A Project Report

on

PREDICTIVE SYSTEM ON THE CAR MARKET TREND

Submitted in partial fulfillment of the

requirement for the award of the degree of

Bachelor of Technology in Computer Science and Engineering

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
Ms. Pushpa Singh
Assistant Professor
Department of Computer Science and Engineering

Submitted By

20SCSE1180088 – ANSH SHANKAR
20SCSE1180139 – DHRUV VARSHNEY

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING /
DEPARTMENT OF COMPUTERAPPLICATION
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA

DECEMBER – 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

We hereby certify that the work which is being presented in the project, entitled **"PREDICTIVE SYSTEM ON THE CAR MARKET TREND"** in partial fulfillment of the requirements for the award of the **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of **JULY-2021 to DECEMBER-2021**, under the supervision of **Ms. Pushpa Singh, Assistant Professor,** Department of Computer Science and Engineering of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the project has not been submitted by me/us for the award of any other degree of this or any other places.

20SCSE1180088 – ANSH SHANKAR

20SCSE1180139 – DHRUV VARSHNEY

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Supervisor

(Ms. Pushpa Singh, Assistant Professor)

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of **20SCSE1180088 – ANSH SHANKAR , 20SCSE1180139 – DHRUV VARSHNEY** has been held on _________________ and his/her work is recommended for the award of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING**

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: December, 2021

Place: Greater Noida

Abstract

Automobile Manufacturing is one of the most sophisticated sets of processes in the World. And for the automobile to be Successful in the market, it requires an extensive amount of work to be done in the field of Market Analysis. This is an area of major concern for companies.

So, we decided to use a Machine Learning model that would predict based on the history of the Cars Manufactured in India. We analyzed the trend of Market demand and build a predictive model that would predict that whether a car would be successful in the market or not. Many other factors could be predicted using the analysis such as which car colour and type of car should the manufacturer build to maximize sales. There is also a "User Section", wherein the customer can check if the car is value for money, or wait for a more suitable car to be available in the market, which can also predict the price of a new car based on the brand name and other features.

Here we have used a Linear Regression technique to build a Machine Learning model. This model is trained on multiple datasets collected from different sources which are then analyzed and processed to obtain desired results. Thus this project is a very handy tool for both manufacturers and customers.

Since this business of car manufacturing is unceasing, hence the process of data generation is also never-ending. This project's accuracy will get better with time as more data is available.

List of Figures

Acronyms

B.Tech.	Bachelor of Technology
SCSE	School of Computing Science and Engineering
MPC	Model Predictive Control
FV	Follower Vehicle

Table of Contents

CHAPTER-1
Introduction

Automobile Manufacturing is one of the most painful process for the car Manufactures in terms of both labour and financial. But if automobile is not well received in the market, the pain and the treasure is lost. For example, for production of new car the company has to setup whole production unit along with its lifeline i.e. Workers. But what if its sales are low or what if the car doesn’t match the standards of the population for which the car is manufactured, the whole ecosystem and the hard work goes to vain.

So we decided to come up with a model that would be a boon to this type situation. We have collected automobiles dataset of different companies from different sources, analyzed and processed it to obtain the information, upon which our machine learning model is trained. The model is based on Linear Regression technique. The manufacturer could extract from it which type of car to be manufactured, what would be the most prominent colour, what would be transmission type, fuel type, etc.

Apart this there is a “Customer Section”, wherein customer can check if the car present in the market is value for money or not. Weather he/she is buying the car at the perfect price or not, or weather if he should wait for some more exciting deal.

The Whole Project is divided into the 5 parts :-

- Data Collection
- Data understanding and exploration
- Data cleaning
- Data preparation
- Model building and evaluation

CHAPTER-2
Literature survey

[1] discussed that a predictive system for car fuel consumption using a back-propagation neural network is proposed in this paper. The suggested system has three components: an information gathering system, a fuel consumption forecasting algorithm, and a performance assessment system. Although there are many factors that influence a car's fuel consumption in a practical drive procedure, the impact factors for fuel consumption in the current system are simply determined as the make of the car, engine style, vehicle type, weight of the car, and transmission system type. An artificial neural network with back-propagation neural network has a learning capability for automobile fuel consumption prediction to test the effect of the proposed predictive system in the fuel consumption forecasting. The results of the prediction showed that the proposed neural network system is successful in predicting fuel consumption and that its performance is adequate..

[2] discussed that A high-interest study topic has been automobile price prediction, as it necessitates significant effort and expertise on the part of the field expert. For a trustworthy and accurate forecast, a large number of unique attributes are considered. They utilised three machine learning approaches to create a model for forecasting the price of secondhand automobiles in Bosnia and Herzegovina (Artificial Neural Network, Support Vector Machine and Random Forest). However, the approaches suggested were used in a group setting.

[3] discussed that Their study looks at how sentiment analysis and Google trends data may be used to anticipate automobile sales. Previous study has proven the utility of both approaches for sales forecasting, but the findings of current research for forecasting the sales of high-involvement items such as vehicles are more equivocal. In this study, linear regression models are used to evaluate over 500,000 social media postings for eleven automobile models on the Dutch market. In addition, the results of this study are compared to the prediction capacity of Google Trends. The findings reveal that while social media emotions have limited predictive value when it comes to automobile sales, Google Trends data and social mention volume show substantial results and may be combined into a useful prediction model. The automotive industry may utilise decision tree regression to build a prediction model with temporal delays that can be employed in addition to standard forecasting approaches.

[4] discussed that because of its safety and operational efficiency, automobile following control is critical. This study uses a linear and continuous model of automobile following for this goal, as well as a Model Predictive Control (MPC) controller. This form of control has the capacity to cope with control limitations, which is a significant benefit. We hired this sort of controller in this study to deal with these restrictions since safety and operational efficiency are constraints for automobile following. The MPC predicts the future behaviour of the leader vehicle (LV) based on the relative distance and relative acceleration of each instant, and the acceleration of the follower vehicle (FV) is regulated based on this behaviour. The MPC aims to keep the relative distance within a safe range by controlling the acceleration. The outcome of the system is compared to the behaviour of real drivers with comparable beginning conditions to evaluate the performance of the developed controller. The simulation findings demonstrate that the MPC controller behaves considerably more safely than real drivers and can give passengers with a pleasant ride.

[5] discussed that Because in order to stay successful in a competitive market, a leasing firm must provide a competitive lease price. It is important to forecast the future price of a used automobile in order to establish the correct pricing. The lease price might be set to meet the car's value degradation if the depreciation is known. Multiple linear regression analysis is a frequently used method for price prediction. However, there are several elements that influence the pricing, making this critical duty difficult. For high-dimensional data, the conventional regression technique may not be appropriate. Support Vector Regression, a contemporary data mining approach that is independent of input dimension, will be used to solve this possible problem. The accuracy of the predictions will next be compared to the statistical regression model. In specifically, using principles from the field of evolutionary search, a fully automated technique for adjusting and implementing SVR is created. The entire machine learning experiment is based on real-world data from a major German automobile manufacturer.

[6] discussed that for long-term forecasting of automobile ownership, an econometric technique is developed. It is compared to other well-known techniques. It is based on estimates of the percentage of family income spent on automobile purchases, as well as an analysis of car pricing and stock. The technique provides a reasonable approximation of previous levels of automobile ownership in the United Kingdom, as well as a prediction that is comparable to Tanner's for the next 15 years. However, Mogridge forecasts a higher saturation level and, as a result, a larger eventual automobile population in the future.

[7] discussed that a design support system is developed in this work that can be integrated into the car side silhouette design tools and can estimate the drag coefficient of a given silhouette. This task is typically performed via two manners: namely wind tunnel testing and computational fluid dynamics (CFD) simulations. Due to the high computational cost for these two approaches, it is impractical to employ them during the silhouette conceptual design stage in a real time. Therefore, a mathematical model is obtained in this study for the drag coefficient estimation of a given silhouette. First, the desired number of silhouettes are generated via a generative design (silhouette sampling) technique so that the silhouettes are evenly distributed in the silhouette design space. Each silhouette is then tested via computational fluid dynamics simulations, and their corresponding drag coefficients (CDs) are obtained. A training dataset is formed with the silhouette geometries and CDs of the silhouettes, and a mathematical model that can estimate the drag coefficient (CD) of a silhouette is finally obtained via principal component analysis (PCA) followed by regression/neural network methods. These three steps are repeated until a desired level of reliable mathematical model is obtained. Finally, three generative design test cases are illustrated based on the mathematical model obtained to predict CD of a given silhouette.

CHAPTER-3
Data Set

A data set (or data set) is a collection of interrelated data, usually presented in tabular form. Each variable's details are stored in columns and each row represents the corresponding record. The dataset can contain data for one or more items, depending on the number of rows. Each entity is called DATUM.

For example, in our dataset we had Manufacturer name, Model, Vehicle type, Sale number, Fuel type as our columns, which basically represents our variable's detail. and each record/data is accommodated in the single row.

For this project, we had employed datasets from many different sources and many different authors, analyzed it, processed it into a single dataset upon which our model is trained.

Manufactu	Model	Sales_in_t	__year_re	Vehicle_ty	Price_in_tl	Engine_siz	Horsepow	Wheelbase	Width	Length	Curb_weig	Fuel_capa	Fuel_effici	Latest_Lau	Power_perf_factor
Acura	Integra	16.919	16.36	Passenger	21.5	1.8	140	101.2	67.3	172.4	2.639	13.2	28	########	58.28015
Acura	TL	39.384	19.875	Passenger	28.4	3.2	225	108.1	70.3	192.9	3.517	17.2	25	########	91.37078
Acura	CL	14.114	18.225	Passenger		3.2	225	106.9	70.6	192	3.47	17.2	26	########	
Acura	RL	8.588	29.725	Passenger	42	3.5	210	114.6	71.4	196.6	3.85	18	22	########	91.38978
Audi	A4	20.397	22.255	Passenger	23.99	1.8	150	102.6	68.2	178	2.998	16.4	27	########	62.77764
Audi	A6	18.78	23.555	Passenger	33.95	2.8	200	108.7	76.1	192	3.561	18.5	22	########	84.56511
Audi	A8	1.38	39	Passenger	62	4.2	310	113	74	198.2	3.902	23.7	21	2/27/2012	134.6569
BMW	323i	19.747		Passenger	26.99	2.5	170	107.3	68.4	176	3.179	16.6	26	6/28/2011	71.19121
BMW	328i	9.231	28.675	Passenger	33.4	2.8	193	107.3	68.5	176	3.197	16.6	24	1/29/2012	81.87707
BMW	528i	17.527	36.125	Passenger	38.9	2.8	193	111.4	70.9	188	3.472	18.5	25	########	83.99872
Buick	Century	91.561	12.475	Passenger	21.975	3.1	175	109	72.7	194.6	3.368	17.5	25	########	71.18145
Buick	Regal	39.35	13.74	Passenger	25.3	3.8	240	109	72.7	196.2	3.543	17.5	23	########	95.6367
Buick	Park Avenu	27.851	20.19	Passenger	31.965	3.8	205	113.8	74.7	206.8	3.778	18.5	24	3/23/2012	85.82841
Buick	LeSabre	83.257	13.36	Passenger	27.885	3.8	205	112.2	73.5	200	3.591	17.5	25	7/23/2011	84.25453
Cadillac	DeVille	63.729	22.525	Passenger	39.895	4.6	275	115.3	74.5	207.2	3.978	18.5	22	2/23/2012	113.8546
Cadillac	Seville	15.943	27.1	Passenger	44.475	4.6	275	112.2	75	201		18.5	22	4/29/2011	115.6214
Cadillac	Eldorado	6.536	25.725	Passenger	39.665	4.6	275	108	75.5	200.6	3.843	19	22	11/27/201	113.7659
Cadillac	Catera	11.185	18.225	Passenger	31.01	3	200	107.4	70.3	194.8	3.77	18	22	9/28/2011	83.48309
Cadillac	Escalade	14.785		Car	46.225	5.7	255	117.5	77	201.2	5.572	30	15	4/17/2012	109.5091

Figure 1: Source datasets 1

	Make	Model	Variant	Ex-Showrc	Displacem	Cylinders	Valves_Pe	Drivetrain	Cylinder_(	Emission_I	Engine_Lo	Fuel_Syste	Fuel_Tank	Fuel_Type	Height	Length	Width	Body_Typ	Doors	City_Milea	Highway_
0	Tata	Nano Gen	Xt	Rs. 2,92,66	624 cc	2	2	RWD (Rea	In-line	BS IV	Rear, Tran	Injection	24 litres	Petrol	1652 mm	3164 mm	1750 mm	Hatchback	5	?23.6 km/litre	
1	Tata	Nano Gen	Xe	Rs. 2,36,44	624 cc	2	2	RWD (Rea	In-line	BS IV	Rear, Tran	Injection	24 litres	Petrol	1652 mm	3164 mm	1750 mm	Hatchback	5	?23.6 km/litre	
2	Tata	Nano Gen	Emax Xm	Rs. 2,96,66	624 cc	2	2	RWD (Rea	In-line	BS IV	Rear, Tran	Injection	15 litres	CNG	1652 mm	3164 mm	1750 mm	Hatchback	4		
3	Tata	Nano Gen	Xta	Rs. 3,34,76	624 cc	2	2	RWD (Rea	In-line	BS IV	Rear, Tran	Injection	24 litres	Petrol	1652 mm	3164 mm	1750 mm	Hatchback	5	?23.6 km/litre	
4	Tata	Nano Gen	Xm	Rs. 2,72,22	624 cc	2	2	RWD (Rea	In-line	BS IV	Rear, Tran	Injection	24 litres	Petrol	1652 mm	3164 mm	1750 mm	Hatchback	5	?23.6 km/litre	
5	Tata	Nano Gen	Xma	Rs. 3,14,81	624 cc	2	2	RWD (Rea	In-line	BS IV	Rear, Tran	Injection	24 litres	Petrol	1652 mm	3164 mm	1750 mm	Hatchback	5	?23.6 km/litre	
6	Datsun	Redi-Go	D	Rs. 2,79,65	799 cc	3	4	FWD (Fron	In-line	BS IV	Front, Trar	Injection	28 litres	Petrol	1541 mm	3429 mm	1560 mm	Hatchback	5	21.38 km/	24 km/litr
7	Datsun	Redi-Go	T	Rs. 3,51,85	799 cc	3	4	FWD (Fron	In-line	BS IV	Front, Trar	Injection	28 litres	Petrol	1541 mm	3429 mm	1560 mm	Hatchback	5	21.38 km/	24 km/litr

Figure 2: Source datasets 2

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0

Figure 3: Source datasets 3

0	0	Tata	Nano Genx	Xt	Rs. 2,92,667	624 cc	2.0	2.0	RWD (Rear Wheel Drive)	In-line	...
1	1	Tata	Nano Genx	Xe	Rs. 2,36,447	624 cc	2.0	2.0	RWD (Rear Wheel Drive)	In-line	...
2	2	Tata	Nano Genx	Emax Xm	Rs. 2,96,661	624 cc	2.0	2.0	RWD (Rear Wheel Drive)	In-line	...
3	3	Tata	Nano Genx	Xta	Rs. 3,34,768	624 cc	2.0	2.0	RWD (Rear Wheel Drive)	In-line	...
4	4	Tata	Nano Genx	Xm	Rs. 2,72,223	624 cc	2.0	2.0	RWD (Rear Wheel Drive)	In-line	...

Figure 4: Source datasets 4

	Distance(km)	Fuel Type	Location	Manufacturing Year	Price in INR	Make	Model
0	48200.0	Diesel	NaN	2016.0	3200000.0	Audi	Q3 35 TDI Technology
1	55310.0	Diesel	Mumbai	2013.0	245000.0	Audi	Q3 2.0 TDI quattro
2	30120.0	Diesel	Nadia	2013.0	255000.0	Audi	Q3 2.0 TDI quattro
3	98000.0	Diesel	New Delhi	2011.0	249900.0	Audi	A8 L- 2013 3.0 TDI quattro
4	42000.0	Diesel	New Delhi	2009.0	220000.0	Audi	A8 (2003 - 2010) 3.0 TDi

Figure 5: Source datasets 5

0	0	Maruti Wagon R LXI CNG	Maruti	Mumbai	2010	72000	CNG	Manual	First	998	58.16	5	26.60	1.75
1	1	Hyundai Creta 1.6 CRDi SX Option	Hyundai	Pune	2015	41000	Diesel	Manual	First	1582	126.20	5	19.67	12.50
2	2	Honda Jazz V	Honda	Chennai	2011	46000	Petrol	Manual	First	1199	88.70	5	18.20	4.50
3	3	Maruti Ertiga VDI	Maruti	Chennai	2012	87000	Diesel	Manual	First	1248	88.76	7	20.77	6.00
4	4	Audi A4 New 2.0 TDI Multitronic	Audi	Coimbatore	2013	40670	Diesel	Automatic	Second	1968	140.80	5	15.20	17.74

Figure 6: Source datasets 6

CHAPTER-4
Linear Regression

Linear Regression, a field of Statistics, is a Linear Approach for modeling the relationship between the Input and Output Variables. This is a widely used Concept of Machine Learning yet simple and effective. In this the relationship are modeled using a linear predictor function and Line of Best Fit is then generated. This line is then used to predict the output/unknown parameter.

There are two types of Linear Regression Technique:-

1. *Simple Linear Regression*
2. *Multiple Linear Regression*

In **Simple Linear Regression**, we find the relationship between a single independent variable (input) and a corresponding dependent variable (output). This can be expressed in the form of a straight line, also called line of Best Fit.
The Equation of Simple Linear regression is :

$$Y=B0+B1X+C$$

Y represents the Output/Dependent Variable.
B0 & B1 represents the intercept and slope coefficient respectively
C represents the Error term.

In **Multiple Linear Regression**, we find the relationship between 2 or more independent variables (inputs) and the corresponding dependent variable (output) as per the best fit. The independent variables can be continuous or categorical depending on the user's need.

The Equation of Mutiple Linear regression is :

$$Y=B0+B1X1 +B2X2 + B3X3+ \ldots.+C$$

Y represents the Output/Dependent Variable.
B0, B1 represents slope coefficient respectively
X1,X2 represents Predictor Variable
C represents the Error term.

CHAPTER-5
Model Design

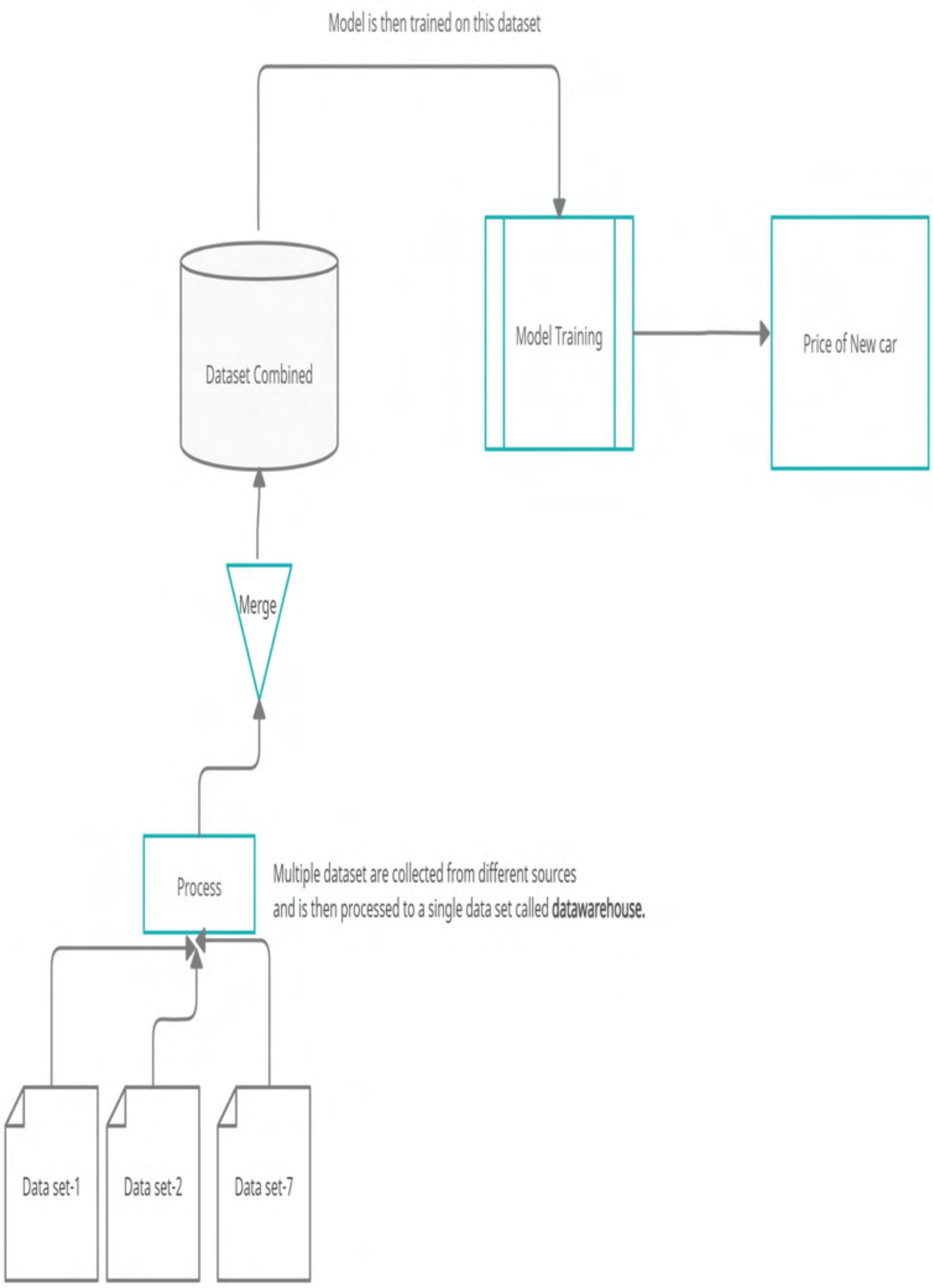

Figure 7: Model Design

CHAPTER-6
Module Description

In the Project, to train my model I am using two Regression Models, the Linear Regression model and the Lasso Regression Model, which are part of the linear model of the sklearn library. The other libraries used in the project are pandas, NumPy, Matplotlib, Seaborn, and metrics and preprocessing from sklearn. Now let's see each of these libraries one by one:-

NumPy is a Python library used for operating with multi-dimensional arrays. It additionally has capabilities for operating in the area of linear algebra, Fourier transform, and matrices.

Pandas is a software program library written for the Python programming language for statistics manipulation and analysis. In particular, it gives statistics systems and operations for manipulating numerical tables and time series.

Seaborn is a remarkable visualization library for statistical picture plotting in Python. It gives stunning default patterns and color palettes to make statistical plots more attractive. It is constructed at the pinnacle of the Matplotlib library and additionally intently included in the data structures from pandas. Seaborn targets to make visualization the principal component of exploring and know-how statistics.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Lasso
from sklearn import metrics
from sklearn import preprocessing
```

Figure 8 : Libraries Used in the Project

Matplotlib is a popular visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform statistics visualization library constructed on NumPy arrays.

Lasso stands for Least Absolute Shrinkage and Selection Operator. It is a type of Linear Regression Model that shrinks the data value towards a central point like mean.

2. Lasso Regression

```
lasso_reg_model=Lasso()
```

```
lasso_reg_model.fit(X_train,Y_train)
```

```
Lasso()
```

Model Training

```
# Prediction of Training Data
training_data_prediction=lasso_reg_model.predict(X_train)
```

```
# R square Error
error_score= metrics.r2_score(Y_train, training_data_prediction)
print("R square error ", error_score)
```

```
R square error  0.17130651180690104
```

Figure 9 : Training the Model with Lasso Regression

1. Linear Regression

```
lin_reg_model=LinearRegression()
```

```
lin_reg_model.fit(X_train,Y_train)
```

```
LinearRegression()
```

Model Evaluation

```
# Prediction of Training Data
training_data_prediction=lin_reg_model.predict(X_train)
```

```
# R square Error
error_score= metrics.r2_score(Y_train, training_data_prediction)
print("R square error ", error_score)
```

```
R square error  0.17130651180824064
```

Figure 10 : Training the Model with Linear Regression

Here in the above pictures of code, we first initialized the model into the respective variable which is named lasso_reg_model and lin_reg_model. We then fit the training input and output into the data. After the model is trained (i.e fitted) with the training data, it's time to evaluate the model. First, we evaluate the model with the training data and then with the test data. To assess our model for accuracy, we find the R^2 error of the model. In a regression model, the R2 score also called the coefficient of determination is the statistical measure of the accuracy of the model that shows how much variance in a dependent variable is explained by the independent variable(s).

Exploratory Analysis

Now coming to the data preprocessing step of our Project, as earlier mentioned that in our project we had taken datasets from multiple sources so integrating them into a single entity was one of the most painful and time-consuming steps of the project. Here, we analyzed all the different aspects that could be considered so that our model could predict with greater accuracy.

Importing all datasets

```
df1=pd.read_csv("Datasets/Car_sales.csv")
```

```
df2=pd.read_csv("Datasets/car data.csv")
```

```
df3=pd.read_csv("Datasets/CAR DETAILS FROM CAR DEKHO.csv")
```

```
df4=pd.read_csv("Datasets/Car details v3.csv")
```

```
df5=pd.read_csv("Datasets/cars_ds_final.csv")
```

```
df6=pd.read_csv("Datasets/cars_ds_final_2021.csv")
```

```
df7=pd.read_csv("Datasets/datasets3.csv")
```

```
df8=pd.read_csv("Datasets/indian-auto-mpg.csv")
```

Figure 11 : Importing the Datasets

Now Exploring each dataset

In [10]:
```
df1.head(10)
```

Out[10]:

	Manufacturer	Model	Sales_in_thousands	__year_resale_value	Vehicle_type	Price_in_thousands	Engine_size	Horsepower	Wheelbase	Width	Length	Curt
0	Acura	Integra	16.919	16.360	Passenger	21.50	1.8	140.0	101.2	67.3	172.4	
1	Acura	TL	39.384	19.875	Passenger	28.40	3.2	225.0	108.1	70.3	192.9	
2	Acura	CL	14.114	18.225	Passenger	NaN	3.2	225.0	106.9	70.6	192.0	
3	Acura	RL	8.588	29.725	Passenger	42.00	3.5	210.0	114.6	71.4	196.6	
4	Audi	A4	20.397	22.255	Passenger	23.99	1.8	150.0	102.6	68.2	178.0	
5	Audi	A6	18.780	23.555	Passenger	33.95	2.8	200.0	108.7	76.1	192.0	
6	Audi	A8	1.380	39.000	Passenger	62.00	4.2	310.0	113.0	74.0	198.2	
7	BMW	323i	19.747	NaN	Passenger	26.99	2.5	170.0	107.3	68.4	176.0	
8	BMW	328i	9.231	28.675	Passenger	33.40	2.8	193.0	107.3	68.5	176.0	
9	BMW	528i	17.527	36.125	Passenger	38.90	2.8	193.0	111.4	70.9	188.0	

In [11]:
```
df1.shape
```

Out[11]: (157, 16)

In [12]:
```
df2.head()
```

Out[12]:

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0

In [13]:
```
df2.shape
```

Out[13]: (301, 9)

Figure 12 :Exploring datasets from different sources

Out[12]:

	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0

```
In [13]: df2.shape
```

Out[13]: (301, 9)

```
In [14]: df3.head()
```

Out[14]:

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner
0	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner
1	Maruti Wagon R LXI Minor	2007	135000	50000	Petrol	Individual	Manual	First Owner
2	Hyundai Verna 1.6 SX	2012	600000	100000	Diesel	Individual	Manual	First Owner
3	Datsun RediGO T Option	2017	250000	46000	Petrol	Individual	Manual	First Owner
4	Honda Amaze VX i-DTEC	2014	450000	141000	Diesel	Individual	Manual	Second Owner

```
In [15]: df3.shape
```

Out[15]: (4340, 8)

```
In [16]: df4.head()
```

Out[16]:

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	torque	seats
0	Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4 kmpl	1248 CC	74 bhp	190Nm@ 2000rpm	5.0
1	Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owner	21.14 kmpl	1498 CC	103.52 bhp	250Nm@ 1500-2500rpm	5.0
2	Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7 kmpl	1497 CC	78 bhp	12.7@ 2,700(kgm@ rpm)	5.0
3	Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner	23.0 kmpl	1396 CC	90 bhp	22.4 kgm at 1750-2750rpm	5.0
4	Maruti Swift VXI BSIII	2007	130000	120000	Petrol	Individual	Manual	First Owner	16.1 kmpl	1298 CC	88.2 bhp	11.5@ 4,500(kgm@ rpm)	5.0

Figure 13 : Exploring datasets from different sources

After a great deal of time, we concluded that on these parameters, we could train our model so that our model could generate the expected results. These attributes are the name of the car, year of manufacturing, Fuel type it uses such as petrol, diesel, etc, what is the Market Price of the car, who is Manufacturer of the car, what is the average mileage of the car, and what is the transmission mode- automatic or manual.

Attributes selected for the project are:-

Car_name

Year of Manufacture

Fuel Type

Price

Manufacturer

Mileage

Transmission

Figure 14 : Selected Attributes for our model

So, after jotting down the parameters around which our model would be built, we moved forward with the preprocessing of the data.

Dataset Preprocessing

Dataset preprocessing is the process of cleaning up the data so that it can be given to our model and the model trains on it without any hindrance. In this step, we adapted many different techniques to get the desired data. Among them, was to segregate the year of manufacture from the date and month of manufacture as shown below. We did this by splitting the column using the str. split() function and splitting it whenever we found the '/' symbol. Similarly, we split many different columns to extract the meaningful data from it.

```
df1split = df1["Latest_Launch"].str.split("/", n = 2, expand = True)
```

```
df1split.head()
```

	0	1	2
0	2	2	2012
1	6	3	2011
2	1	4	2012
3	3	10	2011
4	10	8	2011

Figure 15 : Splitting the Year columns

```
df3split = df3["name"].str.split(" ", n = 1, expand = True)
```

```
df3split[1].unique()
```

```
array(['800 AC', 'Wagon R LXI Minor', 'Verna 1.6 SX', ...,
       'Verito 1.5 D6 BSIII', 'Innova 2.5 VX (Diesel) 8 Seater BS IV',
       'i20 Magna 1.4 CRDi'], dtype=object)
```

```
l=[]
for i in df3split[1].unique():
    if i.split()[0]=="Wagon":
        i="".join(i.split()[0:2])
        l.append(i)
    else:
        i="".join(i.split()[0])
        l.append(i)
```

Figure 16 : Splitting the Car Name column

Another analysis part was to drop the columns which was not an essential part of our prediction so that we get what is required by the model.

```
df1.drop(columns =["Latest_Launch","Vehicle_type","Engine_size","Horsepower","Wheelbase","Width","Length","Curb_weight","Fuel_ca
```

Figure 17 : Deleting the columns which are not required

```
df2.drop(columns =["Selling_Price","Kms_Driven","Seller_Type","Owner"], axis=1,inplace = True)
```

Figure 18 : Deleting the columns which are not required

Next, we renamed the columns to a common name so that there might not be any mismatch of columns during concatenating all the data frames into a single data frame. The common names of all the columns are – Name, Manufacturer, Price, Mileage, Year, Fuel, Transmission.

	Manufacturer	Name	Price	Mileage	Year	Fuel	Transmission
0	Acura	Integra	215000.0	28.0	2012	NaN	NaN
1	Acura	TL	284000.0	25.0	2011	NaN	NaN
3	Acura	RL	420000.0	22.0	2011	NaN	NaN
4	Audi	A4	239900.0	27.0	2011	NaN	NaN
5	Audi	A6	339500.0	22.0	2011	NaN	NaN

Figure 19 : Sample Final Dataset made

```
df7.rename(columns= {'Make':'Manufacturer','Manufacturing Year':'Year','Price in INR':'Price','Fuel Type':'Fuel'},inplace=True)
```

Figure 20 : Renaming the df7 column

```
df6.rename(columns= {'Make':'Manufacturer','Fuel_Type':'Fuel','Ex-Showroom_Price':'Price','Model':'Name'},inplace=True)
```

Figure 21 : Renaming the df6 column

```
df4.drop(columns =["name","selling_price","km_driven","seller_type","owner","engine","max_power","torque","seats"],inplace = True
```

```
df4.rename(columns= {'year':'Year','fuel':'Fuel','transmission':'Transmission','mileage':'Mileage'},inplace=True)
```

```
df4.head(1000)
```

Figure 22 : Operations on df4

Now inline is one of the most important steps to solve for the categorical value, which is called Encoding the Categorical values. In this step, we convert all the categorical values to the corresponding numerical values. For example in the code snippet below, I have encoded the Fuel categorical values, according to different values, such as petrol is coded as 0, diesel is encoded as 1, so this value is substituted wherever these names are present. There are two methods to perform this operation, first by the manual method as seen in figure 19, other is my using LabelEncoder() function of preprocessing from sklearn as in figure 20.

Encoding the Categorical Data

```
# Encoding the Fuel Column
```

```
df.replace({'Fuel':{'petrol':0, 'diesel':1,'cng':2,'hybrid':3,'electric':4}},inplace=True)
```

```
df.head()
```

	Manufacturer	Name	Fuel	Price	Year
0	Audi	Q3	1	3200000	2016
1	Audi	Q3	1	2450000	2013
2	Audi	Q3	1	2550000	2013
3	Audi	A8	1	2499000	2011
4	Audi	A8	1	2200000	2009

```
label_encoder = preprocessing.LabelEncoder()
df['Manufacturer']= label_encoder.fit_transform(df['Manufacturer'])
```

Figure 23 : Encoding Categorical value by Manual Method

```
label_encoder = preprocessing.LabelEncoder()
df['Manufacturer']= label_encoder.fit_transform(df['Manufacturer'])
```

```
df['Manufacturer'].unique()
```

```
array([ 1,  2,  4,  5,  6,  7,  8,  9, 10, 11, 13, 14, 15, 17, 18, 19, 20,
       21, 22, 24, 25, 26, 27, 29, 30, 32, 33, 34, 35, 36, 23, 28, 31,  0,
       12,  3, 16])
```

Figure 24 : Encoding Categorical value by LabelEncoder() function.

Data Set Split

This is the last step before the model is trained, it is called so because in this step we split the dataset into two halves- the training and testing dataset. We have kept the testing data 20% of the original dataset which is 23237 rows. This Train dataset is used to train the model and the test part is used to check the accuracy of the model.

Splitting data into Train and test data

```
X=df.drop(['Price'],axis=1)
Y=df["Price"]
```

```
X_train,X_test,Y_train,Y_test= train_test_split(X,Y,test_size=0.2, random_state=2)
```

Figure 25 : Train- Test Split.

CHAPTER-7
Results

The model predicted quite well in the case of Linear Regression the training data with an R^2-score *0.17130651180824064* while in the case of test data it was *0.17836986214106343*. The Graph of the prediction and the Z-Score is as follows:

```
# R square Error
error_score= metrics.r2_score(Y_train, training_data_prediction)
print("R square error ", error_score)
```

```
R square error  0.17130651180824064
```

```
# Visualising the actual price and predicted prices
```

```
plt.scatter(Y_train, training_data_prediction)
plt.xlabel("Actual Price")
plt.ylabel("Predicted Price")
plt.title("Actual Price vs Predicted Prices")
plt.show()
```

Figure 26 : Result of Linear Regression on Training Data

```
error_score= metrics.r2_score(Y_test, testing_data_prediction)
print("R square error ", error_score)
```

```
R square error  0.17836986214106343
```

```
plt.scatter(Y_test, testing_data_prediction)
plt.xlabel("Actual Price")
plt.ylabel("Predicted Price")
plt.title("Actual Price vs Predicted Prices")
plt.show()
```

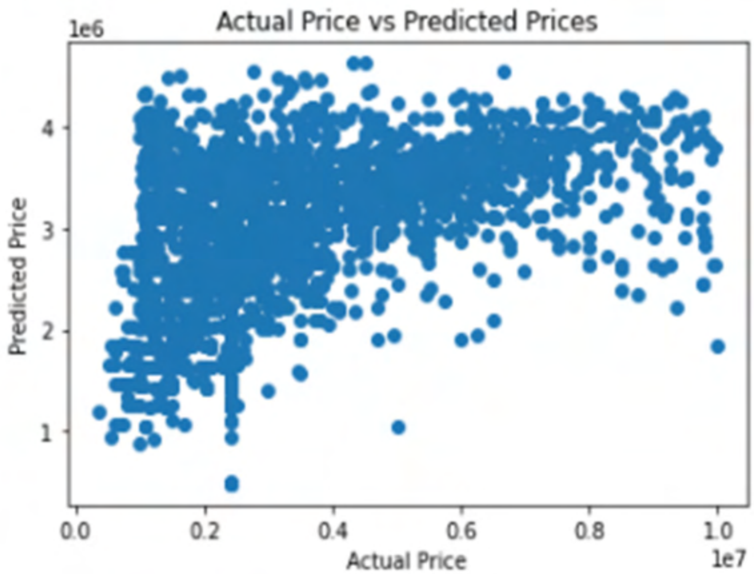

Figure 27 : Result of Linear Regression on Testing Data

In the case of Lasso Regression, the R^2-score is *0.17130651180690104* on training data and *0.17836978200409426* on the test data which is quite good.

```
# R square Error
error_score= metrics.r2_score(Y_train, training_data_prediction)
print("R square error ", error_score)
```

```
R square error  0.17130651180690104
```

```
plt.scatter(Y_train, training_data_prediction)
plt.xlabel("Actual Price")
plt.ylabel("Predicted Price")
plt.title("Actual Price vs Predicted Prices")
plt.show()
```

Figure 28 : Result of Lasso Regression on Training Data

```
error_score= metrics.r2_score(Y_test, testing_data_prediction)
print("R square error ", error_score)
```

```
R square error  0.17836978200409426
```

```
plt.scatter(Y_test, testing_data_prediction)
plt.xlabel("Actual Price")
plt.ylabel("Predicted Price")
plt.title("Actual Price vs Predicted Prices")
plt.show()
```

Figure 29 : Result of Lasso Regression on Testing Data

CHAPTER-8
Conclusion

It was an exciting Project to work on, through this project I was able to understand what it takes for a data scientist of a car manufacturing firm to be able to analysis the market for a good car to be released as per the demands. Although poor dataset availability in this field, I was able to extract the present dataset of second hand cars and use it to predict the market value of the new car.

Though it was all from my side, but there are many scope of improvements in this project. If in future a more reliable and variety of dataset is available to us, we would also analyze different aspects like colour of car, state of sale of car, type of car, top speed of car, total sales of car and many more.

Also this project can be integrated to front end section to create a user friendly interface, where users can find the right car for themselves or for the firm owners where they can see if the car which they are planning to release would be a boom or crash in the market. The model in future will become more accurate as more datasets will be provided to it. The model can also be used to learn itself from the queries of the customer as well as can conduct the important analysis of the market demand which would a major boost for the manufacturing firm.

Through this Project, I was able to learn many new concepts which is commonly used in the field of Data Analysis. I was able to implement all of the features on my own and it is first major success in my career as a Data Scientist.

References

1) Jian-DaWuJun-ChingLiu published title “Development of a predictive system for car fuel consumption using an artificial neural network”

2) Enis Gegic, Becir Isakovic, Dino Kečo, Zerina Mašetić, Jasmin Kevrić published in TEM Journal with title “Car Price Prediction using Machine Learning Techniques”

3) Fons Wijnhoven, Olivia Plant published in ICIS 2017 Proceeding with title “Sentiment Analysis and Google Trends Data for Predicting Car Sales”

4) A. Khodayari; A. Ghaffari; M. Nouri; S. Salehinia; F. Alimardani published in IEEE with title “Model Predictive Control system design for car-following behavior in real traffic flow”

5) Mariana Listiani in Master Thesis Proceeding with title “Support Vector Regression Analysis for Price Prediction in a Car Leasing Application”

6) M. J. H. Mogridge published in Journal of Transport Economics and Policy Proceeding with title “The Prediction of Car Ownership”

7) ErkanGunpinara,Umut CanCoskuna, MustafaOzsipahia, SerkanGunpinarb published their article tiltled ” A Generative Design and Drag Coefficient Prediction System for Sedan Car Side Silhouettes based on Computational Fluid Dynamics”

PREDICTIVE SYSTEM ON THE CAR MARKET TREND USING AI & ML

Ansh Shankar[1], *Vishal Kaushik* [2], *Pushpa Singh*[3], *Amit Shukla*[4] , Dhruv Varshney[5]
anshshankar@gmail.com[1], vkaushik@ddn.upes.ac.in [2],
pushpa.singh@galgotiasuniversity.edu.in[3], amitshukla@galgotiasuniversity.edu.in [4], dhruvvarshney999@gmail.com [5]
[1,3,4,5] *School of Computing Science and Engineering, Galgotias University, India*
[2] *University of Petroleum and Energy Studies, India*

Abstract:

Automobile Manufacturing is one of the most sophisticated sets of processes in the World. And for the automobile to be Successful in the market, it requires an extensive amount of work to be done in the field of Market Analysis. This is an area of major concern for companies. So, we decided to use a Machine Learning model that would predict based on the history of the Cars Manufactured in India. We analyzed the trend of Market demand and build a predictive model that would predict that whether a car would be successful in the market or not. Many other factors could be predicted using the analysis such as which car colour and type of car should the manufacturer build to maximize sales. There is also a "User Section", wherein the customer can check if the car is value for money, or wait for a more suitable car to be available in the market, which can also predict the price of a new car based on the brand name and other features.

Here we have used a Linear Regression technique to build a Machine Learning model. This model is trained on multiple datasets collected from different sources which are then analyzed and processed to obtain desired results. Thus this paper is a very handy tool for both manufacturers and customers. Since this business of car manufacturing is unceasing, hence the process of data generation is also never-ending. This paper's accuracy will get better with time as more data is available.

Keywords— PREDICTIVE SYSTEM, CAR MARKET, Machine Learning,

1. Introduction

Automobile Manufacturing is one of the most painful process for the car Manufactures in terms of both labour and financial. But if automobile is not well received in the market, the pain and the treasure is lost. For example, for production of new car the company has to setup whole production unit along with its lifeline i.e. Workers. But what if its sales are low or what if the car doesn't match the standards of the population for which the car is manufactured, the whole ecosystem and the hard work goes to vain.

So we decided to come up with a model that would be a boon to this type situation. We have collected automobiles dataset of different companies from different sources, analysed and processed it to obtain the information, upon which our machine learning model is trained. The model is based on Linear Regression technique. The manufacturer could extract from it which type of car to be manufactured, what would be the most prominent colour, what would be transmission type, fuel type, etc.

Apart this there is a "Customer Section", wherein customer can check if the car present in the market is value for money or not. or if any other car is due in the market as per the history of the automobiles manufactured.

The Whole Paper is divided into the 5 parts :-

- Data Collection
- Data understanding and exploration
- Data cleaning
- Data preparation
- Model building and evaluation

2. Literature survey

A predictive system for car fuel consumption using a back-propagation neural network has three components: an information gathering system, a fuel consumption forecasting algorithm, and a performance assessment system. Although there are many factors that influence a car's fuel consumption in a practical drive procedure, the impact factors for fuel consumption in the current system are simply determined as the make of the car, engine style, vehicle type, weight of the car, and transmission system type. An artificial neural network with back-propagation neural network has a learning capability for automobile fuel consumption prediction to test the effect of the proposed predictive system in the fuel consumption forecasting. The results of the prediction showed that the proposed neural network system is successful in predicting fuel consumption and that its performance is adequate [1].

A high-interest study topic has been automobile price prediction, as it necessitates significant effort and expertise on the part of the field expert. For a trustworthy and accurate forecast, a large number of unique attributes are considered. They utilised three machine learning approaches to create a model for forecasting the price of secondhand automobiles in Bosnia and Herzegovina (Artificial Neural Network, Support Vector Machine and Random Forest). However, the approaches suggested were used in a group setting [2].

Their study looks at how sentiment analysis and Google trends data may be used to anticipate automobile sales. Previous study has proven the utility of both approaches for sales forecasting, but the findings of current research for forecasting the sales of high-involvement items such as vehicles are more equivocal. In this study, linear regression models are used to evaluate over 500,000 social media postings for eleven automobile models on the Dutch market. In addition, the results of this study are compared to the prediction capacity of Google Trends. The findings reveal that while social media emotions have limited predictive value when it comes to automobile sales, Google Trends data and social mention volume show substantial results and may be combined into a useful prediction model. The automotive industry may utilise decision tree regression to build a prediction model with temporal delays that can be employed in addition to standard forecasting approaches [3].

The MPC predicts the future behaviour of the leader vehicle (LV) based on the relative distance and relative acceleration of each instant, and the acceleration of the follower vehicle (FV) is regulated based on this behaviour. The MPC aims to keep the relative distance within a safe range by controlling the acceleration. The outcome of the system is compared to the behaviour of real drivers with comparable beginning conditions to evaluate the performance of the developed controller. The simulation findings demonstrate that the MPC controller behaves considerably more safely than real drivers and can give passengers with a pleasant ride [4].

Because in order to stay successful in a competitive market, a leasing firm must provide a competitive lease price. It is important to forecast the future price of a used automobile in order to establish the correct pricing. The lease price might be set to meet the car's value degradation if the depreciation is known. Multiple linear regression analysis is a frequently used method for price prediction. However, there are several elements that influence the pricing, making this critical duty difficult. For high-dimensional data, the conventional regression technique may not be appropriate. Support Vector Regression, a contemporary data mining approach that is independent of input dimension, will be used to solve this possible problem. The accuracy of the predictions will next be compared to the statistical regression model. In specifically, using principles from the field of evolutionary search, a fully automated technique for adjusting and implementing SVR is created. The entire machine learning experiment is based on real-world data from a major German automobile manufacturer [5].

The Prediction of Car Ownership" discussed that for long-term forecasting of automobile ownership, an econometric technique is developed. It is compared to other well-known techniques. It is based on estimates of the percentage of family income spent on automobile purchases, as well as an analysis of car pricing and stock. The technique provides a reasonable approximation of previous levels of automobile ownership in the United Kingdom, as well as a prediction that is comparable to Tanner's for the next 15 years. However, Mogridge forecasts a higher saturation level and, as a result, a larger eventual automobile population in the future [6].

3. Design & Implementation

3.1 Data Set

A data set (or data set) is a collection of data, usually presented and each row represents the corresponding record. The dataset can contain data for one or more items, depending on the number of rows. Each entity is called DATUM.

For example, in our dataset we had Manufacturer name, Model, Vehicle type, Sale number, Fuel type as our columns, which basically represents our variable's detail. and each record/data is accommodated in the single row.

For this paper, we had employed datasets from many different sources and many different authors, analyzed it, processed it into a single dataset upon which our model is trained.

Manufactu	Model	Sales_in_t	__year_re	Vehicle_ty	Price_in_tl	Engine_siz	Horsepow	Wheelbas	Width	Length	Curb_weig	Fuel_capa	Fuel_effici	Latest_La	Power_perf_factor
Acura	Integra	16.919	16.36	Passenger	21.5	1.8	140	101.2	67.3	172.4	2.639	13.2	28	########	58.28015
Acura	TL	39.384	19.875	Passenger	28.4	3.2	225	108.1	70.3	192.9	3.517	17.2	25	########	91.37078
Acura	CL	14.114	18.225	Passenger		3.2	225	106.9	70.6	192	3.47	17.2	26	########	
Acura	RL	8.588	29.725	Passenger	42	3.5	210	114.6	71.4	196.6	3.85	18	22	########	91.38978
Audi	A4	20.397	22.255	Passenger	23.99	1.8	150	102.6	68.2	178	2.998	16.4	27	########	62.77764
Audi	A6	18.78	23.555	Passenger	33.95	2.8	200	108.7	76.1	192	3.561	18.5	22	########	84.56511
Audi	A8	1.38	39	Passenger	62	4.2	310	113	74	198.2	3.902	23.7	21	2/27/2012	134.6569
BMW	323i	19.747		Passenger	26.99	2.5	170	107.3	68.4	176	3.179	16.6	26	6/28/2011	71.19121
BMW	328i	9.231	28.675	Passenger	33.4	2.8	193	107.3	68.5	176	3.197	16.6	24	1/29/2012	81.87707
BMW	528i	17.527	36.125	Passenger	38.9	2.8	193	111.4	70.9	188	3.472	18.5	25	########	83.99872
Buick	Century	91.561	12.475	Passenger	21.975	3.1	175	109	72.7	194.6	3.368	17.5	25	########	71.18145
Buick	Regal	39.35	13.74	Passenger	25.3	3.8	240	109	72.7	196.2	3.543	17.5	23	########	95.6367
Buick	Park Avenu	27.851	20.19	Passenger	31.965	3.8	205	113.8	74.7	206.8	3.778	18.5	24	3/23/2012	85.82841
Buick	LeSabre	83.257	13.36	Passenger	27.885	3.8	205	112.2	73.5	200	3.591	17.5	25	7/23/2011	84.25453
Cadillac	DeVille	63.729	22.525	Passenger	39.895	4.6	275	115.3	74.5	207.2	3.978	18.5	22	2/23/2012	113.8546
Cadillac	Seville	15.943	27.1	Passenger	44.475	4.6	275	112.2	75	201		18.5	22	4/29/2011	115.6214
Cadillac	Eldorado	6.536	25.725	Passenger	39.665	4.6	275	108	75.5	200.6	3.843	19	22	11/27/201	113.7659
Cadillac	Catera	11.185	18.225	Passenger	31.01	3	200	107.4	70.3	194.8	3.77	18	22	9/28/2011	83.48309
Cadillac	Escalade	14.785		Car	46.225	5.7	255	117.5	77	201.2	5.572	30	15	4/17/2012	109.5091

Figure 1: Source datasets 1

	Make	Model	Variant	Ex-Showro	Displacem	Cylinders	Valves_Pe	Drivetrain	Cylinder_C	Emission_	Engine_Lo	Fuel_Syste	Fuel_Tank	Fuel_Type	Height	Length	Width	Body_Typ	Doors	City_Miles	Highway_
0	Tata	Nano Gen	Xt	Rs. 2,92,6	624 cc	2	2	RWD (Rea	In-line	BS IV	Rear, Tran	injection	24 litres	Petrol	1652 mm	3164 mm	1750 mm	Hatchback	5	23.6 km/litre	
1	Tata	Nano Gen	Xe	Rs. 2,36,4	624 cc	2	2	RWD (Rea	In-line	BS IV	Rear, Tran	injection	24 litres	Petrol	1652 mm	3164 mm	1750 mm	Hatchback	5	23.6 km/litre	
2	Tata	Nano Gen	Emax Xm	Rs. 2,96,6	624 cc	2	2	RWD (Rea	In-line	BS IV	Rear, Tran	injection	15 litres	CNG	1652 mm	3164 mm	1750 mm	Hatchback	4		
3	Tata	Nano Gen	Xta	Rs. 3,34,7	624 cc	2	2	RWD (Rea	In-line	BS IV	Rear, Tran	injection	24 litres	Petrol	1652 mm	3164 mm	1750 mm	Hatchback	5	23.6 km/litre	
4	Tata	Nano Gen	Xm	Rs. 2,72,2	624 cc	2	2	RWD (Rea	In-line	BS IV	Rear, Tran	injection	24 litres	Petrol	1652 mm	3164 mm	1750 mm	Hatchback	5	23.6 km/litre	
5	Tata	Nano Gen	Xma	Rs. 3,14,8	624 cc	2	2	RWD (Rea	In-line	BS IV	Rear, Tran	injection	24 litres	Petrol	1652 mm	3164 mm	1750 mm	Hatchback	5	23.6 km/litre	
6	Datsun	Redi-Go	D	Rs. 2,79,6	799 cc	3	4	FWD (Fron	In-line	BS IV	Front, Tran	injection	28 litres	Petrol	1541 mm	3429 mm	1560 mm	Hatchback	5	21.38 km/	24 km/lit
7	Datsun	Redi-Go	T	Rs. 3,51,8	799 cc	3	4	FWD (Fron	In-line	BS IV	Front, Tran	injection	28 litres	Petrol	1541 mm	3429 mm	1560 mm	Hatchback	5	21.38 km/	24 km/lit

Figure 2: Source datasets 2

3.2 Linear Regression

Linear Regression, a field of Statistics, is a Linear Approach for modeling the relationship between the Input and Output Variables. This is a widely used Concept of Machine Learning yet simple and effective. In this the relationship are modeled using a linear predictor function and Line of Best Fit is then generated. This line is then used to predict the output/unknown parameter.

There are two types of Linear Regression Technique:-

1. Simple Linear Regression
2. Multiple Linear Regression

In Simple Linear Regression, we find the relationship between a single independent variable (input) and a corresponding dependent variable (output). This can be expressed in the form of a straight line, also called line of Best Fit.

$$Y=B0+B1X+C$$

Y represents the Output/Dependent Variable.
B0 & B1 represents the intercept and slope coefficient respectively
C represents the Error term.

In Multiple Linear Regression, we find the relationship between 2 or more independent variables (inputs) and the corresponding dependent variable (output) as per the best fit. The independent variables can be continuous or categorical depending on the user's need.

The Equation of Mutiple Linear regression is :

$$Y=B0+B1X1 +B2X2 + B3X3+ \ldots.+C$$

Y represents the Output/Dependent Variable.
B0, B1 represents slope coefficient respectively
X1,X2 represents Predictor Variable
C represents the Error term.

3.3 Model Design

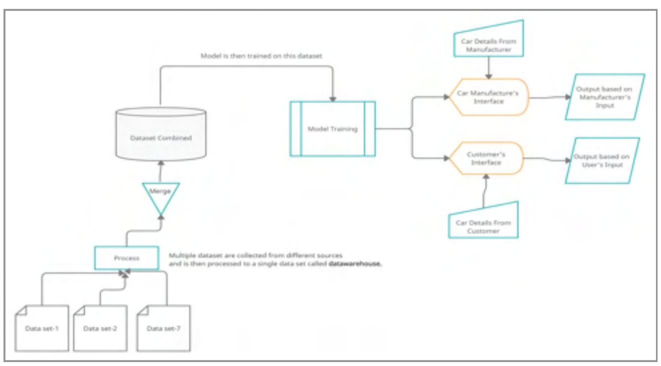

Figure 3: Model Design

4. Module Description

In the paper, to train my model I am using two Regression Models, the Linear Regression model and the Lasso Regression Model, which are part of the linear model of the sklearn library. The other libraries used in the paper are pandas, NumPy, Matplotlib, Seaborn, and metrics and preprocessing from sklearn. Now let's see each of these libraries one by one:-

NumPy is a Python library used for operating with multi-dimensional arrays. It additionally has capabilities for operating in the area of linear algebra, Fourier transform, and matrices.

Pandas is a software program library written for the Python programming language for statistics manipulation and analysis. In particular, it gives statistics systems and operations for manipulating numerical tables and time series.

Seaborn is a remarkable visualization library for statistical color palettes to make statistical plots more attractive. It is constructed at the pinnacle of the Matplotlib library and additionally intently included in the data structures from pandas. Seaborn targets to make visualization the principal component of exploring and know-how statistics.

Matplotlib is a popular visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform statistics visualization library constructed on NumPy arrays.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Lasso
from sklearn import metrics
from sklearn import preprocessing
```

Figure 4 : Libraries Used in the Paper

Lasso stands for Least Absolute Shrinkage and Selection Operator. It is a type of Linear Regression Model that shrinks the data value towards a central point like mean.

2. Lasso Regression

```
lasso_reg_model=Lasso()
```

```
lasso_reg_model.fit(X_train,Y_train)
```

```
Lasso()
```

Model Training

```
# Prediction of Training Data
training_data_prediction=lasso_reg_model.predict(X_train)
```

```
# R square Error
error_score= metrics.r2_score(Y_train, training_data_prediction)
print("R square error ", error_score)
```

```
R square error  0.17130651180690104
```

Figure 5 : Training the Model with Lasso Regression

1. Linear Regression

```
lin_reg_model=LinearRegression()
```

```
lin_reg_model.fit(X_train,Y_train)
```

```
LinearRegression()
```

Model Evaluation

```
# Prediction of Training Data
training_data_prediction=lin_reg_model.predict(X_train)
```

```
# R square Error
error_score= metrics.r2_score(Y_train, training_data_prediction)
print("R square error ", error_score)
```

```
R square error  0.17130651180824064
```

Figure 6 : Training the Model with Linear Regression

Here in the above pictures of code, we first initialized the model into the respective variable which is named lasso_reg_model and lin_reg_model. We then fit the training input and output into the data. After the model is trained (i.e fitted) with the training data, it's time to evaluate the model. First, we evaluate the model with the training data and then with the test data. To assess our model for accuracy, we find the R^2 error of the model. In a regression model, the R2 score also called the coefficient of determination how much variance in a dependent variable is explained by the independent variable(s).

Exploratory Analysis

Now coming to the data preprocessing step of our Paper, as earlier mentioned that in our paper we had taken datasets from multiple sources so integrating them into a single entity was one of the most painful and time-consuming steps of the paper. Here, we analyzed all the different aspects that could be considered so that our model could predict with greater accuracy.

Importing all datasets

```
df1=pd.read_csv("Datasets/Car_sales.csv")
df2=pd.read_csv("Datasets/car data.csv")
df3=pd.read_csv("Datasets/CAR DETAILS FROM CAR DEKHO.csv")
df4=pd.read_csv("Datasets/Car details v3.csv")
df5=pd.read_csv("Datasets/cars_ds_final.csv")
df6=pd.read_csv("Datasets/cars_ds_final_2021.csv")
df7=pd.read_csv("Datasets/datasets3.csv")
df8=pd.read_csv("Datasets/indian-auto-mpg.csv")
```

Figure 7 : Importing the Datasets

5. Dataset Preprocessing

Dataset preprocessing is the process of cleaning up the data so that it can be given to our model and the model trains on it without any hindrance. In this step, we adapted many different techniques to get the desired data. Among them, was to segregate the year of manufacture from the date and month of manufacture as shown below. We did this by splitting the column using the str. split() function and splitting it whenever we found the '/' symbol. Similarly, we split many different columns to extract the meaningful data from it.

	Manufacturer	Name	Price	Mileage	Year	Fuel	Transmission
0	Acura	Integra	215000.0	28.0	2012	NaN	NaN
1	Acura	TL	284000.0	25.0	2011	NaN	NaN
3	Acura	RL	420000.0	22.0	2011	NaN	NaN
4	Audi	A4	239900.0	27.0	2011	NaN	NaN
5	Audi	A6	339500.0	22.0	2011	NaN	NaN

Figure 8 : Sample Final Dataset made

Data Set Split

This is the last step before the model is trained, it is called so because in this step we split the dataset into two halves- the training and testing dataset. We have kept the testing data 20% of the original dataset which is 23237 rows. This Train dataset is used to train the model and the test part is used to check the accuracy of the model.

Splitting data into Train and test data

```
X=df.drop(['Price'],axis=1)
Y=df["Price"]
```

```
X_train,X_test,Y_train,Y_test= train_test_split(X,Y,test_size=0.2, random_state=2)
```

Figure 9 : Train- Test Split.

6. Results

The model predicted quite well in the case of Linear Regression the training data with an R^2-score *0.17130651180824064* while in the case of test data it was *0.17836986214106343*. The Graph of the prediction and the Z-Score is as follows:

```
# R square Error
error_score= metrics.r2_score(Y_train, training_data_prediction)
print("R square error ", error_score)
```

```
R square error  0.17130651180824064
```

```
# Visualising the actual price and predicted prices
```

```
plt.scatter(Y_train, training_data_prediction)
plt.xlabel("Actual Price")
plt.ylabel("Predicted Price")
plt.title("Actual Price vs Predicted Prices")
plt.show()
```

Figure 10 : Result of Linear Regression on Training Data

```
error_score= metrics.r2_score(Y_test, testing_data_prediction)
print("R square error ", error_score)
```

```
R square error  0.17836986214106343
```

```
plt.scatter(Y_test, testing_data_prediction)
plt.xlabel("Actual Price")
plt.ylabel("Predicted Price")
plt.title("Actual Price vs Predicted Prices")
plt.show()
```

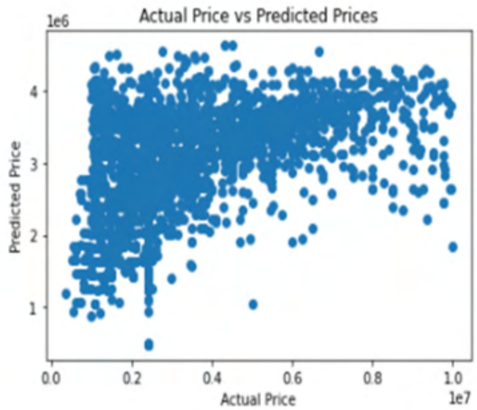

Figure 11 : Result of Linear Regression on Testing Data

In the case of Lasso Regression, the R2-score is 0.17130651180690104 on training data and 0.17836978200409426 on the test data which is quite good.

```
# R square Error
error_score= metrics.r2_score(Y_train, training_data_prediction)
print("R square error ", error_score)
```

```
R square error  0.17130651180690104
```

```
plt.scatter(Y_train, training_data_prediction)
plt.xlabel("Actual Price")
plt.ylabel("Predicted Price")
plt.title("Actual Price vs Predicted Prices")
plt.show()
```

Figure 12 : Result of Lasso Regression on Training Data

```
error_score= metrics.r2_score(Y_test, testing_data_prediction)
print("R square error ", error_score)
```

```
R square error  0.17836978200409426
```

```
plt.scatter(Y_test, testing_data_prediction)
plt.xlabel("Actual Price")
plt.ylabel("Predicted Price")
plt.title("Actual Price vs Predicted Prices")
plt.show()
```

Figure 13 : Result of Lasso Regression on Testing Data

7. Conclusion

In this paper I was able to understand what it takes for a data scientist of a car manufacturing firm to be able to analysis the market for a good car to be released as per the demands. Although poor dataset availability in this field, I was able to extract the present dataset of second hand cars and use it to predict the market value of the new car.

Though it was all from my side, but there are many scope of improvements in this paper. If in future a more reliable and variety of dataset is available to us, we would also analyze different aspects like colour of car, state of sale of car, type of car, top speed of car, total sales of car and many more.

Also this paper can be integrated to front end section to create a user friendly interface, where users can find the right car for themselves or for the firm owners where they can see if the car which they are planning to release would be a boom or crash in the market. The model in future will become more accurate as more datasets will be provided to it. The model can also be used to learn itself from the queries of the customer as well as can conduct the important analysis of the market demand which would a major boost for the manufacturing firm.

Through this Paper, I was able to learn many new concepts which is commonly used in the field of Data Analysis. I was able to implement all of the features on my own and it is first major success in my career as a Data Scientist.

References

1. Jian-DaWuJun-ChingLiu published title “Development of a predictive system for car fuel consumption using an artificial neural network”

2. Enis Gegic, Becir Isakovic, Dino Kečo, Zerina Mašetić, Jasmin Kevrić published in TEM Journal with title “Car Price Prediction using Machine Learning Techniques”

3. Fons Wijnhoven, Olivia Plant published in ICIS 2017 Proceeding with title “Sentiment Analysis and Google Trends Data for Predicting Car Sales”

4. Khodayari; A. Ghaffari; M. Nouri; S. Salehinia; F. Alimardani published in IEEE with title “Model Predictive Control system design for car-following behavior in real traffic flow”

5. Mariana Listiani in Master Thesis Proceeding with title “Support Vector Regression Analysis for Price Prediction in a Car Leasing Application”

6. M. J. H. Mogridge published in Journal of Transport Economics and Policy Proceeding with title “The Prediction of Car Ownership”

7. M. Dehghani, R. M. Dangelico, “Smart wearable technologies: Current status and market orientation through a patent analysis”, Proc. IEEE Int. Conf. on Industrial Technology: 1570-1575, 2017.

8. L. F. Lai, J. F. Jiang, H. Y. Wei and K. S. Hsu, “Depth sensor used in vehicle-related patent analysis” Int. Conf. IEEE-ICAMSE, 2016.

9. Z. Jing, “Patent Analysis on New Energy Auto Industry” Int. Conf. on Computer Science and Information Processing, 2012.

10. H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok., Interpreting TF–IDF Term Weights as Making Relevance Decisions. ACM Transactions on Information Systems 26(3):1-37, 2008.

11. A. Buja, D. Swayne, M. Littman, N. Dean, H. Hofmann, and L. Chen, “Data Visualization with Multidimensional Scaling” Journal ofComputational and Graphical Statistics17(2):444-472, 2008.

Gmail **Ansh Shankar <anshshankar@gmail.com>**

Acceptance:ICIRSMT2021

1 message

ICIRSMT2021 <icirsmt2021@easychair.org> 9 December 2021 at 21:12
To: Ansh Shankar <anshshankar@gmail.com>

Dear Ansh Shankar,

Paper ID: T035, Track : TECHNOLOGY

Congratulations!
On behalf of the review committee and conference organizing committee, I am happy to inform you that your Paper/Abstract entitled

"PREDICTIVE SYSTEM ON THE CAR MARKET TREND USING AI & ML "

has been accepted for presentation at the 2021 ABVU-AIMT: International Conference on Innovative Research in Science, Management and Technology (ICIRSMT 2021) to be held in hybrid mode during December 27-28, 2021 organized by Department of Computer Science and Application, Atal Bihari Vajpayee University, Bilaspur (C.G.), India in association with American Institute of Management and Technology (AIMT), USA
You are requested to do the following:
1. Pay your registration fee in any one of the following mode
i. Online: https://icirsmt2021.gkfusa.org/registration/ (Only if you have international VISA card)
ii. UPI payment: 9926709807(Phone pay) For Indian authors only.
Fee detail are as follow:

Category	INR	USD
Faculty/Industry	1500	20
Research Student	1200	16

2. Pay registration fee and send detail of payment at using following link https://forms.gle/Cxq1FqNEJtdHb2W17.

Please note the following information regarding the registration in conference and publication:
1. Your paper presented in conference will be published Compulsorily in Conference proceedings : AIMTCP (Visit https://gkfusa.org/aimtcp/ for more detail) and may also be published in any one of the journal/Edited book in subjected to review and scope of the journal/edited book after the conference. You need to submit your paper as per the desired template soon after the conference, We will let you know the decision of reviewer]decision:
i. CRC press edited Book: Taylor and Francis Groups- Publication within 3 to 6 months
ii. Scopus Indexed Journal: Publication within a month (In process ,not confirmed).
iii. UGC care listed journal- Publication within a month (In process , not confirmed).
iv. International journal of Decision Science and Information Technology (IJDSIT): A journal of Atal Bihari Vajpayee University, Bilaspur, India in association with Modern Technology and Management Institutions, Inc. , USA – Publication within a month.
v. Global Journal of Review and Business Technology (GRBT): Published by The Global Knowledge Foundation, Inc. USA (Web site www.gkfusa.org).
vi. Global Journal of Modeling and Intelligent Computing (GJMIC): Published by The Global Knowledge Foundation, Inc. USA (Web site www.gkfusa.org).
vii. Global Journal of Computer and Engineering Technology (GJCET): Published by The Global Knowledge Foundation, Inc. USA (Web site www.gkfusa.org).
viii. Advanced Computing Techniques: Informatics and Emerging Technologies-Volume I, Edited Book by Bentham Publication (https://benthambooks.com/book/9789814998451/).

Special Note
Completing the registration and paying the registration fee is not the guarantee of the paper publication. The registration fee is for only to present the paper(s) in ICIRSMT 2021.The papers will be considered for publications as per the respective journal/Edited book policies. The paper publication will depend on the reviewer's reports and editor's decision during the journal review process. However abstract/Full length paper will positively publish in conference proceedings (AIMTCP) with ISSN No.

2. At least one author of an accepted paper must register and pay the registration fee to be included in the program and the conference publication.
3. If more than one authors are going to register for the same paper than each author has to pay registration fee separately and to register himself/herself as mentioned above.

4. An author submitted more than one paper has to pay full registration fee for the first paper and 50% registration fee for remaining papers.
5. The schedule of the conference will be available after 20th December 2021 on the conference website(http://icirsmt2021.gkfusa.org/) and will be sent to you in your E-mail ID.
6. Please note that your paper presentation should be of 10 minutes including question answer session.

Thank you for your support and we look forward to the pleasure of meeting either virtually or offline and interacting with you during the conference.

Feel free to contact us at icirsmt2021@gmail.com or through mobile No. 9425222658 in case of any queries. Also refer your Paper ID for further communication.

Keep in touch with conference website http://icirsmt2021.gkfusa.org/

Thanks and Regards
Programme Chair
ICIRSMT2021

2021
ABVU-AIMT

3rd International Conference on Innovative Research In Science, Management and Technology

(ICIRSMT 2021)

Conference Website : www.icirsmt2021.gkfusa.org

27-28 December, 2021

Venue

Hybrid Mode (Online and offline)

ORGANIZED BY

Department of Computer Science and Application

Atal Bihari Vajpayee University, Bilaspur (C.G.), India

IN ASSOCIATION WITH

American Institute of Management and Technology (AIMT), USA

(A Subsidiary of The Global Knowledge Foundation, Inc, USA)

www.gkfusa.org

SUPPORTED BY

Computer Society of India (CSI), Chhattisgarh Chapter

Publication Partners

CRC Press
Taylor & Francis Group

ABVU-MTMI'S JOURNAL

BENTHAM

GKF, USA

Paper submission through Easy chair

Important Dates

Abstract/Full length paper submission deadline

~~25 November, 2021~~, 10 December 2021

Notification of Acceptance and Registration

~~08 December, 2021~~, 15 December 2021

Program Chairs

Dr. H.S Hota
Atal Bihari Vajpayee University, Bilaspur (C.G.), India

Dr. D.K. Sharma
University of Maryland, Eastern Shore, USA

Contact

E-mail : icirsmt2021@gmail.com
Mob. : +91 9425222658

Chief Patron

Acharya Arun Diwakar Nath Bajpai
Hon'ble Vice Chancellor,
Atal Bihari Vajpayee University, Bilaspur, India

Patrons

Dr. Deva Sharma,
President, The GKF, Inc. USA

Dr. Sudheer Sharma,
Registrar,
Atal Bihari Vajpayee University, Bilaspur, India

Conference Chairs

Dr. Daniel Okunbor,
Fayetteville University, USA

Dr. Madhu Jain,
Indian Institute of Technology (IIT) Roorkee, India

Dr. Aaron Rababaah,
American University of Kuwait, Kuwait

Gmail

Ansh Shankar <anshshankar@gmail.com>

ICIRSMT2021 (Kindly fill all details carefully as same will reflect in your certificate.)
1 message

Google Forms <forms-receipts-noreply@google.com> 9 December 2021 at 21:50
To: anshshankar@gmail.com

Google Forms

Thanks for filling in ICIRSMT2021 (Kindly fill all details carefully as same will reflect in your certificate.)

Here's what was received.

ICIRSMT2021 (Kindly fill all details carefully as same will reflect in your certificate.)

Email *

anshshankar@gmail.com

ICIRSMT2021


2021
ABVU-AIMT
3rd International Conference
on
Innovative Research In Science, Management and Technology
(ICIRSMT 2021)
Conference Website : www.icirsmt2021.gkfusa.org
27-28 December, 2021
Venue
Hybrid Mode (Online and offline)
Publication Partners
CRC Press
Taylor & Francis Group
ABVU-MTMI'S
JOURNAL
BENTHAM
GKF, USA
ORGANIZED BY
Department of Computer Science and Application
Atal Bihari Vajpayee University, Bilaspur (C.G.), India
IN ASSOCIATION WITH
American Institute of Management and Technology (AIMT), USA
(A Subsidiary of The Global Knowledge Foundation, Inc, USA)
www.gkfusa.org
SUPPORTED BY
Computer Society of India (CSI), Chhattisgarh Chapter
Paper submission through Easy chair
Important Dates
Abstract/Full length paper submission deadline
25 November, 2021
Notification of Acceptance and Registration
08 December, 2021
Chief Patron
Acharya Arun Diwakar Nath Bajpayee
Hon'ble Vice Chancellor,
Atal Bihari Vajpayee University, Bilaspur, India
Patrons
Dr. Deva Sharma,
President,
The GKF, Inc. USA
Dr. Sudheer Sharma,
Registrar,
Atal Bihari Vajpayee University, Bilaspur, India
Program Chairs
Dr. H.S Hota
Atal Bihari Vajpayee University, Bilaspur (C.G.), India
Dr. D.K. Sharma
University of Maryland, Eastern Shore, USA
Conference Chairs
Dr. Daniel Okunbor,
Fayetteville University, USA
Dr. Madhu Jain,
Indian Institute of Technology (IIT) Roorkee, India
Dr. Aaron Rababaah,
American University of Kuwait, Kuwait
Contact
E-mail : icirsmt2021@gmail.com
Mob. : +91 9425222658

Paper ID *

T035, Track : TECHNOLOGY

Paper Title *

PREDICTIVE SYSTEM ON THE CAR MARKET TREND USING AI & ML

Name of Registered Author *

Ansh Shankar

Institute/Organization *

Galgotias University

City *

Greater Noida

Country *

India

Contact Number *

8287504975

Preferred Mode of Presentation *

- [x] Online

Offline

Amount Paid (If you have more than one paper then fill separate google form for that) *

1200

Transaction Receipt *

Submitted files

Paper Fees Payment_Paper-id T035 - Ansh Shankar.jpeg

Create your own Google Form
Report Abuse