

SECOND-HAND CAR PRICE MONITOR SYSTEM

BY

SCOTT LAI YONG LUO

A REPORT

SUBMITTED TO

UNIVERSITI TUNKU ABDUL RAHMAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

BACHELOR OF INFORMATION SYSTEMS (HONOURS)

BUSINESS INFORMATION SYSTEMS

(IB)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

(KAMPAR CAMPUS)

JUNE 2023

REPORT STATUS DECLARATION FORM

Title: SECOND HAND CAR PRICE MONITOR SYSTEM

Academic Session: JUNE 2023

I SCOTT LAI YONG LUO
(CAPITAL LETTER)

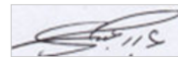
declare that I allow this Final Year Project Report to be kept in
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

Verified by,



(Author's signature)



(Supervisor's signature)

Address:

NO 177, KAMPUNG BALI
31750, TRONOH
PERAK

Abdulkarim M. Jamal Kanaan Jebna
Supervisor's name

Date: 14/9/2023

Date: 14/09/2023

Universiti Tunku Abdul Rahman			
Form Title : Sample of Submission Sheet for FYP/Dissertation/Thesis			
Form Number: FM-IAD-004	Rev No.: 0	Effective Date: 21 JUNE 2011	Page No.: 1 of 1

FACULTY/INSTITUTE* OF INFORMATION AND COMMUNICATION TECHNOLOGY

UNIVERSITI TUNKU ABDUL RAHMAN

Date: 14/9/2023

SUBMISSION OF FINAL YEAR PROJECT /DISSERTATION/THESIS

It is hereby certified that **Scott Lai Yong Luo** (ID No: **19ACB01919**) has completed this final year project/ dissertation/ thesis* entitled “**SECOND-HAND CAR PRICE MONITOR SYSTEM**” under the supervision of **Dr. Abdulkarim Kanaan Jebna** (Supervisor) from the Department of **Information System**, Faculty/Institute* of **Information and Communication Technology**.

I understand that University will upload softcopy of my final year project / dissertation/ thesis* in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

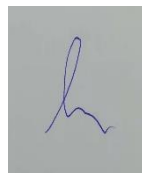
Yours truly,



(*SCOTT LAI YONG LUO*)

DECLARATION OF ORIGINALITY

I declare that this report entitled “**SECOND-HAND CAR PRICE MONITOR SYSTEM**” is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.



Signature : _____

Name : SCOTT LAI YONG LUO

Date : 12/9/2023

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my Final Year Project supervisor, Dr. Abdulkarim Kanaan Jebna, for his unwavering and exceptional support that extended far beyond the traditional role of a supervisor. Dr. Abdulkarim Kanaan Jebna not only guided me through the academic aspects of my project but also actively engaged with me in numerous enriching experiences. Under his mentorship, I had the privilege to accompany him to conferences, where I had the opportunity to expand my horizons and connect with experts in our field. We even participated together in competitions, where his expertise and encouragement were instrumental in our success. Furthermore, Dr. Abdulkarim Kanaan Jebna consistently acknowledged my achievements, no matter how small they may have seemed to me. His generous praise during moments of progress served as a constant source of motivation and confidence. Conversely, when faced with challenges or moments of self-doubt, his words of encouragement were a steady guiding light, instilling in me the determination to overcome obstacles. In addition to his commendable mentorship, Dr. Abdulkarim Kanaan Jebna frequently provided invaluable insights and recommendations to enhance the progress and quality of my project. He was always approachable, and I knew that I could turn to him for guidance and clarification at any time, regardless of the situation. Moreover, even during breaks and holidays, our journey of learning did not cease. Dr. Abdulkarim Kanaan Jebna and I continued to explore new knowledge and skills together, transcending the typical teacher-student relationship to become friend. Dr. Abdulkarim Kanaan Jebna has been a remarkable mentor, friend, and constant source of inspiration throughout this project. His unwavering support, combined with his dedication to my growth, has been a pivotal force behind my achievements. I am profoundly grateful for his patience, wisdom, and encouragement, and I remain committed to meeting and exceeding the high expectations he has set for me.

ABSTRACT

In Malaysia, the used car market frequently lacks transparency and fair pricing, making it difficult for buyers and sellers to make wise decisions. In-depth research on the creation of a machine learning-based pricing prediction model created exclusively for the Malaysian used car market is presented in this paper. The main goal of this project was to develop a reliable and open model that predicts used car pricing based on essential variables, including age, condition, location, and the availability of comparable models on the market. Numerous techniques, such as neural networks and regression models, were employed to accomplish this purpose. The scope of the project also encompassed the identification of challenges and limitations in the current used car market in Malaysia, along with proposed solutions to improve transparency and fairness for all stakeholders. Methodologically, the project involved data collection, preprocessing, feature selection, model training, and evaluation. The results demonstrated that the developed model provided precise and transparent pricing information, empowering buyers, and sellers to make informed decisions regarding their transactions. Subsequently, the project findings were translated into practical implementation with the development of a comprehensive dashboard system. This dashboard serves as a user-friendly interface, enabling real-time access to accurate pricing data. It allows buyers and sellers in the Malaysian used car market to make well-informed decisions efficiently. This project holds significant potential for enhancing transparency and fairness in the Malaysian used car market, while also serving as a valuable reference for similar initiatives in other countries and markets. The fusion of advanced machine learning techniques with an accessible dashboard interface represents a substantial step towards creating a fair and equitable used car market ecosystem.

TABLE OF CONTENTS

SECOND-HAND CAR PRICE MONITOR SYSTEM	i
REPORT STATUS DECLARATION FORM	ii
SUBMISSION OF FINAL YEAR PROJECT /DISSERTATION/THESIS	iii
DECLARATION OF ORIGINALITY	iv
ACKNOWLEDGEMENTS	v
ABSTRACT.....	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES	x
LIST OF TABLES	xi
LIST OF ABBREVIATIONS	xii
CHAPTER 1: INTRODUCTION.....	1
1.1 Problem Statement.....	2
1.2 Objectives	3
1.3 Project Scope and Direction.....	4
1.4 Contributions.....	5
1.5 Report Organization.....	6
CHAPTER 2: LITERATURE REVIEW	7
2.1 Dashboard	7
2.1.1 AirDNA.....	7
2.1.2 Seattle Housing Market.....	10
2.1.3 CarBase	12
CHAPTER 3: SYSTEM METHODOLOGY/APPROACH	13

3.1	CRISP-DM.....	13
3.1.1	Business Understanding Phase	13
3.1.2	Data Understanding Phase	14
3.1.3	Data Preparation Phase	15
3.1.4	Modelling Phase.....	16
3.1.5	Evaluation Phase.....	17
3.1.6	Deployment Phase	18
CHAPTER 4: SYSTEM DESIGN.....		19
4.1	Use Case Diagram.....	19
4.2	System Design	20
CHAPTER 5: SYSTEM IMPLEMENTATION.....		22
5.1	Hardware Setup.....	22
5.2	Software Setup	24
5.3	Setting and Configuration	26
5.4	CRISP-DM.....	27
5.4.1	Data Understanding	27
5.4.2	Data Preparation.....	40
5.4.3	Modelling.....	43
5.5	System Operation (with Screenshot)	45
5.6	Implementation Issues and Challenges.....	49
CHAPTER 6: SYSTEM EVALUATION AND DISCUSSION.....		50
6.1	System Testing and Performance Metrics	50
6.2	Testing Setup and Result	51
6.3	Project Challenges	58
6.4	Objectives Evaluation	60
CHAPTER 7: CONCLUSION AND RECOMMENDATION		62
7.1	Conclusion	62

7.2 Recommendation	64
ACHIEVEMENT	66
REFERENCES.....	67
FINAL YEAR PROJECT WEEKLY REPORT	69
FINAL YEAR PROJECT WEEKLY REPORT	70
FINAL YEAR PROJECT WEEKLY REPORT	71
FINAL YEAR PROJECT WEEKLY REPORT	72
FINAL YEAR PROJECT WEEKLY REPORT	73
POSTER 74	
PLAGIARISM CHECK RESULT	75
CHECKLIST FOR FYP2 THESIS SUBMISSION	77

LIST OF FIGURES

Figure 1. AirDNA Dashboard	7
Figure 2. Seattle Housing Market Dashboard	10
Figure 3. CarBase Dashboard	12
Figure 4. CRISP-DM	13
Figure 5. Use Case Diagram	19
Figure 6. System Design	20
Figure 7. E-commerce Platform	27
Figure 8. Browser Console of Website	28
Figure 9. JSON Formatter	29
Figure 10. URL.CSV	30
Figure 11. Used_Car_Info.CSV	31
Figure 12. Error URL	31
Figure 13. Heatmap	41
Figure 14. General Page of Dashboard	45
Figure 15. Clustering Page of Dashboard	46
Figure 16. Upload File	46
Figure 17. Data Page of Dashboard	47
Figure 18. Sorting	47
Figure 19. Searching	48

LIST OF TABLES

Table 1. Hardware Specification	22
Table 2. Software Specification	24
Table 3. Important Feature of Random Forest vs XGBoost	43
Table 4. Learning Algorithms Performance Using R2 on Test Set	51
Table 5. Summarize of metrics for different model	53
Table 6. Comparison of Different Models	55

LIST OF ABBREVIATIONS

<i>CRISP-DM</i>	Cross-Industry Standard Process for Data Mining
<i>MSE</i>	Mean Squared Error
<i>MAE</i>	Mean Absolute Error
<i>RMSE</i>	Root Mean Squared Error
<i>R²</i>	R-Squared
<i>BPNN</i>	Back Propagation Neural Network

CHAPTER 1: INTRODUCTION

In Malaysia, there has been a significant increase in the number of registered passenger vehicles in recent years, with approximately 459,000 registered vehicles as of September 2022[1]. Additionally, car registrations in Malaysia have been increasing steadily over time, with an average of 72,428 units registered annually between 1988 and 2022[2]. The number of registered cars in Malaysia is growing at a steady rate, and this trend is expected to continue in the future[3]. It is worth noting that in 2021, the number of registered cars in Malaysia had already exceeded the country's human population, with 33.3 million registered cars nationwide versus a human population of 32.6 million[1]. This trend suggests that almost everyone in Malaysia owns a car, which is due in part to the underdevelopment of public transportation in the country[4].

On the other hand, the epidemic has also boosted the used car market. The used car market has seen a huge increase in sales due to factors such as declining income, lack of money, and increasing preference for private cars to maintain social distancing [5]. The shortage of new car supply also drives consumers to consider used cars [6]. As the new vehicle industry takes a hit, the used car industry has risen during the pandemic.

The used car market in Malaysia will be a booming industry in the future, and many Malaysians choose used cars for their affordability and practicality. However, the used car market in Malaysia faces challenges related to pricing transparency and fairness, which can make it difficult for buyers and sellers to make informed decisions [7]. To address this problem, a second-hand car price monitoring system has been developed in Malaysia. This system aims to provide accurate and real-time insights into the dynamic nature of the used car market, empowering users to make well-informed decisions when buying or selling second-hand vehicles. By leveraging machine learning algorithms and data analytics, the system predicts and monitors price fluctuations. For example, it can also compare prices with market averages, similar vehicles, or specific regions and regularly update its data repository and analytical models to provide a comprehensive analysis. At the same time data visualization tools enhance user experience and facilitate data analysis. The implementation of this price monitoring system contributes to improving transparency and fairness in the used car market, benefiting both buyers, and sellers, and promoting a more efficient and competitive marketplace.

1.1 Problem Statement

Market participants are irritated and distrustful due to the prevalence of unfair pricing practices such as overpricing and under-pricing. Unfair pricing makes it challenging for sellers to draw in buyers because astronomical prices may turn away potential buyers. When buying a used car, buyers typically do their homework and do pricing comparisons [8] to be sure they are getting a fair deal. Pricing that is perceived as unjust or excessive may make it challenging for sellers to draw in potential buyers. If a used car is overpriced, it could take longer to sell. The sale term is also prolonged, carrying costs are increased, and a financial burden is produced. A seller's reputation can also be harmed by unscrupulous pricing practices, which can lead to unfavourable reviews and bad word of mouth [9]. If a customer feels they have been overcharged, they may provide bad feedback online or speak out against the seller's pricing practices. The seller's reputation and future business prospects are impacted by this.

Unfair pricing places a financial burden on the buyer, particularly when they overpay for a used automobile, which may impact on their budget and restrict their capacity to pay for the costs connected with owning a car. When purchasers discover they overpaid for a second-hand car, they may feel tricked or exploited, which erodes trust between buyers and sellers[10]. This may make customers wary and less likely to make more purchases or have faith in other suppliers in the industry. In addition, unfair pricing lowers consumers' purchasing power[11], limiting them from obtaining options with higher quality. If customers are unable to discover an affordable option, they can be forced to make concessions or choose a subpar one.

Buyers and sellers are compelled to rely on questionable sources of information or perform time-consuming manual research because there is no single system for accessing the correct pricing, features, and conditions of used vehicles. This consumes resources and makes it more likely that decisions will be made poorly because of missing or incorrect information. The effects of this lack of justice and openness go beyond just the buyers and sellers. Additionally, it harms the economy. Without a trustworthy used-car market, consumer confidence deteriorates, which lowers the demand for used cars. This, in turn, influences the sales and revenue of car dealerships and restricts the market's growth. The eventual benefits of a healthy used car market are lost on the economy.

1.2 Objectives

To develop a machine learning-based price prediction model for used cars in Malaysia, with the aim of providing buyers and sellers with more accurate and transparent pricing information and enabling them to make more informed decisions about the value of their transactions. Machine learning techniques such as Backpropagation neural network and regression model can be used to predict the prices of used cars with high accuracy and that such models can be integrated into a monitoring system to provide buyers and sellers with better information about pricing.

- To develop a machine learning-based price prediction model that accurately predicts the price of used cars.
- To integrate the developed model into a monitoring system that provides buyers and sellers with better information about pricing.
- To identify challenges and limitations in the current second-hand car market in Malaysia.
- To propose potential solutions to improve transparency and fairness for both buyers and sellers.
- To determine the most important features that influence the price of used cars in Malaysia.
- To improve the index performance of the model and increase the accuracy to over 95%
- To create a dashboard that enables users to compare prices of different car models in real-time.

1.3 Project Scope and Direction

The scope of this project entails the development of a comprehensive dashboard designed to empower users in Malaysia to effectively monitor and navigate the dynamic used car market. This dashboard will aggregate data from diverse sources, including established car dealerships and prominent online marketplaces. It will employ state-of-the-art machine learning and deep learning algorithms to provide predictive insights into the pricing of specific car models. These predictions will be founded on a robust analysis of critical factors, such as the vehicle's age, mileage, condition, and other pertinent variables. Currently, the project's scope is centred around the Malaysian region and the domain of used cars. In the future, the system may expand to more regions. The primary audience for this project includes prospective car buyers and sellers in Malaysia. However, the benefits of this platform extend to various stakeholders:

Car Buyers: Individuals seeking to purchase used cars can use this platform to gain insights into fair market prices, enabling them to make informed and cost-effective decisions.

Car Sellers: Sellers can use the dashboard to determine competitive and fair pricing for their vehicles, attracting potential buyers more effectively.

Car Dealerships: Established dealerships can utilize the platform to optimize their pricing strategies, improving their competitiveness in the market.

Marketplace Users: Users of online marketplaces can benefit from enhanced transparency and more accurate pricing information, reducing the risk of overpaying or underselling.

The overarching direction of this project is to craft an intuitively designed and user-friendly platform, complemented by a dynamic and informative dashboard. The platform aims to offer prospective car buyers and sellers in Malaysia a streamlined approach to accessing and comprehending valuable insights within the used car market. By harnessing the power of data visualization, the platform's dashboard will present intricate market data in a comprehensible and actionable format. Through the amalgamation of data derived from multiple sources, the platform aspires to empower users with the knowledge required to make judicious decisions concerning the fair market value of specific cars. This endeavour will substantially enhance transparency in the used car market, fostering equitable and efficient transactions for all participants. Furthermore, the platform has been meticulously designed to be scalable and adaptable. This inherent flexibility positions it for potential expansion into other countries and markets, thereby elevating its impact and utility on a global scale.

1.4 Contributions

The contributions of this project are multifield and impactful. Firstly, it addresses a critical issue in the Malaysian used car market by introducing a machine learning-based pricing prediction model. This model provides precise and transparent pricing information, benefiting both buyers and sellers and fostering a fairer marketplace, reducing the risk of fraud and misrepresentation by unscrupulous merchants or intermediaries, and creating trust within the market. Secondly, the development of a web-based platform, in the form of a user-friendly dashboard, for monitoring used car prices offers users a valuable tool for making informed decisions. This web-based dashboard is easily accessible to users via the Internet, ensuring convenience and widespread availability. It would enable buyers and vendors to readily compare prices, features, and conditions, saving time and effort. This platform serves as a valuable resource for navigating the complexities of the market and empowers users to make choices based on real-time data. Thirdly, the implementation of a user-friendly dashboard through data visualization enhances the accessibility of market insights. This dynamic visualization tool equips stakeholders with the means to explore market trends, brand and model distributions, pricing patterns, and other critical information effortlessly. Its user-friendly interface and real-time data updates empower users to stay informed, identify emerging market trends promptly, and adeptly navigate the dynamic market landscape. Moreover, the scalability of the platform positions it for potential expansion to other countries, thereby extending its positive impact beyond Malaysia. Lastly, this project serves as a valuable reference for similar initiatives in the automotive industry and other markets seeking transparency and fairness in pricing mechanisms.

1.5 Report Organization

This report comprises six integral chapters, each contributing to a comprehensive understanding of the project's development and outcomes. In Chapter 1, we introduce the project, addressing the problem statement, background, motivation, objectives, scope, contributions, achievements, and the report's organizational structure. Chapter 2 conducts a thorough review of the existing dashboard, evaluating its strengths and weaknesses. Chapter 3 elucidates the system's design, providing a blueprint for subsequent development. Chapter 4 details system implementation and testing procedures. Chapter 5 presents the system's outcomes, followed by a comprehensive discussion of their significance. Finally, Chapter 6 offers a concise summary of the report, including recommendations for future research and development.

CHAPTER 2: LITERATURE REVIEW

This chapter provides a comprehensive review of the literature on existing dashboards and relevance research papers in the field of data analytics. The purpose of this chapter is to gain a deeper understanding of the current state of the art in dashboard design, development, and implementation. The chapter begins by introducing the concept of dashboards, their applications, and their importance in data analytics. Next, it presents an overview of the key features and characteristics of effective dashboards. Finally, it reviews the existing literature on dashboard design, development, and implementation.

2.1 Dashboard

2.1.1 AirDNA

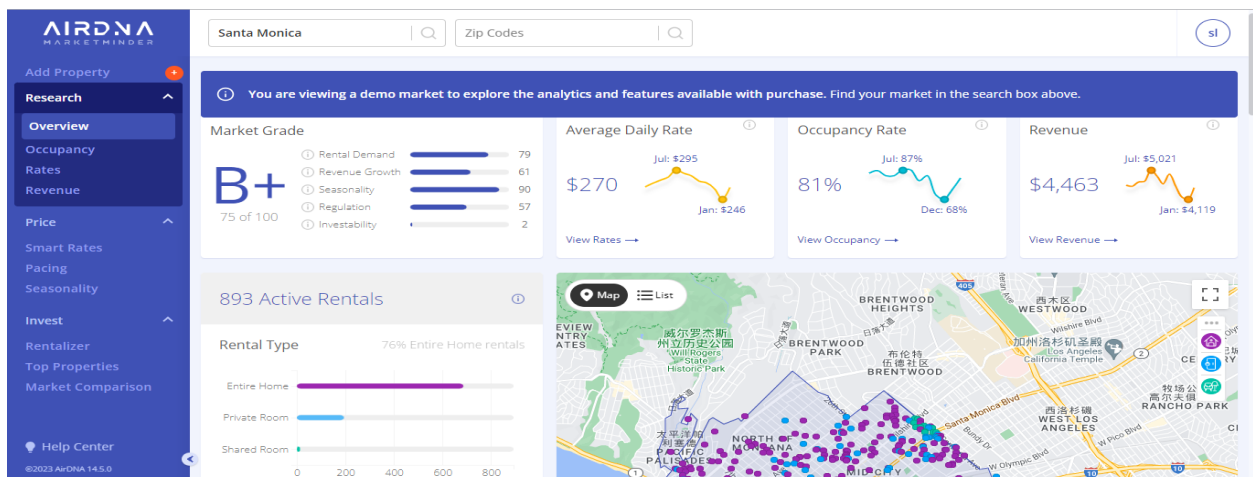


Figure 1. AirDNA Dashboard

AirDNA is a platform that provides data analysis on short-term vacation rentals, including Airbnb and Vrbo. The platform tracks performance data of over 10 million vacation rentals, allowing users to analyse occupancy rates, pricing, and investment research. The specific link provided focuses on seasonality data for vacation rentals in Santa Monica, California.

The website offers an interactive dashboard that allows users to explore the data visually. The dashboard includes a map of Santa Monica with rental properties marked, a chart of monthly occupancy rates, and a table displaying average daily rates for various types of rentals. Users can also filter the data by property type, bedrooms, and other criteria.

CHAPTER 2: LITERATURE REVIEW

AirDNA provides a comprehensive overview of the industry, including occupancy rates, pricing, and investment research. The user-friendly interface features charts and graphs that make it easy to understand the data. Additionally, AirDNA boasts a vast database of over 10 million Airbnb and Vrbo vacation rentals, which users can search by location to focus on specific areas of interest. The website offers a free trial to give users a chance to test the service before committing to a paid subscription.

In addition, they combine data from multiple sources, including Airbnb, Vrbo, and other booking platforms, with data on local supply and demand trends to generate insights on the short-term rental market. According to their website, they use advanced artificial intelligence and machine learning technology to accurately identify blocks and unavailable days on Airbnb and Vrbo [12]. They have also developed a proprietary algorithm called MarketMinder. However, it is possible that they use a combination of various models such as time series analysis, regression analysis, and clustering algorithms to perform data analysis and make predictions.

However, there are also some weaknesses to consider when using AirDNA. The data provided may not be entirely accurate since it is based on algorithms that may not capture all of the nuances of the vacation rental market. Additionally, the website requires a paid subscription to access most of the data, which may be a barrier for some users. AirDNA also does not provide personalized support or consulting services to help users interpret the data and make strategic decisions based on it. Lastly, it is important to note that AirDNA primarily focuses on the Airbnb and Vrbo markets, which may not be relevant to users who are interested in other vacation rental platforms.

Conclusion, AirDNA's vacation rental data platform offers a wealth of information for individuals and businesses looking to invest in the vacation rental market in Santa Monica or other areas. The interactive dashboard makes it easy to visualize and explore the data, and the platform provides a range of tools for in-depth analysis.

Based on the AirDNA dashboard for Santa Monica, California, it provides data and insights into the vacation rental market's occupancy rates, revenue, and other key metrics. In contrast, our system's machine learning-based price prediction model can forecast the price of used cars accurately based on various factors such as age, condition, location, and availability of similar

Bachelor of Information Systems (Honours) Business Information Systems
Faculty of Information and Communication Technology (Kampar Campus), UTAR

CHAPTER 2: LITERATURE REVIEW

models on the market. Moreover, our system can identify the most important features that influence the price of used cars in Malaysia using feature selection techniques, which is not provided by the AirDNA dashboard.

Additionally, our system includes a clustering feature that allows users to group similar data points and identify patterns, which is not available in the AirDNA dashboard. Overall, our system differs from the AirDNA dashboard in terms of its predictive capabilities, feature selection techniques, and clustering feature capabilities.

2.1.2 Seattle Housing Market

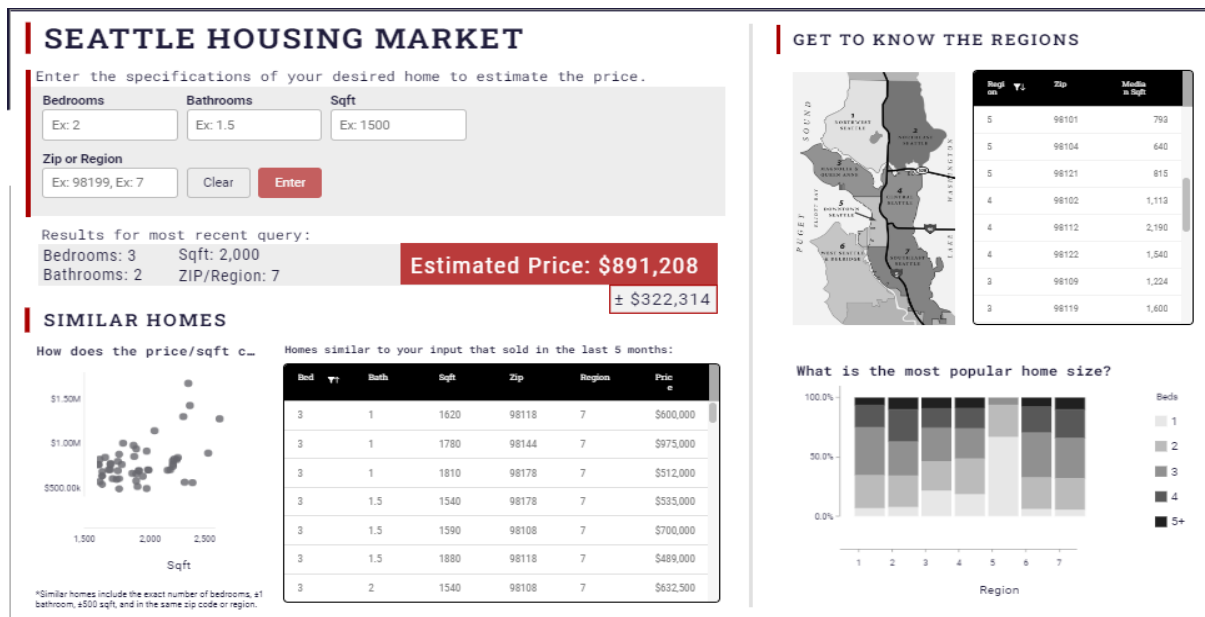


Figure 2. Seattle Housing Market Dashboard

The Seattle Housing Market is an interactive dashboard that displays data for homes sold in Seattle, Washington, USA between August, and December 2022. The dashboard showcases Astrato's Input form functionality, which enables the user to predict the price of their desired home based on their requirements using a sophisticated filtering option. Astrato is a company that provides solutions for data management and analysis, and they offer a range of services including data apps, data warehousing, and data analytics [13].

The Seattle Housing Market Dashboard is one of the data apps offered by Astrato. It allows users to explore the Seattle housing market by filtering data based on various parameters. The dashboard also includes a feature that enables users to predict the price of their desired home based on their requirements.

Astrato provides best practices for creating workbooks (also known as dashboards) and data apps on their blog. These best practices include guidelines for data visualization, dashboard design, and user experience [14]. They also offer a guide on what data apps are and how they can be used to improve business operations, including logistics, marketing, and financial health [15].

CHAPTER 2: LITERATURE REVIEW

To create workbooks and data apps, Astrato provides a Data View Editor (DVE) that allows users to select data, add customized SQL, fields, dimensions, and measures in their workbooks. The DVE can also be used to edit joins, create dimensions, and measures [16]. Astrato also offers a range of best practices guides to help users make the most out of their platform.

In terms of strengths, the dashboard provides a user-friendly interface with intuitive data visualization tools and filtering options, allowing users to easily explore and analyse housing market data in Seattle. Additionally, the dashboard includes a feature that displays similar homes to the user's desired home, and a supplemental data sheet that allows the user to explore different regions and how house prices and availability have changed over time. Regarding the type of model used in the dashboard, the dashboard showcases Astrato's Input form functionality and filtering option, suggesting that the dashboard may utilize some form of machine learning or predictive modeling to provide price predictions based on user requirements.

Overall, the Seattle Housing Market Dashboard is a data app provided by Astrato that showcases their capabilities in data management and analysis. The dashboard allows users to explore the Seattle housing market by filtering data based on various parameters and predicting the price of their desired home based on their requirements. Astrato offers a range of services and best practices guides to help users make the most out of their platform.

Our system is focused on the used car market and provides users with the ability to predict car prices, select clustering, and view data visualizations. The Astrato Seattle Housing Market Dashboard, on the other hand, is focused on the housing market and allows users to predict home prices based on their desired attributes. While both systems utilize predictive modeling to help users make informed decisions, the attributes and data used for predictions are different. Our system focuses on car-specific attributes, while the Astrato dashboard focuses on home-specific attributes such as number of bedrooms, bathrooms, and square footage. In terms of important features, our system identifies important features for predicting car prices through the process of data preparation and feature engineering. The Astrato dashboard does not mention any specific methods for identifying important features, although it likely involves some form of feature selection or dimensionality reduction. Overall, both systems serve different markets and provide users with different predictive capabilities based on the specific attributes and data available.

2.1.3 CarBase

The screenshot shows the CarBase.my website interface. At the top, there is a navigation bar with three main sections: 'search', 'compare', and 'decide'. Below this is a search bar with dropdown menus for 'Any Brand', 'Any Model', 'Any Price (Min)', 'Any Price (Max)', 'Any Body Type', and 'Any Segment', followed by a 'Search' button. The main content area is titled 'Car Market Value Guide' and includes a form for entering vehicle details. The form fields are: Brand (BORWARD), Engine Capacity (1395), Model (BX5), Transmission (5 SP AUTOMAT), Year Manufactured (2021), and Variant (Series) (T MY19). A 'Get Valuation' button is located below the form. Below the form, there is a table showing the vehicle description and market value for different regions.

Vehicle Description	Peninsular Malaysia	Sabah / Sarawak
2021 BORWARD BX5 T 4D WAGON 1395	RM 100,600	RM 107,200
Original Price *	RM 127,890	-

Figure 3. CarBase Dashboard

CarBase.my is a popular automotive website in Malaysia that offers an extensive database of new cars on sale, including prices, specifications, warranty details, high-resolution photos, expert and user reviews, and more, in a user-friendly layout [17]. The website also features a car comparison tool and a market value guide dashboard that allows users to estimate the value of their car based on various factors such as make, model, year of production, and condition. However, the Car Market Value Guide lacks some advanced features, such as machine learning algorithms and location-based data, that are available in other car price prediction systems.

Although CarBase.my provides a reliable estimate of a car's market value based on relevant factors, our system offers more advanced features and a more user-friendly interface for users who require detailed and accurate car price prediction information in Malaysia. Our system utilizes machine learning algorithms to predict car prices, considering a wider range of features, including car specifications, market trends, and location-based data. Additionally, our system provides advanced data visualization features that make it easier for users to analyse and compare car prices. While CarBase.my is an excellent resource for researching car prices in Malaysia, our system provides more comprehensive and accurate information that can assist users in making informed decisions when buying or selling cars in the Malaysian market. Our user-friendly interface and advanced data visualization features set us apart from other car price prediction systems and make our system the superior choice for those seeking reliable and detailed car price prediction information in Malaysia.

CHAPTER 3: SYSTEM METHODOLOGY/APPROACH

The methodology employed in this project follows CRISP-DM framework, which provides a structured approach to data analysis. The CRISP-DM methodology consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment [22]. One of the advantages of using the CRISP-DM method is the ability to return to previous stages at any time instead of blindly progressing forward. This allows for the revisiting of previous steps to gain new insights or revise goals and processes if needed.

3.1 CRISP-DM

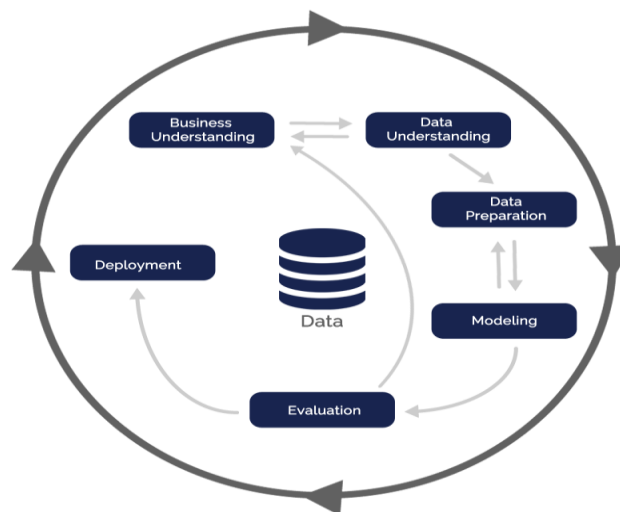


Figure 4. CRISP-DM

3.1.1 Business Understanding Phase

The Business Understanding phase was centred on comprehensively grasping the problem domain, ensuring the development of effective regression models for predicting used car prices. This required delineating key aspects of the research problem. A thorough exploration of the used car market was conducted, delving into the factors affecting car prices. Factors like economic conditions, and consumer preferences were meticulously considered. The dataset collected encompassed a diverse set of attributes pertinent to used cars, including brand, model, manufacturing year, mileage, and engine specifications. A comprehensive understanding of stakeholders integral to the used car market, including buyers, sellers, traders, and investors. Notably, the accuracy of price predictions holds immense value for traders and investors, equipping them with a strategic tool to navigate a rapidly evolving market landscape.

3.1.2 Data Understanding Phase

The data understanding phase involves defining that the collected data are relevant attributes of used cars, their meaning, value, and potential relevance to the research objectives must be understood, and then it is necessary to find out where these required data can be obtained. The primary tool used in this phase is a crawler program that collects data from various sources. After acquiring the data, check the quality of the data such as missing values or inconsistencies, and view the data distribution.

Web scraping is a widely used technique for data collection, and BeautifulSoup is a powerful Python library that enables users to extract structured data from HTML and XML files. BeautifulSoup offers several parsers and efficient ways of navigating, searching, and modifying parse trees, which can save programmers significant time and effort. By parsing the source code of a webpage, the library can extract any data from any website, such as HTML table headings or all the links on the webpage.

In this project, we will use BeautifulSoup to extract data from the website. It allows us to parse data from HTML and XML files and modify parse trees, making web scraping tasks more manageable and efficient. With BeautifulSoup, we can retrieve the price of the car using the `find_all()` method, among other features.

3.1.3 Data Preparation Phase

The data preparation phase is a critical step in the overall machine learning pipeline, where the raw data is cleaned, transformed, and formatted to enable efficient analysis. It is crucial to lay a solid foundation for the subsequent modelling stage, as the quality of the data will significantly impact the accuracy and effectiveness of the machine learning models.

During the data preparation phase, various techniques are used to clean and preprocess the data, such as removing irrelevant or duplicated data, handling missing values, and addressing outliers. The data is also transformed into a suitable format for the model to process, such as converting categorical data into numerical values, scaling the data, and performing feature engineering to extract meaningful information.

It is essential to ensure that the processed data is of high quality and is suitable for the specific use case. By carefully cleaning and preparing the data, the machine learning models can be trained on accurate and relevant information, which can improve the accuracy and effectiveness of the models. In contrast, if the data is not properly processed, it can lead to poor model performance and inaccurate predictions.

Therefore, in the data preparation phase, the project team should prioritize data cleaning, transformation, and formatting to ensure a robust and reliable foundation for subsequent modelling. This includes performing data exploratory analysis to understand the data and detect any data quality issues, applying appropriate data preprocessing techniques to clean and transform the data, and validating the processed data to ensure it meets the quality standards required for the project.

3.1.4 Modelling Phase

In the Modelling phase, various regression algorithms are employed to develop predictive models. The algorithms used in this research include Linear Regression, Lasso Regression, Ridge Regression, ElasticNet Regression, Decision Tree Regression, Random Forest Regression, XGBoost Regression, and a deep neural network (Multilayer Perceptron regressor). Each model is set up as a pipeline, incorporating a preprocessor and a regressor. Hyperparameter fine-tuning is performed to identify the optimal hyperparameters for each model.

Linear regression is a simple, yet powerful modelling technique used to predict continuous numerical values based on one or more input variables. Lasso, ridge, and elastic net are regularization techniques used to reduce overfitting in linear regression models by adding a penalty term to the cost function.

Decision trees are a popular modeling technique for both classification and regression problems. They work by recursively splitting the data into subsets based on the input variables until a stopping criterion is met. Random forests are an ensemble method that combines multiple decision trees to improve the model's accuracy and reduce overfitting.

XGBoost is a popular gradient boosting library used to train decision trees in a way that maximizes the model's predictive performance. By using multiple modeling techniques, we can compare the results of each model and choose the best one for our business problem.

3.1.5 Evaluation Phase

The evaluation phase is a crucial step in the machine learning process as it helps to determine the effectiveness of the developed models and whether they meet the project objectives. During this phase, we will assess the performance of each model and select the most suitable one for the intended application.

To begin the evaluation process, we will first identify the evaluation metrics that are appropriate for the given problem. These metrics could include MSE, MAE, RMSE, R2, and others, depending on the specific nature of the problem. The chosen metrics will serve as a basis for comparing the performance of different models.

Next, we will evaluate each model's performance using the chosen metrics and then split the available data into training and testing sets. Furthermore, cross-validation is utilized to evaluate the model's performance more robustly, iteratively training and evaluating the model based on alternative fold combinations. This aids in evaluating the model's generalization capabilities and performance on various data sets. Train each model on the training set and evaluate its performance on the testing set. This process will help identify the models that perform well on the testing set.

After evaluating the performance of each model, we will select the most suitable one based on the evaluation metrics and other criteria such as computational complexity, interpretability, and ease of deployment. The selected model will be further tested and validated before it is deployed in the real-world application.

3.1.6 Deployment Phase

Finally, in the Deployment phase, the selected model is applied to new data for real-world predictions. The deployment process involves integrating the model into a practical system or providing opinions based on the model's predictions.

To ensure a smooth and effective deployment, the model needs to be thoroughly tested and validated before it is released into production. This process involves conducting a series of tests and simulations to evaluate the model's performance and ensure that it is accurate and reliable in real-world scenarios.

In addition to testing the model, the deployment phase also involves developing a user interface and designing a dashboard to visualize the model's predictions and insights. This user interface serves as a means for users to interact with the model and interpret the results generated by it.

Once the user interface and dashboard have been designed and developed, they are deployed online and made accessible to end-users. It is crucial to monitor the model's performance and user feedback during the initial stages of deployment to ensure that it is working as intended and to address any issues that may arise.

CHAPTER 4: SYSTEM DESIGN

4.1 Use Case Diagram

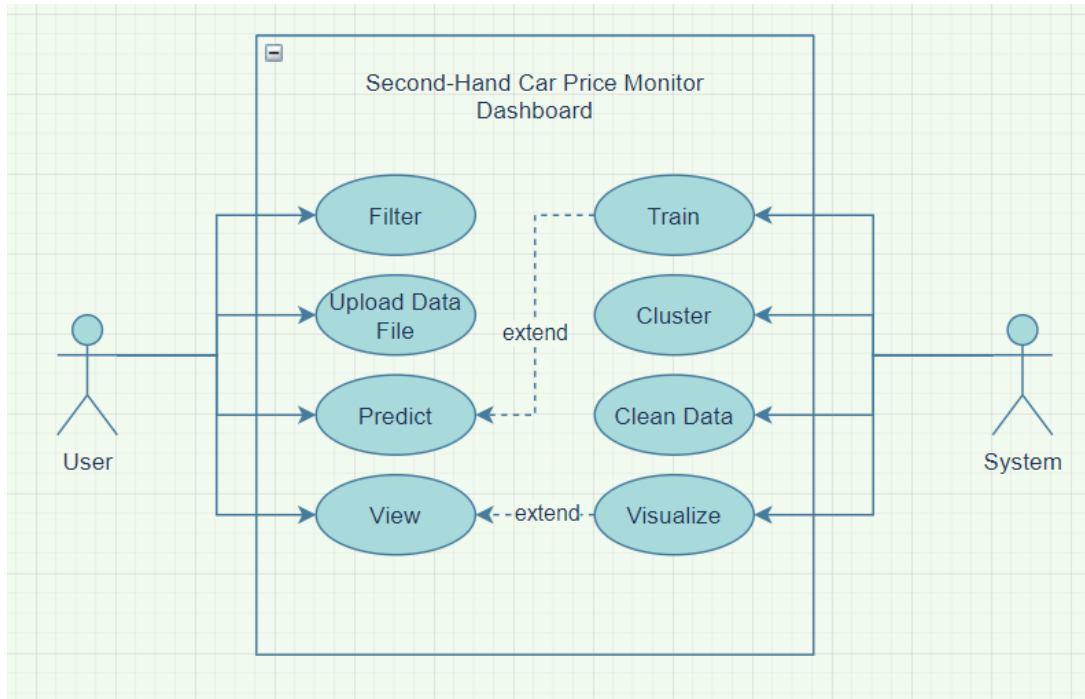


Figure 5. Use Case Diagram

In the use case diagram, two primary actors are identified: the 'User' and the 'System.' The 'User' actor represents individuals interacting with the system, while the 'System' actor denotes the core functionalities of the application. Users have several key interactions with the system, including the ability to 'Filter' data to refine the dataset according to specific criteria. They can also 'Predict' used car prices, 'View' data visualizations and statistics, and 'Upload Data Files' for analysis. On the system side, it is responsible for 'Training' machine learning models, conducting 'Preprocessing' tasks to prepare the data, 'Visualizing' data through charts and graphs, and performing 'Clustering' analysis. These use cases collectively depict the various functionalities and interactions within the system, facilitating a comprehensive understanding of the application's capabilities.

4.2 System Design

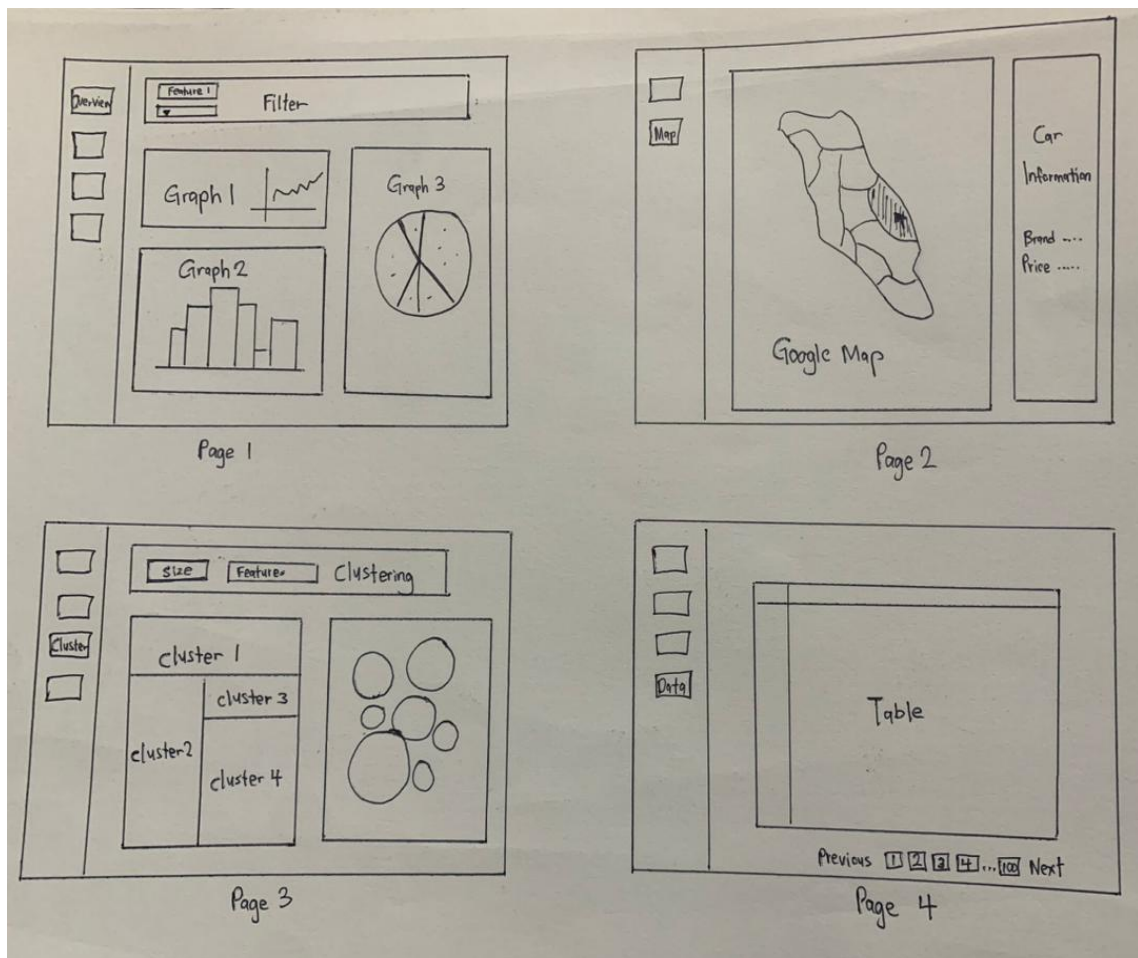


Figure 6. System Design

The system design for the car price prediction dashboard consists of four pages accessible through a navigation panel located on the left side of the screen. The four buttons on the navigation panel correspond to the four pages: Overview, Map, Cluster, and Data.

The first page, Overview, displays an overview of the car price prediction results. At the top of the page, there is a filter navigation panel that allows users to filter the data by various factors such as brand, model, and location. Below the filter navigation panel, the page displays several visualizations such as bar charts and histograms that summarize the data.

The second page, Map, displays a Google Map of Malaysia that allows users to point to a particular area on the map to view information about the cars in that location. The information could include the average price of cars in that area, the number of cars available, and the most popular brands and models in that area.

CHAPTER 4: SYSTEM DESIGN

The third page, Cluster, is the main page for clustering. The top navigation panel on this page allows users to choose the cluster size and features for clustering. The page displays the clustering results in a visualization such as a scatter plot or a heat map. Users can interact with the visualization by selecting a cluster to view more detailed information about the cars in that cluster.

The fourth and final page, Data, displays the data used to train the machine learning model that powers the car price prediction. Users can navigate through the data using the previous and next buttons to view the different tables in the data set. This page is intended to give users an understanding of how the model was built and what data was used.

Overall, the system design is intended to provide users with a comprehensive and user-friendly interface for exploring and understanding car price prediction results. The navigation panel on the left allows users to easily switch between different pages and functionalities, while the visualizations and filters on each page provide users with a clear and informative view of the data.

CHAPTER 5: SYSTEM IMPLEMENTATION

The successful development and implementation of a web-based platform that allows users to monitor the prices of used cars in Malaysia requires careful consideration of the hardware and software requirements. In this chapter, we will discuss the specific hardware and software requirements needed to ensure the proper functioning and optimal performance of the system. Adequate hardware and software resources are essential to support the efficient data processing, analysis, and storage that are required for machine learning algorithms to predict the price of used cars accurately. We will provide a detailed overview of the hardware and software requirements, including their specifications and justifications for their selection. Additionally, we will explain how to truly implement the CRISP-DM framework mentioned in our methodology and discuss the challenges encountered during the implementation process.

5.1 Hardware Setup

Table 1. Hardware Specification

Description	Specifications
Model	Asus Aspire E 14
Processor	Intel Core i5-8250U 1.6GHz with Turbo Boost up to 3.4GHz
Operating System	Windows 11
Graphic	NVIDIA GeForce MX150 with 2 GB VRAM
Memory	8GB DDR4 RAM
Storage	512GB HDD

The hardware specifications of a computer play a crucial role in the performance and efficiency of a machine learning system. In the case of our system, the Asus Aspire E 14 laptop with an Intel Core i5-8250U processor, 8GB DDR4 RAM, and NVIDIA GeForce MX150 with 2 GB VRAM is well-suited for the training of our machine learning algorithms.

The processor, with its Turbo Boost capability of up to 3.4GHz, allows for fast and efficient computation of complex machine learning models. The 8GB DDR4 RAM provides sufficient memory to handle large datasets, while the NVIDIA GeForce MX150 graphics card enhances the system's ability to handle complex visualizations and image processing tasks.

CHAPTER 5: SYSTEM IMPLEMENTATION

Furthermore, the 512GB HDD provides ample storage space for the various data sets and software tools required in the training process. Overall, this hardware configuration ensures that our machine learning system can run smoothly and efficiently and provide accurate predictions and insights for the users.

5.2 Software Setup

Table 2. Software Specification

Description	Specifications/ Version
Python	Python 3.8.8
R	R 4.3.1
RStudio	RStudio 1.4
Jupyter	IPython: 7.29.0 ipykernel: 6.4.1 ipywidgets: 7.6.5 jupyter_client: 6.1.12 jupyter_core: 4.8.1 jupyter_server: 1.4.1 jupyterlab: 3.2.1 nbclient: 0.5.3 nbconvert : 6.1.0 nbformat: 5.1.3 notebook : 6.4.5 qtconsole: 5.1.1 traitlets: 5.1.0
Anaconda	Conda 4.10.3

In the pursuit of creating a robust and efficient used car price monitoring system for Malaysia, the project relies on a well-thought-out selection of software tools. Each of these tools plays a pivotal role in enhancing the project's functionality, data analysis capabilities, and overall effectiveness.

Python is chosen as the primary programming language for its exceptional versatility and extensive libraries. It serves as the foundation for various data processing tasks, including data collection, preprocessing, and the development of machine learning models. Python's flexibility allows for seamless integration with other tools, making it a vital component of this project.

CHAPTER 5: SYSTEM IMPLEMENTATION

R complements Python by bringing specialized statistical functions and packages into the project's arsenal. This integration enhances the project's analytical capabilities, particularly in data exploration, statistical analysis, and data visualization. R's statistical prowess is harnessed to derive valuable insights from the dataset.

RStudio provides an integrated development environment (IDE) that streamlines the development process when working with R. It offers a user-friendly interface for writing, executing, and debugging R scripts. RStudio's efficiency in managing R projects ensures that the statistical aspects of the project run smoothly.

JupyterLab, a powerful and interactive environment, forms the backbone of the project's data exploration and analysis efforts. This platform, equipped with IPython, ipywidgets, and various Jupyter components, facilitates real-time data manipulation and visualization. It allows for interactive data exploration, making it easier to uncover trends and patterns within the dataset.

Anaconda, a comprehensive data science platform, plays a vital role in simplifying software package management and environment setup. Its package management system ensures that all project dependencies are consistent and compatible. Anaconda's robustness in creating and managing isolated environments ensures reproducibility throughout the project's lifecycle.

This meticulously selected ensemble of software tools enables efficient development, thorough data analysis, and insightful data visualization. By leveraging the strengths of both Python and R, combined with the user-friendly interfaces of RStudio and JupyterLab, the project ensures that data processing, model development, and analysis are carried out seamlessly. Anaconda's role in managing dependencies and environments ensures that the project remains consistent and reproducible. Altogether, this software combination forms the foundation for the creation of a reliable used car price monitoring system that addresses the needs of the Malaysian market.

5.3 Setting and Configuration

The successful execution of this project relies on carefully configured software and library ecosystem. In the R environment, a suite of libraries including “Shiny”, “shinydashboard”, “DT”, “ggplot2”, “dplyr”, “d3Tree”, “plotly”, “caret”, “randomForest”, and “shinyjs” were harnessed. These libraries were instrumental in creating an interactive web application that facilitates data visualization, machine learning, and user interaction. The Shiny framework, combined with shinydashboard, provided the structural foundation for the application, while DT enabled the creation of interactive data tables. The powerful ggplot2 and dplyr libraries enhanced data visualization and manipulation capabilities, ensuring that data was presented effectively. Plotly was used to generate interactive charts and graphs, and caret streamlined machine learning model development. The randomForest library was crucial for implementing random forest models, a cornerstone of machine learning within this project. Additionally, shinyjs extended interactivity through JavaScript functions.

In the Python environment, “pandas”, “numpy”, “Scikit-learn”, “XGBoost”, “matplotlib”, seaborn, “requests”, “BeautifulSoup”, “schedule”, and “re” played pivotal roles. Pandas and numpy formed the backbone for data manipulation and numerical computations. Scikit-learn empowered the development, training, and evaluation of machine learning models. XGBoost enriched machine learning capabilities, particularly for regression and classification tasks. Matplotlib and seaborn provided diverse options for data visualization. Requests and BeautifulSoup facilitated web scraping, allowing the collection of data from external sources. The schedule library enabled task automation, ensuring regular updates and data collection. Lastly, the re library supported text processing and pattern matching, making textual data extraction efficient.

These libraries, each with their unique configuration and settings, were expertly integrated to create a dynamic and robust system. Their configuration parameters were adjusted to meet project-specific needs, such as fine-tuning machine learning models or specifying web scraping rules. The effective utilization of these libraries, coupled with their tailored settings, contributed significantly to the success of this project by enabling data analysis, visualization, and machine learning within a seamless user interface.

5.4 CRISP-DM

5.4.1 Data Understanding

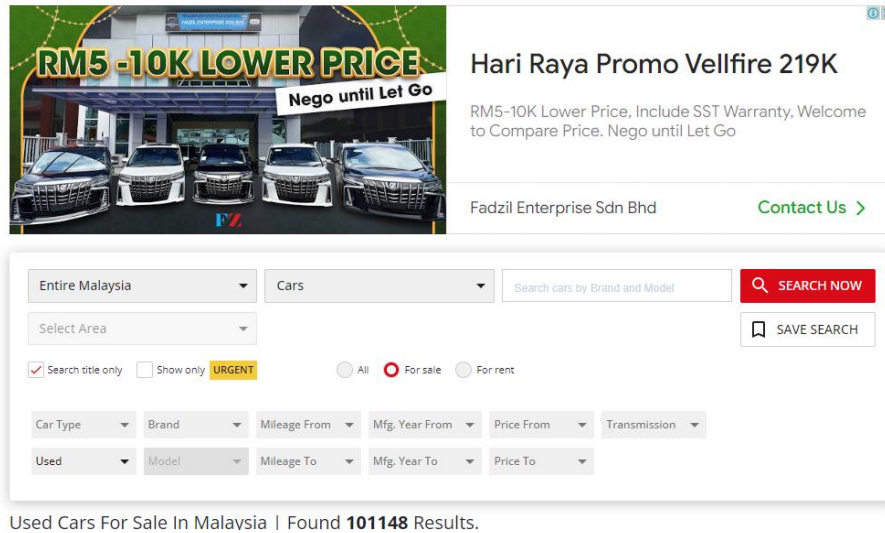


Figure 7. E-commerce Platform

Firstly, our data is sourced from a **Figure 7** well-known online classifieds website in Malaysia that provides a platform for buying and selling a wide range of second-hand goods and services. The website offers extensive categories, including real estate, cars, electronics, home furnishings, pets, personal items, job opportunities, home services, and more. Therefore, we chose to obtain data on second-hand cars in Malaysia through this website for our subsequent data analysis. We used the website's filtering function to find the list of second-hand cars, and we observed that the website had more than 100000 of second-hand car results, with 40 results per page. In each result, we can see some general information such as the car name, price, condition, year, engine capacity, mileage, location, and time. However, this data is not detailed enough, so we need to access the data in each result's internal link. By clicking on the car name of each result, we can enter the internal link and obtain the car's specifications. The specifications include eight major categories such as general information, transmission, engine, dimensions and weight, brakes, suspension, steering, and tires and wheels. The general information category includes information such as the car's brand, model, variant, series, manufacturing year, mileage, type, seating capacity, and country of origin. The dimensions and weight category includes information such as the car's length, width, height, wheelbase, curb weight, and fuel tank capacity. The engine category includes information such as

CHAPTER 5: SYSTEM IMPLEMENTATION

the car's engine displacement, compression ratio, peak power, peak torque, engine type, and fuel type. The remaining categories of transmission, brakes, suspension, steering, and tires and wheels record information related to their respective categories.

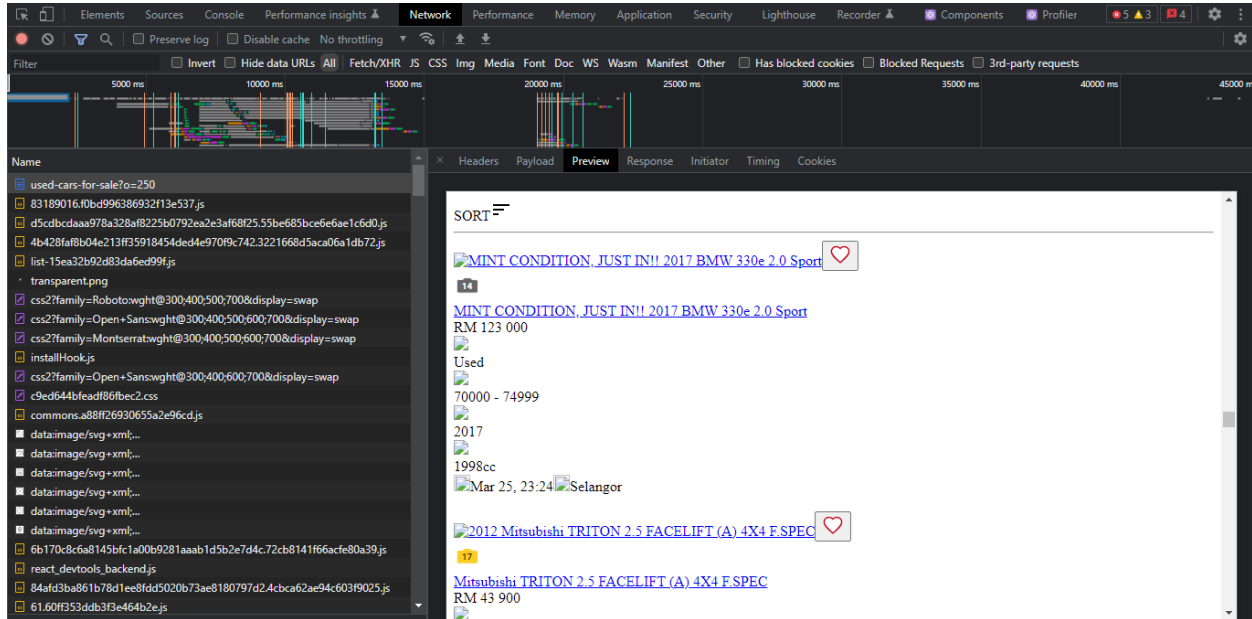


Figure 8. Browser Console of Website

However, we later discovered a problem with the website, which is that after page 251, the results cannot be displayed properly, and the website shows "Sorry, No Results Found!". Therefore, theoretically, we can only obtain results from the first 250 pages, which is 10,000 results. Therefore, our early data analysis focused on these 10,000 results. We then clicked the F12 key to open the **Figure 8** Chrome browser console and found the data interface we wanted by refreshing the webpage while looking for the desired data content. By observing, we found that the data interface for the 250th page is in the form of `https://www.xxx/malaysia/used-cars-for-sale?o=250`, and the data interfaces for other pages also follow this pattern

`https://www.xxx/malaysia/used-cars-for-sale?o={i}`. Next, we entered this interface on the browser and right clicked to view the source code and found that the data we wanted was in JSON format on the JavaScript. Therefore, our next task was to write the web scraper code to obtain the data.

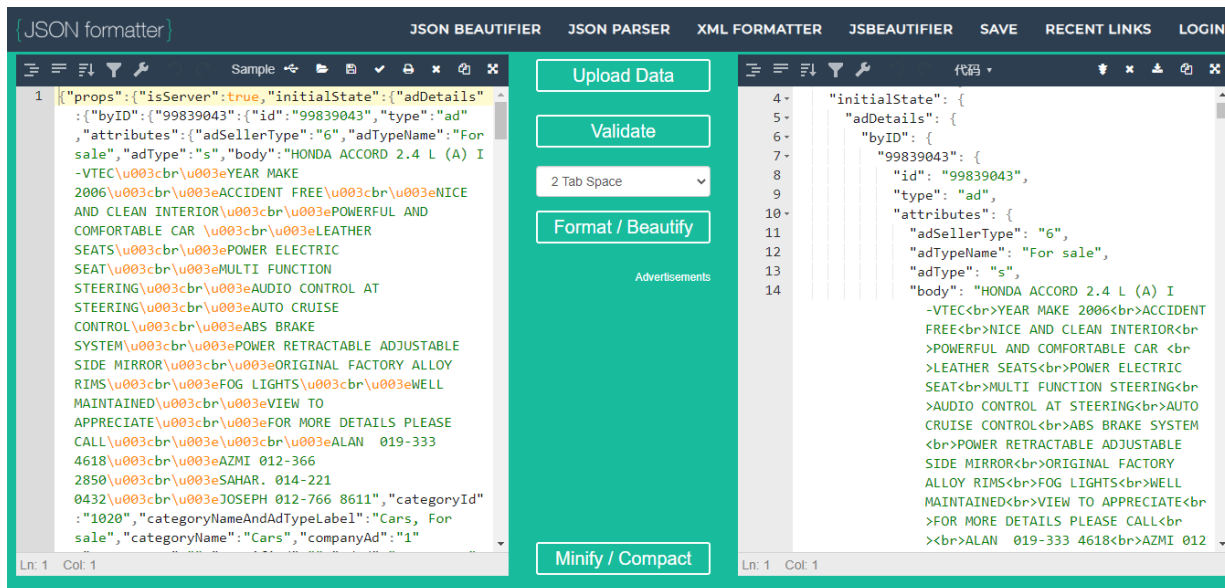


Figure 9. JSON Formatter

During the data scraping process, we used a tool called Figure 9 "JSON formatter" to make the JSON code easier to read and locate the position of the JSON keys.

Headers play a crucial role in the web scraping process as they provide valuable information about the client, such as the browser and operating system being used, as well as the type of content being requested. When making requests to websites, headers can be utilized to indicate the intention of the request and provide additional details like the user agent, referrer, and language preferences.

Certain websites may restrict or block access to their content if they detect that requests are originating from automated scripts or bots. By setting the appropriate headers, the request can appear to come from a legitimate browser or client, thereby decreasing the chances of being blocked or flagged as suspicious. As a result, the first step in web crawling is to establish the header. Failing to set header information can lead to false data or being banned since some websites utilize anti-crawler mechanisms.

Therefore, while conducting web crawling, it is necessary to set appropriate request headers to emulate browser behaviour, giving the request a more authentic and trustworthy appearance.

CHAPTER 5: SYSTEM IMPLEMENTATION

Otherwise, it may result in a 401 Unauthorized error, indicating that the request requires authentication, or a 403 Forbidden error, indicating that the server has rejected the request, and the client lacks permission to access the resource.

The code starts by opening a CSV file named "data.csv" in write mode then creating a csv.writer object to write data to the file. The first row of the file contains the headers "Name", "Price", and "URL". The for loop iterates through page numbers 1 to 250, and for each page, it sends a GET request to the corresponding URL using the requests library and a header variable. As it was mentioned before that the website can only be effective until the 250th page, so we did not continue to crawl the data after the 251st page. The HTML content of the page is then parsed using BeautifulSoup, and the script tag containing JSON data is retrieved. The JSON data is cleaned up and converted to a Python dictionary using the json library, and the relevant information (car name, price, and URL) is extracted and written to the CSV file using the writer object. If there is an error in decoding the JSON data, the loop continues to the next page without writing any data.

	A	B	C	D	E	F
1	Name	Price	URL			
2	2015 Honda CITY 1.5 E (A)	47800	https://www.mudah.my/2015+Honda+CITY+1+5+E+A+-99617931.htm			
3	2019 Perodua BEZZA 1.3 PREMIUM X (A)	35800	https://www.mudah.my/2019+Perodua+BEZZA+1+3+PREMIUM+X+A+-99434884.htm			
4	2018 Perodua MYVI 1.5 H (A)	44800	https://www.mudah.my/2018+Perodua+MYVI+1+5+H+A+-100224272.htm			
5	2018 Mazda CX-5 2.0G GL 2WD FACELIFT (A)	110800	https://www.mudah.my/2018+Mazda+CX+5+2+0G+GL+2WD+FACELIFT+A+-99617542.htm			
6	2018 Nissan SERENA 2.0 S-HYBRD PREMIUM HGHWY STAR	96800	https://www.mudah.my/2018+Nissan+SERENA+2+0+S+HYBRD+PREMIUM+HGHWY+STAR-100416959.htm			
7	2001 Toyota SOARER 4.3 V8 (A)	69900	https://www.mudah.my/2001+Toyota+SOARER+4+3+V8+A+-100528782.htm			
8	2014 Mitsubishi TRITON 2.5 VGT GS FACELIFT (A) 4X4	60800	https://www.mudah.my/2014+Mitsubishi+TRITON+2+5+VGT+GS+FACELIFT+A+4X4-100528779.htm			
9	Isuzu D-MAX 3.0(A) Z-PRESTIGE VGS 4X4 PICKUP TRUCK	81800	https://www.mudah.my/Isuzu+D+MAX+3+0+A+Z+PRESTIGE+VGS+4X4+PICKUP+TRUCK-99944830.htm			
10	DEC 2014 MINI COOPER 2.0 S (A) F56 Local High Spec	127800	https://www.mudah.my/DEC+2014+MINI+COOPER+2+0+S+A+F56+Local+High+Spec-100326957.htm			
11	FEB 2021 BMW 320i (A)G20 Sport DA High spec lowner	218800	https://www.mudah.my/FEB+2021+BMW+320i+A+G20+Sport+DA+High+spec+lowner-100329664.htm			
12	MAR 2019 MERCEDES E200 (A)W213 CKD CAR KING 20k KM	259880	https://www.mudah.my/MAR+2019+MERCEDES+E200+A+W213+CKD+CAR+KING+20k+KM-100037051.htm			
13	REG 2016 MERCEDES A250 AMG (A) CBU Facelift 39k KM	149500	https://www.mudah.my/REG+2016+MERCEDES+A250+AMG+A+CBU+Facelift+39k+KM-99978921.htm			
14	2009 Perodua MYVI 1.3 EZ FACELIFT (A)	15800	https://www.mudah.my/2009+Perodua+MYVI+1+3+EZ+FACELIFT+A+-98844779.htm			
15	2006 Honda ACCORD 2.4 VTI-L (A)	18800	https://www.mudah.my/2006+Honda+ACCORD+2+4+VTI+L+A+-99839043.htm			
16	2017 Isuzu D-MAX 2.5 STANDARD 4X4 FACELIFT (A)	81600	https://www.mudah.my/2017+Isuzu+D+MAX+2+5+STANDARD+4X4+FACELIFT+A+-99083630.htm			
17	2017 Nissan GRAND LIVINA 1.8 CLASSIC/COMFORT (A)	53800	https://www.mudah.my/2017+Nissan+GRAND+LIVINA+1+8+CLASSIC+COMFORT+A+-100069373.htm			
18	2014 Honda CIVIC 1.8 S (A)	47800	https://www.mudah.my/2014+Honda+CIVIC+1+8+S+A+-100069291.htm			
19	ALL-IN 2016 Honda CIVIC 1.5 TC-PREMIUM (A)	93800	https://www.mudah.my/ALL+IN+2016+Honda+CIVIC+1+5+TC+PREMIUM+A+-99087806.htm			
20	2012 Honda CR-Z 1.5 (HYBRID) (A)	41800	https://www.mudah.my/2012+Honda+CR+Z+1+5+HYBRID+A+-100528728.htm			
21	ALL-IN 2019 Proton X70 1.8 PREMIUM 2WD (A)	80800	https://www.mudah.my/ALL+IN+2019+Proton+X70+1+8+PREMIUM+2WD+A+-99088333.htm			

Figure 10. URL.CSV

Figure 10 is the CSV file generated by the first program after crawling the relevant data. The file contains information such as the car's name, price, and corresponding URL. These URLs are critical and may serve as valuable references in the future, as we plan to utilize them for further data extraction using the second crawler program.

CHAPTER 5: SYSTEM IMPLEMENTATION

Then create a Python script for web scraping information about used cars in Malaysia from a website. The script extracts data from a list of URLs in a CSV file and writes the information to a new CSV file. First, the code defines the header row for the output CSV file, which includes various details about the car such as its name, state, city, condition, brand, model, variant, series, manufacturing year, mileage, type, seat capacity, country of origin, transmission, engine type, fuel type, dimensions, weight, brakes, suspension, steering, and tyre and wheel specifications. Then, initializes the output CSV file and writes the header row. After that, loops through the input CSV file and extracts data for each URL from the previous scraping program. For each URL, the script sends an HTTP request to the server and obtains the response. It then extracts the car's ad ID from the URL and uses it to obtain the relevant data from the HTML page using BeautifulSoup.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
1	name	state	city	condition	brand	model	variant	series	mfg_year	mileage	type	seat_capa	country_o	transmissi	engine_cc	compressi	peak_pow	peak_torc	engine_ty	fuel_type	length	
2	2015 Honc	Selangor	Seri Kemt	Used	HONDA	CITY	E	MY2014	2015	170 000	- 14D SEDAN	5	MALAYSIA	CONTINU	1497	10.3	88	145	MULTI POI	PETROL	1	
3	2019 Pero	Selangor	Seri Kemt	Used	PERODUA	BEZZA	PREMIUM	-	2019	75 000	- 754D HATCH	-	MALAYSIA	4 SP AUTC	1329	-	-	-	-	PETROL	-	
4	2018 Pero	Selangor	Seri Kemt	Used	PERODUA	MYVI	H	MY2017	2018	45 000	- 454D HATCH	5	MALAYSIA	4 SP AUTC	1496	10	76	136	NATURALI	PETROL	1	
5	2018 Mazc	Selangor	Seri Kemt	Used	MAZDA	CX-5	2.0G GL 2V	FACELIFT	2018	50 000	- 544D WAGO	5	MALAYSIA	6 SP AUTC	1998	13	114	200	NATURALI	PETROL	1	
6	2018 Nissz	Selangor	Seri Kemt	Used	NISSAN	SERENA	S-HYBRID	C27	2018	85 000	- 854D VAN	8	MALAYSIA	1 SP AUTC	1997	0	106	210	NATURALI	PETROL	1	
7	2001 Toyo	Sabah	Penampai	Used	TOYOTA	SOARER	SCY430	V8-	2001	75 000	- 752D COUPE	4	JAPAN	5 SP AUTC	4292	11	206	430	MULTI POI	PETROL	1	
8	2014 Mitsi	Kuala Lum	Sungai Be	Used	MITSUBISI	TRITON	VGT GS	KB4TGJYX	2014	85 000	- 85DUAL CAB	5	THAILAND	5 SP AUTC	2477	16.5	131	350	DIESEL	TUI	DIESEL	1
9	Isuzu D-M	Selangor	Kajang	Used	ISUZU	D-MAX	Z PRESTIG	MY2013 FJ	2017	100 000	- 1DUAL CAB	5	MALAYSIA	5 SP AUTC	2999	17.3	130	380	TURBO IN	DIESEL	1	
10	DEC 2014	I Penang	Greenlant	Used	MINI	COOPER	S	F56	2014	65 000	- 652D HATCH	4	UNITED KI	6 SP AUTC	1998	11	141	280	TURBO DIH	PETROL	1	
11	FEB 2021	E Penang	Greenlant	Used	BMW	3 SERIES	201 SPORT	MY19 G20	2021	50 000	- 544D SEDAN	5	MALAYSIA	8 SP AUTC	1998	10,2	135	300	TURBO IN	PETROL	4	
12	MAR 2019	Penang	Greenlant	Used	MERCEDES	E-CLASS	200 AVAN	W213	2019	20 000	- 244D SEDAN	5	MALAYSIA	9 SP AUTC	1991	9.8	135	300	TURBO IN	PETROL	4	
13	REG 2016	I Penang	Greenlant	Used	MERCEDES	A-CLASS	250 AMG	L W177	2015	30 000	- 344D HATCH	-	GERMANY	7SP DUAL	1991	-	-	-	-	PETROL	-	
14	2009 Pero	Kuala Lum	Old Klang	Used	PERODUA	MYVI	EZ	FACELIFT	2009	150 000	- 14D HATCH	5	MALAYSIA	4 SP AUTC	1298	10	64	116	ELECTRON	PETROL	1	
15	2006 Honc	Kuala Lum	Old Klang	Used	HONDA	ACCORD	VTI-L	-	2006	160 000	- 14D SEDAN	5	MALAYSIA	5 SP AUTC	2354	9.3	125	218	MULTI POI	PETROL	1	
16	2017 Isuzu	Selangor	Petaling J	Used	ISUZU	D-MAX	STANDARI	MY2013 FJ	2017	140 000	- 1DUAL CAB	5	MALAYSIA	5 SP AUTC	2499	18.1	100	320	TURBO IN	DIESEL	1	
17	2017 Nissz	Selangor	Seri Kemt	Used	NISSAN	GRAND LE	CLASSIC/C	MY2013	2017	70 000	- 744D WAGO	7	MALAYSIA	4 SP AUTC	1798	9.9	93	174	ELECTRON	PETROL	1	
18	2014 Honc	Selangor	Seri Kemt	Used	HONDA	CIVIC	S	-	2014	150 000	- 14D SEDAN	5	MALAYSIA	5 SP AUTC	1798	10.6	104	174	FUEL INIEJ	PETROL	1	
19	ALL-IN 201	Selangor	Petaling J	Used	HONDA	CIVIC	TC-PREMI	-	2016	60 000	- 644D SEDAN	5	MALAYSIA	CONTINU	1498	10.6	127	220	TURBO F/I	PETROL	1	
20	2012 Honc	Melaka	Batu Bere	Used	HONDA	CR-Z	(HYBRID)	-	2012	80 000	- 842D HATCH	2	JAPAN	CVT AUTO	1497	10.4	90	167	FUEL INIEJ	PETROL	1	
21	ALL-IN 201	Selangor	Petaling J	Used	PROTON	X70	PREMIUM	-	2019	90 000	- 944D SUV	-	CHINA	6 SP AUTC	1799	-	-	-	-	PETROL	-	

Figure 11. Used_Car_Info.CSV

The script then extracts various car details from the JSON data on the page, such as the car's name, price, state, city, condition, brand, model, variant, series, manufacturing year, mileage, type, seat capacity, country of origin, transmission, engine type, fuel type, dimensions, weight, brakes, suspension, steering, and tyre and wheel specifications. It then writes all the information to the Figure 11 output CSV file.

```
Error extracting data from https://www.mudah.my/2010+Proton+SAGA+1+3+FL+H+-99608945.htm: '99608945'
Error extracting data from https://www.mudah.my/2015+Nissan+ALMERA+1+5+E+FACE+LIFT+A+-99556084.htm: '99556084'
Error extracting data from https://www.mudah.my/2009+Proton+SAGA+1+3+BASE+LINE+H+-99609760.htm: '99609760'
Error extracting data from https://www.mudah.my/2016+Ford+RANGER+2+2+XLT+FACE+LIFT+A+-99608613.htm: '99608613'
Error extracting data from https://www.mudah.my/2016+Ford+RANGER+2+2+XLT+FACE+LIFT+A+-99608613.htm: '99608613'
```

Figure 12. Error URL

CHAPTER 5: SYSTEM IMPLEMENTATION

If an error occurs during the extraction process, the script logs the error and continues with the next URL. It appears that Figure 12 some URLs in the final output have errors extracting data from the URL and are no longer valid. There are a couple of potential reasons for this. One possibility is that the sales list for the used car has been offline for an extended period. Alternatively, the car may have already been sold, causing the corresponding URL to no longer be accessible in the usual manner.

Data Quality Report**Summary Statistics:**

Used_Car_Info.csv	
Number of Rows	9684
Number of Columns	37
Missing Value	14980
Duplicate Rows	26
Column Name	Missing Value
name	0
state	0
city	0
condition	0
brand	0
model	0
variant	554
series	3152
mfg_year	0
mileage	0
type	15
seat_capacity	307
country_origin	39
transmission	0
engine_cc	0
compression_ratio	940
peak_power	313
peak_torque	306
engine_type	316
fuel_type	0
length	315
width	315
height	315
wheel_base	317

CHAPTER 5: SYSTEM IMPLEMENTATION

kerb_weight	349
fuel_tank	323
front_brakes	516
rear_brakes	516
front_suspension	553
rear_suspension	553
steering	1514
front_tyres	518
rear_tyres	541
front_rims	1185
rear_rims	1208
price	0
url	0

Column Name	Data Quality Issues
name	The names columns in the dataset contain additional information such as warranty offers, engine specifications, and other details, which could make it difficult to use the column for analysis purposes.
state	No issue
city	No issue
condition	The condition column has no missing value and only has one unique values. This has nothing to do with and helps with analyzing the model. So decided to drop this column.
brand	No issue
model	No issue
variant	The main problem with these two columns of data is that there are too many missing values, which are 554 and 3152 respectively. The missing value of the Series column has already accounted for 1/3 of the total data. So, decided to drop the series column
series	

CHAPTER 5: SYSTEM IMPLEMENTATION

mfg_year	The column mfg_year does not have any missing values and there is a class of data showing '1995 or older' which should be converted to 1995 and the data type is an object. So the data type needs to be converted to a numeric type.
mileage	The data in the mileage column seems to be binned into ranges rather than continuous values. This can be a problem if we are trying to use the mileage column for analysis or modeling.
type	The type column is inconsistent, as some entries contain extra information like the number of doors while others do not. Similarly, "DUAL CAB PICKUP" and "DOUBLE CAB PICKUP" could both refer to a pickup truck with a double cab, but it's not clear what the difference is between them.
seat_capacity	There are 307 missing values in the seat_capacity column, and the data type is the Object, which should be converted to int64 type. And choose to drop the remaining missing value.
country_origin	<p>The dataset on countries has several issues that need to be addressed before it can be analyzed accurately. One issue is the inconsistent spelling and capitalization of country names, such as "JERMANY" and "GREAT BRITIAN." Another issue is the presence of entries that are too specific or ambiguous, such as "Not Sufficient Data" and "AMERICA," which can make it difficult to analyze the data accurately. Additionally, there are multiple entries for the same country, such as "UNITED STATES OF AMERICA", "AMEICA" and "USA," which should be combined to avoid redundancy.</p> <p>To preprocess the data, several steps need to be taken. First, a mapping dictionary can be created to map different names for the same country to a standard name. For example, "UNITED STATES OF AMERICA", "AMERICA", and "USA" can be mapped to "UNITED STATES". This will standardize the names of the countries and make it easier to analyze the data accurately. Second, after the country names are standardized, they can be encoded as numerical values using label encoding</p>

CHAPTER 5: SYSTEM IMPLEMENTATION

	<p>or one-hot encoding. This will allow the data to be used in machine learning models.</p> <p>With these preprocessing steps, the dataset on countries will be clean, consistent, and ready for analysis. By standardizing the country names and encoding them as numerical values, the data can be used to gain valuable insights into the distribution of countries in the dataset.</p>
transmission	<p>Based on the provided data, the main data quality issue is the lack of consistency in the values of the "transmission" column. There are different spellings, abbreviations, and descriptions used to indicate the same type of transmission. For example, "CONTINUOUS VARIABLE" and "CONTINUOUS VARIABLE TRANSMISSION" both refer to the same thing. Similarly, "CVT" is used in diverse ways such as "AUTOMATIC CONSTANTLY VARIABLE (CVT)" and "CONSTANTLY VARIABLE (CVT)". These inconsistencies make it difficult to analyze the data and can lead to errors and biases in the results. To address this issue, data preprocessing methods such as standardization, normalization, or mapping can be applied to ensure that the values are consistent and can be interpreted accurately.</p>
compression_ratio	<p>Some values use commas instead of periods as decimal separators, which may cause issues when using the data in a regression model and the data type is an object. To preprocess this column, we can replace commas with periods to ensure that all values use the same decimal separator. We can then convert the column to a numerical data type, such as a float, to use in a regression model.</p>
engine_type	<p>The data quality issue with the engine_type column is that there are multiple formats used to describe the same type of engine, such as "MULTI POINT F/INJ" and "MULTI POINT F/INJ DIRECT INJECTION". This can make it difficult to analyze the data and draw meaningful insights.</p>

CHAPTER 5: SYSTEM IMPLEMENTATION

	<p>One approach to preprocess this data would be to standardize the format of the engine type names. This could involve merging similar engine types (e.g. "TURBO F/INJ" and "TURBO MPFI" could be combined into "TURBO MULTI POINT FUEL INJECTION") and removing redundant or unnecessary information (e.g. "ELECTRIC" and "1986" do not provide any useful information about the engine type). It would also be helpful to create a new column that categorizes the engines into broad groups such as gasoline, diesel, electric, hybrid, etc., to allow for more high-level analysis.</p>
fuel_type	The issue of this column is the unbalanced distribution data.
length	<p>These eight columns of data have 315, 315, 315, 317, 349, 313, 306, and 0 missing values respectively. The total number of these missing values is very close, so it is inferred that they are all collectively missing values. In addition, their data type is also an Object and must be converted to int64 type. Here we intend to drop missing values instead of replacing it with the mean value.</p>
width	
height	
wheel_base	
kerb_weight	
peak_power	
peak_torque	
engine_cc	
fuel_tank	<p>A data quality issue in the fuel_tank column is the presence of missing values (NaN). This means that some vehicles in the dataset do not have information about the fuel tank size. Also, there are values in this column that are not numbers, so we need to datatype them. Finally, some values in the column have decimal points, such as "42.8", which can be a problem if the data is used for calculations that require integer values.</p>
front_brakes	<p>The dataset on brake types has several issues that need to be addressed before it can be used for machine learning models. One issue is inconsistent capitalization, where the same brake types are written in different ways. For example, "Ventilated Disc" and "Ventilated discs" are both used. Another issue is the use of abbreviations like "NSD" and "DSABS," which can make it difficult to understand the meaning of the data.</p>
rear_brakes	

CHAPTER 5: SYSTEM IMPLEMENTATION

	<p>Additionally, there are typos like "VD" and "DS" that do not correspond to any known brake type.</p> <p>To preprocess the data, several steps need to be taken. First, standardizing the capitalization of all brake types, such as all lowercase or all uppercase, can solve the inconsistency issue. Second, expanding abbreviations like "NSD" and "DSABS" with their full meaning, if known, can make the data more understandable. Third, removing typos like "VD" and "DS" can eliminate data that has no meaning. Fourth, duplicate categories like "Ventilated Discs" and "Ventilated discs" can be merged into a single category.</p> <p>Finally, after preprocessing the data, the brake types can be encoded using numeric values or one-hot encoding to be used in machine learning models. With these preprocessing steps, the dataset on brake types will be clean, consistent, and ready for use in machine learning models.</p>
front_suspension	There are multiple values for the same type of suspension, and these values may not be consistent.
rear_suspension	
steering	In terms of preprocessing, we can consider grouping the different steering types into broader categories such as rack & pinion, recirculating ball, ball & nut, and worm & roller.
front_tyres	The data quality issue with the rear_tyres column is that the tire sizes are in different formats. Some are separated by a space (e.g. "175/65 R14"), while others are separated by a slash (e.g. "205/45R17"). This inconsistency in the format can cause problems when trying to analyze the data or build a model.
rear_tyres	
front_rims	To preprocess the data, we can standardize the tire sizes by separating the numbers and letters using regular expressions and then creating separate columns for each element of the tire size (e.g. width, aspect ratio, and rim size). We can also convert any missing values to a standard format (e.g. "NA") and potentially use imputation techniques to fill in missing values.

CHAPTER 5: SYSTEM IMPLEMENTATION

rear_rims	<p>The data quality is that the rim sizes are given in different formats, such as only the rim diameter (e.g., 17), the width and diameter separated by "x" (e.g., 5.5Jx15), and only the width (e.g., 8JJx18). This makes it difficult to process the data directly for use in a regression model.</p> <p>To preprocess this column, we can extract the rim diameter and width separately and store them in two new columns. For example, we can create a new column called front_rim_diameter and extract the numerical values from the front_rims column. Similarly, we can create another new column called front_rim_width and extract the width values</p>
price	<p>The price column does not have any missing values, but the data type is Object. Therefore, RM needs to be deleted, and only the following numbers are reserved as digital types.</p>
url	<p>The url column is used as a data reference and is convenient for querying when problems are found in the future. This column has nothing to do with the model. So, choose to drop the column</p>

5.4.2 Data Preparation

Then data preprocessing script for a dataset containing information about used cars, stored in a CSV file called "Used_Car_Info.csv". The script starts by reading the CSV file into a Pandas dataframe object using the `pd.read_csv()` function. The data in the dataframe is then cleaned and processed in several steps.

First, any missing values or placeholders (such as "-") in the data are replaced with NaN using the `replace()` function from NumPy. The 'name', 'condition', 'series', and 'url' columns are then removed from the dataframe using the `drop()` function. Next, the 'price' column is manipulated to remove the currency symbol ("RM") and any spaces and converted to a numeric data type using the `str.split()`, `str.join()`, and `pd.to_numeric()` functions.

The 'compression_ratio' column is cleaned by replacing any commas with periods and converting the resulting strings to float data types. Rows containing missing or invalid values are then removed using the `dropna()` function and the resulting data is filtered to include only values greater than or equal to 1.

Z-scores are then calculated for the 'compression_ratio' column and any rows with z-scores outside of a certain threshold (3 in this case) are removed. Similar cleaning steps are performed for the 'fuel_tank', 'peak_power', 'peak_torque', 'kerb_weight', 'wheel_base', 'width', 'length', and 'height' columns. Invalid or missing values are replaced or removed, and the resulting data is filtered to include only values within a certain threshold of the mean using z-scores.

Then the 'seat_capacity' column is converted to an integer data type, and the 'brand' column is filtered to only include the top 10 brands by count. Then, data rows with values of 'Terengganu', 'Perlis', and 'Putrajaya' in the 'state' column are dropped from the dataset, as are rows with values of certain countries in the 'country_origin' column. Rows with missing values in the 'country_origin' column are also dropped, and certain country names are standardized using a dictionary mapping.

Next, four lists of similar values for the 'steering' column are merged into a single value using the `replace` method. The 'mfg_year' column is converted to integers and any values of '1995 or older' are replaced with '1995'. The 'mileage' column is processed to remove spaces and replace

CHAPTER 5: SYSTEM IMPLEMENTATION

one value ('More than 500 000') with a range of values. Mileage values are then converted to a categorical variable by dividing them into ranges defined by a list of tuples. Labels for each range are assigned using a list of integers. A function is created to assign each mileage value to a level based on which range it falls into, and this function is applied to create a new 'mileage_level' column.

Finally, a dictionary is created to standardize certain values in the 'engine_type' column using regular expressions. The 'engine_type' column is then filtered to only include values that match specific keywords, and missing values are dropped. The resulting dataset is now cleaned and ready for further analysis.

After cleaning the data, the next crucial step is featuring selection, which plays a vital role in determining the performance of the final model. The principle of "garbage in, garbage out" holds true here, as including irrelevant features can negatively impact the model's performance while having too many features can lead to model complexity and overfitting [17]. Therefore, careful selection of features is of utmost importance.

To guide the feature selection process, a heatmap is utilized to visualize the relationships between different features [18]. This heatmap analysis provides insights into the correlations between features and the target variable, which is the price of used cars. By examining the heatmap, we can identify features that exhibit strong correlations with the price, indicating their significant impact on the target variable.

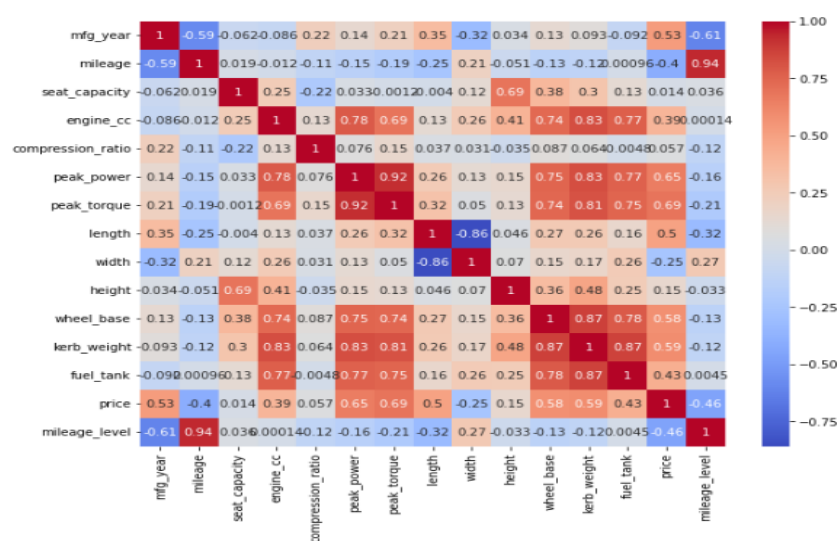


Figure 13. Heatmap

CHAPTER 5: SYSTEM IMPLEMENTATION

Upon analysing Figure 13, it becomes apparent that certain characteristics have correlations exceeding 0.5, indicating their strong influence on the car's price. These influential features include manufacturing year, peak power, peak torque, wheelbase, and curb weight. Moreover, features such as engine displacement, fuel tank, mileage level, and length show correlations close to 0.5, indicating their relevance in predicting the price of a used car. It is important to note that the heatmap primarily visualizes continuous variables, so the focus is shifted toward selecting prominent categorical features that also contribute to price prediction. These categorical features include the state, country of origin, brand, model, engine type, and fuel type.

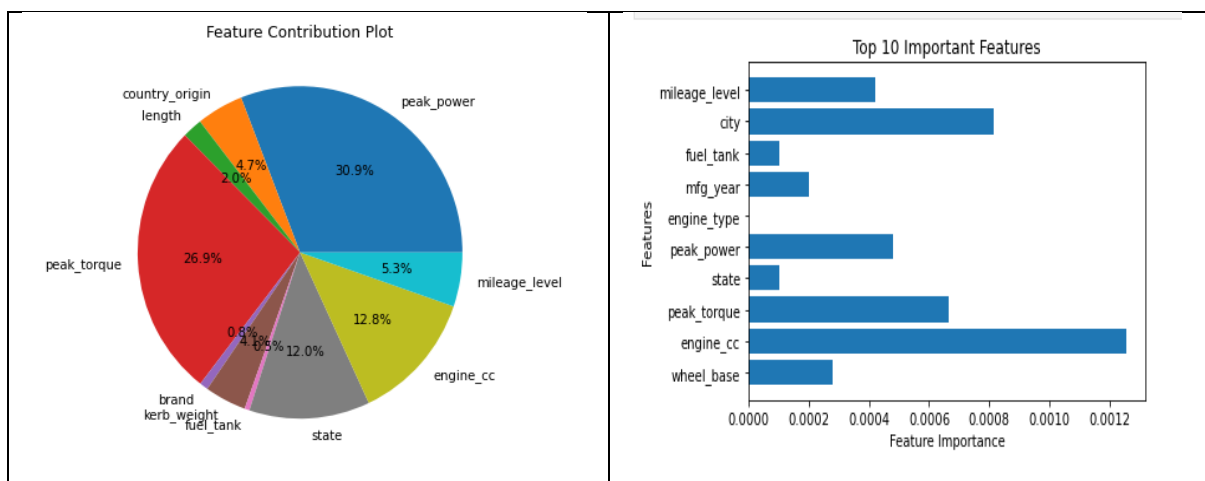
In addition to the heatmap analysis, feature importance analysis is performed using the Random Forest algorithm to identify the top 10 important features [19]. These models highlight that kerb weight, peak power, state, country of origin, engine displacement, peak torque, mileage level, fuel tank capacity, length, and brand are the most influential features in determining the price of used cars. This finding emphasizes the significance of considering these 10 characteristics when assessing the value of a car in the market. By paying attention to these influential factors, prospective buyers and sellers can make more informed decisions and gain valuable insights into the pricing dynamics of the used car market. Considering these influential features can provide meaningful guidance for individuals involved in the buying and selling processes, leading to more transparent and informed decision-making.

5.4.3 Modelling

In the modeling phase, our dataset, comprising a blend of categorical and continuous features, undergoes meticulous data preprocessing to harmonize and equip it for machine learning. Categorical features like 'state,' 'city,' 'brand,' 'model,' 'country_origin,' 'engine_type,' and 'fuel_type' are one-hot encoded, preserving their categorical nature without introducing erroneous ordinal relationships. Meanwhile, continuous features such as 'mfg_year,' 'mileage_level,' 'engine_cc,' and others are standardized to ensure a level playing field for our models. Then the data is split into training and testing datasets in the ratio 80:20, and the random state equal to 42.

Our modeling arsenal encompasses a suite of regression algorithms, each tailored to uncover unique patterns and relationships within the dataset. Linear Regression, Lasso Regression, Ridge Regression, ElasticNet Regression, Decision Tree Regressor, Random Forest Regressor, and an advanced Deep Neural Network (DNN) model make up our toolkit. The hyperparameters of each model undergo meticulous tuning using GridSearchCV, ensuring peak performance. For instance, Lasso, Ridge, and ElasticNet regressors fine-tune regularization through hyperparameter tuning, while Decision Tree Regressor and Random Forest Regressor explore different tree depths and quantities, respectively.

Table 3. Important Feature of Random Forest vs XGBoost



Understanding the driving forces behind our predictions is essential. To achieve this, we employ the Random Forest Regressor and XGBoost, renowned for its feature importance analysis. This technique ranks and calculates the significance of each feature in predicting used

CHAPTER 5: SYSTEM IMPLEMENTATION

car prices. Table 3 shows the top 10 most influential features are presented, providing valuable insights into the factors shaping our price predictions.

Within our toolkit, an advanced Deep Neural Network (DNN) model is a standout contender. Its hyperparameters are optimized to achieve peak performance, considering factors such as hidden layer sizes, activation functions, solvers, learning rates, and alpha values. The DNN's adaptability and capacity for handling intricate relationships make it a formidable asset in predicting used car prices.

5.5 System Operation (with Screenshot)

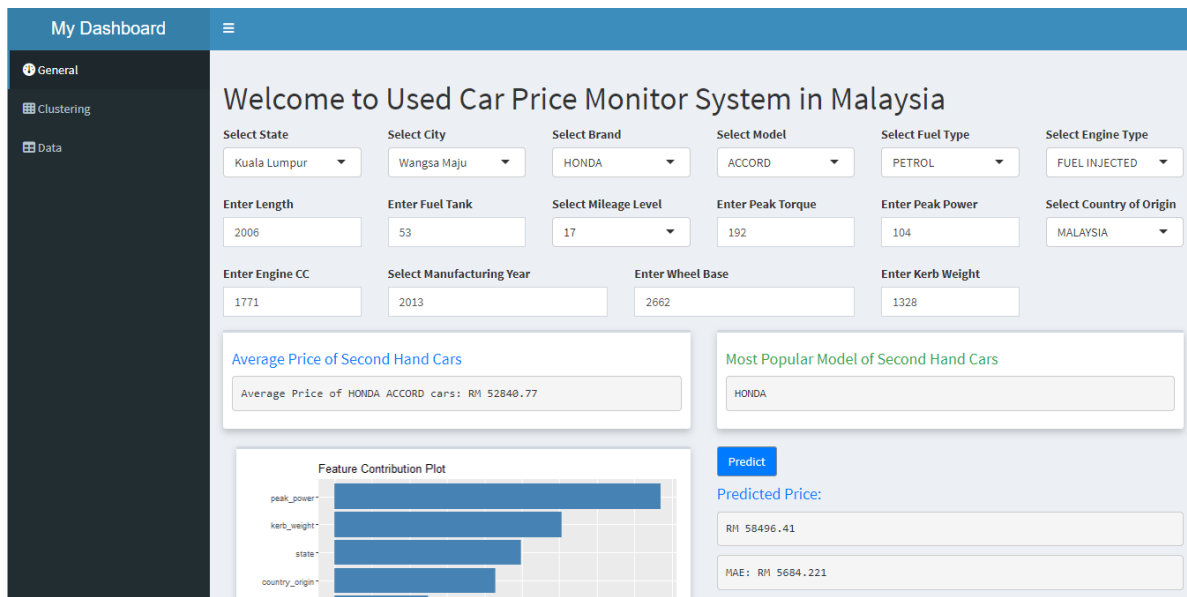


Figure 14. General Page of Dashboard

The system's operation is orchestrated through its main dashboard, as illustrated in the accompanying screenshots. The dashboard, divided into two key sections, presents an accessible and user-friendly interface for seamless navigation.

Upon entering the Figure 14 dashboard, users are warmly welcomed by a friendly greeting. The central feature of the dashboard is the filter functionality, which allows users to input a wide array of car features to facilitate price prediction. These encompass 16 crucial attributes, including state, city, brand, model, fuel type, and more. Importantly, these attributes are classified into numerical and categorical inputs, ensuring versatility in user interaction.

Noteworthy is the inclusion of default values for numerical inputs. These default values are dynamically generated based on mean dataset values, providing users with initial estimates for each category, grounded in the dataset's average metrics.

The dashboard also offers valuable statistical insights. Users can discover the brand with the highest occurrence in the dataset and the average car price. Initially, the average price mirrors the entire dataset. However, it dynamically adapts to reveal the average price specific to the chosen brand and model.

CHAPTER 5: SYSTEM IMPLEMENTATION

For users seeking deeper insights, a graphical chart visually displays the top 10 features that wield the most substantial influence on car prices. This chart offers an intuitive and informative overview of the primary factors influencing pricing decisions.

Facilitating predictive capabilities, the dashboard integrates a 'Predict' button. Once engaged, this button triggers the price prediction process based on user input. Following this action, the system promptly provides the estimated car price, Mean Absolute Error (MAE), and a price range. This range is determined by considering the predicted value and MAE.

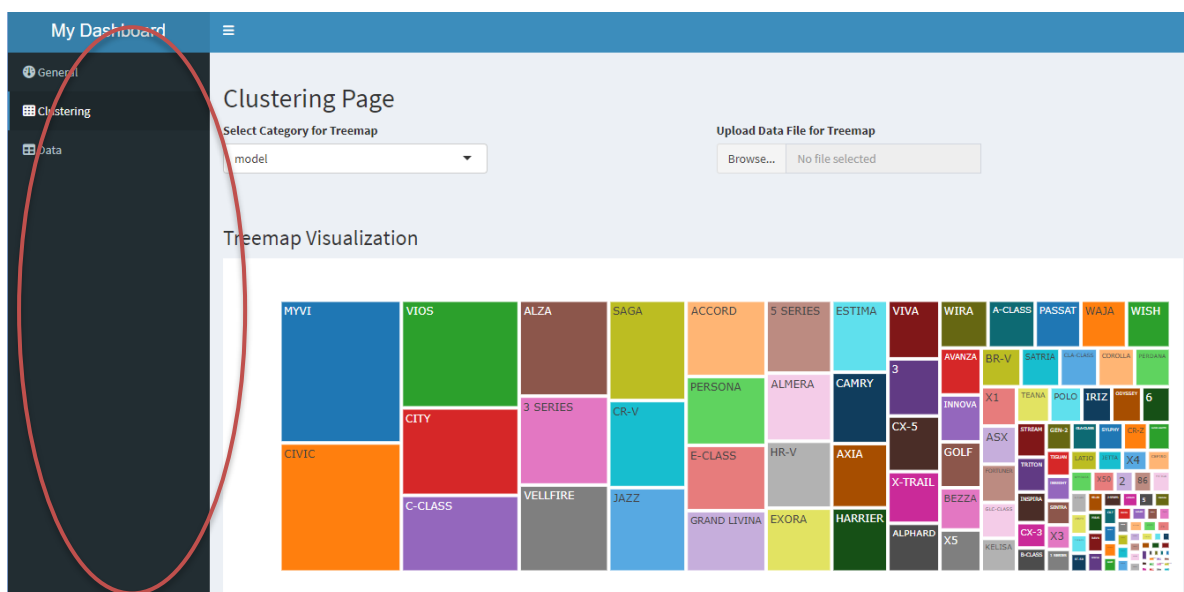


Figure 15. Clustering Page of Dashboard

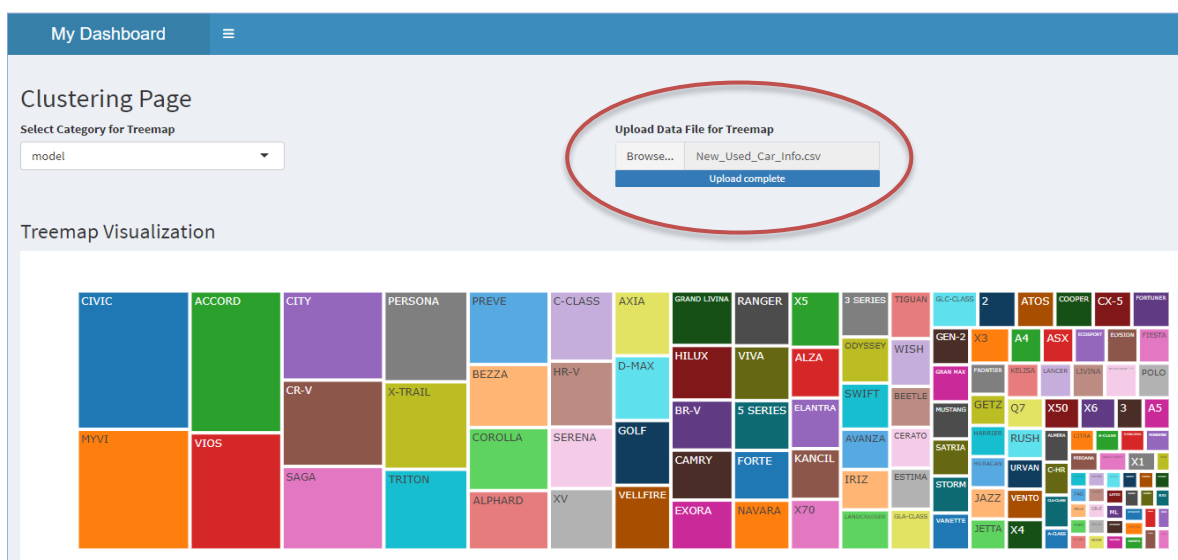


Figure 16. Upload File

CHAPTER 5: SYSTEM IMPLEMENTATION

Continuing to the Figure 15 second page, users can easily switch to the 'Clustering' section by clicking on the navigation bar. Here, the system employs treemaps to present a visual representation of various clusters within the dataset. Users have the flexibility to select the features they want to explore, leading to the display of corresponding data clusters. Furthermore, Figure 16 for users interested in assessing the distribution of their unique datasets, there's an option to upload their files. Upon uploading, the system initiates preprocessing procedures and subsequently visualizes the data using treemaps. To maximize the screen real estate for data exploration, users can conveniently collapse the navigation bar by clicking on 'Dashboard' at the top of the page.

My Dashboard

Data Page

Upload CSV File

Browse... No file selected

Show 10 entries

Search:

	state	city	brand	model	variant	mfg_year	mileage	type	seat_capacity	country_origin	transmission	engin
1	Kuala Lumpur	Wangsa Maju	HONDA	ACCORD	VTI-L	2001	2499	4D SEDAN	5	MALAYSIA	5 SP AUTOMATIC	
2	Kedah	Alor Setar	TOYOTA	VIOS	G	2008	2499	4D SEDAN	5	MALAYSIA	4 SP AUTOMATIC	
3	Selangor	Glenmarie	BMW	5 SERIES	30I M-SPORT (CKD)	2018	2499	4D SEDAN	5	MALAYSIA	8 SP AUTOMATIC CONVENTIONAL	
4	Kedah	Sungai Petani	PROTON	WAJA	CAMPRO VERSION A	2006	2499	4D SEDAN	5	MALAYSIA	4 SP AUTOMATIC	
5	Kedah	Sungai Petani	HONDA	CITY	I-DSI	2008	2499	4D SEDAN	5	MALAYSIA	CVT AUTO 7 SP SEQUENTIAL	
6	Selangor	Ampang	PERODUA	BEZZA	X	2021	2499	4D SEDAN	5	MALAYSIA	4 SP AUTOMATIC	

Figure 17. Data Page of Dashboard

My Dashboard

Data Page

Upload CSV File

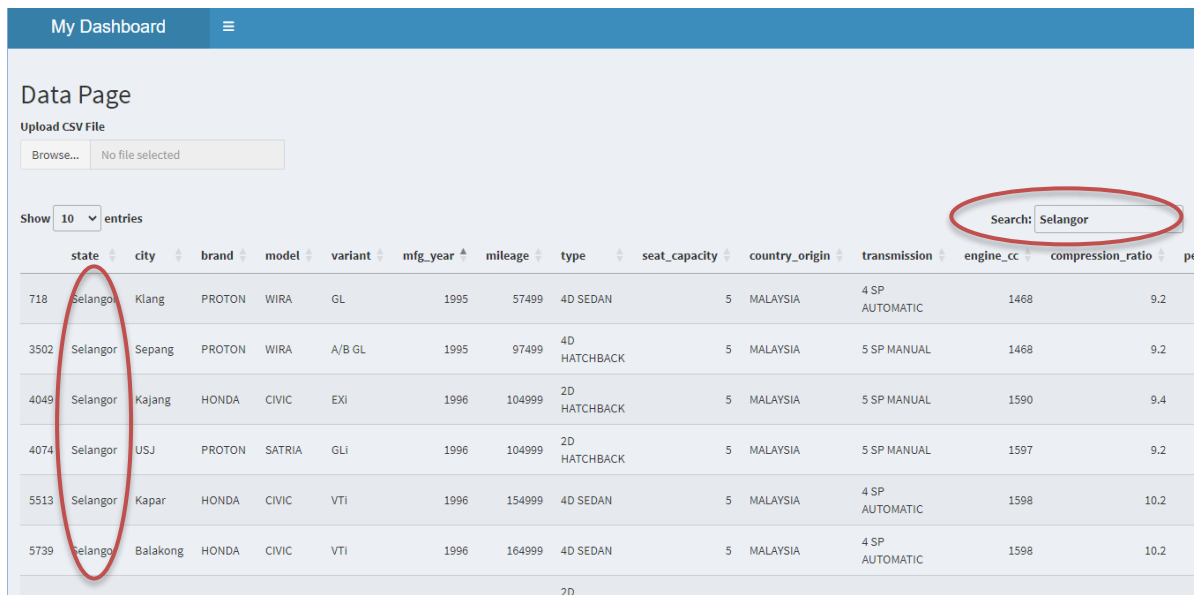
Browse... No file selected

Show 10 entries

Search:

	state	city	brand	model	variant	mfg_year	mileage	type	seat_capacity	country_origin	transmission	engine_cc	compression_ratio
16	Perak	Ipoh	TOYOTA	ESTIMA		1995	2499	4D WAGON	7	JAPAN	4 SP AUTOMATIC	2438	8.9
156	Kuala Lumpur	Gombak	PROTON	WIRA	GL	1995	27499	4D SEDAN	5	MALAYSIA	5 SP MANUAL	1468	9.2
718	Selangor	Klang	PROTON	WIRA	GL	1995	57499	4D SEDAN	5	MALAYSIA	4 SP AUTOMATIC	1468	9.2
1852	Kelantan	Pasir Puteh	PROTON	WIRA	GL	1995	77499	4D SEDAN	5	MALAYSIA	5 SP MANUAL	1468	9.2
3502	Selangor	Sepang	PROTON	WIRA	A/B GL	1995	97499	4D HATCHBACK	5	MALAYSIA	5 SP MANUAL	1468	9.2
3523	Pahang	Kuantan	HONDA	CIVIC	EXI	1995	97499	4D SEDAN	5	JAPAN	5 SP MANUAL	1797	9.4

Figure 18. Sorting



My Dashboard

Data Page

Upload CSV File

Browse... No file selected

Show 10 entries

Search: Selangor

	state	city	brand	model	variant	mfg_year	mileage	type	seat_capacity	country_origin	transmission	engine_cc	compression_ratio
718	Selangor	Klang	PROTON	WIRA	GL	1995	57499	4D SEDAN	5	MALAYSIA	4 SP AUTOMATIC	1468	9.2
3502	Selangor	Sepang	PROTON	WIRA	A/B GL	1995	97499	4D HATCHBACK	5	MALAYSIA	5 SP MANUAL	1468	9.2
4049	Selangor	Kajang	HONDA	CIVIC	EXI	1996	104999	2D HATCHBACK	5	MALAYSIA	5 SP MANUAL	1590	9.4
4074	Selangor	USJ	PROTON	SATRIA	GLI	1996	104999	2D HATCHBACK	5	MALAYSIA	5 SP MANUAL	1597	9.2
5513	Selangor	Kapar	HONDA	CIVIC	VTI	1996	154999	4D SEDAN	5	MALAYSIA	4 SP AUTOMATIC	1598	10.2
5739	Selangor	Balakong	HONDA	CIVIC	VTI	1996	164999	4D SEDAN	5	MALAYSIA	4 SP AUTOMATIC	1598	10.2

Figure 19. Searching

Moving on to the Figure 17 third page, users can access the 'Dataset View,' offering valuable insights into the training data. The interface allows users to gain an understanding of the dataset's structure. For further convenience, users can sort the data features Figure 18 in ascending or descending order by clicking on the adjacent up and down arrows. Moreover, a search feature Figure 19 enables users to find specific data points within the dataset swiftly. Additionally, users have the option to upload their datasets, seamlessly integrating their own data with the current dataset for comprehensive analysis and comparison.

This multi-page system not only simplifies the process of exploring and analyzing car price data but also empowers users to interact with the data, customize their views, and integrate their own datasets, making it a versatile and user-centric tool for used car price monitoring and analysis.

5.6 Implementation Issues and Challenges

In the process of implementing our system using R and RStudio to create a Shiny app, we have encountered several notable challenges and issues. One significant challenge revolves around the need to grasp a new programming language. For individuals more accustomed to Python, the transition to R requires a considerable investment of time and effort to become proficient in R's syntax and conventions.

A pivotal decision was made to opt for R over Python due to the availability of a richer set of resources, tutorials, and documentation for building Shiny apps within the R ecosystem. Although this choice may expedite the learning curve, it presents a conundrum for those who have a stronger foundation in Python.

Another noteworthy challenge lies in the need to familiarize ourselves with various R packages, such as “ggplot” for data visualization. These packages offer robust features, but mastering them takes time, making this a formidable challenge, especially for those less experienced with R.

One of the most significant hurdles is the transition from Python-centric workflows to an R-based environment. Tasks such as data preprocessing, cleaning, and model development, which were previously executed in Python, need to be adapted to R's idiosyncrasies.

Furthermore, as data preprocessing and cleaning are critical stages in machine learning, transferring these tasks from Python to R involves rewriting code, ensuring data consistency, and addressing language-specific nuances.

The development of machine learning models in R also presents challenges, necessitating a profound understanding of R's libraries and tools. This involves either converting existing Python models or crafting new ones, which can be a complex endeavour.

CHAPTER 6: SYSTEM EVALUATION AND DISCUSSION

Chapter 6 serves as an evaluation and reflection on the project's development and its outcomes. It discusses the challenges faced and how they were overcome, presents the results of the machine learning models, and assesses the fulfilment of the project's objectives.

6.1 System Testing and Performance Metrics

The project's success hinges on rigorous system testing and the judicious use of performance metrics, especially given its focus on predicting used car prices through supervised machine learning techniques. Several key performance metrics guide the evaluation of the prediction model's accuracy and effectiveness. First, the R-squared (R^2) metric measures the model's ability to explain the variance in used car prices. A higher R^2 score indicates a better fit to the data. Second, the Mean Absolute Error (MAE) quantifies the average magnitude of prediction errors, providing insight into the model's precision. Third, the Mean Squared Error (MSE) computes the average of squared prediction errors, making it sensitive to larger deviations. Lastly, the Root Mean Squared Error (RMSE), derived from MSE, gauges prediction error in the same units as the target variable.

System testing plays a pivotal role in evaluating the functionality and reliability of the dashboard, which serves as the project's user interface. Testing efforts encompass multiple aspects. First and foremost, price prediction functionality undergoes meticulous testing to ascertain the model's accuracy in forecasting used car prices. Additionally, data visualization features, including treemaps and charts, are scrutinized to ensure they provide users with meaningful insights into price distributions and trends. File upload functionality is rigorously tested to guarantee the smooth processing and integration of external datasets. Extensive user interface testing ensures that navigation, buttons, and filters are intuitive and user-friendly. Finally, integration testing ensures seamless data integration when users upload their datasets, maintaining data integrity. By focusing on these metrics and conducting comprehensive system testing, the project ensures an effective used car price monitoring system, providing accurate predictions and robust data analysis tools for users.

6.2 Testing Setup and Result

Table 4. Learning Algorithms Performance Using R2 on Test Set

Learning Algorithm	R2	Cross-validated R2
Linear Regression	0.87	0.82
Lasso	0.87	0.82
Ridge	0.87	0.82
ElasticNet	0.87	0.81
Decision Tree	0.89	0.85
Random Forest	0.97	0.94
XGBoost	0.97	0.95
Deep Neural Network	0.96	0.96

According to **Error! Reference source not found.**, Linear Regression, Lasso, Ridge, and ElasticNet are linear regression-based models that assume a linear relationship between the input features and the target variable. These models have shown reasonably good performance with an R2 score of 0.87 and a cross-validated R2 score of 0.82. This indicates that the linear regression models can capture a significant portion of the variation in the target variable based on the given input features. The assumption of linearity holds well for the dataset, suggesting that these models can provide useful predictions.

On the other hand, the non-linear Decision Tree model, which has an R2 score of 0.89 and a cross-validated R2 score of 0.85, has fared marginally better. This shows that more intricate connections between the input data and the target variable can be captured by the Decision Tree model. Decision trees can capture non-linear patterns and interactions by segmenting the feature space, improving predictions.

However, the Random Forest ensemble model has demonstrated significantly higher performance with an R2 score of 0.97 and a cross-validated R2 score of 0.94. By combining multiple decision trees and aggregating their predictions, Random Forest reduces overfitting and captures more intricate relationships in the data. This results in highly accurate predictions and a better fit for the underlying patterns in the dataset.

CHAPTER 6: SYSTEM EVALUATION AND DISCUSSION

Similarly, the XGBoost model, another ensemble model based on gradient boosting, has achieved impressive performance with an R² score of 0.97 and a cross-validated R² score of 0.95. XGBoost iteratively improves the model's performance by focusing on the instances that are harder to predict. This boosting technique optimizes the model and enables it to handle complex relationships and interactions effectively, leading to superior predictive performance. Lastly, the Deep Neural Network (DNN) model has demonstrated strong performance with an R² score of 0.96 and a cross-validated R² score of 0.96. DNNs are highly flexible and can capture intricate non-linear relationships and hierarchies within the data. This makes them well-suited for modelling complex datasets and extracting meaningful patterns, resulting in accurate predictions.

In conclusion, the higher-performing models such as Random Forest, XGBoost, and DNN leverage their capabilities to handle complex relationships and interactions, surpassing the performance of linear regression-based models. These advanced models provide superior predictive accuracy and are better equipped to capture the nuances of the used car pricing data, ultimately enabling more accurate predictions and informed decision-making.

CHAPTER 6: SYSTEM EVALUATION AND DISCUSSION

Table 5. Summarize of metrics for different model

	Linear	Lasso	Ridge	Elastic Net	Decision Tree	Random Forest	XGBoost	Deep Neural Network
MSE	3174044 59.98485 66	3168006 17.81992 8	3174705 20.38153 005	3167839 84.76413 98	2799952 46.81478 3	84384496. 34709692	96214780. 47644088	98577018. 97351588
MAE	11680.86 5200130 995	11623.27 2444657 587	11737.41 8639877 233	11659.94 1840241 61	10721.20 6812748 644	5221.6189 47963116	5435.1461 98512878	5743.1862 60177717 5
RSME	17815.84 8562020 743	17798.89 3724609 065	17817.70 2443960 894	17798.42 6468767 957	16733.05 8501504 827	9186.1034 36555508	9808.9133 17816652	9928.5960 2227404
R2	0.873031 6168816 632	0.873273 1663020 568	0.873005 1913495 387	0.873279 8198701 882	0.887996 0799209 036	0.9662444 47031534 2	0.9615120 87417542 1	0.9605671 42905647 7

The

CHAPTER 6: SYSTEM EVALUATION AND DISCUSSION

Table 5 shows the results of the evaluation metrics for different machine-learning models used to predict car prices. The models evaluated are Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression, Decision Tree, Random Forest, and XGBoost, Deep Neural Network.

For the MSE metric, the Random Forest model has the lowest value of 84,384,496.35, indicating the model's ability to accurately predict car prices. For the MAE metric, the Random Forest model also has the lowest value of 5,221.62, indicating that the model's predictions are closer to the actual values.

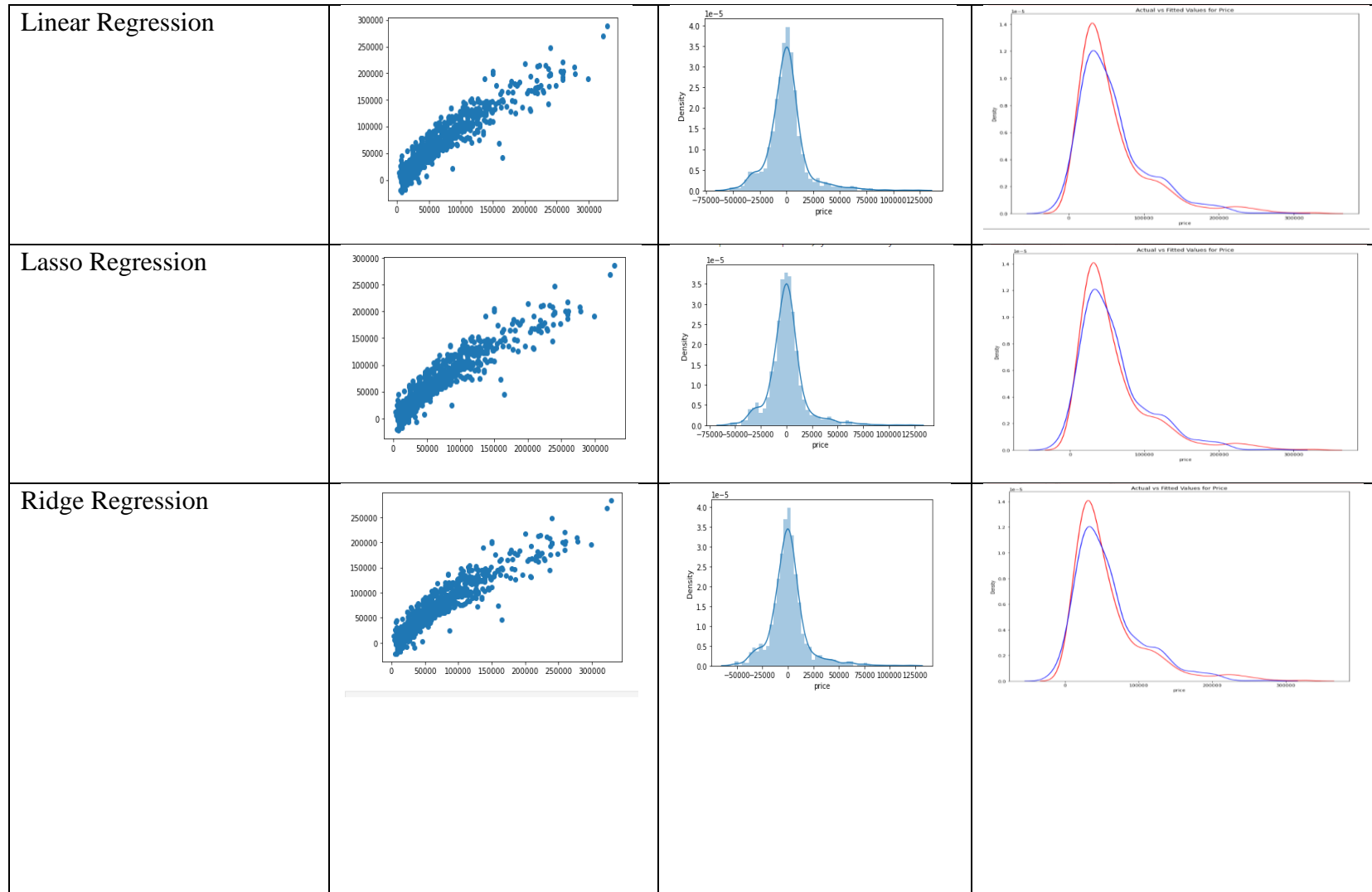
For the RSME metric, the Random Forest model has the lowest value of 9,186.10, indicating that the model has the smallest error between the predicted and actual car prices. For the R2 metric, the Random Forest model also performs the best, with a score of 0.9662, indicating that the model explains 96.62% of the variance in the data.

In summary, the Random Forest model outperforms the other models in all the evaluation metrics, making it the most suitable model for predicting car prices. However, it's worth noting that the other models also perform reasonably well.

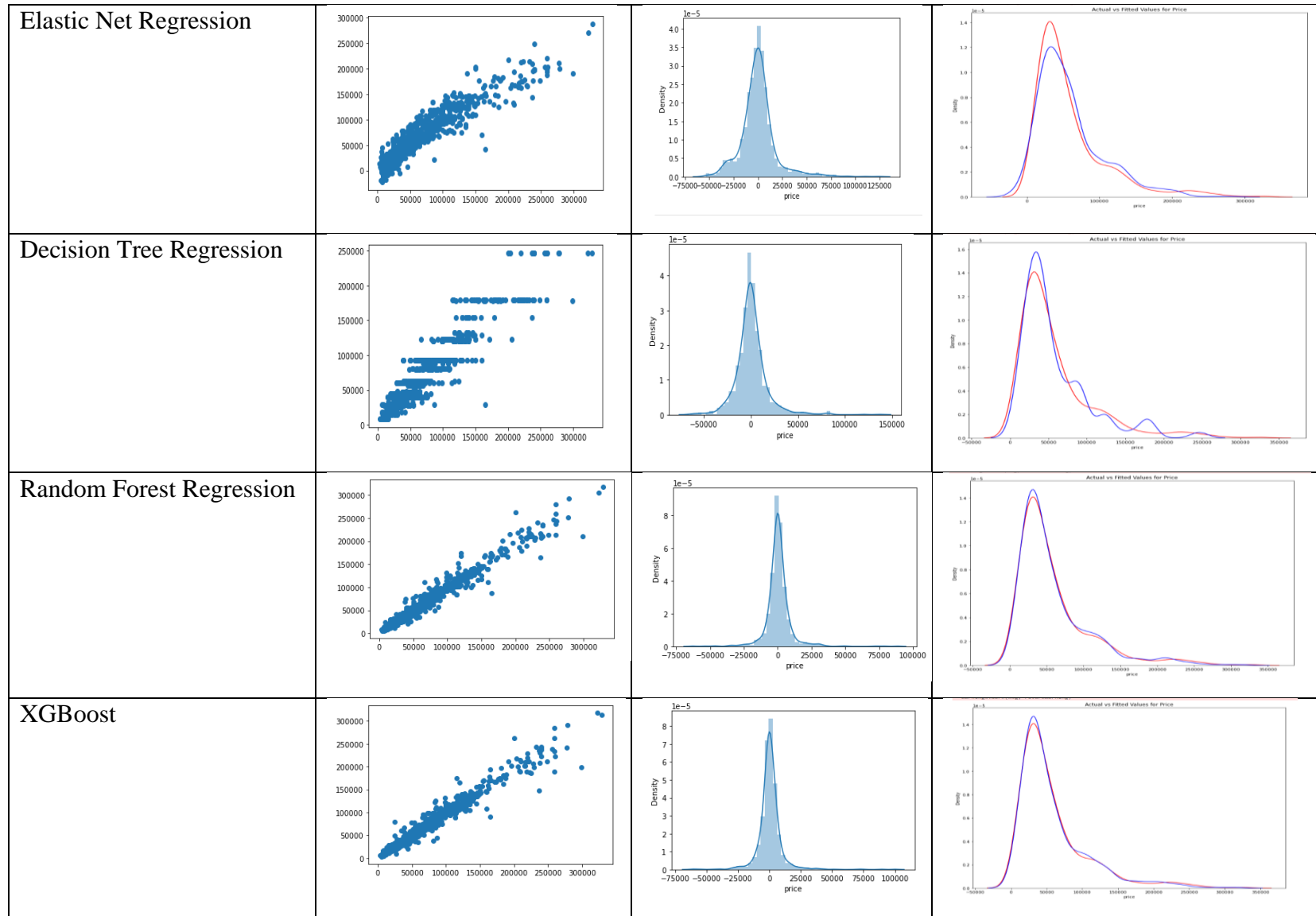
Ranking of Training Model (from best to worse by observing data):

1. Deep Neural Network
2. Random Forest Regression
3. XGBoost
4. Decision Tree Regression
5. Elastic Net Regression
6. Lasso Regression
7. Linear Regression
8. Ridge Regression

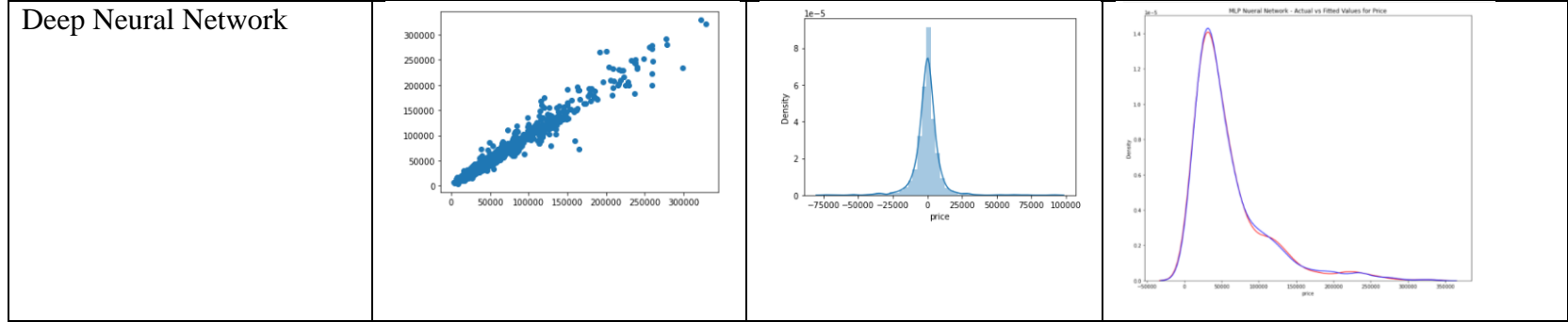
Table 6. Comparison of Different Models



CHAPTER 6: SYSTEM EVALUATION AND DISCUSSION



CHAPTER 6: SYSTEM EVALUATION AND DISCUSSION



6.3 Project Challenges

Throughout the development of the project, several significant challenges were encountered, each of which played a pivotal role in shaping the system's overall performance and effectiveness.

Firstly, the project grappled with the issue of Limited Data Availability. The process of collecting data through web scraping proved to be not only time-intensive but also yielded a dataset of relatively modest proportions. This shortage of extensive data can potentially hinder the model's ability to generalize effectively, which, in turn, might restrict the accuracy of its predictions.

Additionally, the project confronted the obstacle of an Imbalanced Dataset. This dataset exhibited an uneven distribution of classes, with certain categories being disproportionately represented, while others remained underrepresented. This imbalance can introduce biases into the model's predictions, necessitating the application of specialized techniques such as resampling or advanced algorithms to rectify.

Furthermore, the project encountered Data Quality Issues within the dataset. These issues encompassed missing data points, the presence of outliers, and inconsistencies in the dataset. Addressing these issues and performing comprehensive data cleaning and preprocessing demanded a substantial investment of effort and attention.

Another noteworthy challenge was the presence of High-Dimensional Features in the dataset. Certain features contained an extensive array of categories, leading to increased dimensionality. Managing and processing these high-dimensional features posed computational challenges, mandating careful consideration and skilful handling.

Hyperparameter tuning was yet another obstacle to be surmounted. Optimizing the model's hyperparameters was a demanding task, necessitating multiple iterations and substantial computational resources to identify the optimal combination.

Lastly, the introduction of a new programming language, R, and the transition from Python introduced a layer of complexity into the project's implementation. Gaining proficiency in R

CHAPTER 6: SYSTEM EVALUATION AND DISCUSSION

and addressing its language-specific nuances added both time and effort to the development process.

In overcoming these challenges, the project not only succeeded in achieving its objectives but also demonstrated adaptability and problem-solving prowess. The experience gained in addressing data limitations, improving data quality, and fine-tuning the model serves as a valuable resource for future endeavours in the realms of machine learning and data analysis.

6.4 Objectives Evaluation

Machine Learning-Based Price Prediction Model

The primary objective of this project was to develop a machine learning-based price prediction model that accurately predicts the price of used cars. To assess the achievement of this objective, various machine learning algorithms were deployed, including linear regression, lasso, ridge, elastic net, decision tree, random forest, XGBoost, and a deep neural network. The model performance was evaluated using the R-squared (R^2) metric, with the most successful models achieving an R^2 of over 0.95. This demonstrates that the project successfully met its goal of creating an accurate price prediction model.

Integration into a Monitoring System

Another key objective was the integration of the developed model into a monitoring system to provide buyers and sellers with improved pricing information. The deployment of a web-based dashboard allowed users to access real-time price comparisons for different car models. This objective was successfully achieved, enhancing transparency, and facilitating better decision-making in the used car market.

Identification of Challenges and Limitations

The project aimed to identify challenges and limitations in the current second-hand car market in Malaysia. This objective was effectively addressed through a thorough problem statement analysis. Challenges such as pricing opacity and information asymmetry were identified, setting the stage for potential solutions.

Proposal of Potential Solutions

Building on the identification of market challenges, the project proposed potential solutions to improve transparency and fairness for both buyers and sellers. While the actual implementation of these solutions might require further action, the project succeeded in presenting viable strategies to address market issues.

Feature Importance Analysis

To determine the most critical features influencing the price of used cars in Malaysia, feature importance analysis was conducted. Factors such as the state, country of origin, engine cc, peak power, peak torque, fuel tank capacity, brand, mileage level, length, and kerb weight were

identified as the most significant contributors to pricing. This analysis provides valuable insights for market participants and stakeholders.

Model Performance

The project's machine learning models exhibited high performance, with Random Forest and XGBoost models achieving R-squared values of 0.97 and cross-validated R-squared values of 0.94 and 0.95, respectively. These results indicate the robustness of the predictive models. The Deep Neural Network model also performed exceptionally well, achieving an R-squared value of 0.96 and a cross-validated R-squared value of 0.96. This underscores the versatility of machine learning techniques in accurately predicting used car prices.

Dashboard Deployment

The project successfully deployed a user-friendly and interactive dashboard that allows users to compare real-time prices of different car models. This achievement enhances market monitoring and aids users in making informed decisions.

CHAPTER 7: CONCLUSION AND RECOMMENDATION

This concluding chapter serves as the culmination of the Used Car Price Monitoring System project. Its primary purpose is to provide a comprehensive summary of the project's outcomes, achievements, and the contributions it brings to the domain of used car trading in Malaysia. Additionally, this chapter outlines a set of strategic recommendations for future work, aiming to further enhance and expand upon the project's successes.

7.1 Conclusion

In conclusion, the development and implementation of the Used Car Price Monitoring System have yielded a valuable tool for individuals in Malaysia looking to buy or sell used cars. This project leveraged the power of machine learning, particularly supervised learning, to predict used car prices accurately, providing users with a reliable estimation of a car's market value.

In the final iteration of this project, a Neural Network-based price prediction model was implemented, resulting in an impressive R2 score of approximately 96%. This performance underscores the system's remarkable accuracy in predicting the prices of used cars in Malaysia. Furthermore, the model's cross-validated R2 score, which also reached 96%, validates its robustness and reliability.

Throughout this project, the machine learning model has consistently demonstrated its effectiveness in providing highly accurate price predictions. The feature selection process identified critical factors, including peak power, kerb weight, state, country of origin, engine cc, fuel tank capacity, mileage level, peak torque, vehicle length, and brand, as the primary determinants influencing the prices of second-hand cars.

The project's foundation lies in a user-friendly and interactive dashboard, which not only predicts prices but also offers data visualization capabilities through treemaps and charts. This enables users to explore the dataset comprehensively and gain insights into pricing trends and distributions.

Despite encountering several challenges, including data collection, data quality, and the need to transition from Python to R for specific functionalities, the project successfully addressed

CHAPTER 7: CONCLUSION AND RECOMMENDATION

these issues and delivered a functional system. Moreover, system testing ensured that the dashboard operated seamlessly, delivering on its promise to provide accurate predictions, robust data visualization, and user-friendly functionality.

7.2 Recommendation

The successful execution of this project has established a solid foundation for enhancing the transparency and efficiency of the Malaysian used car market. To further refine and expand upon these achievements, several comprehensive recommendations are proposed for future work.

First and foremost, an imperative step is the expansion of our dataset, which currently comprises approximately 9,000 entries. To significantly bolster the accuracy and representativeness of our predictive models, the dataset should be substantially augmented. This augmentation necessitates the acquisition of data from an array of sources, encompassing not only online platforms but also direct market surveys. By incorporating information from a diverse range of channels, our dataset will attain a heightened level of comprehensiveness. This, in turn, will enable us to capture an even broader spectrum of market trends and patterns, thus amplifying the reliability and robustness of our predictive models. Additionally, this enriched dataset will facilitate the identification of emerging market dynamics and further empower stakeholders in their decision-making processes.

Addressing potential data imbalances within our dataset, with a particular focus on rectifying underrepresentation issues concerning certain car manufacturers and models, is of paramount importance. To mitigate the inherent biases and constraints that may be introduced by such imbalances, proactive efforts should be made to collect additional data for these underrepresented entities. By deliberately seeking out and acquiring data related to these underrepresented segments, our dataset will evolve into a more balanced and representative resource. This balanced representation will significantly reduce the likelihood of underfitting in our models, ensuring they effectively encapsulate the nuanced intricacies of the market. This, in turn, will bolster the credibility of insights derived from our models, enabling all stakeholders to make more informed and trustworthy decisions.

In addition to expanding and balancing the dataset, the integration of online learning techniques into our models is recommended as a forward-looking strategy. The adoption of online learning algorithms, such as incremental learning or partial fit, will endow our models with the capability to continuously adapt and evolve. This adaptability is essential for staying synchronized with the dynamic and ever-evolving nature of the Malaysian second-hand car

CHAPTER 7: CONCLUSION AND RECOMMENDATION

market. Through online learning, our models can effectively absorb, and process newly acquired data while preserving previously acquired knowledge. This progressive approach will ensure that our models remain up-to-date and relevant, providing stakeholders with real-time insights into the market's changing dynamics.

Moreover, the continuous maintenance and enhancement of the deployed dedicated and interactive dashboard are of utmost importance. This dashboard, already in operation, serves as a pivotal tool for market monitoring and data visualization. Ensuring its seamless functionality and relevance is critical. The dashboard should be regularly updated to accommodate new data sources, integrate advanced data visualization techniques, and remain aligned with evolving user requirements. By proactively maintaining and upgrading the dashboard, stakeholders will perpetually benefit from a user-friendly interface that offers real-time access to comprehensive market-related information. This ongoing commitment to dashboard optimization will enable stakeholders to stay well-informed, promptly detect emerging trends, and adeptly navigate the ever-dynamic market landscape.

In conclusion, the project's current accomplishments have set the stage for an even more transparent, efficient, and effective Malaysian used car market. The proposed recommendations, including dataset expansion, data balance enhancement, online learning integration, and dashboard deployment, collectively represent a multifaceted strategy to further elevate the precision, comprehensiveness, and utility of our market data. Implementation of these measures will not only advance the credibility and representativeness of our models but also provide invaluable support to buyers and sellers in their decision-making processes. Ultimately, this will foster a more informed, competitive, and equitable second-hand automobile market in Malaysia, benefiting all stakeholders and contributing to the market's sustainable growth.

ACHIEVEMENT

The successful presentation of our paper, titled "Revolutionizing the Malaysian Used Car Market: A Machine Learning Approach to Transparent Pricing," at the prestigious 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS2023) marks a significant milestone in our project's journey. This paper is accepted for publication by IEEE Xplore, and we also gave a presentation last week on September 7th.

Revolutionizing the Malaysian Used Car Market: A Machine Learning Approach to Transparent Pricing

Scott Lai Yong Luo
Faculty of Information and
Communication Technology
Universiti Tunku Abdul Rahman
Kampar, Malaysia
scottlaiyongluo1109@gmail.com

*Abdulkarim M. Jamal Kanaan
Faculty of Information and
Communication Technology
Universiti Tunku Abdul Rahman
Kampar, Malaysia
a.kanaan@msn.com
abdulkarim@utar.edu.my

Ramesh Kumar Ayyasamy
Faculty of Information and
Communication Technology
Universiti Tunku Abdul Rahman
Kampar, Malaysia
rameshkumar@utar.edu.my

Goh Chuan Meng
Faculty of Information and
Communication Technology
Universiti Tunku Abdul Rahman
Kampar, Malaysia
gohcm@utar.edu.my

Boon-Yaik Goh
Faculty of Information and
Communication Technology
Universiti Tunku Abdul Rahman
Kampar, Malaysia
coibye@utar.edu.my

Chai Meei Tyng
Faculty of Information and
Communication Technology
Universiti Tunku Abdul Rahman
Kampar, Malaysia
chaimti@utar.edu.my

Abstract— In Malaysia, the used car market frequently lacks transparency and fair pricing, making it difficult for buyers and sellers to make wise decisions. In-depth research on the creation of a machine learning-based pricing prediction model created exclusively for the Malaysian used car market is presented in this paper. The main goal of this project is to develop a reliable and open model that predicts used car pricing based on essential variables including age, condition, location, and the availability of comparable models on the market. Numerous techniques, like the neural network and regression model, are used to accomplish this purpose. The scope of the project also encompasses the identification of challenges and limitations in the current used car market in Malaysia, along with proposed solutions to improve transparency and fairness for all stakeholders. The methodology involves data collection, preprocessing, feature selection, model training, and evaluation. The results demonstrate that the developed model provides precise and transparent pricing information, empowering buyers and sellers to make informed decisions regarding their transactions. This project holds significant potential for enhancing transparency and fairness in the Malaysian used car market, while also serving as a valuable reference for similar initiatives in other countries and markets.

Keywords—Price Prediction, Used Car Market, Data Mining, Machine Learning, Neural Network

REFERENCES

- [1] Statista. (2022, January 26). Malaysia: passenger vehicle registrations 2022. <https://www.statista.com/statistics/672219/malaysia-passenger-vehicle-registrations/>
- [2] Trading Economics. (2023, March). Malaysia New Vehicles Registered - February 2023 Data - 1988-2022 Historical. <https://tradingeconomics.com/malaysia/car-registrations>
- [3] “Automotive industry in Malaysia,” Statista, <https://www.statista.com/topics/5040/automotive-industry-in-malaysia/>
- [4] Grunsven, L. van G. (2020) Urban development in Malaysia: Towards a new systems paradigm - think city. Available at: <https://thinkcity.com.my/wp/wp-content/uploads/2020/04/Issue-2.pdf>
- [5] “Malaysia used car market size & share analysis - industry research report - growth trends,” Malaysia Used Car Market Size & Share Analysis - Industry Research Report - Growth Trends, <https://www.mordorintelligence.com/industry-reports/malaysia-used-car-market>
- [6] E. Mahalingam, “Chip shortage continue to hamper vehicle sales in Malaysia,” CarSifu, <https://www.carsifu.my/news/chip-shortage-continue-to-hamper-vehicle-sales-in-malaysia>
- [7] BusinessToday, By, and BusinessToday, “The Renaissance of the user car industry in the post-pandemic,” BusinessToday, <https://www.businesstoday.com.my/2021/10/30/the-renaissance-of-the-user-car-industry-in-the-post-pandemic-a-fresh-perspective-from-mytukar/>
- [8] Mohamed, S. S., Koo, V. C., Jusoh, A., & Ho, C. S. (2016). Automotive Consumerism: A Study of Car User's Practices & Behaviour in Klang Valley, Malaysia. ResearchGate. https://www.researchgate.net/publication/301886100_Automotive_Consumerism_A_Study_of_Car_User's_Practices_Behaviour_in_Klang_Valley_Malaysia
- [9] L. Cabral and L. Xu, “Seller reputation and price gouging: Evidence from the COVID-19 pandemic,” Economic inquiry, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8251445/>

- [10] D. Grewal, D. M. Hardesty, and G. R. Iyer, "The effects of buyer identification and purchase timing on consumers' perceptions of trust, Price Fairness, and repurchase intentions," *Journal of Interactive Marketing*, vol. 18, no. 4, pp. 87–100, 2004. doi:10.1002/dir.20024
- [11] H. Zhao, X. Yao, Z. Liu, and Q. Yang, "Impact of pricing and product information on consumer buying behavior with customer satisfaction in a mediating role," *Frontiers*, <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.720151/full>
- [12] AirDNA. (n.d.). AirDNA Data - How It Works. <https://www.airdna.co/airdna-data-how-it-works>
- [13] Astrato. (n.d.). Astrato. <https://astrato.io/>
- [14] Astrato. (2021, July 7). Building a Data Dashboard with Astrato: Best Practices. <https://astrato.io/blog/building-a-data-dashboard-with-astrato-best-practices/>
- [15] Astrato. (n.d.). What is a Data App? <https://astrato.io/blog/what-is-a-data-app/>
- [16] Astrato. (n.d.). Data Sources: Getting Started in Astrato. <https://help.astrato.io/en/articles/5214403-data-sources-getting-started-in-astr>
- [17] CarBase.my. [Online]. Available: <https://www.carbase.my/>

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	Study week no.: 4
Student Name & ID: Scott Lai Yong Luo & 19ACB01919	
Supervisor: Dr. Abdulkarim Kanaan Jebna	
Project Title: Second-Hand Car Price Monitor System	

1. WORK DONE

- Build the neural network model

2. WORK TO BE DONE

- Deploy the Shiny App

3. PROBLEMS ENCOUNTERED

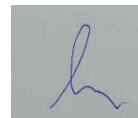
- Time to learn neural networks
- Parameter adjustment and calculation time

4. SELF EVALUATION OF THE PROGRESS

- I feel that the current progress is a bit delayed and need to work harder



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	Study week no.: 6
Student Name & ID: Scott Lai Yong Luo & 19ACB01919	
Supervisor: Dr. Abdulkarim Kanaan Jebna	
Project Title: Second-Hand Car Price Monitor System	

1. WORK DONE

- Build an auto web scraping script
- Optimize the previous scraping code
- Deploy the simple shiny app
- Visualize the feature contribution plot

2. WORK TO BE DONE

- Complete the predict function

3. PROBLEMS ENCOUNTERED

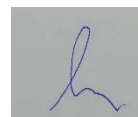
- Difficulty of learning a new language
- Try to understand the model server logic that combines UI and machine learning
- The predict function encounters bugs and cannot run.

4. SELF EVALUATION OF THE PROGRESS

- I feel good about the current progress, but it still need to improve.



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	Study week no.: 8
Student Name & ID: Scott Lai Yong Luo & 19ACB01919	
Supervisor: Dr. Abdulkarim Kanaan Jebna	
Project Title: Second-Hand Car Price Monitor System	

1. WORK DONE

- Using shinydashboard library to makeup the UI
- Update the user selectInput button with more feature
- Create the Treemap in clustering page

2. WORK TO BE DONE

- Complete the predict funtion

3. PROBLEMS ENCOUNTERED

- Encountered problems when creating Multi-level treemap
- Still encountered problems to run the model behind the server logic and take user input as test data

4. SELF EVALUATION OF THE PROGRESS

- Feeling frustrated and not making much progress



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	Study week no.: 10
Student Name & ID: Scott Lai Yong Luo & 19ACB01919	
Supervisor: Dr. Abdulkarim Kanaan Jebna	
Project Title: Second-Hand Car Price Monitor System	

1. WORK DONE

- Complete the predict function
- Create dynamic statistic based on user input
- Added display of the predict price range and mae value
- Add the file input function in clustering page

2. WORK TO BE DONE

- Reduce the mae value
- Find out the suitable user input as input features

3. PROBLEMS ENCOUNTERED

- There is a problem running the neural network. Temporarily use the random forest model instead

4. SELF EVALUATION OF THE PROGRESS

- I feel happy with the progress so far, but there still need to improve



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	Study week no.: 12
Student Name & ID: Scott Lai Yong Luo & 19ACB01919	
Supervisor: Dr. Abdulkarim Kanaan Jebna	
Project Title: Second-Hand Car Price Monitor System	

1. WORK DONE

- Functional testing of shinyapp

2. WORK TO BE DONE

- Complete the report
- Try deploy out the dashboard

3. PROBLEMS ENCOUNTERED

- None

4. SELF EVALUATION OF THE PROGRESS

- I feel that the current progress is a bit insufficient and need to speed up the progress.



Supervisor's signature



Student's signature

POSTER

**SECOND-HAND CAR
PRICE MONITOR
SYSTEM**

**Drive the Car of Your Dreams without Breaking
the Bank !!!**

- **Real-time data from various sources**
- **Accurate and up-to-date market value estimates**
- **User-friendly interface for easy searches**
- **Customizable search criteria to fit your needs**

Looking to buy a used car
but not sure if you're getting
a fair price ?

Get the Best
Deals on Used
Cars with Real-
Time Data

PLAGIARISM CHECK RESULT

Second Hand Car Price Monitor System

ORIGINALITY REPORT

6%

SIMILARITY INDEX

5%

INTERNET SOURCES

2%

PUBLICATIONS

4%

STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

1%

★ Submitted to Asia Pacific University College of
Technology and Innovation (UCTI)

Student Paper

Exclude quotes Off

Exclude matches < 8 words

Exclude bibliography Off

PLAGIARISM CHECK RESULT

Universiti Tunku Abdul Rahman			
Form Title: Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)			
Form Number: FM-IAD-005	Rev No.: 0	Effective	Date: Page No.: 1 of 1



FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Full Name(s) of Candidate(s)	Scott Lai Yong Luo
ID Number(s)	19ACB01919
Programme / Course	IB
Title of Final Year Project	Second-Hand Car Price Monitor System

Similarity	Supervisor's Comments (Compulsory if parameters of originality exceed the limits approved by UTAR)
Overall similarity index: <u>6</u> % Similarity by source Internet Sources: <u>5</u> % Publications: <u>2</u> % Student Papers: <u>4</u> %	
Number of individual sources listed of more than 3% similarity: <u>0</u>	
The parameters of originality required, and limits approved by UTAR are as Follows: (i) Overall similarity index is 20% and below, and (ii) Matching of individual sources listed must be less than 3% each, and (iii) Matching texts in the continuous block must not exceed 8 words <i>Note: Parameters (i) – (ii) shall exclude quotes, bibliography, and text matches which are less</i>	

Note: Supervisor/Candidate(s) is/are required to provide a softcopy of the full set of the originality report to the Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.

Signature of Supervisor

Name: Dr. Abdulkarim M. Jamal Kanaan

Date: 15/09/2023

Signature of Co-Supervisor

Name: _____

Date: _____

PLAGIARISM CHECK RESULT



UNIVERSITI TUNKU ABDUL RAHMAN

FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY

(KAMPAR CAMPUS)

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	19ACB01919
Student Name	Scott Lai Yong Luo
Supervisor Name	Dr. Abdulkarim Kanaan Jebna

TICK (✓)	DOCUMENT ITEMS
	Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
✓	Title Page
✓	Signed Report Status Declaration Form
✓	Signed FYP Thesis Submission Form
✓	Signed form of the Declaration of Originality
✓	Acknowledgement
✓	Abstract
✓	Table of Contents
✓	List of Figures (if applicable)
✓	List of Tables (if applicable)
	List of Symbols (if applicable)
✓	List of Abbreviations (if applicable)
✓	Chapters / Content
✓	Bibliography (or References)
✓	All references in bibliography are cited in the thesis, especially in the chapter of literature review
	Appendices (if applicable)
✓	Weekly Log
✓	Poster
✓	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)
✓	I agree 5 marks will be deducted due to incorrect format, declare wrongly the ticked of these items, and/or any dispute happening for these items in this report.

*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report.

(Signature of Student)

Date: 12/9/2023