Research Study on predicting Used car Sales using Big Data

Aditya Trivedi
Department of Systems and Computer Engineering
Carleton University
Ottawa, Canada
adityatrivedi@cmail.carleton.ca

***Abstract*—The world is facing the effects of Pandemic since the year 2020, there has been a toll on many manufacturing plants all around the world who are aligned in automotive sector due to shortage of semi-conductor chips. Due to this prime reason, the sale of new cars has seen a dip and the sale of used cars has been steadily increasing. Looking at this phenomenon, we can label it under time series domain of Data intensive application as there is a trend seen in the sales of cars which is temporary and based on various factors. This paper presented here is a research project which focuses on the Time series analysis domain of Data intensive application specifically leveraging the capabilities of Auto Regressive Moving Average(ARIMA) model which is a statistical method used for forecasting. The paper has achieve the concept of combining external factors in predicting the sales of used cars which has not gathered much attention in the research domain. The paper will also talk about different insights apart from forecasting the sales which can be helpful in various businesses through the platforms R Python respectively.**

I. Introduction

The project "A research study on predicting used car sales due to impact of chip shortage using Big Data" is a collaborative study by employing quantitative and qualitative research methods in order to predict the sales trend of the used cars due to the global shortage of the semiconductor's chips in the United States of America. With the analysis of big data datasets and applying time series prediction models with the help of available tools and techniques, the project has present a forecast for the sales of the used cars and gain insights on the market trend of the used automobiles industry. This research is unique and important as the world is still in the clutches of COVID-19 pandemic and because of it since last 2 years, countries have been facing issues which has led to shortage of semiconductor chips manufacturing and its successive challenges in the domain of supply chain management[1]. As the demand of automobile vehicles is increasing day by day it becomes difficult to meet it with the shortage of semiconductor chips. Thus, this research project could help in identifying the future trend of the sales of used card available in the market. For a broader perspective of the research, it has be carried out for answering the following question.

- Q1 Effect on sales of used cars due to semi-conductor chip shortage and how to forecast the trend of future sales?

II. Study Design

The study of this research has been carried out in a phase wise manner. Intensive research matter artifacts has be searched from various sources along with data used for developing the model and using it for prediction. The phase of the study is mentioned as below

- The author have searched for subject matter related research
- papers from platforms like IEEE Explore, ACM Digital Library, Google Scholar, and International Journal for Data Science and Analytics etc.
- The author has work on the "Used Cars in the United States of America" Dataset in terms of data preprocessing and data modeling aspect.
- Data training and analysis has be conducted on various data analysis and visualization tools.
- The last step involved implementing quantitative and descriptive analysis on the processed data in order to answer the research question.

III. Motivation

In the recent times due to impact of COVID-19 many industries including the auto manufacturing has suffered huge loss due to intermittent shortage of chips and procurement of the produced chips due to supply chain management challenges. Also, United States of America is considered as the car hub of the world with being 2 nd Largest car manufacturer and distributor after China. Due to the posing issues mentioned above, the market of used cars in the USA had surged exponentially in 2021 and that the prices of used cars have increased up to 32% compared to 2020. This rise was seen due to increasing supply shortage of chips and forcing the car manufacturing plants to pause or shut down the plants for an indefinite period. Thus, the motivation is to understand

the underlying insights of how the sales and prices of used cars in future has look like due to the present situation.

IV. Related Work

A. Paper 1

One of the most volatile and un-predictable domains of business is that of Crude Oil. Due to regular changes in the world economy, the prices of Crude Oil are dynamically changing because of the impact of various factors. This paper further provides an overview of predicting the prices of Crude oil by considering external factors which are responsible for pivoting the prices of Crude. The authors have implemented the Auto regressive Moving Average (ARIMA) model for predicting the prices of Crude oil. The authors have briefly mentioned the working and functionality of ARIMA model in the Section-II of the paper. The authors constructed an ARIMA Model and combined with a Radial Basis Function (RBF) neural network. After computing the time-series linear regression with ARIMA model, the results were combined with the RBF and prediction analysis was carried out by the authors. The aspect that was notable was the context of topic selection by the authors as this field related to crude oil is exceptionally dynamic and the authors fairly managed to bridge the gap between ARIMA model's application and a real-world business stream where the concepts of Artificial Intelligence and Data Mining can be applied. Also, the authors explained the whole procedure of how they have computed the prices right from constructing the time sequence scatter plot till prediction of the prices through the developed combination of ARIMA and RBF Neural network. This type of explanation often helps the readers to grasp the insights of the research being done by the authors and also helps in visualising the process of developing hypothesis till reaching to the actual solutions. Also, the statistical analysis performed by the authors is evident of the fact that the authors have greatly invested their time and efforts in studying the dynamics of ARIMA model and applying it in the research study. Since the authors have not completely shown the actual calculations and statistics of the ARIMA and RBF Neural network so it is difficult to understand the whole calculation in terms of providing weights to the data parameters, what has be the error factor in the ARIMA model, how the errors are encapsulated etc. Authors should have provided this information for better transparency in understanding. Although, the authors have considered the time period of January,1980 to December 2005 but they haven't mentioned the factors impacting the prices of crude oil in the observation time period. Also, to replicate this model in further research and applications, an individual won't be able to apply it with the correct information because there lacks the application method or usage method of this research; whereas in the introduction section it is mentioned by the authors that this model has be helpful in predicting crude oil prices with time-series analysis. Overall, the paper was insightful and provided important aspect on the application of ARIMA model in predicting numerical or analytical parameters for various business and activities which are linked with time-series phenomenon where prices or sales are affected over a certain period due to changes in the deciding parameters of prices[3].

B. Paper 2

This paper the authors used the ARIMA modeling algorithm in order to predict the air quality of a city in Columbia. Due to its novel concept, this type of research papers are useful because it helps in planning important strategies to combat the daily life problems. The authors took observation from different stations(places in the city) and combined together under pollutants such as O3, PM2.5, and PM10. Further, they analysed the pollutants over a time of day, month, and year and external factors such as wind direction, and season. Then they predicted the precipitation of this pollutants and air quality based on ARIMA modeling analysis. This paper falls under the prediction category of data analysis of big data. Also, one impressive thing about this paper is that the authors have mentioned the overall process of Data collection, Data pre-processing, Data analysis, and Predictive analysis of the collected data. This falls in line with the Big Data analysis methodology principles and also as a reader it is evident in the paper. As many tools are available in the market, authors have used IBM SPSS V.24 which is one of the advanced statistical tool used for performing complex calculations, which is noteworthy. The authors in this paper actually identified the relationships between the pollutants and the places where they are found, the frequency of time period in which they are found, the observation stations where they are found in considerable manner, and how the direction of wind is affecting the pollutants spread across the city of Bogota. Few things that could have been improved are that the authors could have included more graphical representation of SPSS tool as the reader can come to know about the tool more, along with the distribution maps of the pollutants for all the observation stations for more intuitive information. Also, empirical information in the paper was missing In the paper which is required when the authors are working with data but it was not present. Further, the authors could also have used clustering techniques in order to cluster the pollutants prediction data based on observation stations[4].

C. Paper 3

The authors in this paper have predicted the car sales tied with the Google Search keywords. This type of prediction is dependent on factors such as accuracy, homogeneity of keywords, correlation of keywords with the data source. Authors of this paper have presented in a very detailed manner how they have predicted the car sales by considering various factors. The depth of paper was evident while reading

it, all the information were presented in a nuance manner and useful for the readers. Also, the use of graphs, process flow diagrams, bar-charts of results made it more convenient for the reader to understand the concept. The authors have used univariate linear regression models in the paper for predicting the data. Whereas, considering the seasonality of the data with the keywords, Holt Winter's exponential smoothing can also be used which takes prior time steps, seasonality, and trends in account while predicting the data[5].

D. Paper 4

Electric vehicles in consumer segment are becoming much popular many countries, and this research paper published by the authors is related to predicting the sales of Electric vehicles along with prediction of CO2 gases in the environment due to EVs. The paper was a considerably large and had an in-depth analysis because of its highly content rich research domain and problem statement. The paper in the first few sections talked about the current scenario relate to production and sale of EVs in 26 countries across 5 continents. Furthermore, it also discussed the emissions of CO2 gases due to gasoline and diesel-powered vehicles. In the next few sections there was a discussion about the methodology in which the authors selected the logistic growth function which is a part of logistic regression model. The paper is written in a deeply analytical manner where the authors have taken the data analysis period of 2005-2018 and the prediction of sales of EVs is done till 2035. As a reader, it was easy for me to grasp the content of the paper as it had sufficient amount of diagrams and charts which were explaining the logistic growth model, amount of CO2 gas being emitted by countries currently and sales of EVs per country. Also, the graphs related to Logistic growth function mentioned about the saturation and inflection point which denoted the start of peak sales growth for EVs from the year 2023-2031. I liked the description where the countries were clustered according to their probable EVs sales and CO2 emissions till the year of 2035 as it provides a general idea of the whole paper in a gist. The cons of paper were less as it was presented in a perfect manner, but if they authors could have included more information about the logistic growth model and the empirical or statistical tools used for the analysis purpose then it would have been beneficial for the readers[6].

E. Paper 5

The authors of this paper have conducted a research study on predicting the car sales of Audi by using the Web search and Social Network data in the period of 2011-2019. Automotive companies can rely on insights of this type of data which is generally considered as implicit or indirect data generated from various sources. As pe the results obtained by deploying various models of Machine Learning like Support Vector Machine, Random Forest, and Deep Learning it was found that Support Vector Machine has greatest accuracy out of all the models. The paper was having a strong backing of various tools as the authors have used numerous tools such as RapidMiner for deploying Machine Learning Models, Selenium for writing Python scripts to crawl data, Weibo and Baidu API tool for performing sentiment analysis and search indexing respectively. Due to such a large usage of tools, the reader can also come to know about different tools available in the market to solve Big Data problems which is noteworthy. Also, the authors have depicted the required features of a research study like methodology, analysis, and results in a very short but effective manner. Due to depiction of results obtained through different deployments of model in a graphical chart manner along with tabular manner, it becomes easy for the reader to understand the results. The most interesting part in the research paper was regarding segregation of different sentiments i.e positive and negative and labelling them as 1 and -1 in the data analysis part as it shows that authors have treated the sentiments in the way they should be treated and so that it doesn't affect the validity of the results. The only disadvantage observed in the paper was that there is not enough information on the empirical formulas and maths related to the ML models deployed which could have been an added advantage. Also, the images taken from the tools should also be applied into the paper for better understanding of the tools usage[7].

F. Paper 6

The research paper published talks about predicting the car sales using the methodology of Fuzzy Logic by applying Sturges formula. The paper talks about applying FTS and ANFIS algorithms which falls under Fuzzy logic method to determine the car sales. The paper was more inclined towards statistical aspect of applying Fuzzy logic in predicting car sales, which is a good aspect whenever the authors are looking for a research paper with strong mathematical background. Also, there is a decent amount of usage in terms of charts and graphs for depiction of results obtained by applying Fuzzy logic through MATLAB software. Although, the paper title shows relevance about predicting car sales, but the paper is more focussed about theory related to Fuzzy Logic. Also, the authors have not mentioned the data analysis period for predicting the car sales, along with information related to types of cars, companies of cars, attributes related to predicting cars, on what basis the prediction is done is missing in the paper due to which is became difficult to understand the content flow of paper. In the conclusion the authors have just mentioned about the accuracies of algorithms such as FTS and ANFIS which does not portray anything about the research subject as the authors have not clearly specified that on what basis the prediction is being made[8].

G. Paper 7

The paper published by the author talks about a adaptable and user friendly traditional approach of predicting car sales which is related to data mining in general and calculating the prices of used cars in the Romanian Market. The authors have developed a tool using backend and front-end languages to calculate the price of cars. It's a decent paper which talks about an unique aspect of Big Data application in the domain of predicting car sales with the help of web interface. The explanation provided by the authors is also effectively written as they have mentioned in detailed about how they have used MS Access for maintaining a database by web scrapping the Romanian website for car listing, developing backend module in C and front-end module in Angular JS. They have mentioned each aspect of the web application with sufficient graphs and charts to support their research. The one thing I learnt in this paper which is noteworthy is regarding backend architecture of storing and utilizing the information stored in database for data mining. Also, the authors have provided information through sequence diagrams of how each submodule of backend and front-end application communicates with each other. The disadvantage of this paper is that the author has not the captured the coding part of the whole application which could have been useful in studying the architecture in a more useful manner. Also, the paper is lacking information about time in which the data is mined as it is useful in this kind of applications. Also, they have not considered any limitations related to data being mined for this paper[9].

H. Paper 8

The paper published by the authors is about mapping the customer behaviours with the car sales in the Croatian Market which is obtained by applying various statistical functions on the data set. The data collected by the authors was collected between 2014 to 2019 and it was collected in real time through a questionnaire for a large automotive company. The authors have effectively and uniquely presented the flow of the paper as it becomes easy to understand the different aspects of the paper while reading it. The authors have predicted the sales of the cars for a large automobile manufactures by the method of logistic regression and they have provided some extremely useful insights which are unique and useful. The paper has enough information in terms of theory and graphical representation to provide information on what is being researched and how the research is being conducted. Some useful finding suggests that the distance between the customer and the dealer is greatly significant in the process of purchasing a car. Also, the males traveled more compared to females for buying a car. This kind of insights gathered by the authors through regression technique is specifically useful for the marketing strategists and analysts for preparing offers and selling strategies. This paper showed the prowess of regression model. The authors could have mentioned which company was being analysed in the paper and what models were being analysed as it could have provided more transparency in the research. Also, the authors could have mentioned the use of tools and techniques employed in the research paper for better understanding[10]

I. Paper 9

The paper published is another application of ARIMA model used for predicting the stock prices. This paper is unique in nature as it shows the capability of ARIMA model for predicting the time series problems such as stocks. The authors have employed a hybrid model by combining ARIMA and XGBoost ML models for predicting the stocks prices of 10 stocks in the period of 2015-2018. The paper was a delight to read as it provided information about the ARIMA and XGBoost models in detail, also the authors comprehensively presented the information collected by them and also pictured them in various graphs and charts for the ease of the reader. The interesting part of the paper was when the tabular forms of ARIMA and XGBoost results were depicted for all the 10 stocks as it gives a bird eye view of the whole research paper. The authors have not mentioned the training and testing dataset, also the basic information of the dataset is missing as from where it is collected I.e Stockexchange, what are the parameters present in the dataset, also which tools have been used by them in order to obtain the results. Overall, it was a results driven paper which portrayed the application of ARIMA time series model as it was proven that ARIMA model has greater accuracy over XGBoost ML model in predicting stock prices[11].

J. Paper 10

This is an unique paper which is published by authors to predict the car prices of models which are never made by BMW based on the features and specifications of the already developed cars by BMW with the help of Business Intelligence and regression models. The algorithm used is Feed forward Back propagation method. The paper was a nice paper to read and it had positive impact as it was one of the most easy to read and reader-friendly paper where a layman can understand about BI, regression and Neural networks used in the paper for predicting the prices of BMW cars. This type of paper are always in demand as it generates curiosity amongst the readers. The authors have comprehensively provided information on the period of dataset, the method of training and testing neural network, and knowledge about Feed Forward Back propagation method as it is not being used very frequently. The paper explained in a detailed manner how the features which are already developed by BMW in its cars are impacting the prices and how they can be considered while predicting the prices of future cars. Also the information related to error handling and correction

is also mentioned in a detailed manner which provides a sense of transparency and acceptance of in accuracy by the trained neural network. In conclusion, it was a thorough paper where researchers focussed on just one company to have an overall insights of it and they effectively produced the results through Feed Forward Back propagation method[12].

K. Paper 11

The paper published by the authors is complete research which is produced for predicting the prices of used cars in the Indian Market. The authors have employed Machine learning technique and various algorithms under it to predict the prices of used cars. This paper has used dataset obtained from Kaggle which is an open-source platform for accessing large datasets across different domains. The authors have tried to include everything in the paper and it is also evident on reading as they have performed literature review of previously published papers and presented in a tabular format, performed data pre-processing and analysis of dataset at feature level, performed outlier and error analysis and omitted the extra features presented in the dataset, performed categorical feature encoding during the data analysis step, performed training and testing of the dataset in the proportion of 80:20. There are many positive aspects of the paper apart from mentioned above as looking at the paper it is obvious that the research might have taken a long time to complete and authors might have gave a large effort. The authors have mentioned each and everything in detailed manner and they have also create a web interface for the end-users to predict the user car prices. They have also mentioned the pseudo code of all the models such as SVM, Linear Regression, and Decision tree. Adequate graphical representation about the research flow, ML model execution flow, code snippets, and analysis results is available in the research paper. The amount of knowledge obtained from the paper can be utilized in future research as its an easy and content driven paper aiming at providing maximum knowledge about the ML models for predicting the used car prices in India[13].

L. Paper 12

The paper published by the authors talks about the impact of cost and revenue on the profitability of a car sale with the auto major companies of India. The authors have used Radial Basis function(RBF) to achieve results. Although, the authors have taken the data for 5 major automotive companies of India but they have haven't mentioned the time period in which the data was captured. Also, there is information only about the methodology of RBF and how it is being utilized in the research and not on the various features of dataset so it becomes a na¨ıve study related to studying impact of cost and model in a normal scenario on the profitability of the car sale. Also, the research is performed assuming normal scenario but impact of various external factors is not considered which can be considered as outliers or impactful factors depending on the situation, thus it becomes more of a static study[14].

M. Paper 13

The issue observed in the previous literature review of paper has been successfully addressed in this paper where the authors are considering external factors such as Covid-19 and sales of cars In India by top 5 auto manufacturers during the period of 2019-2020 through Random Forest regression. The authors presented a small and content rich paper wherein they have mentioned the different aspects related to impacting the prices and sales of car. It is evident from the paper that the authors tired their hands on different models before arriving on Random Forest Regression as it shows that they were not satisfied with the previous results. The graphical representation in the paper is also direct and informative which can be understood by the readers. They have precisely weighed the coefficients from the datasets in order to fit under the training and testing data sets. The authors could have included more information about the Random Forest Regression and how they have achieved the results like code or pseudocode, tools etc[15].

N. Paper 14

The paper published by the authors is a cumulative analysis of the sales of new products, specifically car in this case without any historical data of the product. The paper is a comprehensive way of analysing the sales of new cars which are having new features and don't have any historical data. The authors have proposed an unique approach based on Eucledian Similarity measurement in which the authors have analysed random features for upto 3 months and predicting the data on the same. The authors have also provided information on the features considered for predicting the sales. Also, the authors have compared the results with other models such as KNN, SVM, Random Forest Regression which is an effective way of testing the accuracy of the proposed model. Although, the authors have considered the features of new car but they have not discussed the impacts the new features are having on the car sale which could have been beneficial. Overall, the paper was decent and content driven with strong results[16]

O. Paper 15

The paper is an application of the ARIMA model in the field of predicting COVID-19 cases for India and the USA. The authors have used the dataset available from open source platform and analysed it by analysing it under ARIMA model. This paper is only considering the cases and factors related to COVID-19 but it misses the factors related to geography, number of treated cases etc as it could be helpful in such data analysis. Also, the cases reported by online platforms

are often mispronounced because they are being fetched from Government records and as we have seen in the past that Government records were often falsified. The paper has strong mentioning of empirical formulas related to ARIMA which is essential[17].

P. Paper 16

The paper published is talking about the prediction of export and wholesales car sales in Indonesia using the HoltWinters exponential smoothing technique which is a kind of forecasting method for time series analysis similar to ARIMA from the time period 2011-2019. The paper was new for me as I have not read and explored about the Holt-Winters exponential smoothing technique, but the authors made it very easy to understand through their paper. Also, they have considered 4 top performers of Indonesian car market which helps in having an overall picture of the market. The paper has large amount of graphical representation which is easy to understand and also emough information on the theoretical portion of Holt-Winters exponential smoothing as well. Although, the authors could have only considered only export or wholesale market car sale data as mixing it both can be subjected to erroneous prediction because the export sales is happening outside the country and it is dependent on different scale of demands. Also, the authors have not mentioned by using which tools or language this research was carried out because it might be possible that functions and features for Holt-Winters exponential smoothing might not be available in all the programming languages. Overall, it was an informative paper which can be used in predicting the sales of cars for other companies and countries as well, but can the method proposed handle external factors affecting the sales is not a answer with conviction [16].

V. Problem Statement

It is well known that when there is a global crisis, every area of the market, including the automobile industry, is severely impacted. The catastrophe described here is not a common occurrence, but rather a once-in-a-generation occurrence that has the capacity to break or collapse an entire company all at once. When it comes to the effects of this pandemic, the automobile industry has gotten a lot of attention. This is one of the industries that has been severely damaged since the outbreak. Automobiles have been the least worrisome commodity for the population due to the country's lockdown and curfew. People are not considering anything other than the essentials, which is completely understandable and natural.

However, looking back at the history of the automobile business, it's been pretty fascinating to see how the industry would fare in the aftermath of the outbreak. Since many businesses have been harmed by the lockdown, a few have begun to develop future plans to entice clients. Businesses do not want their customers to get bored with them. They're coming up with new techniques and ideas to boost their sales once more. It has undoubtedly be a lengthy process, but the motor company has undoubtedly recover. Because everyone in the vehicle sector has been impacted by this situation, the burden of loss is not evenly distributed. In comparison to them, some corporations have to absorb significant losses, while others do not. A prediction model has shown to be an invaluable tool in determining how to save a corporation from a pandemic catastrophe. That particular service has been served by this model.

The model has show the organisation how to save itself in a crisis situation and evolve quickly, among other things. Such a forecasting technique has shown to be extremely effective in allowing such businesses to re-emerge powerfully by insuring their worth while keeping an eye on production requirements and sales volume. The research began with a basic concept: anticipating used cars sales outcomes during Covid-19. However, in reality, there are other elements that influence a car sales. It is easy to forecast what has happen next in terms of production and sale of a company based on this premise (if every other thing goes well). Some of the dependent components are: natural resources, labour, capital, all sorts of goods and normative and positive affirmations. All of the world's increasingly complicated electrical and digital devices are powered by hips, or semiconductor devices.This includes well-known devices such as computers and cellphones, as well as other "smarter" products such as appliances, watches, and, most notably, automobiles. With current cars having smart and complicated entertainment systems, navigation, and sensors, the automotive industry accounts for a considerable share of global chip consumption. A modern car's various features can be powered by anywhere from 500 to 1,500 separate chips. Converging causes have resulted in a semiconductor shortage. Before the COVID-19 epidemic, demand for microprocessors was increasing to enable the development of new markets including as 5G, self-driving vehicles, artificial intelligence, and the Internet of Things. Automobile plants around the world were compelled to shut down in the early weeks of the COVID19 lockdown, and sales were dramatically decreased. As a result, the automotive industry cut back on semiconductor procurement. However, due to an expansion in online schooling and people working from home, the lockdown facilitated an increase in PC demand; however, this need was unmet, keeping PC growth in the single digits. Predictive analysis is often pivoted on certain factors which contribute in predicting the data for future. Automotive sales by considering external factors using predictive analysis is having a substantial research present in real world but there are few papers which talks about using ARIMA as the tool for predicting the car sales. Along with this, whenever the time series analysis has been used, the authors have considered available machine learning models such as Random Forest Tree, Support Vector Machine, and Decision Tree etc and they have not considered the external factors which contributes

to the sales of automobiles. Whereas, the research works produced as an application of ARIMA analysis are not having have only considered the static data confined in a close time interval by taking in account few important factors and leaving the rest.

This project has try to combine the missing portions from both the types of current research work and has accommodate the differences observed. Thus, the main goal of implementing ARIMA analysis is to consider a specific time interval for observing the data and consider the effects of chips shortage on the used cars in the observation period. By exercising the analysis on specific time interval, it has help to predict the sales of used cars in future when similar catastrophe is occurring. The main gap in the available research and the proposed research in this course is taking in account the impact of the external factors which is having on the automotive sales as it is not covered in any of the available research. Time-series forecasting is a method for forecasting events over a period of time. Weather forecasting, earthquake prediction, astronomy, statistics, econometrics, signal processing, and other fields of study employ the technique.

Timeseries data prediction has been a hot topic of interest in the industry and academia since its inception, especially with the emergence of computer technologies that can process large amounts of data, due to the large number of completely different fields in which getting an accurate prediction is a fundamental piece. Computer technologies are now used in all modern time-series forecasting applications, which employ various models such as ARIMA, artificial neural networks, support vector machines, hidden Markov models, and so on. All of these diverse time-series forecasting models have one thing in common: they all rely on the assumption that future trends has be similar to previous patterns. To put it another way, they use historical data to forecast future data. As a result, they're sometimes referred to as pattern-based time-series forecasting models. Some studies have looked into employing external elements as features in the model instead of standard patternbased models for these applications. In this paper, we continue to investigate the idea of including external inputs in models while keeping pattern-based models as the foundation. With this method, we believe we can strike a reasonable compromise between model interpretability and predicting accuracy.

VI. ARIMA Implementation

ARIMA models are a type of statistical model that can be used to analyse and forecast time series data. Although it is quite simple to use, this model is really strong. Auto-Regressive Integrated Moving Average (ARIMA) is an acronym for Auto-Regressive Integrated Moving Average.

The following are the parameters of the ARIMA model: p: The lag order, or the amount of lag observations incorporated within the model. d: The degree of difference is the number of times the raw observations are differenced. q: The order of moving average, also referred to as the dimensions of the moving average window As the project has implement Predictive analysis hence, the below figure shows the block diagram of predictive analysis methodology. Wherein, the dataset has been undergone the pre-processing procedure which includes cleaning, handling missing data, and feature selection[20].

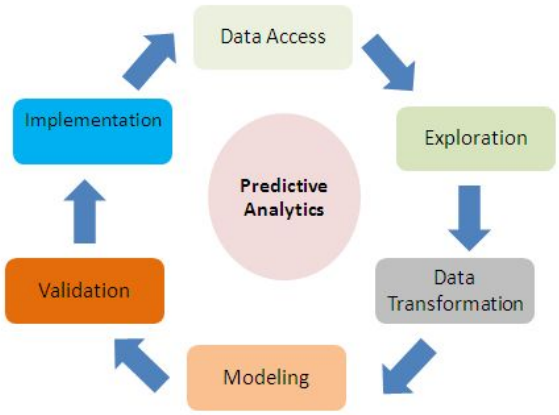

Fig. 1. Predictive Analytics

Also, the ARIMA model is a part of time series analysis and below diagram shows the flow of how ARIMA has be implemented in the project[21].

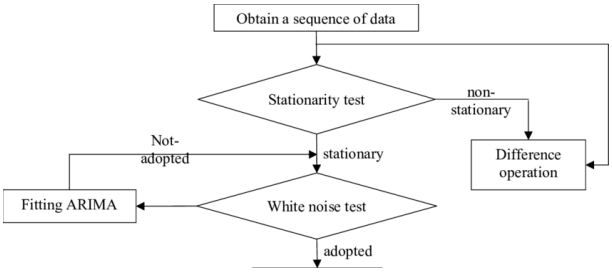

Fig. 2. Flow Chart of ARIMA Model

Thus, this section frames the technical aspect of the project by scoping it with the help of content related to the Machine Learning model such as ARIMA, also it mentions what are the steps required to implement the ARIMA model. In other words, laying down the foundation for the successive stages of the project. The next steps has be covering the implementation of aforementioned methodologies on R Language, as it gives higher efficiency and features in order to implement ARIMA models.

VII. Time Series Modeling

Most businesses use time series data to examine sales figures for the coming year, website traffic, compet-

itiveness, and other factors. It is, nevertheless, one of the areas that many experts are unfamiliar with. Methods for studying time series data in order to extract useful statistics and other aspects of the data are referred to as time series analysis. The employment of a model to predict future values based on previously observed values is known as time series forecasting. While regression analysis is frequently used to assess links between one or more separate time series, it is not always referred to as "time series analysis," which refers to relationships between different points in time within a single series. Interrupted time series analysis is used to find changes in a time series' evolution from before to after some action that may alter the underlying variable. Temporal ordering means events occurring in a series of time.

The prediction of used car sales during the impact of shortage of semi conductor chip is being considered as function of time because the data collected is during the period of COVID19 i.e 2020-2021 in which due to supply chain management issues, the procurements were delayed and specifically the semiconductor chips were not delivered to the manufacturing units which resulted in halting of the auto manufacturing plants, and as a successive effect the sale of used cars was impacted considerably.

Amongst various types of models like Classification, Curve Fitting, Descriptive analysis, etc; this research project is focussed on the Forecasting model under Time series analysis methodology.

The forecasting methods involves analysing the historical data as a model for the future data, and predicting scenarios that could happen along future points. The data classification is done with the help of Flow Time Series data which means determining the activity of the qualities during a specific time period, which is usually a part of the overall picture and contributes to a portion of the outcomes.

VIII. ARIMA Implementation

A generalisation of an autoregressive moving average (ARMA) model is an autoregressive integrated moving average (ARIMA) model. Both of these models are used to fit time series data in order to better understand or anticipate future points in the series (forecasting). When data shows signs of non-stationarity in the sense of mean (but not variance/autocovariance), an initial differencing step (equivalent to the "integrated" element of the model) can be applied one or more times to eliminate the non-stationarity of the mean function, ARIMA models are used (i.e., the trend). When a time series exhibits seasonality, seasonal-differencing can be used to remove the seasonal component. Because the ARMA model is theoretically sufficient to describe a regular (a.k.a. purely nondeterministic) wide-sense stationary time series, according to Wold's decomposition theorem, we are motivated to make stationary a non-stationary time series, e.g., by using differencing, before we can use the ARMA model. Note that in the ARIMA framework, if the time series contains a predictable sub-process (a.k.a. pure sine or complex-valued exponential process), the predictable component is handled as a non-zero-mean but periodic (i.e. seasonal) component and is eliminated via seasonal differencing. The AR in ARIMA refers to the fact that the evolving variable of interest is regressed on its own lagged (i.e. prior) values. The MA section of the equation denotes that the regression error is a linear combination of error terms whose values happened simultaneously and at different times in the past. The letter I (for "integrated") denotes that the data values have been replaced with the difference between their current values and their former values (and this differencing process may have been performed more than once). Each of these features is designed to help the model fit the data as closely as feasible. ARIMA(p,d,q) is a non-seasonal ARIMA model in which the parameters p, d, and q are non-negative integers, p is the autoregressive model's order (number of time lags), d is the degree of differencing (the number of times the data has had past values subtracted), and q is the moving-average model's order. The capital P,D,Q refers to the autoregressive, differencing, and moving average terms for the seasonal portion of the ARIMA model, while the lowercase P,D,Q refers to the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model. A non-stationary time series is rendered stationary in ARIMA models by applying finite differencing to the data points. The ARIMA(p, d, q) model's mathematical formulation is as follows[22]:

$$\varphi(L)(1-L)^d y_t = \theta(L)\varepsilon_t$$

$$\left(1 - \sum_{i=1}^{p} \varphi_i L^i\right)(1-L)^d y_t = \left(1 + \sum_{j=1}^{q} \theta_j L^j\right)\varepsilon_t$$

Fig. 3. ARIMA Equation

L i s the lag operator, w is the parameters of the autoregressive part of the model, theta the parameters of the moving average part and epsilon(t) are error terms. • p,d,q are integers greater than or equal to zero and refer to the order of the autoregressive, integrated, and moving average parts of the model respectively. • The integer d controls the level of differencing. Generally, d=1 is enough in most cases. When d=0, then it reduces to an ARMA(p,q) model. • An ARIMA(p,0,0) is initially the AR(P) model and ARIMA(0,0,q) model is the MA(q) model. • ARIMA(0,1,0) which mathematically is yt= yt-1 + epsilon(t) is a special one and known as the Random Walk model. It is widely used for nonstationary data, like economic and stock price series.

Also, the ARIMA model is a part of time series analysis and below diagram shows the flow of how ARIMA

has be implemented in the project.

IX. Purpose of Using ARIMA

The ARIMA models requires huge amount of data to train the model, hence this research project embodies a huge dataset which consists of 9.98 GB of data of used cars listed on Cargurus Inventory in the US. The main benefit of having such a huge data is that the structure of the data remains same for a certain portion in the dataset and there is not much deviation in the features of the data. This has helped in extending the focus data part to a considerable amount and the moving average is not disturbed because of this reason for a huge portion of the data.

The prediction of events with the help of historical data using ARIMA modeling has maximum efficiency when the prediction is performed on shorter time ranges. As the data frequency is not changing in a shorter time ranges so it becomes convenient to identify, treat and consider the features of the data while prediction as they have not changed considerably. They also have the advantage of being less sensitive than many other systems to the underlying assumptions regarding the nature of the data fluctuations.

As the research project is having consideration of external factors in order to predict the data, hence with latest developments in the field of ARIMA, studies shows that external features can be combined while determining the prediction of the data. This means they employ all external elements as model inputs and the variable they want to predict as the target variable.

ARIMA modeling is available to implement on various platforms and tools like Python, R, SPSS as there are inbuilt library functions available in these platforms so it becomes easy to use the tool support. For the purpose of cross verification and in order to explore the data in two different ways. The analysis is being done in 2 ways by first executing the Box-Jenkins Model on the data using simple libraries available on various tools and secondly by developing a neural network which has take in account the external factors related to chip shortage and predict the data with underlying ARIMA model.

X. Architecture

Below is the architecture diagram for the project which outline the processes that has be carried out in order to execute the research project.

As there is no separate algorithm for the research project, hence the architecture shown below works as a defacto process flow. As per the below figure, the dataset has be treated for cleaning in order to have useful data to train, and further it has be undergoing model analysis. Once the forecasting has been done, the error handling has be followed and it has be followed by visualizations of the results. The detailed technical architecture is mentioned below.

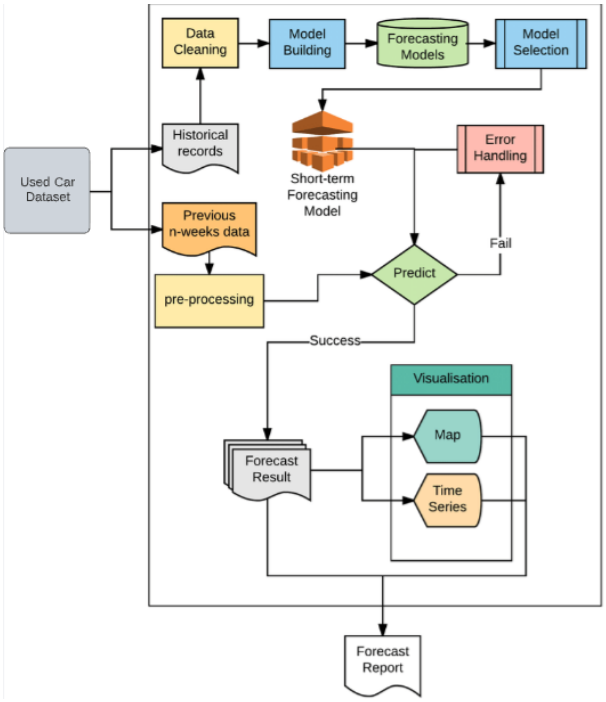

Fig. 4. Project Architecture

As the dataset has been sourced from a public platform hence there are numerous ways for data to be duplicated or mislabeled when merging multiple data sources. Even if the data is right, outcomes and algorithms are untrustworthy if the data is erroneous. However, data cleaning procedure is important in order to process the data in an efficient way before analysing it. The data cleaning procedure was completed by removing the erroneous and redundant entries in the dataset.

Along with handling erroneous data, there was data found in the dataset under structural error domain having N/A as many cell values, thus the handling of this type of data was done by considering them under same categorization of null values as 0. Categorical Imputation was done in categories such as Condition of the car, Transmission where the data was not available and kept as “Other”. The missing values in the date column was handled by scraping the data from the website through Python script.

There were many outliers observed in the dataset specifically in the category of Speedometer Prices of the used car in the dataset thus to refine it, I have used exponential smoothing to remove the outliers and smoothen the data. Also, there were many cell which were missing data in the dataset, the handling of such empty cells was done by inputting observations done manually. For example, the price of one used car was not available in the dataset, then according to

the Manufacturing Year, Kilometers driven, and conditions specified by the owner; an explicit price value was given after comparing with similar car listed on the platform. Once the outlier handling was completed, the normalization of the data was performed wherever required by ranging specific columns of data in the range of 0 to 1. The next step involved fitting the pre-processed data in the ARIMA model. As the analysis is being done in two ways i.e through inbuilt functions on data analysis platforms on R Python thus the dataset was divided into training and testing set with equal distribution of 80percentage and 20 percentage respectively. Whereas, the inbuilt library such as pmdarima in Python and arima() function in R available has be leveraged for predicting the data through ARIMA model.

Once the results are obtained from both the methods, the results has be compared with the actual results through Root Mean Square Error. Since I was not able to find the efficiency of RMSE over MAE or vice versa in the literature domain specifically for the Time series analysis, but I wish to employ RMSE method for result validation as it is more inclined towards telling that how concentrated the data is around the best fit in the dataset.

Once the results are obtained, they has be visualised by using the libraires like matplotlib

$$\sqrt{\frac{\sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2}{n}}$$

Fig. 5. RMSE

XI. Dataset

The dataset which is the Used cars listing on Craigslist in the US has been publicly made available from Kaggle. This dataset contains 426681 rows which contains information about the listed used cars such as Model No., Manufacturer, VIN No., Kilometers driven, Engine Condition, Ask Price, Location, No. of Cylinders, Listed date etc. Although there was few of the important data missing in the dataset which was scraped using available Python web scrapping scripts from Cargurus Inventory. Along with this dataset, information related to chip shortage was also scrapped through web scrapping tool as no direct data set capturing this information was publicly available. The web scrapping was done on various websites and the information related to chip shortages, manufacturing, and procurement was captured for the period of January,2020- October,2020 as the shortage was significant in this time period.

XII. Results and Analysis

The procedure for performing the prediction was carried on onto two platforms, namely R Python. With the help of inbuilt libraries available in R Python, the data set was preprocessed and further the model was developed by training and testing the data with bifurcation ratio of 80:20 for training testing respectively. The below snapshot is taken by plotting the values from the dataset which shows that for how many days the car has been listed on the website. This can be interpreted in a way that no. of days unsold is inversely proportional to the total sale of the used cars. In simple words, is the number of total unsold days is higher than on collective basis we can say that the sale of used cars was low for a particular period.

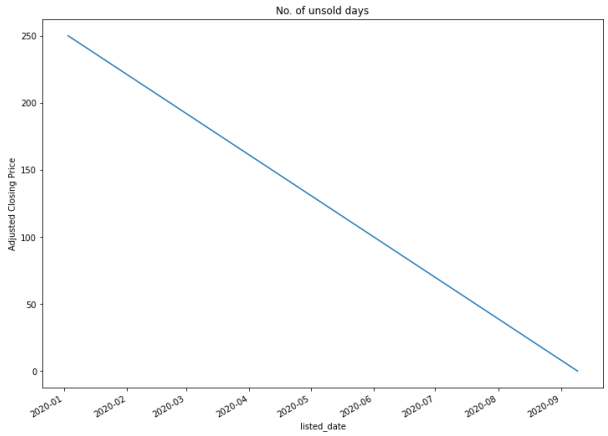

Fig. 6. Data Distribution

The dataset exhibits a distinct trend, as can be shown. This indicates that the time series is not stationary and has require differencing (at least a difference order of 1) to make it stationary. Let's take a quick glance at the time series' autocorrelation plot. Pandas have this built-in as well. The autocorrelation for a large number of lags in the time series is plotted in the example below. Further, in order to find the density distribution of the no of unsold days on the market and the snap shot is shown as below.

Running the example as shown in below snap shot, it was observed that the first 15-20 lags have a positive correlation, which is possibly significant for the first 0 lags. A suitable starting point for the model's AR parameter is 0.

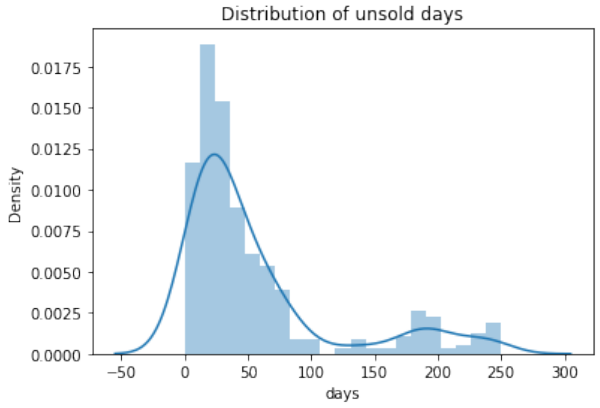

Fig. 7. Density distribution

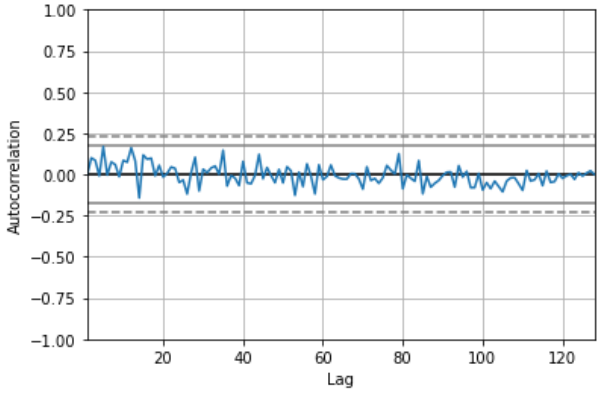

Fig. 8. AutoCorrelation

Since the autocorrelation values are in acceptable range, the next step was to fit the model in the dataset. The fitting of the model is done by using the inbuilt ARIMA function in Python and to understand the summary of the model fitting procedure, the below snap shot shows the summary of SARIMAX results which stands for SARIMAX stands for Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors.

We're trying to anticipate the dependent variable, which is the daysonmarket. Constant beta is one of the independent variables. In our equation above, the error term is sigma2 or epsilon. We can proceed to the following stage and assess the term importance after ensuring that we did not make any simple errors with our model.

The snap shot summarizes the fitting of ARIMA(0,1,0) model. The basic information is pretty self-explanatory:

- Dep. Variable – What we're trying to predict.
- Model – The type of model we're using. AR, MA, ARIMA.
- Date – The date we ran the model

SARIMAX Results

Dep. Variable:	daysonmarket	No. Observations:	128
Model:	ARIMA(0, 1, 0)	Log Likelihood	-770.482
Date:	Tue, 19 Apr 2022	AIC	1542.963
Time:	16:03:10	BIC	1545.807
Sample:	0	HQIC	1544.119
	- 128		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
sigma2	1.089e+04	1411.964	7.713	0.000	8123.605	1.37e+04

Ljung-Box (L1) (Q):	38.79	Jarque-Bera (JB):	0.09
Prob(Q):	0.00	Prob(JB):	0.96
Heteroskedasticity (H):	0.57	Skew:	-0.02
Prob(H) (two-sided):	0.07	Kurtosis:	2.87

Fig. 9. Model Fitting Summary

- Time – The time the model finished
- Sample – The range of the data
- No. Observations – The number of observations

The log-likelihood function finds the distribution that fits the sampled data the best. While useful, AIC and BIC penalise the model for complexity, which aids in the parsimoniousness of our ARIMA model. The Akaike Information Criterion (AIC) is used to determine the linear regression model's strength. Because adding more parameters always increases the maximum likelihood value, the AIC penalises a model for adding them. Like the AIC, the Bayesian Information Criterion (BIC) penalises a model for its complexity, but it additionally considers the number of rows in the data.

The Hannan-Quinn Information Criterion (HQIC), like the AIC and BIC, is a model selection criterion that isn't used as frequently in practise.

In conclusion, the ARIMA fitting model summary data shows that how much close is the model selected to the dataset under the forecasting subject. As per the observed ranges, the values obtained in the SARIMAX results are in the satisfactory range which suggested that ARIMA(0,1,0) is a suitable model in order to predict the data.

Also, the above mentioned model ARIMA(0,1,0) is a model selected after parsing the data through many iterations. This means there were many models of ARIMA(p,d,q) with different values of p=lags, d=degree of differencing, and q=moving average model's order. Also, the dataset being large in size with approximately 400K rows so the computation time for each row was extremely high. Since the model ARIMA(0,1,0) is yielding favorable results hence the next step was involved in forecasting the sales of Used cars in terms of "No of Days unsold on market".

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a commonly used metric for comparing predicted and observed values (sample or population values) by a model or estimator. The square root of the second sample moment of the discrepancies between anticipated and observed values, or the quadratic mean of these differences, is represented by the RMSD. When the

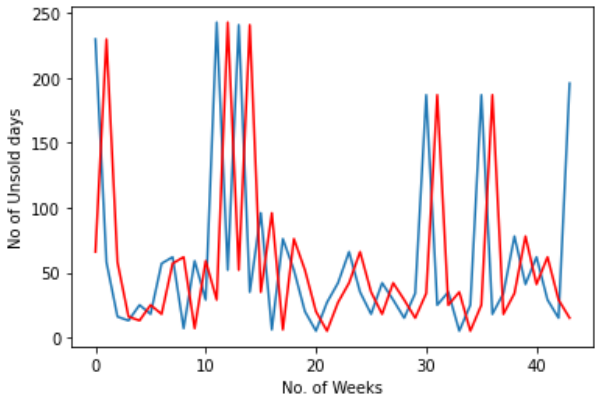

Fig. 10. Prediction Results

```
predicted=42.000000, expected=29.000000
predicted=29.000000, expected=15.000000
predicted=15.000000, expected=34.000000
predicted=34.000000, expected=187.000000
predicted=187.000000, expected=25.000000
predicted=25.000000, expected=35.000000
predicted=35.000000, expected=5.000000
predicted=5.000000, expected=25.000000
predicted=25.000000, expected=187.000000
predicted=187.000000, expected=18.000000
predicted=18.000000, expected=34.000000
predicted=34.000000, expected=78.000000
predicted=78.000000, expected=41.000000
predicted=41.000000, expected=62.000000
predicted=62.000000, expected=29.000000
predicted=29.000000, expected=15.000000
predicted=15.000000, expected=196.000000
Test RMSE: 94.550
```

Fig. 11. RMSE Validation

computations are performed over the data sample that was used for estimate, these deviations are referred to as residuals, and when they are computed out-of-sample, they are referred to as errors (or prediction errors). The RMSD is used to combine the magnitudes of prediction errors for numerous data points into a single predictive power measure. Because it is scale-dependent, RMSD is used to evaluate predicting errors of different models for a specific dataset rather than between datasets.

Because the errors are squared before being averaged, the RMSE gives huge mistakes a lot of weight. As a result, the RMSE is most beneficial when huge errors are most unwelcome.

Here, X-axis is the number of unsold days on the market and Y-axis is the number of weeks in the observant year 2020. The red line depicted is the forecasted values of the sales of the used cars in the US. As it can be seen that the ARIMA(0,1,0) has predicted the values with 1.5% of error from the original values. The prediction accuracy of this model is 98% for ARIMA in Python which was attained after multiple iterations of computing the forecasted values for the dataset. Whereas the RMSE of the testing dataset is 94.55 as per below figure.

We can say from the above figure that in the initial months of the year, the shortage of the semiconductor chips was not that much prevalent up till week 15(Mid of April,2020) but beyond that the semiconductor chips production and procurement increased drastically which is evident in the below figure. The concurrent impact was observed on the manufacturing of the new cars in the automobile units and production of new cars was either halted or stopped by major automobile manufacturing plants in the US and around the world. Thus from the figure above we can say that during mid 2020 due to chip shortage the demand of used cars increased which ultimately decreased their total days in the market as it can be seen in the week range 18-40 (April,2020-October,2020). The forecast predicted by the ARIMA(0,1,0) is in line with the actual scenario which happened during the Covid-19 pandemic in the year 2020 and there was a huge demand of used cars in sale as a result of shortage of semiconductor chips in the global market[1].

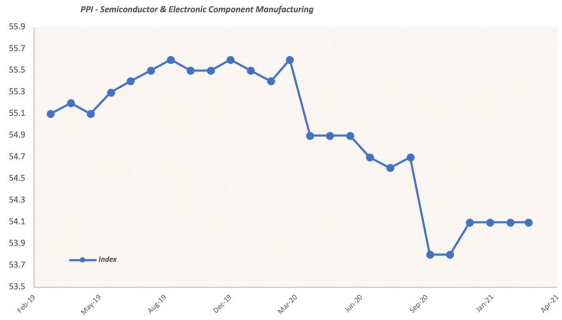

Fig. 12. Global Chip Shortage trend

Also, another insight of the result is plotted in below snap shot where it shows the region of the US where the cars are listed from, the density is distributed on the basis of year of manufacturing for a particular car. With the help of this insights business decision can beb taken by further tuning the input parameters of the model.

XIII. ARIMA Analysis with R

The second portion as mentioned in Stage-4 was carried out in R with the same dataset and this section has be discussing the results obtained by performing the data analysis in R. The below snap shot shows the distribution of the additive trend series analysis. The components of an additive time series are added together to form the time series. If you have a rising trend, you'll see that the peaks and troughs are generally the same magnitude across the time

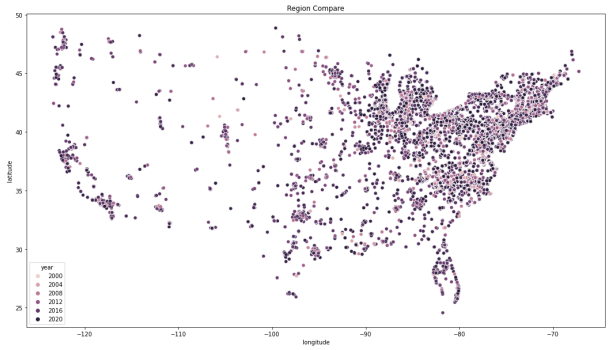

Fig. 13. Region Wise trend

series. This is common in indexed time series, where the absolute value grows while the changes remain relative. Trend- It shows how the No. of days unsold in the market is behaving Seasonality- It shows how the trend is changing based on days, months, year, week. Observed – The trend which is generally observed by exploring the dataset. We must compare the residuals after decomposing our data. We don't need to do anything particularly complicated because we're only trying to classify the time series – a large purpose of this exercise is to create a short function that might be used to make an initial classification in a batch processing environment, so the simpler the better.We'll see if the residuals still include information about how much correlation exists between data points. This is the Auto-Correlation Factor (ACF), which may be calculated using a function.

As per the below snap shot, we can see that in the ACF graph has positive trend in 0 lag which implies our p value in ARIMA(p,d,q) model should be 0.

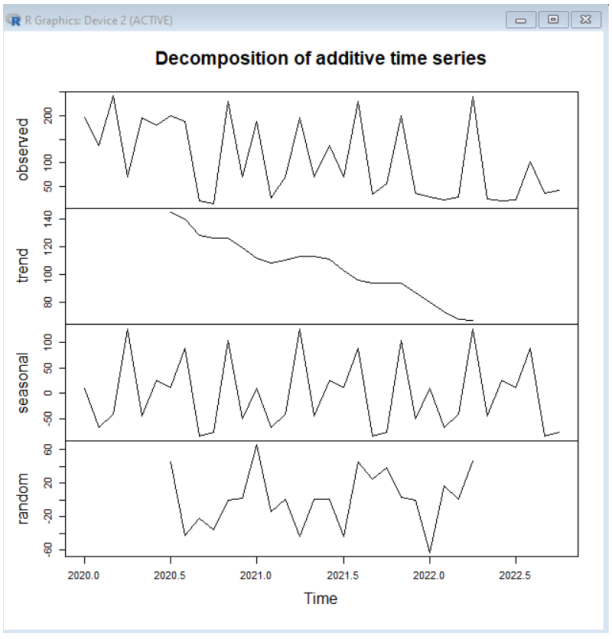

Fig. 14. Decomposition of Additive Time Series

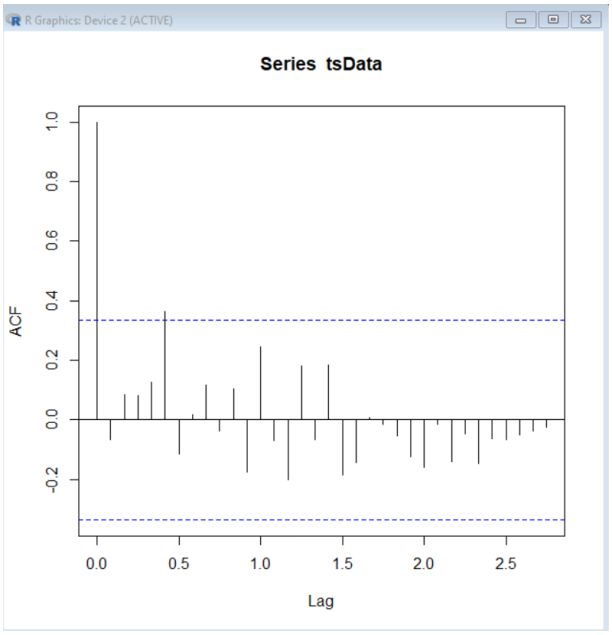

Fig. 15. AutoCorrelation Function

After fitting the ARIMA model in R with the help of ARIMA function under FitAR package, we concluded that the ARIMA(0,1,0) model is suitable for the dataset and the result was predicted after fitting this model. The predicted result is as shown below where we can see that the model predicted the future values using the available value for the No. of unsold days in the market. As it is visible that predicted results are in line with the ARIMA model in Python where it was shown that till the month of October,2020 the sales of used cars were increasing. In the below figure as well it is evident that the ARIMA(0,1,0) model has predicted the sales of used cars till 2023 by forecasting the total No. of Unsold days on the market. As the total No. of Unsold days on market for a single car decreases the sales increases which is forecasted in the below figure.The accuracy was measured to be 98.71%.

XIV. Efficiency Analysis

In order to have more features, the dataset which is downloaded from Kaggle platform was modified and now it consists the data of used cars listed in US. The metrics of the dataset was 104876R X 61C which is a huge dataset to compute and perform the analysis. Thus, for the purpose of the project 426681 rows were selected in order to decrease the computing complexity of the dataset and also to utilize the available resources in its limit. The results obtained for such a large data took enormous amount of time and resources in terms of RAM and Memory.

The below snap shot taken during the operation of ARIMA function in Python in order to predict the results suggests that entire computation power of CPU(100%), around

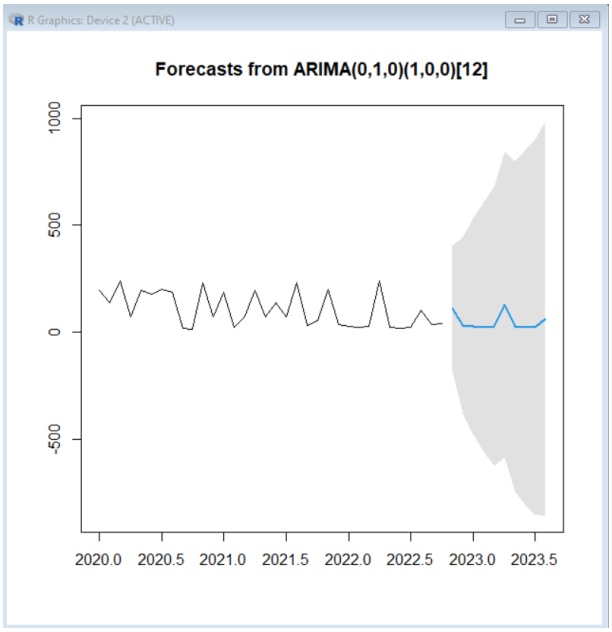

Fig. 16. Forecasted results in R

92% of available memory was being employed in order to perform the predictions. Since, this snap shot was taken when the first 3 minutes of the operation elapsed, it took around 4 hours to compute the results and plot it in Graphical manner.

The code written for this project was as optimised as it can be in order to execute this operations, but it is generally observed that the libraries such as statsmodels, sklearn are used for performing complex statistical operations which generally requires high amount of computation power.

In order to execute the code with such a huge dataset, it is advisable to utilise the Virtual Machines provided by Google Collab which is a cloud repository used to execute Machine Learning and Data Intensive operations.

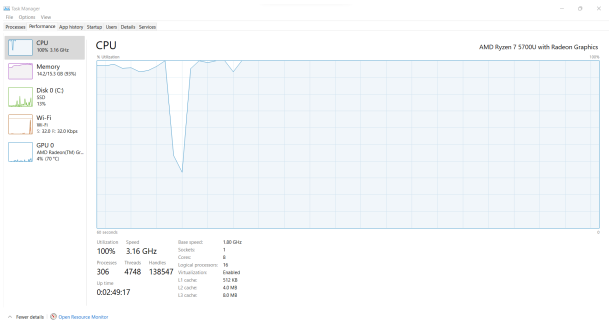

Fig. 17. Computation Metrices

XV. Complexity Analysis

The time and space complexity for the Python code is measured in this section. The complexity of an algorithm is a measure of the number of resources required to run the algorithm to completion for a given input. Typically, an algorithm's resources are either space or time, which are two different sorts of resources complexities. The overall amount of time required is the time complexity to complete its execution via an algorithm. The time it takes for an algorithm to run depending on the type of data, may differ for the same input sizes. The procedure for measuring the time and space complexity was done manually without any usage of tool or library available in public domain. Since the ARIMA inbuilt function is used for this project hence as per available literature in the public domain, the computational complexity of ARIMA function is by default O(n).

As an inbuilt function is leveraged in this project hence the complexity doesn't change drastically because the available function is used without changing its original structure and parameters, so we can say that the complexity remains O(n).

XVI. Related Approaches Comparison

Two approaches namely, ARIMA implementation in Python and ARIMA implementation in R were employed in order to predict the sales of used cars for the US. The problem domain relates to time series analysis and there are different approaches available in order to predict the data having time series trend. Various Machine Learning approaches like Long Short Term Model (LSTM) which is a sort of recurrent neural network that can learn how elements in a sequence are ordered. It's frequently used to tackle difficulties involving time series forecasting.

Autoregressive (AR): An autoregressive (AR) model forecasts future behaviour using data from the past. When there is a correlation between the values in a time series and the values that before and succeed them, it is useful for predicting. There is a variety of literature published by scholars where they are using combination of multiple models in order to predict time series. Yan Wang et al published a paper in 2019 where in they predicted the prices of 10 stocks of Chinese Stock Exchange by developing a hybrid model of ARIMA along with XGBoost technique, with this approach they were able to produce an accuracy of 86.87%. Thus it shows that when ARIMA is combined with different Deep Learning models, there is an increase in the cumulative accuracy.

The primary reason of this project is to publish a project which works about appending external factors in order to predict the results like shortage of semi-conductor chips in this case. Hence, due to this reason only individual analysis on platforms i.e Python and R is carried out. Another approach is proposed by Huang et al as single hidden layer feedforward neural networks known as extreme learning machines in 2004. They have a lightning rapid learning rate, excellent

generalisation abilities, and universal approximation abilities. The weights and biases of the hidden layer nodes are chosen at random, which saves time in backpropagation hyper tuning. A Moore Penrose inverse operation can be used to determine the weights of the output layer analytically[23].

Extreme learning machines have a significant advantage over regular back-propagated neural networks in terms of training time. The weights and biases of the hidden layers in typical feedforward neural network models are iteratively modified using slow gradient-based learning methods like backpropagation. ELMs are computationally significantly faster than typical deep neural network models since the hidden layer weights and biases are randomly assigned. In addition to their computational economy, ELMs have been demonstrated to outperform standard Neural Network techniques in terms of efficiency and generalisation performance on a variety of benchmark issues. For this paper, simple neural

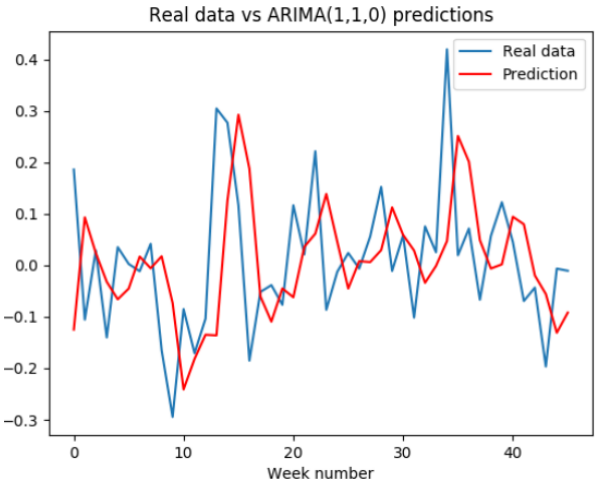

Fig. 18. Neural Network Approach

network approach with 2 layers was proposed but during the implementation, the accuracy was not improved due to which the approach can be further improved as a part of future scope of the study. The below snap shot is the output of Neural Network approach where the accuracy was 72% but even after epoch=20 the accuracy was not improved hence it is not considered in the analysis section but it is viable to invest more resources to further improve this accuracy.

XVII. Conclusion

The project has achieved its goal and is now able to forecast all future market behaviour and trends based on the current circumstance. This prediction ARIMA model makes it simple to determine the likelihood of a used cars sales, and to take appropriate action. This model is ready to assist in the prediction of sales and production following the impact of the corona virus outbreak. Furthermore, if something similar to this or any other pandemic occurs again, this model has recommend how to save the business value by delivering the necessary information through its projections. The model aids an individual and a businesses financial planning by forecasting future results. The automobile business has been in a precarious position; they must listen to their customers in order to move forward cautiously. Further, with inclusion of various other parameter and hyper parameters, the model can be more strengthen and it can be employed in predicting various other features related to car sales. This model is also beneficial in studying the domain of how external factors impact current scenarios and how various features can be predicted by combining this ARIMA model with certain Neural network and Deep Learning methodologies to yield fruitful results for individuals and businesses

References

[1] https://www.consumerreports.org/buying-a-car/when-to-buy-a-used-car-a6584238157/

[2] https://www.factorywarrantylist.com/car-sales-by-country.html

[3] https://www.ctvnews.ca/autos/u-s-car-sales-plunge-as-chip-shortages-choke-off-supply-1.5607459

[4] Z. Mo and H. Tao, "A Model of Oil Price Forecasting Based on Autoregressive and Moving Average", 2016 International Conference on Robots Intelligent System (ICRIS), 2016. Available: 10.1109/icris.2016.4.

[5] . Wu, T. Yuan, K. Qie and J. Luo, "Geographical distribution of extreme deep and intense convective storms on Earth", Atmospheric Research, vol. 235, p. 104789, 2020. Available: 10.1016/j.atmosres.2019.104789

[6] . Wachter, T. Widmer and A. Klein, "Predicting Automotive Sales using Pre-Purchase Online Search Data", Proceedings of the 2019 Federated Conference on Computer Science and Information Systems, 2019. Available: 10.15439/2019f239.

[7] N. Rietmann, B. Hugler and T. Lieven, "Forecasting the trajectory ¨ of electric vehicle sales and the consequences for worldwide CO2 emissions", Journal of Cleaner Production, vol. 261, p. 121038, 2020. Available: 10.1016/j.jclepro.2020.121038

[8] J. Jiang, "Traffic demand Forecast of online car-hailing based on BP Neural Network", E3S Web of Conferences, vol. 214, p. 02035, 2020. Available: 10.1051/e3sconf/202021402035.

[9] A. Falih Zuhdi, Aripriharta, A. Rakhmat Taufani, A. Firmansah and G. Jiun Horng, "Car Sales Prediction System Based on Fuzzy Time Series and Adaptive Neuro Fuzzy Inference System", 2020 International Computer Symposium (ICS), 2020. Available: 10.1109/ics51289.2020.00061

[10] "Data mining for identifying trends in markets", 2022

[11] Digital Transformation of monitoring customer behaviour in the car sales", 2022.

[12] Y. Wang and Y. Guo, "Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost", China Communications, vol. 17, no. 3, pp. 205-221, 2020. Available: 10.23919/jcc.2020.03.017

[13] N. et al, "Predicting the Selling Price of Cars Using Business Intelligence with the Feed-forward Backpropagation Algorithms", 2022

[14] C. Valluri, S. Raju and V. Patil, "Customer determinants of used auto loan churn: comparing predictive performance using machine learning techniques", Journal of Marketing Analytics, 2021. Available: 10.1057/s41270-021-00135-6.

[15] C. et al, "redicting and Modelling Costs and its impact on Profitability using Artificial Neural Networks: A Case in Automotive Passenger Car Industry in India", 2022.

[16] , "Sales forecast of manufacturing companies during covid19", 202

[17]]Y. et al, "Sales Forecast Method for Products with No Historical Data", 2021.

[18] P. et al, "ARIMA-Based Forecasting of Total COVID-19 Cases in the USA and India", 2021.

[19] .M Hani'Ah, I. K. Putri and A. R. T. H. Ririd, "Parameter Optimization Of Holt – Winters Exponential Smoothing Using Golden Section Method for Predicting Indonesian Car Sales," 2021 International Conference on Electrical and Information Technology (IEIT), 2021, pp. 21-26, doi: 10.1109/IEIT53149.2021.9587379.
[20] https://www.pinterest.com/pin/341288477984434743/
[21] https://www.researchgate.net/figure/Flow-chart-of-ARIMA-model
[22] "What is the role of ICT in security sector?," Safety is our everything, Sep. 14, 2021.https://gsi-alarme-securite.com/protectionof-information/what-is-the-role-of-ict-in-security-sector.html (accessed Apr 07, 2022).
[23] https://www.analyticsvidhya.com/blog/2021/12/time-series-forecasting-with-extreme-learning-machines/: :text=There%20are%20two%20main%20approaches,%2C%20ARIMAX%2C%20and%20SARIMAX%20methods..

Comparative Analysis of Literature Review

Paper title	Summary	Pros	Cons
Zhuo Mo and Han Tao presented paper on oil prediction using Auto regressive and Moving average	Overall, the paper was insightful and provided important aspect on the application of ARIMA model in predicting numerical or analytical parameters for various business and activities which are linked with time-series phenomenon where prices or sales are affected over a certain period due to changes in the deciding parameters of prices.	Also, the authors explained the whole procedure of how they have computed the prices right from constructing the time sequence scatter plot till prediction of the prices through the developed combination of ARIMA and RBF Neural network	To replicate this model in further research and applications, an individual won't be able to apply it with the correct information because there lacks the application method or usage method of this research; whereas in the introduction section it is mentioned by the authors that this model will be helpful in predicting crude oil prices with time-series analysis
Angel Pinzon Hassan et al presented paper on air pollutants prediction using ARIMA	This paper the authors used the ARIMA modeling algorithm in order to predict the air quality of a city in Columbia. Due to its novel concept, this type of research papers are useful because it helps in planning important strategies to combat the daily life problems	The authors in this paper actually identified the relationships between the pollutants and the places where they are found, the frequency of time period in which they are found, the observation stations where they are found in considerable manner, and how the direction of wind is affecting the pollutants spread across the city of Bogota	Few things that could have been improved are that the authors could have included more graphical representation of SPSS tool as the reader can come to know about the tool more, along with the distribution maps of the pollutants for all the observation stations for more intuitive information. Also, empirical information in the paper was missing In the paper which is required when the authors are working with data but it was not present. Further, the authors could also have used clustering techniques in order to cluster the pollutants prediction data based on observation stations
Phillip Wachter et al presented paper on car sales using pre-purchase data	The authors in this paper have predicted the car sales tied with the Google Search keywords. This type of	The depth of paper was evident while reading it, all the information were presented in a nuance manner and useful for the readers. Also, the use of	Whereas, considering the seasonality of the data with the keywords, Holt Winter's exponential smoothing can also be

	prediction is dependent on factors such as accuracy, homogeneity of keywords, correlation of keywords with the data source. Authors of this paper have presented in a very detailed manner how they have predicted the car sales by considering various factors.	graphs, process flow diagrams, bar-charts of results made it more convenient for the reader to understand the concept	used which takes prior time steps, seasonality, and trends in account while predicting the data
Nele Retman et al presented paper on EV sales prediction and CO2 cosequences	The paper was a considerably large and had an in-depth analysis because of its highly content rich research domain and problem statement.	As a reader, it was easy for me to grasp the content of the paper as it had sufficient amount of diagrams and charts which were explaining the logistic growth model, amount of CO2 gas being emitted by countries currently and sales of EVs per country. Also, the graphs related to Logistic growth function mentioned about the saturation and inflection point which denoted the start of peak sales growth for EVs from the year 2023-2031. I liked the description where the countries were clustered according to their probable EVs sales and CO2 emissions till the year of 2035 as it provides a general idea of the whole paper in a gist.	The cons of paper were less as it was presented in a perfect manner, but if they authors could have included more information about the logistic growth model and the empirical or statistical tools used for the analysis purpose then it would have been beneficial for the readers.
Yang Zipei and Zhang Chong presented paper on Automobile forecast with Social Media data	Automotive companies can rely on insights of this type of data which is generally considered as implicit or indirect data generated from various sources.	The most interesting part in the research paper was regarding segregation of different sentiments i.e positive and negative and labelling them as 1 and -1 in the data analysis part as it shows that authors have treated the sentiments in the way they should be treated and so that it doesn't affect the validity of the results.	The only disadvantage observed in the paper was that there is not enough information on the empirical formulas and maths related to the ML models deployed which could have been an added advantage
Abid Falih Zubdi et al presented paper on car prediction using Fuzzy Logic System	The research paper published talks about predicting the car sales using the methodology of Fuzzy Logic by	The paper was more inclined towards statistical aspect of applying Fuzzy logic in predicting car sales, which is a good aspect whenever the	In the conclusion the authors have just mentioned about the accuracies of algorithms such as FTS and ANFIS

	applying Sturges formula. The paper talks about applying FTS and ANFIS algorithms which falls under Fuzzy logic method to determine the car sales.	authors are looking for a research paper with strong mathematical background.	which does not portray anything about the research subject as the authors have not clearly specified that on what basis the prediction is being made
Adele Puscasiu et al Presented paper on Data mining software for car sales	The paper published by the author talks about a adaptable and user friendly traditional approach of predicting car sales which is related to data mining in general and calculating the prices of used cars in the Romanian Market.	It's a decent paper which talks about an unique aspect of Big Data application in the domain of predicting car sales with the help of web interface. The explanation provided by the authors is also effectively written as they have mentioned in detailed about how they have used MS Access for maintaining a database by web scrapping the Romanian website for car listing, developing backend module in C# and front-end module in Angular JS.	the paper is lacking information about time in which the data is mined as it is useful in this kind of applications. Also, they have not considered any limitations related to data being mined for this paper.
Ivan Rados et al presented paper on Digital transformation monitoring in car sales	The authors have effectively and uniquely presented the flow of the paper as it becomes easy to understand the different aspects of the paper while reading it. The authors have predicted the sales of the cars for a large automobile manufactures by the method of logistic regression and they have provided some extremely useful insights which are unique and useful.	Insightful paper with some beneficial real world marketing strategies through regression	The authors could have mentioned which company was being analysed in the paper and what models were being analysed as it could have provided more transparency in the research.
Yan Wang et al presented paper on stock prediction with ARIMA and XGBoost	The authors have employed a hybrid model by combining ARIMA and XGBoost ML models for predicting the stocks prices of 10 stocks in the period of 2015-2018	The interesting part of the paper was when the tabular forms of ARIMA and XGBoost results were depicted for all the 10 stocks as it gives a bird eye view of the whole research paper.	The authors have not mentioned the training and testing dataset, also the basic information of the dataset is missing as from where it is collected I.e Stockexchange, what are the parameters present in

			the dataset, also which tools have been used by them in order to obtain the results.
Nur Oktavin Idris presented paper on Feedforward back propogation algorithm for prediction car sales	This is an unique paper which is published by authors to predict the car prices of models which are never made by BMW based on the features and specifications of the already developed cars by BMW with the help of Business Intelligence and regression models.	The paper explained in a detailed manner how the features which are already developed by BMW in its cars are impacting the prices and how they can be considered while predicting the prices of future cars	NA
Chejarla Venkat Narayana et al presented car sales prediction approach through BI	The amount of knowledge obtained from the paper can be utilized in future research as its an easy and content driven paper aiming at providing maximum knowledge about the ML models for predicting the used car prices in India.	The authors have tried to include everything in the paper and it is also evident on reading as they have performed literature review of previously published papers and presented in a tabular format, performed data pre-processing and analysis of dataset at feature level, performed outlier and error analysis and omitted the extra features presented in the dataset, performed categorical feature encoding.	NA
C Samuel Joseph et al presented paper on predicting and modelling cost and impact on profitability of car sales in India	The paper published by the authors talks about the impact of cost and revenue on the profitability of a car sale with the auto major companies of India. The authors have used Radial Basis function(RBF) to achieve results.	There is information only about the methodology of RBF and how it is being utilized in the research	The research is performed assuming normal scenario but impact of various external factors is not considered which can be considered as outliers or impactful factors depending on the situation, thus it becomes more of a static study
Prabhat Sharma et al presented paper on Sales forecast during Covid-19 in India	The issue observed in the previous literature review of paper has been successfully addressed in this paper where the authors are considering external	Small and content rich research paper.	The authors could have included more information about the Random Forest Regression and how they have achieved the results like code or pseudocode, tools etc.

	factors such as Covid-19 and sales of cars In India by top 5 auto manufacturers during the period of 2019-2020 through Random Forest regression.		
Yun Dai et al presented paper on car prediction with no historical data	The paper published by the authors is a cumulative analysis of the sales of new products, specifically car in this case without any historical data of the product.	The paper was decent and content driven.	Although, the authors have considered the features of new car but they have not discussed the impacts the new features are having on the car sale which could have been beneficial.
Pinar Chihan presented paper on COVID-19 cases in India and USA through ARIMA	The authors have used the dataset available from open source platform and analysed it by analysing it under ARIMA model	Addressed the problem with good accuracy.	The dataset could have been more broader and verified.
Mamluatul Hani'ah et al presented paper on Indonesian car sales using Holt-Winters exponential smoothing technique	Overall, it was an informative paper which can be used in predicting the sales of cars for other companies and countries as well, but can the method proposed handle external factors affecting the sales is not a answer with conviction.	The paper has large amount of graphical representation which is easy to understand and also emough information on the theoretical portion of Holt-Winters exponential smoothing as well	Export and Whole sales data could be dependent on various factprs which is not considered.

Appendix

Platforms	R & Python
IDE	Spyder(Python) & RGUI(R)
Libraries - Python	Pandas, Numpy, Scikit learn, Matplotlib, Math, Statsmodels, ARIMA, Seaborn
Libraries - R	fUnitRoots, lmtest, forecast, FitAR
Language	R & Python
Dataset Link	https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset