

CAR PRICE PREDICTION PROJECT

USE CASE REPORT



Submitted By :-

Kirti Ohlan

**ACKNOWLEDGMENT**

I would like to express my deepest gratitude to my SME (Subject Matter Expert) Shwetank Mishra as well as Flip Robo Technologies who gave me the opportunity to do this project on Car Price Prediction, which also helped me in doing lots of research wherein I came to know about so many new things.

Also, I have utilized a few external resources that helped me to complete the project. I ensured that I learn from the samples and modify things according to my project requirement. All the external resources that were used in creating this project are listed below:

1. <https://www.google.com/>
2. <https://www.youtube.com/>
3. <https://scikit-learn.org/stable/user_guide.html>
4. <https://github.com/>
5. <https://www.kaggle.com/>
6. <https://medium.com/>
7. <https://towardsdatascience.com/>
8. <https://www.analyticsvidhya.com/>

**INTRODUCTION**

* Business Problem Framing
* Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle’s price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models . Conventional linear regression also yielded satisfactory results, with the advantage of a significantly lower training time in comparison to the aforementioned methods.
* Conceptual Background of the Domain Problem

We will compare the performance of various machine learning algorithms like Linear Regression, Ridge Regression, Decision Tree Regressor , XGBoost and choose the best out of it. Depending on various parameters we will determine the price of the car. Regression Algorithms are used because they provide us with continuous value as an output and not a categorized value because of which it will be possible to predict the actual price a car rather than the price range of a car. User Interface has also been developed which acquires input from any user and displays the Price of a car according to user’s inputs.

You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

1. Which variables are important to predict the price of a variable?
2. How do these variables describe the price of the car?

* Review of Literature
* Based on the sample data provided to us from our client database where we have understood that the company is looking at prospective properties to buy car to enter the market. The data set explains it is a regression problem as we need to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. Also, we have other independent features that would help to decide which all variables are important to predict the price of the variable and how do these variables describe the price of the car.

* Motivation for the Problem Undertaken

Deciding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used car prices.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem
* We are building a model in Machine Learning to predict the actual value of the prospective properties and decide whether to invest in them or not. So, this model will help us to determine which variables are important to predict the price of variables & also how do these variables describe the price of the car. This will help to determine the price of cars with the available independent variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns.
* Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features'). The most common form of regression analysis is linear regression, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion. For specific mathematical reasons this allows the researcher to estimate the
* conditional expectation of the dependent variable when the independent variables take on a given set of values.
* Regression analysis is also a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.
* Data Sources and their formats
* The dataset is in CSV (Comma Separated File) format. Dataset is scraped by cars24 website.
* Size of dataset is 5963 rows and 9 columns
* The model will learn from this file. It contains all the independent variables and the target variable.

* Data Preprocessing Done

Data pre-processing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. Data pre-processing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we pre- process our data before feeding it into our model. Therefore, it is the first and crucial step while creating a machine learning model. I have used some following pre-processing steps:

1. Loading the training dataset as a dataframe
2. Used pandas to set display I ensuring we do not see any truncated information
3. Checked the number of rows and columns present in our training dataset
4. Checked for missing data and the number of rows with null values
5. Dropped all the unwanted columns and duplicate data present in our dataframe
6. Checked the unique values information in each column to get a list for categorical data
7. Performed imputation to fill missing data using mean on numeric data and mode for categorical data columns
8. Used Pandas Profiling during the visualization phase along with distplot,barplot, count plot, scatter plot and the others
9. With the help of ordinal encoding technique converted all object datatype columns to numeric datatype
10. Thoroughly checked for outliers and skewness information
11. With the help of heatmap, correlation bar graph was able to understand the Feature vs Label relativity and insights on multicollinearity amongst the feature columns
12. Separated feature and label data to ensure feature scaling is performed avoiding any kind of biasness
13. Checked for the best random state to be used on our Regression Machine Learning model pertaining to the feature importance details
14. Finally created a regression model function along with evaluation metrics to pass through various model formats

* Data Inputs- Logic- Output Relationships

When we loaded the training dataset, we had to go through various data pre processing steps to understand what was given to us and what we were expected to predict for the project. When it comes to logical part the domain expertise of understanding how real estate works and how we are supposed to cater to the customers came in handy to train the model with the modified input data. In Data Science community there is a saying “Garbage In Garbage Out” therefore we had to be very cautious and spent almost 80% of our project building time in understanding each and every aspect of the data how they were related to each other as well as our target label.

With the objective of predicting car sale prices accurately we had

to make sure that a model was built that understood the customer

priorities trending in the market imposing those norms when a relevant

price tag was generated. I tried my best to retain as much data possible

that was collected but I feel discarding columns that had lots of missing

data was good. I did not want to impute data and then cause a biasness in

the machine learning model from values that did not come from real

people.

* Hardware and Software Requirements and Tools Used

Hardware Used:

1. RAM: 8 GB
2. CPU: AMD Ryzen 5 3550H with Radeon Vega Mobile Gfx 2.10 GHz
3. GPU: AMD Radeon ™ Vega 8 Graphics andNVIDIA GeForce GTX 1650 Ti

Software Used:

1. Programming language: Python
2. Distribution: Anaconda Navigator
3. Browser based language shell: Jupyter Notebook

Libraries/Packages Used:

Pandas, NumPy, matplotlib, seaborn, scikit-learn and pandas\_profiling

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* I have used both statistical and analytical approaches to solve the problem which mainly includes the pre-processing of the data and EDA to check the correlation of independent and dependent features. Also, before building the model, I made sure that the input data is cleaned and scaled before it was fed into the machine learning models.
* For this project we need to predict the sale price of cars, means our target column is continuous so this is a regression problem. I have used various regression algorithms and tested for the prediction. By doing various evaluations I have selected XGBRegressor as best suitable algorithm for our final model as it is giving good r2-score and least difference in r2-score and CV-score among all the algorithms used.Other regression algorithms are also giving me good accuracy but some are over-fitting and some are with under-fitting the results which may be because of less amount of data.
* In order to get good performance as well as accuracy and to check my model from over-fitting and under-fitting I have made use of the K-Fold cross validation and then hyper parameter tuned the final model.
* Once I was able to get my desired final model I ensured to save that model before I loaded the testing data and started performing the data pre-processing as the training dataset and obtaining the predicted sale price values out of the Regression Machine Learning Model.

* Testing of Identified Approaches (Algorithms)

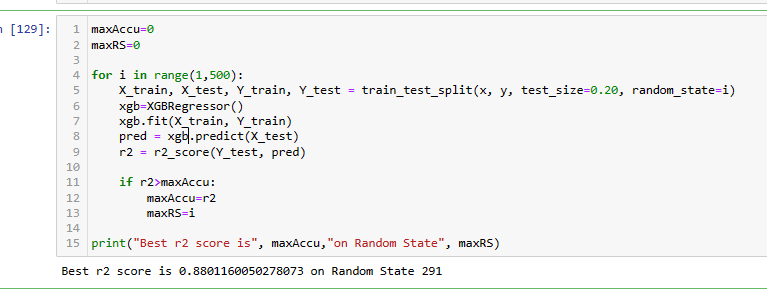
The algorithms used on training and test data are as follows:

1. Linear Regression Model
2. Ridge Regularization Regression Model
3. Support Vector Regression Model
4. Decision Tree Regression Model
5. Random Forest Regression Model
6. K Nearest Neighbours Regression Model
7. XGBBoost Regression Model

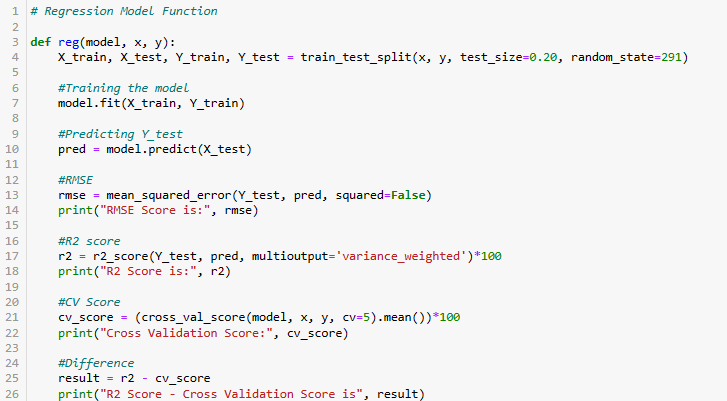
* Run and Evaluate selected models

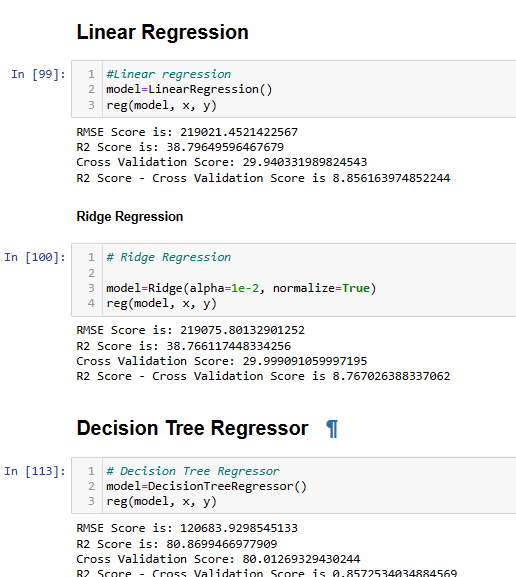
I used a total of 10 Regression Models after choosing the random state amongst 1-500 number. Then I even defined a function for getting the regression model trained and evaluated. The code for the models is listed below.

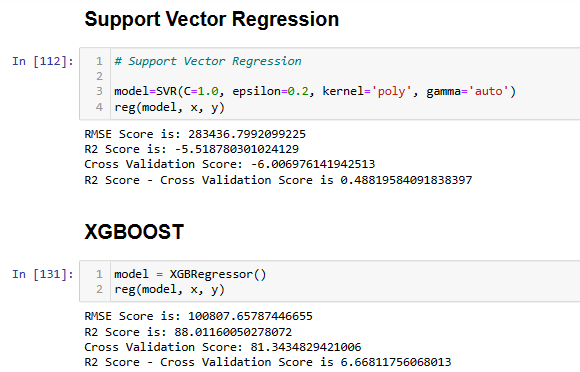
* Random state:-



* Regression model function







* Key Metrics for success in solving problem under consideration

The key metrics used here were r2\_score, cross\_val\_score, MAE, MSE and RMSE. We tried to find out the best parameters and also to increase our scores by using Hyperparameter Tuning and we will be using GridSearchCV method.

* 1. Cross Validation:

Cross-validation helps to find out the over fitting and under fitting of the model. In the cross validation the model is made to run on different subsets of the dataset which will get multiple measures of the model. If we take 5 folds, the data will be divided into 5 pieces

where each part being 20% of full dataset. While running the Cross-validation the 1st part (20%) of the 5 parts will be kept out as a holdout set for validation and everything else is used for training data. This way we will get the first estimate of the model quality of the dataset.

In the similar way further iterations are made for the second 20% of the dataset is held as a holdout set and remaining 4 parts are used for training data during process. This way we will get the second estimate of the model quality of the dataset. These steps are repeated during the cross-validation process to get the remaining estimate of the model quality.

* 2. R2 Score:

It is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted.

* 3. Mean Squared Error (MSE):

MSE of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors — that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. RMSE is the Root Mean Squared Error.

* 4. Mean Absolute Error (MAE):

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

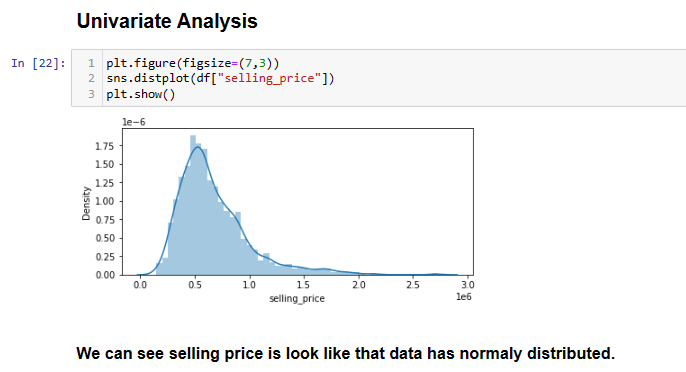
* 5. Hyperparameter Tuning:

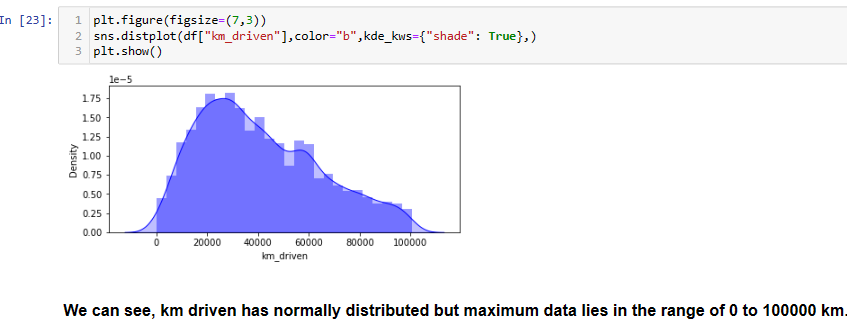
There is a list of different machine learning models. They all are different in some way or the other, but what makes them different is nothing but input parameters for the model. These input parameters are named as Hyperparameters. These hyperparameters will define the architecture of the model, and the best part about these is that you get a choice to select these for your model. You must select from a specific list of hyperparameters for a given model as it varies from model to model.

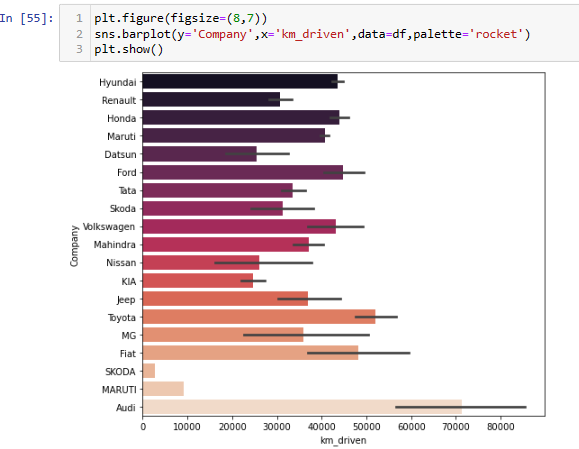
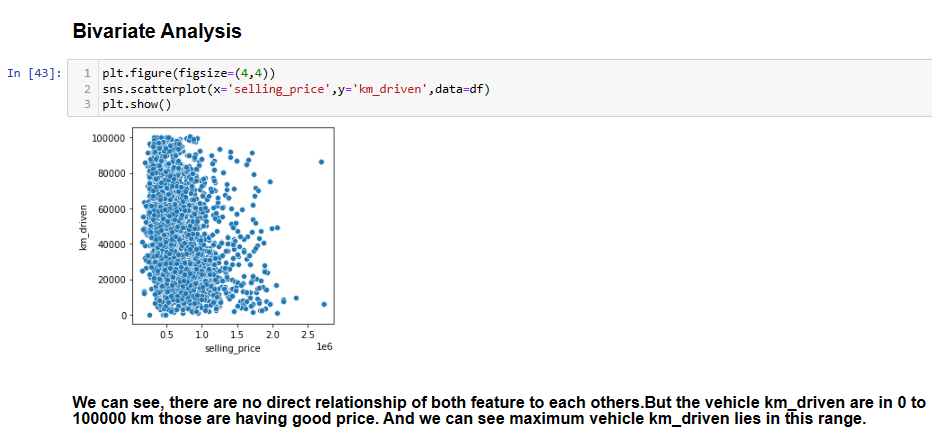
We are not aware of optimal values for hyperparameters which would generate the best model output. So, what we tell the model is to explore and select the optimal model architecture automatically. This selection procedure for hyperparameter is known as Hyperparameter Tuning. We can do tuning by using GridSearchCV.

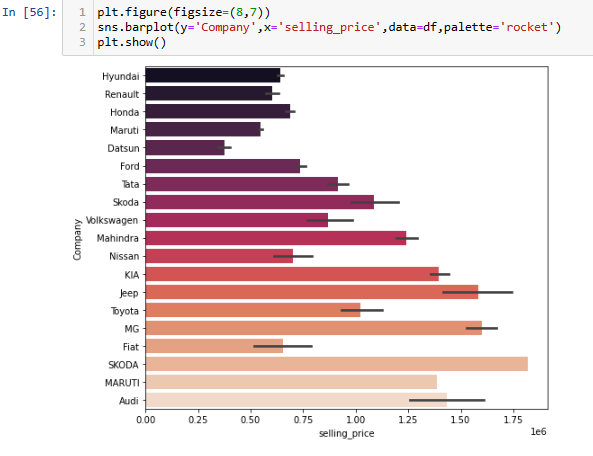
* GridSearchCV is a function that comes in Scikit-learn (or SK-learn) model selection package. An important point here to note is that we need to have Scikit-learn library installed on the computer. This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, we can select the best parameters from the listed hyperparameters.
* Visualizations

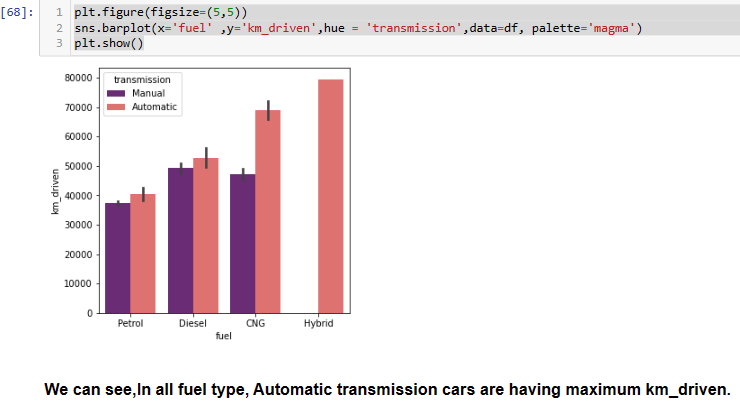
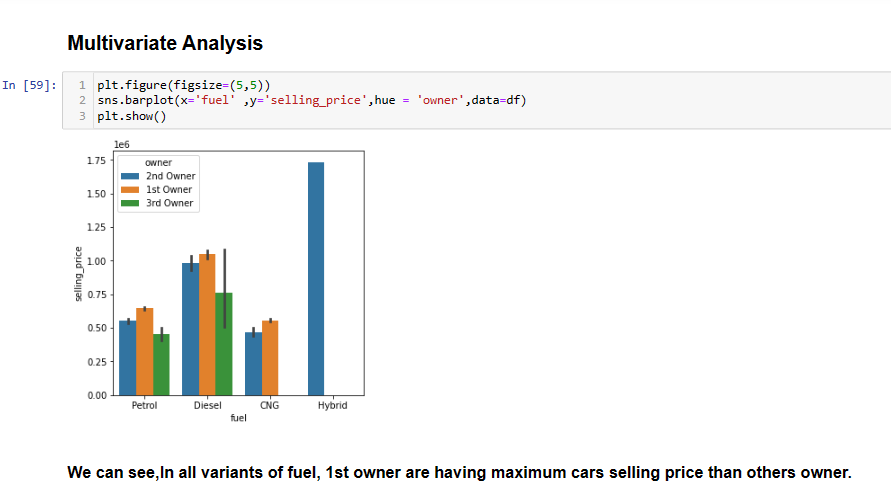
I used pandas profiling to get the over viewed visualization on the pre-processed data.pandas-profiling is an open-source Python module with which we can quickly do an exploratory data analysis with just a few lines of code. It generates interactive reports in web format that can be presented to any person, even if they don’t know programming. It also offers report generation for the dataset with lots of features and customizations for the report generated. In short, what pandas-profiling does is save us all the work of visualizing and understanding the distribution of each variable. It generates a report with all the information easily available.







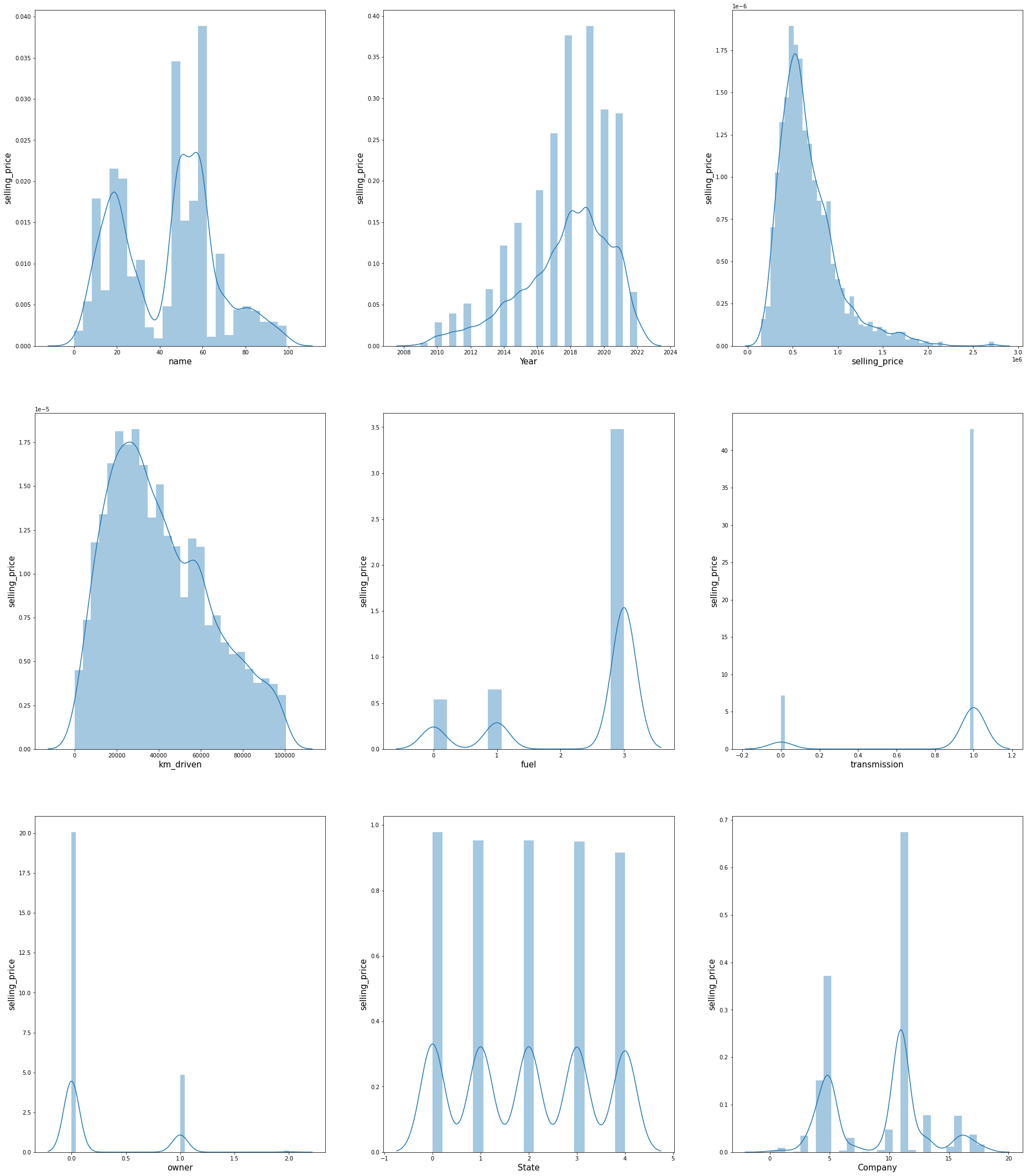


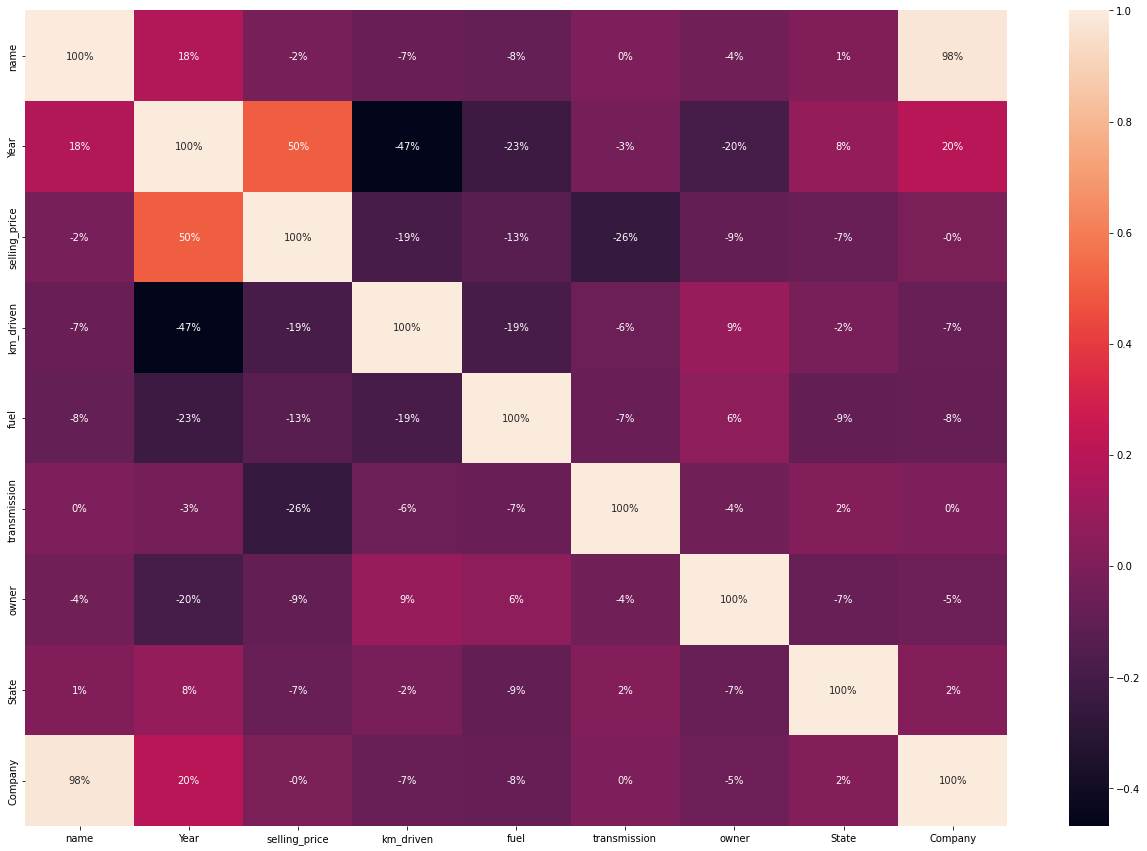


* Interpretation of the Results
* Visualizations: It helped me to understand the correlation between independent and dependent features. Also, helped me with feature importanceand to check for multi collinearity issues. Detected outliers/skewness with the help of boxplot and distribution plot. I got to know the count of a particular category for each feature by using count plot and most importantly with predicted target value distribution as well asscatter plot helped me to select the best model.
* Pre-processing: Basically, before building the model the dataset should be cleaned and scaled by performing few steps. As I mentioned above in the pre-processing steps where all the important features are present in the dataset and ready for model building.
* Model Creation: Now, after performing the train test split, I have x\_train, x\_test, y\_train&y\_test, which are required to build Machine learning models. I have built multiple regression models to get the best R2 score, MSE, RMSE & MAE out of all the models.

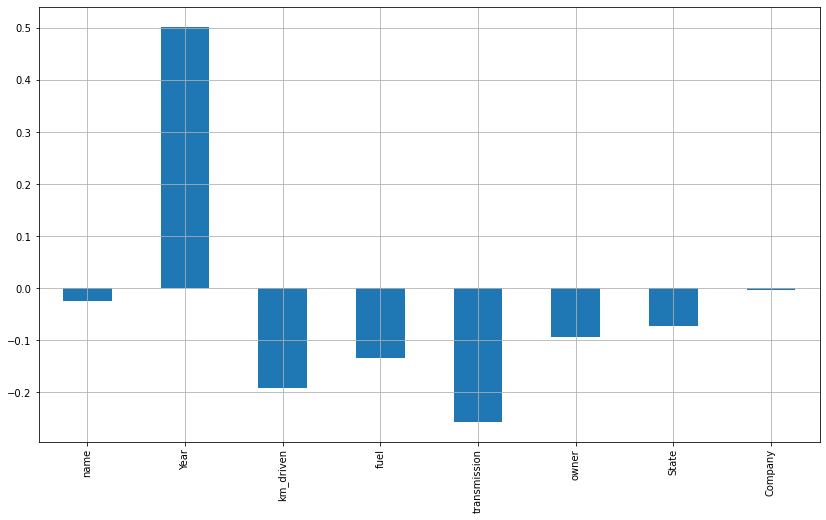
**CONCLUSION**

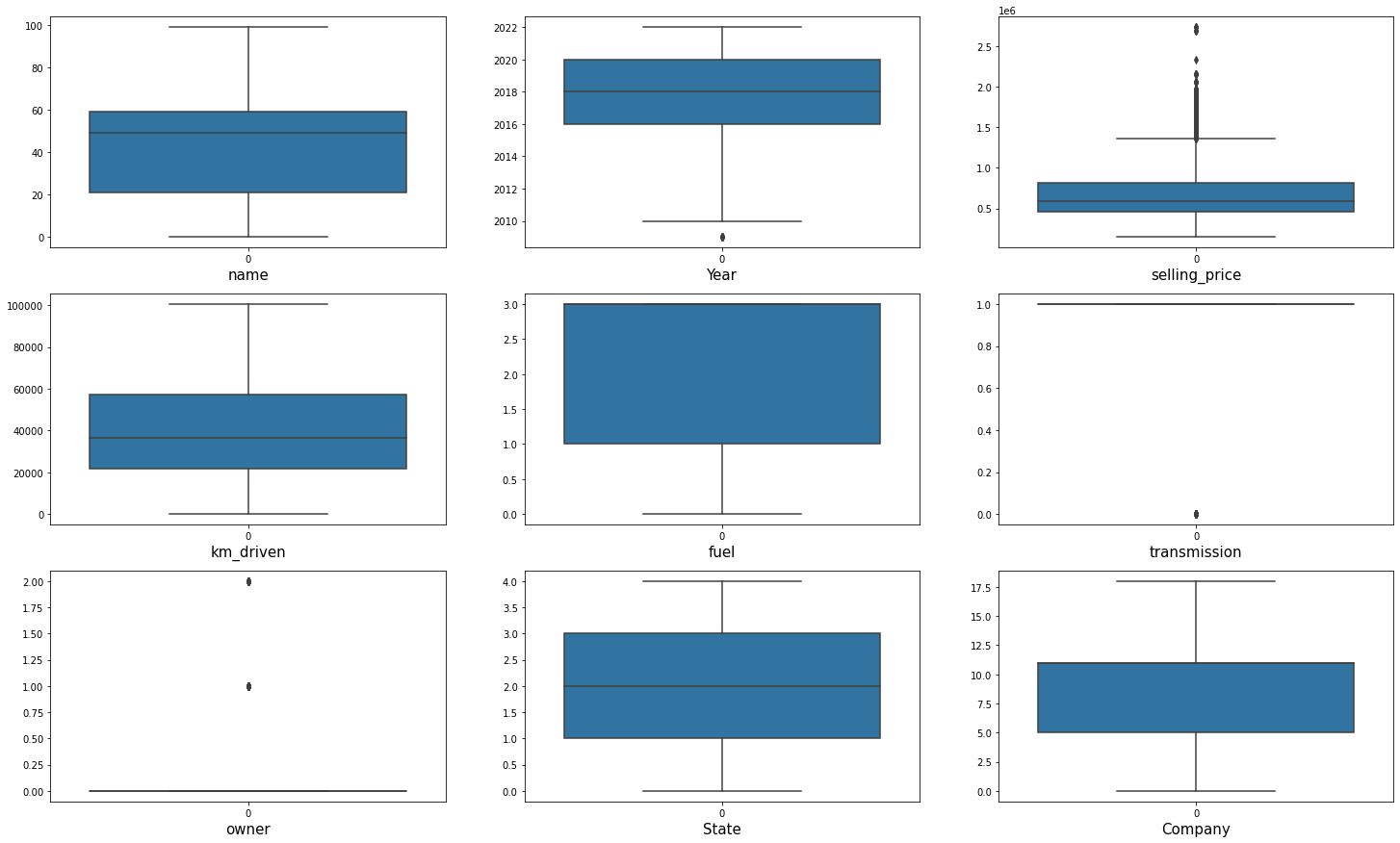
* Key Findings and Conclusions of the Study
* I observed all the encoded dataset information by plotting various graphs and visualised further insights

**DISTPLOT:-**

**HEATMAP :**-

**CORRELATION :**-



BOXPLOT:-

* Learning Outcomes of the Study in respect of Data Science

Since India’s used-car market is booming as buyers have a wide range of options, easy financing, convenient digital sales channels, and a growing preference for personal mobility in the COVID-19 era, car prediction can be a challenging task due to the high number of attributes that should be considered for accurate prediction. The main weakness of Gradient boosting is that it sacrifices intelligibility and interpretability. The main limitation of this study is the low number of records that have been used. In future work, we intend to collect more data related to electric vehicles [22] and combustion vehicles and to use more advanced techniques.

* Limitations of this work and Scope for Future Work

In future this machine learning model may bind with various website which can provide real time data for price prediction. Also we may add large historical data of car price which can help to improve accuracy of the machine learning model. We can build an android app as user interface for interacting with user. For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset

