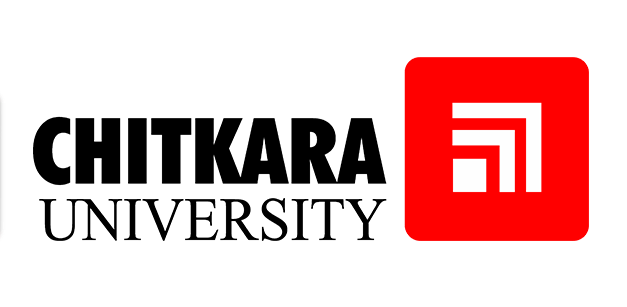
**Artificial Intelligence and Machine Learning**

**Project Report**

**Semester-IV (Batch-2022)**



**Predicting Used Car Prices with Artificial Intelligence**

**Supervised By: Submitted By:**

**Jatin Arora Dhruv Bansal(2210990271)G7**

# Department of Computer Science and Engineering

**Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab**

# ABSTRACT

The aim of this project is to develop a predictive model for estimating the selling price of used cars based on various features such as the car's age, mileage, brand, model, and condition. The project utilizes machine learning techniques to analyze historical data and build a regression model that can accurately predict the selling price of a car given its attributes.

The motivation behind this project stems from the growing demand for reliable pricing information in the used car market. Potential buyers and sellers often struggle to determine fair market prices due to the complex interplay of factors influencing a car's value. By developing a robust pricing model, this project seeks to provide a valuable tool for both buyers and sellers, helping them make informed decisions and facilitating fair transactions in the used car market.

The methodology involves data collection from various sources, data preprocessing to handle missing values and outliers, feature engineering to extract relevant information, and model training and evaluation using machine learning algorithms such as linear regression, decision trees, or ensemble methods. The project will focus on optimizing the model's performance metrics such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) to ensure accurate price predictions.

The project's significance lies in its potential to contribute to the automotive industry by offering a data-driven approach to pricing used cars, reducing information asymmetry, and improving market transparency. The developed model can be deployed as a web-based tool or integrated into existing platforms to provide users with real-time price estimates based on current market trends and car attributes.

# Table of Contents

|  |  |  |
| --- | --- | --- |
| **Sr.No** | **TITLE** | **Page no.** |
| 1 | Introduction   1. Background 2. Objective 3. Introduction to Blockchain Technology 4. Overview of Machine Learning in Fraud Detection 5. Significance | 1 |
| 2 | Problem Definition and Requirement   1. Problem Statement 2. Software Requirements 3. Hardware Requirements 4. Data Set | 4 |
| 3 | Proposed Design/Methodology   1. Project Directory 2. Methodology 3. Algorithms Used | 8 |
| 4 | Results   1. Data Preprocessing 2. Model Training 3. Model Evaluation 4. Result Analysis | 11 |
| 5 | References   1. Data Set 2. Study Material | 30 |

# 1. INTRODUCTION

**1.1. BACKGROUND:**

The automotive industry has witnessed a significant surge in the demand for used cars in recent years. Factors such as economic considerations, changing consumer preferences, and the availability of reliable pre-owned vehicles have fueled this trend. However, one of the key challenges faced by both buyers and sellers in the used car market is accurately determining the fair selling price of a vehicle. This challenge is compounded by the diverse range of factors that can influence a car's value, including its age, mileage, brand, model, condition, and market trends.

**1.2. OBJECTIVE:**

**. Develop a predictive model to estimate the selling price of used cars based on various attributes.**

**. Collect and preprocess a diverse dataset of used car listings to train the model.**

**. Perform feature engineering to extract relevant features that impact car pricing.**

**. Train and evaluate machine learning algorithms such as regression models to predict selling prices accurately.**

**1.3. INTRODUCTION TO ARTIFICIAL INTELLIGENCE :**

Artificial Intelligence, often abbreviated as AI, is a branch of computer science that focuses on creating intelligent systems capable of performing tasks that typically require human intelligence. The goal of AI is to develop algorithms and technologies that enable machines to simulate human-like cognitive functions such as learning, reasoning, problem-solving, perception, and decision-making.

The field of AI encompasses a wide range of techniques, methodologies, and applications, each aimed at replicating different aspects of human intelligence. From basic rule-based systems to advanced machine learning algorithms and neural networks, AI has evolved significantly over the years, driven by advancements in computing power, data availability, and algorithmic innovation.

**1.4. OVERVIEW OF MACHINE LEARNING IN PRICE PREDICTION :**

Overview of Machine Learning in Price Prediction:

Machine learning plays a crucial role in price prediction across various domains such as finance, e-commerce, real estate, and automotive industries. The ability of machine learning algorithms to analyze large datasets, identify patterns, and make predictions based on historical data makes them valuable tools for price forecasting and decision-making. Here's an overview of how machine learning is utilized in price prediction tasks:

1. Data Collection and Preprocessing:

-Price prediction models require relevant historical data related to the items or assets being forecasted. This data may include features such as past prices, market trends, economic indicators, product attributes, and customer behavior .

-Data preprocessing steps involve cleaning the data, handling missing values, encoding categorical variables, scaling numerical features, and splitting the dataset into training and testing sets.

2. Feature Engineering:

- Feature engineering is a critical step where meaningful features are extracted or created from raw data to improve the predictive power of the model. This may involve transformations, aggregations, or combinations of existing features to capture important patterns and relationships.

- In price prediction, engineered features might include moving averages, price volatility measures, seasonality indicators, sentiment scores (in sentiment analysis-based models), or derived features specific to the domain.

3. Model Selection and Training:

- Various machine learning algorithms can be used for price prediction, including linear regression, decision trees, random forests, support vector machines (SVM), gradient boosting methods (such as XGBoost or LightGBM), and neural networks (deep learning).

- The choice of model depends on factors such as the complexity of the data, the interpretability of predictions, scalability requirements, and the trade-off between bias and variance.

- Models are trained on the training dataset using optimization techniques to minimize prediction errors (e.g., minimizing Mean Squared Error for regression tasks).

4. Model Evaluation and Validation:

- Trained models are evaluated using appropriate metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), or R-squared (coefficient of determination).

- Cross-validation techniques such as k-fold cross-validation or time-series cross-validation are used to assess the model's generalization performance and detect overfitting or underfitting issues.

5. Hyperparameter Tuning and Optimization:

- Hyperparameters are parameters that control the learning process of machine learning algorithms (e.g., learning rate in neural networks, depth of decision trees).

- Hyperparameter tuning involves searching for the optimal combination of hyperparameters that yields the best model performance. Techniques like grid search, random search, or Bayesian optimization can be employed for hyperparameter tuning.

Machine learning-based price prediction models can provide valuable insights and assist decision-makers in making informed pricing strategies, optimizing inventory management, detecting anomalies or market trends, and improving overall business performance. However, it's important to regularly update and retrain models as data distributions and market conditions change over time.

1.5. **SIGNIFICANCE:**

The significance of a price prediction project using machine learning techniques extends across various domains and industries, including finance, e-commerce, real estate, retail, and more. Accurate price prediction models empower businesses and individuals to make data-driven decisions regarding pricing strategies, investment opportunities, inventory management, and resource allocation. Price prediction models help businesses optimize their pricing strategies by balancing competitiveness, profitability, and market demand. They can identify price-sensitive segments, recommend dynamic pricing adjustments, and optimize pricing tiers for different customer segments or market conditions.

# 2. PROBLEM DEFINITION AND REQUIREMENT

**2.1. PROBLEM STATEMENT:**

Predicting the selling price of used cars involves a systematic approach that starts with collecting a diverse dataset encompassing key features such as make, model, year of manufacture, mileage, condition, and location. Subsequently, data preprocessing steps are essential to clean the dataset, handle missing values, outliers, and convert categorical variables into numerical representations. Feature selection and engineering are crucial for identifying the most pertinent features and potentially creating new ones, like car age or depreciation rate. Splitting the dataset into training and testing sets enables the evaluation of model performance. Model selection entails choosing an appropriate regression algorithm, like linear regression or decision trees, followed by training the model on the training dataset. Evaluation metrics such as MAE, MSE, and RMSE gauge the model's performance on the testing dataset. Hyperparameter tuning further refines the model's performance through techniques like grid search or random search. Deploying the model into production allows it to predict selling prices based on input features. Continuous monitoring and maintenance are paramount to ensure the model's accuracy and effectiveness over time, accounting for changes in data distribution and external factors. Iterative experimentation and consideration of data quality are fundamental for building a robust predictive model**.**

**2.2. SOFTWARE REQUIREMENT:**

Our solution will be implemented using Python, a versatile and widely adopted programming language suitable for data analysis and machine learning tasks. In addition to Python, we will leverage several libraries to facilitate various aspects of our project:

1. **NumPy:** Utilised for efficient numerical operations and array manipulation.
2. **Matplotlib:** Employed for creating visualisations and graphs to analyse data.
3. **Seaborn:** Used to enhance the aesthetics of visualisations and statistical data exploration.
4. **Pandas:** Utilised for data manipulation and analysis, offering powerful data structures and functions.
5. **Scikit-learn (sklearn):** Utilised for implementing machine learning algorithms, model evaluation, and preprocessing.

**Methodologies:-**

1. Linear Regression: The linear relationship between independent variables and a continuous dependent variable, making it a fundamental and interpretable technique in predictive modeling and data analysis.
2. Gradient Boosting Trees (GBT): It use sequential decision trees to make powerful predictions by focusing on errors from previous trees.
3. **Random Forest:** Utilized as an ensemble learning method, constructing multiple decision trees to improve accuracy and robustness.

These software components provide the essential framework for developing our fraud detection algorithms effectively. By leveraging these libraries, we can streamline the development process and ensure the scalability and reliability of our solution.

* 1. **HARDWARE REQUIREMENT:**

In terms of hardware, we require a computer with sufficient processing power and memory to support the computational demands of our algorithms.

While the specific hardware specifications may vary depending on the scale of the project and the size of the dataset, a standard computer with a multi-core processor and ample RAM should suffice for development and testing purposes.

* 1. **DATA SET:**

The dataset includes fundamental car details such as make, model, and year of manufacture.

Mileage provides insight into usage patterns and overall condition, impacting perceived value.

Assessment of condition (excellent, good, fair, poor) influences market desirability and selling price.

Transmission type and fuel variation shape consumer preferences and pricing dynamics.

Ownership details like the number of previous owners reflect on reliability and maintenance history.

Insights into accident and service histories impact perceived safety and maintenance requirements.

Geographical location affects supply, demand, and pricing dynamics.

Market demand reflects consumer preferences and trends for specific makes and models.

Economic indicators such as inflation rates and GDP growth shape purchasing power and market dynamics.

The selling price reflects the culmination of various intrinsic and extrinsic factors.

# 3. PROPOSED DESIGN / METHODOLOGY

Our proposed design and methodology focus on utilising linear regression, GBT and random forest algorithms to develop a fraud detection system tailored for Ethereum transactions. As we have only one file containing all the work and a CSV file containing the data, our approach will be streamlined and contained within this single file.

**3.1. PROJECT DIRECTORY:**

UsedCarPricePrediction.ipynb

car\_price\_data1.csv

**usedcarpricepredection.ipynb:** This Python script contains all the code for data preprocessing, model training, evaluation, and deployment. It encompasses the entire workflow of our price prediction model .

car\_price\_data1.csv**:** The price prediction dataset sourced from Kaggle, containing car data.

**3.2. METHODOLOGY:**

**Data Preprocessing:**

 Split the dataset into training, validation, and test sets for model training, tuning, and evaluation.

 Perform data scaling and normalization on numerical features to ensure consistency in model training

 Set up a data processing pipeline to automate these steps and maintain consistency across

experiments.

**Model Training:**

* Implement logistic regression, GBT and random forest algorithms using scikit-learn.
* Train each model on the training dataset.

**Model Evaluation:**

* Evaluate the performance of each model using metrics such as accuracy, ,r2,rmse.
* Compare the performance of logistic regression, decision tree, and random forest algorithms to identify the most effective approach for fraud detection .

**3.3. ALGORITHM USED:**

Linear Regression: Linear regression is a simple yet powerful statistical method for modelling the relationship between variables, making it suitable for tasks such as price prediction based on features like age, mileage, and brand in the used car market. Its interpretability and ease of implementation make it a popular choice in data analysis and predictive modelling.

Random Forest: It is an ensemble learning method that uses multiple decision trees to make predictions, combining their outputs through averaging or voting, resulting in robust and accurate predictions across various domains, including regression and classification tasks.

By utilising Linear regression, GBT and random forest algorithms within a single Python script, we ensure a cohesive and efficient approach to developing our price prediction model . This streamlined methodology enables us to leverage the dataset effectively and identify the most suitable algorithm for predicting price within car data .

# RESULTS

In this section, we present the detailed results of our price prediction model using linear regression, GBT and random forest algorithms. We outline the steps taken in data preprocessing, discuss the handling of missing values, class imbalance correction, and the creation of correlation matrices. Then, we delve into the model training process and provide comprehensive evaluations of each model's performance on the car dataset .

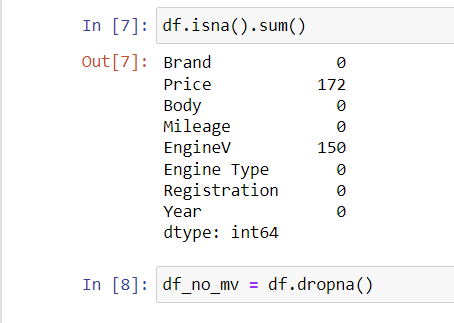
**4.1. Data Preprocessing:**

**Handling Missing Values:**

* We addressed missing values in the dataset using appropriate techniques such as imputation or deletion. The goal was to ensure that the dataset was complete and suitable for model training.
* We created a heat map to visualise the missing values in the dataset.

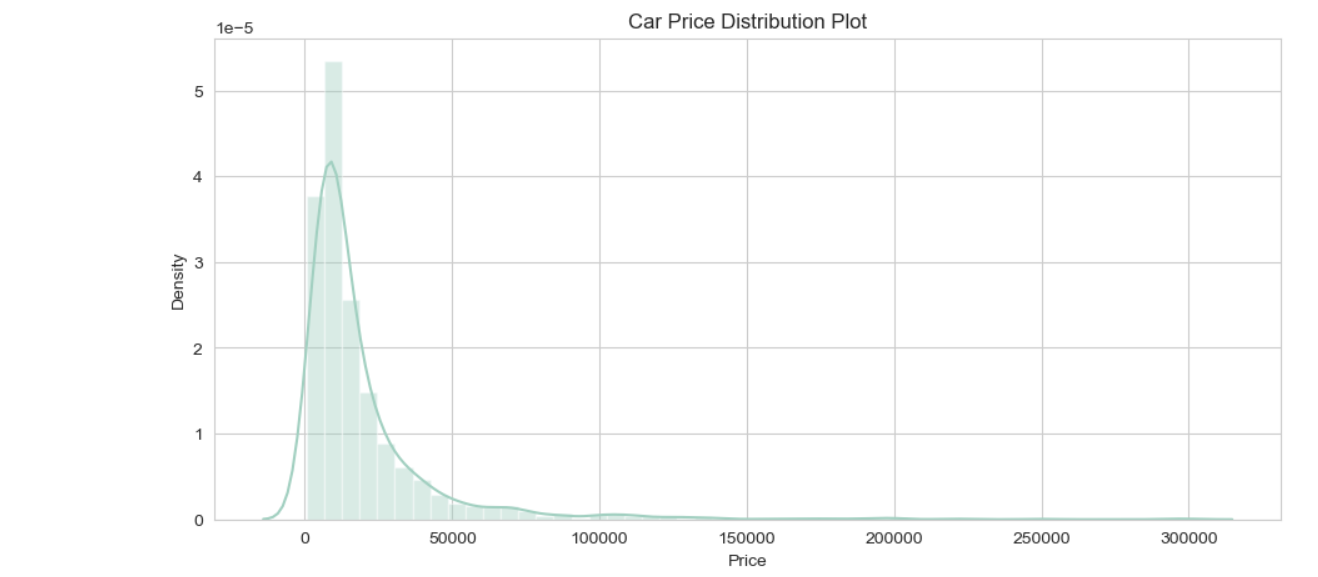
**-**

* The null values are indicated by the red colour, to handle these values we dropped the null values using the drop na() function.
* We then again visualised the dataset with the help of a heat map, to check whether all the missing values have been handled.



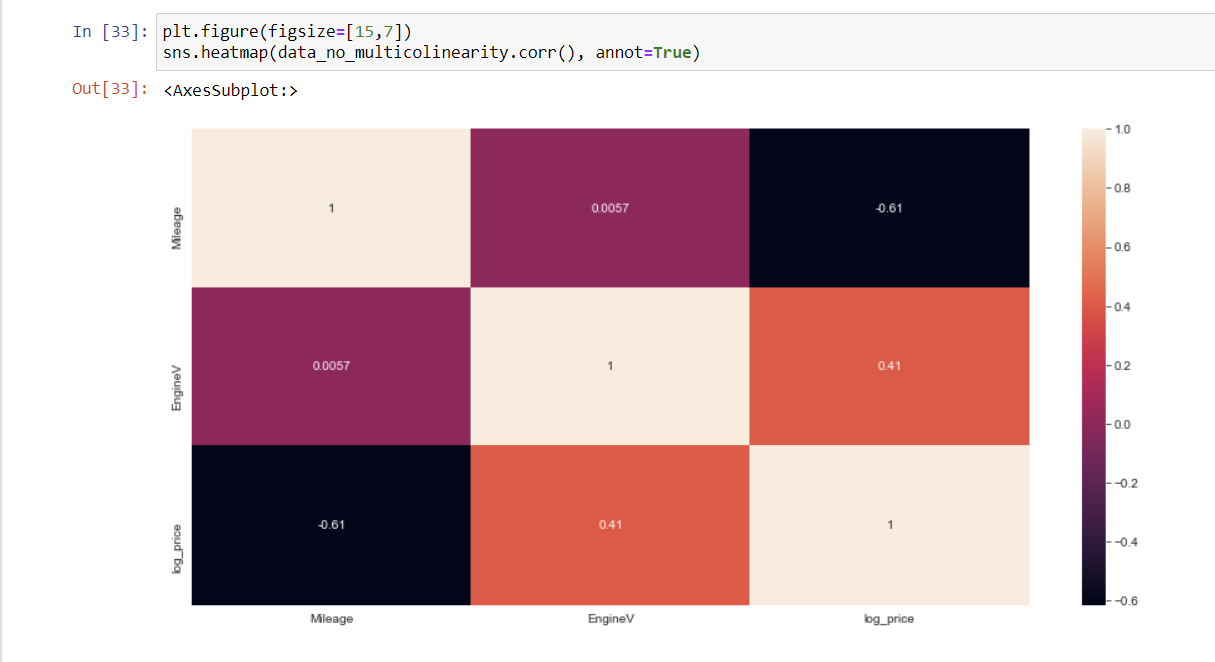
**Bar plot to visualise the distribution of car price categorised as normal versus fraudulent:**

* We created a bar plot to check the visual representation of the Density versus car prices .



-The above distribution plot shows how prices are distributed among different density levels . The highest prices of cars are in range of 0-50000 .

**Creating Correlation Matrix:**

* We created a correlation matrix to identify relationships between features and determine which features were most strongly correlated with car price . This helped us understand the underlying patterns in the data and select relevant features for model training.
* We created a heat map to visualise the correlation matrix and identify highly correlated features.
* 

* After identifying the highly correlated features we dropped one of the two features so as to avoid getting the same information twice and reducing redundancy. This ensures that our analysis remains clear, accurate, and efficient.

**4.2. Model Training:**

**Linear Regression**:

Linear regression is a simple yet powerful statistical method for modeling the relationship between variables, making it suitable for tasks such as price prediction based on features like age, mileage, and brand in the used car market. Its interpretability and ease of implementation make it a popular choice in data analysis and predictive modelling.

**Random Forest:**

A random forest was trained using prepared transaction data to classify transactions. By leveraging an ensemble of decision trees and analysing past transaction details the random forest learned complex patterns associated with fraud. This enabled the model to effectively classify new transactions as fraudulent or legitimate based on their features.

**Gradient Booster :**

Gradient boosting works by sequentially adding decision trees to the model, each tree correcting the errors of the previous ones. This iterative process focuses on minimizing the residual errors and optimizing the overall prediction performance. Features like age, mileage, brand, and condition can be effectively utilized by gradient boosting algorithms to capture nonlinear relationships and make accurate price predictions.

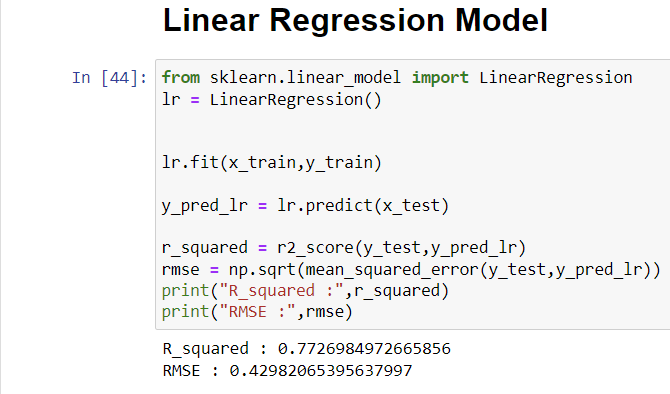
**4.3. Model Evaluation:**

**Metrics Calculation:**

We evaluated each model's performance using standard metrics such as accuracy, precision, r\_squared and RMSE. These metrics provided insights into the models' ability to correctly classify car data and predict price.

**Linear Regression:**

**- Prediction Report :**



Key Takeaways:

**-**Linear Relationship: Linear regression assumes a linear relationship between the independent variable(s) and the dependent variable. This relationship is represented by a straight line in a two-dimensional space or a plane in a multi-dimensional space.

**-**Interpretability: Linear regression provides interpretable coefficients that indicate the strength and direction of the relationship between each independent variable and the dependent variable. For example, a positive coefficient indicates a positive correlation, while a negative coefficient indicates a negative correlation.

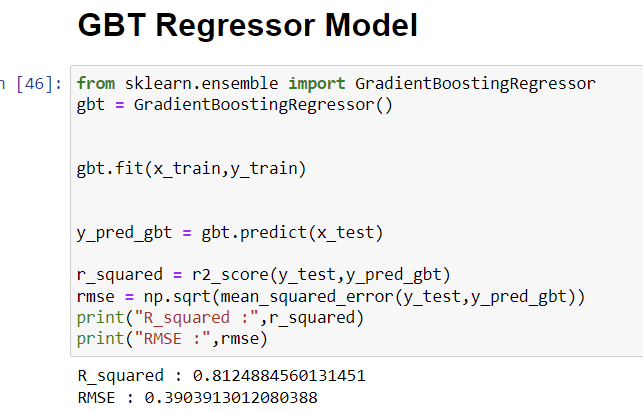
-Prediction: Linear regression can be used for prediction tasks, where the model predicts the value of the dependent variable based on the values of independent variables. It's particularly useful for making continuous predictions, such as predicting house prices, stock prices, or sales volumes.

-Assumptions: Linear regression relies on several assumptions, including linearity (relationship between variables), independence of residuals (errors), homoscedasticity (constant variance of residuals), and normality of residuals (residuals are normally distributed).

-Model Evaluation: Common metrics for evaluating linear regression models include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and R-squared (R2) score. These metrics help assess the accuracy and goodness of fit of the model.

* + - **GBT:**

**- Prediction Report :**



Key Takeaways:

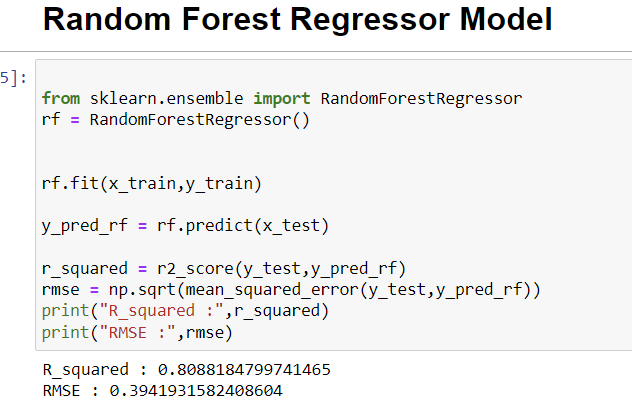
-Ensemble Learning: GBT is an ensemble learning technique that combines the predictions of multiple decision trees to create a strong predictive model. It leverages the strengths of individual trees while mitigating their weaknesses, leading to improved predictive accuracy.

-Sequential Learning: GBT builds trees sequentially, where each new tree corrects the errors (residuals) of the previous ones. This iterative process focuses on reducing prediction errors and optimizing the overall model performance.

-Gradient Descent Optimization: GBT uses gradient descent optimization to minimize a loss function (e.g., mean squared error) during tree construction. This optimization approach ensures that the model learns from the gradients of the loss function, leading to better convergence and predictive performance.

1. **Random Forest Regressor:**

**- Prediction Report:**



Key Takeaways:

-Ensemble Learning: Random Forest is an ensemble learning technique that builds multiple decision trees during training and combines their predictions to make more accurate and robust predictions. It leverages the wisdom of crowds by aggregating the outputs of individual trees.

-Bagging Technique: Random Forest uses a bagging (bootstrap aggregating) technique where each tree is trained on a random subset of the training data with replacement. This randomness in data sampling and feature selection helps reduce overfitting and improve generalization.

-Feature Importance: Random Forest provides a measure of feature importance based on how much each feature contributes to reducing impurity or increasing information gain in the trees. This information can be valuable for understanding the factors that influence price predictions in the used car market.

**4.4. Results Analysis:**

a) Evaluation Metrics:

-Calculate and analyse performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and R-squared (R2) score .

-MAE and RMSE measure the average magnitude of errors, with lower values indicating better accuracy.

-R2 score quantifies the proportion of variance in the target variable explained by the model, with values closer to 1 indicating a better fit.

b) Model Comparison:

-Compare the performance of different machine learning algorithms used in our project **i.e** Linear Regression, Gradient Boosting Trees, Random Forest.

## -Evaluating the model(s) providing the most accurate and reliable predictions for selling prices of used cars based on the chosen evaluation metrics is Gradient booster (GBT) .

## c) Feature Importance Analysis:

## -If applicable (e.g., for models like Gradient Boosting Trees or Random Forest), analyse the feature importance scores to understand which features contribute significantly to price predictions.

## -The key factors such as age, mileage, brand, model, condition, and others that have a strong influence on the predicted prices.

## d) Residual Analysis:

## -Examine the residuals (differences between actual and predicted prices) to check for patterns or biases in the model's predictions.

## -Plotting residuals against predicted prices or actual prices can help identify areas where the model may be underperforming or overestimating/underestimating prices.

## -Cross-Validation Results: If you used cross-validation techniques during model training, analyse the cross-validation results to ensure that the model's performance generalizes well to new data.

## -Look for consistency in performance metrics across different folds or validation sets.

## e) Business Impact Analysis:

## -Translate the model's performance into business impact by quantifying potential benefits such as improved pricing strategies, revenue gains, cost savings, or customer satisfaction improvements.

## -Consider the practical implications of using the model in real-world scenarios and assess its alignment with business objectives.

## f) Iterative Improvement:

## -Use the insights gained from result analysis to iteratively improve the model. This may involve fine-tuning hyperparameters, adding or modifying features, or exploring alternative algorithms.

## -Continuously monitor the model's performance over time and update it as new data becomes available or business requirements evolve.

## 5. REFERENCES

**5.1. Data Set:**

-<https://www.kaggle.com/code/mohaiminul101/car-price-prediction>

**5.2. Study Material:**

-NUMPY <https://www.youtube.com/watch?v=aYmcRnmZVGQ&list=PL9n0l8rSshSnragNblKDBsT8Xu3otp3jA>

* RANDOM FOREST-https://www.youtube.com/watch?v=3NdH3egUjpM&list=PLBSCvBlTOLa8Xp1tjW0C3So94zhDbTTvh