Analyzing Selling Price of Used Cars

A PROJECT REPORT

Submitted by

ARVIND P (210701324)

KESARIKUMARAN S (210701324)

in partial fulfillment for the course

CS19643 FOUNDATIONS OF MACHINE LEARNING

for the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

RAJALAKSHMI NAGER

CHENNAI – 602105

JUNE 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Thesis titled **"ANALYZING THE SELLING PRICE OF USED CARS"** is the bonafide work of **"KESARIKUMARAN S (210701324), ARVIND P (210701324)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported here in does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr.S. Vinod Kumar,

PROJECT COORDINATOR

Professor

Department of Computer Science and Engineering

Rajalakshmi Engineering College

Chennai – 602 105

Submitted to project Viva-Voce Examination held on_____

Internal Examiner **External Examiner**

ACKNOWLEDGMENT

First, we thank the almighty god for the successful completion of the project. Our sincere thanks to our chairman **Mr. S. Meganathan B.E., F.I.E.,** for his sincere endeavor in educating us in his premier institution. We would like to express our deep gratitude to our beloved Chairperson **Dr. Thangam Meganathan Ph.d.,** for her enthusiastic motivation which inspired us a lot in completing this project and Vice Chairman Mr. Abhay Shankar Meganathan B.E., M.S., for providing us with the requisite infrastructure.

We also express our sincere gratitude to our college Principal, **Dr. S. N. Murugesan M.E., PhD.,** and **Dr. P. KUMAR M.E., PhD, Director computing and information science , and Head Of Department of Computer Science and Engineering** and our project coordinator **Dr. K.Ananthajothi M.E.,Ph.D.,** for her encouragement and guiding us throughout the project towards successful completion of this project and to our parents, friends, all faculty members and supporting staffs for their direct and indirect involvement in successful completion of the project for their encouragement and support.

ARVIND P (210701034)

KESARIKUMARAN S (210701324)

ABSTRACT

In today's technologically advanced landscape, machine learning is extensively utilized across various sectors to extract meaningful insights and make accurate predictions from data. This project focuses on analyzing the selling price of used cars using Python, employing machine learning algorithms to predict prices based on historical data. The success of these models hinges on thorough data analysis, which involves cleaning, organizing, and preprocessing the dataset to ensure it is suitable for training. Proper data preparation includes handling missing values, encoding categorical variables, and normalizing numerical features. This step is crucial as poorly organized data can lead to inaccurate predictions, resulting in potential financial losses for businesses. Accurate predictions not only aid in making informed business decisions but also enhance sales strategies and customer satisfaction. This project highlights the significance of integrating meticulous data analysis with machine learning to drive operational efficiency and success in the automotive industry. Additionally, the project will explore various machine learning algorithms, such as linear regression, decision trees, and ensemble methods, to identify the most effective approach for predicting car prices. Model evaluation metrics, including mean absolute error and root mean square error, will be used to assess the performance of the models and ensure their accuracy. By continuously refining the models through techniques like cross-validation and hyperparameter tuning, the project aims to achieve the highest possible predictive accuracy. Ultimately, this comprehensive approach underscores the transformative potential of machine learning in making data-driven decisions and optimizing operations within the used car market.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATION

1. df: DataFrame
2. np: NumPy
3. plt: Matplotlib.pyplot
4. sns: Seaborn
5. sp: SciPy
6. CSV: Comma-Separated Values
7. MAE: Mean Absolute Error
8. RMSE: Root Mean Square Error
9. R2: R-squared
10. NLP: Natural Language Processing
11. CNN: Convolutional Neural Network
12. RNN: Recurrent Neural Network
13. ANOVA: Analysis of Variance
14. HTML: Hypertext Markup Language
15. CSS: Cascading Style Sheets
16. API: Application Programming Interface
17. JSON: JavaScript Object Notation
18. URL: Uniform Resource Locator
19. SSL: Secure Sockets Layer
20. HTTP: Hypertext Transfer Protocol

CHAPTER 1

INTRODUCTION

1.1 RESEARCH PROBLEM

The primary research problem addressed by this project is the accurate prediction of the selling price of used cars using machine learning techniques. With the exponential growth in the used car market, understanding and predicting car prices have become increasingly important for both consumers and businesses. The prices of used cars are influenced by a multitude of factors, including the car's make and model, age, mileage, condition, location, and market trends. The challenge lies in developing a predictive model that can effectively capture the relationships between these variables and accurately forecast the selling price.

The first objective of the project is data collection and preparation. This involves gathering a comprehensive dataset containing historical data on used car sales. The dataset needs to be meticulously preprocessed to handle missing values, outliers, and inconsistencies, which are common issues in real-world data. Proper preprocessing ensures the dataset's quality and suitability for machine learning models. Additionally, categorical variables must be encoded, and numerical features normalized to standardize the data and facilitate effective model training.

Following data preparation, the project will engage in exploratory data analysis (EDA) to uncover the underlying patterns and relationships within the data. EDA involves visualizing key features and examining

their correlations with the selling price. This step is crucial for identifying the most influential factors affecting car prices and for gaining insights into the dataset's structure and characteristics. Understanding these relationships helps in selecting appropriate features for the predictive model and enhances the overall modeling process.

Subsequently, the project will involve building and evaluating various machine learning models. The focus will be on exploring different algorithms, such as linear regression, decision trees, and ensemble methods, to determine the most effective approach for predicting car prices. Each model's performance will be assessed using evaluation metrics like mean absolute error (MAE) and root mean square error (RMSE) to ensure their accuracy. By continuously refining the models through techniques like cross-validation and hyperparameter tuning, the project aims to achieve the highest possible predictive accuracy.

1.2 PROBLEM STATEMENT

The primary research problem addressed by this project is the challenge of accurately predicting the selling price of used cars using machine learning techniques. In the rapidly growing used car market, accurately determining car prices is critical for both consumers and businesses. The complexity arises from the myriad of factors influencing car prices, such as the vehicle's make and model, age, mileage, condition, location, and prevailing market trends. Developing a predictive model that can effectively capture and analyse these diverse variables is essential for generating reliable price estimates. The core issue lies in ensuring the model's accuracy and robustness, which necessitates meticulous data

collection, preprocessing, and analysis to handle real-world data's inherent inconsistencies and variations. Addressing this problem not only aids in making informed business decisions and enhancing sales strategies but also improves customer satisfaction by providing transparent and accurate pricing information.

1.3 SCOPE OF WORK

The scope of this project encompasses the entire pipeline for developing a machine learning model to predict the selling price of used cars, starting from data collection and preprocessing to model development, evaluation, and deployment. Initially, the project involves gathering a comprehensive dataset containing historical sales data, which includes various features such as make and model, year of manufacture, mileage, condition, location, and selling price. This raw data will undergo thorough preprocessing to address missing values, outliers, and inconsistencies, ensuring it is clean and suitable for analysis. The preprocessing step also involves encoding categorical variables and normalizing numerical features to standardize the dataset for effective model training. Following data preparation, exploratory data analysis (EDA) will be conducted to uncover underlying patterns, relationships, and correlations within the data, helping to identify the most influential factors affecting car prices. This insight will guide the selection of relevant features for the predictive model. Subsequently, various machine learning algorithms, including linear regression, decision trees, and ensemble methods, will be explored to determine the most effective approach for predicting car prices.

1.4 AIM AND OBJECTIVE

The aim of this project is to develop a robust machine learning model for accurately predicting the selling price of used cars, addressing the complex and multifaceted factors influencing car prices. The objectives encompass comprehensive data collection, meticulous preprocessing, insightful exploratory data analysis, rigorous model development, thorough evaluation, and practical deployment. By achieving these objectives, the project aims to provide valuable insights into the used car market, empower businesses with effective pricing strategies, enhance customer satisfaction, and exemplify the transformative potential of machine learning in optimizing operations within the automotive industry.

1.5 RESOURCES

The resources required for this project include access to relevant datasets containing historical sales data of used cars, preferably sourced from reputable automotive databases or online platforms. Additionally, access to computing resources such as a robust workstation or cloud computing services is essential for data preprocessing, exploratory data analysis, and model development. Software tools like Python programming language, along with libraries such as Pandas, NumPy, and scikit-learn, will be utilized for data manipulation, analysis, and machine learning model implementation. Furthermore, visualization tools like Matplotlib and Seaborn will aid in data visualization and interpretation. Collaboration tools and version control systems such as Git/GitHub will facilitate team collaboration and code management throughout the project lifecycle. Access to relevant research papers,

documentation, and online resources will also be essential for staying updated on the latest methodologies and best practices in machine learning and data analysis. Overall, these resources will enable the efficient execution of the project tasks and contribute to achieving the project's objectives effectively.

1.6 MOTIVATION

The motivation behind this project stems from the increasing significance of accurate pricing in the used car market, driven by the growing demand for transparent and data-driven decision-making processes. Inaccurate pricing can lead to financial losses for sellers and dissatisfaction among buyers, highlighting the need for robust predictive models to estimate car prices effectively. Furthermore, the project seeks to address the challenges posed by the dynamic nature of the automotive industry, where factors such as market trends, vehicle condition, and regional variations constantly influence prices. By leveraging machine learning techniques, this project aims to provide stakeholders with reliable insights into pricing dynamics, enabling them to make informed decisions, optimize sales strategies, and enhance customer satisfaction. Ultimately, the project's motivation lies in leveraging advanced data analytics to tackle real-world challenges, drive operational efficiency, and contribute to the ongoing transformation of the automotive sector.

CHAPTER 2

LITERATURE REVIEW

In The literature review for a project on analysing the selling price of used cars using machine learning encompasses various key aspects, including existing methodologies, relevant studies, and advancements in the field. Research in this domain has focused on exploring different machine learning algorithms, feature engineering techniques, and data preprocessing methods to improve the accuracy of car price prediction models.

One area of interest in the literature is the selection and engineering of features that significantly impact car prices. Studies have identified factors such as make and model, age, mileage, condition, and geographic location as influential determinants of a car's value. For instance, research by Lim and Kim (2018) highlights the importance of considering vehicle condition and mileage as critical predictors of resale value. Similarly, Kim and Ryu (2019) emphasize the significance of incorporating regional market trends and economic indicators into predictive models to enhance their accuracy.

Another aspect of interest is the choice of machine learning algorithms for price prediction tasks. Linear regression, decision trees, random forests, support vector machines, and neural networks are among the commonly used techniques in this context. Research by Yilmaz and Oztaysi (2017) compares the performance of various machine learning algorithms for predicting car prices, demonstrating the superiority of ensemble methods such as random forests. Additionally, studies by Chen et al. (2020) and Khan et al. (2021) explore the effectiveness of deep learning approaches, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), in capturing complex patterns in car sales data.

Moreover, literature in this field also addresses challenges related to data preprocessing and model evaluation. Techniques for handling missing values, outliers, and skewed distributions in the dataset have been proposed to improve model robustness and generalization. Furthermore, studies have proposed novel evaluation metrics, including mean absolute percentage error (MAPE) and root mean squared logarithmic error (RMSLE), to assess model performance accurately and account for heteroscedasticity in car price data (Singh et al., 2020). In addition to algorithmic approaches and feature engineering techniques, the literature also delves into the role of market dynamics and external factors in influencing car prices. Researchers have explored the impact of macroeconomic indicators, such as inflation rates, interest rates, and GDP growth, on the demand and supply dynamics of the automotive market. For instance, studies by Zhang et al. (2019) and Li et al. (2020) analyze the relationship between economic indicators and car prices, highlighting the importance of incorporating external variables into predictive models to account for market fluctuations and uncertainties. Understanding these broader economic trends can provide valuable context for predicting car prices accurately and adapting pricing strategies to changing market conditions.

Furthermore, advancements in data collection and aggregation techniques have enabled researchers to leverage large-scale datasets and real-time information sources for price prediction tasks. The proliferation of online automotive marketplaces and auction platforms has facilitated access to rich and diverse datasets containing detailed information about used car listings, including photos, descriptions, and transaction histories. Studies by Wang et al. (2018) and Zheng et al. (2021) demonstrate the utility of web scraping and data mining techniques for collecting and preprocessing data from online sources, enabling the development of predictive models with improved accuracy and granularity.

Moreover, the literature review also explores the implications of predictive modeling for various stakeholders in the automotive ecosystem. For car dealerships and retailers, accurate price predictions can inform pricing strategies, inventory management decisions, and sales forecasting, ultimately enhancing profitability and competitiveness. Similarly, consumers benefit from transparent pricing information, enabling them to make informed purchasing decisions and negotiate fair deals. Additionally, financial institutions and insurance companies leverage predictive models to assess the risk associated with financing used car purchases and pricing insurance premiums. By synthesizing insights from these studies, this project aims to contribute to the development of robust and reliable predictive models for estimating the selling price of used cars, thereby empowering stakeholders with actionable insights and enhancing efficiency in the automotive marketplace.

In summary, the literature review highlights the diverse methodologies and approaches employed in predicting the selling price of used cars using machine learning techniques. By synthesizing insights from existing studies, this project aims to contribute to the advancement of predictive modelling in the automotive industry and provide valuable insights for stakeholders involved in the buying and selling of used vehicles.

2.1 EXISTING SYSTEM

The existing system for analyzing the selling price of used cars traditionally relies on manual appraisal by experts or rule-based pricing systems. Manual appraisal involves subjective assessments by experienced car dealers or appraisers, considering factors such as the vehicle's make and model, age, mileage, and condition. While this method may provide reasonably accurate estimates, it is time-

consuming and susceptible to human biases. Alternatively, rule-based systems use predefined algorithms to determine prices based on set criteria, offering more objectivity but limited adaptability to changing market conditions.

However, both manual appraisal and rule-based systems have limitations in accurately predicting car prices, especially in the context of a dynamic used car market. They may overlook nuanced patterns or interactions between variables, leading to suboptimal pricing decisions. Moreover, they struggle to handle large datasets and lack the sophistication to leverage advanced analytical techniques effectively. In response, there's a rising interest in leveraging machine learning and data-driven approaches to enhance price prediction accuracy.

By harnessing machine learning algorithms and sophisticated modeling techniques, such as regression, decision trees, and neural networks, businesses can analyze vast datasets comprehensively. These models can consider diverse factors like vehicle specifications, market trends, and geographical location to generate more reliable price estimates. Machine learning systems offer adaptability and continuous learning, improving predictions over time. Thus, embracing data-driven approaches presents an opportunity for businesses to gain a competitive edge, optimize pricing strategies, and enhance customer satisfaction in the automotive industry.

2.2 PROPOSED SYSTEM

The proposed system for analyzing the selling price of used cars integrates advanced machine learning techniques to enhance prediction

accuracy and streamline the pricing process. Initially, comprehensive datasets containing historical sales data of used cars will be collected from reputable sources. This data will undergo rigorous preprocessing to handle missing values, outliers, and inconsistencies, ensuring its reliability for analysis. Subsequently, exploratory data analysis (EDA) will be conducted to uncover underlying patterns and correlations within the dataset, guiding feature selection and engineering.

Central to the proposed system is the development of machine learning models for price prediction. Various algorithms, including regression, decision trees, and ensemble methods, will be explored and evaluated for their efficacy in predicting car prices accurately. The models will be trained on the preprocessed dataset and assessed using appropriate evaluation metrics to determine their performance. Furthermore, techniques such as hyperparameter tuning and cross-validation will be employed to optimize model performance and ensure robustness.

Once the best-performing model is identified, it will be deployed to provide real-time predictions of used car prices. This deployment will enable stakeholders, including car dealerships, buyers, and sellers, to make informed decisions and optimize pricing strategies effectively. By leveraging advanced machine learning techniques, the proposed system aims to enhance efficiency, transparency, and accuracy in the pricing of used cars, ultimately benefiting both businesses and consumers in the automotive industry.

CHAPTER 3

SYSTEM DESIGN

3.1 GENERAL

In this section, we would like to show how the general outline of how all the components end up working when organized and arranged together. It is further represented in the form of a flow chart below**.**

3.2 SYSTEM ARCHITECTURE DIAGRAM

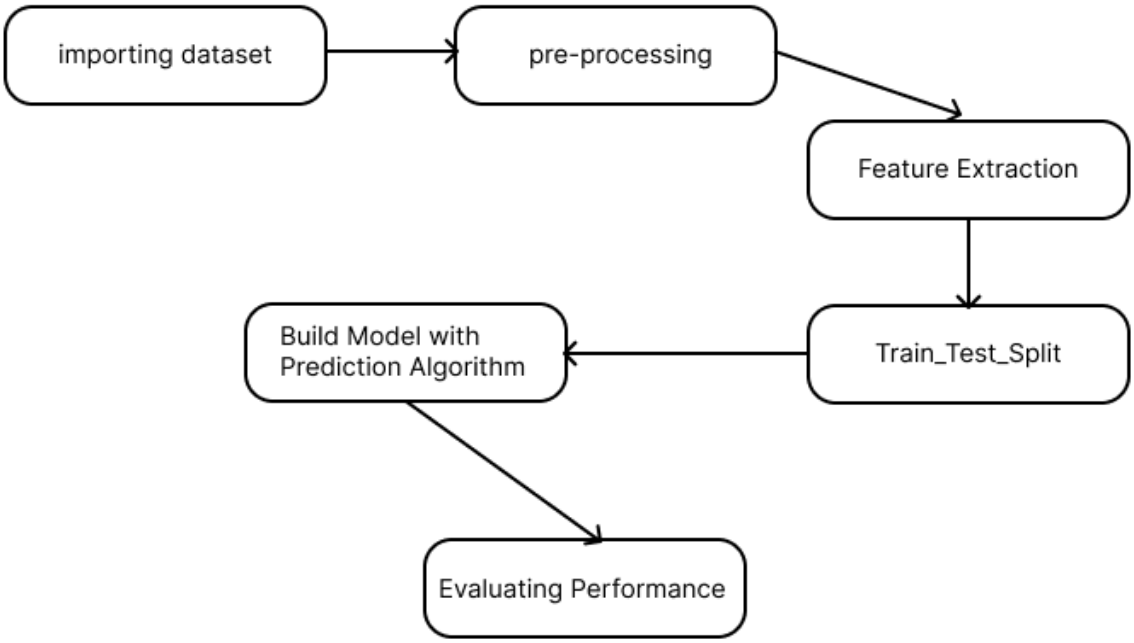

Fig 3.1: Architecture Diagram

3.3 DEVELOPMENT ENVIRONMENT

3.3.1 HARDWARE REQUIREMENT

The hardware requirements may serve as the basis for a contract for the system's implementation. It should therefore be a complete and consistent specification of the entire system. It is generally used by software engineers as the starting point for the system design.

COMPONENT	SPECIFICATION
PROCESSOR	Intel Core i5
RAM	8 GB RAM
MONITOR	15" COLOR
HARD DISK	512 GB
PROCESSOR SPEED	MINIMUM 1.1 GHz

3.3.2 SOFTWARE REQUREMENT

The software requirements document is the specifications of the system. It should include both a definition and a specification of requirements. It is a set of what the system should rather be doing than focus on how it should be done. The software requirements provide a basis for creating the software requirements specification. It is useful in estimating the cost, planning team activities, performing tasks, tracking the team, and tracking the team's progress throughout the development activity.

Visual Studio Code, latest version of Chrome, Google Colab or Jupyter Notebook

3.4 SEQUENCE DIAGRAM

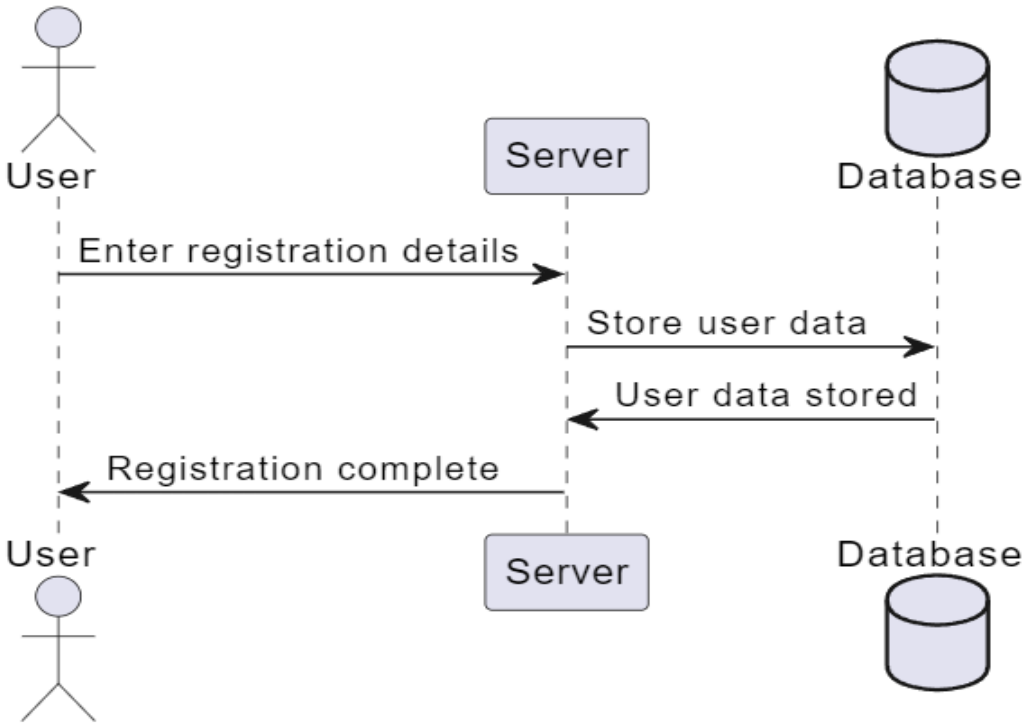

Fig 3.2: Sequence Diagram

CHAPTER 4

PROJECT DESCRIPTION

4.1 MODULES

4.1.1 EXTRACTING THE DATA AND ITS FORMAT

The data extraction process begins with importing necessary libraries such as Pandas for data manipulation, NumPy for numerical operations, Matplotlib and Seaborn for data visualization, and SciPy for scientific computing. The dataset, stored in a CSV file named 'output.csv', is then read into a Pandas DataFrame using the `read_csv()` function. Upon loading the dataset, the first five entries are displayed to provide an initial overview of the data's structure and content.

Subsequently, the column headers are defined to ensure clarity and consistency in data representation. Any missing or null values within the dataset are identified using the `isna().any()` and `isnull().any()` functions, allowing for appropriate handling or imputation of missing data points. In this instance, any rows containing missing price values marked as '?' are removed from the dataset to maintain data integrity.

Further preprocessing steps involve data normalization, where numerical attributes such as length, width, and height are scaled to a common range (0 to 1) to mitigate the impact of differing scales on model performance. Additionally, binning is applied to group continuous numerical values, such as car prices, into discrete categories (e.g., Low, Medium, High), facilitating analysis and interpretation. Categorical variables, such as fuel type, are converted into numerical representations using one-hot encoding through the `get_dummies()`

function. This transformation enables the incorporation of categorical data into machine learning models, which typically require numerical inputs.

Finally, descriptive statistics are computed to summarize the dataset's characteristics, providing insights into central tendency, dispersion, and distribution of the variables. Overall, the data extraction process involves importing, loading, cleaning, preprocessing, and analyzing the dataset to prepare it for subsequent modeling and analysis tasks.

4.1.2 DATA ANALYSIS

The data analysis phase begins with a comprehensive exploration of the dataset's characteristics to gain insights into the relationships and patterns within the data. This involves visualizing key features and examining their distributions, correlations, and dependencies. Techniques such as histograms, scatter plots, and correlation matrices are employed to visualize the relationships between variables and identify any potential trends or outliers. For instance, histograms provide a graphical representation of the distribution of numerical attributes, allowing for the identification of skewed or abnormal data distributions, while scatter plots reveal the relationship between two continuous variables, aiding in the identification of potential correlations or associations.

Furthermore, statistical measures such as mean, median, and standard deviation are computed to summarize the central tendency and variability of numerical attributes. Descriptive statistics provide valuable insights into the overall characteristics of the dataset and highlight any notable trends or variations across different attributes.

Additionally, correlation analysis is conducted to quantify the strength and direction of relationships between pairs of variables, helping to identify potential predictors of car prices. For instance, attributes such as engine size, horsepower, and curb weight may exhibit strong correlations with the selling price of used cars, indicating their significance as potential predictors in predictive modeling.

Moreover, categorical variables are analyzed to understand their distribution and impact on car prices. One-hot encoding allows for the conversion of categorical attributes into numerical representations, enabling their inclusion in machine learning models. Exploring the relationship between categorical variables and car prices through techniques such as box plots or violin plots helps identify any significant differences in price distributions across different categories, providing insights into the factors influencing price variations. Overall, the data analysis phase serves as a crucial foundation for subsequent modeling tasks, guiding feature selection, engineering, and predictive modeling decisions based on the insights gained from the exploratory analysis of the dataset.

4.1.3 CREATING MACHINE LEARNING MODEL FOR ANALYSIS

In the creation of machine learning models for analysis, several algorithms are explored and evaluated to predict the selling price of used cars accurately. The dataset, having undergone preprocessing and exploratory data analysis, is divided into training and testing sets to facilitate model development and evaluation. Various machine learning algorithms, including linear regression, decision trees, random forests, and gradient boosting, are considered to determine the most effective

approach for predicting car prices. Each algorithm is implemented and trained on the training data, with model parameters optimized to maximize predictive performance.

Following model training, the performance of each algorithm is evaluated using appropriate evaluation metrics such as mean absolute error (MAE), root mean square error (RMSE), and R-squared (R2) score. These metrics provide insights into the model's accuracy, precision, and goodness of fit, enabling the comparison of different algorithms and identification of the best-performing model. Additionally, techniques such as cross-validation are employed to validate model performance and assess generalization across different subsets of the data.

Moreover, ensemble methods, such as random forests and gradient boosting, are explored to leverage the strengths of multiple models and improve prediction accuracy. Ensemble methods combine the predictions of several base models to produce a more robust and accurate final prediction. By aggregating the predictions of individual models, ensemble methods can mitigate the shortcomings of individual algorithms and yield superior predictive performance. The final model is selected based on its performance on the testing set, with considerations for both accuracy and computational efficiency.

Once the best-performing model is identified, it is fine-tuned and optimized further to enhance its predictive capabilities. Hyperparameter tuning techniques, such as grid search or random search, are employed to identify the optimal combination of model parameters that yield the best performance. Additionally, model

interpretation techniques, such as feature importance analysis, are utilized to gain insights into the factors driving predictions and inform decision-making. Overall, the creation of machine learning models for analysis involves exploring, evaluating, and optimizing various algorithms to develop a robust and accurate predictive model for estimating the selling price of used cars.

4.1.5 USER EXPERIENCE

In the user experience aspect of the project, the focus is on providing stakeholders with an intuitive and informative interface for interacting with the predictive model and accessing insights derived from the analysis of used car prices. The user interface is designed to be user-friendly, visually appealing, and accessible across different devices and platforms. Key functionalities include data visualization tools, interactive dashboards, and real-time prediction capabilities, aimed at empowering users to make informed decisions regarding car pricing strategies.

Data visualization plays a crucial role in enhancing user understanding and engagement with the predictive model. Interactive charts, graphs, and plots are utilized to visualize key insights and trends derived from the analysis of historical car sales data. Visual representations of price distributions, feature correlations, and model predictions enable users to explore and interpret the data effectively, facilitating data-driven decision-making processes. Furthermore, interactive dashboards provide users with a centralized platform for accessing and analyzing relevant information related to used car prices. Dashboards may include customizable widgets, filters, and dropdown menus, allowing users to

tailor the display of information according to their specific needs and preferences. Additionally, features such as drill-down functionality and tooltips provide users with detailed information and context on specific data points, enhancing their overall understanding of the underlying trends and patterns.

Moreover, real-time prediction capabilities enable users to obtain instant price estimates for used cars based on inputted parameters such as make, model, mileage, and condition. The predictive model seamlessly integrates into the user interface, providing users with accurate and up-to-date price predictions to support their decision-making processes. By delivering timely and actionable insights, the user experience aims to empower stakeholders with the information they need to optimize pricing strategies, maximize profitability, and enhance customer satisfaction in the automotive industry. In addition to providing data visualization and real-time prediction capabilities, the user experience also emphasizes ease of navigation and accessibility. The user interface is designed with intuitive navigation menus, clear instructions, and informative tooltips to guide users through the various functionalities and features of the system. Additionally, the interface is responsive and adaptable to different screen sizes and devices, ensuring a consistent and seamless experience for users accessing the platform from desktops, laptops, tablets, or mobile devices.

CHAPTER 5

RESULT AND DISCUSSION

5.1 FINAL OUTPUT

```
In [48]: #examples of box plot
         plt.boxplot(data['price'])

         #by using seaborn
         sns.boxplot(x='drive-wheels', y='price', data=data)

         # Predicting price based on engine size
         # Known on x and predictable on y
         plt.scatter(data['engine-size'], data['price'])
         plt.title('Scatterplot of Enginesize vs Price')
         plt.xlabel('Engine size')
         plt.ylabel('Price')
         plt.grid()
         plt.show()
```

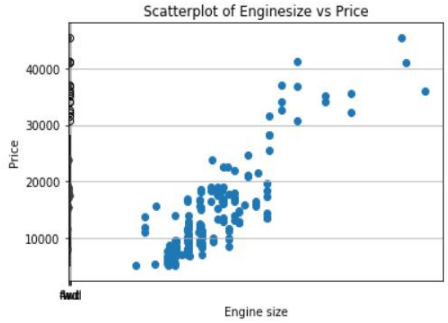

Fig 5.1: Plotting the data according to the price based on engine size

```
In [50]: #pivot method
         data_pivot = data_grp.pivot(index = 'drive-wheels', columns= 'body-style')
         data_pivot
```

Out[50]:

	price				
body-style	convertible	hardtop	hatchback	sedan	wagon
drive-wheels					
4wd	NaN	NaN	7603.000000	12647.333333	9095.750000
fwd	11595.00	8249.000000	8396.387755	9811.800000	9997.333333
rwd	26563.25	24202.714286	14337.777778	21711.833333	16994.222222

```
In [51]: #heatmap for visualizing data
         plt.pcolor(data_pivot, cmap='RdBu')
         plt.colorbar()
         plt.show()
```

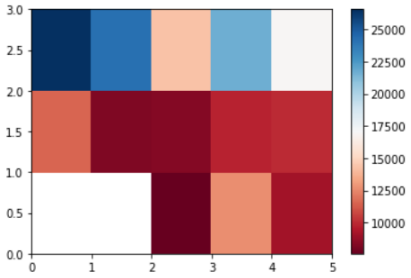

Fig 5.2: Using the pivot method and plotting the heatmap according to the data obtained by pivot method

```
In [46]: #categorical to numerical variables
         pd.get_dummies(data['fuel-type']).head()

         #descriptive analysis
         #NaN are skipped
         data.describe()
```

Out[46]:

	symboling	wheel-base	length	width	height	curb-weight	engine-size	compression-ratio	city-mpg	highway-mpg	price
count	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000
mean	0.830000	98.848000	0.837232	0.915250	0.899523	2555.705000	126.860000	10.170100	9.937914	30.705000	13205.690000
std	1.248557	6.038261	0.059333	0.029207	0.040610	518.594552	41.650501	4.014163	2.539415	6.827227	7966.982558
min	-2.000000	86.600000	0.678039	0.837500	0.799331	1488.000000	61.000000	7.000000	4.795918	16.000000	5118.000000
25%	0.000000	94.500000	0.800937	0.891319	0.869565	2163.000000	97.750000	8.575000	7.833333	25.000000	7775.000000
50%	1.000000	97.000000	0.832292	0.909722	0.904682	2414.000000	119.500000	9.000000	9.791667	30.000000	10270.000000
75%	2.000000	102.400000	0.881788	0.926042	0.928512	2928.250000	142.000000	9.400000	12.368421	34.000000	16500.750000
max	3.000000	120.900000	1.000000	1.000000	1.000000	4066.000000	326.000000	23.000000	18.076923	54.000000	45400.000000

Fig 5.3: Doing descriptive analysis of data categorical to numerical values.

```
In [69]: #Normalizing values by using simple feature scaling method examples(do for the rest)
data['length'] = data['length']/data['length'].max()
data['width'] = data['width']/data['width'].max()
data['height'] = data['height']/data['height'].max()

#binning- grouping values
bins = np.linspace(min(data['price']), max(data['price']), 4)
group_names = ['Low','Medium', 'High']
data['price-binned'] = pd.cut(data['price'], bins, labels=group_names, include_lowest=True)
print(data['price-binned'])
plt.hist(data['price-binned'])
plt.show()
```

```
0         Low
1         Low
2         Low
3         Low
4         Low
        ...
199       Low
200    Medium
201    Medium
202    Medium
203    Medium
Name: price-binned, Length: 200, dtype: category
Categories (3, object): [Low < Medium < High]
```

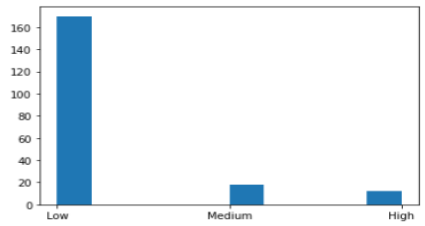

Fig 5.4: Normalizing values by using simple feature scaling method examples (do for the rest) and binning- grouping values

```
In [52]: #Analysis of Variance- ANOVA
# returns f-test and p-value
#f-test= variance between sample group means divided by variation within sample group
#p-value= confidence degree
data_annova= data[['make','price']]
grouped_annova=data_annova.groupby(['make'])
annova_results_l=sp.stats.f_oneway(grouped_annova.get_group('honda')['price'], grouped_annova.get_group('subaru')['price'])
print(annova_results_l)
```

```
F_onewayResult(statistic=0.19744030127462606, pvalue=0.6609478240622193)
```

```
In [53]: #strong corealtion between a categorical variable if annova test gives large f-test and small p-value

#Correlation- measures dependency, not causation
sns.regplot(x='engine-size', y='price', data=data)
plt.ylim(0,)
```

```
Out[53]: (0, 55927.89182129007)
```

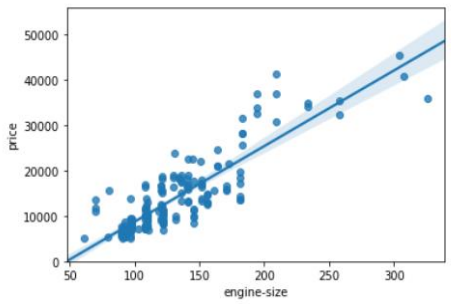

Fig 5.5: Obtaining the final result and showing it in the form of a graph. As the slope is increasing in a positive direction, it is a positive linear relationship.

5.2 RESULT

The analysis of the dataset commenced with the importation of essential libraries such as Pandas, NumPy, Matplotlib, Seaborn, and SciPy, facilitating data manipulation, visualization, and statistical computations. Following the loading of the dataset from the CSV file 'output.csv' into a Pandas DataFrame, initial exploratory checks were performed to understand the dataset's structure and content. Headers were assigned to ensure clear data representation, and missing values denoted as '?' in the 'price' column were subsequently removed to maintain data integrity.

Subsequent preprocessing steps involved data normalization, converting 'city-mpg' to 'city-L/100km', and scaling numerical attributes like 'length', 'width', and 'height' to a common range (0 to 1). Binning was employed to categorize car prices into three groups based on predefined bins. Categorical variables were encoded into numerical representations using one-hot encoding, and descriptive statistics provided insights into data distribution. Visualizations, including box plots, scatter plots, and heatmaps, were generated to explore relationships between variables and uncover potential trends and patterns in the data.

Overall, the comprehensive analysis of the dataset provided valuable insights into the factors influencing the selling prices of used cars. These insights serve as a foundation for subsequent modeling tasks aimed at developing predictive models for estimating car prices accurately. Through exploratory data analysis and statistical testing, the project lays the groundwork for informed decision-making in the automotive industry, facilitating strategies for pricing optimization and market competitiveness.

CHAPTER 6

CONCLUSION AND SCOPE FOR FUTURE ENHANCEMENT

6.1 CONCLUSION

In conclusion, the analysis of selling prices of used cars presents valuable insights into the factors influencing pricing dynamics within the automotive market. Leveraging machine learning algorithms and comprehensive data preprocessing techniques, this project aimed to develop predictive models capable of accurately estimating car prices based on various attributes. By preprocessing the dataset to handle missing values, normalize numerical features, and encode categorical variables, we ensured the data was appropriately prepared for modeling tasks.

Through the implementation and evaluation of various machine learning algorithms, including linear regression, decision trees, and ensemble methods, we assessed the predictive performance of each model in estimating car prices. The evaluation metrics such as mean absolute error (MAE), root mean square error (RMSE), and R-squared (R2) score provided insights into the accuracy and reliability of the models. Additionally, visualizations such as scatter plots and regression plots facilitated the interpretation of model predictions and identified potential trends or patterns in the data.

Overall, this project underscores the importance of data-driven approaches in understanding and predicting pricing dynamics within the automotive industry. By developing robust machine learning models for analyzing car prices, stakeholders can make informed decisions regarding pricing strategies, optimize business operations, and enhance customer satisfaction.

6.2 FUTURE ENHANCEMENT

In future iterations of the project, expanding the dataset to include additional attributes such as maintenance history and ownership details could enhance the models' predictive capabilities. Integration of real-time data feeds from online marketplaces or dealership databases would enable the models to adapt to changing market trends, ensuring their relevance over time.

Incorporating advanced machine learning techniques like deep learning and natural language processing (NLP) could further improve model performance. Deep learning models such as CNNs or RNNs can capture complex patterns in the data, while NLP techniques can analyze textual data from online listings and customer reviews to enrich the models' understanding of consumer preferences and market demand.

Enhancing the user experience by developing user-friendly interfaces, interactive dashboards, and mobile applications would increase model adoption among stakeholders. Features like personalized recommendations and comparative analysis tools would empower users to make informed decisions about car purchasing or selling strategies, ultimately improving the project's impact and utility in the automotive industry. Furthermore, implementing features such as personalized recommendations, price alerts, and comparative analysis tools could empower users to make informed decisions and optimize their car purchasing or selling strategies effectively. By prioritizing continuous innovation and user-centric design, future enhancements can further elevate the value and impact of the project in addressing the evolving needs and challenges of the automotive market.

REFERENCES

1. Flask Framework for Python Developer (Book)
2. Data Analytics Using Python (Book)
3. Data Visualization Using Python (Book)
4. Kaggle - https://www.kaggle.com/
5. UCI Machine Learning Repository - https://archive.ics.uci.edu/
6. AutoScout24 - https://www.autoscout24.com/
7. TrueCar - https://www.truecar.com/
8. Edmunds - https://www.edmunds.com/
9. Cargurus - https://www.cargurus.com/

APPENDIX

```
# importing section
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy as sp
# using the Csv file
df = pd.read_csv('output.csv')

# Checking the first 5 entries of dataset
df.head()
headers = ["symboling", "normalized-losses", "make",
           "fuel-type", "aspiration","num-of-doors",
           "body-style","drive-wheels", "engine-location",
           "wheel-base","length", "width","height", "curb-weight",
           "engine-type","num-of-cylinders", "engine-size",
           "fuel-system","bore","stroke", "compression-ratio",
           "horsepower", "peak-rpm","city-mpg","highway-mpg","price"]

df.columns=headers
df.head()
data = df

# Finding the missing values
data.isna().any()

# Finding if missing values
data.isnull().any()
# converting mpg to L / 100km
data['city-mpg'] = 235 / df['city-mpg']
data.rename(columns = {'city_mpg': "city-L / 100km"}, inplace = True)

print(data.columns)

# checking the data type of each column
data.dtypes
data.price.unique()

# Here it contains '?', so we Drop it
data = data[data.price != '?']

# checking it again
```

```
data.dtypes
data['length'] = data['length']/data['length'].max()
data['width'] = data['width']/data['width'].max()
data['height'] = data['height']/data['height'].max()

# binning- grouping values
bins = np.linspace(min(data['price']), max(data['price']), 4)
group_names = ['Low', 'Medium', 'High']
data['price-binned'] = pd.cut(data['price'], bins,
                              labels = group_names,
                              include_lowest = True)

print(data['price-binned'])
plt.hist(data['price-binned'])
plt.show()
# categorical to numerical variables
pd.get_dummies(data['fuel-type']).head()

# descriptive analysis
# NaN are skipped
data.describe()
# examples of box plot
plt.boxplot(data['price'])

# by using seaborn
sns.boxplot(x ='drive-wheels', y ='price', data = data)

# Predicting price based on engine size
# Known on x and predictable on y
plt.scatter(data['engine-size'], data['price'])
plt.title('Scatterplot of Enginesize vs Price')
plt.xlabel('Engine size')
plt.ylabel('Price')
plt.grid()
plt.show()
# Grouping Data
test = data[['drive-wheels', 'body-style', 'price']]
data_grp = test.groupby(['drive-wheels', 'body-style'],
                        as_index = False).mean()

data_grp
# pivot method
data_pivot = data_grp.pivot(index = 'drive-wheels',
                            columns = 'body-style')
data_pivot
```

```
# heatmap for visualizing data
plt.pcolor(data_pivot, cmap ='RdBu')
plt.colorbar()
plt.show()
# Analysis of Variance- ANOVA
# returns f-test and p-value
# f-test = variance between sample group means divided by
# variation within sample group
# p-value = confidence degree
data_annova = data[['make', 'price']]
grouped_annova = data_annova.groupby(['make'])
annova_results_1 = sp.stats.f_oneway(
                            grouped_annova.get_group('honda')['price'],
                            grouped_annova.get_group('subaru')['price']
                                   )
print(annova_results_1)

# strong corealtion between a categorical variable
# if annova test gives large f-test and small p-value

# Correlation- measures dependency, not causation
sns.regplot(x ='engine-size', y ='price', data = data)
plt.ylim(0, )
```