National College of Ireland

Defect Detection and Price Estimation of Used Cars Using Deep Learning with Regression Models

MSc Research Project
MSc in Data Analytics

Kishore Lakshmanan
Student ID: x20253583

School of Computing
National College of Ireland

Supervisor: Mr. Hicham Rifai

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Kishore Lakshmanan
Student ID:	x20253583
Programme:	MSc in Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Mr. Hicham Rifai
Submission Due Date:	15/08/2022
Project Title:	Defect Detection and Price Estimation of Used Cars Using Deep Learning with Regression Models
Word Count:	7210
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	12th August 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Defect Detection and Price Estimation of Used Cars Using Deep Learning with Regression Models

Kishore Lakshmanan
x20253583

Abstract

Through various internet platforms, buying goods is becoming simple and faster in the digital age. Despite the advantages it offers consumers, there are a range of disadvantages as well, including the spread of fraudulent items and the fraud which is carried out by online vendors and sellers. These concerns can be addressed with the use of deep learning with regression technologies. Applying deep learning and machine learning methods, this research examined the identification and categorization of damage along with price estimation of secondhand cars. The most significant method for determining the faults is through deep learning models. In this domain of detecting damage, considerable research has been conducted. First, using deep learning techniques like CNN and MobileNet models, study will assess how well defects can be found in various vehicle images. After that, car price estimator is performed by applying machine learning techniques (Linear Regression and Decision Tree). For our research, the efficiency of the decision tree algorithm and convolutional neural network is excellent i.e. accuracy of 94.8 % and 95.5 % achieved respectively. Finally, the combined output of both outputs assists the consumer choose the appropriate car for purchasing. This research will also assist manufacturing firms automate inspection, government organizations categorize the severity of vehicle damage post accidents, and loan management and insurance claims.

Keywords- Computer Vision, Deep Learning, Machine Learning, CNN, MobileNet, Linear Regression and Decision Tree Algorithm.

1 Introduction

Defect detection is one of the main concept in the machine learning and computer vision technologies. It is a broad area of research. The objective is to determine the type and severity of the product defects from images. Although Vehicles are becoming more autonomous, the administrative procedures related to insurance, renting, selling and buying cars can be automated using machine vision potential for advancement in lending activities and challenges regarding vehicle insurance. Particularly, Vehicle defect detection can be used for insurance company instead of manual inspection for vehicle to assess the state of vehicle after accidents and minimize the mistakes. Mistakes in insurance disputes lead to financial loss. The difference between an optimal settlement and the realistic settlement of a claim is referred as claims leakage. The primary cause of claim leakage is failures in existing processes, both manual and automated. Similarly, organizations who can provide auto loans, particularly for used cars, have issues in deciding the

level of the damage to the vehicle. Inspection and validations are widely used especially, which takes a lot of time in order to assess the damages and determine whether to make claims or loans. For instance, one of our studies makes use of a dataset made up of images sampled from Figure 1 Figure 1 Waqas et al. (2020). This would facilitate defect detection in a classified approach.

Figure 1: Different kind of car damage images

Another application for defect detection is to detect anomalies for identifying problems pertaining to the manufacturing and automotive industries. The majority of researches employed deep learning algorithms for identification and classification of vehicle damage.

1.1 Research Background and Motivation

1.1.1 Research Based on Defect Detection Model

The majority of research conducted vehicle defect detection is based on deep learning methods. CNN-based vehicle defect detection technique is general and respond to several types of car-damage images. The algorithm produced effective predictive outcomes in a number of incidents. Regardless of the advantages of the illumination, the damaged area of various cars, or a scenario with a high exposure, the fitting impact is stronger and also has the best durability, Zhang et al. (2020). The combined distributions of local binary pattern (LBP) and rotation-invariant measure of the local variability (VAR) technique that was devised was effective in identifying a variety of defects. This technology may be used by automobiles to analyze, discover, and categorize a variety of problems in their vehicles. The samples used in the research had an average classification accuracy of 96.2 percent, Kamani et al. (2011).

1.1.2 Research Based on Price Evaluation Model

Every business knows that making intelligent and difficult decisions is an essential responsibility. Making the incorrect choice might result in significant loss or possibly the closure of a corporation. The retail industries, namely the used automobile service company, in an aim to offer a creative solution to this issue. The recommended study shows that predictive analytical models will be an important addition to company, particularly

for assisting the judgment process. Utilizing statistical tools and technology, companies uses predictive analysis to examine previous data in order to provide fresh insights and make future plans appropriately. Narayana et al. (2021) primary goal is to develop a prediction model, or fair pricing mechanism, to forecast the selling price of a car based on data including its model, age, fuel systems, vendor type, gearbox type, and mileage. By estimating the selling price of a used automobile based on its qualities, this article lowers the risk for both the seller and the customer. The suggested model accomplishes this objective by combining statistical regression methods including linear, decision tree, and random forest regressions with machine learning algorithms.

1.2 Research Question

RQ: To what extent can deep learning model detect the level of damage in a car and how it can be used to estimate price of the car as a parameter in machine learning model?

By include this research question in the report, the public is supplied with the appropriate platform to purchase and sell secondhand products without any hurdles and also this application will be useful for the insurance company and to the government from analyzing severity of the vehicles involved in the accidents.

1.3 Contribution

The main objective of this research work is to build a model that can both predict vehicle price using machine learning models and detection of defects using a variety of deep learning approaches. Also experiment analysis of the model to check how efficient the developed application can find damage and estimate price of used cars for helping sellers and buyers in online portal by using various machine learning and deep learning models. This project's major goal is to assist common people in educating the automotive industry about the efficient buying and selling of their products. This provides a thorough analysis of machine learning or deep learning techniques employed.

The remaining parts of the paper is divided into the different sections. The research of all past relevant works is described in section 2 followed by section 3 which shows data cleaning and preprocessing for model development, section 4 which illustrates the design architecture of the model, section 5 which includes the model building process, section 6 discussed the various experiments and performance analysis of the model and section 7 which outlines the conclusion of our paper and explains how it will continue to be used in further work.

2 Literature Review

Depending on the area of research, several items are used for investigation in the subject of defect detection. Different mechanisms, including deep learning, image processing, and machine learning, played a more important part in each and every defect's diagnosis. This section addressed and provided a summary of all existing papers, organized into the following subsections: 3.1 Defect Detection in Various Fields Using Deep Learning Models, 3.2 Defect Detection in Vehicle and Automobile Industries Using Different Algorithms, 3.3 Review of Price Evaluation Techniques in Vehicle Sales and 3.4 Review of Price Prediction Techniques in Vehicle Sales .

2.1 Defect Detection in Various Fields Using Deep Learning Models

Deep Learning methodology is the important techniques for detecting defects in every field. The related work done in our research includes various fields such as manufacturing, railway, surface, textile and fabric industries. Out of all, railway and manufacturing industry is the main research for many of the publishers. Here, safety to the public is the essential thing in these two area. Hence, fault analysis is the crucial segment to avoid accidents and any malfunction of parts due to inspection failure in the machinates. Attard et al. (2019) has proposed a model using deep learning model to automating the process to identify cracks on the concrete surface. Using Mask R-CNN, cracks on cement floor may be located and their associated mask can be formed to help extracting relevant attributes that are helpful for inspection. By automating crack diagnosis, lowering the amount of time required to do this operation, cutting expenditures, and raising employee safety, such a technology could help in eliminating the negatives of human inspection. They prepared a corresponding labels collection of masks on pictures from a subset of a standardized crack dataset in order to train Mask R-CNN for crack identification. A accuracy value of 93.94 percent and a recall value of 77.5 percent were obtained during tests using the trained model.

Le et al. (2021) developed a deep learning (DL) based model for crack fracture detection on concrete surfaces. A Deep Learning Convolutional Neural Network served as the foundation for the created image categorization model (CNN). The existing literature was mined for a collection of 40,000 photos of concrete surfaces (with and without cracks) to train and verify the CNN model. Several factors, including lighting and surface quality, were addressed while examining the hard surfaces (i.e., exposed, plastering, and paint). The accuracy of the constructed model was examined using a variety of error assessment criteria, including accuracy, precision, recall, specificity, and F1-score. The accuracy was determined to be 99.7 percent for the training dataset, which composed half of the database. As a result, the constructed CNN model may be regarded as legitimate because it successfully classifies fractures using the testing data. It is also proven that the proposed DL-based model was reliable and effective since it can account for a variety of situations on concrete surfaces.

Maeda et al. (2018) research offers three ways to address issues in identifying road damages. Firstly, the first massive road damage data gathering is produced. It consists of 9,053 road damage photographs which have been taken with a smartphone fitted on a car, and it encompasses 15,435 incidents of road surface damage. After training the fault detection model with the created data set using cutting-edge object detection methods with convolutional neural networks, they compared the accuracy and execution performance on both using a GPU system and a smartphone. By using the recommended object classification approach, they show that the damage type can be accurately divided into eight categories. They produced recalls and precision's better than 71 percent and 77 percent utilizing MobileNet and Inception V2 with a close intervals of 1.5 seconds on a smartphone, respectively.

For the analysis of visual data for the discovery of rail surface flaws, Faghih-Roohi et al. (2016) suggested a deep convolutional neural networks approach. The photos are selected from several hours of automatic video recordings. Due to the overwhelming volume of data, it is not possible to manually review the photos and find rail surface faults. As a result, automated rail fault detection can assist in lowering time and expenditures

even while ensuring the safety of rail transit. The extraction of relevant attributes for the diagnosis of rail surface cracks, however, it is a challenging and non-trivial process, which poses a significant obstacle. So, they suggest that a practical method for feature learning be the use of CNN. Recently, several related fields have successfully used DCNN in their analysis. The author contrast the outcomes of several network designs with various dimensions and activation functions.

A multi-feature fusion network is developed by Tao Ye (2021) to simultaneously identify three common mechanical component failures. The feature extraction module is employed to enrich the underlying semantic features of the deep feature extraction, and three modules are embedded in the network to increase the detecting effects of tiny mechanical equipment failure. A compression block is inserted into the network to improve the channel characteristics, and a multi-branch expanded convolution module is used to obtain the fusion characteristics of multi-scale visual field. For learning on the Tensorflow platform, Intel 1080Ti Processors were employed in all studies. The experimental study showed that all three of the network's modules assist in detecting mechanical faults in railroad cargo trucks, and that Encoding / decoding network's performance is stronger to that of the most of other extensively used SSD-based one-stage object detectors. MFF-net can reach 0.8872 mAP and 33 frames per second with an input image size of 300 x 300 pixels.

Wang et al. (2020) did the detailed analysis of the most recent ML developments across a range of AM fields. ML may be used in the DfAM to produce improved topological designs and novel, high-performance nanocomposites. Modern ML algorithms may assist in AM manufacturing by optimizing processing parameters, investigating particle distribution, and doing in-process deficiency monitoring. In terms of AM operation, ML may help professionals with pre-manufacturing planning, product quality evaluation, and control. Additionally, given data breaches may occur with the use of ML algorithms, there has been growing concerns regarding data security in AM. It finishes with a section that highlights the key research results from the literature and offers opinions on a few chosen, promising ML applications in AM research and development.

The following table 2.1 describes the related paper who worked on particular application and their accuracy of the developed model.

Table 1: Literature Review: Comparison of Various Defect Detection Approaches

Authors	Techniques	Results
Attard et al (2019)	Mask RCNN	Accuracy of 93.94% achieved
Le et al (2021)	CNN	Accuracy of 99.7% achieved
Maeda et al (2018)	MobileNet and Inception V2	Accuracy of 71% and 77% achieved respectively
Tao Ye (2021)	Multi-feature Fusion Network	MFF-net can reach 0.8872 mAP and 33 frames per second

Jing et al. (2020) developed a method for fault identification using an enhanced

YOLOv3 model. They have included enhanced YOLOv3 to the research to improve efficiency and the frequency at which fabric product defects are discovered. First, dimensionality grouping of the target frames is carried out using YOLOv3 by merging the anomaly based size with the k-means approach to determine the number and size of preceding frames. Second, to better apply the YOLO detection layer to defect identification in gray textile and lattice cotton, low-level features are merged with confidential information, and feature maps of different size are added to the detection layer. The improved network model delivers a lower than 5-percent error detection rate for both gray and checked fabric. According to experiments, the proposed method is more effective than YOLOv3 at identifying and marking fabric issues as well as minimizing mistake detection. It must be improved how well the method performs in real time, which is the downside.

2.2 Defect Detection in Vehicle and Automobile Industries Using Different Algorithms

Khanal et al. (2021) One of the key application areas to improve quality at a lower cost is quality automotive inspection, which may be accomplished with the use of computer vision technology. Automatic identification of vehicle components is crucial for both quality checks and automatic assembly of auto parts. Khanal et al. (2021) produced a deep neural network framework built on Convolution layer is used to classify automobile parts. Dataset is trained and evaluated using a VGG16 deep learning architecture with an output layer made comprised of 8 fully connected neurons. The collection includes 20,439 RGB pictures of eight exterior and interior automobile components that were taken from the front. The dataset was originally split into training and testing segments, and then the training dataset was further segmented into training and validation segments. The individual components identification accuracy was 93.75 percent on average and 97.2 percent on average. Automotive production, model checking, and car assessment systems are only a few of the applications that the classification of car parts supports.

Park et al. (2020) also research in the same field and his work named as detecting defects and dents on the surface of the vehicles using Region-Based Convolutional Neural Networks. To achieve the visual quality of a new car, the detection and isolation of dents on a vehicle body that happen during manufacturing are necessary. This work were using an illumination that can reveal dents on an images by reflecting the Mach waves in order to improve feature extraction. Heat-maps were generated using the classification scores derived from the Region Based Convolutional Neural Networks to localize dents after the model was trained using the image highlighted by the waves. The performance of the suggested R-CNN was statistically assessed in this research when it was applied to the assessment of dents on the exterior of a vehicle body. The proposed approach had a 98 percent classification performance for typical and abnormal patches, and its MAE was 13.7 pixels, demonstrating a minor difference between the sites of the estimated dents and the actual damage.

Around the world, numerous car accidents happen every single minute, causing traffic congestion that interferes with regular travel. By using a car damage detection segmentation technique based on an enhanced mask regional convolutional neural network and transfer learning, the accident concerns were attempted to be resolved by Zhang et al. (2020) research methodology. The self-made dataset demonstrated better precision and accuracy for the mask regional convolutional neural networks by collecting a damaged image and creating a dataset labeled as test and training set. Since the collection is tiny

and contains few images, data segmentation has not been conducted.

In the automobile sector, loan handling and insurance clearance are extremely important. In order to categorize the automotive damage into three categories—high destruction, moderate harm, and less harm. Authors. Waqas et al. (2020) constructed a model using techniques of deep learning and a mobilenet model. Furthermore, a hybrid strategy was used to provide the model for the defect classifier the processed images in order to limit user uploads of fraudulent images to the internet. They employed research and the moire technique in this study to find fake photos. For fault classification and morie's detection, they respectively attained accuracy of 95 and 100 percent.

Zhao et al. (2017) study proposes a manufacturer, large data statistical modeling problem diagnosis technique for battery systems in electric cars. A battery pack's unexpected fluctuations in cell terminal voltage may be identified and quantified as a probability using a machine learning model and the 3-level multi-level screening method. In order to provide a more thorough model for battery system damage detection, this study employed the neural network technique to integrate the findings of fault and defect detection with large data statistical regulation. Researchers can identify design defects in battery storage and provide suggestions for the flow of designing by studying the anomalies that are lying beneath the surface. The outcomes of the calculations are also checked using the clustering outlier diagnostic method and the local outlier factors (LOF) algorithm. A related study between the statistical diagnostic findings and the genuine vehicle is provided to further confirm the efficacy of the diagnosis approach.

Automatic inspection is crucial for industrial quality management. Kamani et al. (2011) presents an industrial solution for automated vehicle body painting assessment. In order to detect various types of issues, the new strategy analyzes images progressively taken from the automotive body. Observed defects are categorized into several types of defects using a Bayesian classifier. Self-driving cars and autonomous vehicles (AVs), which are increasingly becoming a regular part of life, are used by a number of people today. Giant firms utilize two AV software systems, namely Baidu Apollo and Autoware, according to Garcia et al. (2020). 499 AV bugs and 16,851 commits were investigated from this program, and the bugs were then categorized. This led to the identification of 16 key results, which were then used to provide a framework for future research on locating and fixing software bugs. Both in terms of detection and classification, the average accuracy was 96 percent.

Liqun et al. (2020) has contributed to the repository of knowledge about deep learning-based defect detection for auto parts. Machine vision technologies have developed from human classification of old methods to computer vision techniques for identifying flaws in car parts. In this study, a VGG16 network structure model is created using deep learning to identify automobile part issues initially. The accuracy rate is now 94.36 percent. Updates have also been made to the VGG16 network structure model. By using the Inceptionv3 module, the model's span is increased based on its depth, and its accuracy rating of 95.29 leads to a more accurate recognition of the image. While both strategies are more effective than the previous method, the traditional HOG+SVM classification methodology only achieves an accuracy of 93.88 percentage.

The following table 2.2 describes the related paper who worked on particular application and their accuracy of the developed model.

Table 2: Comparison of Previous Research Work on Defect Detection

Authors	Techniques	Results
Garcia, J., (2020)	Improved Mask RCNN & Mask RCNN	For minority injury, it gives 99 and 96 percent respectively
Kamani P (2011)	joint distribution of local binary pattern (LBP) and rotation invariant measure of the local variance (VAR)	Average classification accuracy was 96.2%
Jing, J.,(2020)	Yolo, YoloV2, YoloV3	Accuracy of 91, 95 and 97 percentage respectively
Liqun (2020)	VGG Network	Accuracy of 96.68%
Waqas (2020)	MobileNet and Moire effect	Accuracy of 95 and 99 percentage respectively

2.3 Review of Price Evaluation Techniques in Vehicle Sales

Salim and Abu (2021) occupied with researching the s curve model to predict the cost of the car. In this study, his model will be shown as a non-linear alternative for forecasting used car prices. On the foundation of a dynamically S-shaped Classifier, this research study constructs an S-curve pricing model (SMF). A well-known website provided information on actual used car costs. Comparisons between linear and nonlinear regression are available. While the residue of an S-curve model is nearer to that of linear regression, the error is smaller than that of linear regression. It is anticipated that their methodology would, overall, provide a more precise and accurate projection of used car pricing in Malaysia.

Yin et al. (2021) created a system for used vehicle prediction using Mean Encoding and PCA-based Deep Factorization Machine. They provide a deep learning method that integrates Mean Encoding, Principal Component Analysis, and Deep FM to assess used car pricing. In in order to reduce the number of attributes, the technique first preprocesses data composed of used cars using Mean Encode and PCA. The process uses DeepFM to determine the lowest and maximum values of the resampled attributes, and the output of the network is utilized to determine the used-car cost. The proposed method may more effectively produce greater performance in the evaluation of used-car price when compared to LR, FM, and DNN.

Sun et al. (2017) suggested a system to assess the cost of secondhand vehicles using the BP neural network theory. Its objective is to provide a process for assessing used car pricing in order to determine the best price for the car. The more recent study was determined by the analysis performed by the research, which employs a vast quantity of automobile data and widely dispersed car information to assess costing for each type of vehicle using an optimized neural network model. In such systems, whether the vendor and buyer may have more effective market information depends on how accurately used car pricing evaluations are made.

Bilen (2021) has been developed by study on how prices are determined in the retail market for used cars. In spite of holding the car's actual price, in-depth qualities, the environment, and a significant number of other effects constant, his paper concludes that a negative or positive deviation of a vehicle's retail value either increases or decreases its everyday sale constraints significantly more than the car's resulting value. The predicted assumption values are substantially higher for the cars that have been exposed to more,

and they are sized similarly to the influence of the usual value. Inflation trends and the effects of the flexible pricing method on automotive distributors are also explored.

2.4 Review of Price Prediction Techniques in Vehicle Sales

Jin (2021) developed a forecasting model for the fair values of used automobiles based on a number of characteristics, such as the mileage, year of manufacture, fuel consumption, gearbox, road tax, fuel usage, and engine capacity. In the used automobile market, this approach can help vendors, purchasers, and automakers. Based on the data that users submit, it can eventually generate a quite accurate cost estimate. Data science and machine learning are used during the model development process. The utilized dataset was obtained from used automobile listings. To attain the highest level of accuracy, a number of regression techniques were used in the study, including linear regression, support vector regression and random forest regression. This study used data visualization to fully understand the dataset before beginning the model-building process. To train the regression, the dataset was split up and changed, ensuring the model's performance. R-square was obtained in order to assess each regression's efficiency. Random forest has the highest R-square, coming in at 0.90416 out of all regression analysis in this research.

Hankar et al. (2022) employed a variety of regression approaches based on supervised machine learning to forecast the resale value of old automobiles given a variety of parameters like mileage, fuel type, financial power, brand, model, and the year the car was produced. A high R-squared score and a small root mean square error were displayed by the gradient boosting in each of the evaluated models. Özçalıcı (2017) use decision trees to forecast the second-hand automobile sales price. To choose the most pertinent characteristics, genetic algorithm is utilized. 252645 posters are scanned for this study in order to serve this objective. There are 139 features offered for each advertising. With the selection of few characteristics, several models are explored using genetic algorithms. In the out-of-sample experiment, percent 65,67 had the best forecasting performance. The suggested model can be utilized as a decision-support tool for people who work in the used automobile industry.

Monburinon et al. (2018) performed a comparison of the effectiveness of regression using supervised machine learning models. Data about the used automobile market obtained from a German e-commerce website is utilized to train each model. The highest performance is therefore provided by gradient enhanced regression trees, with a mean absolute error of 3D 0.28. followed by the multiple linear regression and random forest modeling, respectively, with MSE values of 3D 0.35 and 0.55. Gajera et al. (2021) predicting old car prices using machine learning. They constructed a statistical model that will be able to forecast the cost of a used automobile using supervised machine learning techniques like linear-regression, KNN, Random Forest, XG boost, and Decision tree. It will be helped by a collection of characteristics and past customer data in this situation. Accuracy is the parameter referred to find the best model.

3 Research Methodology

This section will describe the each stages of the methodology is used for defect detection and evaluation of the used car prices. The best research will follow the data mining processes in the methodology to give certain shape of the data for efficient result and analysis. In this project, CRISP-DM (Cross Industry Standard Process for Data Mining)

is used for the data mining and data analysis methodology of defect detection and price estimation of the car. The below figures 2 illustrates the working flow of the CRISPDM which includes the overall stages involved in the methodology. Understanding the business requirement is the first stage to focus in this section which is the main objective of the research. To create and assess the results for a better outcome, this study comprises two stages of research. Finding the fault in a used vehicle is the initial stage. To build the perfect model, the second stage is estimating the valuation of the vehicle based on its properties, such as the car model, production year, etc. Then, the outcomes of both the stages provides the user with reliable information to purchase the vehicles. Following the other stages involved are collection of data from the trusted source, data preprocessing, transformation of data, data augmentation and normalization of data. Finally, the evaluating the result is taken place based on the data processed in this explanatory data analysis technique.

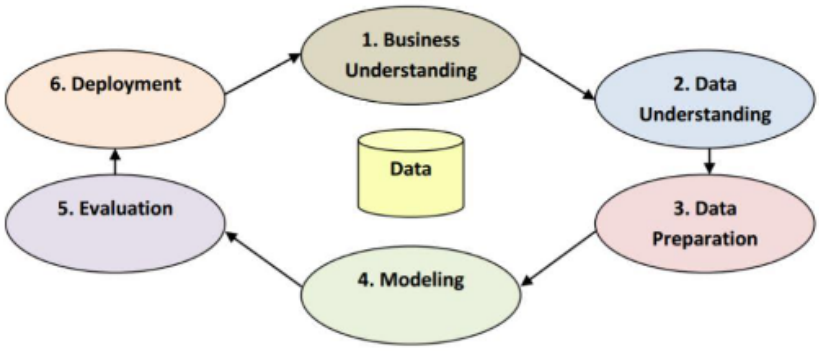

Figure 2: Process involved in CRISP-DM Methodology

3.1 Data collection

In this project, two datasets is used for efficient outcome based on the research objective, and the results will be linked to provide accurate insights. The preliminary phase of the study used a dataset from Kaggle [1] that included more than 2000 images of a mix of severely damaged, partially damaged, and undamaged automobiles. Therefore, using these images will assist in designing our defect detecting section effectively. Dataset from Cardekho which is available in kaggle [2]was applied for the second section. It is the website used as an online marketplace for vehicle sales and purchases. This includes all essential information about the vehicle, including the selling price, the distance travelled, the type of seller, the number of ownership have changed happened for that vehicle etc. This damage assessment and the vehicle status will be helpful in assessing a better outcome.

3.2 Data Preprocessing

Data cleaning or the preprocessing is the important steps to convert the data into more clear and understandable format with enormous insights from the features selected for

[1]https://www.kaggle.com/datasets/anujms/car-damage-detection
[2]https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho

our research. The dataset used here are images and continuous variables available for vehicle information. The images are initially divided into training and validation data which was stored in the variable before implementing the model. All the images were already cleaned and no action required to validate the images. only data augmentation and normalization needs to be done before developing the model which will be explained later on this section.

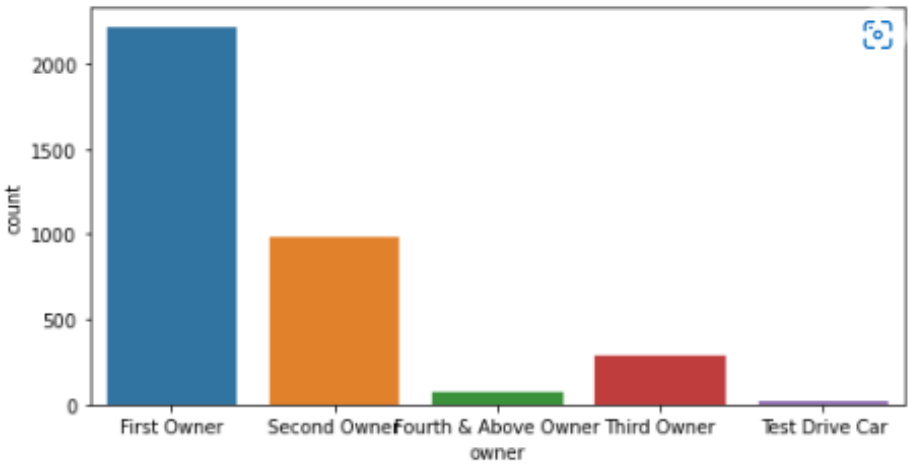

Figure 3: Bar chart includes different owner types

The dataset which includes car details is now checked for any null values, removing unnecessary columns, dropping the duplicate values etc as a data cleaning stages. Then heat map is used to check the missing values of each features in the dataset. Some of the plots are used to indicate the importance of the particular columns and their count of different label in that variables. These plot are demonstrated through the in-build packages such as matplotlib and plotly express from the python module. The above figure 3 shows the owner details in the dataset and below figure 4 explains the number of second hand vehicle available for sales in the specific year.

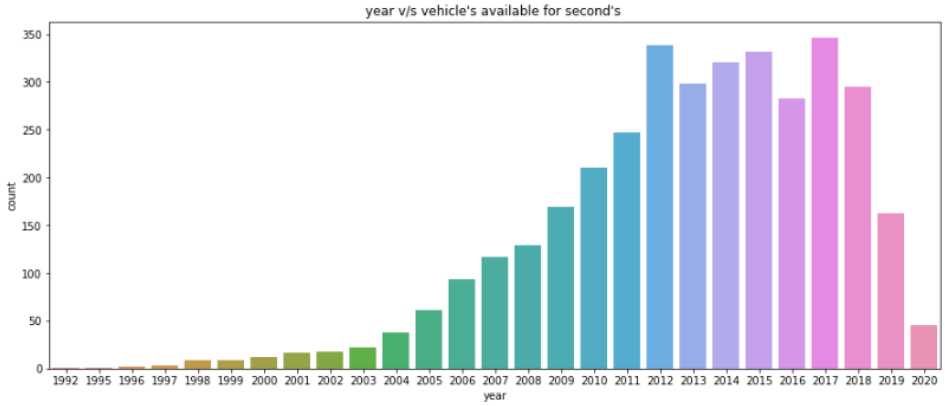

Figure 4: Number of Vehicle Available for Second's in Particular Year

3.3 Data Transformation

Data transformation is used to convert the data from one form to another form to smooth execution of the model. i.e. conversion of alphanumeric variable to numerical one. First dataset contains car images which does not include any transformation to the data. second dataset contains words as a value in some of the columns such as seller type, transmission, owner and fuel which was converted to categorical numbers. Fuel column contains petrol, diesel, LPG, electric and CNG are replaced with 0 to 4 respectively. Seller type column contains dealer, individual and trustmark dealer are replaced with 0 to 2 respectively. Transmission attribute contains manual and automatic are replaced with 0 and 1. The below table 3 shows the data-set after transmission.

Table 3: Values After Data Transmission Process

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner
0	Maruti 800 AC	2007	60000	70000	0	1	0	0
1	Maruti Wagon R LXI Minor	2007	135000	50000	0	1	0	0
2	Hyundai Verna 1.6 SX	2012	600000	100000	1	1	0	0
3	Datsun RediGO T Option	2017	250000	46000	0	1	0	0
4	Honda Amaze VX i-DTEC	2014	450000	141000	1	1	0	1

3.4 Data Augmentation and Normalization

Data Augmentation is the process of adding additional data in order to expand the original data. Also, the regularized function in data augmentation assists in controlling the over-fitting of data [3]. With the help of tensorflow keras library, **ImageDataGenerator** module is used to rescale the image size. Re-scaled image is now set a target size of 150X150, batch size of 12 and the binary class to test and train dataset. This normalized image is inputted to the model building process for standard output without over fitted data.

4 Design Specification

Techniques for defect detection and cost estimation are crucial in the model's development. Combining two dataset that comprise different techniques and make it challenging to construct the model has enhanced the complexity of the research. This model was developed using robust model building tools and methodologies after analyzing all of the previous research papers. The user will receive both the results of the defect detection and an estimate of the price of the vehicle. The user will then benefit from having a complete view of the automobile in one place. Thus, the user makes a decision by determining whether to buy the vehicle or not and, if not, by exploring alternative assessed vehicles. The below figure 5 shows the overall architecture of the car resale model which includes three layers.

[3]https://www.analyticssteps.com/blogs/data-augmentation-techniques-benefits-and-applications

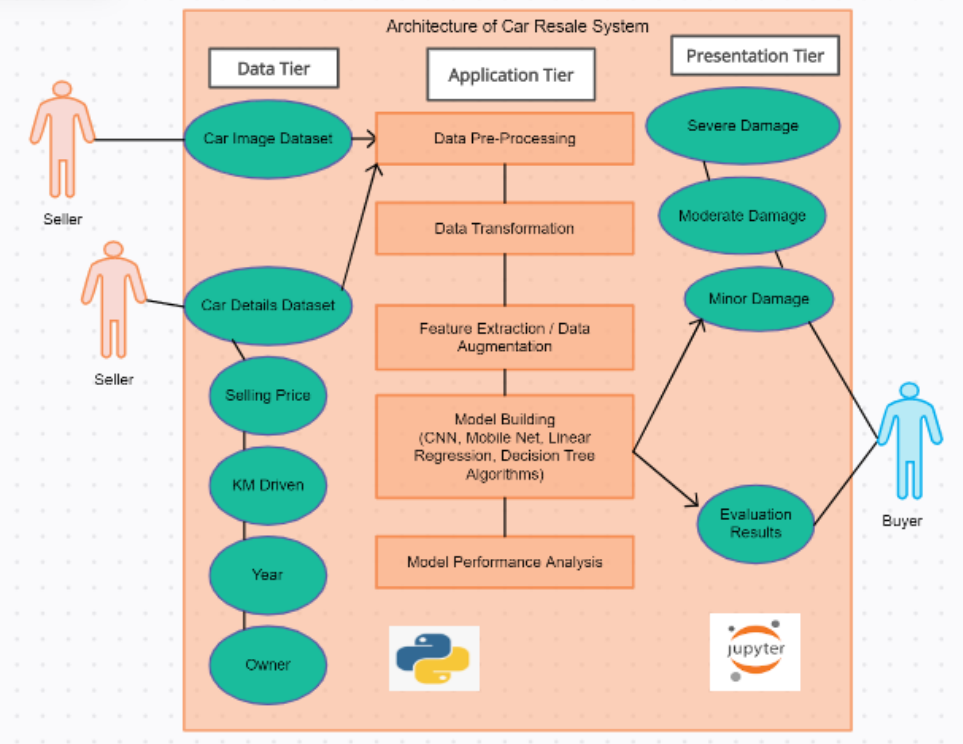

Figure 5: Design Architecture of the Overall Model

- **Data Tier:**
  Data are necessary at this point in order to develop the model. The two datasets were gathered via public source data available online. Images and text are both contained within the two types of datasets. The collected dataset has now been added to the separate Jupyter Notebook with the installed Python module. To perform the actions, it transforms the data into a dataframe as a system readable format.

- **Application Tier:**
  In this tier, data is pre-processed, transformed to build the model for visualization. Pre-processing steps includes the removal of irrelevant columns and dropping the null values etc. The processed data is used for defect detection and price prediction of a car. CNN, MobileNet, Linear Regression and Decision Tree algorithms are used for building the models. All the process is carried out through Jupyter Notebook with all the libraries imported for the particular model to run. Then the results are evaluated through the parameters such as R square value, MSE and accuracy.

- **Presentation Tier:**
  Car defect detection is divided into three segments and the pricing is evaluated independently for the user's view. The Plotly package included with Python is applied to illustrate the model results. It demonstrates a the model's accuracy of all the models involved with other visualizations useful for the research.

5 Implementation

5.1 Introduction

In this section, the model implementation, evaluation and results are demonstrated for the defect detection and price estimation of the used cars. Defect detection is the important part of the project. Since the dataset contains image which needs to be categorized into 3 segments (minor, moderate and severe). The price prediction dataset contains the categorical data which has to be replaced into binary for building the model. After all the pre-processing is done previously, now implementation of all the four models used in the project were executed and also with the process of feature extraction or data augmentation is discussed in this section. The evaluation of the model is performed with the parameters such as accuracy, R square value and MSE. The best fit model is selection based on the performance achieved by each model.

5.2 Deep Learning Models

5.2.1 Convolutional Neural Network

Many challenging computer vision issues have been solved in recent years using CNN. Utilizing a pretrained network is another popular and very efficient Deep Learning technique for small picture datasets. CNN is made up of numerous layers, each of which is made up of neural nodes with distinct functions. It's important to note that nodes within the same layer of the model are not linked to one another, Le et al. (2021). Here, the CNN model is constructed using the **Sequential** library using the "Tensorflow Keras" model. Then, a specific parameter is added to the model to enhance damage detection. The below figure 6 shows the overview of the CNN model.

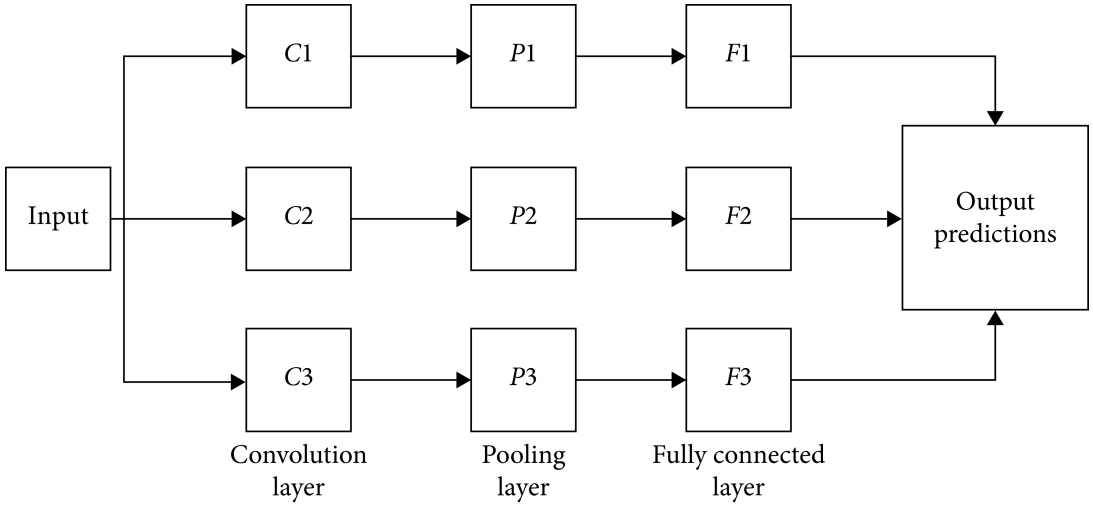

Figure 6: Process Involved in the CNN Model

The first layer is the input layer, which holds the image input data. The features from images are identified using a convolutional layer. The feature map that arises from the filtering of the activation map is then created. Technically, the classifier may receive the convolutional layer's output directly. By compiling the existence of features in patches, the pooling layer offers a method for downsampling the feature maps. A

nonlinear activation function is used to send the convolution layer's findings to the pooling layer. In this case, Maxpool2D is employed with relu activation and a 150x150 input shape. Upto 4 Maxpool layers are used. Then the model is added with flatten function to convert the image array into 1D array. Dense function is also added with the activation function of relu and sigmoid. Fully connected layer refers to the output of the pooling layer after weights have been applied to anticipate the proper labels. The output layer, which includes the problem's anticipated results, comes lastly. The below figure 7 shows the summary of the build CNN model.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 148, 148, 32)      896

 max_pooling2d (MaxPooling2D  (None, 74, 74, 32)       0
 )

 conv2d_1 (Conv2D)           (None, 72, 72, 64)        18496

 max_pooling2d_1 (MaxPooling  (None, 36, 36, 64)       0
 2D)

 conv2d_2 (Conv2D)           (None, 34, 34, 128)       73856

 max_pooling2d_2 (MaxPooling  (None, 17, 17, 128)      0
 2D)

 conv2d_3 (Conv2D)           (None, 15, 15, 128)       147584

 max_pooling2d_3 (MaxPooling  (None, 7, 7, 128)        0
 2D)

 flatten (Flatten)           (None, 6272)              0

 dense (Dense)               (None, 256)               1605888

 dense_1 (Dense)             (None, 1)                 257

=================================================================
Total params: 1,846,977
Trainable params: 1,846,977
Non-trainable params: 0
_________________________________________________________________
```

Figure 7: Summary of the Build CNN Model

5.2.2 MobileNet Model

Specifically designed for mobile vision applications, MobileNet is a simple yet efficient convolutional neural network that needs minimal computational power. Numerous practical benefits, such as object detection, fine categorization, face characteristics, and positioning, make extensive use of MobileNet. When compared to convolutional neural networks, it just employs a relatively small number of parameters. Imagenet is used in this instance to create the model's weights and average pooling. It makes a dense parameter prediction with softmax activation. It uses the 224x224 default picture format. **MobileNet** library from "tensorflow keras application" was used to build this model. Following the execution of certain hyperparameter tuning, the outcome is then anticipated.

5.3 Price Estimation Models

5.3.1 Linear Regression Model

With the use of linear regression, the link between the two variables may be represented by constructing a model that fits the obtained results. The first component is regarded

as an explained variable, whereas the second variable is regarded as a predictor variables. Here, the model relates to the selling price of the car to the kilometer driven using a linear regression model. **LinearRegression** from "scikit-learn" library was used to build the model

5.3.2 Decision Tree Model

The supervised learning category comprises the decision tree approach. Unlike other supervised learning algorithms, the decision tree method can address classification and regression problems simultaneously. Depending on the target variable we pick, there are two different forms of decision tree algorithms. There are two types of variables: continuous and categorical. Due to the fact that our prediction model is based on a continuous variable. The decision tree regressor model is used in this research. **DecisionTreeRegressor** from "scikit-learn" library was used to implement the model.

5.4 Hyperparameter Tuning

The ability to adapt a machine learning model for a particular task or dataset is made possible by the use of hyperparameters, which are points of selection or configuration. For the deep learning models to be better trained, the hyperparameter values were essential. It took several iterations and setups of the model parameters until the optimal model for defect identification of the different automobile images was obtained. The **Callback** library from keras was used in this project for hyperparameter tuning. Model were tunned with 3 parameters such as early stopping, learning rate reduce and model checking. This project makes use of the Adam optimizer along with certain additional parameters, which are shown below 4.

Table 4: Hyperparameter Tuning of the Model

Model	Cost Function	Optimizer	No of Epochs	Batch Size	Lr Reduce	Early Stopping
CNN	Binary Cross Entropy	Adam Optimizer	20	32	Patience=2	Patience=4
Mobile Net	Sparse Categorical Cross Entropy	Adam Optimizer	20	32	Patience=2	Patience=4

6 Evaluation

In this section, evaluation and result of the model is provided for the defect detection and price estimation of used car. **pyplot** library from "matplotlib" were used to plot the

result as graph or some other visualizations. The performance measure of an implemented model was analysed through some of the evaluation metrics described below.

- **Accuracy**
  The model's performance across all classes is often described by its accuracy metric. It is beneficial w hen every class is equally important. It is determined by the total number of estimates divided by the number of predictions that were correct.

- **R Square Value**
  R square is a measure of how much of the variance within the dependent variable can be predicted based on the independent variables. The coefficient of determination is another name for it.

- **Mean Square Error**
  The statistical model's mean squared error calculates the degree of error. It evaluates the mean squared difference between the outcomes that were forecasted and those that were obtained.

6.1 EXPERIMENT 1: Defect Detection Model

6.1.1 Convolutional Neural Networks

CNN model is trained and evaluated using the accuracy parameter. The accuracy of the model determines the model performance. It has the accuracy of 95.5 % with the 20 epochs set. With early stop parameter, it breaks at 10. It is the best fit model of our defect detection model. The below graph shows the Accuracy and Loss graph for training-validation set 8.

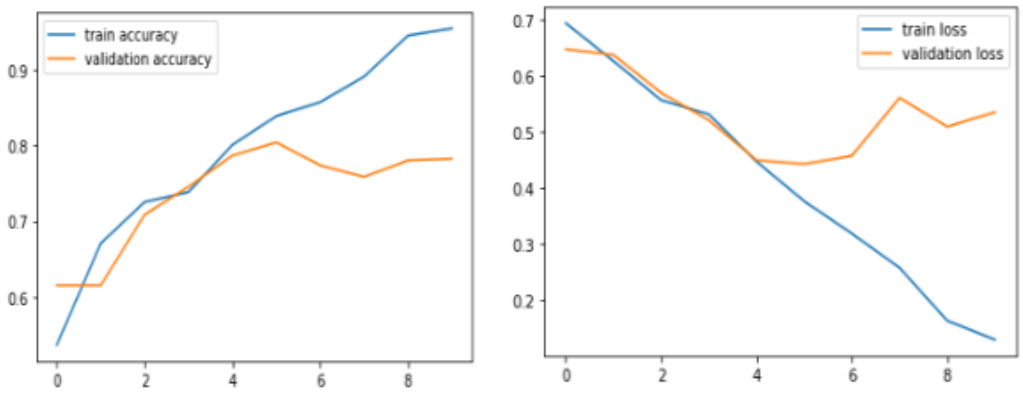

Figure 8: CNN:Accuracy and Loss Graph for Training-Validation Set

6.1.2 MobileNet

MobileNet model also one of the deep learning model to detect the faults. It is also evaluated using the accuracy parameter. The Model performance of MobileNet has 47.8 % accuracy with the 20 epochs settings. With early stop parameter, it breaks at 6th

epoch. The below graph shows the Accuracy and Loss graph for training-validation set 9.

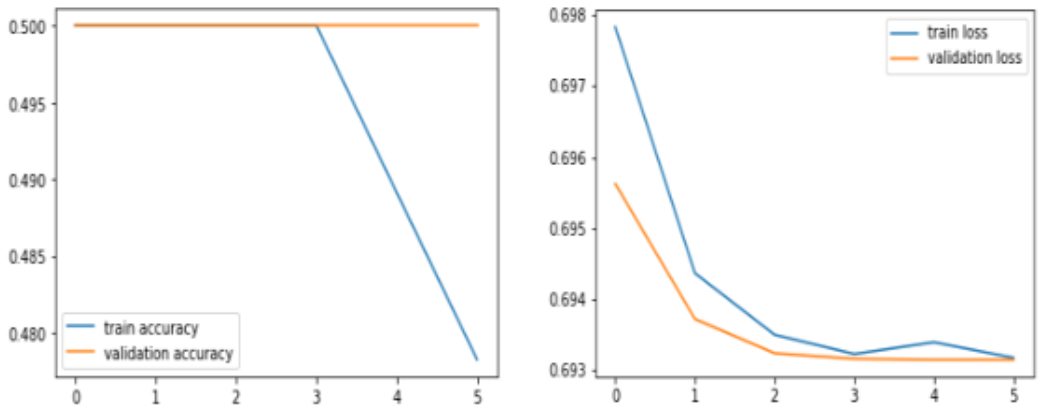

Figure 9: MobileNet:Accuracy and Loss Graph for Training-Validation Set

6.1.3 Classification of defects

In this project, vehicle defect is classified into three categories which includes severe, moderate and minor defects. Each image has assessed as arrays and dimensions which is then predicted through the models acquired after the pre-processing steps. Now the pricted value is checked with the conditions for separating the damage and not damaged cars. The value less than 0.5 is damaged one and more than 0.5 is not damaged vehicle. This parameter was set after several iterations for evaluation of the model. The below figure shows the classification of car damages 10. Once the result gets damaged car result, it then assessed for severity check 11.

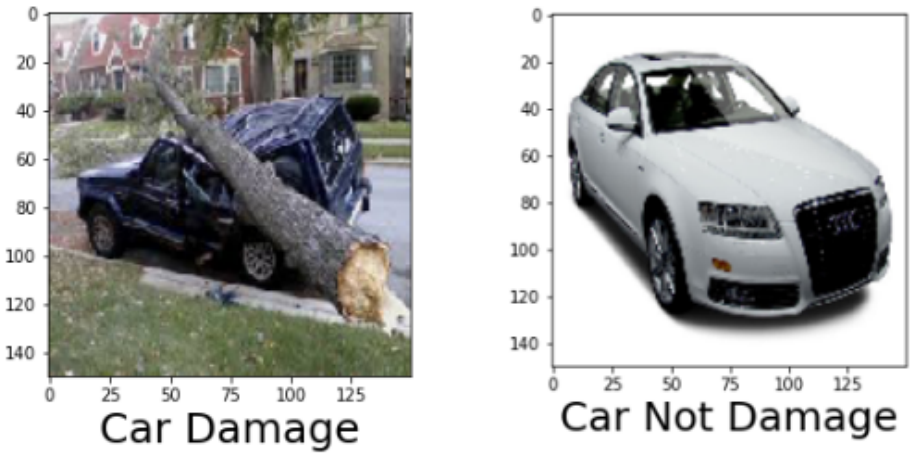

Figure 10: classification of Car Damage Results

```
1/1 [==============================] - 0s 39ms/step
[[0.29291642]]
[0]
Assessment: Severe damage to vehicle
Severity assessment complete.
```

Figure 11: Car Damage Assessment

6.2 EXPERIMENT 2: Price Evaluation Model

6.2.1 Decision Tree

Decision tree regression model is used for evaluating the results as a continuous variable. Since, the output of model is required to provide the estimation of selling value of the secondhand cars. For the prediction model, data is divided into 70 % and 30 % of training and test sets respectively. Here the R square and MSE is considered as performance metrics for calculating the results. It is the best fit model for price estimation of car and having the accuracy of 96.8 %

6.2.2 Linear Regression

Linear Regression is the simplest model to be evaluated. The dataset is split into 70 % and 30 % of training and test sets respectively for predicting the output. Here also the R square and MSE is considered as performance metrics for calculating the results. The below table shows the performance measure of both decision tree and linear regression models 5.

Table 5: Performance Metrics of Price Evaluation Model

Model	R Square Value		MSE Value	
	Train	Test	Train	Test
Linear Regression	0.417	0.37	1.55E+11	1.5E+11
Decision Tree	0.947	0.968	1.38E+10	7.7E+09

6.3 Discussion and Comparison of Implemented Model

In this section, detailed discussion of the evaluated models and comparison between each algorithms. As discussed in the evaluation section, convolutional neural networks and the decision tree algorithm both performed well and produced better accuracy results than the other algorithms used in the defect detection and price estimation models respectively. The comparison of performance analysis between all models used in this research were

plotted in the graph 12. In the defect detection model, it should be better to add certain conditions for checking the originality of the images from the client to avoid inserting fraud data into the model development. The work done by Waqas et al. (2020) focuses on finding the fake images from the dataset and failed to implement the evaluation model for building a tremendous platform for the users to purchase the right automobiles. This model can be efficient when adding the extra feature of image recognition used in Waqas et al. (2020).

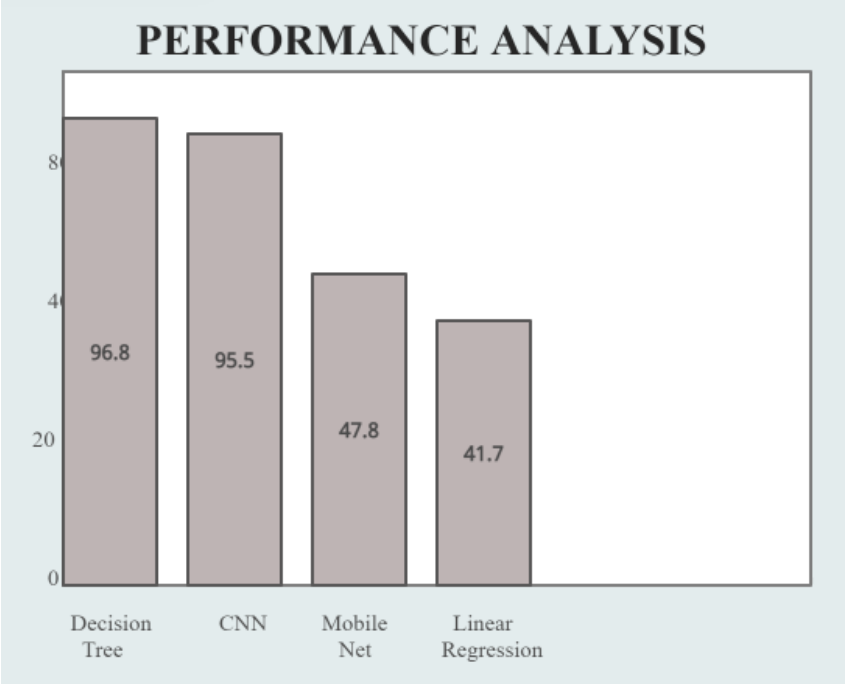

Figure 12: Performance Analysis of the Developed Models

7 Conclusion and Future Work

Considering the customer trust in purchasing of second-hand product, it's crucial to make the existing system perfect to increase the business growth and happiness of the customers. Defect detection has developed an effective application to find the damages in the vehicle and helped several other application as well through the research work. The regression algorithm has produced the best prediction of the cars through the decision tree algorithm. Other side, Convolutional neural network performed better compared to other deep learning models for our research work. The model produced great accuracy than the existing model also has the additional feature of price evaluation techniques.

Although the research was completed successfully, there were a difficulties which solved during the model building process paved way for better execution. The problem includes the unstructured text data which transformed for producing the regression model. The plot function also used to analyse the important of the features before exporting data into the implementation part. The image dataset was small. For making good model, data augmentation was achieved to elaborate the dataset with the help of in-build libraries. In conclusion, better the data exploratory analysis better the model

development. So, the hyperparameter tunning made in this project impacts the model accuracy and also explicit analysis through the graphical output.

Due to the development of technology, there were many software available online to generate fake images. In future, the validation of the fraud images should be taken into consideration and also developing the model with larger dataset to tune the defect detection approaches. It would be excited to see various problems due to the existence of more data. Also, the defect detection of vehicle is being used by various fields and audience such as loan, insurance management, rental companies to verify the car when return back, government agencies for analysing the severity of the damage and also for common people to sell and buy the second-hand vehicles. Taking these things in account, best idea to develop the model with all facilities in single application.

8 Acknowledgement

I would like to appreciate and thank to my mentor Mr. Hicham Rifai for his great effort and allocating valuable time to help me throughout the research which makes me to complete the project within the deadline. It's my pleasure to be part of National College of Ireland and also thank to my family and friends for their constant motivation and support throughout my studies.

References

Attard, L., Debono, C. J., Valentino, G., Di Castro, M., Masi, A. and Scibile, L. (2019). Automatic crack detection using mask r-cnn, *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)* pp. 152–157.

Bilen, M. (2021). Predicting used car prices with heuristic algorithms and creating a new dataset, *Journal of Multidisciplinary Developments* **6**(1).

Faghih-Roohi, S., Hajizadeh, S., Núñez, A., Babuska, R. and De Schutter, B. (2016). Deep convolutional neural networks for detection of rail surface defects, *2016 International Joint Conference on Neural Networks (IJCNN)* pp. 2584–2589.

Gajera, P., Gondaliya, A. and Kavathiya, J. (2021). Old car price prediction with machine learning - irjmets, *International Research Journal of Modernization in Engineering Technology and Science* **03**.
URL: *https://www.irjmets.com/uploadedfiles/paper/volume3/*

Garcia, J., Feng, Y., Shen, J., Almanee, S., Xia, Y. and Chen, Q. A. (2020). A comprehensive study of autonomous vehicle bugs, *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)* pp. 385–396.

Hankar, M., Birjali, M. and Beni-Hssane, A. (2022). Used car price prediction using machine learning: A case study, *2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC)* .

Jin, C. (2021). Price prediction of used cars using machine learning, *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)* pp. 223–230.

Jing, J., Zhuo, D., Zhang, H., Liang, Y. and Zheng, M. (2020). Fabric defect detection using the improved yolov3 model, *Journal of Engineered Fibers and Fabrics* **15**: 1558925020908268.

Kamani, P., Noursadeghi, E., Afshar, A. and Towhidkhah, F. (2011). Automatic paint defect detection and classification of car body, *2011 7th Iranian Conference on Machine Vision and Image Processing* pp. 1–6.

Khanal, S. R., Amorim, E. V. and Filipe, V. (2021). Classification of car parts using deep neural network, *CONTROLO 2020* pp. 582–591.

Le, T.-T., Nguyen, V.-H. and Le, M. V. (2021). Development of deep learning model for the recognition of cracks on concrete surfaces, *Applied Computational Intelligence and Soft Computing* .

Liqun, W., Jiansheng, W. and Dingjin, W. (2020). Research on vehicle parts defect detection based on deep learning, *Journal of Physics: Conference Series* **1437**: 012004.

Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T. and Omata, H. (2018). Road damage detection using deep neural networks with images captured through a smartphone, *Computer-Aided Civil and Infrastructure Engineering* **abs/1801.09454**.

Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S. and Boonpou, P. (2018). Prediction of prices for used car by using regression models, *2018 5th International Conference on Business and Industrial Research (ICBIR)* pp. 115–119.

Narayana, C. V., Likhitha, C. L., Bademiya, S. and Kusumanjali, K. (2021). Machine learning techniques to predict the price of used cars: Predictive analytics in retail business, *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)* pp. 1680–1687.

Park, S. H., Tjolleng, A., Chang, J., Cha, M., Park, J. and Jung, K. (2020). Detecting and localizing dents on vehicle bodies using region-based convolutional neural network, *Applied Sciences* **10**(4).

Salim, F. and Abu, N. (2021). Used car price estimation: Moving from linear regression towards a new s-curve model, *International Journal of Business and Society* **22**: 1174–1187.

Sun, N., Bai, H., Geng, Y. and Shi, H. (2017). Price evaluation model in second-hand car system based on bp neural network theory, *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)* pp. 431–436.

Tao Ye, Zhihao Zhang, X. Z. Y. C. . F. Z. (2021). Fault detection of railway freight cars mechanical components based on multi-feature fusion convolutional neural network, *2021 International Journal of Machine Learning and Cybernetics* p. 1789–1801.

Wang, C., Tan, X., Tor, S. and Lim, C. (2020). Machine learning in additive manufacturing: State-of-the-art and perspectives, *Additive Manufacturing* **36**: 101538.

Waqas, U., Akram, N., Kim, S., Lee, D. and Jeon, J. (2020). Vehicle damage classification and fraudulent image detection including moiré effect using deep learning, *2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)* pp. 1–5.

Yin, X., Liu, L., Xu, X. and Xiao, W. (2021). Used-car price evaluation using mean encoding and pca based deepfm, *2021 China Automation Congress (CAC)* pp. 3578–3582.

Zhang, Q., Chang, X. and Bian, S. B. (2020). Vehicle-damage-detection segmentation algorithm based on improved mask rcnn, *IEEE Access* **8**: 6997–7004.

Zhao, Y., Liu, P., Wang, Z., Zhang, L. and Hong, J. (2017). Fault and defect diagnosis of battery for electric vehicles based on big data analysis methods, *Applied Energy* **207**: 354–362. Transformative Innovations for a Sustainable Future – Part II.

Özçalıcı, M. (2017). Predicting second-hand car sales price using decision trees and genetic algorithms, *Alphanumeric Journal* pp. 103 – 114.