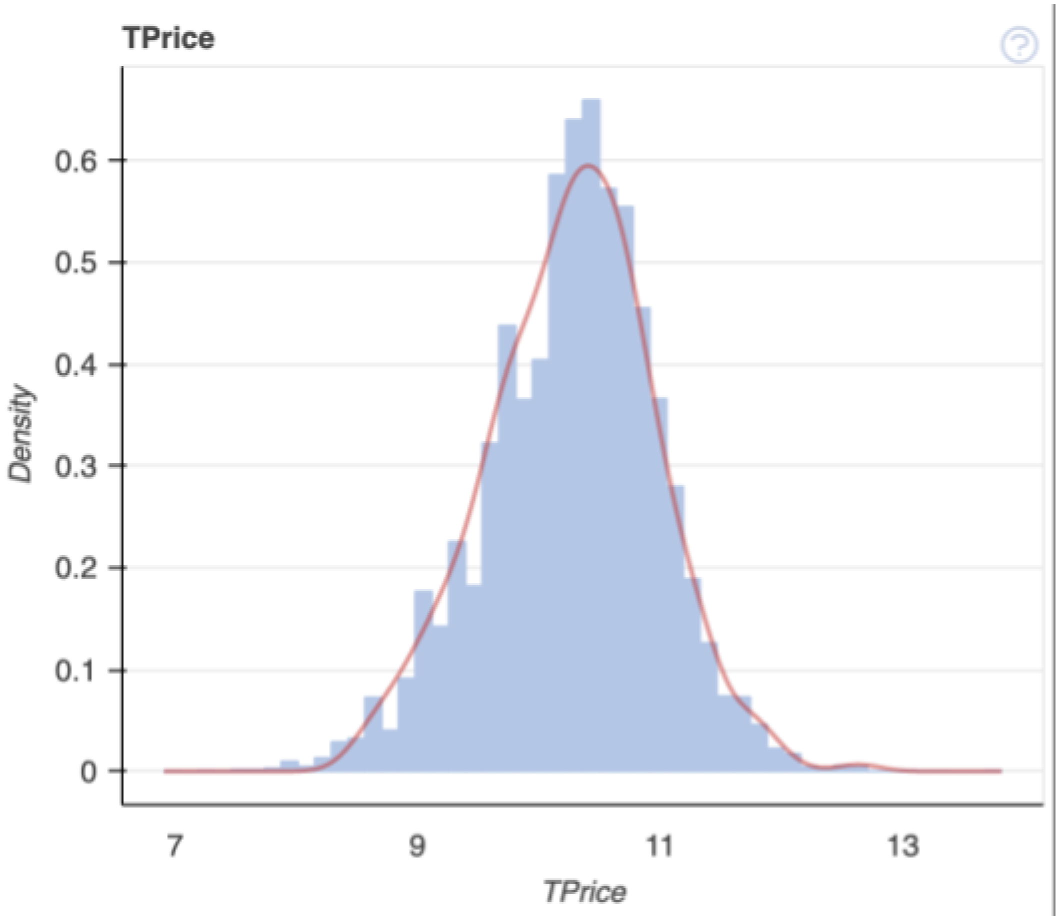
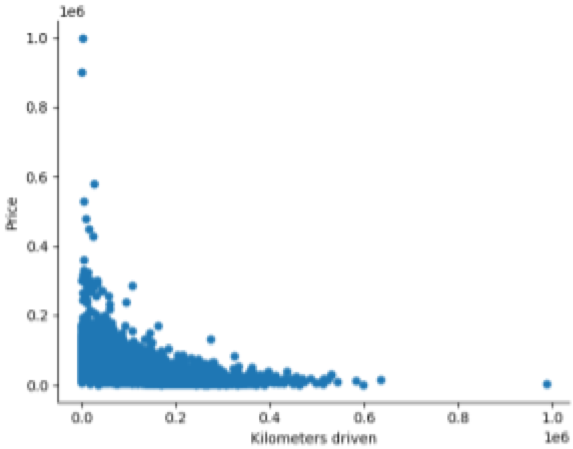
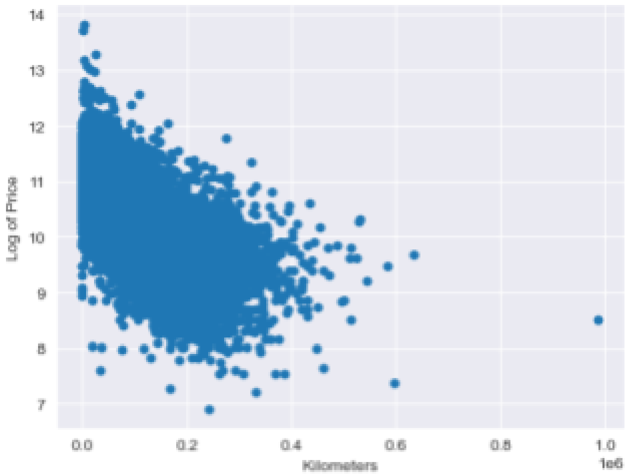


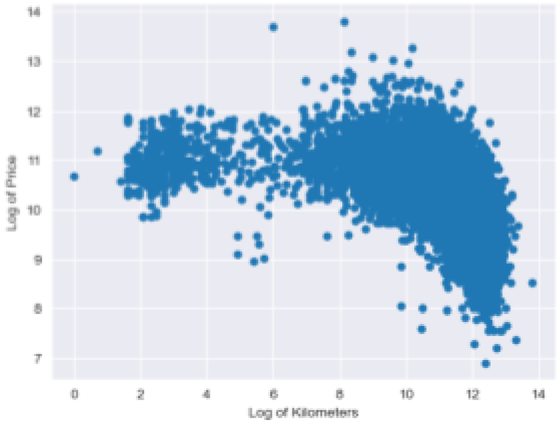
Figure 3: Response variable after transformation

After transforming the variables, the variable shows improvement in linear relation with the predictors compared to the original variable.





(clockwise):
Figure 4: Kilometer vs Price (before transformation)
Figure 5: Kilometer vs transformed "Price" variable
Figure 6: Transformed Kilometer vs Transformed Price



Further, to include the categorical predictor variables for training the model, we performed one-hot encoding on the predictors - "Brand", "Gearbox", "Fuel" and "Status" to merge it with the other numeric predictors. This expanded the number of predictor variables to 75.

APPENDIX B

MODELING

After the data exploration and transformation, the next step performed was choosing the model to train. However, the predictor variables ID, Name were not considered for model training as using them would lead the model to overfit as it would lead to model learning very specific trends instead. Since the response variable was numeric, we chose a regression model for training. For this purpose, the dataset was divided into training, validation and testing sets. The dataset was split with 20% (3405 rows, 75 columns) for testing set, 30% (4086 rows, 75 columns) for validation set and training set containing 9533 rows. Since the dataset consisted of 17025 rows, splitting the validation set with 20% data would leave with very small number of rows to test the trained model. Hence the validation set was split to have 30% data instead.

For measuring the model performance, Root mean squared error and r-square functions were used.

Model 1: Linear regression

- For the model training, we chose Linear Regression as the first model from the linear_model of scikit-learn library. It is used for performing regression when multiple features are provided for predictions.
- The model was trained using the training data and make predictions for baseline model. The RMSE score for the baseline model was 0.2904. However, the model could not perform well on the validation set and had the RMSE score of 1200716205.4814 while the r-squared value for linear regression model was negative which showed that the model was highly underfitting.

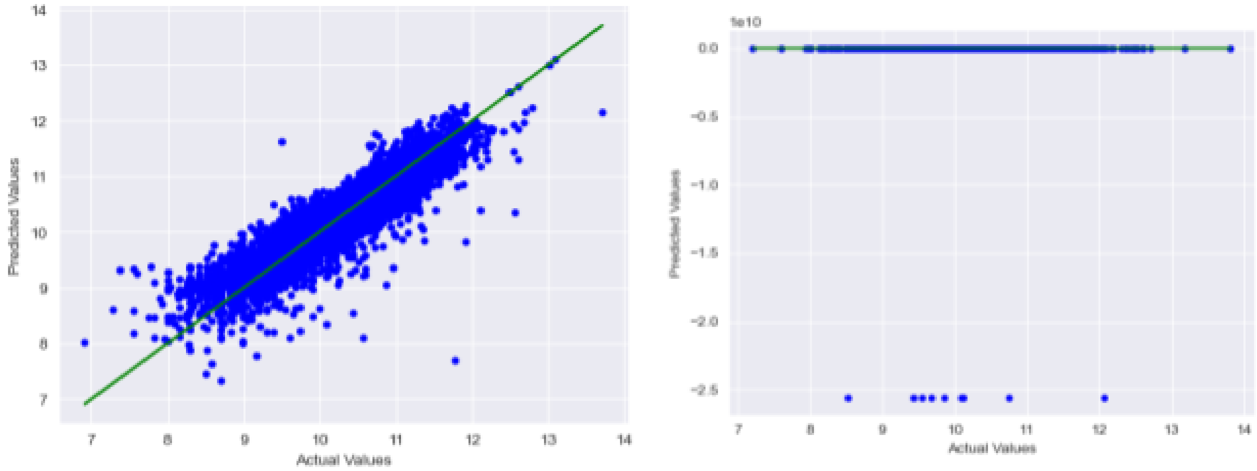


Figure 7: Linear Regression model - training set vs validation set performance

Model 2: Ridge Regression

- The second model chosen was Ridge Regression from the linear_model module of scikit-learn library. It is usually considered to solve the regression problem and the estimator has a built-in support for multivariate regression.
- The model was trained using the training data and made predictions for the baseline model. The RMSE score for the baseline model was 0.29079 and the RMSE for the validation set was 0.29561 with the r-squared value for validation set to be 0.825.

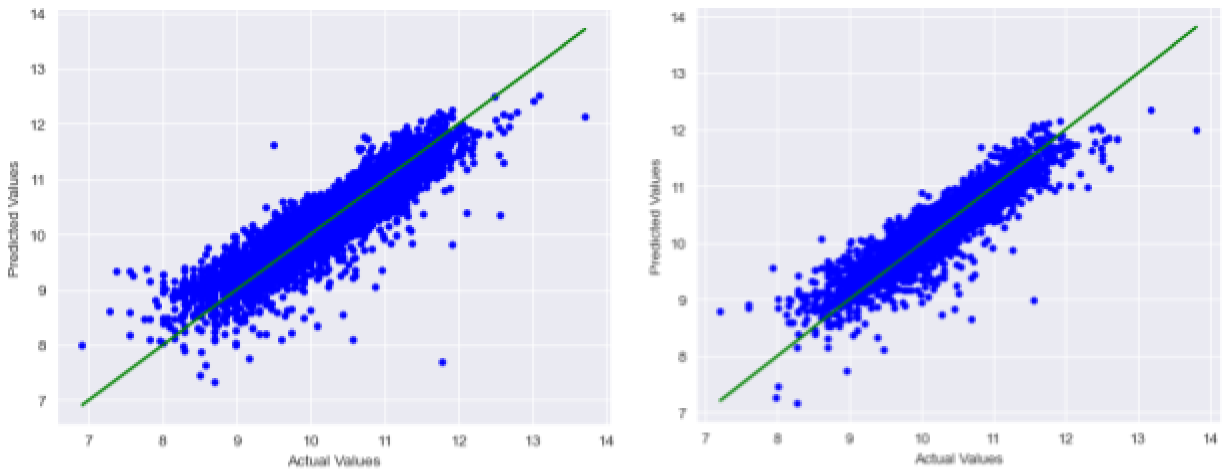


Figure 8: Ridge regression training vs validation set performance

Model 3: RandomForestRegressor

- The final model chosen was Random Forest Regressor from the ensemble module of scikit-learn library. It uses sub-samples for multiple decision tree and averages the results.
- The model was trained using the training data and made predictions for the baseline model. The RMSE score for the baseline model was0.086336. The validation set had the RMSE score of 0.22292 which showed that the model was underfitting to some level for the outliers and the r-squared value was 0.9014.

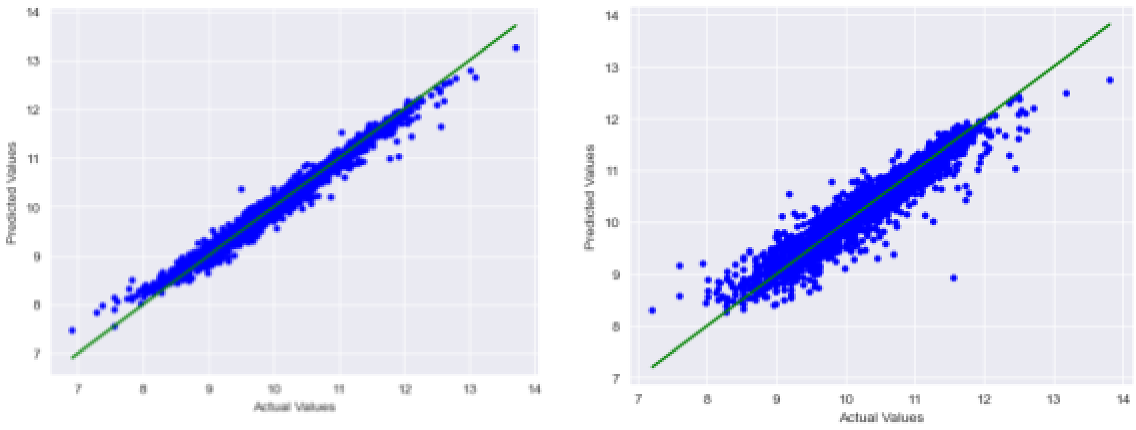


Figure 10: RandomForestRegressor - training data vs validation set performance

- Hence, looking at the r-squared value we analyzed that the RandomForestRegressor performed better for predicting the response variable.
- The RandomForestRegressor was used further on the test data to analyze the performance. The RMSE score for the test set was 0.22149 and the r-squared value was 0.9018 which implied the model was able to make accurate predictions.