NAME OF THE PROJECT

CAR PRICE PREDICTION

Submitted by:

MD AKIF PERWEZ

ACKNOWLEDGMENT

I would like to thank our SME (Sajid Chaudhary) for his expert advice and encouragement throughout this project.

INTRODUCTION

- Business Problem Framing

  In this project, we have to make car price valuation model using new machine learning models from new data. Because with the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models

- Conceptual Background of the Domain Problem

  1. Firstly, we will prepare our own dataset using web scraping.

  2. After that we will check whether the project is a regression type or a classification type.

  3. We will also check whether our dataset is balanced or imbalanced. If it is an imbalanced one, we will apply sampling techniques to balance the dataset.

  4. Then we will do model building and check its accuracy.

  5. Our main motto is to build a model with good accuracy and for that we will also go for hyper parameter tuning.

  Review of Literature

  I am summarizing my research done on the topic.
  ▪ I have created my own dataset using web scraping and imported important libraries for my project. ▪ I have created the data frame.
  ▪ I have analysed my data by checking its shape, number of columns, presence of null values if any and checking the datatypes.
  ▪ Then I have done some data cleaning steps, e.g. Checking the value counts of the target variable, dropping some irrelevant columns from the dataset, checking correlation between the dependant and independent variables using heat map,

visualizing data using distribution plots, detecting and removing skewness in my data if any, outliers detection using boxplots and removing them, balancing dataset using

Motivation for the Problem Undertaken

I have been working towards many project, but this project give me immense pleasure to do.

We all know how different cars are used in the market and the price are quite different for all of them. My work is to make best model and suggest our client to use this which help them to make their business more and more profit.

After all we engineers can provide best solution to the clients to improve their business.

In this Project I used my potential to scrap data from the different website like OLX, CarsDekho, and Cars24 which help to get the actual price of the market.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  If you look at data science, we are actually using mathematical models to model (and hopefully through the model to explain some of the things that we have seen) business circumstances, environment etc. and through these model, we can get more insights such as the outcomes of our decision undertaken, what should we do next or how shall we do it to improve the odds. So mathematical models are important, selecting the right one to answer the business question can bring tremendous value to the organization. Here I am using Random Forest Regressor with accuracy 100% after hyper parameter tuning.

- Data Sources and their formats

  Data Source: The read_xlsx function of the pandas library is used to read the content of an Excel file into the python environment as a pandas Data Frame. The function can read the files from the OS by using proper path to the file. Data description: Pandas describe() is used to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values

  Data Pre-processing Done

  - I have checked for null values and there are some null values present, I have removed it using interpolate method.

- I have label encoded the object type columns in the dataset.
- I have checked the correlation between dependant and independent variables using heat map. I have seen most of the independent variables are correlated with each other and the target variable is positively correlated with a very few independent variables.
- I have done some visualization using histogram.
- I have checked outliers using boxplots, outliers are present in price column.
- I also have checked for skewness in my data, but the skewness present is very negligible, so I don't consider it.
- I have splitted the dependant and independent variables into x and y.
- I have scaled the data using StandardScaler method and made my data ready for model building

- Hardware and Software Requirements and Tools Used
  - Hardware requirements: Processor: Intel(R) Core (TM) i3-4030U CPU @ 1.90GHz RAM: 3.98 GB System type: 64-bit operating system, x64-based processor
  - Software requirements:

    Python: One of the most used programming languages Tools used: Jupyter notebook: Jupyter is a free, open-source, interactive web tool known as a computational notebook where I have written my python codes. NumPy: NumPy is an open-source numerical Python library. NumPy contains a multi-dimensional array and matrix data structures. Pandas: Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support

for multi-dimensional arrays. Matplotlib: It is a Python library used for plotting graphs with the help of other libraries like Numpy and Pandas. It is a powerful tool for visualizing data in Python. Seaborn: It is also a Python library used for plotting graphs with the help of Matplotlib, Pandas, and Numpy.

Scikit-learn:
It is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, and clustering and dimensionality reduction. Scipy.stats: This module contains a large number of probability distributions as well as a growing library of statistical functions

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  1. To check the correlation among the data, I have used heat map to visualize it.
  2. To get a clear view of the columns visually, I have used distribution plots.
  3. For checking outliers, I have used boxplots.
  4. For scaling the data, I have used StandardScaler method.
  5. For training and testing the data, I have imported train_test_split library from scikit-learn. 6. For model building, I have used 5 Regressor models(DecisionTreeRegressor(),KNeighborsRegressor(),AdaBo

ostRegr
essor(),LinearRegression(),GradientBoostingRegressor()),out of which AdaBoostRegressor model is the best model for my dataset.
7. For better accuracy of the model, I have used hyper parameter tuning

Testing of Identified Approaches (Algorithms)

I have used 5 algorithms for testing

- **DecisionTreeRegressor**- Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
- **RandomForestRegressor**- A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.
- **AdaBoostRegressor**- An AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction.
- **LinearRegression**- Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).
- **GradientBoostingRegressor**- Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier).

- Run and Evaluate selected models
  Finding the Random State

```
from sklearn.metrics import r2_score
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
maxAucc=0
maxRS=0
for i in range(1,100):
    x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.20,random_state=i)
    le=LinearRegression()
    le.fit(x_train,y_train)
    pred=le.predict(x_test)
    acc=r2_score(y_test,pred)
    if acc>maxAucc:
        maxAucc=acc
        maxRS=i
print("Best Accuracy is ",maxAucc,"on random_state",maxRS)
```

```
Best Accuracy is  0.07684807400038496 on random_state 84
```

- Key Metrics for success in solving problem under consideration

Random Forest

```
parameters={"n_estimators":np.arange(2,20),
            "criterion":["mse","mae"]},
GCV=GridSearchCV(RandomForestRegressor(),parameters,cv=5)
GCV.fit(x_trains,y_train)
GCV.fit(x_trains,y_train)
```

```
GridSearchCV(cv=5, estimator=RandomForestRegressor(),
             param_grid=({'criterion': ['mse', 'mae'],
                          'n_estimators': array([ 2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16, 17, 18,
       19])},))
```

```
GCV.best_params_
```

```
{'criterion': 'mae', 'n_estimators': 2}
```

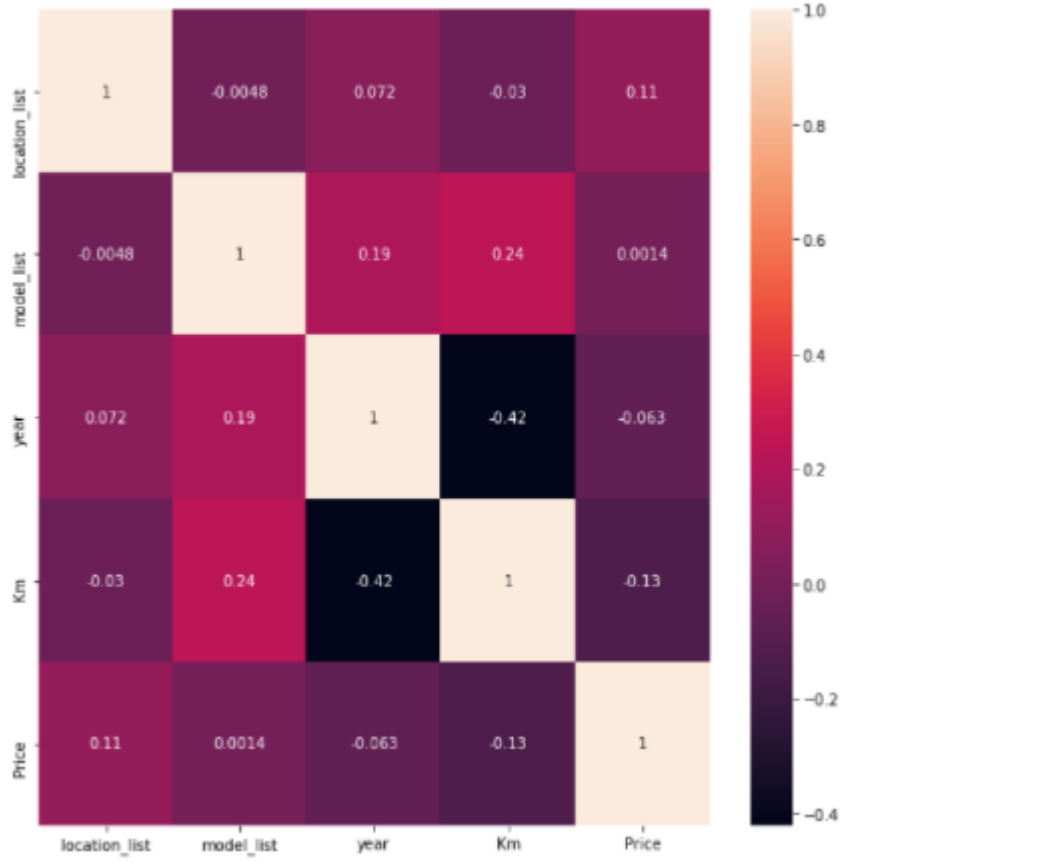
```
rf_mod=RandomForestRegressor(criterion="mae",n_estimators=2)

rf_mod.fit(x_trains,y_train)
pred=rf_mod.predict(x_tests)
print(r2_score(y_test,pred)*100)
```

```
100.0
```

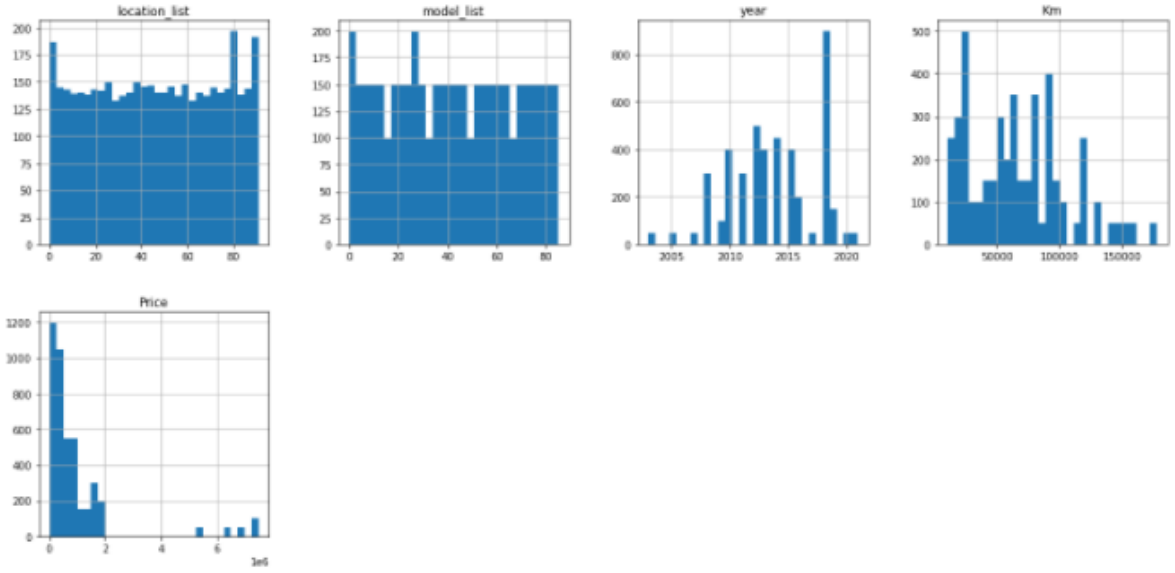

- Visualizations

```
plt.figure(figsize=(10,10))
sns.heatmap(df.corr(),annot=True)
plt.show()
```



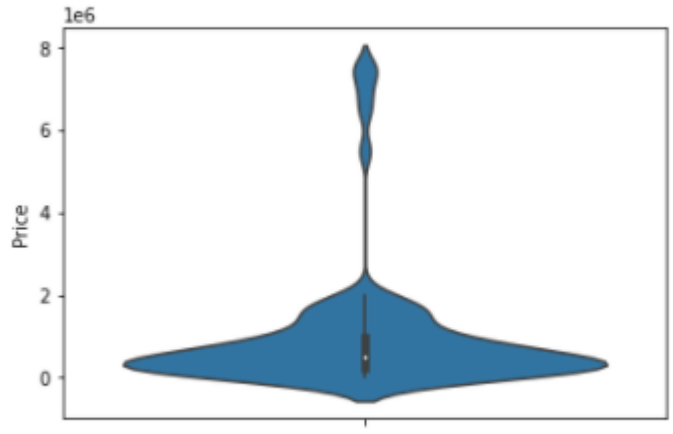
```
#Plotting histogram
#A histogram shows the frequency on the vertical axis and the horizontal axis in another dimension.
# In this graph, we can also check whether the graph is right skewed, left skewed or the graph is normally distributed graph
df.hist(figsize=(20,20),grid=True,layout=(4,4),bins=30)
```

```
array([[<AxesSubplot:title={'center':'location_list'}>,
        <AxesSubplot:title={'center':'model_list'}>,
        <AxesSubplot:title={'center':'year'}>,
        <AxesSubplot:title={'center':'Km'}>],
       [<AxesSubplot:title={'center':'Price'}>, <AxesSubplot:>,
        <AxesSubplot:>, <AxesSubplot:>],
       [<AxesSubplot:>, <AxesSubplot:>, <AxesSubplot:>, <AxesSubplot:>],
       [<AxesSubplot:>, <AxesSubplot:>, <AxesSubplot:>, <AxesSubplot:>]],
      dtype=object)
```




```
sns.violinplot(y="Price",data=df)
```

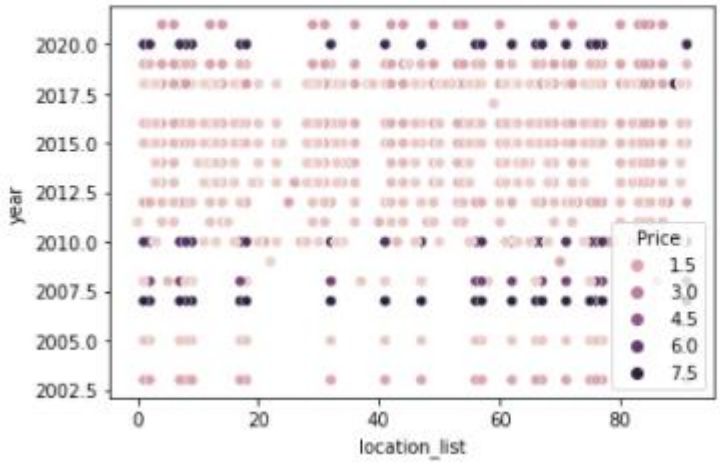
```
<AxesSubplot:ylabel='Price'>
```



Here most price are lying in between 0 to 3

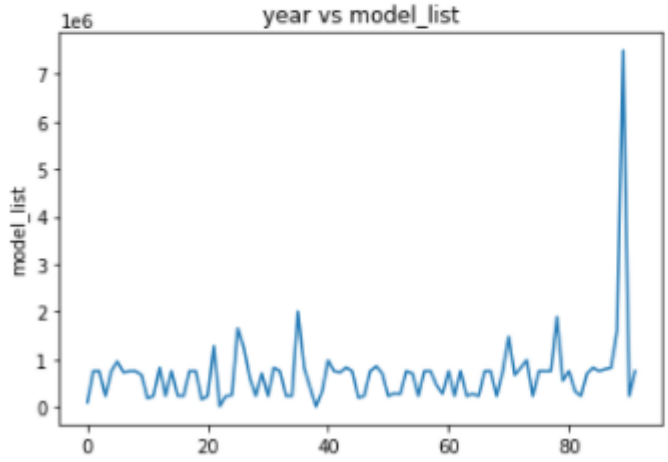
```
sns.scatterplot(x="location_list",y="year",data=df,hue="Price")
```

```
<AxesSubplot:xlabel='location_list', ylabel='year'>
```



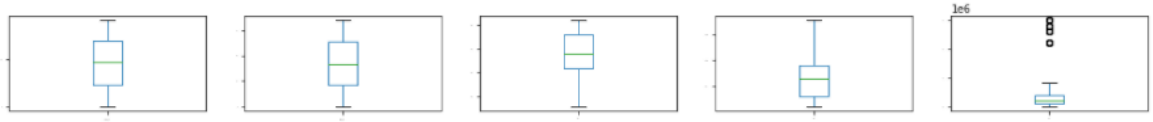

```
df.groupby('location_list')['Price'].median().plot()
plt.xlabel('year')
plt.ylabel('model_list')
plt.title("year vs model_list")
```

```
Text(0.5, 1.0, 'year vs model_list')
```



Checking Outliers

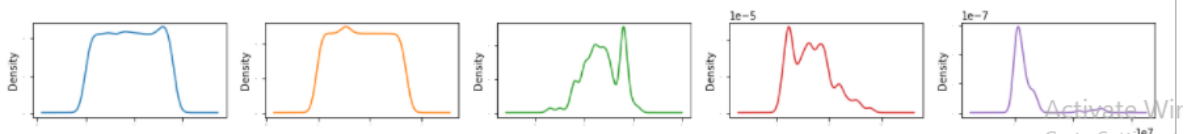
```
df.plot(kind="box",subplots=True,layout=(10,10),sharex=False,legend=False,fontsize=1,figsize=(40,20))
plt.show()
```



outliers are in price column

Skewness

```
df.plot(kind="density",subplots=True,layout=(10,10),sharex=False,legend=False,fontsize=1,figsize=(40,20))
plt.show()
```



- Interpretation of the Results

  In the visualization part, I have seen how my data looks like using heat map, boxplot, distribution plots, histogram, Scatter plot etc. In the pre-processing part, I have cleaned my data using many methods like interpolate, Label Encoder etc. In the modelling part, I have designed our model using algorithm like AdaBoostRegressor. The accuracy, Mean Absolute Error, Mean Squared Error, Root Mean Absolute Error are achieved for the model.

CONCLUSION

- Key Findings and Conclusions of the Study

  The key findings are we have to study the data very clearly so that we are able to decide which data are relevant for our findings. The techniques that I have used are heat map, interpolate, Label Encoder etc. The conclusion of our study is we have to achieve a model with good accuracy and f1-score

- Learning Outcomes of the Study in respect of Data Science

  We will develop relevant programming abilities. We will demonstrate proficiency with statistical analysis of data. We will develop the ability to build and assess data-based models. We will execute statistical analysis with professional statistical software. The best algorithm for this project according to my work is AdaBoostRegressor because the accuracy that I have achieved is quite satisfactory than the other model.

- Limitations of this work and Scope for Future Work

  The scope for future work is to collect as many data as we can so that the model can be built more efficiently.