Industrial Training Report On

Predicting selling price of used cars by using various Supervised learning algorithms an comparing them.

Submitted in partial fulfilment of requirements for the award of the Degree of

Bachelor of Technology
In
ELECTRONICS & COMMUNICATION ENGINEERING

Submitted By

Shubham kumar
04751202819

Under the guidance of

Dr. Rachna Jain
Assistant Professor, CSE

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING
BHARATI VIDYAPEETH'S COLLEGE OF ENGINEERING
(AFFILIATED TO GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY, DELHI)
NEW DELHI – 110063
JAN 2022

BHARATI VIDYAPEETH PUNE

BHARATI VIDYAPEETH'S COLLEGE OF ENGINEERING

A-4, Paschim Vihar, Main Rohtak Road, New Delhi-110 063

(AFFILIATED TO GGSIPU DELHI, APPROVED BY AICTE, NEW DELHI)

(AN ISO 9001:2008 CERTIFIED INSTITUTION)

CERTIFICATE

This Certificate Accredits that Mr./Ms. Shubham Kumar

Completed the Four weeks In-House Training Programme on Machine Learning & Deep Learning

from 26 July *to* 22 August 2021 *He/She has obtained* A *Grade*

Rachna Jain,
Course Coordinator

Principal

Legend:	A Grade (100-90%)	B+ Grade (89-80%)	B Grade (79-70%)	C Grade (69-60%)	D Grade (59-50%)

CANDIDATE'S DECLARATION

It is hereby certified that the work which is being presented in the B. Tech Industrial/In-house training Report entitled **" Predicting selling price of used cars by using various Supervised learning algorithms and comparing them "** in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** and submitted in the **Department of Electronics & Communication Engineering** of **BHARATI VIDYAPEETH'S COLLEGE OF ENGINEERING, New Delhi (Affiliated to Guru Gobind Singh Indraprastha University, Delhi)** is an authentic record of our own work carried out during a period from **July 2021 to December 2021** under the guidance of **Dr Rachna Jain, Assistant Professor, CSE.**

The matter presented in the B. Tech Major Project Report has not been submitted by me for the award of any other degree of this or any other Institute.

Shubham kumar
04751202819

This is to certify that the above statement made by the candidate is correct to the best of my knowledge. He/She/They are permitted to appear in the External Industrial/In-house training Examination.

Dr. Rachna Jain	**Prof. Kirti Gupta**
Assistant professor, CSE	**Head, ECE**

ABSTRACT

The prediction of car price has always been the area of high research interest, as it requires a large amount of effort to be made, and the knowledge of field experts. There is a large number of different attributes to be evaluated so that the prediction can be both trustworthy and accurate. In order to build a model so that it can predict the price of used cars in India, we have applied three machine learning techniques (i.e., Random Forest, Linear Regression, Lasso Regression, Ridge regression). Despite that, we applied these three approaches to work as an ensemble. The data which we gathered from the cardekho.com website is used to predict the price of used cars.

After having respective performances on all these algorithms on the given dataset were compared, we found out which one suits the best on the available dataset. Moreover, the model was evaluated over the test dataset and very good accuracy was obtained.

ACKNOWLEDGEMENT

I express my deep gratitude to **Dr Rachna Jain,** for her valuable guidance and suggestion throughout my project work**.** We are thankful to **Dr. Preeti Nagrath, Dr Rachna Jain, Dr Ashish Gupta, Mrs Nitika Sharma**, for their valuable guidance.

We would like to extend my sincere thanks to **Head of the Department, Prof. Kirti Gupta** for his time-to-time suggestions to complete my project work. I am also thankful to **Prof. Dharmender Saini, Principal** for providing me the facilities to carry out my project work.

Shubham kumar

04751202819

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Overview of our Work

Car price prediction is one of the popular and interesting topics. As per the information that we got from the websites of cardekho that more than 50000 cars are currently registered under cardekho. Mainly the reselling of cars can be done in two ways either online or offline. In offline mode there is a mediator present between the buyer and seller who is making a huge number of profits over the transaction. The second alternative is through online mode where the customers can get the best price for their price. We have developed a machine learning system that can predict the price of old cars. On the basis of this, some of the objectives have been formulated. That’s why this is going to be a real time project.

1.2 Dataset Used

1.2.1 Cardekho.com dataset:

This dataset consists of details of 301 used cars. A total of 9 features like car name, kms driven, fuel type, transmission type is also given. Each column represents a very important feature which is must know information for the buyer if he wants to purchase a used car. We used 30% of data for testing purpose and rest of the data for training purpose.

1.3 Motivation

The Automobile business has been dominated by a few multinational corporations and a number of retailers across the world. Mainly the multinational corporation are manufacturers by trade but the retailer deals with both the new and secondhand cars. The used automobile market has increased significantly, the value increased contribute by a bigger percentage of the total market. In India the secondhand vehicle in accounts for around 3.5 million of vehicle annually.

1.4 Objective

- To prepare a supervised learning model, so that it can estimate the selling price of cars based on various aspects.

- The prepared model must be capable of making predictions based on various features.

- Provide the comparison on the behalf of the graph for better visualization.

Chapter 2

Literature Review

Machine Learning, in the simplest of terms, is teaching your machine about something. You collect data, clean the data, create algorithms, teach the algorithm essential patterns from the data and then expect the algorithm to give you a helpful answer. If the algorithm lives up to your expectations, you have successfully taught your algorithm. If not, just scrap everything and start from scratch. That is how it works here. Oh, and if you are looking for a formal definition, Machine Learning is the process of creating models that can perform a certain task without the need for a human explicitly programming it to do something**.**

Supervised Learning – You supervise the learning process, meaning the data that you have collected here is labelled and so you know what input needs to be mapped to what output. This helps you correct your algorithm if it makes a mistake in giving you the answer. At its most basic sense, machine learning uses programmed algorithms that learn and optimize their operations by analyzing input data to make predictions within an acceptable range. With the feeding of new data, these algorithms tend to make more accurate predictions. Although there are some variations of how to group machine learning algorithms, they can be divided into three broad categories according to their purposes and the way the underlying machine is being taught.

Applications of Supervised Learning:

- **Bioinformatics**
- **Speech Recognition**
- **Spam Detection**
- **Object-Recognition for Vision**

We conducted an analysis on the core concepts of this project and utilized those notions to collect information such as the technological stack, algorithm, and flaws of this project, which lead us to develop a better working model.

CarDekho:

CarDekho is a website where sellers may sell old vehicles or cars.
This startup has a simple UI that asks sellers for information such as model of car, distance traveled, year of registration, and type of vehicle (petrol, diesel, CNG). This enables online model to execute certain algorithms on Applied parameters in order to estimate the selling price.

GET VEHICLE PRICE:

This is a website that predicts the same as that of the CarDekho Android app, which works on some algorithm that predicts on the basis of a kilometer traveled, fuel type, age of the car. The site uses a ML approach to estimate the price of the vehicle. It will help you to compare the price with different sellers.

2.1 TECHNOLOGY USED

Here, we have used python for machine learning as a major technology for implementing the concepts because of its inbuilt function and libraries package available in python. Some of the major libraries used here are: -

NUMPY

This is a Python-based array-processing package. It comes with a high-performance multidimensional array object and utilities for manipulating them. It's a Python module that's required for scientific computing. NumPy may be used as a multi-dimensional container of general data in addition to its normal applications.
NumPy can define any data type, allowing it to interact with a wide range of databases fast and easily.

SCIKIT-LEARN

It provides a variety of supervised and unsupervised learning methods through a uniform Python interface. It is provided under various Linux distributions and is licensed under a liberal simplified BSD license, promoting academic, commercial use. The library is being constructed.

MATPLOTLIB

Matplotlib is a library used in a python programming language, which is being used for plotting the graph with having a mathematical numerical extension NumPy, which serves as visualization utility.

SEABORN

Seaborn is a Python data visualization framework built on top of matplotlib and tightly integrated with pandas data structures. Seaborn's visualization is a critical component that aids in data exploration and comprehension.

PANDAS

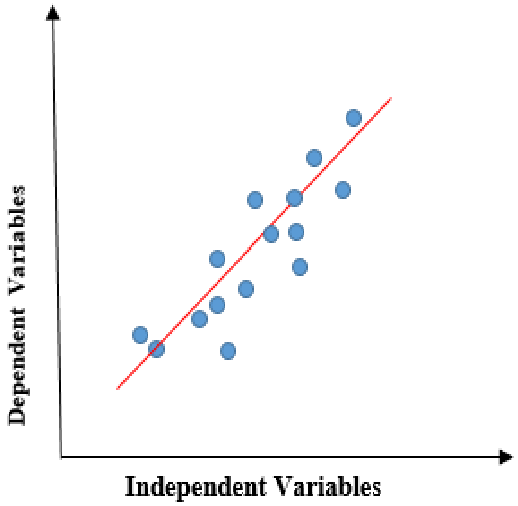
Pandas are generally based on two fundamental Python libraries: matplotlib for data visualization and NumPy for mathematical calculations. Pandas functions as a wrapper around these libraries, enabling you to use fewer lines of code to access numerous matplotlib and NumPy methods.
Pandas are also used for importing the data from the files.

2.2 REGRESSION TECHNIQUES USED:

In Regression, we plot a graph between the variables which best fit the given data points. The machine learning model can deliver predictions regarding the data. Regression shows a line or curve that passes through all the data points on a target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum**.**
It is used principally for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

2.2.1 LINEAR REGRESSION

Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.



To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$y = mx+b \Longrightarrow y = a_0+a_1x$$

y= Dependent Variable.

x= Independent Variable.

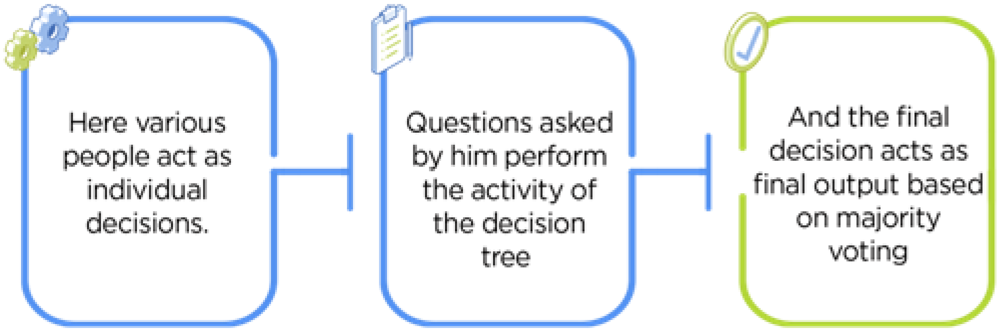
a0= intercept of the line.

a1 = Linear regression coefficient.

2.2.2 RANDOM FOREST REGRESSION

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.


Here various people act as individual decisions.
Questions asked by him perform the activity of the decision tree
And the final decision acts as final output based on majority voting

L1 and L2 regularization:

If a regression model uses the L1 Regularization technique, then it is called Lasso Regression. If it used the L2 regularization technique, it’s called Ridge Regression.

L1 regularization adds a penalty that is equal to the absolute value of the magnitude of the coefficient. This regularization type can result in sparse models with few coefficients. Some coefficients might become zero and get eliminated from the model. Larger penalties result in coefficient values that are closer to zero (ideal for producing simpler models).

On the other hand, L2 regularization does not result in any elimination of sparse models or coefficients. Thus, Lasso Regression is easier to interpret as compared to the Ridge.

2.2.3 LASSO REGRESSION

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e., models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

Lasso Regression uses L1 regularization technique. It is used when we have a greater number of features because it automatically performs feature selection.

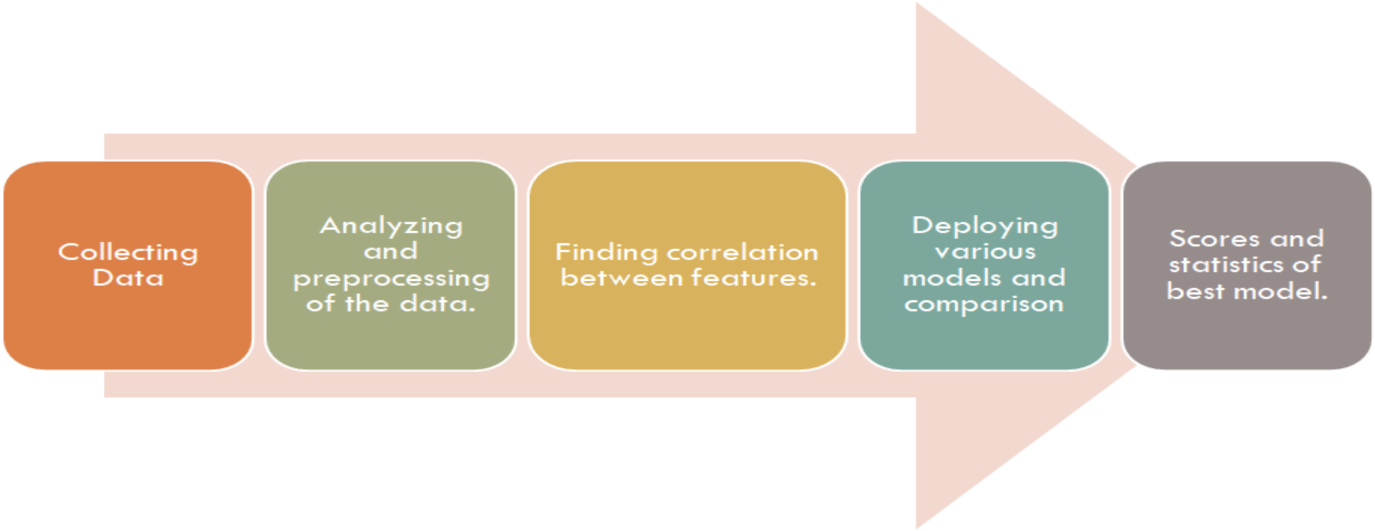
2.2.3 RIDGE REGRESSION

Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

In ridge regression, the first step is to standardize the variables (both dependent and independent) by subtracting their means and dividing by their standard deviations. This causes a challenge in notation since we must somehow indicate whether the variables in a particular formula are standardized or not. As far as standardization is concerned, all ridge regression calculations are based on standardized variables. When the final regression coefficients are displayed, they are adjusted back into their original scale. However, the ridge trace is on a standardized scale.

Chapter 3

Methodology


Collecting Data
Analyzing and preprocessing of the data.
Finding correlation between features.
Deploying various models and comparison
Scores and statistics of best model.

3.1 Preprocessing Data

The given was checked if it contains and null values. Age of car or number of years that the car has been used is a very important feature so we made a column ‘no_year’ and calculated the age of car by subtracting, it from the column ‘current_year’. Column for car name was dropped. We used to get dummies function to convert non-numerical data into numerical data.

3.2 Visualizing Data

Using matplotlib and seaborn library we plotted various features on graph like countplot, barplot to understand the dataset properly. We used corr() function to find out correlation between various feature in dataset and then plotted heatmap for the same.

3.3 Feature importance

We imported ExtraTreesRegressor from sklearn library to plot the most important features in our dataset and the we plotted the bar graph to find out which are the most important features in our dataset.

3.4 Implement various Model

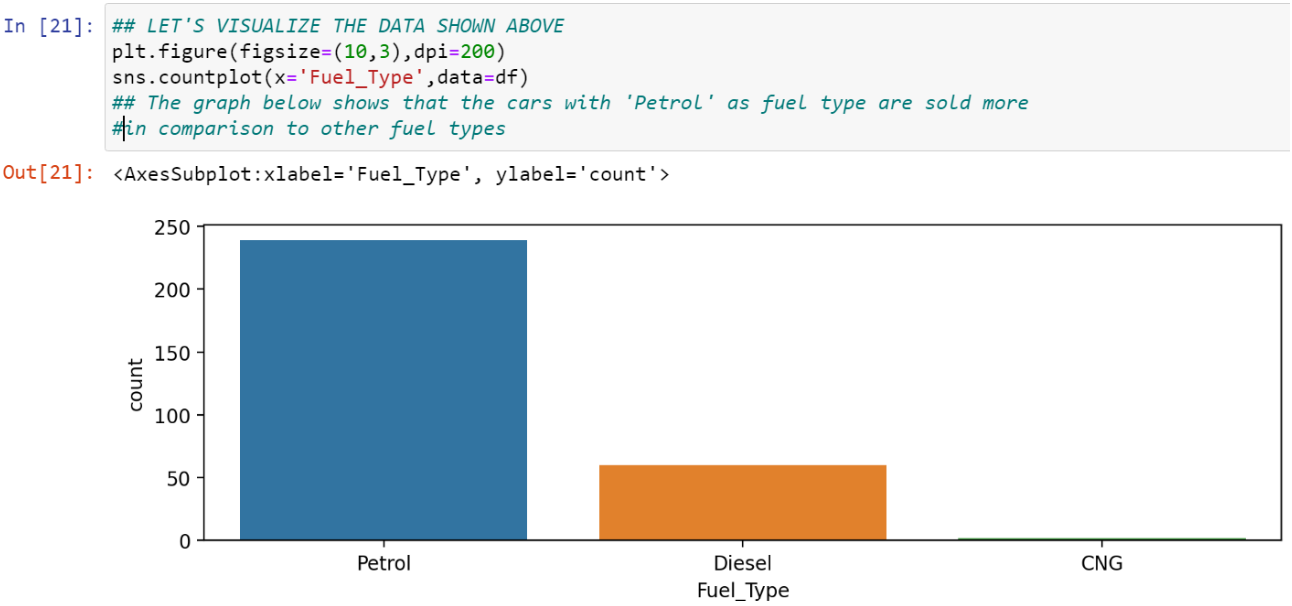
We first imported train test split from sklearn library and the splitted our dataset into training set and testing set. Minmaxscaler was imported from sklearn to normalize the data.

For implementing Random Forest model, we first used RandomizedSearchCV to find out the best hyperparameters for the random forest algorithm and then we implemented the algorithm. Similarly, all the algorithms were implemented and finally we found out the error and accuracy for each algorithm and plotted scatter plot to find out which algorithm is most suitable for implementing and calculating the selling price of used cars.

Chapter 4

Results and Model Performance

1. Data visualization plot

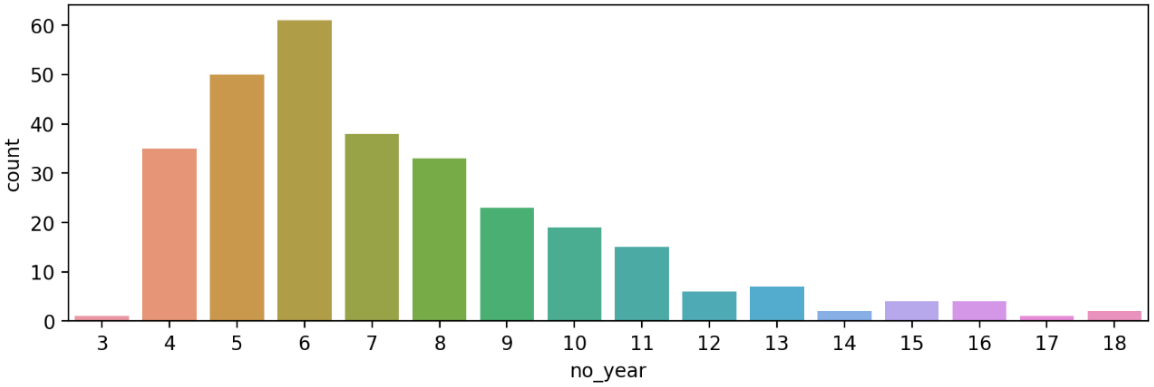


```
In [21]: ## LET'S VISUALIZE THE DATA SHOWN ABOVE
         plt.figure(figsize=(10,3),dpi=200)
         sns.countplot(x='Fuel_Type',data=df)
         ## The graph below shows that the cars with 'Petrol' as fuel type are sold more
         #in comparison to other fuel types

Out[21]: <AxesSubplot:xlabel='Fuel_Type', ylabel='count'>
```

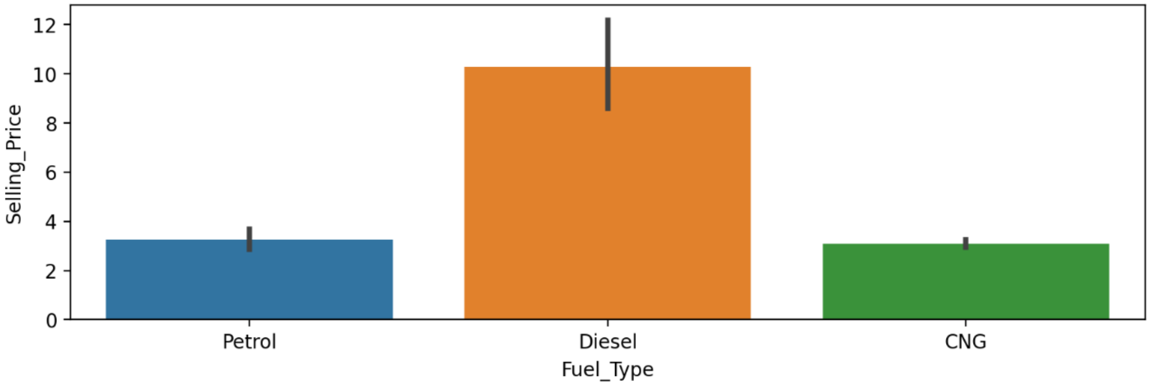
```
In [23]: plt.figure(figsize=(10,3),dpi=200)
         sns.countplot(x='no_year', data=final_dataset)
         #Graph below shows the number of cars of specific age sold on this website
```

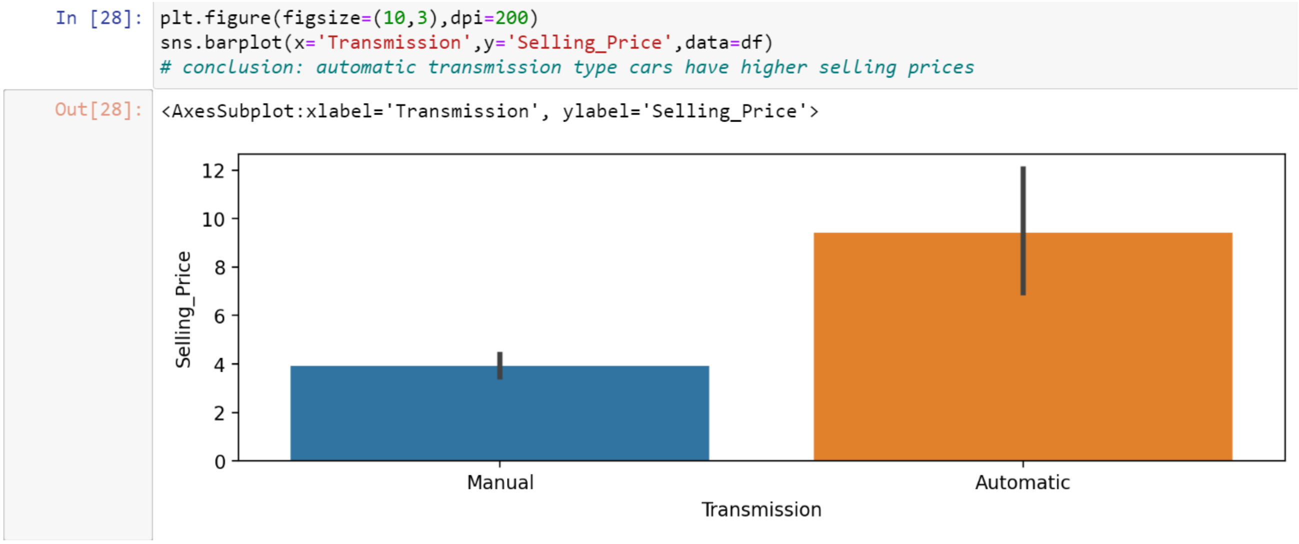
```
Out[23]: <AxesSubplot:xlabel='no_year', ylabel='count'>
```



```
In [27]: plt.figure(figsize=(10,3),dpi=200)
         sns.barplot(x='Fuel_Type',y='Selling_Price',data=df)
         # from the graph below we concluded that the cars with diesel as fuel type has higher selling price
```

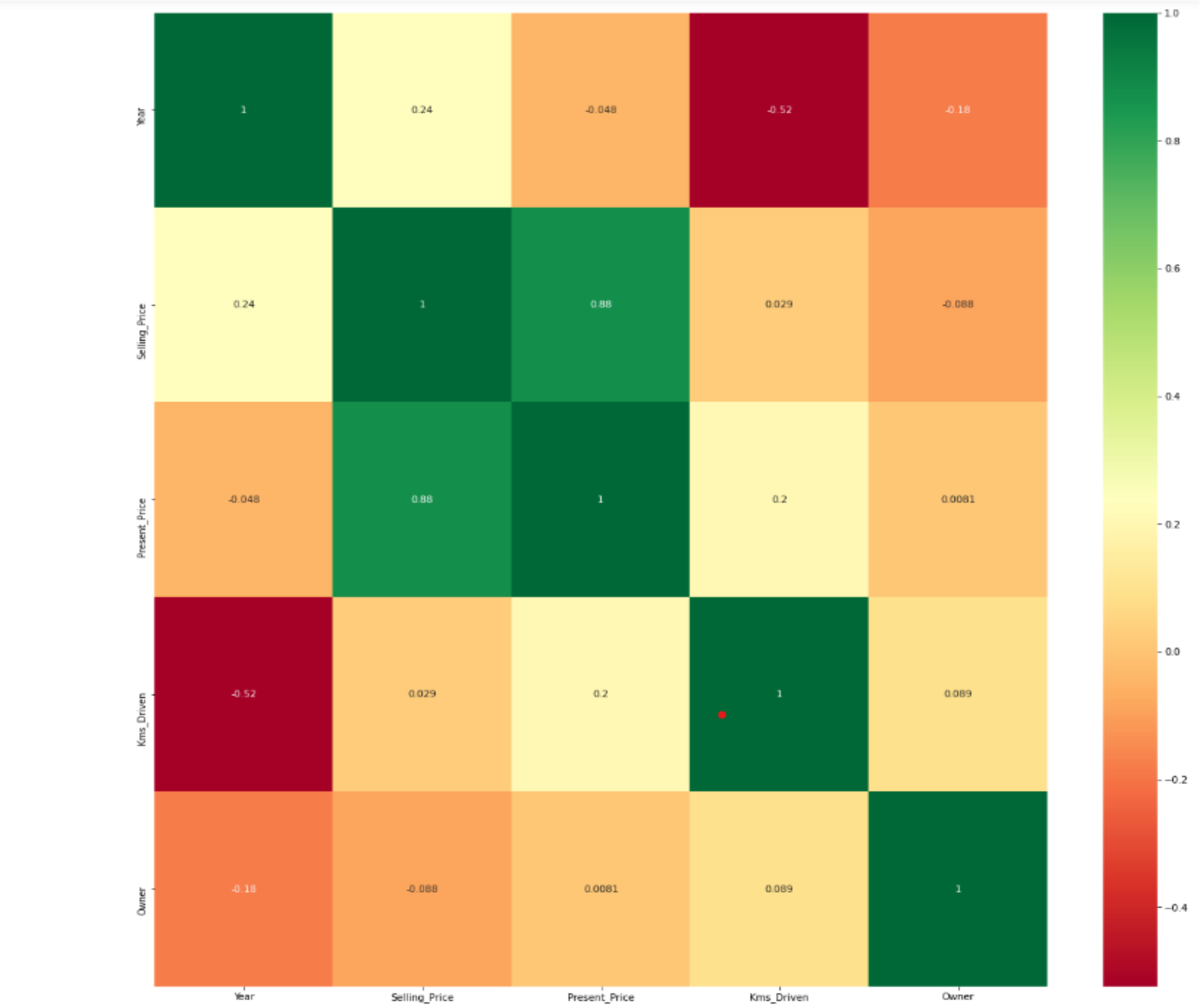
```
Out[27]: <AxesSubplot:xlabel='Fuel_Type', ylabel='Selling_Price'>
```



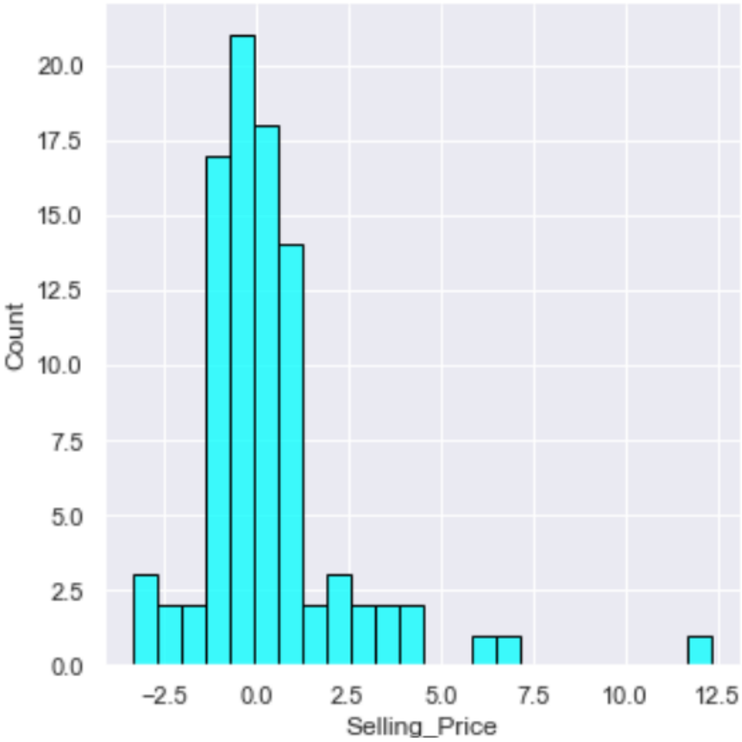


```
In [28]: plt.figure(figsize=(10,3),dpi=200)
         sns.barplot(x='Transmission',y='Selling_Price',data=df)
         # conclusion: automatic transmission type cars have higher selling prices
```

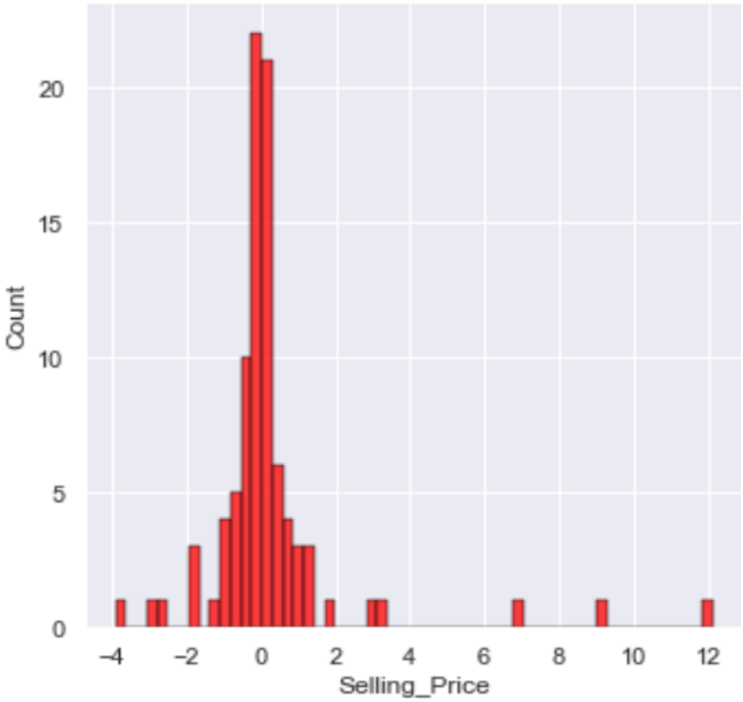
```
Out[28]: <AxesSubplot:xlabel='Transmission', ylabel='Selling_Price'>
```



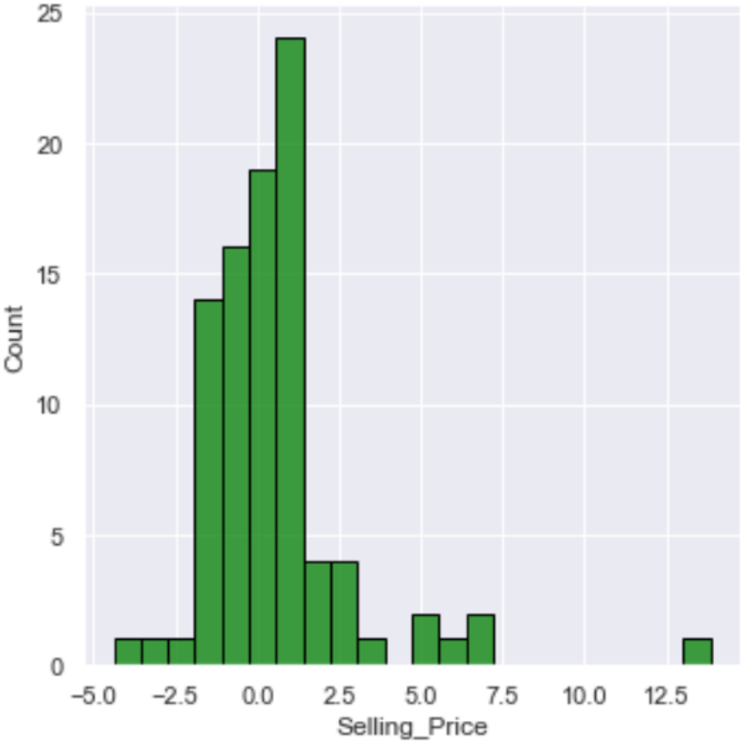
2. Model accuracy:



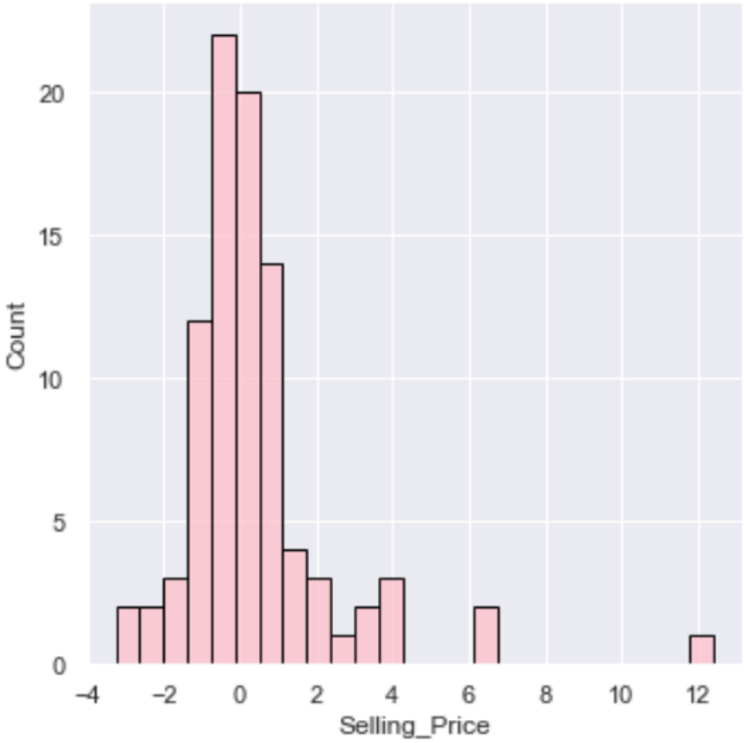
For linear regression



For random forest

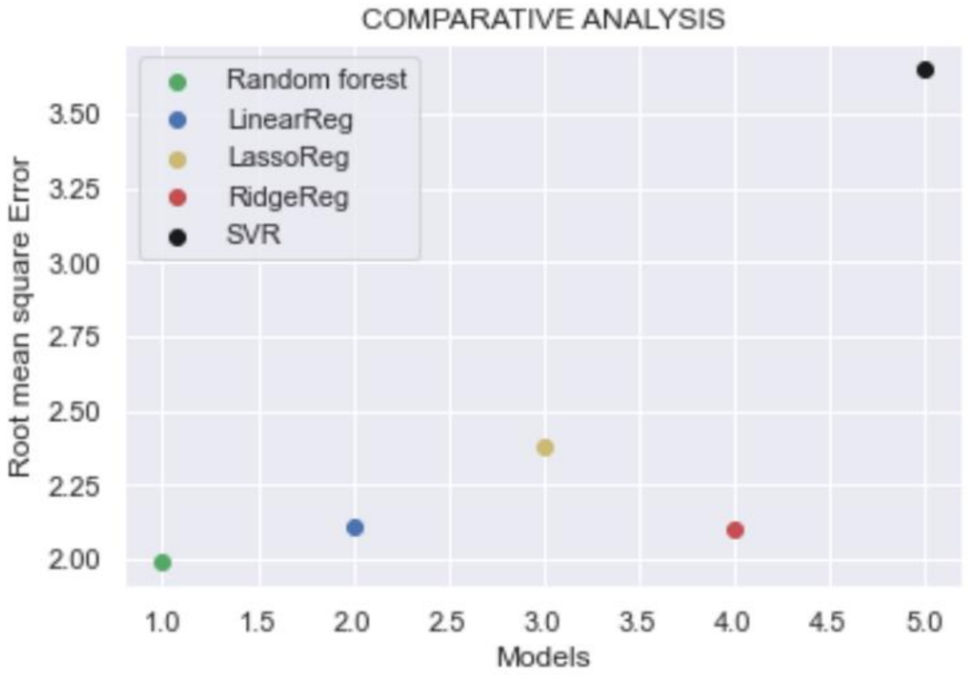


For lasso regression



For ridge regression

3. Models' error plot: -



Algorithm's performance analysis: -

Algorithm	MAE	RMSE	Accuracy
Random forest	0.87	1.98	-4.025
Linear regression	1.24	2.10	85.17%
Lasso regression	1.44	2.37	81.11%
Ridge regression	1.23	2.10	85.20%
SVR	1.67	3.65	55.35%

Chapter 5

Conclusion and Future work

The number of people who wants to sell their used car and to find out the best price for their cars is increasing very rapidly. We discussed various research published over the recent years.

We used and analyzed various supervised learning models to implement our task. For each part, we modified or replaced the component to see the influence on the final result. We used the dataset from cardekho.com for training and evaluated our model using various algorithms. We found out that random forest regression is the best technique that can predict the accurate results. The amount of data will never stop increasing and new information will keep appearing, so future studies should consider if static models are good enough when thinking of long-term application or if lifelong learning should be increasingly thought of.

Of course, this is just a first-cut solution and a lot of modifications can be made to improve this solution like:

- We can enlarge the datasets by combining it with the data from various online platforms like OLX, so that we can get more accurate predictions.
- Making the front-end for the python file so as to create a website where a customer can enter the important details and find out the accurate selling price of the car for entered details.
- Doing more hyper parameter tuning (learning rate, batch size, number of layers, number of units, dropout rate, batch normalization etc.). We can implement this on SVR and find out if the accuracy improves.
- Use the cross validation set to understand overfitting.
- Writing the code in a proper object-oriented way so that it becomes easier for others to replicate.

REFERENCES

[1].Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani, ―Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization‖, 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, January 2018

[2].Hossein Hadian Jazi, Hugo Gonzalez, Natalia Stakhanova, and Ali A. Ghorbani. "Detecting HTTP-based Application Layer DoS attacks on Web Servers in the presence of sampling." Computer Networks, 2017

[3]. A. Shiravi, H. Shiravi, M. Tavallaee, A.A. Ghorbani, Toward developing a systematic approach to generate benchmark datasets for intrusion detection, Comput. Security 31 (3) (2012) 357–374.

[4].Z. He, T. Zhang, and R. B. Lee, ―Machine Learning Based DDoS Attack Detection from Source Side in Cloud,‖ in Proceedings of the 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud), pp. 114–120, New York, NY, USA, June 2017

[5].R. Doshi, N. Apthorpe, and N. Feamster, "Machine Learning DDoS Detection for Consumer Internet of Things Devices," 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, 2018, pp. 29-35.

[6].Jerome H. Friedman, (2002), Stochastic gradient boosting, Computational Statistics & Data Analysis, 38, (4), 367-378