**Mini Project Report on**



**CAR PRICE PREDICTOR MODEL USING MACHINE LEARNING**



**Submitted in partial fulfillment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Student Name -** **University Roll No.**

Suchika Raturi 2021478

***Under the Mentorship of***

**Dr. Jyoti Agarwal**

**Associate Professor**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**January-2025**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Car Price Predictor Model using Machine Learning”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. Jyoti Agarwal, Associate Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

**Name: University Roll no:**

Suchika Raturi 2021478

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | **1-2** |
| Chapter 2 | Literature Survey | **3-5** |
| Chapter 3 | Methodology | **6-7** |
| Chapter 4 | Result and Discussion | **8-9** |
| Chapter 5 | Conclusion and Future Work | **10** |
|  | References |  |

**Chapter 1**

**Introduction**

One of the first inquiries someone has when looking to purchase or sell an automobile is, "Exactly how much is this car worth?" To gain a sense of car prices in the past, customers would have perused newspaper ads, ask friends, or visit dealerships. However, there is a more intelligent way to determine this now that we have access to computers' power and data.

The project involves predicting car prices using machine learning. It can be challenging to tell if the advertised price for a used or old car is accurate due to the many factors that influence a used car's cost on the market. In order to make informed purchases, this project focuses on creating machine learning models that can properly guess an old car's cost on the basis its qualities.

In order to solve this problem, this project aims to predict the Price of a used Car by taking it's Company name, its Model name, Year of Purchase, and other parameters. Particularly, for this analysis use the linear regression method, a powerful tool for predictive modeling, and Python's complex ML environment.

A complete collection of data about numerous car brands is included in the dataset utilized for this project. It contains key information about the vehicle, such as the car company, models, year of purchase, fuel type, and, most importantly, the kilometers that the car has travelled, which can play big role in predicting prices of used cars

The following are the parameters used:

1. Name of Car: The Car name with its brand.
2. Year: Year in which car has been purchased.
3. Selling Price: The price in which car has been sold.
4. Present Price: The current price of the car.
5. KM driven: It shows how much car has been driven.
6. Fuel Type: Type of fuel which is used in car like petrol, diesel, CNG, electric.
7. Selling type: It basically shows that the car has been sold by individual owner or by the dealer.
8. Transmission: It shows car type like Automatic or Manual.

Other elements like if the dealer is the vehicles owner or a dealer, whether the owner is the first or second owner of a vehicle, whether the gear is manual or automatic, and whether gasoline or diesel is used as energy are also taken into account while creating the model.

Therefore, this model was chosen improve methods and account for the dataset's huge number of features. When determining a pre-owned vehicle's price, a variety of factors are taken into account, including the moment of purchase, the mileage of the car, its exchange value, its availability, and the car's machine and seating capacities.

The goal of this model is to anticipate car pricing simply, these are differentiated with the use of various machine-learning algorithms. They separated the prices into various, previously determined price categories.

This report details the steps undertaken in the project and discusses the results, challenges, and potential future improvements. The car price prediction model aims to provide a robust and practical solution for real-world application.

**Chapter 2**

**Literature Survey**

Over the years, several methodologies have been employed to address this challenge, ranging from traditional statistical approaches to modern Machine Learning techniques.

Traditionally, price prediction relied on method like **Linear Regression**, which attempts to establish a relationship between car attributes such as age, mileage, brand, and fuel type with the car's price. This approach was simple but faced many challenges in handling non-linear relationships.

In addition to Linear Regression, Multiple Regression was used to consider more variables, but it still struggled with complex datasets.

Over time, more advanced Machine Learning (ML) techniques emerged as a powerful solution to address the limitations of traditional methods. These techniques involve Decision Tree, Random Forests and Gradient Boosting Machines. In addition to these techniques Deep Learning models have been applied to car price prediction. These models can automatically learn complex patterns from large datasets, making them powerful tools for high-accuracy predictions.

By using LinearRegression, this project aims to predict car prices based on features such as age, transmission, fuel type, etc. ensuring a balance between accuracy and transparency.

**2.1 Terminologies used in this project:**

**2.1.1 Regression Machine Learning**

Regression analysis is a type of supervised machine learning wherein the model is trained with both input features and output labels. It helps in establishing a relationship among the variables. The input variables are called independent variables and correspond to features in the dataset, while the output variable is called the dependent variable.

**2.1.2 Overfitting**

Overfitting is an undesirable machine learning behavior that occurs when the machine learning model gives accurate predictions for training data but not for new data.

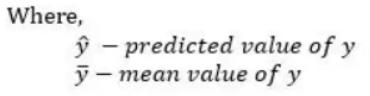
This condition can affect all supervised machine learning models. In the case of regression models, overfitting can occur when there many terms for the number of observations.

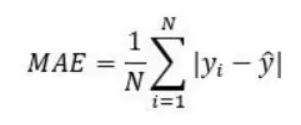
**2.1.3 Linear Regression**

Linear Regression is a technique to estimate the linear relationship between each of a number of independent variables and a dependent variable. The objective of Linear Regression is to find a line that minimizes the prediction error of all the data points.

**2.1.4 Mean Absolute Error (MAE)**

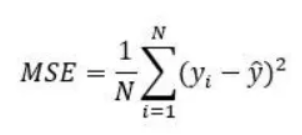
The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.





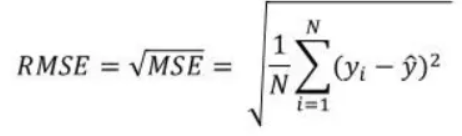
**2.1.5 Mean Squared Error (MSE)**

Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals.



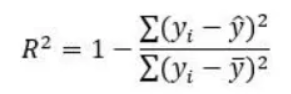
**2.1.6 Root Mean Squared Error (RMSE)**

Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals.



* + 1. **R Squared (R2)**

The coefficient of determination or R-squaredrepresents the proportion of the variance in the dependent variable which is explained by the linear regression model. It is a scale-free score i.e. irrespective of the values being small or large, the value of R square will be less than one.



Both RMSE and R- Squared quantifies how well a linear regression model fits a dataset. The RMSE tells how well a regression model can predict the value of a response variable in absolute terms while R- Squared tells how well the predictor variables can explain the variation in the response variable.

**Chapter 3**

**Methodology**

This section includes the steps followed in developing the Car Price Prediction model, including details of the dataset, preprocessing techniques, model development, and tools used in the project.

**3.1 Project requirements**

**3.1.1 Hardware Components**

1. Adequate hardware specifications suitable for running Python applications smoothly.

**3.1.2 Software Components**

1. Operating System: Windows 11

2. Programming Language: Python,

3. Back-End: Python

**3.2 Tools and Technologies Used**

The following tools and technologies were use during the project:

1. Python
2. Jupyter Notebook
3. scikit-learn
4. Pandas
5. NumPy
6. Matplotlib and Seaborn
7. Pickle

**3.3 Module Division**

**3.3.1 Dataset description**

The dataset for this project was sourced from Kaggle, a popular platform for datasets and machine learning competitions. The dataset comprises information about various factors that influence car prices.

**3.3.2 Dataset pre-processing**

It involves handling missing values, feature scaling, encoding categorical variables using one-hot encoding, and data splitting for the training and testing of the data.

**3.3.3 Algorithm and Model Development**

The chosen algorithm for this project is Linear Regression.

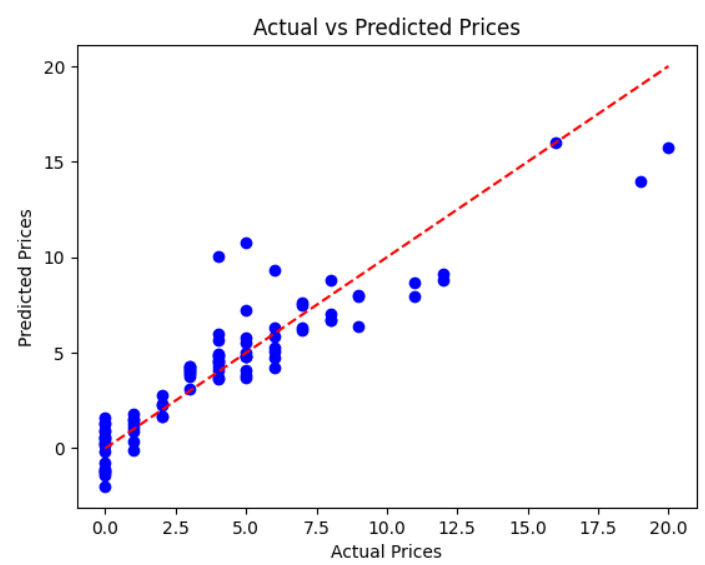
Firstly, the model was trained on the training dataset using the **scikit-learn** library.

Then the model's performance was evaluated using metrics such as: Mean absolute error, Mean squared error, Root mean squared error and r-squared.

**Chapter 4**

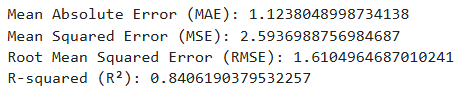
**Result and Discussion**

This section evaluates the performance of the Car Price Prediction model. It visualizes the predicted prices versus the actual prices.



A scatter plot was created comparing the predicted car prices to the actual car prices. A straight diagonal line represents perfect predictions, and the closer the points are to this line, the better the model's performance.

To assess the effectiveness of the Linear Regression model, the following performance metrics were calculated on the testing dataset:



By analyzing the model's results and performance, it is clear that the Linear Regression model provides a solid foundation for predicting car prices.

**Chapter 5**

**Conclusion and Future Work**

**5.1 Conclusion**

In conclusion, our project successfully developed a model for predicting car prices using machine learning algorithms, particularly Linear Regression. Through data analysis and model training, we achieved promising results in accurately predicting car prices based on various features such as car model, fuel type, transmission and age of the car. It helps people decide if they’re paying the right amount for a car.

**5.2 Future Work**

To enhance the Car Price Prediction Model, several avenues for future development can be explored.

**-** We can experiment with advanced Machine Learning algorithms like Random Forests, Gradient Boosting Machinesto capture more complex relationships in the data. Then compare the performance of these models against Linear Regression to identify the most suitable algorithm.

**-** Secondly, we can create a web page or a web application using python libraries so that it may provide a user-friendly interface for the buyers.

**-** We can also deploy the application on a cloud platform for wider accessibility and scalability.

**-** Thirdly, we can develop an API to enable real-time predictions, making the system suitable for integration with e-commerce platforms and other automotive services.

**-** Lastly, we can perform feature importance analysis to identify and rank the most influential factors affecting car prices. This could help in improving the model.

**References**

1. Annina S, Mahima SD, Ramesh B, “An Overview of Machine Learning and its Applications”. International Journal of Electrical Sciences & Engineering (IJESE), 2015. pp. 22-24.
2. Sakshi Gupta, “Regression vs. Classification in Machine Learning”, October 6, 2021 [Online] Available: <https://www.springboard.com/blog/data-science/regression-vsclassification/>
3. Jim Frost, “Overfitting Regression Models,” Statistics by Jim, 2022. [Online], Available: <https://statisticsbyjim.com/regression/overfitting-regressionmodels/>
4. Skikit-learn, supervised-learning. [Online] Available: https://scikitlearn.org/stable/supervised\_learning.html#supervised-learning
5. Used cars database. Retrieved from: <https://www.kaggle.com/>
6. Jupyter notebook: <https://jupyter.org/>
7. Github: <https://github.com/>
8. YouTube Reference: <https://www.youtube.com/>
9. Linear regression in machine learning: <https://www.geeksforgeeks.org/ml-linearregression/>
10. Mean Squared Error or R-Squared – Which one to use?”. [Online] https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e