**AN ENGINEERING PROJECT REPORT**

On

**IMPLEMENTATION OF LINEAR REGRESSION IN CAR PRICE PREDICTION SYSTEM**

**Submitted By:**

**ATUL POKHAREL** (170308)

**NABIN RAJ DHUNGEL** (170322)

**NABIN JOSHI**  (170323)

**RAKSHYA BHETWAL** (170334)

Under the Supervision of:

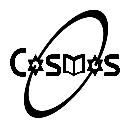
**Mr. Ankit Bhattarai**

**Submitted To:**

**The Department of IT and Computer Engineering**

**In Partial fulfillment of requirement for the degree of**

**Bachelor of Engineering in Computer**



**Cosmos College of Management & Technology**

**(Affiliated to Pokhara University)**

**Tutepani, Lalitpur, Nepal**

**Date of Submission:**

**2021/09/22**

# ACKNOWLEDGEMENT

The warm response received by this project, both from teachers and students, has encouraged the learning group to present this project within time. Every topic of this project has been written in simple and lucid language so that any one going through it will be clearly able to understand the subject. Similarly, the context has been revised thoroughly to make it more comprehensive.

We would like to express our sincere gratitude to our supervisor Mr. Ankit Bhattarai for his guidance throughout this work. Our deepest appreciation to the entire team of Cosmos College of Management And Technology, for granting us this valuable opportunity and helping us accomplish this project successfully by giving valuable feedbacks and suggestions for the improvement of the subject matter of this project Constructive criticism, suggestions and feedbacks for the improvement of this project would be highly appreciated.

**-PROJECT TEAM**

# ABSTRACT

The artificial intelligence (AI) concepts started millennia before the rise of modern computers from ancient myths to early advances in logic and mathematics. Today, modern AI has an influence on almost every area of human activity-from sectors such as healthcare, transport, business, customer care to science fiction and popular culture with current trends and studies conducted in this sector it has a long way to go and certainly has a scope in the future. Recent studies have shown that the density of adaptation of predictive features by a company has exponentially increased day by day. So our team decided to implement this concept which focuses on the implementation of a price prediction model. Thereupon we agreed to build a project which will predict the price of used cars in Nepal. A considerable number of distinct attributes are examined for reliable and accurate predictions. To build a model for predicting the price of used cars in Nepal, we applied machine learning technique Linear Regression. The price prediction model focused on the used automobile domain. The data used for the prediction was collected from the web portal kaggle.com as well as web scraped. The Technical Aspects of this project consist of heavy calculations of a huge dataset and the implementation of machine learning algorithm. Furthermore, the model was evaluated using test data and the accuracy above 75% was obtained.

**Keywords: Artificial Intelligence, Machine learning, Python, Prediction Model, Linear Regression**

**Table of Content**

[ACKNOWLEDGEMENT i](#_Toc83499754)

[ABSTRACT ii](#_Toc83499755)

[LIST OF ABBREVIATATIONS vi](#_Toc83499756)

[1. INTRODUCTION 1](#_Toc83499757)

[**1.1 Background** 1](#_Toc83499758)

[**1.2 Problem Definition** 2](#_Toc83499759)

[**1.3 Objectives** 2](#_Toc83499760)

[**1.4 Report Structure** 3](#_Toc83499761)

[2. LITERATURE REVIEW 4](#_Toc83499762)

[3. FEASIBILITY STUDY 6](#_Toc83499763)

[**3.1 Technical Feasibility** 6](#_Toc83499764)

[**3.2 Financial Feasibility** 6](#_Toc83499765)

[**3.3 Legal Feasibility** 6](#_Toc83499766)

[4. METHODOLOGY 7](#_Toc83499767)

[**4.1** **Software Development Approach** 7](#_Toc83499768)

[**4.2 System Workflow** 8](#_Toc83499769)

[**4.2.1 Creating Environment and Installing Libraries** 8](#_Toc83499770)

[**4.2.2 Data Collection** 10](#_Toc83499771)

[**4.2.3 Processing Categorical Variables** 11](#_Toc83499772)

[**4.2.4 Train – Test Split** 13](#_Toc83499773)

[**4.2.5 Linear Regression Model** 14](#_Toc83499774)

[**4.3** **Important concepts for understanding the algorithm** 16](#_Toc83499775)

[ **Cost Function** 16](#_Toc83499776)

[ **Gradient Descent** 16](#_Toc83499777)

[ **Underfiting** 16](#_Toc83499778)

[ **Overfitting** 16](#_Toc83499779)

[ **Bias** 16](#_Toc83499780)

[ **Cross-validation bias** 17](#_Toc83499781)

[5. SYSTEM DESIGN AND ARCHITECTURE 18](#_Toc83499782)

[**4.1 Block Diagram** 18](#_Toc83499783)

[**5.2 Use Case Diagram** 19](#_Toc83499784)

[**5.3 Roadmap of Overall System** 20](#_Toc83499785)

[6. RESULT AND ANALYSIS 21](#_Toc83499786)

[6.1 Achievements 21](#_Toc83499787)

[**6.1.1 Successfully imported the necessary libraries** 21](#_Toc83499788)

[**6.1.2 Successfully loaded the excel file** 21](#_Toc83499789)

[**6.1.3 Successfully implemented the data visualization techniques** 22](#_Toc83499791)

[**6.2 Result of training and predicting data** 25](#_Toc83499792)

[GANTT- CHART 26](#_Toc83499793)

[REFERENCES 27](#_Toc83499794)

[APPENDIX 28](#_Toc83499795)

**List of Figures**

[Figure 1: Software Development Approach 7](file:///E:\Machine%20Learning\report\final%20report.docx#_Toc83499796)

[Figure 2: Workflow of Study 8](#_Toc83499797)

[Figure 3: Sample of collected Data 10](#_Toc83499798)

[Figure 4: Equation of the Model 14](file:///E:\Machine%20Learning\report\final%20report.docx#_Toc83499799)

[Figure 5: Prediction of small dataset for simple linear regression 15](#_Toc83499800)

[Figure 6: System Architecture 18](file:///E:\Machine%20Learning\report\final%20report.docx#_Toc83499801)

[Figure 7: UML Use Case Diagram 19](#_Toc83499802)

[Figure 8: Roadmap used for building machine learning systems 20](#_Toc83499803)

[Figure 9: Importing necessary libraries 21](file:///E:\Machine%20Learning\report\final%20report.docx#_Toc83499804)

[Figure 10: Successfully loaded data 21](file:///E:\Machine%20Learning\report\final%20report.docx#_Toc83499805)

[Figure 11: Count of cars based on manufacturers 22](#_Toc83499806)

[Figure 12: Effect on Price due to manufacture year shown in scatter plot 23](#_Toc83499807)

[Figure 13: Relationship between price and kilometers driven shown in scatter plot 23](#_Toc83499808)

[Figure 14: Sample data taken to plot power vs mileage using line chart 24](#_Toc83499809)

[Figure 15: Result of training and predicting data 25](#_Toc83499810)

[Figure 16: Gantt Chart showing time duration of the project 26](#_Toc83499811)

# LIST OF ABBREVIATATIONS

AI: Artificial Intelligence

RMSE: Root Mean Square Error

ML: Machine Learning

MSE: Materials Science and Engineering

ODAV: Optimal Distribution of Auction Vehicles

ANN: Artificial Neural Networks

IP: Intellectual Property

# INTRODUCTION

# 

## **1.1 Background**

Used Car Price Prediction is a popular problem as determining the resale value of a car is very challenging due to several factors that drive a used vehicle’s price on market. The total number of motor vehicles plying across the country reached nearly 3.1 million among which 7,651 units of cars were registered as of mid-May 2018, per the information provided by the Department of Transport Management [1]. Similarly from 2018-2019 3.54% rise was observed and this number has been increasing with each passing month thanks to surge in number of private vehicles and vehicles employed by construction industry. The figure is five times over that of the last decade and it is likely that this trend will continue, and the numbers of car will increase in the future. However, the manufacture fixes the costs of recent cars within the industry together with some additional costs that are incurred by the govt. majorly within the type of various taxes so as to ensure the customers that buy a replacement car of the money that they invest to be righteous. But because of such increase in prices of the new cars and therefore the inability of the many customers to shop for a new car due to the shortage of adequate funds, they like used cars which has resulted into a world increase within the sales of used cars. Hence this adds an additional significance to possess a second hand car price prediction model to accurately determine the worth of the car considering a range of features.

Accurate price prediction requires knowledge about the numerous number of things which are vital in determining the worth of the used car. Attributes such as years that the car is employed, its model and build, manufacturing country, fiscal power, mileage, etc. is very prominent to determine its worth. The physical appearance of the car also majorly influences the value of the car. Similarly, the kind of fuel, acceleration, the inside style, the quantity of its cylinders (measured in cc or cubic centimeters), the braking system, its size, safety index, paint color, reputation of the automaker, customer reviews, whether it's a sports car, its physical state, transmission, cooling system, cosmic wheels, steering mechanism, GPS navigator all may influence the worth furthermore. Therefore, we can easily conclude that the worth of car depends on numerous factors. However, information about all the factors mentioned above isn't available often, thus the customer needs to make his/her decision to get based on only a few attributes provided. On that account, during this project only a low subset of the factors that are mentioned above are taken into consideration for the prediction model.

To be able to predict the worth of used cars can offer assistance to both the buyer and sellers.

**Used car sellers (dealers):** Used car sellers will be benefited from this model as the model will help them better understand the important features of a second hand car and what makes a car desirable. This information which if they will consider shall help them offer a far better price to their customers.

**Online pricing services:** The model developed during this study might be helpful for existing online web services predicting car prices to help them get better ends up in prediction. Having another price prediction system might just be handy for them to provide their users with better prices.

**Individuals:** The system might give individuals a platform to estimate the worth of any used car they need to sell or buy and come up with an accurate value of their car in accordance to its condition.

## **1.2 Problem Definition**

In this project, we focus on the second-hand car market of our country. The user of second-hand cars is increasing rapidly in our country for a few years. But most of the buyers are in a dilemma whether their decision of buying a second-hand car is right or wrong. The buyers mostly feel confused about the price, whether that recommended car is suitable for them or not and so on. This project highlights those confusions of the buyers and they can search cars of their interest by entering the name, model, kilometers run, etc. If there is any car in such a category, they can easily get it and perform the deal.

## **1.3 Objectives**

* To predict the price of second-hand cars on the basis of given attributes,
* To provide overview of available car model to the buyers.

## **1.4 Report Structure**

This paper is organized in the following manner:

Chapter 2 contains related work in the field of price prediction of used cars.

Chapter 3 includes the research methodology of our study is explain.

Chapter 4 covers the system design and architecture

Chapter 5 includes different types of feasibility studies

Chapter 6 incorporates analysis and result of the overall process

Other than the chapters, the report also comprises of:

* Gantt Chart that shows time duration taken in various phase of our project.
* References which includes different journals we had studied and sites we have taken help from.
* Appendix

# LITERATURE REVIEW

Used Car Price Prediction has been studied thoroughly in various researches.

**Listian** discussed, in her paper written for Master thesis [2], that regression model that was built using Support Vector Machines (SVM) can predict the price of a car that has been leased with better precision than multivariate regression or some simple multiple regression. The weakness of this research is that a change of simple regression with more advanced SVM regression was not shown in basic indicators like mean, variance or standard deviation.

**Richardson** in his thesis work [3]. He applied multiple regression analysis and demonstrated that hybrid cars retain their value for longer time than traditional cars. This has roots in environmental concerns about the climate and it gives higher fuel efficiency.

**Wu et al.** [4] conducted car price prediction study, by using neuro-fuzzy knowledge-based system. They took into consideration the following attributes: brand, year of production and type of engine. Their prediction model produced similar results as the simple regression model. Moreover, they made an expert system named ODAV (Optimal Distribution of Auction Vehicles) as there is a high demand for selling the cars at the end of the leasing year by car dealers. This system gives insights into the best prices for vehicles, as well as the location where the best price can be gained. Regression model based on k-nearest neighbor machine learning algorithm was used to predict the price of a car. This system has a tendency to be exceptionally successful since more than two million vehicles were exchanged through it [5].

**Gonggie** [6] proposed a model that is built using ANN (Artificial Neural Networks) for the price prediction of a used car. He considered several attributes: miles passed, estimated car life and brand. The proposed model was built so it could deal with nonlinear relations in data which was not the case with previous models that were utilizing the simple linear regression techniques. The non-linear model was able to predict prices of cars with better precision than other linear models.

Furthermore, **Pudaruth** [7] applied various machine learning algorithms, namely: k-nearest neighbors, multiple linear regression analysis, decision trees and naïve bayes for car price prediction in Mauritius. The dataset used to create a prediction model was collected manually from local newspapers in period less than one month, as time can have a noticeable impact on price of the car. He studied the following attributes: brand, model, cubic capacity, mileage in kilometers, production year, exterior color, transmission type and price. However, the author found out that Naive Bayes and Decision Tree were unable to predict and classify numeric values. Additionally, limited number of dataset instances could not give high classification performances, i.e. accuracies less than 70%.

**Noor and Jan** [8] build a model for car price prediction by using multiple linear regression. The dataset was created during the two-months period and included the following features: price, cubic capacity, exterior color, date when the ad was posted, number of ad views, power steering, mileage in kilometer, rims type, type of transmission, engine type, city, registered city, model, version, make and model year. After applying feature selection, the authors considered only engine type, price, model year and model as input features. With the given setup authors were able to achieve prediction accuracy of 98%.

# 3. FEASIBILITY STUDY

## **3.1 Technical Feasibility**

Machine Learning techniques as Linear Regression, Random Forest are the backbone for this project kind of projects. Various libraries as Pandas, Numpy, Matplotlib, etc are used in this project.

## **3.2 Financial Feasibility**

The estimated costs to develop this project would have generally included resource cost and advertising cost. Since we developed this project on a web-based platform, expense has been reduced to none.

## **3.3 Legal Feasibility**

From developing this project to the distribution of this project, we have addressed all the legal requirements. All the data required for this project was taken only by permission. We have considered protecting our Intellectual Property (IP) rights and minimizing the risk of infringing the IP rights of others. This project has been developed without breaching any law in the country.

# METHODOLOGY

## **Software Development Approach**

As the development process includes constant improvement with the code as well as Dataset we will be using an incremental model for software development. In this model, we will design the system and test the system then implementation will be done incrementally until the final product is built.

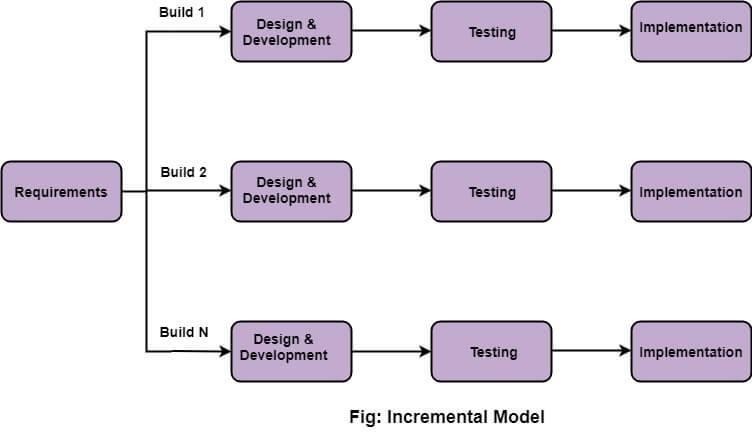


Figure 1: Software Development Approach

(Incremental Model)

## **4.2 System Workflow**



Figure 2: Workflow of Study

### **4.2.1 Creating Environment and Installing Libraries**

For the development of this project, we used different software tools. Those are:

* **PyCharm**

The platform to be used for the development process is PyCharm. We have selected the IDE for the development of the project. The main reason to select PyCharm over the anaconda jupyter notebook is that we will implement the principle of separation of concerns in our project that is the program will be broken down into pieces which will act as a chain reaction upon execution. The separation of the code will be very much efficient to address the errors and to easily access the code when needed.

* **The Jupyter Notebook**

The Jupyter Notebook is an open-source web application that was used for data cleaning and transformation, statistical modeling, data visualization, machine learning, and many more.

* **Excel**

Excel is a phenomenal tool for data analysis. The visualization of the data in the initial phase will be done by using excel. The find, replace and filter features provided by excel is very much helpful to understand the data and is easy to perform data manipulation techniques. We have used excel in the initial phase of data analysis to perform cleaning of the dataset.

* **Libraries are:**
* **Pandas**

Pandas is an open-source, library that provides high-performance, and easy-to-use data structures and data analysis tools for the Python programming language. It is used to read the csv file as a data frame and to cast the string value of a column to integer value.

* **NumPy**

NumPy can be referred to as the fundamental package for scientific computing for the medium of Python. It is imported to support and manipulate the large data with multidimensional arrays and matrices.

* **Matplotlib**

Matplotlib is a visualization library in Python for 2D plots of arrays. It is a multi- platform data visualization library built on NumPy arrays. It is used for data analysis and for embedding plots into applications.

* **Sklearn**

Sklearn is used for implementing various features such as classification, regression and clustering algorithms including support vector machines for implementing prediction algorithm. It is used to split the data frame into training and testing data sample and to implement the linear regression algorithm. The accuracy and the value of root mean square error is also calculated with the help of sklearn library.

* **Flask**

It is a tool that we used to deploy our model.

### **4.2.2 Data Collection**

Since manual collection of data is a time consuming task, data was collected from the web portal of kaggle.com. The attributes taken into consideration were: company, bought year, price, kilometers driven, fuel type, etc. The collected dataset contained 863 samples.

Since 863 samples were very less, we web scraped to increase the dataset to 6020 samples. After the raw data was collected further processing was done.

Further modification of the data was done as the data with null entries were removed to maintain homogeneity of the dataset as this could have affected our prediction model.

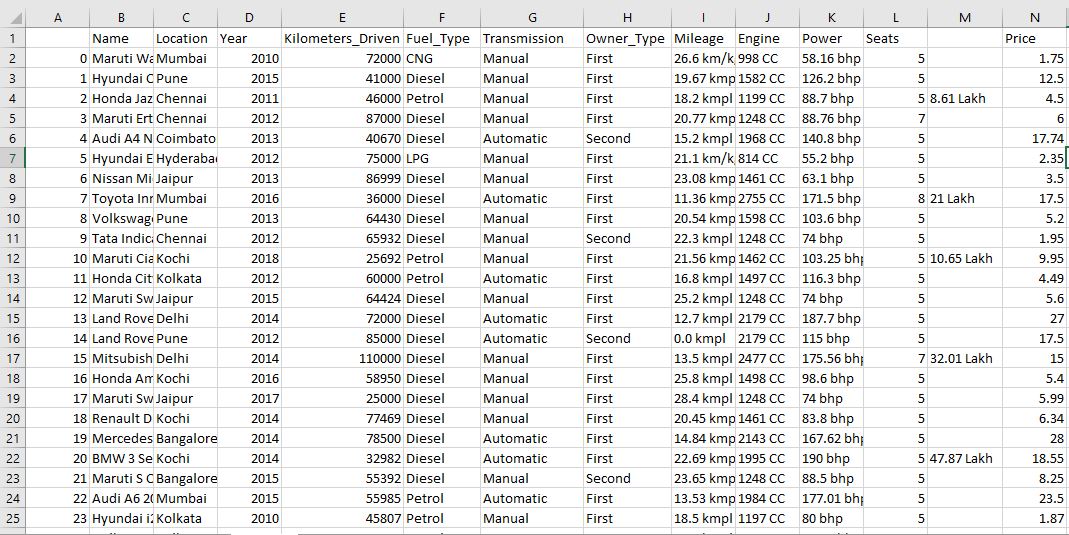


Figure 3: Sample of collected Data

### **4.2.3 Processing Categorical Variables**

The libraries and modules were then employed to read and visualize the data for the better understanding of our datasets. We used One-Hot data encoding technique, column transformer and pipeline to pre-process the data and feed it to the machine. Pipelining is a very handy technique when we are not sure about the pre-processing steps to be used in our pre-production environment.

#### **4.2.3.1 One-Hot Encoder**

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. The machine learning models perform 23 best with numerical values but it is not uncommon that real-world datasets contain one or more categorical features columns.

When we are talking about categorical data, we have to further distinguish between nominal and ordinal features. In our project we had datas such as name, location, fuel\_type and transmission which can be classified as nominal data and other datas such as owner\_type (first, second) is our ordinal data. Simply saying, nominal data do not imply any order or natural ranking whereas ordinal data defines the rank.

One-Hot Encoder was imported as:

from sklearn.preprocessing import OneHotEncoder

And implemented as:

ohe.fit(x[['name','company','fuel\_type']])

OneHotEncoder(categories = 'auto', drop = None, dtype = <class 'numpy.float64'>,

handle\_unknown='error', sparse=True)

As shown above we fed OneHotEncoder with categorical values such as name, company and fuel\_type and then created a column transformer, applied OneHotEncoder through the columns and the remainder was pass-through. The column transformer implemented OneHotEncoder to the above mentioned columns of every input data.

#### **4.2.3.2 Column Transformer**

It applies transformers to columns of an array or pandas Data Frame.

This estimator allows different columns or column subsets of the input to be transformed separately and the features generated by each transformer will be concatenated to form a single feature space. This is useful for heterogeneous or columnar data, to combine several feature extraction mechanisms or transformations into a single transformer.

It was imported as:

from sklearn.compose import make\_column\_transformer

And implemented by making object of class make\_column\_transformer as:

column\_trans = make\_column\_transformer((OneHotEncoder(categories = ohe.categories\_),

{list2}), remainder = 'passthrough')

#### **4.2.3.3 Pipeline**

The purpose of pipeline is to assemble several steps that can be cross-validated together while setting different parameters. For this, it enables setting parameters of the various steps using their names and the parameter name separated by a ‘\_\_’, as in the example below. A step’s estimator may be replaced entirely by setting the parameter with its name to another estimator, or a transformer removed by setting it to ‘passthrough’.

It was imported as:

from sklearn.pipeline import make\_pipeline

And implemented by making object of class make\_pipeline as:

pipe = make\_pipeline(column\_trans,lr)

pipe.fit(x\_train ,y\_train)

### **4.2.4 Train – Test Split**

The whole data set collected was then split into training (80%) and testing (20%) subsets. The splitting was done by picking at random which results in a balance between the training data and testing data amongst the whole dataset. This was done to avoid overfitting and enhance generalization. After the dataset was split linear regression algorithm was imported from scikit learn which is a popular 19 dependency of collection of machine learning libraries. The test and train dataset were loaded to the algorithm and the accuracy was determined from the output it generated.

* The data was manipulated using the pandas’ library.
* The splitting of the train and test data set was done by importing train\_test\_split from sklearn.model\_selection.
* The data set was loaded to the linear regression algorithm by importing the linear\_model.LinearRegression()
* The accuracy of the model was calculated by importing the r2\_score from sklearn.metrics..

### **4.2.5 Linear Regression Model**

The goal of linear regression is to model the relationship between one or multiple features and a continuous target variable. In this project we have used Multiple Linear Regression.

The equation of a multilinear regression is defined as follows:

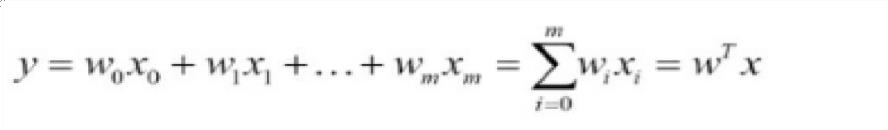


Figure 4: Equation of the Model

where,

* y is the predicted value
* w₀ is the bias term.
* w₁,…,wₙ are the model parameters
* x₁, x₂,…,xₙ are the feature values.

The mathematical representation of linear regression is a linear equation that combines a specific set of input data (in our case features such as name, company, fuel\_type) to predict the output value (in our case price) for that set of input values. To sum it up for a given predictor variable x (in our case features) and a response variable y (in our case price), the distance is minimized my fitting the values in a straight line most commonly the average squared distance between the sample points and the fitted line. The value for intercept and slope is learned from the data to predict the outcome variable of new data.

For simplicity let us take a simple example of simple linear regression plot for demonstration of its mechanism as simple linear regression and multiple linear regression follows the same logic behind the algorithm.

Here in the figure shown below we have simply plotted Price vs Year of purchase. In the plot the price is our response variable whereas the year in which the model was bought is our predicting variable. The red line in the graph below is referred to as the best fit straight line, also called the regression line.

**Y-axis: Price**

**X-axis: Year of Purchase**

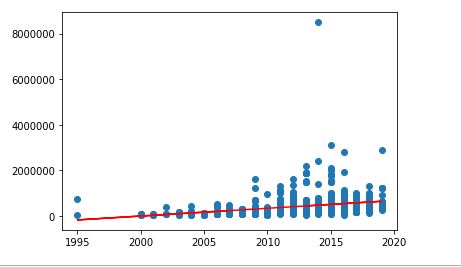


Figure 5: Prediction of small dataset for simple linear regression

The regression model is a line defined by coefficients estimated from training data. Once the coefficients are estimated, we can use them to make predictions.

## **Important concepts for understanding the algorithm**

### **Cost Function**

The cost function helps us to find out the best possible values for a\_0 and a\_1 which would provide the best fit line for the data points. Since we want the best values for a\_0 and a\_1, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted value and the actual value. [9]

### **Gradient Descent**

Gradient descent is a method of updating a\_0 and a\_1 to reduce the cost function (MSE). The idea is that we start with some values for a\_0 and a\_1 and then we change these values iteratively to reduce the cost. Gradient descent helps us on how to change the values. [9]

### **Underfiting**

If a machine learning model is not able to predict with a decent level of accuracy, then we say that the model underfits. This could be because of a variety of reasons, including that the forecast does not have the right features, or the problem statement is too complicated for the algorithm chosen for machine learning.

### **Overfitting**

Overfitting occurs when the model fits the data too well or simply put when the model is too complex. The Overfitting model learns the detail and noise in the training data to such an extent that it negatively impacts the performance of the model on new data/test data.

### **Bias**

A machine learning model is said to have a low bias if its predictability level is high. Bias plays an important role when we have to compare two machine learning algorithms for the same problem statement.

### **Cross-validation bias**

Cross-validation in machine learning is a technique that provides an accurate measure of the performance of a machine learning model. This performance will be closer to what you can expect when the model is used in a future unseen dataset.

# 5. SYSTEM DESIGN AND ARCHITECTURE

## **4.1 Block Diagram**

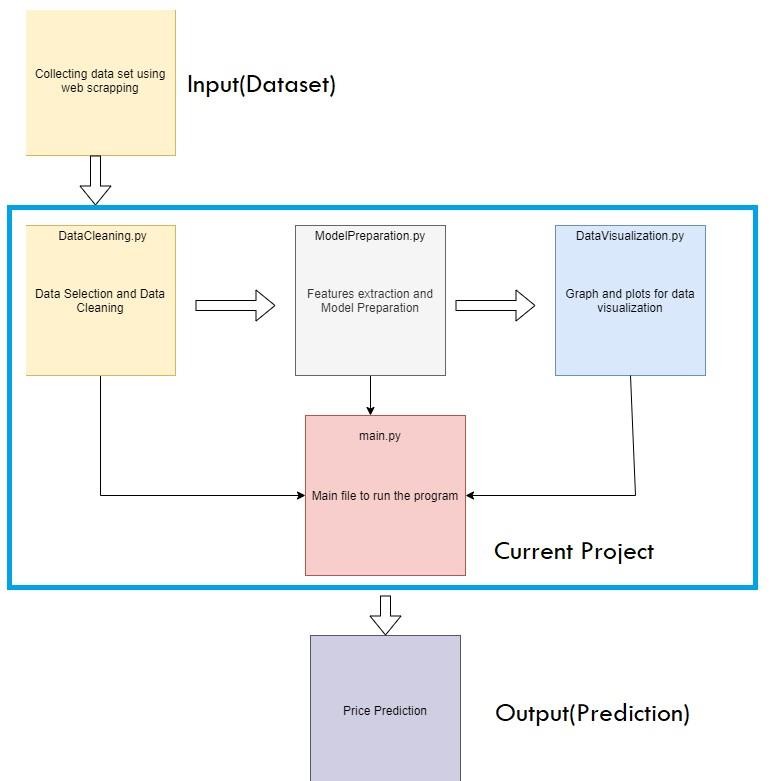


Figure 6: System Architecture

## **5.2 Use Case Diagram**

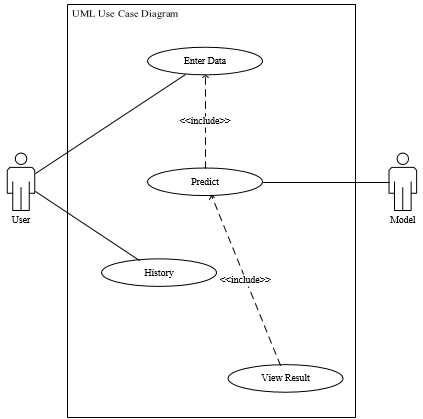


Figure 7: UML Use Case Diagram

## **5.3 Roadmap of Overall System**

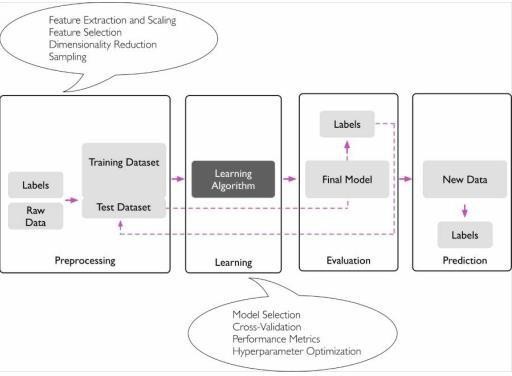


Figure 8: Roadmap used for building machine learning systems

The overall process can be generalized as:

* First, we collected the data about used cars, identified important features that reflect the price.
* Second, we preprocessed our datasets and removed entries with NAN values.
* Third, we applied a linear regression model on the preprocessed dataset with features as inputs and the price as output.

All the processes have been discussed in detail in Chapter 4.

# RESULT AND ANALYSIS

# 6.1 Achievements

## **6.1.1 Successfully imported the necessary libraries**

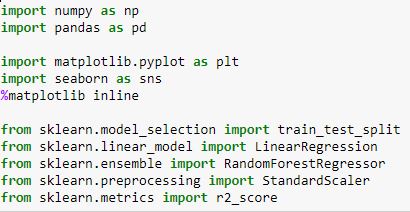


Figure 9: Importing necessary libraries

The necessary libraries required for the project is install using pip install command. The libraries were successfully installed and did not cause any errors during the development phase.

## **6.1.2 Successfully loaded the excel file**

## 

Figure 10: Successfully loaded data

The excel file with the data is successfully loaded in the python using pandas library.

## **6.1.3 Successfully implemented the data visualization techniques**

* **Bar Graph**

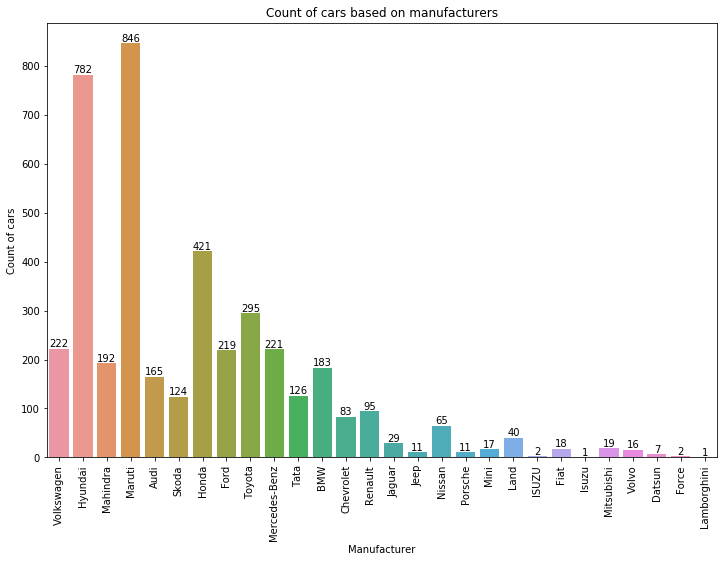


Figure 11: Count of cars based on manufacturers

The above bar graph as in Figure 11 shows the number of cars of each manufacturing brand. The different colors indicate different manufacturers.

* **Scatter Plot**



Figure 12: Effect on Price due to manufacture year shown in scatter plot

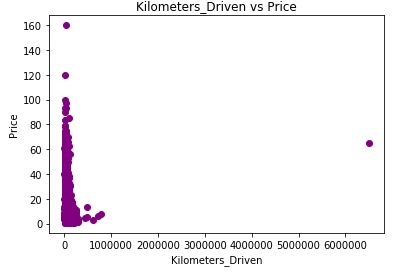


Figure 13: Relationship between price and kilometers driven shown in scatter plot

The scatter plot as in Figure 12 clearly shows that the year of manufacture and price are inversely proportional to each other. The older the car is in terms of years, the value of car depreciates rapidly.

For example: If we have two cars A and B, A is 6 months old and B is 1 years old then the price of car B is lesser in comparison to the price of car A.

Similarly the scatter plot shown in Figure 13 shows the relationship between price (in lakhs) and kilometers driven. Kilometers travelled is one of the major factor that results in price depreciation.

* **Line Chart**



Figure 14: Sample data taken to plot power vs mileage using line chart

In the above chart, we have taken sample of few data that shows relationship between mileage and power of a car. More the power of the car, lesser is its mileage. Also more the power of the car, lesser is its depreciation rate.

## **6.2 Result of training and predicting data**

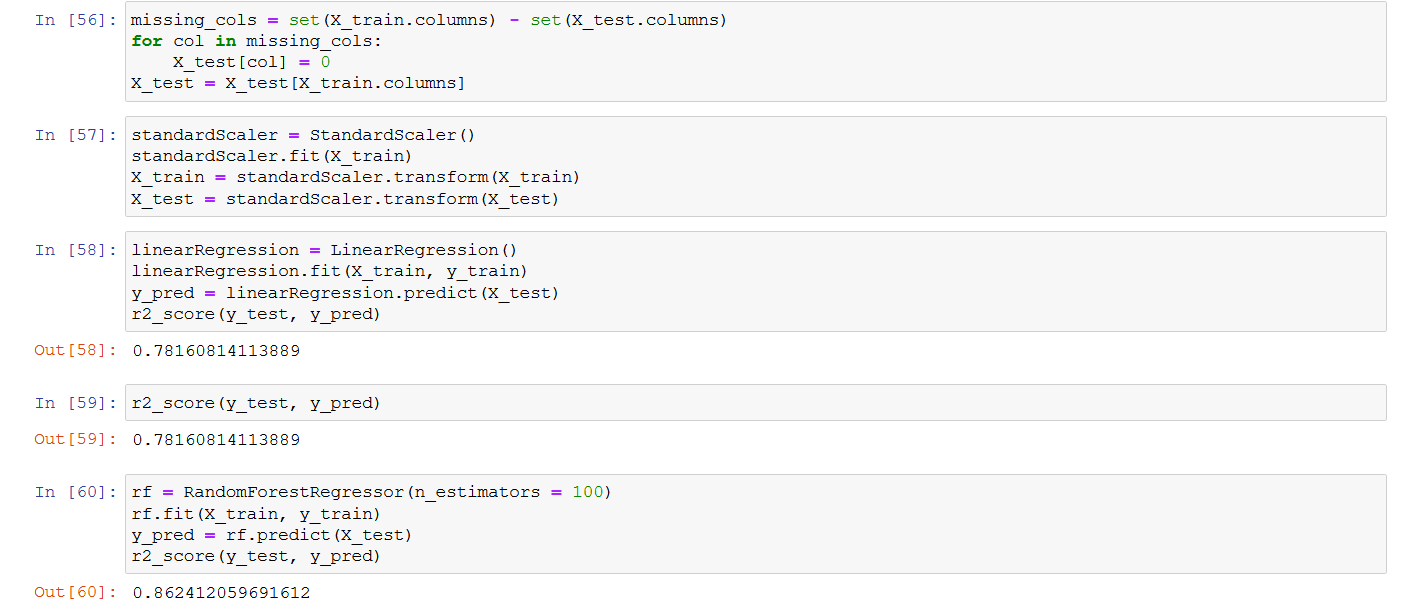


Figure 15: Result of training and predicting data

As shown in above figure, at first we split the data into train-test modules: x-train and x-test respectively. The y-pred (dependent variable) then predicted the tested data with respect to x-test (independent variable). The variance between the dependent and independent variable is given by r2 score which range varies from 0 to 100%. As shown in figure, while applying linear regression the correlation between the two variables on training was found to be 78%. It means that the regression model is mostly valid, but not in all cases.

# GANTT- CHART

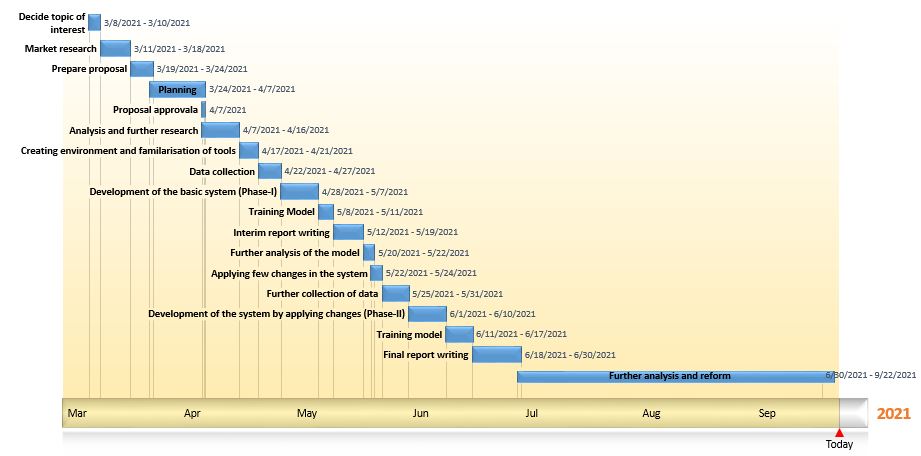


Figure 16: Gantt Chart showing time duration of the project

# REFERENCES

[1]*https://myrepublica.nagariknetwork.com/news/3-1-million-motor-vehicles-on-nepali-roads-dotm/*

[2] *Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application (Doctoral dissertation, Master thesis, TU Hamburg-Harburg).*

[3]*Richardson, M. S. (2009). Determinants of used car resale value. Retrieved from: https://digitalcc.coloradocollege.edu/islandora/object /coccc%3A1346*

[4] *Wu, J. D., Hsu, C. C., & Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications, 36(4), 7809-7817.*

[5] *Du, J., Xie, L., & Schroeder, S. (2009). Practice Prize Paper—PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation, and Genetic Algorithms to Used-Vehicle Distribution. Marketing Science, 28(4), 637-644.*

[6] *Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit. In Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on (Vol. 2, pp. 682-685). IEEE.*

[7] *Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764.*

[8] *Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications, 167(9), 27-31*

[9] *Gandhi, R., (2018). https://towardsdatascience.com/introduction-to-machine-learningalgorithms-linear- regression-14c4e325882a*

# APPENDIX

**How the solution addresses real world problems?**

As per the increase in the price of new cars and the incapability of customers to buy new cars mainly due to the lack of funds, used cars sales are on a global increase. Predicting the prices of used cars is a much-needed problem to be addressed. The proposed system Used car prediction model will help to address the real-world problems in the following ways:

* By utilization of the abundant data. There is a lot of data on the internet related to use of automobile and by implementing the prediction model, the data can be used for a better purpose.
* By delivering a new feature for the users. The prediction model after its implementation can be deployed on the web which will provide a new feature for the users.
* By studying the data and generating previously unknown insights from it. New ideas and insights can be drawn during the implementation process.
* The model can also minimize the fraud transaction to some extent as people can be alert about the deserving price of the automobile.