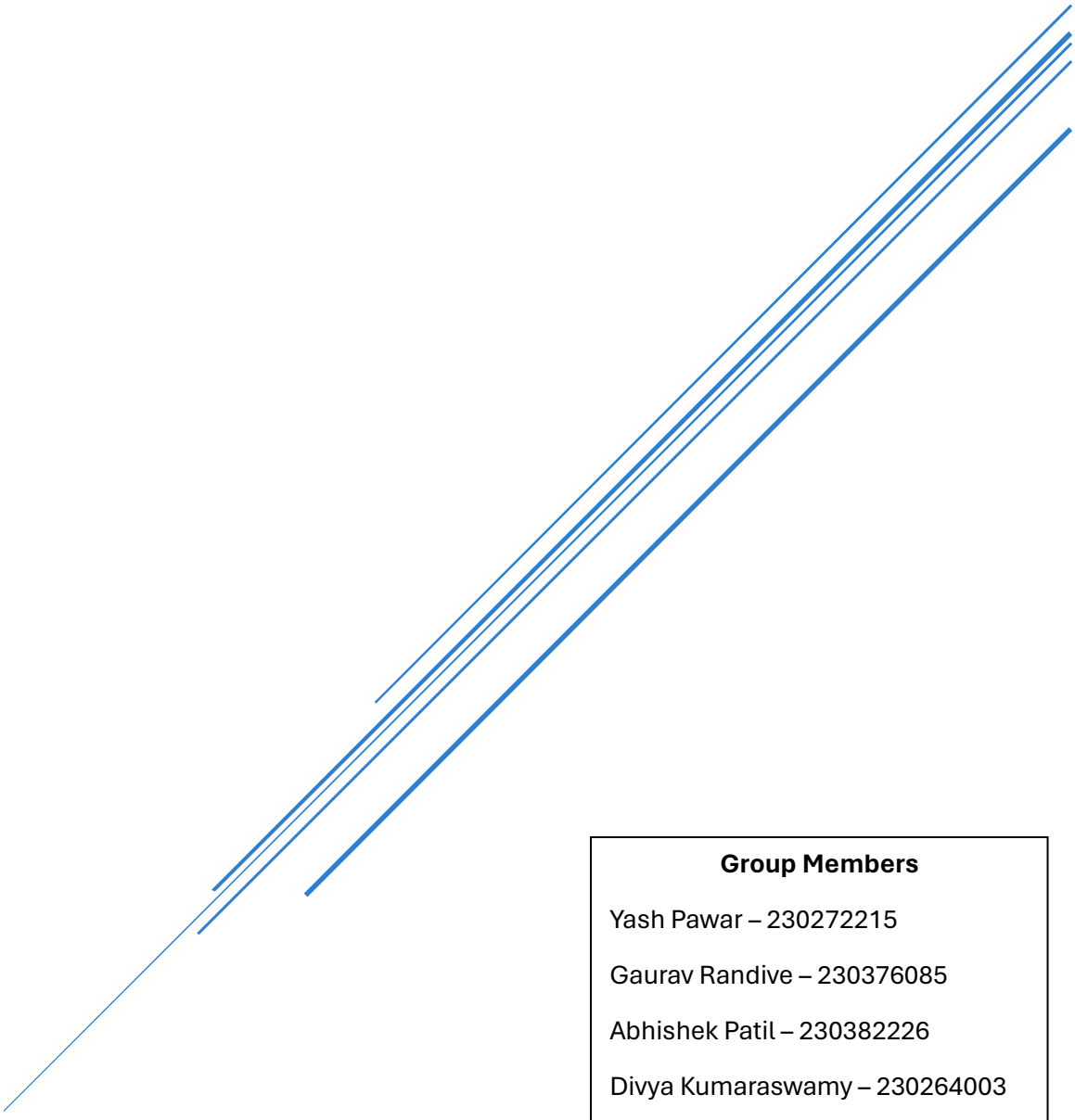
DATA MINING AND WEB ANALYTICS

USED CAR VALUATION TOOLS



Group Members

Yash Pawar – 230272215

Gaurav Randive – 230376085

Abhishek Patil – 230382226

Divya Kumaraswamy – 230264003

Nisha Jangir – 230355105

Table of Contents

Introduction

The Problem

Accurately projecting automobile pricing in the dynamic automotive market is a critical challenge that impacts a wide range of stakeholders, from individual consumers to major automotive firms. This is a complex issue that is impacted by a wide range of variables, such as market trends, economic situations, and vehicle specs. This study focuses on the CW dataset, a large collection of Cars-related data, to discover the most important predictors of car pricing to improve the prediction accuracy. In this study we have narrowed down the list of 33 potential predictors to the top 10 variables that have the biggest effects on car costs after doing an initial modelling analysis. These include the following: monthly_mileage, Cylinder_Numbers, Engine_Size, Color, Owners, Insurance, carwidth, Year, peakrpm, Credit_History. Notably, in our predictive model we have chosen "Price" as our target variable.

Stakeholders

Customers: Prospective buyers are directly impacted by forecasts on car prices. Accurate projections can ensure customers receive a just return on their investments and assist them in making educated decisions. **Automobile Companies:** Manufacturers and dealerships rely on accurate price estimates to set competitive prices, manage inventory, and develop marketing campaigns that resonate with their target market**. Financial Institutions:** Based on estimates of car pricing, banks and financing companies create insurance policies and auto loans. Accurate values are crucial for calculating loan-to-value ratios and assessing risk. **Market analysts:** Professionals that track developments in the automotive sector use price projections to advise clients, publish studies, and provide insights into potential future changes in the sector.

Significance of the Problem

Knowing a car's reasonable value helps buyers make wise financial decisions and could result in savings of thousands of dollars. Accurate pricing can mean the difference between an automobile company's leading position in the market and its trailing competitors. It affects brand impression, profitability, and stock prices. Financial institutions gain from optimized loan and insurance offers as well as decreased risk. Market analysts view accuracy in pricing prediction models as a sign of credibility and efficaciousness in predicting market trends. The goal of this research is to use the predictive capability of the chosen variables to give stakeholders information that can help with strategy creation and decision-making in the automobile industry.

Data Set and Visualization

Type & dimensions of the dataset.

The Dataset tabular and structured, which consists of records in which the features of cars are mentioned for e.g.: Model, Year, Price, Monthly Mileage etc. There are 2230 entries (Rows) and 32 Variables (Columns).

Variables, definitions, their types, and their roles

We have chosen the following 10 Predictors (Variables) for our study after performing a feature selection test on our Dataset. To get the most important variables, we did a Random Forest Modelling on the dataset and chose the top 10 Variables for our Study. Below table shows the list of variables, their Definitions, Type and Role.

Variable	Definition	Type	Role
monthly_mileage	The average monthly mileage of the car.	Float	Independent Variable/ Input
Cylinder_Numbers	The number of cylinders in the car's engine.	String	Independent Variable/ Input
Engine_Size	The size of the car's engine.	Float	Independent Variable/ Input
Color	The color of the car.	String	Independent Variable/ Input
Owner	Number of people who owned the car	Integer	Independent Variable/ Input
Insurance	Kind of Insurance of car	String	Independent Variable/ Input
carwidth	width of the car.	Float	Independent Variable/ Input
Year	The manufacture year of the car.	Integer	Independent Variable/ Input
peakrpm	Car peak revolutions per minute (RPM)	Integer	Independent Variable/ Input
Credit_History	Credit history of the previous owner.	Float	Independent Variable/ Input
Price	The price of the car.	Float	Dependent Variable/ Output

Verbal presentation

Honda CR-V (2004): This record contains information on a 2004 Honda CR-V, a dependable and useful small crossover SUV. It runs on petrol, fits in with the trend towards more conventional fuel sources, and has a manual gearbox that allows for engaged driving. With 4 doors, the SUV suggests adequate accessibility. Its 'Fair' condition reflects its age and use. It is brown, which could be appealing to anyone looking for a car that looks natural and modest. It is positioned as a more practical model, lacking luxury options like heated seats, leather upholstery, and cruise control. At about $6,553.60, it is a priced choice for consumers looking for a reliable and useful SUV. The car is appropriate for both city driving and longer trips because of its engine size and weight, which suggest a balance between performance and fuel efficiency.

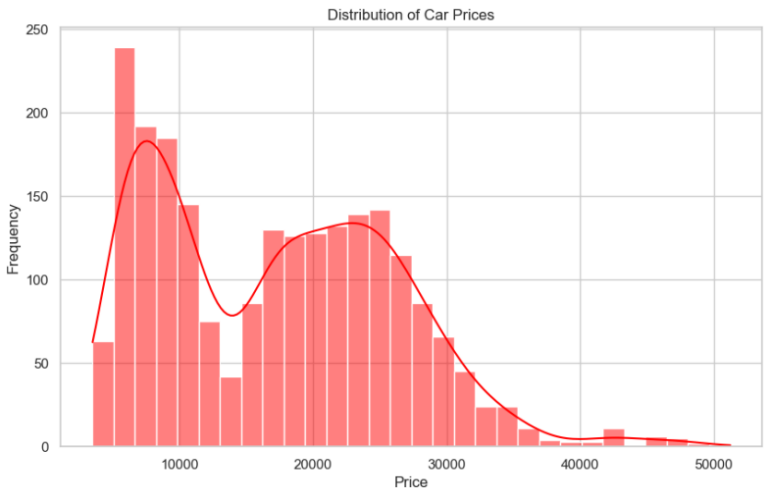
Toyota Corolla (2020): The Toyota Corolla 2020 Model has an electric engine with Manual gear transmission and 4 doors in a really good condition. It is brown in colour with cruise control, Leather & Heated seats, Navigation and only one owner. It has a collision kind of insurance with full-service history and 3-star safety with features like premium sound, Multimedia except Bluetooth. Speaking about the outer parts, it has a forged wheel with a sunroof at the top. With 3 Cylinders and a price of $24381.08 the Toyota corolla has an 8-year warranty and positive credit history score of 0.0152. The Engine size is 2.0443Litres, weighs around 3340 pounds and has a car length of 169.7 inches and a width of 65.2 inches. It has a monthly mileage of 1307 miles, with a peak RPM of 5720, and an estimated mileage of 12446.97 miles.

Level of the data

Each record in the dataset represents a single car's information, including its specifications, features, condition, and price. When estimating the price of a car these are the key aspects to be considered.

Uni-variate visualization and commenting

Univariate visualization for the Distribution of Car Prices



The histogram and kernel density line depict a right-skewed car price distribution, with most cars priced below the mean and a few significantly pricier ones. The tallest histogram bars indicate the most common price range, and the skewness is highlighted by a tail extending toward the higher prices. The density line smoothly outlines this skewness toward lower-priced cars.

Measures of Centrality and Spread:

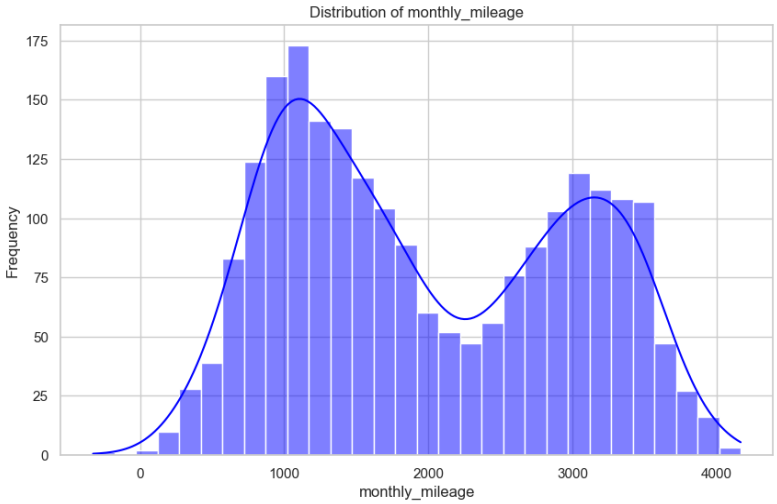
Price Mean = $17241.51: The mean is drawn towards the more expensive cars rather than at the top of the histogram due to the skewed distribution.

Price Median = $17393.38: It is typical of a right-skewed distribution that the median is marginally higher than the mean. Because it is less impacted by extreme values than the mean, it indicates a more "typical" pricing in this context.

Price Std = $8961.47: It represents the variance in car prices from the mean. The standard deviation suggests that there is a large variety of prices and that several car prices deviate significantly from the average.

Univariate visualization for the Distribution of monthly_mileage

The histogram and density plot display a bimodal distribution of cars' monthly mileage, indicating two primary usage groups. The distribution is asymmetrical, suggesting varied driving habits among the car owners.



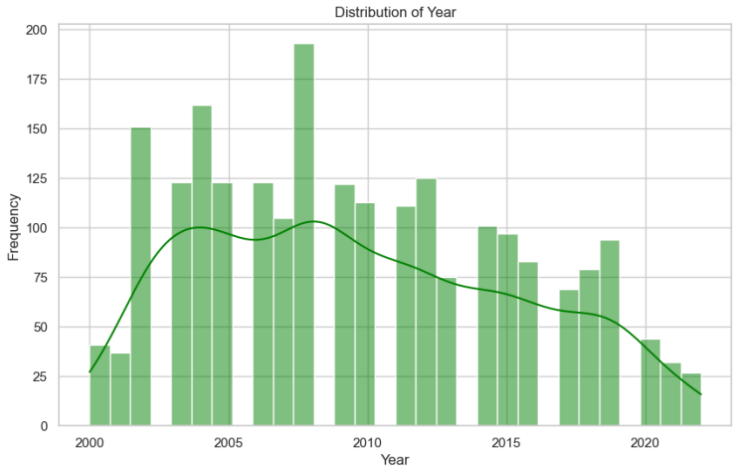
Measures of Centrality and Spread:

mean monthly mileage = 1986.59: The monthly mileage average is approximately 1986.59 miles. Although this is a measure of central tendency, it might not fairly depict the most common values in a bimodal distribution.

monthly_mileage median = 1758.51: At 1758.51 miles, the monthly mileage is less than the mean. For skewed or bimodal distributions, the value that divides the higher half from the lower half of the mileage data is frequently seen as a more accurate indicator of central tendency.

monthly_mileage std = 994.40: The monthly mileage numbers deviate significantly from the average, as seen by this value, which shows a significant dispersion around the mean.

Univariate visualization for the Distribution of Year



The histogram and kernel density curve reveal a multi-peaked distribution of car manufacturing years, highlighting surges in production or sales, particularly in the mid-2000s and near 2010. The spread of years shows a mix without a predominant single year of manufacturing.

Measures of Centrality and Spread for the Year:

Year Mean (2009.70): The cars were manufactured on average in 2009 or earlier. This implies that, on average, the cars in the sample date back to the late 1990s and early 2000s.

Year Median (2009.0): 2009 is also the median manufacture year, indicating that half of the vehicles are 2009 or older. There is less of an impact from outliers or data skewness on this central tendency measure.

Year Standard Deviation (5.68): The fact that the standard deviation is so low suggests that most of the years when cars were manufactured are concentrated around the mean. Less fluctuation in the years the cars were manufactured is indicated by a smaller standard deviation.

Bi-variate visualization and commenting (target vs predictor)

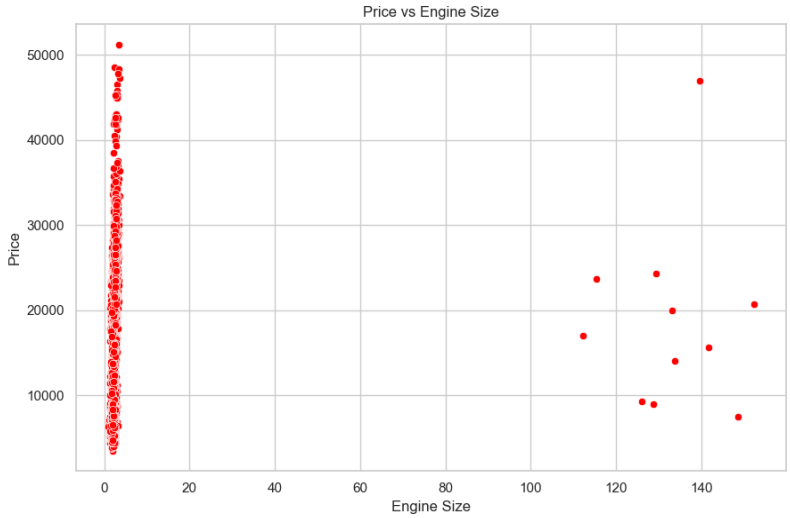
Price vs monthly Mileage

The association between car prices and monthly milage is depicted in this scatter plot. There seems to be a negative correlation: the car's price tends to go down as monthly mileage goes up. According to this tendency, cars with less mileage driven are worth more, which is consistent with the widespread belief that cars that are driven less often retain more of their value.

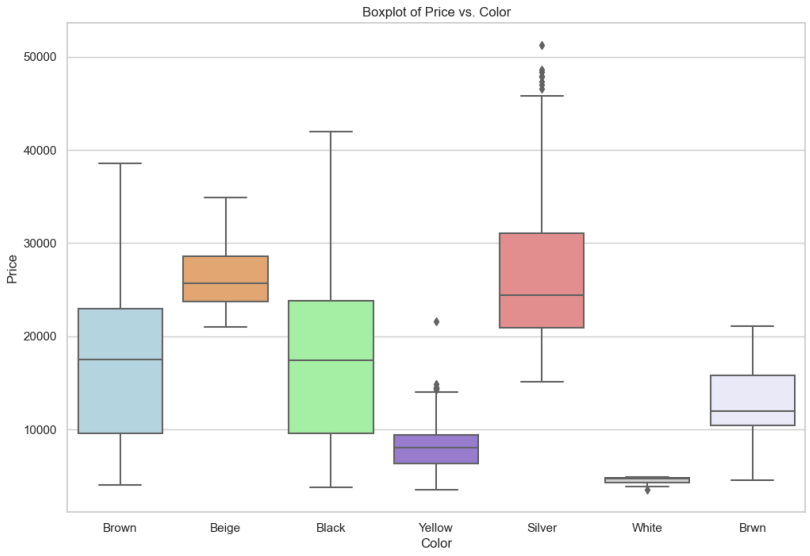
Monthly mileage is a significant predictor of car prices, with the scatter plot showing a consistent trend were price decreases with increasing mileage. Data points cluster more densely at lower mileages, indicating these have a greater impact on price. As mileage goes up, the data points disperse, suggesting additional variables may affect the price of high-mileage cars.

Price vs Engine Size

The plot shows car prices versus engine sizes, with a dense cluster at smaller engine sizes and sparse data for larger engines. This indicates engine size may influence price but is not the sole determinant, especially for cars with bigger engines where other factors could be at play.

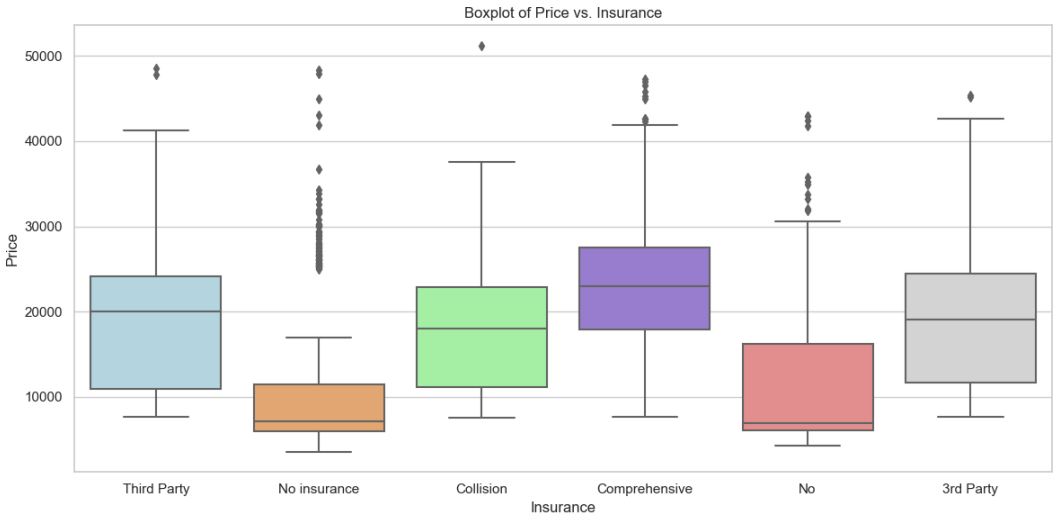


Boxplot for Price vs. Color



The boxplot displays car price differences by color, with black and brown commanding higher median prices and broader ranges, indicating a preference for these colors in higher-end cars. Yellow and white show the lowest median prices. Despite some outliers, especially in the yellow and silver categories, color influences car prices but is likely not a decisive factor without other car attributes.

Boxplot for Price vs. Insurance



The boxplot compares car prices across different insurance categories. Vehicles with collision insurance tend to have a higher median price and more varied pricing, indicating that pricier cars often carry this type of insurance. The lowest median price is observed in cars without any insurance, reflecting the lower value of these vehicles. Although insurance type does seem to relate to car price, with comprehensive and collision insurance associated with higher values, it is not an absolute indicator of price due to the considerable price range overlaps among the categories. Insurance type can be a partial indicator of a car's worth, but it should be combined with other variables for accurate price predictions.

Data Quality Assessment and treatment

Outliers and extremes

Definition: Outliers are data points that stand out markedly from the bulk of a dataset. They have the potential to skew analysis and results.

Treatment: We detected the outliers using the Interquartile Range (IQR) for numerical data and handled them according to their impact on the dataset.

Variable	No. of outliers
Engine Size	22
Owners	199
Credit History	20

Some of the numbers in the owner's column were 22 and 44. We attributed it to human mistake by taking the factors' importance into account. We therefore changed 22 to 2 and 44 to 4, as that could be the solution given human error. We used the capping and flooring procedure to eliminate the outliers for Engine size and Credit history.

Missing values

Definition: Missing values arise when data for a particular variable is absent in an observation.

Treatment: Missing values were addressed during the initial data preparation phase. It's important to evaluate how the chosen method of filling in these values affects the performance of the model. We examined all variables for missing entries and found that only the 'Color' variable had two missing entries. We imputed these missing values by using the most frequent color in the dataset.

Predictive Modelling Formulation

By creating a model that accepts the given data as input and produces a valuation (price) for the car as output, we can forecast the value of used cars depending on the variables supplied.

In our predictive modelling formulation, we will be training and testing several models, including the Random Forest, Decision Tree, and Linear Regression.

Type of the problem

In this report, our target variable "Price" is identified as a continuous, numeric variable, indicating that the problem is a regression problem. This falls under the category of supervised learning, as the target variable is predetermined and provided in the dataset.

Data Partitioning

The data is divided such that 80% is used for training and 20% for testing. Initially, the model learns from the training data, and then its performance is assessed using the test data.

Performance metrics

We will be assessing the performance of our model using the following performance metrics.

- Mean Absolute Error (MAE): This metric averages out the absolute differences between the forecasted and true values, offering an easy-to-understand gauge of prediction accuracy.
- Root Mean Square Error (RMSE): This measure takes the square root of the average squared deviations between the predicted and actual figures. It tends to highlight larger errors more than MAE.
- R-squared (R^2): This statistic indicates the proportion of variance in the dependent variable that is predictable from the independent variable(s).

When selecting a metric, the specific needs and context of the issue at hand should guide the decision. For this situation, RMSE might be preferable because it amplifies and penalizes larger discrepancies between predicted and actual car prices, which is particularly critical in the context of pricing used vehicles.

Baseline Model

The variables were fed into the Linear Regression model, which was used as the baseline Model. The model predicted the price with an **RMSE** of about **2239** on the testing set. An 80:20 split of the data was made into training and testing datasets. It offers an understandable, basic model that can be used as a benchmark for more intricate models.

Models Evaluation

Model	RMSE	R-square
Linear Regression	2239	93%
Random Forest	1731	96%
Decision Tree	2417	92%

Feature engineering efforts

The initial steps of feature engineering involved simple preprocessing tasks like converting categorical variables into numerical format and scaling numerical variables. There was no implementation of advanced feature engineering techniques.

Improving the model's performance

SUMMARY TABLE

Model	Hyper-parameters	Changes in Features	Performance
Linear Regression	Default	Training = 70%, Testing = 30%	RMSE = 2548, R-square = 93%
Random Forest	**Default**	**Small grid – n_estimators = [100,200] , depth = [10,20]**	**RMSE = 1911, R-square = 96%**
Decision Tree	Default	Default	RMSE = 2417 R-square = 92%

Introduction of supervised/unsupervised methods

Linear Regression

Linear regression is a fundamental predictive analysis method that investigates how well a set of independent variables can predict a dependent variable and identifies which specific variables significantly influence the outcome. (Solutions, n.d.)

Random Forest

Random forest ML is an ensemble of decision trees that repeatedly dichotomize a data set based on their determination of the most informative feature. (Canadian Journal of Cardiology, n.d.)

Decision Tree

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical tree structure, which consists of a root node, branches, internal nodes, and leaf nodes. (IBM, n.d.)

Error Cost analysis

Errors in predicting car prices can significantly impact all parties involved:

Underestimating Car Prices: Sellers may incur financial losses, and buyers might question the car's condition.

Overestimating Car Prices: Can repel buyers, extend the time to sell, and cause financial institutions to offer loans on inflated values, heightening default risks.

The Random Forest model, demonstrating the smallest RMSE and highest R^2, effectively reduces both overestimation and underestimation errors, indicating it aligns well with the diverse interests of all stakeholders.

Conclusion

The aim of this project is to develop tools that will enable used car sellers, purchasers, automakers, and financial institutions to all have their demands met. Ten important variables, such as mileage, engine size, and car colour, were found to be significant predictors of car pricing through the examination of a large dataset. Models such as Random Forest, Decision Tree, and Linear Regression were used in the study; Random Forest performed the best because of its lowest error rates and better balance of interests among stakeholders.

Non-technical Summary

In simple terms, this project explored the factors that influence used car prices, such as mileage, Engine size, and color etc. by analyzing data from numerous car records. The goal was to discover what contributes to a car's market value. Various prediction techniques were evaluated, and the Random Forest method, which integrates multiple smaller decisions, emerged as the most effective in accurately forecasting car prices.

Next Steps for Improvement

- Gathering Additional Information: Enhancing predictions with more comprehensive data, such as maintenance history, could refine accuracy.
- Experimenting with Advanced Techniques: Investigating cutting-edge data analysis methods could yield deeper insights.
- Adapting Models for Unique Requirements: Creating specialized models to cater to specific groups, like devising more accurate valuation models for luxury vehicles, could benefit niche markets.

New Ideas for Relevant Projects

Predictive Maintenance Models: Like car price predictions, models could be created to forecast maintenance needs, preventing unexpected failures and expensive fixes.

Car Depreciation Models: Tools could be developed to estimate how car values decline over time, aiding buyers in selecting vehicles that depreciate less.

Environmental Impact Analysis: By examining how various car types affect the environment, a guide could be created to assist buyers in choosing more environmentally friendly options.

This project lays the groundwork for informed decision-making in the used car market, with possibilities for further improvements to car buying and ownership experiences.

Referencing

Statistics Solutions (2013). What is Linear Regression? [online] Statistics Solutions. Available at: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/.

www.sciencedirect.com. (2022). Random Forest - an overview | ScienceDirect Topics. [online] Available at: https://www.sciencedirect.com/topics/nursing-and-health-professions/random-forest.

IBM (2023). What Is a Decision Tree | IBM. [online] www.ibm.com. Available at: https://www.ibm.com/topics/decision-trees.

Machine Learning Models - https://scikit-learn.org/stable/supervised_learning.html#supervised-learning