**Car sales price prediction**

A Project Report in partial fulfillment of the degree

# Bachelor of Technology

in

# Computer Science & Engineering

## By

**2203A51141** M.REETHIKA

**Under the Guidance of**

# D. Ramesh

# Assistant Professor, Department of CSE.

**Submitted to**



# DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

## SR UNIVERSITY, ANANTHASAGAR, WARANGAL





**DEPARTMENTOFCOMPUTERSCIENCE&ENGINEERING**

**CERTIFICATE**

This is to certify that the Project Report entitled “Car sales price prediction ”is a record of Bonafide work carried out by M.Reethika**,** bearing RollNo**, 2203a51141** during the academic year 2023-2024 in partial fulfillment of the award of the degree of ***Bachelor of Technology*** in **Computer Science Engineering** by the SR UNIVERSITY, WARANGAL.

# Supervisor HeadoftheDepartment

Mr. D. Ramesh Dr. M. Sheshikala

Asst. Professor, Assoc. Prof .& HOD (CSE)

SR University SR University

**ACKNOWLEDGEMENT**

We express our thanks to course coordinator Mr. D. Ramesh, Asst. prof. for guiding us from the beginning through the end of the course project. We express our gratitude to head of the department CS&AI, Dr. M. Sheshikala, Associate Professor for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, Dr C. V. Guru Rao, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thank to all teaching and non-teaching staff of the department for their suggestions and timely support.

# ABSTRACT

This study proposes a predictive model for estimating car purchase amounts based on customer attributes such as name, email, country, gender, age, annual salary, credit card debt, and net worth.

Cars of a particular make, model, year, and set of features start out with a price set by the manufacturer. Using Machine Learning algorithms to better utilize data on all the less common features of a car can more accurately assess the value of a vehicle. This study compares the performance of Linear Regression, k Nearest-Neighbour, Decision tree, Support vector machine ,Random Forest Regressor and Adaboost ML algorithms in predicting the price of used cars Utilizing machine learning techniques, the model aims to provide vehicle salespersons with accurate predictions of consumer spending, enabling targeted marketing strategies and personalized customer interactions.

The results show that Random Forest Regression demonstrates the highest price prediction performance across all metrics used. The model finds relevant variables and uses regression analysis to forecast possible car expenses by examining a dataset of client profiles and associated purchase quantities. The outcomes show that the suggested method is effective in forecasting car sales prices, enabling sales teams to improve customer satisfaction levels and maximize their sales tactics.

This study puts forward a detailed plan to predict car prices. It examines various customer details like their names, email addresses, location, gender, age, income, credit card debt, and wealth. Using advanced computer methods, it aims to forecast how much people are likely to spend on buying a car. By analyzing a large dataset of customer information and purchase amounts, the plan identifies important factors and how they relate to buying behavior.

# Table of Contents

|  |  |  |
| --- | --- | --- |
| **S.NO** | **Content** | **PageNo** |
| 1 | Introduction | 1 |
| 2 | Literature review | 2 |
| 3 | Design | 4 |
| 4 | Methodology | 5 |
| 5 | Data Pre-processing | 8 |
| 6 | Results | 9 |
| 7 | Conclusion | 19 |
| 8 | Future scope | 20 |
| 9 | References | 20 |

**1.INTRODUCTION:**

My project is about CAR SALES PRICE PREDICTION. The plan figures out important things and uses a method called regression analysis to guess how much money people might spend on cars. It looks at a bunch of information about customers and how much they buy. The results say that the plan works well in guessing car prices, which helps sales teams make customers happier and sell more cars.

The plan uses regression analysis to guess how much people might spend on cars. It looks at lots of information about customers and how much they buy. The results show that the plan is good at guessing car prices, which helps sales teams make their approaches more personal and make customers happier.

The model examines important factors and estimates how much people might spend on cars using regression analysis. It looks at various details about customers and their purchases. The results suggest that this approach is effective in predicting car prices. This helps sales teams enhance customer satisfaction and improve their sales tactics..

# 2. LITERATURE REVIEW

[1] Various studies have been conducted in order to predict the price of used cars. Researchers regularly anticipate product prices using past data. Pudaruth predicted car prices in Mauritius, and these cars were not new, but rather used to predict the prices, he employed multiple linear regression, knearest neighbours, Naive Bayes, and decision tree techniques.

[2] Kuiper, S. (2008) demonstrated a multivariate regression model that helps in classifying and predicting values in numeric format. It demonstrates how to apply this multivariate regression model to forecast the price of 2005 General Motors (GM) vehicles.

[3] In 2019, Pal et al discovered as a methodology for predicting used cars prices using Random Forest. The paper evaluated usedcar price prediction using Kaggle data set which gave an accuracy of 83.62% for test data and 95% for train-data.

[4 ] Gegic, E. et al. (2019) demonstrate the need to create a model to forecast the cost of second hand cars in Bosnia and Herzegovina. They used machine learning techniques such as artificial neural networks, support vector machines, and random forests.

[5] The goal of the system that Dholiya , M., et al. developed is to give the user a realistic estimation of how much the vehicle might cost them. Based on the specifics of the automobile the user is looking for, the system, which is a web application, may also offer the user a list of options for various car kinds. It assists in providing the buyer or seller with useful information on which to base their decision.

[6] Richardson conducted his analysis under the presumption that automakers are more inclined to produce cars that don't lose value quickly. He demonstrated, in particular, that hybrid cars are better equipped to maintain their value than conventional vehicles by utilising multiple regression analysis. This is perhaps because there are increasing concerns about the environment and the climate, as well as because it uses less gasoline. In this study, the significance of additional variables including age, mileage, make, and MPG (miles per gallon) was also taken into account. All of his information was gathered from several

| **SINO** | **DATEOF** | **AUTHORS** | **NAME** | **METHODOLOGY** | **ERROR** | **PUBLICATION** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2013 | Johnson | Car Purchase Price Prediction | Multiple linear regression, neural networks | 7.7 | Scikit-learn |
| 2 | 2013 | Romeo et al. | Car Price Classification | K-nearest neighbor, decision tree | 14 | Springboard Blog |
| 3 | 2014 | Cakir et al. | Car Purchase Price Prediction | Artificial neural networks, multivariate polynomial regression | 6.5 | Statistics by Jim |
| 4 | 2016 | Gandhi et al. | Car Price Prediction using SVM | Support vector machine | 23.43 | Frontiers in Plant Science |
| 5 | 2016 | Sujatha and Isakki | Car Price Prediction using Classification Techniques | Combination of MLSR and PLS Regression | 32.24 | n/a |
| 6 | 2017 | Ying-xue et al. | Car Price Prediction using SVM | Support vector machine-based open crop model (SBOCM) | 22 | n/a |
| 7 | 2018 | Crane-Droesch | Car Price Prediction using Neural Networks | Neural networks | 17 | n/a |

# 3.DESIGN:

**RequirementSpecifications**

## Hardware Requirements

## System

## RAM

## HardDisk

## Input

## Output

## Software Requirements

* + - **OS**
    - **Platform**
    - **Program Language**

# 4. METHODOLOGY:

After Data pre-processing and data visualization the next step is to apply the models on the dataset. Our dataset comes under supervised learning as it contains the labeled data (target variables, feature variables). First the dataset is splitted into training set and testing set. Then the model is trained on training set and then tested on testing set.

**4.1 Linear Regression**

Linear regression is a supervised learning algorithm for predicting continuous values by modeling a linear relationship between independent variables and a dependent variable. It minimizes the difference between predicted and actual values by estimating coefficients during training. It's widely used in fields like economics and finance for tasks such as sales forecasting and price prediction. Implementation in Python is straightforward using libraries like scikit-learn. The basic equation involves coefficients representing the weights assigned to each feature. Its simplicity and interpretability make it popular for regression tasks. Model performance is assessed using metrics like mean squared error (MSE) or coefficient of determination (R-squared). Overall, linear regression provides a simple yet effective approach for continuous value prediction.

* from sklearn.linear\_model import LinearRegression
* lr = LinearRegression()
* lr.fit(X\_train, y\_train)
* predictions = lr.predict(X\_test)

**4.2 K-Nearest Neighbor algorithm:**

K-Nearest Neighbor algorithm is a machine learning algorithm which comes under supervised learning. This is used for both classification and regression. This algorithm is non parametric. This is also called as lazy learning algorithm. This algorithm works by first selecting the k value which is an integer value and less than the number of rows. When a new data point is given, KNN finds the nearest neighbors to that data point based on the distance using various methods like Euclidean distance or Manhattan distance. And assigns the data point to that class.

* from sklearn.neighbors import  KNeighbors Classifier
* classifier=KNeighborsClassifier(n\_neighbors=5,metric='minkowski',p=2)
* classifier.fit(x\_resem\_train,y\_resem\_train

**4.3Naive Bayes algorithm:**

# Naive Bayes algorithm is a machine learning algorithm which comes under supervised learning. This is used for both classification and regression.This algorithm is non parametric. This algorithm works based on the bayes theorem. Naive Bayes algorithm is a probabilistic classifier. It predicts the probability of an object. And also it does not require much training data.

* from sklearn.naive\_bayes import GaussianNB
* gnb=GaussianNB()
* gnb.fit(x\_resem\_train,y\_resem\_train)

# 4.4 Desicion Tree algorithm:

# Decision tree algorithm is a machine learning algorithm which comes under supervised learning. This is used for both classification and regression problems. This algorithm is also known as ID3 algorithm. This algorithm is non parametric method. It forms a tree from the given dataset. It has two nodes decision nodes and leaf nodes. Decision nodes are used for taking decisions and leaf nodes are the output of that decisions. The attribute selection happens by entropy and information Gini.

* from sklearn.tree import DecisionTreeClassifier
* classifier=DecisionTreeClassifier(criterion='entropy',random\_state=0)
* mm=classifier.fit(x\_resem\_train,y\_resem\_train)

# 4. 5support vector machine algorithm:

# Support vector machine algorithm is a machine learning algorithm which comes under supervised learning. This is used for both classification and regression problems. SVM works by constructing a hyperplane or a line that separates the different classes of data points. SVM has support vectors. The distance between positive hyperplane and negative hyperplane is called margin. from sklearn.svm import SVC

* svm\_model=SVC(kernel='linear')
* svm\_model.fit(x\_resem\_train,y\_resem\_train)

**4.6 RANDOM FOREST Algorithm:**

An ensemble learning method for regression and classification applications is called Random Forest. Decision Trees are used by the algorithm. They are made up of a number of separate binary trees that have been randomly trained on different subsets of data. Even if each of these trees may have been overtrained on its own, the randomness of the training process causes the trees to generate separate estimates, which are then added together to yield a conclusion. It has been demonstrated that Random Forests work well in a variety of classification and regression issues. When the number of trees in a forest increases, the generalization error for forests converges asymptotically to a limit. The strength of each individual tree in the forest of Decision Tree Regressors determines the generalization error of the forest.

**4.7 ADABOOST:**

AdaBoost is an ensemble learning method that combines multiple weak learners, like simple decision trees, into a strong classifier. It sequentially trains each weak learner, focusing more on the data points that were previously misclassified. By giving more weight to these difficult instances, AdaBoost improves overall accuracy. The final prediction is a weighted sum of the weak learners' outputs, where each learner's weight depends on its accuracy. This technique often results in highly accurate models for both classification and regression tasks.

# 5.DATASETPREPROCESSING:

# DATASET DESCRIPTION

# Attributes:

# customer name

# customer email

# country

# gender

# age

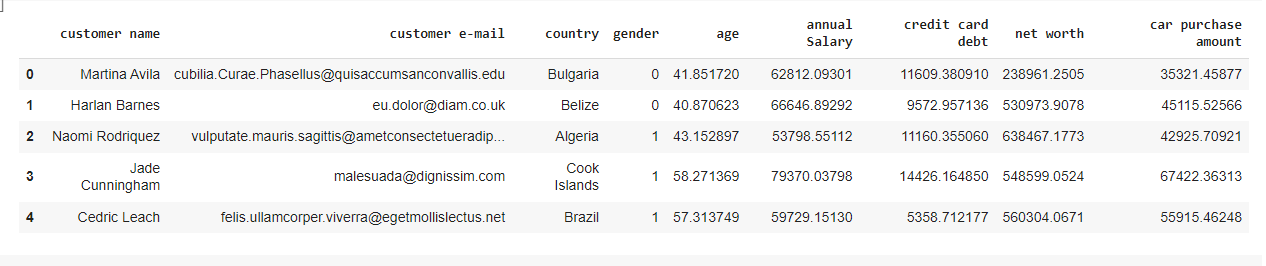
# annual salary

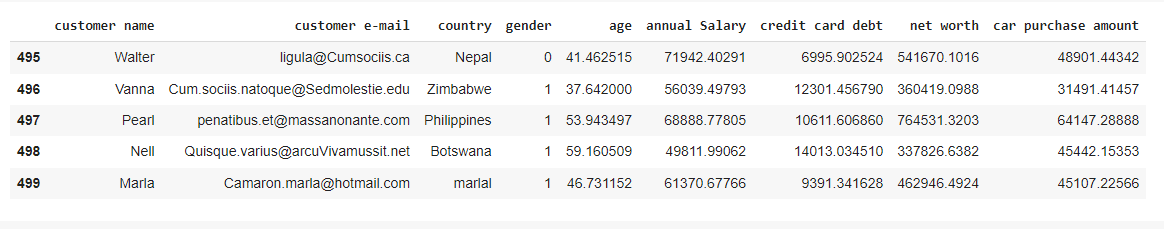
# credit card det

# net worth

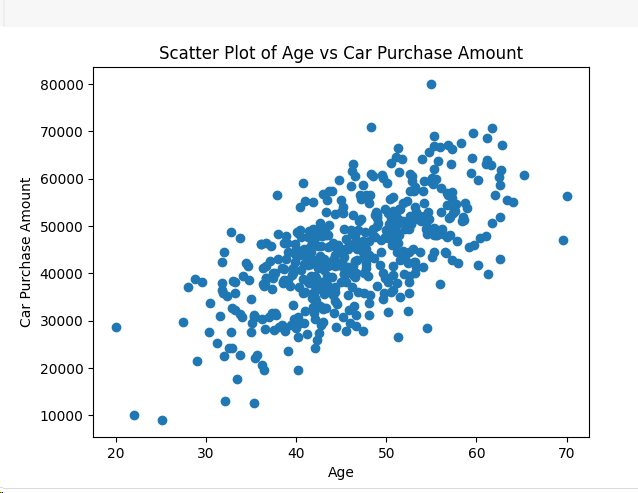
# car purchase amount

**Dataset**

****

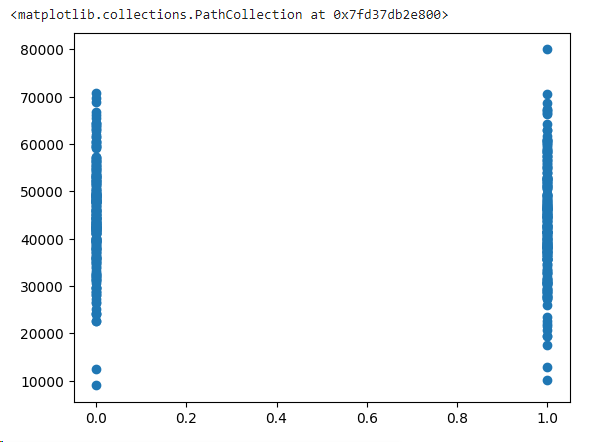
****

**GRAPHS PLOTTED BETWWEN FEATURE AND TARGET VARIABLES:**



**Explanation:** This graph illustrates that people aged between 40 to 60 purchased more no of cars between the range of 30000 to 70000.

**GENDER VS CAR PURCHASE AMOUNT:**



**Explanation**: This graph illustrates car purchase amounts by gender, using separate bars for males and females (and potentially other genders).from the above graph we could conclude That men spent more.

# 

# Explanation:

# This scatter plot shows the relationship between annual salary and car purchase amount. people having Annual salary between 50000 to 80000 purchased more no of cars within range of 30000 to

# 700000.

# 

# Explanation:

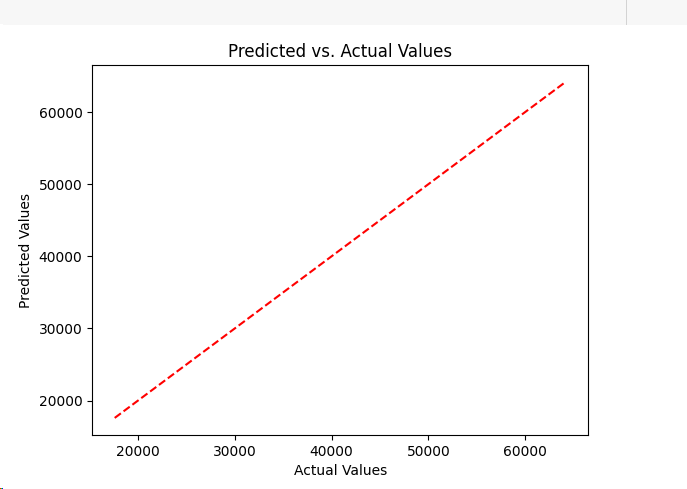
# This graph illustrates that people having networth between 0.2-0.6 are interested in purchasing cars in range of amount 30000 to 60000.

# 6. RESULTS:

**CODE**

**Dataset:**

**Linear Regression:**

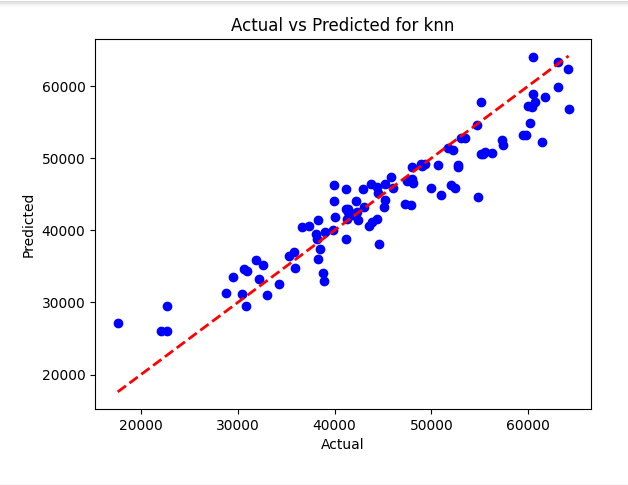


**Explanation:**

this graph Illustate that This scatter plot visualizes a linear regression model's performance in predicting car purchase amount. Ideally, data points scatter randomly around a diagonal line from bottom left to top right, indicating accurate predictions. Points above the line represent underestimations, while those below overestimations. A wider spread suggests higher prediction variability, essential for understanding the model's accuracy and inherent errors. This lack of alignment suggests that the model may not accurately capture the relationship between the features and the target variable.

The diagonal line in a scatter plot represents the predicted purchase amount. A positive slope indicates a higher purchase amount with increasing independent variable val

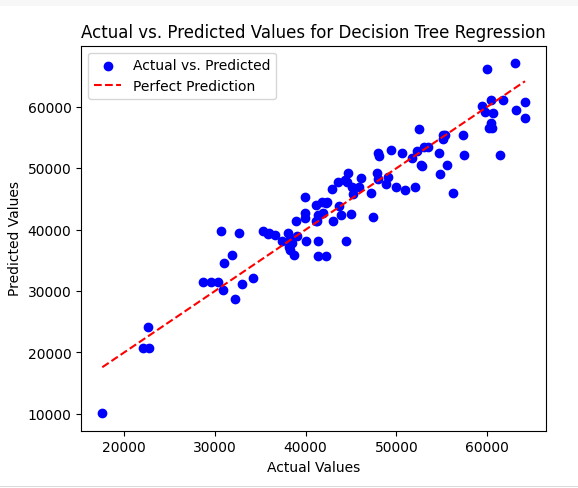
**K NEAREST NEIGHBOUR:**

****

**EXPLANATION:**

The graph shows actual versus predicted car purchase amounts using KNN. Points clustered around the diagonal line suggest accuracy, while scattered points indicate less precision. Model accuracy is assessed by how tightly points cluster around the diagonal line. In the above graph points are not so close to the line so it is less accurate. KNN relies on nearest neighbors for predictions, not extending the line. between 40000 to 50000 on both actual values(x-axis) and predicted values (y-axis) datapoints are so close to the line, this shows that only this part of predictions are correct and accurate.

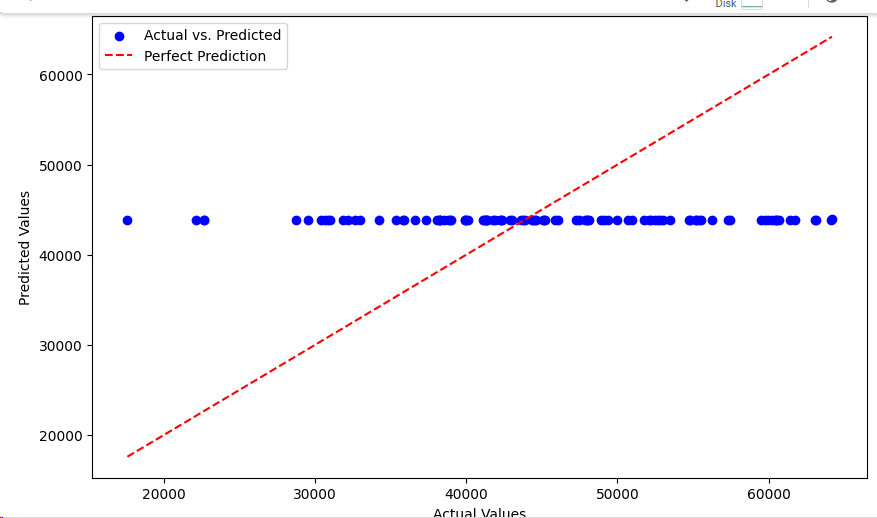
**Decision tree:**

****

**EXPLANATION:**

The structure of the graph provides insights into the relationship between data points and predicted price ranges. When data points tightly cluster within predicted price categories, closely aligning with actual purchase amounts, it suggests a well-performing model. In the above graph within the predicted range data points are less . wide dispersion of data points across various predicted ranges signifies lower model precision. Such scattering implies that the decision tree might struggle to capture the nuanced relationships within the data.

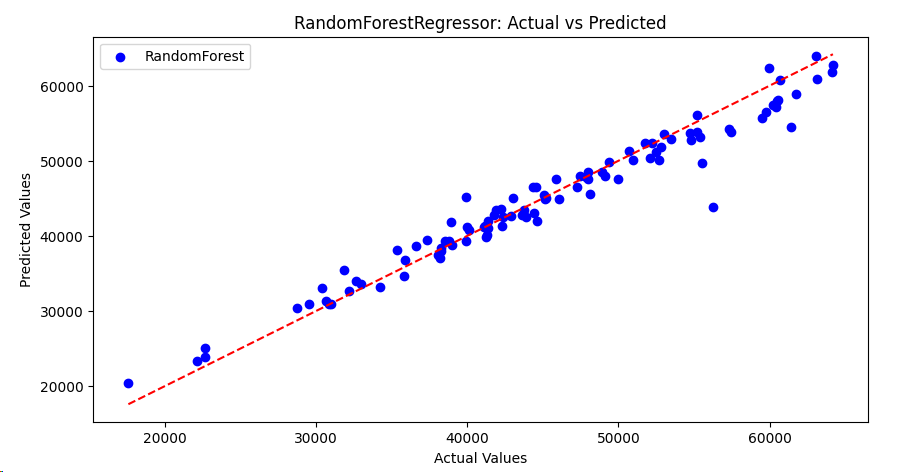
**Support Vector machine:**

****

**ERXPLANATION:**

## This is a line drawn by the SVM model to best separate the data points into two classes (e.g., high purchase amount vs. low purchase amount).If the data points cluster tightly around the separation line, it suggests the SVM model is making good predictions. So in the above graph datapoints are on the line but far from the line in between predicted values 40000-50000,this shows it is not best fit.The predicted purchase amounts are not too close to the actual values

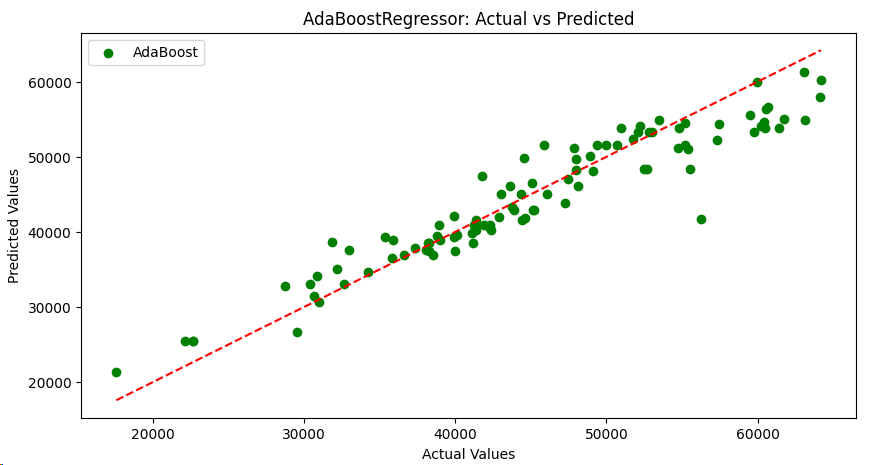
**Random Forest Regressor:**

****

**EXPLANATION:**

This graph illustrates relation between actual and predicted values of car purchase amount.If the data points are concentrated in a specific area of the Y-axis, it suggests the model is making consistent predictions and capturing the general relationship between features and purchase amount .in the above graph points are connected to the line hence this shows that model is making accurate predictions . A large spread of data points indicates the model's predictions might be more scattered and less precise. Data points are not widespread proves this model is not less precise.

**AdaBoost Regressor:**

****

**EXPLANATION:**

AdaBoost, an ensemble method, doesn't generate a single line on a prediction graph. Data points represent individual car purchases plotted against predicted amounts. Tight clustering suggests consistent predictions, indicating a good model, while a wide spread may indicate less precision. Error metrics like mean squared error provide additional insights into model performance.

**COMPARISION**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.NO** | **MACHINE LEARNING MODEL** | **Mean square error** | **Mean absolute error** | **Root mean square** |
| 1 | Linear regression | **1747061761.1378946** | **40738.23251092467** | **41797.867901818805** |
| **2** | K-Nearest Neighbor | **13191565.563198535** | **2845.638711** | **3632.0194882735054** |
| **3** | Decision Tree | **13207881.480235** | **2878.5941712** | **3634.264916077967** |
| **4** | Support vector machine | **109624906.46186401** | **8418.546282827978** | **10470.191328808849** |
| **5** | Random forest | **5260674.502036504** | **1586.586976386999** | **293.6160319540** |
| **6** | AdaBoost | **12294036.2883615** | **2607.2126715088234** | **506.2852548475717** |

# 7. CONCLUSION:

In conclusion, I have selected the required used car prices dataset with needed features and parameters from Kaggle. Kaggle is an open source Machine learning and data science platform which offers data and notebooks for data scientists and data analysts. The required data is cleaned and pre-processed used machine learning techniques before applying any algorithm for predicting the price. Then after pre-processing and cleaning the data first we need apply train test split to keep the data into two parts for training and validation using train and test data respectively. Then we must apply a simple linear regression model and predict the output and test its test and train error then again, we need to train and test it with multiple linear regression model and validate its accuracies. Then i used adaboost and Random forest methods and knn, Decision tree, support vector machine methods for predicting the output of car price. At last, after comparing all the algorithms the best one to predict is **Random Forest** as it gave less error.

**8. FUTURE SCOPE :**

This work will focus on answering the research questions. They all entail a comparison of different ML algorithms for price prediction. This will be accomplished by sourcing and preparing a dataset on which all the algorithms can be trained on and compared fairly. The algorithms selected must therefore be similar enough for the same dataset to be used for all of them. This also means that no large optimization efforts on the dataset will be made to boost the performance, if these changes do not benefit the other models. Maximizing price prediction performance of any one algorithm in ways that do not offer better comparisons is outside the scope of this work.

# 9. REFERENCES:

<https://www.diva-portal.org/smash/get/diva2:1674070/FULLTEXT01.pdf>

<https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26612934.pdf>