**AN ENGINEERING PROJECT PROGRESS 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:

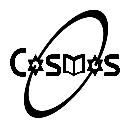
**Mr. Ankit Bhattarai**

**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:**

**2078/02/14**

# 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 has decided to implement this concept which focuses on the implementation of a price prediction model. So we decided 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. The Technical Aspects of this project consist of heavy calculations of a huge dataset and the implementation of machine learning algorithm. Furthermore, the model will evaluated using test data and the accuracy above 80% will obtained.

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

**Table of Content**

[ABSTRACT i](#_Toc73054894)

[1. INTRODUCTION 1](#_Toc73054895)

[**1.1 Background** 1](#_Toc73054896)

[**1.2 Problem Definition** 2](#_Toc73054897)

[**1.3 Objectives** 2](#_Toc73054898)

[**1.4 Report Structure** 3](#_Toc73054899)

[2. LITERATURE REVIEW 4](#_Toc73054900)

[3. SYSTEM ANALYSIS AND DESIGN 6](#_Toc73054901)

[**3.1 Methodology:** 6](#_Toc73054902)

[**3.1.1 Creating Environment and Installing Libraries** 6](#_Toc73054903)

[**3.1.2 Data Collection** 8](#_Toc73054904)

[**3.1.3 Processing Categorical Variables** 9](#_Toc73054905)

[**3.1.4 One-Hot Encoder** 9](#_Toc73054906)

[**3.1.5 Train – Test Split** 10](#_Toc73054907)

[**3.1.6 Linear Regression Model** 10](#_Toc73054908)

[4. Work Progress 12](#_Toc73054909)

[**4.1 Model Implementation** 12](#_Toc73054910)

[**4.2 Summary** 12](#_Toc73054911)

[**4.3 Snapshot** 13](#_Toc73054912)

[5. Work Volume 14](#_Toc73054913)

[6. REFERENCES 15](#_Toc73054914)

**List of Figures**

[Figure 1: Workflow of Study 6](#_Toc73054811)

[Figure 2: Sample of collected Data 8](#_Toc73054812)

[Figure 3: Implementation of OneHotEncoder 9](#_Toc73054813)

[Figure 4: Equation of the Model 10](file:///D:\REPORT.docx#_Toc73054814)

[Figure 5: Prediction of small dataset for simple linear regression 11](#_Toc73054815)

[Figure 6: Roadmap used for building machine learning systems 12](#_Toc73054816)

[Figure 7: Graphical User Interface of our Model 13](#_Toc73054817)

[Figure 8: Gantt Chart showing volume of work completed 14](#_Toc73054818)

# 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 method we used to continue our work and the work completed

Chapter 5 includes a gant chart showing the volume of work completed

Chapter 6 includes references of the journals we studied to implement our model

# 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%.

# SYSTEM ANALYSIS AND DESIGN

## **3.1 Methodology:**



Figure 1: Workflow of Study

### **3.1.1 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. I 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.

### **3.1.2 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 and fuel type.

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.

The collected dataset contained 863 samples. After the raw data was collected further processing was done.

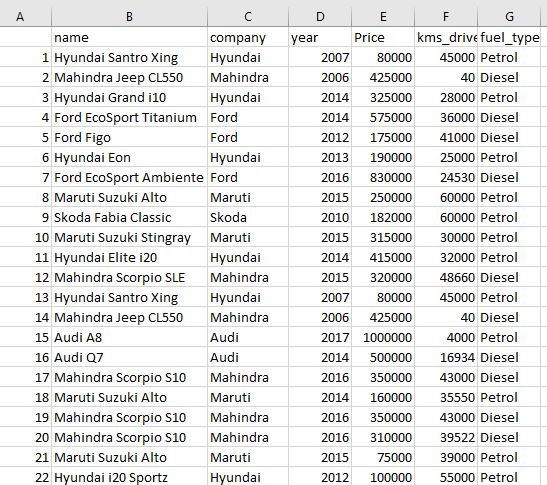


Figure 2: Sample of collected Data

### **3.1.3 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.

### **3.1.4 One-Hot Encoder**

We feed OneHotEncoder with categorical values such as name, company and fuel\_type. We then created a column transformer, applied OneHotEncoder through the columns and the remainder was passthrough. The column transformer implements OneHotEncoder to the above mentioned columns of every input data.

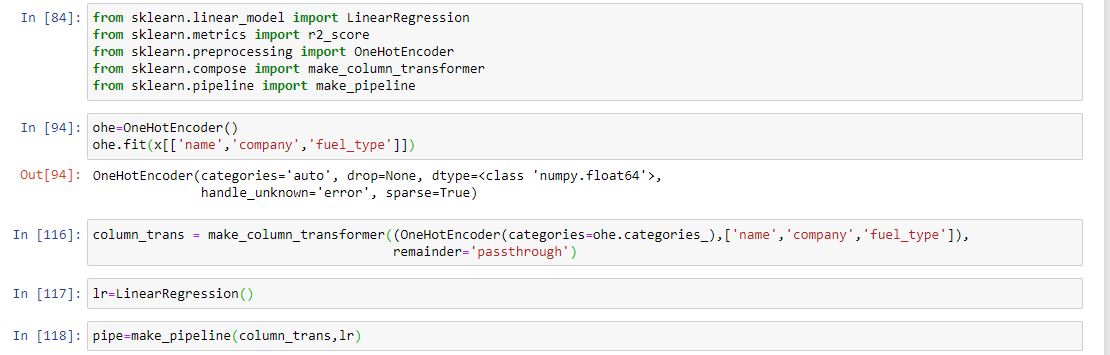


Figure 3: Implementation of OneHotEncoder

We implement pipeline and column transformer as shown in ln [118] in the above figure. Whatever data we fed to the pipeline as input, was applied with OneHotEncoder through column transformer and then the transformed columns along with the previous columns was given to the linear regression model.

### **3.1.5 Train – Test Split**

Our dataset was then split such that 80% of the dataset was used to train our model and 20% of the dataset was employed to train our model and make predictions. We employed a Linear Regression model for this study.

### **3.1.6 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:

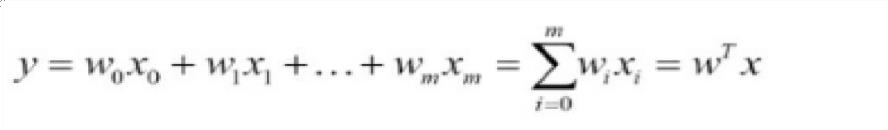


Figure 4: Equation of the Model

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**

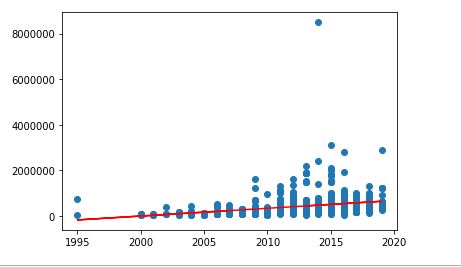


Figure 5: Prediction of small dataset for simple linear regression

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.

# Work Progress

## **4.1 Model Implementation**

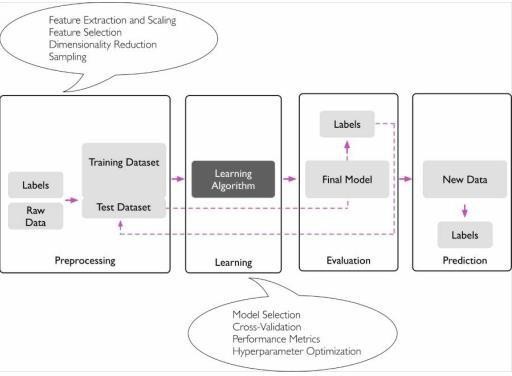


Figure 6: Roadmap used for building machine learning systems

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 have applied a linear regression model on the preprocessed dataset with features as inputs and the price as output.

## **4.2 Summary**

Car price prediction can be a challenging task due to the high number of attributes that should be considered for the accurate prediction. The major step in the prediction process is collection and preprocessing of the data which has already been fulfilled. However, our aim for the future research is testing this system to work successfully with various data sets. We will extend our test data and also validate the proposed approach.

## **4.3 Snapshot**

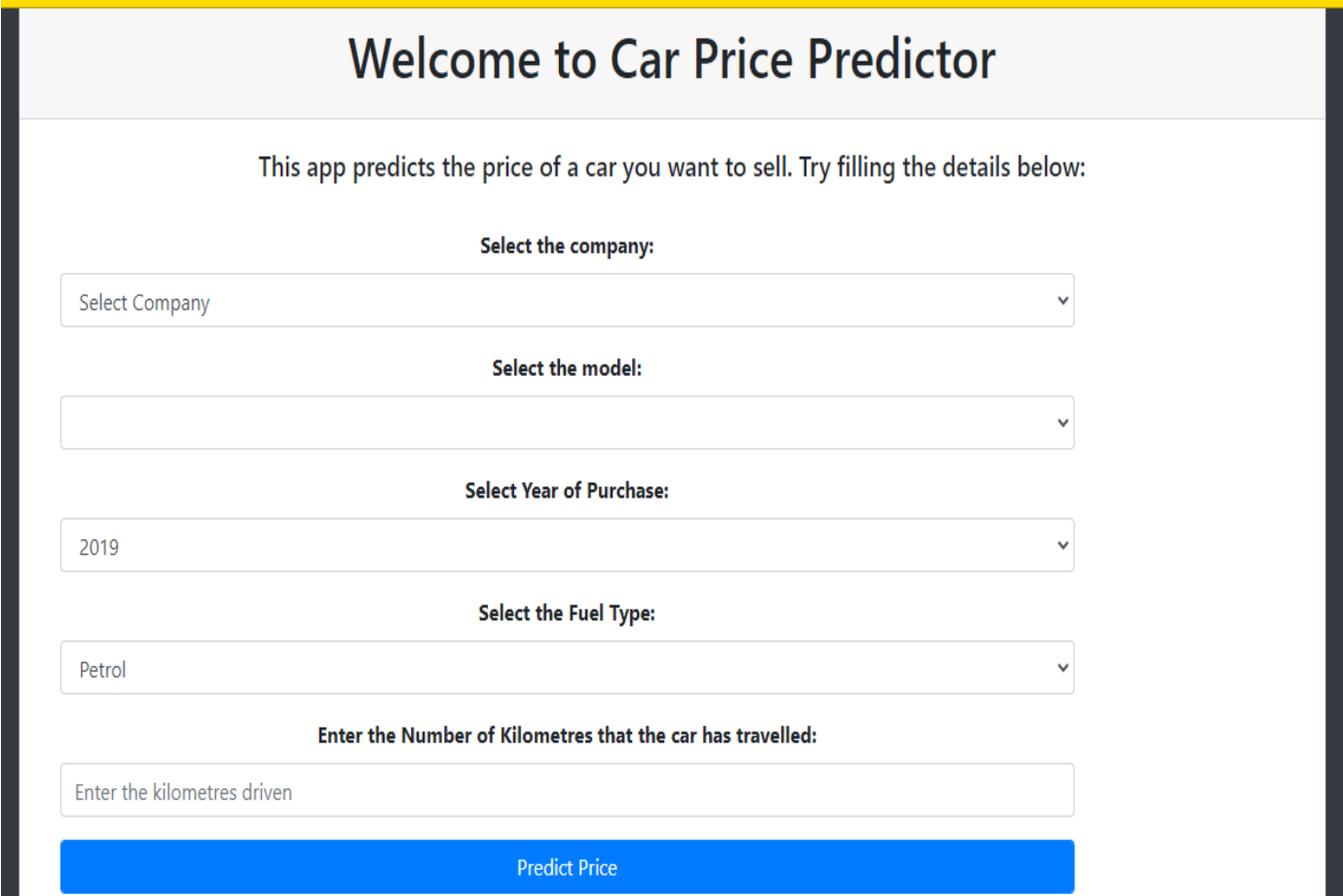


Figure 7: Graphical User Interface of our Model

# Work Volume

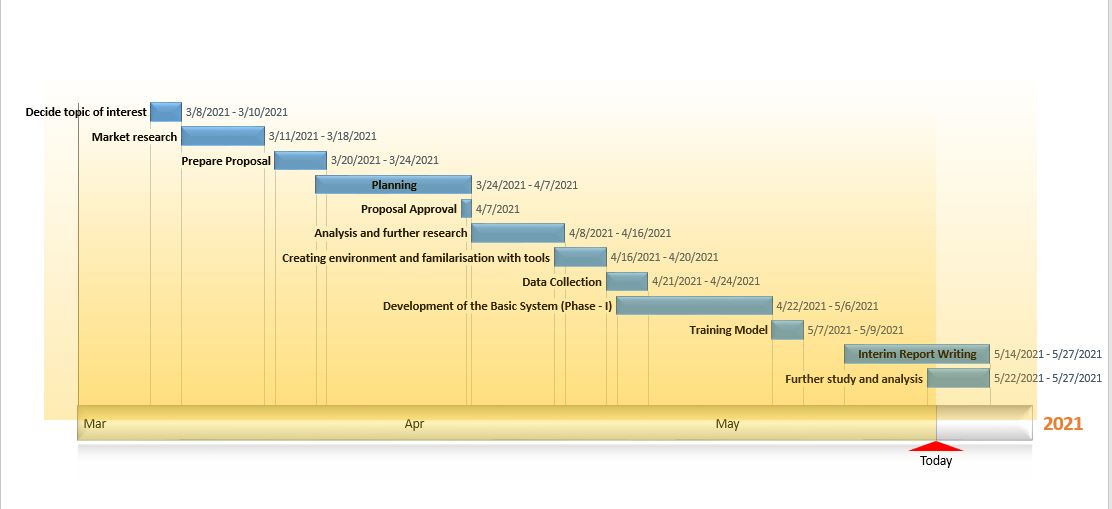


Figure 8: Gantt Chart showing volume of work completed

# 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*