SQIT3063 DATA MINING (GROUP C)

SECOND SEMESTER SESSION 2022/2023 (A222)

REPORT ASSESSMENT:

REPORT OF AUTOMOBILE PRICE PREDICTION USING DATA MINING TECHNIQUES

PREPARED FOR:

DR. CH'NG CHEE KEONG

PREPARED BY: GROUP 2

No.	Name	Matric Number
1	KAREN KUAN JIA YING **(Leader)**	285565
2	TEH YI QING	285572
3	KH'NG WEI XIN	285602
4	LAI MEI KUN	287675
5	LOW YIN FEI	285459

CHAPTER ONE

INTRODUCTION

1.0 Introduction

In the following phase, we will introduce the study's history, problem statement, importance, and scope. Our issue is car prices. First, a broad overview. To conclude, almost every household has a scooter due to vehicle popularity, but do you know the automobile price composition? Are car quotes and guide pricing different? In the Excel sheet, odometer mileage, gasoline type, horsepower, metallic colour, and so on affect car costs.

1.1 Background of the study

The global automobile sales have been rising steadily around the world, however precise numbers may change from year to year. Global automobile sales were typically rising before the COVID-19 epidemic. The pandemic, however, had a substantial effect on the automobile sector, which saw a fall in sales as a result of lockdowns, financial instability, and interrupted supply networks. Numerous elements, including the state of the economy, consumer confidence, governmental initiatives, and market trends, have an impact on the revival of the auto sector. Car sales are anticipated to gradually increase as the global economy continues to stabilise and recover from the epidemic, while unique regional differences may arise.

One of the major car production and sales centres in Southeast Asia, Malaysia boasts a substantial automotive market. Both domestically produced cars and cars imported from other countries make up Malaysia's automotive sector. In the past, Malaysian vehicle sales have increased due to factors including population expansion, rising income, and accelerating urbanisation. However, due to lockdown measures and economic uncertainty, the COVID-19 epidemic drastically decreased automobile sales in Malaysia, following the global pattern.

Numerous variables, such as ease, adaptability, perceived comfort, and the availability of reasonable auto finance choices, might be linked to the desire for private vehicles. An increased reliance on private automobiles may result from problems with public transit systems in some places, like insufficient coverage, congestion during peak hours, or

scheduling issues. It's important to note, though, that Malaysia's public transit has improved recently. Public transport networks, including urban rail systems like the Mass Rapid Transit (MRT) and the Light Rail Transit (LRT), as well as bus services, have been expanded and improved thanks to investments made by the government. By addressing the requirements of the expanding population and reducing traffic on the roadways, these efforts seek to offer more effective and dependable public transit choices.

Many university students and employees drive, especially when the campus is situated in an area with few or difficult access points for public transit. Owning a car allows you greater freedom and convenience when travelling to and from school as well as for running other personal errands. Universities frequently have parking policies and procedures for applying for stickers to control the parking of these cars on campus. These procedures often entail students or employees requesting a parking ticket or permit from the university's parking office or other appropriate authorities.

To guarantee orderly and effective use of parking spots on campus, parking stickers and permits are used. Universities can control the number of parking spots available and deter unauthorised parking by designating certain parking sections or zones to those with valid permits. Students or employees may be required to give information for the sticker application procedure, such as car registration information and identification. It's crucial to check the university's parking department or website for correct and up-to-date information on sticker application criteria and processes as particular procedures and rules for parking stickers might different between colleges.

1.2 Problem Statement

In this section, we will discuss the problem statement and issues from several perspectives, as these issues will have a significant impact on the automobile industry. As a consequence, we will discuss how COVID-19 affects the automobile business, where it shows declines in vehicle export rates, technological advancement, and how government subsidies influence the automobile sector by increasing sales of electric vehicles (EV) in their countries, as well as how people react to that policy.

COVID has become a popular topic among the public in 2019, and it affects a wide range of sectors. One of the industries that has a considerable impact on it is the automobile sector. People are suffering from state-wide lockdowns in some countries, such as China, United States, Malaysia, and others; as a result, the inclination of people to acquire an automobile is lower, as most businesses require them to work from home to comply with the country's policy. Not to mention their goal to get a car for trip purposes, they are obliged to stay at home due to the state-wide lockdown implementation. In terms of automotive, COVID-19 reduced the South African manufacturing industry's output by 49.4 percent, with motor vehicles, transport equipment, vehicle parts, and accessories accounting for 7.8 percent of total output (Liedtke, 2020). A drop in car production might have a substantial influence on the total quantity of automobiles exported from South Africa to worldwide markets.

Next, with the advancement of technology, the automobile industry has become more advanced. autonomous vehicles, which can drive themselves and perform critical activities without the need for human interaction. A lot of technology has been implemented for autonomous vehicles to ensure individual safety, such as vision-guided systems like LIDAR, radar, GPS, and computer vision because they wish to recognize objects and paths without conflict. However, there are still many difficulties for a customer to consider when purchasing an autonomous vehicle, including dependability and safety, testing and validity, legal issues, orientation, data privacy, and technical barriers. (RGS Deemantha, B Hettige, 2023).

Besides that, government incentives are also one of the issues in response to environmental concerns. Many countries have implemented incentive policies to stimulate the electric vehicles (EV) industry and consumer uptake (Zhu et al., 2019; Ye et al., 2021). The German government, for example, offers a subsidy of up to 6,000 EUR to buyers of EVs costing less than 60,000 EUR. Manufacturers are given R&D subsidies to encourage them to invest more in EV technical advancement. For instance, in 2019, Innovate UK provided 20 million pounds of R&D funding to advance low-carbon automobile propulsion technology.

Although other nations have established subsidy policies to stimulate EV sales, this strategy does not function effectively in Malaysia, and the result of EV sales is not promising. This is because, among prospective EV users, environmental concerns are the lowest priority in purchasing EVs, and other factors such as purchase and maintenance costs, regulatory

policies, vehicle service points, and road infrastructure have become matters of concern in diverting Malaysians' car-purchasing behaviors from ICEs to EVs (Muzir et al., 2022).

In short, the aforementioned issue should be addressed in the automobile sector. These challenges are critical for policymakers and manufacturers because officials design policies to make people's lives happier and more harmonious, whilst manufacturers produce goods to meet the wants of prospective purchasers. We feel that a corporation formed by two parties will undoubtedly benefit the rest of the world's people.

1.3 Research Objectives

1. To determine the factors that affect the prices of cars.

2. To develop various predictive models to predict the car prices.

3. To assess the models' performance by comparing their average squared error.

4. To select the best model to estimate the prices of cars using the provided dataset.

1.4 Significance of Study

The importance of this study is that it will provide realistic car pricing estimates. The created prediction model will assist a wide range of stakeholders, including regulators, automotive manufacturers, and buyers.

Policymakers can use this model to forecast the price of an imported car in order to avoid overpricing by the importer. They may use this model to determine the price of a car by simply putting all of the needed characteristics into the model, and the model will reveal the price range of that automobile. Customers will gain indirectly from this model since they will be able to purchase important automotive at the lowest possible price. Furthermore, from the standpoint of policymakers, this model can be used to determine an appropriate rate for taxes on imported cars, which can then be used to determine the suitable pricing for the imported car.

Automobile makers can also use this model to anticipate the price of their automobile, knowing the functionality of the car's components that effect the most of the car that is selling at market price. Then they can add the components so that the car can sell for the highest possible price. It will establish a virtuous cycle in the automobile sector, allowing Americans to buy automobiles at their income level. In summary, it will benefit the United States of America's economy by increasing sales in the vehicle sector.

Customers will profit from this study as well. This is because customers are the ones who purchase the car. With our understanding of the vehicle sector, we discovered that most Americans rely heavily on cars as their primary source of mobility, as public transportation is not always reliable. As a result of the model that will be built, clients will be able to purchase a car based on their income level.

Finally, the model that will be built will benefit individuals, whether they are country politicians, local manufacturers, or customers. It will undoubtedly benefit the US economy if the prices of all autos remain low enough that everyone can afford at least one car.

1.5 Scope of Study

The study focuses on the use of data mining technology in automotive price projections based on past trend analysis, with a dataset of 1367 pieces of collected data from Canada, United States of America. It includes the predictions that we are going to make which are their corresponding car pricing that can be our dependent variables and a variety of car-related qualities as well as the automobile year of manufacture, the accumulated Kilometers on odometer, the horsepower of the automobile and more.

1.6 Conclusion

This section covers the issue statement, COVID-19, technology, and government incentives for electric automobiles. COVID-19 has reduced vehicle exports and consumer automobile purchases. Self-driving vehicles have raised worries about data privacy, safety, and reliability. The government is also subsidising electric car research and development to stimulate customer adoption. Due to consumer concerns about the environment, pricing,

rules, infrastructure, and upkeep, such projects may succeed in different countries. We seek to achieve the research objectives outlined above while also benefiting three parties: policymakers, local manufacturers, and customers. In this study, we will create our models using data mining approaches.

CHAPTER TWO

LITERATURE REVIEW

2.0 Introduction

This chapter begins with a brief on automobile costs and covers a variety of topics, including the variables that affect automobile prices, the price ranges across various vehicle types and categories, as well as the trends and patterns seen in the automotive industry. It reviews several previous studies on literature review and data mining techniques, application of data mining in various areas.

2.1 Literature Review

Numerous studies and research papers have been conducted to predict used car prices worldwide, employing various methodologies and approaches. The accuracy of these predictions has ranged from 50% to 90%.

Sanap et al. (2022) has come out with a solution that can get an appropriate estimation of car prices using Machine Learning Techniques. These techniques can help to save cost of money and time. To predict the car prices, they choose one of the machine learning techniques that is the most suitable for prediction which is linear regression to complete this prediction.

Gegic et al. (2019) gathered data from a Bosnian website for used cars, resulting in 797 car samples after pre-processing. They used three different machine learning techniques, including Support Vector Machine, Random Forest, and Artificial Neural Network. The best results were obtained by combining these algorithms with pre classification of prices using Random Forest, achieving accuracies of up to 87.38%.

In a conference, Rupesh Gupta et al. (2022) have published a paper that is related to automobile price prediction using regression models. Even though they have substantial knowledge on this topic, they think predicting the prices of automobiles will become a popular research topic. They have used five regression models to complete this prediction and

they found out random forest regression has the best performance with values 0.93 for R Squared Score, 1390.9 as Mean Absolute Error and 2139.7 as Root Mean Square Error.

In a study by K. Samruddhi and Kumar (2020), a supervised machine learning model using K-Nearest Neighbors was proposed to predict used car prices. The dataset was obtained from Kaggle and included 14 different attributes. The accuracy reached up to 85% by experimenting with different values of K and changing the percentage of training data to testing data. The model was also cross-validated using 5 and 10 folds with the K-fold method.

Based on the research article by Mehmet Bilen with the title " Predicting Used Car Prices with Heuristic Algorithms and Creating a New Dataset" that published in year 2021, he stated that The Fisher+ANN model got the best performance with MAE 0.01050, MSE 0.000281 error, and R2 0.8958 performance value as a consequence of the prediction procedures employing multiple preprocessing steps and different prediction methods. ANN outperforms other algorithms in all scenarios where alternative preprocessing approaches are applied.

Nikhade and Borde (2023) has published an article that related with car price prediction using machine learning. The main purpose of the project is to predict the used car prices, compare it and estimate the lifespan of certain car. They think the most important factor that will affect the price of a used car is the number of kilometers. They get their dataset from Kaggle and they have used a variety of methods to predict the outcomes which include multiple linear regression, decision trees and K-Nearest Neighbor.

A study of car price prediction using machine learning algorithms are made by B V Raghurami Reddy and Dr. K. Santhi Sree (2022). In this article, they have to find the price of a used car. Other than that, they have to find out which features or variables are significant in predicting the car prices and the weight of each variables in predicting the automobile prices. They have used machine learning algorithms such as linear regression, ridge regression, lasso regression, K-Nearest Neighbors (KNN) regressor, random forest regressor and bagging regression and more to complete this car price prediction.

In 2021, an article that makes research on Claim Amount Forecasting and Pricing of Automobile Insurance based on BP Neural Network will be published by Wenguang Yu et al.. They think there are few researchers who apply BP Neural Network in predicting automobile prices. Through the work, they found out the accuracy of this prediction model to both the data of Shandong Province and to the data of six cities are more than 95%.

Moreover, a study from Aman Kharwal (2021) has made a car price prediction with machine learning. The car prices will vary based on the features that they provided for the car model. As a example, the car brand, model, horsepower are the features that will affect the prices of automobile. In this study, he used machine learning which is Python programming language to complete this prediction of car prices.

According to Sobana Selvaratnam and Nagulan Ratnarajah in the article " Feature selection in automobile price prediction: an integrated approach" that was published in 2021, they have applied the LASSO and stepwise selection regression methods in an integrated way. The findings show that combining embedded and wrapper feature selection to build a hybrid form of feature selection yields better outcomes.

In conclusion, predicting used car prices is an important and actively researched topic. Various models and techniques have been explored, with the best achieved accuracy reported at 83.63% using the Random Forest technique on Kaggle's dataset. Multiple regressors have been tested, and the final model tends to be a linear regression model.

2.2 Data Mining Techniques

i. Decision Tree

Decision tree is a non-parametric supervised learning approach that used for classification and regression applications. Decision tree is organised hierarchically and it contain root node, branches, internal nodes, and leaf nodes. The root node of a decision tree,

which has any incoming branches, is shown below. The internal nodes which also known as decision nodes, are fed by the root node's outgoing branches. Both of the node types undertake evaluations based on the available attributes to create consistent subsets, which are represented by leaf nodes or terminal nodes. All of the outcomes within the dataset are represented by the leaf nodes. (IBM, 2021)

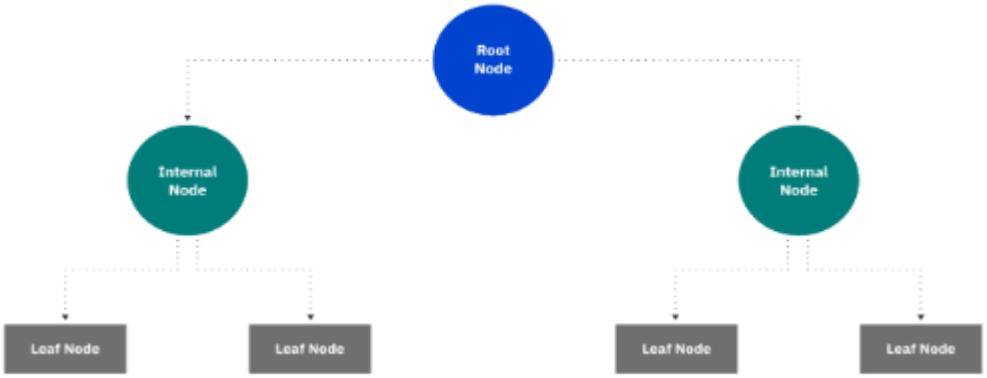


Figure 1 : Decision Tree

ii. Regression

Regression model can explain the relationship between one or more independent variables and provide a response or dependent. Regression analysis comes in many forms, including logistic regression, stepwise regression modelling, multiple, non-linear, and linear. A model that uses a linear regression has an output-input relationship that is a straight line. Even when a relationship isn't particularly linear, our brains still attempt to recognize the

pattern and associate that relationship with a crude linear model. Multiple regression is means the regression model that have more than one input variable that may have an impact on the final result which also known as target variable. Non-linear regression involves fitting data to a model before expressing the result as a mathematical function. (Kenton, 2022) Stepwise regression is more of a method than the other topics we have covered up to this point, which are specific kinds of models. The analyst may begin building a model with the input variable that is most directly correlated if the model has numerous potential inputs. The next step after completing that is to improve the model's accuracy. (*What is a regression model?* 2021)

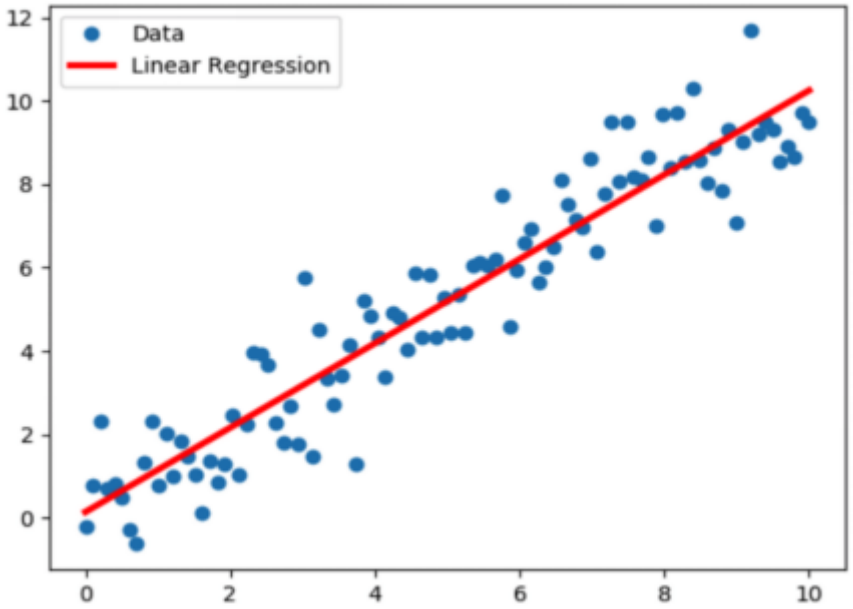


Figure 2 : Linear Regression

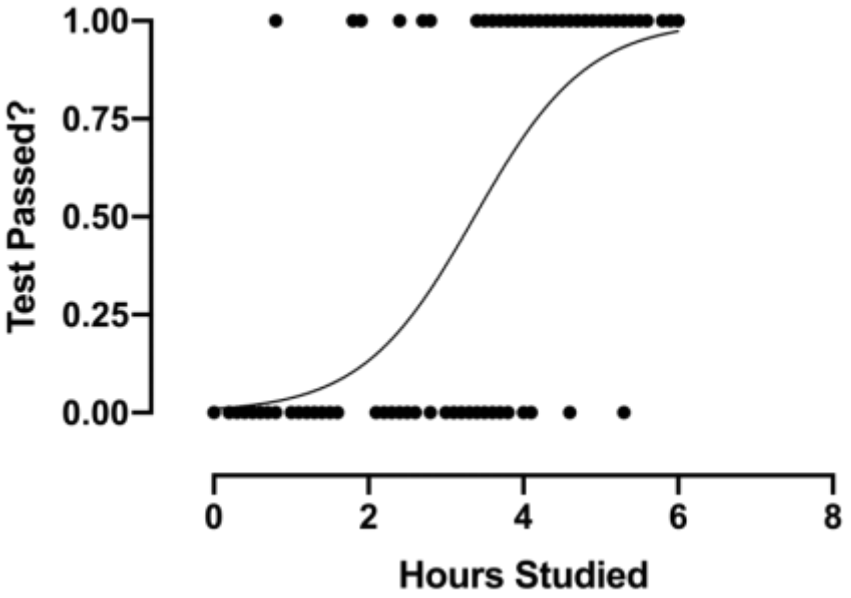


Figure 3 : Logistic Regression

Logistic regression can be used when the dependent variable is discrete. This method mostly can used to calculate the situation when there only have two choice such as pass or fail, true or false and others. So, since the target variable has a range of 0 to 1, the probability will only has a value between 0 or 1. A sigmoid curve depicts its relationship to the independent variable. (Sharma, 2022)

iii. Neural Network

In a neural network, neurons serve as the basic building blocks and are typically arranged in layers, as seen in the following diagram. Neural network is a representation of how human brain function. It simulates a large number of connected processing units that resemble abstract representations of neurons in order to function. It normally have three layers which is the input layer that contains units that represent the input fields, one or more hidden layers, and the output layer, which contains a unit or units that represent the target fields. The weight between the connection are vary. The first layer will receives input data, and each neuron in the subsequent layer receives values propagated from the previous layer's neuron and then the output layer will delivers a result. (*The Neural Networks Model* 2021)

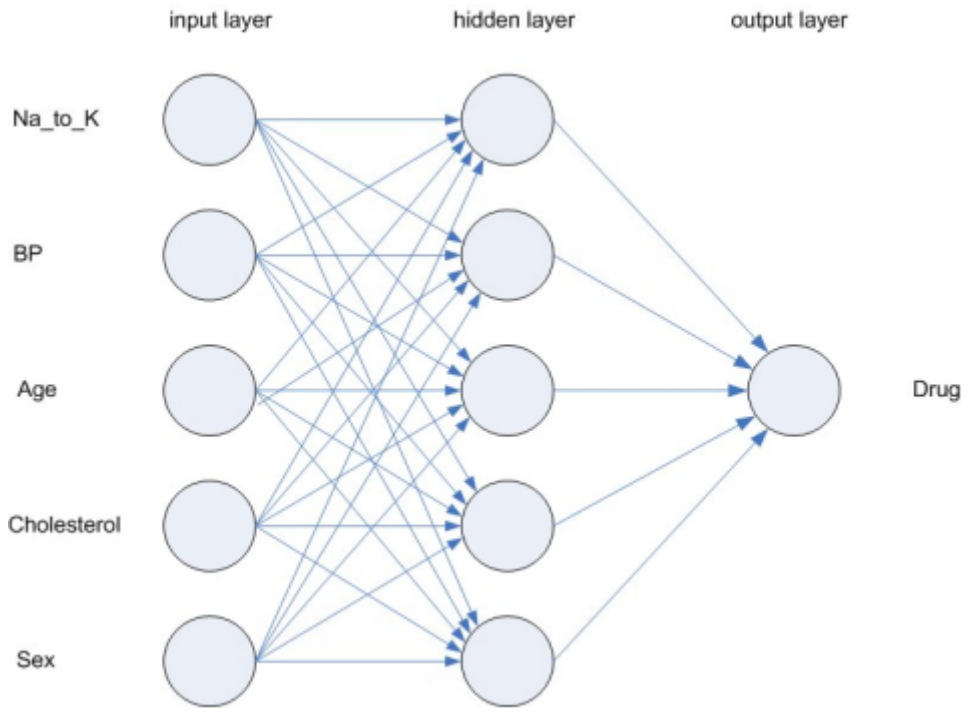


Figure 4 : Neural Network

iv. Random Forest

Random forest combines the output of various decision trees to produce single outcome. It is use by many people by its adaptability and usability because it can solve classification and regression issues. Decision trees and random forests differ primarily in that decision trees take into account all possible feature splits while random forests only choose a subset of those features. (*What is Random Forest?*)

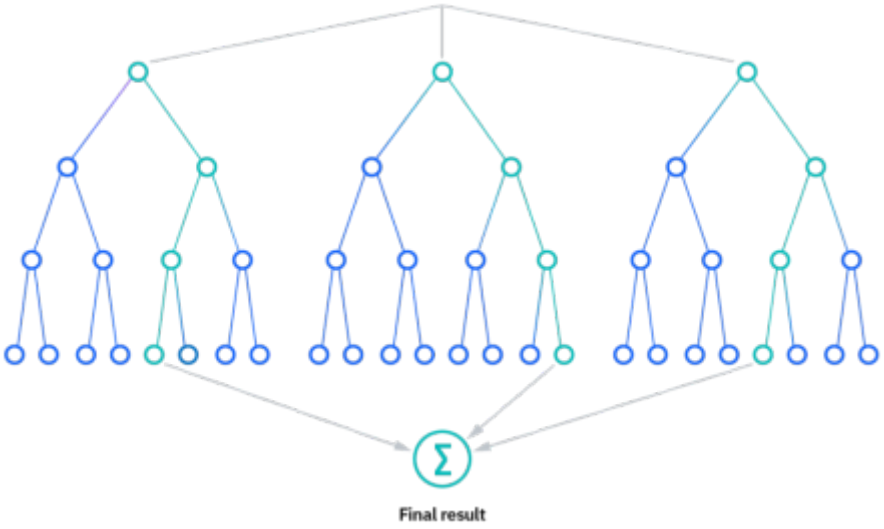


Figure 5 : Random Forest

2.3 Application of Data Mining in Various Area

Today's information era allows practically every department, industry, sector, or business to benefit from data mining. Numerous applications of data mining exist, including fraud detection, retail services, traffic analysis, sports analytics, and pattern analysis of criminal activity.

First of all, data mining is widely used in fraud detection across various industries, such as finance, insurance, and e-commerce. By analysing large volumes of transactional data, data mining algorithms can identify patterns and anomalies that indicate fraudulent activities. These patterns may include unusual purchasing patterns, abnormal transaction amounts, or suspicious network connections. Data mining helps in detecting and preventing fraudulent activities, thereby saving organisations from financial losses and protecting their customers.

Moreover, data mining plays a crucial role in the retail industry by analysing customer data and market trends. Retailers can utilise data mining techniques to gain insights into customer behaviour, preferences, and purchasing patterns. By analysing historical sales data, customer demographics, and other relevant information, retailers can segment customers, personalise marketing campaigns, optimise inventory management, and make data-driven decisions. This enables them to improve customer satisfaction, increase sales, and enhance overall operational efficiency.

Other than that, data mining is also employed in traffic analysis to study and improve transportation systems. By analysing data from various sources such as traffic cameras, sensors, and GPS devices, data mining algorithms can identify traffic patterns, predict congestion, and optimize traffic flow. This information can help urban planners, transportation authorities, and logistics companies make informed decisions regarding route planning, traffic signal optimization, public transportation management, and infrastructure development, leading to reduced congestion and improved transportation efficiency.

Data mining techniques are widely used in sports analytics to extract meaningful insights from vast amounts of sports-related data. By analysing player performance statistics, game footage, and other relevant data, teams and coaches can gain valuable insights into player strategies, opponent weaknesses, and game dynamics. Data mining can also be used to predict player performance, optimise team composition, and develop winning strategies. Additionally, sports organisations can use data mining to understand fan behaviour, personalise marketing efforts, and enhance the overall fan experience.

Lastly, data mining is employed in law enforcement and criminal justice systems to analyse patterns of criminal activity. By examining large volumes of historical crime data, data mining algorithms can identify patterns, trends, and associations that can help in predicting and preventing criminal behaviour. These techniques can be used for various purposes, such as identifying crime hotspots, detecting organised crime networks, and assisting in investigative efforts. Data mining aids law enforcement agencies in allocating resources effectively and proactively combating crime.

2.4 Conclusion

In a nutshell, the problem of accurately estimating the values of used cars is one that receives a significant amount of attention and study, and several models and approaches are investigated. The processes of data mining have a broad variety of applications and may give insightful knowledge in a variety of domains, including the identification of fraudulent activity, retail, traffic analysis, sports analytics, and law enforcement.

CHAPTER THREE

METHODOLOGY

3.0 Introduction

This chapter begins with the prior to evaluating country-specific data, it might be beneficial to predict vehicle costs using global data, which can assist a variety of stakeholders. This chapter also discusses the data collection, data analysis, research design, research process, flow chart, and expected findings. A methodical strategy for analyzing and modeling the many different aspects that have an impact on the cost of automobiles is at the core of the process for estimating future costs for automobiles. Big datasets with automotive attributes and prices help data mining and predictive modeling estimate car costs. This introduction explains the method for anticipating car prices.

3.1 Data Collection

Data Collection is a systematic process of gathering observations or measurements. (Pritha, 2022).

We used secondary data as our dataset for this study. This dataset depicts the chief marketing officer (CMO) of an automobile agency who was reviewing a list of car model attributes that he had received from the production facility. He had to settle on the base prices before delivering the manufacturer's proposed retail prices to dealers the following week. The CMO asked a data scientist at the research lab to forecast costs based on data from previous automobile models. Each car model had unique features that might influence the pricing. The data scientist made the decision to employ feed-forward neural networks to forecast the prices of new models. He wanted to know which prediction model was best for automobile manufacturing plants after comparing numerous methods.

It is published by the Ivey Publishing, with the title “Predicting Automobile Prices Using Neural Networks”. This dataset is prepared by Rasha Kashef and Boya Zhang. They are using the provided dataset to develop a predictive model to estimate the prices of cars.

Link below is the source for this data:

https://www.iveypublishing.ca/s/product/predicting-automobile-prices-using-neural-networks/01t5c00000CwpP3AAJ

In total, there are 1367 pieces of data. There is 1 dependent variable which is based on the excel file provided, the dependent variable has 1 which is price (the offer price in dollar) and the independent variable has 27 (Refer to Table 1).

Variables	Attributes
Price	Offer price in $
Age	Age in months
KM	Accumulated Kilometres on odometer
Ful	Fuel Type (Petrol, Diesel, CNG)
HP	Horsepower
MC	Metallic Color? (Yes=1, No=0)
clr	Color (Blue, Red, Grey, Silver, Black, etc.)
Auto	Automatic (Yes =1, No = 0)
CC	Cylinder Volume in cubic centimeters
Drs	Number of doors
Cyl	Number of cylinders
Grs	Number of gear positions
Wght	Weight in Kilograms
G_P	Guarantee period in months
Mfr_G	Within Manufacturer's Guarantee period (Yes=1, No= 0)
ABS	Anti-Lock Brake System (Yes=1, No = 0)
Abag_1	Driver Airbag (Yes=1, No= 0)
Abag_2	Passenger Airbag (Yes=1, No=0)
AC	Automatic Air Conditioning (Yes = 1, No =0)
Comp	Boardcomputer (Yes=1, No = 0)
CD	CD Player (Yes =1, No = 0)
CLock	Central Lock (Yes =1, No = 0)
Pwin	Powered Windows (Yes =1, No =0)
PStr	Power Steering (Yes =1, No = 0)
Radio	Radio (Yes =1, No =0)
SpM	Sport Model (Yes=1, No = 0)
M_Rim	Metallic Rim (Yes =1, No = 0)
Tow_Bar	Tow Bar (Yes = 1, No =0)

Table 1: each variable and its attributes

For our research, we only used a few independent variables which are age (age in months), G_P(guarantee period in months), KM (Accumulated Kilometres on odometer), HP (Horsepower), CC (Cylinder Volume in cubic centimeters), Wght (Weight in Kilograms) and Auto (Automatic (Yes =1, No = 0)) and our dependent variables is Price (offer price in $).

3.2 Data analysis (including statistical tool)

In statistical analysis, a measurement scale is the type of information offered by numbers. (Jo Ann,2022) There are four measuring scales; however, we only used three of them in our research: nominal scales, interval scales, and ratio scales.

Variables	Attributes	Measurement scales
Price	Offer price in $	Interval Scales
Age	Age in months	Interval Scales
KM	Accumulated Kilometres on odometer	Ratio Scales
Ful	Fuel Type (Petrol, Diesel, CNG)	Nominal Scales
HP	Horsepower	Ratio Scales
MC	Metallic Color? (Yes=1, No=0)	Nominal Scales
clr	Color (Blue, Red, Grey, Silver, Black, etc.)	Nominal Scales
Auto	Automatic (Yes =1, No = 0)	Nominal Scales
CC	Cylinder Volume in cubic centimeters	Ratio Scales
Drs	Number of doors	Nominal Scales
Cyl	Number of cylinders	Nominal Scales
Grs	Number of gear positions	Nominal Scales
Wght	Weight in Kilograms	Ratio Scales
G_P	Guarantee period in months	Interval Scales
Mfr_G	Within Manufacturer's Guarantee period (Yes=1, No= 0)	Nominal Scales
ABS	Anti-Lock Brake System (Yes=1, No = 0)	Nominal Scales
Abag_1	Driver Airbag (Yes=1, No= 0)	Nominal Scales
Abag_2	Passenger Airbag (Yes=1, No=0)	Nominal Scales
AC	Automatic Air Conditioning (Yes = 1, No =0)	Nominal Scales
Comp	Boardcomputer (Yes=1, No = 0)	Nominal Scales
CD	CD Player (Yes =1, No = 0)	Nominal Scales
CLock	Central Lock (Yes =1, No = 0)	Nominal Scales
Pwin	Powered Windows (Yes =1, No =0)	Nominal Scales
PStr	Power Steering (Yes =1, No = 0)	Nominal Scales
Radio	Radio (Yes =1, No =0)	Nominal Scales
SpM	Sport Model (Yes=1, No = 0)	Nominal Scales
M_Rim	Metallic Rim (Yes =1, No = 0)	Nominal Scales
Tow_Bar	Tow Bar (Yes = 1, No =0)	Nominal Scales

Table 2: The relationship between each variables and its measurement scales

Table 2 has shown clearly the relationship between each variables and its measurement scales. For better understanding, nominal scales are employed to name or identify individuals, objects, or events. A nominal scale typically deals with non-numeric variables, or numbers with no values. Because it can quantify the difference between values, an interval scale is classified as a quantitative measurement scale. It is possible to compute the mean and median of the variables. Researchers can use ratio scales to compare differences

or intervals. The ratio scale is distinguished by having an absolute zero and no negative values.

In this study, we employed Excel and SAS as statistical tools. The dataset provided is in xlsx format. In Excel, we may rank each variable from smallest to largest so that we can go over each variable column and identify any outliers, missing values, and so on in the dataset manually. SAS was the software we used to examine the data and create the model in Chapters 4 and 5.

In conclusion, there are three measurement scales that we have identified and applied in our research which are nominal scales, interval scale and ratio scale. Each measurement scale is justified with a clear explanation so that it can help us to better understand the concept. We used Excel and SAS as our main statistical tools to analyse our data as it will display a detailed output for each analysis.

3.3 Research design

Research design is important as it shows the methods and techniques, we used to achieve the objectives that we have listed in chapter 1, research objective. Table below shows the relationship between each problem, methods used, and objective achieved.

Problem	Methods Used	Objectives
COVID 19	• Literature review • Data Collection • Data Analysis	1.To determine the factors that affect the prices of cars.
The sales of Electric Vehicles (EV) are not promising in Malaysia	• Decision Tree • Regression • Neural Network	2. To develop various predictive models to predict the car prices.
Technological advancement	• Decision Tree • Regression • Neural Network	3. To assess the models' performance by comparing their average squared error. 4. To select the best model to estimate the prices of cars using the provided dataset.

Table 2: Research Design

3.4 Research Process

Stage 1: Pre-processing

The purpose at this stage is to guarantee that the data in the dataset is clean. We use Excel's "sort and filter" capabilities to sort each variable in each column to see if the values in the dataset are incomplete, outliers, or inconsistent. By observation, we discovered that one of the values in the CC column, 16000, might be an outlier because the other values in the same column in the given dataset range from 1300 to 2000. Also, the KM column ranges from 5000 to 232942 on average; however, there is one outlier number that ranges from 3 KM to 4002 KM, which is substantially lower than the average value.

Then, we use Excel to helps us to do a box plot and using Excel function to helps us to do calculation for us to validates our data.

For box plot, we need to know its Min, Q1, Median, Q3, Max, Mean, IQR, Lower Limit and Upper Limit, below are the functions for each of it:

<u>For KM column:</u>

Min = MIN (A2: A1368) = 3

Q1 = QUARTILE.EXC(A2:A1368,1) = 43002

Median = MEDIAN(A2:A1368) = 64002

Q3 = QUARTILE.EXC(A2:A1368,3) =97276

Max =MAX(A2:A1368) =232942

Mean = AVERAGE(A2:A1368) = 68605.5

IQR=D6-D4 = 44274

Lower Limit =D4-(D9*1.5) = -23409

Upper Limit = D6+(D9*1.5) = 153687

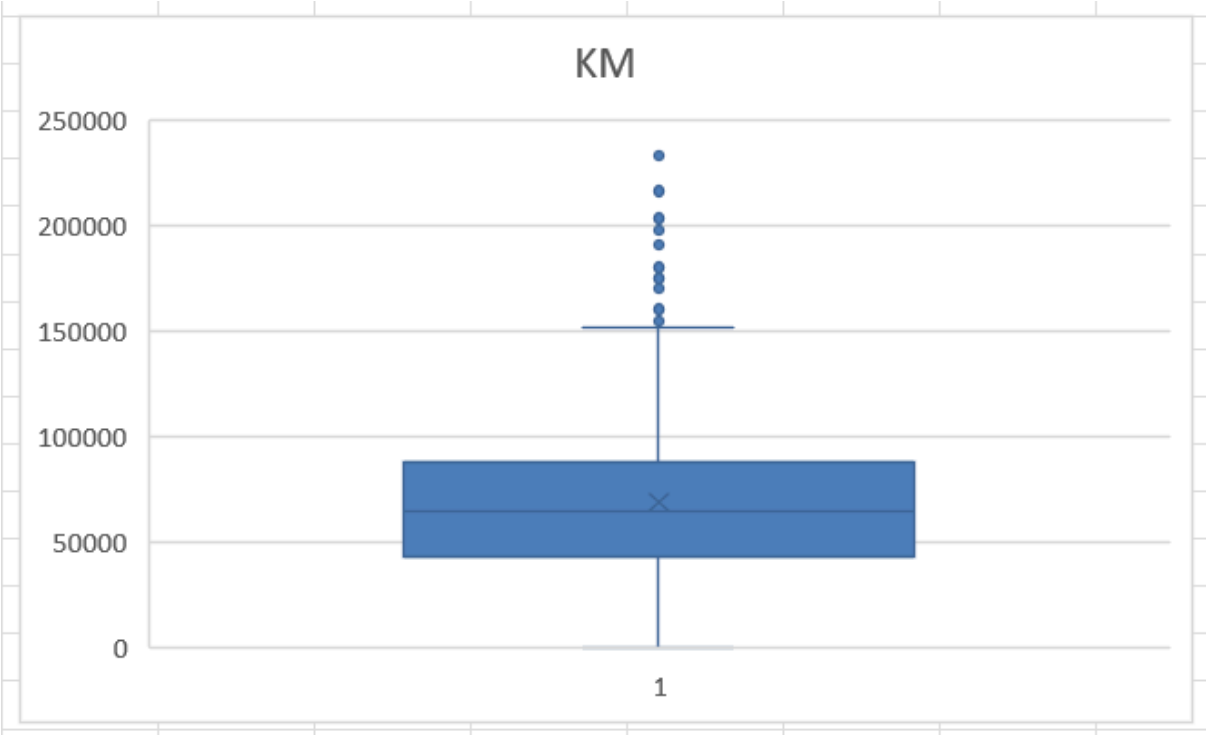


Figure 6 : box plot for KM column

Based on the Figure 1, we can observed that there are some outlier. Also, based on the calculation, we noted that our range for KM column is from -23409(Lower Limit) to 153687 (Upper Limit). Hence, it means that the value that in KM column that exceeds 153687 will considered as outlier, from row 1324 to 1368.

For CC Column:

Min = MIN (L2: L1368) = 1300

Q1 = QUARTILE.EXC(L2:L1368,1) = 1400

Median = MEDIAN(L2:L1368) = 1600

Q3 = QUARTILE.EXC(L2:L1368,3) =1600

Max =MAX(L2:L1368) =16000

Mean = AVERAGE(L2:L1368) = 1574.815

IQR=P22-P20 = 200

Lower Limit =P20-(P25*1.5) = 1100

Upper Limit = P22+(P25*1.5) = 1900

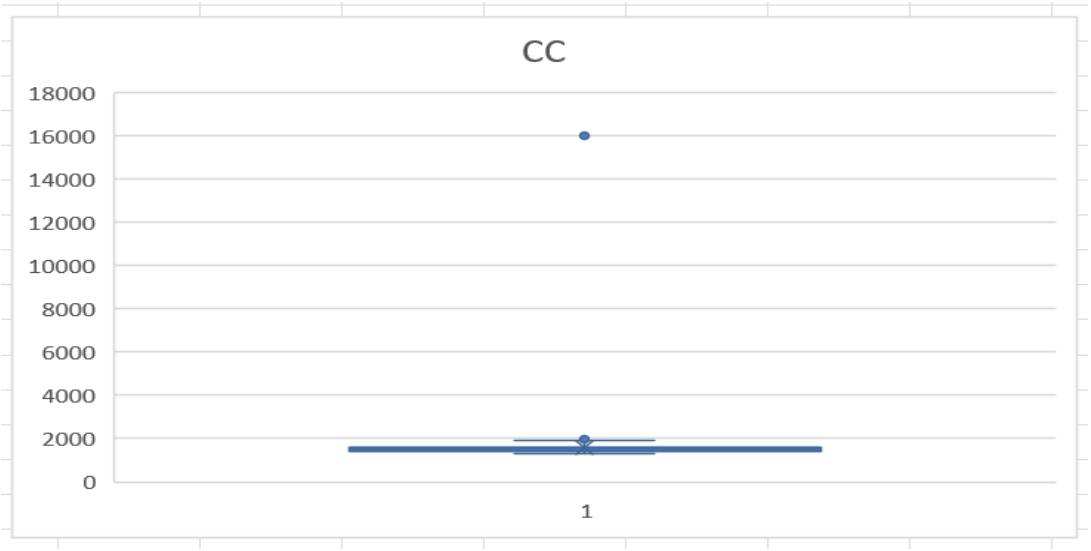


Figure 7 : box plot for CC column

Based on the Figure 2, we can observed that there are some outlier. Also, based on the calculation, we noted that our range for CC column is from 1100(Lower Limit) to 1900(Upper Limit). Hence, it means that the value that in CC column that exceeds 1900 will considered as outlier, from row 1297 to 1368.

Stage 2: Data transformation or normalization

Data transformation is the process of converting data or information from one format to another, usually from the format of a source system into the required format of a new destination system. Basically, it means rescale the data into a suitable range to increase processing speed and reduce memory allocation. There are three methods that can be used which are decimal scaling, min-max normalization and Z-score. In our research, we replace all the outlier value with its median for that column in Excel.

Stage 3: Model construction/ Model development

Based on the 2.2 parts, we can build numerous models for our research, such as decision trees, regression, neural networks, and random forests. SAS is required to create the model construction. For instance, if we need to build a decision tree, we can import the dataset into SAS and specify the goal variable. The target variable for our study is price (offer price in USD). Then, from the toolbar, select "Sample," then locate the "Data Partition" icon, and drag it into the diagram. Then, using the attributes, we must determine the data set allocation, training, and validation, as well as ensure that the training set is greater than the validation set. The data division node must then be linked to the decision node. The data partition node must then be linked to the decision tree node. Finally, when the programming has completed successfully, click the result. The outcome will then be displayed using various charts and tables.

Stage 4: Model comparison and Assessment

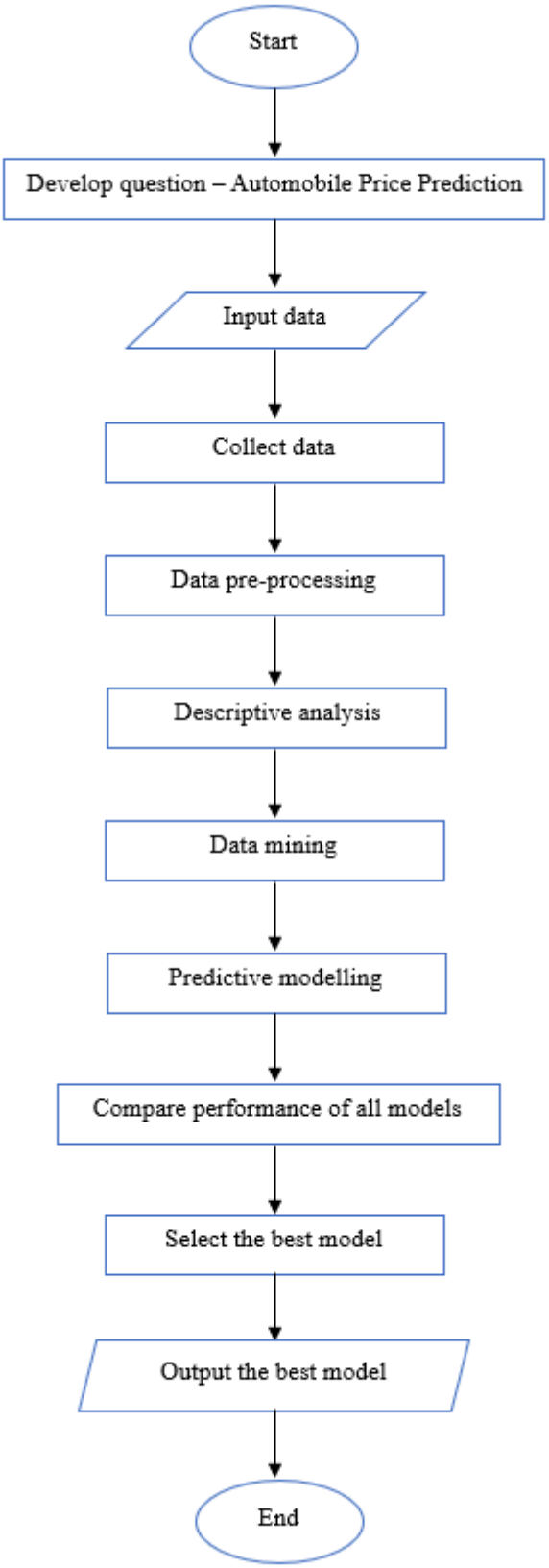
Model comparison entails evaluating different models to identify their strengths, limitations, and overall performance in solving a certain problem. This stage is usually done after the model construction process. For comparison, many models such as decision trees, support vector machines, neural networks, and ensemble approaches may be used. Typically, the following aspects are assessed during model comparison: accuracy, robustness, interpretability, scalability and training and prediction time.

Model evaluation entails a more in-depth examination of the top-performing models found during the comparison phase. The goal is to evaluate the models' capacity to generalise well to new, previously unknown data.

The models are evaluated in this stage on an independent test dataset that was not used during the training or model comparison stages. On this dataset, the evaluation metrics are produced to measure the models' performance on unseen examples. Other factors such as the model's complexity, interpretability, and practical practicality may also be considered.

The results of model evaluation assist in making an informed decision regarding the appropriate model for deployment. The selected model is predicted to perform well and be well-suited for the specific task at hand.

3.5 Flow chart



3.6 Expected Finding

Automobile prices might vary depending on the sand dataset. However, researchers may discover:

1. Factors Influencing Car Prices:

The investigation may show the main elements that affect automobile costs. These criteria include automobile make, model, year, mileage, engine size, fuel type, gearbox, and amenities. The study may reveal which variables correlate most strongly with pricing, helping stakeholders understand automobile price drivers.

2. Non-linear Relationships:

Some qualities may have non-linear associations with automobile costs. In the early years, a car's age may affect price depreciation more than afterwards. To effectively represent these correlations, consider non-linear regression or polynomial features.

3. Model Interpretability:

Researchers may examine model interpretability as well as accuracy. Linear regression and decision trees are transparent methods that show stakeholders how qualities affect pricing. Explaining pricing changes and aiding decision-making may benefit from this information.

The judgements on the price of automobiles will be more accurate and insightful if a full inquiry and topic knowledge are performed.

3.7 Conclusion

Rasha Kashef and Boya Zhang's study, "Predicting Automobile Prices Using Neural Networks," sought to create an automobile price prediction model. The research utilized historical automobile models' data, including elements that may affect costs. To find the best pricing prediction model for vehicle manufacturing facilities, the researchers used feed-forward neural networks.

The dataset has 1367 data points, one dependent variable (offer price in USD), and 27 independent variables indicating automobile model parameters. Age, mileage, fuel type, horsepower, color, number of doors, guarantee duration, and other parameters were considered.

Research has various steps. Pre-processing in Excel identified outliers, missing numbers, and discrepancies. Data were transformed or normalized to fit analysis.

Model development using SAS followed. Based on study needs, decision trees, regression, neural networks, and random forests were created.

The researchers compared and assessed various models after model building. Accuracy, robustness, interpretability, scalability, and training/prediction time were compared. For generalization, the best models were tested on an independent dataset.

The information does not include study results. However, the predicted outcomes might include identifying factors that substantially impact vehicle prices, recognizing non-linear correlations between variables and prices, and selecting the best predictive model for projecting car prices using the dataset.

The study sought to construct a forecast model for automobile pricing and assess several methods. The study strategy was methodical and used Excel and SAS for data analysis.

CHAPTER FOUR

ANALYSIS AND FINDINGS

4.0 Introduction

Data cleaning and exploration would define the actions performed to clean the data and discover variable correlations. The model development and assessment section describes how to create and evaluate the multiple linear regression model. findings and discussion would give analytical findings and explain their consequences.Car buyers and sellers may benefit from advanced vehicle pricing analyses. Buyers may assess automobile value by knowing the aspects that determine pricing. This data may help auto sellers determine appropriate prices.

This part will discuss preamble analysis, predictive modeling analysis, and results.

4.1 Preamble Analysis (Descriptive analysis)

The dataset given is then examined to understand its structure and variables. The dataset that we examined is a dataset from Canada and it is used for automobile price prediction. There are eight important variables that will affect the price of car have chosen. The following variables is chosen according their type of scales, we have chosen one of the nominal scales which is gearbox type (Auto), all of the interval scales which are offer price in dollars, age in months and guarantee term in month (G_P) and all ratio scales which are odometer kilometres (KM), cylinder volume in cubic centimetres (CC), weight in kilograms (Wght) and Horse Power (HP). Hence, the final eight variables that were selected to continue the predictive analysis were gearbox type (Auto), age in months and guarantee term in month

(G_P), odometer kilometres (KM), cylinder volume in cubic centimetres (CC), weight in kilograms (Wght), Horse Power (HP) and the price of automobile in dollars which will be our target for the further process.

After then, data cleaning techniques were used to deal with the missing values, outliers, and inconsistent data. We utilised boxplots to look for outliers in each variable, and the cylinder volume and odometer kilometres variables, CC and KM, both include outliers. The median value was then used to substitute outliers in order to lessen the influence they had on the model.

For the first variable that contained outliers is odometer kilometers (KM), there are 45 data are outliers. Since the Lower Limit is -23409 and the Upper Limit is 153687, we have 45 outliers are more than upper limit, the range of outliers is from 154464 to 232942. Hence, the outliers are replace by median for the variable KM which is 64002.

As an example, for the third column is variable KM. As we can see, from row 989 to row 1001, the value of KM is more than upper limit, hence we will will replace it using the median of the data from KM.

Before replacing outliers:

989	5950	74	232942	75	0	2000	1176	4
990	7000	80	218120	75	0	2000	1154	4
991	6050	79	217766	75	0	2000	1139	4
992	4450	75	203256	75	0	2000	1139	4
993	6800	78	200734	75	0	2000	1104	4
994	8550	70	197503	75	0	2000	1139	4
995	6200	81	194767	75	0	2000	1124	4
996	7000	77	191622	75	0	2000	1154	4
997	5800	77	183279	75	0	2000	1119	4
998	8800	79	180380	75	0	2000	1104	4
999	6550	71	178802	75	0	2000	1119	4
1000	6000	74	176179	113	0	1600	1079	4
1001	6550	79	176002	89	0	1300	1039	4

After replacing outliers:

989	5950	74	64002	75	0	2000	1176	4
990	7000	80	64002	75	0	2000	1154	4
991	6050	79	64002	75	0	2000	1139	4
992	4450	75	64002	75	0	2000	1139	4
993	6800	78	64002	75	0	2000	1104	4
994	8550	70	64002	75	0	2000	1139	4
995	6200	81	64002	75	0	2000	1124	4
996	7000	77	64002	75	0	2000	1154	4
997	5800	77	64002	75	0	2000	1119	4
998	8800	79	64002	75	0	2000	1104	4
999	6550	71	64002	75	0	2000	1119	4
1000	6000	74	64002	113	0	1600	1079	4
1001	6550	79	64002	89	0	1300	1039	4

There are 45 data of odometer kilometers (KM) has replaced by median, 64002.

For the second variable which is cylinder volume in cubic centimeters (CC), there are 72 data are outliers. Since the Lower Limit is 1100 and the Upper Limit is 1900, we have 75 outliers are more than upper limit or lower than lower limit, most of the outliers is 2000. Hence, the outliers are replace by median for the variable CC which is 1600.

As an example, for the sixth column is variable CC. Surprisingly, same as variable KM from row 989 to row 1001, the value of CC is all outliers, hence we will will replace it using the median of the data from CC.

Before replacing outliers:

989	5950	74	64002	75	0	2000	1176	4
990	7000	80	64002	75	0	2000	1154	4
991	6050	79	64002	75	0	2000	1139	4
992	4450	75	64002	75	0	2000	1139	4
993	6800	78	64002	75	0	2000	1104	4
994	8550	70	64002	75	0	2000	1139	4
995	6200	81	64002	75	0	2000	1124	4
996	7000	77	64002	75	0	2000	1154	4
997	5800	77	64002	75	0	2000	1119	4
998	8800	79	64002	75	0	2000	1104	4
999	6550	71	64002	75	0	2000	1119	4
1000	6000	74	64002	113	0	1600	1079	4
1001	6550	79	64002	89	0	1300	1039	4

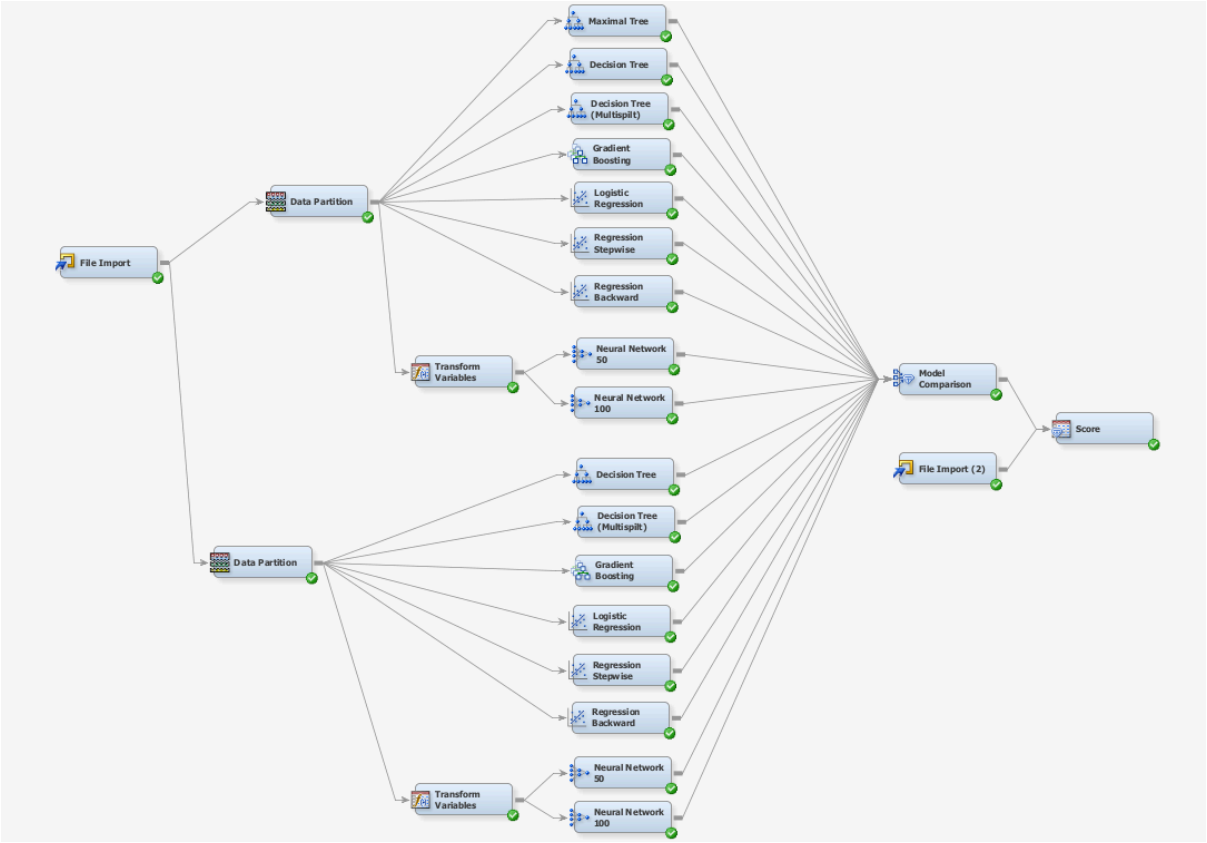
After replacing outliers:

989	5950	74	64002	75	0	1600	1176	4
990	7000	80	64002	75	0	1600	1154	4
991	6050	79	64002	75	0	1600	1139	4
992	4450	75	64002	75	0	1600	1139	4
993	6800	78	64002	75	0	1600	1104	4
994	8550	70	64002	75	0	1600	1139	4
995	6200	81	64002	75	0	1600	1124	4
996	7000	77	64002	75	0	1600	1154	4
997	5800	77	64002	75	0	1600	1119	4
998	8800	79	64002	75	0	1600	1104	4
999	6550	71	64002	75	0	1600	1119	4

There are 72 data of odometer kilometers (CC) has replaced by median, 1600. Finally, the outliers are all clean and the further process which is predictive modelling can be continued.

4.2 Predictive Modelling Analysis and Results

After cleaning and replacing all the outliers, the predictive modelling has been done by following. For this section, we have used SAS Enterprise Miner to continue it. The flow of the predictive modelling is the transformed data that had transform from preprocessed data by using Microsoft Access are import in the system. By connecting the node which have many types, file import which used to import file, data partition which used to decide the partition of data used for training and testing, various type of data mining model which used to process data, model comparison that used to compare the model based on a criteria and score which are the result that can be shown.



First of all, there is a file import, it is use to import the file that we have prepared for automobile price prediction. The transformed data is then be import in the SAS system since the data had already been preprocess at Microsoft Excel by using box-plot. To assess the model's stability, two additional data partitions were created with 70:30 and 80:20 splits for training and testing.

The trained neural network model was then evaluated using the chosen evaluation metrics on the testing set. The performance of the neural network model was analyzed, and areas for improvement were identified. The model selection and evaluation process were reapplied to these partitions to verify the model's consistency and performance across

different data splits. The results from these additional partitions were included in the final evaluation report, providing insights into the model's performance in different scenarios.

The upper part is we use data partition for 70% of result used for training and another 30% of the data used for testing while the second data partition is 80% of result used for training and another 20% of the data used. For the data mining technique node that we have connect with the data partition have decision tree, regression and neural network in different setting. Before connecting data partition and neural network, the node transform variables has used to connect both of them because it is used to improve the fit of neural network to the data given.

Decision Tree has used the model of multispilt decision tree, normal decision tree and maximal tree. By comparing the model between decision tree series, the original decision tree model will be the best model for this prediction. Moreover, the regression we have used in this predictive modelling have gradient boosting, logistic regression, stepwise regression and backward regression. Surprisingly, the average squared error of regression are mostly high for the prediction of automobile price.

Besides that, we also use neural network as one of the data mining model, which have two type of neural network which are iteration with 50 and 100. By comparing the neural network from different maximum iteration, neural network that use 100 iteration has lower average squared error for this dataset. This is because the more times of iteration, the accuracy of the model will be lower.

After that, all the data mining model are connected to the model comparison node to find out which model is the best for this dataset. This is result that have obtained which the most suitable data mining model of the dataset is neural network that use data partition of 80:20 and the maximum iterations was 100. The selection criteria we have chosen to

determine the best model is Valid: Average Squared Error. Neural4 have the lowest average squared error which is 49016.34 hence it is the most suitable data mining for this dataset.

Selected Model	Predecessor Node	Model Node	Model Description	Selection Criterion: Valid: Average Squared Error
Y	Neural4	Neural4	Neural Network 100	49016.34
	Neural	Neural	Neural Network 50	62469.69
	Neural2	Neural2	Neural Network 100	62469.69
	Neural3	Neural3	Neural Network 50	103041.3
	Tree	Tree	Decision Tree	164256.5
	Tree3	Tree3	Maximal Tree	164256.5
	Tree2	Tree2	Decision Tree (Multispilt)	188122.8
	Tree5	Tree5	Decision Tree (Multispilt)	224679.8
	Boost	Boost	Gradient Boosting	224862.6
	Tree4	Tree4	Decision Tree	230040.9
	Boost2	Boost2	Gradient Boosting	251962.7
	Reg	Reg	Logistic Regression	1612310
	Reg2	Reg2	Regression Stepwise	1622041
	Reg3	Reg3	Regression Backward	1622041
	Reg4	Reg4	Logistic Regression	1756207
	Reg5	Reg5	Regression Stepwise	1788647
	Reg6	Reg6	Regression Backward	1788647

The prediction will be continue by using a future value which we have use to import in the SAS enterprise Miner by using the second file import node. This node has allowed us to import the dataset as shown below which only include the variables except for the price because it will be predicted in the next step. This node is connected with the score node while the score node are connecting with the model comparison node.

	A	B	C	D	E	F	G
1	Age	KM	HP	Auto	CC	Wght	G_P
2	15	131463	195	0	1800	1109	10
3	23	43612	89	1	1100	1089	4
4	26	32191	113	0	1600	1069	20
5	32	23002	75	0	1800	1124	4
6	33	74133	195	1	1900	1189	4

Lastly, we can check the prediction result by using the properties of the SCORE part and the result of prediction of the automobile price of this five data has shown as below. The results were presented in a clear and concise manner, using tables as appropriate.

EMWS1.Score_SCORE

Obs #	Age	KM	HP	Auto	CC	Wght	G_P	Prediction for Price
1	15	131463	195	0	1800	1109	10	22742.23
2	23	43612	89	1	1100	1089	4	9410.741
3	26	32191	113	0	1600	1069	20	15684.15
4	32	23002	75	0	1800	1124	4	8668.029
5	33	74133	195	1	1900	1189	4	16572.02

As you can see, the result are shown, the price prediction in dollars of this five data aren22742.33, 9410.741, 15684.15, 8668.029, 16572.02 respectively. The HP variables has affect mostly to the price which we can see the data that have higher HP will own a higher prices comparing with others.

4.3 Conclusion

In this study, we used data cleaning and outlier detection techniques to prepare a dataset for predictive modeling. We then used a neural network model with 80:20 data partition and 100 maximum iterations to predict automobile prices. The results showed that the neural network model was the most accurate model for predicting automobile prices. The HP variable had the greatest impact on automobile prices. The overall results of this study suggest that neural networks can be a powerful tool for predicting automobile prices.

CHAPTER FIVE

CONCLUSION

5.0 Introduction

The automotive business is a dynamic and complicated sector that is greatly impacted by several factors that affect how much cars cost. By supplying useful insights and supporting informed decision-making, an understanding of these components and an accurate estimation of automobile pricing may be advantageous to both car brokers and purchasers. With the use of Excel functions and SAS software, we developed predictive models in this study to explore the factors affecting automobile expenses.

We carried out a thorough assessment of the literature on automobile costs before we started our investigation, looking at a variety of publications. Through this thorough investigation, we were able to pinpoint a number of variables that have been extensively discussed as important elements in influencing automobile costs. The relevance of factors including age, KM (kilometres driven), HP (horsepower), Auto (kind of gearbox), CC (engine displacement), weight, and G_P (gasoline or diesel fuel type) was highlighted by our literature research. These elements were chosen as our independent variables for future analyses that would predict automobile pricing.

We used the robust Excel functions and SAS software tools to create precise prediction models. Three different modelling strategies were used in our study: decision trees, regression, and neural networks. The effectiveness of these methods in data analysis and prediction tasks across several areas has been demonstrated. We attempted to capture the subtleties and complexities of the interactions between the independent variables and automobile pricing by using a variety of modelling methodologies.

To evaluate the effectiveness of our prediction models, we also used two types of data partitioning. We were able to divide the dataset into training and testing subsets using the data partitioning ratios of 70:30 and 80:20, respectively. Using this method, we were able to analyse our models' generalizability and performance on hypothetical data.

Our study's main objective was to provide reliable automobile price predictions. In addition, we wanted to assess how well our prediction models performed in terms of their average squared error. We were able to determine which model performed the best and choose it as the best option for estimating automobile pricing by analysing the average squared error for each model.

The findings of our study have important ramifications for both purchasers and vehicle brokers. Accurate automobile price estimates may speed up talks and reduce waiting times for both parties. Brokers may use these prediction algorithms to give customers accurate and knowledgeable price information, resulting in more seamless transactions. Similar to this, consumers may gain from a clearer grasp of automotive costs, empowering them to make wise choices throughout the purchase process.

Overall, the goal of this study was to investigate the variables affecting automobile expenses, create predictive models using SAS and Excel functions, and assess the efficacy of these models based on their average squared error. By attaining these research goals, our study offers insightful knowledge and practical solutions that have the potential to revolutionise the purchasing and selling of cars, benefiting both brokers and consumers in the automotive sector.

5.1 Benefits/ Contribution of the Study

We can learn more about Excel functions like finding the min, max, IQR, median, upper limit, lower limit, and so on throughout this study. In addition, we understand how to utilise SAS to analyse data and have developed numerous prediction models, such as decision trees, regression, and neural networks, in our research.

To begin, we can establish the elements that influence car costs by reading a large number of publications. Many articles have discussed the various elements that influence car prices. This was also explicitly stated by our researchers in our literature review.

Next, we're using SAS to build predictive models for car prices. We employ three approaches in our study: decision tree, regression, and neural network, using two forms of data partitioning: 70:30 and 80:20. We have 18 models in total that we employ to anticipate car prices.

We can predict the automobile price using a set of independent factors, which are age, KM, horsepower, HP, Auto, CC, weight, and G_P. We can also compare different predictive models based on their average squared error. This is the third goal of our research. We may select the best model to estimate the average squared error by having the result for the model comparison based on the average squared error.

We can predict the automobile price using a set of independent factors, which are age, KM, horsepower, HP, Auto, CC, weight, and G_P. We can also compare different predictive models based on their average squared error. This is the third goal of our research. We may select the best model to estimate automobile pricing using the available dataset by having the result for the model comparison based on the average squared error.

In short, our research objectives have all been achieved. It benefits both the broker and the buyer because these algorithms can estimate the car price. It will save them time as they continue to negotiate with others.

5.2 Overall Conclusion

In conclusion, our research concentrated on examining several factors that affect automobile costs and creating forecasting models utilising SAS and Excel tools. To determine the factors influencing automobile expenses, we reviewed the literature and looked at a lot of publications. With data partitioning ratios of 70:30 and 80:20, our study used three modelling approaches: decision trees, regression, and neural networks.

We successfully forecasted automobile pricing using 18 distinct models using a set of independent variables including age, KM, horsepower, HP, Auto, CC, weight, and G_P. We chose the top model for reliably anticipating automotive cost by comparing these models based on their average squared errors.

Both auto brokers and purchasers stand to gain significantly from the findings of our study. An effective and time-saving technique for calculating automobile pricing, these predictive models support bargaining and decision-making processes. By using our algorithms, brokers and purchasers may get trustworthy price estimates, saving time when haggling with other parties.

As a whole, our research met its goals, shedding light on the variables affecting automobile pricing and creating accurate prediction models. By offering useful insights for pricing strategies and enabling stakeholders to make deft decisions based on precise estimates, our study benefits the automobile sector.

References

Aditya Nickhade, & Borde, R. (2013). *Car Price Prediction using Machine Learning - IARJSET*. IARJSET. https://iarjset.com/papers/car-price-prediction-using-machine-learning/

Bhandari, P. (2022). Data Collection | Definition, Methods & Examples. *Scribbr*. https://www.scribbr.com/methodology/data-collection/

Bilen, M (2021). Predicting Used Car Prices with Heuristic Algorithms and Creating a New Dataset. Journal of Multidisciplinary Developments. 6(1), 29-43, 2021. https://www.researchgate.net/publication/356109326_Predicting_Used_Car_Prices_with_Heuristic_Algorithms_and_Creating_a_New_Dataset

Gupta, R., et.al (2022). 2022 International Conference on Inventive Computation Technologies (ICICT). IEE. https://ieeexplore.ieee.org/abstract/document/9850657/figures#figures

Hettige, B., Deemantha, S. (2023). Autonomous Car: Current Issues, Challenges and Solution: A Review. https://www.researchgate.net/publication/366986201_Autonomous_Car_Current_Issues_Challenges_and_Solution_A_Review

Hu, L., Wang, P., Zhang, Q. (2019). Indirect network effects in China's electric vehicle diffusion under phasing out subsidies. Appl. Energy 251, 113350. https://www.semanticscholar.org/paper/Indirect-network-effects-in-China%E2%80%99s-electric-under-Zhu-Wang/0a1871b99afe0e91002721afed9f81b27499afe7

Kashef,R., Zhang, B (2020). Predicting Automobile Prices Using Neural Network. IVEY Publishing.https://www.iveypublishing.ca/s/product/predicting-automobile-prices-using-neural-networks/01t5c00000CwpP3AAJ

Kenton, W. (2022, October 10). What is nonlinear regression? comparison to linear regression. https://www.investopedia.com/terms/n/nonlinear-regression

Kharwal, A. (2021). Car Price Prediction with Machine Learning. Thecleverprogrammer. https://thecleverprogrammer.com/2021/08/04/car-price-prediction-with-machine-learning/

Lee, J. A. (2014, November 7). *Measurement scale | Statistical Analysis, Types & Uses*. Encyclopedia Britannica. https://www.britannica.com/topic/measurement-scale

Liedtke, S. (2020). South Africa's manufacturing output was down 49.4% on Covid-19 impact (Online). https://www.engineeringnews.co.za/article/south-africas-manufacturing-output-down-494-on-covid-19-impact-2020-07-09/rep_id:4136

Muzir, N. a. Q., Mojumder, M. R. H., Hasanuzzaman, M., & Selvaraj, J. (2022). Challenges of Electric Vehicles and Their Prospects in Malaysia: A Comprehensive Review. *Sustainability*, *14*(14), 8320. https://doi.org/10.3390/su14148320

Sanap, V. C., Rangila, M. M., Rahi, S., & Ospanova, A. (2022). Car Price Prediction using Linear Regression Technique of Machine Learning. ResearchGate. https://doi.org/10.15680/IJIRSET.2022.1104050

Selvaratnam, S., et.al, (2021). 2021 International Research Conference on Smart Computing and Systems Engineering (SCSE). https://www.researchgate.net/publication/355633791_Feature_selection_in_automobile_price_prediction_An_integrated_approach/link/627cd723107cae29199f2d54/

Sharma, P. (2022, January 19). *Different types of regression models*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/

The Neural Networks Model. (2021, March 4). https://www.ibm.com/docs/en/spss-modeler/18.0.0?topic=networks-neural-model

What is a decision tree. (n.d.). https://www.ibm.com/topics/decision-trees

What is a regression model?. IMSL by Perforce. (2021, June 16). https://www.imsl.com/blog/what-is-regression-model

What is Random Forest?. IBM. (n.d.). https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems

Wenguan, Y, et.al. (2021). Claim Amount Forecasting and Pricing of Automobile Insurance Based on the BP Neural Network. *Hindawi Complexity Volume 2021*, 6616121, (17). https://doi.org/10.1155/2021/6616121

Ye, F., Kang, W., Li, X., Wang, Z.(2021). Why do consumers choose to buy electric vehicles? A paired data analysis of purchase intention configurations. Transp. Res. A Policy Pract. 147, 14–27. https://www.sciencedirect.com/science/article/abs/pii/S0965856421000495