Price Prediction for Used Cars : Car Dheko

A Comparison of Machine Learning Regression Models

ALIYAS THOMAS
MD113

Abstract

Cars of a particular make, model, year, and set of features start out with a price set by the manufacturer. As they age and are resold as used, they are subject to supply-and-demand pricing for their particular set of features, in addition to their unique history. The more this sets them apart from comparable cars, the harder they become to evaluate with traditional methods. Using Machine Learning algorithms to better utilize data on all the less common features of a car can more accurately assess the value of a vehicle. This study compares the performance of Linear Regression, Ridge Regression, Lasso Regression, and Random Forest Regression ML algorithms in predicting the price of used cars. An important qualification of a price prediction tool is that depreciation can be represented to better utilize past data for current price prediction. The study has been conducted with a large public dataset of used cars. The results show that Random Forest Regression demonstrates the highest price prediction performance across all metrics used. It was also able to represent average depreciation much more closely than the other algorithms, at 13.7% predicted annual geometric depreciation for the dataset independent of vehicle age.

Keywords: Machine Learning, Price Prediction, Used Cars, Regression Analysis, Depreciation

Table of Contents

Terminology

Depreciation	the change in net present value over time
ML	Machine Learning
Revaluation	the change in value or price of an asset that is caused by everything other than aging

1 Introduction

Chapter 1 will serve to give the reader an understanding of the background, problem motivation, and the overall purpose and importance of the work in this report. In addition, it will outline specific scientific goals and questions which this research seeks to answer.

1.1 Background and problem motivation

New cars of a particular make, model, and year all have the same retail price, excluding optional features. This price is set by the manufacturer. Used car, however, are subject to supply-and-demand pricing. Further, used cars have additional attributes that factor into the price. These include the condition, milage, and repair history, which sets cars that may have shared a retail price apart.

The used car market is generally divided into two categories, retail and wholesale. The retail price is the higher of the two prices and is what an individual should expect when buying a car at a dealership. The wholesale price is the lower price which dealers will pay. Whether the dealer has sourced the car from a trade-in, auction, or another dealer, this price is considerably lower to ensure that the dealer will make a profit on the vehicle. Prices for peer-to-peer car sales generally lie in-between the retail and wholesale price points. Because there is no "middle-man" in peer-to-peer transactions, there is only a single price point, rather than two. A difficulty in peer-to-peer transactions is for both parties to agree on a fair price. There are many tools which provide an approximation, but do not factor in the particularities of the car into the price. Car markets are to some extent local and therefore location also affects the price. There is therefore a need for a valuation method which can make use of more of the features particular to each car, and extract information from all other previous sales of cars with shared features.

Machine learning (ML) is a subfield of Artificial Intelligence (AI) that works with algorithms and technologies to make useful inferences from data. Machine learning algorithms are well suited to problems entailing large amounts of data which would not be possible to process without such algorithms. ML works algorithmically rather than mathematically and permit a machine to "learn" and adapt its predictions to best fit the data it has trained on. [1]

1.2 Overall aim

The purpose of this thesis is to evaluate several different machine learning models for used car price prediction and draw conclusions about how they behave. This will deepen the knowledge of machine learning applied to car valuations and other similar price prediction problems.

1.3 Problem statement

For the purposes of car valuation, popular guides tend not to use machine learning. Instead, they source data from local sales and average the prices of many similar cars. This method works well if you have a common car with a common set of features. The condition of the car is judged very roughly, typically on a scale of one to three. Cars that are "unusual" are therefore hard to evaluate. Effectively, no inferences are drawn from similar cars but from a different make and model, whereas with machine learning, the entirety of the dataset and its features are used to train the model predictions. Using machine learning is a solution to the problem of utilization of all the data and will assist in utilizing all the features of a car to make valuations.

New cars of a particular make, model, location, and feature selection are identical in condition, function, and price. When new cars are sold for the first time they are then classified as used cars. As an asset ages, its price changes because it declines in efficiency in the current and in all future periods. Depreciation reflects the change in net present value over time. Revaluation, on the other hand, is the change in value or price of an asset that is caused by everything other than aging. This includes price changes due to inflation, obsolescence, and any other change not associated with aging [2]. Used cars are subject to depreciation and revaluation. Depreciation can be used as an umbrella term for both of these, and the rest of this report will follow that convention when referring to the loss of value over time. Revaluation plays a part in the depreciation of cars based on the features that they have. Power hungry cars will be less sought after when the price of gasoline is high, for example. A car with the same make, model, year, and geographic region, but this a larger engine than a different car should command a different value at different times.

In addition to the age of the car and the revaluation of its features, used cars have a unique service history that develops over time. Parts will become worn with time and miles driven (mileage). What is replaced, when it is replaced, and by whom, are all to be considered as it relates to the current working condition of the car and its desirability on the market. The particularities are difficult to account for in traditional price-setting models, as it is a major differentiator in vehicles. Generally, it is summarized in the "condition" of the car. The value of repairs or custom modifications to the car are recognized only if they noticeably improve the overall condition of the car.

Using machine learning to better utilize data on all the less common features of a car can more accurately predict the value of a vehicle. This is a clear benefit to consumers, especially those who themselves cannot ascertain the value of the vehicle that they are buying or selling and must rely on a tool. A tool that is more tailored to the non-standard features of the car can provide a more accurate price and make the market fairer for all participants.

There are several machine learning regression models that can be applied to price prediction. This work will investigate which one offers the best performance according to several criteria. The nature of machine learning is to train on past data to predict unseen data. Applied to price prediction of cars, the data is sourced from past sales while the predictions are for the present value of cars. Therefore, a criterion for the selection of a machine learning model it remains accurate in its predictions for future years, not included in the data set.

1.4 Research Questions

The research questions that this study will answer are:

(1) Which ML model and parameters gives the best overall accuracy in making price predictions for used cars?

(2) Which ML model can most accurately assess the depreciation of a car over time?

(3) Which ML model demonstrates the best potential for development of a consumer tool for evaluating used cars or a particular subset of used cars?

These are chosen to satisfy the scientific goals. Research Question 1 will determine which of several algorithms gives the best performance in a verifiable way. Research Question 2 will then examine and compare the behavior of the algorithms to suggest which can best assess depreciation over time, if any. Finally, Research Question 3 will combine the knowledge gained from the previous questions and show which of the algorithms in aggregate demonstrate the best potential for building a consumer tool for price prediction of used cars.

1.5 Scope

This work will focus on answering the research questions. They all entail a comparison of different ML algorithms for price prediction. This will be accomplished by sourcing and preparing a dataset on which all the algorithms can be trained on and compared fairly. The algorithms selected must therefore be similar enough for the same dataset to be used for all of them. This also means that no large optimization efforts on the dataset will be made to boost the performance, if these changes do not benefit the other models. Maximizing price prediction performance of any one algorithm in ways that do not offer better comparisons is outside the scope of this work.

1.6 Outline

Chapter 2 will explain relevant theory and related work to give introductory knowledge of the concepts and related research. Chapter 3 will go over project milestones, motivations for these milestones being chosen, and how they will be accomplished. Chapter 4 will describe the implementation of the research to fulfill the project milestones. Chapter 5 will present the results of the measurements resulting from the implementation with tables and charts. Chapter 6 will discuss the results, the achievement of project milestones, and the societal and ethical implications that this work could have. Chapter 7 will present the conclusions that can be drawn from this work, definitively answer the research questions, and explore the potential for future research.

2 Theory

This chapter will explain relevant theory and related work. This includes concepts related to regression learning, all metrics used for the performance measurement of the models, and related research in the field of machine learning applied to price prediction.

2.1 Regression Machine Learning

Regression analysis is a fundamental concept in the field of machine learning. It is a type of supervised machine learning wherein the model is trained with both input features and output labels. It helps in establishing a relationship among the variables by estimating how they in combination arrive at an estimated output in the form of a continuous variable rather than a discrete label. The input variables are called independent variables and correspond to features in the dataset, while the output variable is called the dependent variable. The simplest of these algorithms is linear regression which assumes that the relationship of each variable is linearly proportional to the output. [3]

2.2 Overfitting

Overfitting a model is a condition where a statistical model begins to describe the random error in the data rather than the relationships between variables. This condition can affect all supervised machine learning models. In the case of regression models, overfitting can occur when there many terms for the number of observations. This leads to the regression coefficients representing the noise rather than the actual relationships in the data. Much better prediction results on the training data is an indication of overfitting. [4]

2.3 Linear Regression

Linear Regression is a technique to estimate the linear relationship between each of a number of independent variables and a dependent variable. Linear Regression fits a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. [5]

2.4 Ridge Regression

Ridge Regression is closely related to linear regression and also assumes a linear relationship between features and the dependent variable (price).

It utilizes a regularization technique that penalizes the use of large coefficients when optimizing the linear relationship. [5] A supplied parameter alpha determines the factor with which large coefficients are penalized. Ridge regression performs L2 regularization meaning that it adds a penalty equal to the square of the magnitude of coefficients. [6]

Minimization Objective: (LR-Obj) + α(sum of square of coefficients)* (1)

2.5 Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) regression performs L1 regularization meaning that it adds a factor of the sum of the absolute value of coefficients in the optimization objective. This penalizes large coefficients when optimizing the linear relationship of each variable, like Ridge Regression. [6]

Minimization Objective: (LR-Obj) + α(sum of absolute value of coefficients)* (2)

2.6 Random Forest Regression

Random Forest is an ensemble learning technique for classification and regression tasks. The algorithm makes use of Decision Trees. They consist of a set of independent binary trees, each stochastically trained on random subsets of data. Although these trees individually may be overtrained, the randomness in the process of training results in the trees producing independent estimates, which are then combined to produce a result. Random Forests have been shown to be effective in a wide range of classification and regression problems. The generalization error for forests converges asymptotically to a limit as the number of trees in the forest becomes large. The generalization error of a forest of Decision Tree Regressors depends on the strength of the individual trees in the forest and the correlation between them [7]. Random Forest Regression is a stochastic process in that each tree is trained on a random subset of data, meaning that the algorithm will behave differently each time it is trained. The algorithm therefore combines the results of many Decision Trees utilizing regression. The sci-kit learn library implements these trees to minimize the objective function MSE (equal to the square of the RSME, see Equation 3).

2.7 Evaluation Metrics

RMSE (root mean squared error) is a commonly used measures for evaluating the quality of predictions in regression ML. It shows how far predictions fall from measured true values using Euclidean distance. Since the error is squared in this method, a few unusually large prediction errors will skew the metric higher than more evenly distributed errors. A lower value indicates higher prediction accuracy. [8]

The equation below shows the formula for calculating RMSE, where "ŷ" is the predicted value, and "y" is the actual value.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{3}$$

R-squared is a another commonly used measure for evaluating the predictions in regression ML. It is also termed the standardized version of MSE (the squared value of RSME) because it is unaffected in magnitude by the scaling of the values in the dataset. That means that the absolute magnitude of the errors doesn't affect the R-squared measure, only the proportion of those errors to the average value. Like RMSE, a few uncommonly large values disproportionately affect the value. R-squared values are always in the range of 0-1, with one being no error (the predictions of the ML model perfectly fit the actual data). Values closer to one means that the ML gives better predictions. [9]

MAPE (mean absolute percentage error) is another measure for evaluating the accuracy of predictions. It is calculated by taking the absolute value of the percentage error between the actual value and the predicted value for each element. The values are then averaged to get the MAPE. This estimated error is not squared unlike for the RMSE and R- squared metrics. The individual errors before averaging depend on the proportion of the magnitude of the actual value to the magnitude of the predicted value. Perfect predictions will give a MAPE of zero, and a lower value signifies better predictions.

2.8 Related Work

2.8.1 House Price Prediction

A previous study using machine learning regression applied to house price predictions compared the performance of five algorithms. It also attempted to analyze the correlation between variables to determine each of their influences on the price of a house. The study concluded that Lasso Regression showed the best overall performance, although ANN achieved a slightly higher RMSE score. [10]

The regression algorithms used in the study were Linear, Lasso, Ridge, Random Forest, and ANN. Similarly, this study will apply Linear, Lasso, Ridge, and Random Forest to price prediction to evaluate and compare each of them with various metrics. The evaluation metrics chosen for their study were RMSE and R-squared. This study will use these same metrics with the addition of MAPE to compare the evaluation performance of the models tested.

This study differs from their thesis in the application of the price prediction. Their study trained the regression algorithms to predict the price of houses, while this study will predict the price of used cars. A notable difference between these is that, over time, houses will increase in value while used cars will decrease. In fact, their study found that the variable representing the year that the house was sold had the highest positive correlation to the price. The dataset used for training of the models included a larger number of continuous variables (features) than datasets for used cars typically include.

2.8.2 Modern Housing Valuation, A Machine Learning Approach

Another study that applied ML algorithms to predict the price of houses to achieve the highest possible accuracy, as well as judge the relative importance of the variables (features) in the dataset. Using an ANN (Artificial Neural Network), with implementation-specific improvements, they were able to achieve a MAPE of 6.37%. This value is lower than the manual real-estate agent’s appraisals of their data set. This demonstrates the potential of machine learning to more accurately assess the price of an asset than existing methods. [11]

Their study is similar to this one in that they applied ML to price prediction. The metric that they used to evaluate and optimize the model performance was MAPE, which is one of the metrics that this study will use.

In contrast to this study, all of the models evaluated in theirs were ANNs. This study will not implement any ANN models.

2.8.3 Comparison of Supervised Learning Models for predicting prices of Used Cars

A study aiming to decide to investigate the optimal ML algorithm for price prediction of used cars elected to consider the algorithms Linear Regression (LR), Light Gradient Boosted Machine (LGBM), Random Forest Regression (RFR), and Decision Tree Regression (DTR). Additionally, they sought to evaluate the relative feature importance of the variables in their dataset. They compared the R-squared performance of these algorithms and found that LGBM scored the best. RFR scored a close second on this metric, and slightly outperformed on other metrics. The three most important features for price prediction were found to be the region, mileage, and manufacturer of the car, in that order. [12]

Their study had similar goals to this one. The also selected various models to train on a dataset and compared the performance with various metrics, including R-squared, as this study will include. This method of this study differs in which models and evaluation metrics will be chosen. Their study focuses on the R-squared metric, which is squared like the loss functions of most ML regression models.

3 Methodology

The chapter will present the method followed in performing the research.

3.1 Scientific method description

This work will use a quantitative method to achieve the scientific goals. The evaluation of models will be done by collecting and comparing various performance metrics for each of the machine learning algorithms to be tested in this work.

Machine learning models need a large amount of data to train on. The first step in performing this study is to source a sufficiently large and reliable dataset. There are several criteria for such a dataset. It must be large enough, include sufficiently many relevant features, have very few null values for those features, have reliable values, and must be distributed over several years.

To ensure the highest possible accuracy for the various models, a result-driven iterative process including data cleaning, model training, and model testing will be used to refine the models.

3.2 Project method description

From the project statement and the scientific goals, the following project milestones were produced:

1. Study previous research into price prediction models with regression and identify the most used and most viable algorithms for the task.
2. Source an appropriate dataset of peer-to-peer car sales to use in the training of the models.
3. Remove any missing or outlier values from the dataset and make appropriate normalizations to the data.
4. Instantiate one of each of the models and make appropriate normalizations to the dataset to boost the performance of each model.
5. Measure the efficacy of the models and compare the performances.
6. Compare the model's predicted depreciation by simulating the aging of the vehicles in the dataset.

The first project milestone is to use previous research on price prediction and identify the most used and viable ML regression models. This milestone is necessary to gain an understanding of which ML models are the best candidates for developing a price prediction tool, and therefore the most relevant to study in this research.

The second project milestone is to source an appropriate dataset of peer- to-peer car sales for the model to be trained on. This milestone was chosen in order to have a sufficiently large and complete collection of car sales data for the models to provide accurate predictions and therefore meaningful comparison of them. The dataset must also span several years for the model to be able to infer prices in years future to the dataset. Keeping these criteria in mind, there are several publicly available datasets to be had from sites such as Kaggle.

The third project milestone is to remove missing and outlier values from the dataset. This milestone was chosen because many ML models are sensitive to outlier values. Datasets that are sourced by means of web- scraping can often have missing, incomplete, or unreasonable values. These need to be identified and removed. A caveat to this is that removing too many infrequently occurring values can reduce the size of the dataset, which will negatively affect the prediction accuracy of the model. Removing infrequent values will also limit the potential of the model to predict similar values. For example, removing rare car makes and models means that the scope of the ML model will not include those makes and models.

The fourth project milestone is to instantiate each of the models and make appropriate normalizations to the dataset in order to boost performance. The training dataset will be used to train the machine learning algorithms chosen to predict the price. From the cleaned dataset, 80% will be randomly selected to be used in training the models while the remaining 20% will be used for testing. Achieving this milestone requires preparing a programming environment which allows access to all the regression ML models chosen. Python3 with the sklearn library provides an easy way to implement, train, and test the models.

The fifth project milestone is to measure the efficacy of the models and compare the performances of each. The metrics used for this will be MAPE, RSME, and R-squared. Since the models are trained and tested on the same data, these metrics can be directly compared. The MAPE metric is the most important for evaluating a future potential consumer tool for valuation of used cars. This is because the formula for calculating MAPE does not square errors, and therefore the relative (percentage) errors are equally considered in calculated the metric. A consumer is likely to consider the average error in the price prediction in deciding how accurate the price prediction for their car valuation.

The sixth and final project milestone is to compare the prediction the model's predicted depreciation by simulating the aging of the vehicles in the dataset and measuring the average percentage change in the new predictions compared to the original. This milestone was chosen to add another evaluation criteria for deciding which of the models are most suited for price prediction. Being able to infer values future to the dataset helps to prevent obsolescence of the model. Used cars are a depreciating asset and the model should reflect that. Furthermore, newer vehicles in aggregate will depreciate faster than older ones. To achieve this milestone, we will simulate the aging of the vehicles in the dataset by incrementing the features *yearsold* and *Year*. The feature *Mileage* must also be increased by the average miles that are driven in a year. According to the Federal Highway Administration, American cars are driven an average of 14,263 miles per year. Thus, for each vehicle in the dataset, these three values will be increased and fed into the model to generate predictions for the aging of the vehicles. Thereafter, the percentage change in the predicted price will be recorded for each of the models. Previous studies show that geometric depreciation is a good approximation of real vehicle depreciation in developed countries for used cars. The annual depreciation rate for this distribution was found to be in the range of 15-31% in one such study [2]. Geometric depreciation means that the percentage decrease in value each year is constant. By measuring the predicted depreciation for cars with different ages, it can be shown whether the models approximate geometric depreciation. The expected result assuming geometric depreciation is that the cars, regardless of their ages, approximately lose the same percentage of their value each year. Additionally, this value can be expected to be in the range of 15-31%. A caveat for this value is that the dataset is not necessarily representative of the population of cars. Cars are not worth anything are not sold, and not represented in the dataset. Additionally, some cars can increase in value and subsets of cars that are sold more often will be overrepresented.

3.3 Evaluation method

This work will be evaluated by how well the results derived from the method description are able to produce satisfactory answers to the research questions. The method should be able to produce conclusive answers to the first two research questions. It will be possible to train and test the Machine Learning models chosen, so long as they are viable for regression analysis. Through the creation of dummy variables, the categorical features in the dataset can be converted to continuous variables to be used as inputs in the regression models. This can however lead to a loss of information and reduced performance of the various ML models to different degrees. This work is contingent on the ability to fairly compare the performances of the algorithms according to several criteria, but not necessarily on achieving a very high performance for any of the algorithms, although if they do all have to achieve performance results that show that they were implemented successfully and can thus be fairly compared.

4 Implementation/Design

This chapter will describe the process of implementing the system. The implementation was divided into five parts titled Data Set, Data Cleaning and Normalization, Machine Learning Algorithms, Measurements, and Inference. Each of these parts are explained in their own sections as part of this chapter and are shown in the UML diagram below (see figure 1). The high level component of the UML diagram without a dedicated section of this chapter, Simulated Aging, is detailed in the measurement section. The entire implementation was written in Python3 in the PyCharm ide. The libraries utilized are pandas, sklearn (sci-kit learn), NumPy, re (regular expressions), matplotlib, and seaborn. (See Appendix C for the entire source code)

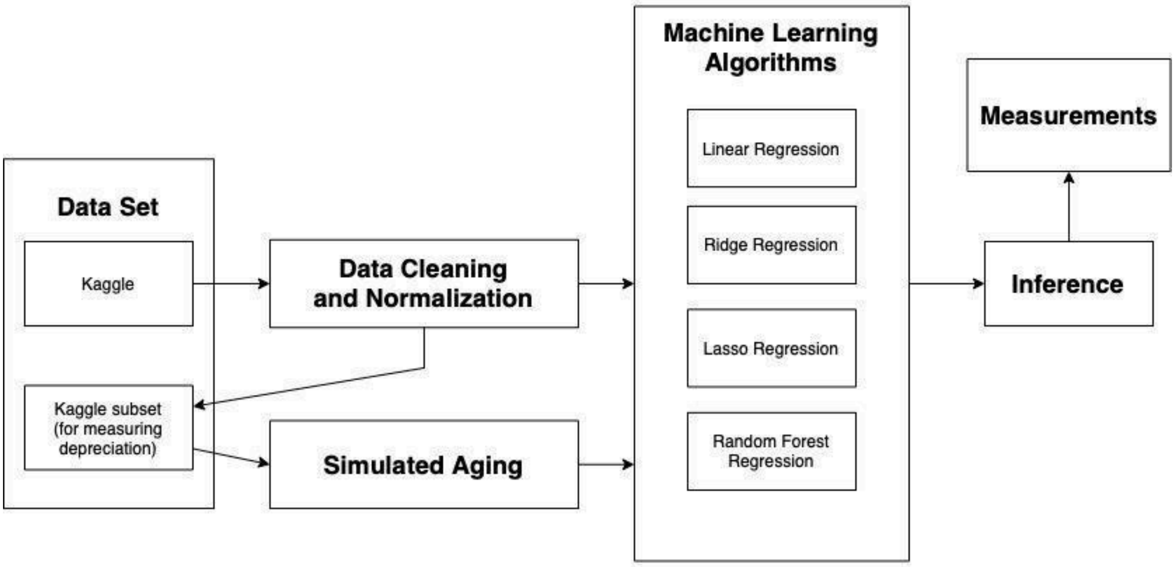


Figure 1: UML Component Diagram, Implementation Overview

4.1 Data Sets

4.1.1 Kaggle Dataset

The dataset was sourced from Kaggle and includes 122,144 car listings from the years 2018, 2019, and 2020 from all areas in the United States. It is available publicly. It includes all types of road-going consumer vehicles, such as vans, pickup-trucks, and cars (See Appendix B for the published data set). This dataset shows listings of used cars and not necessarily the final sales price. The dataset does not have duplicate listings of the same car however, with the previous listings being removed as the sale was most likely unsuccessful. Therefore, the listed price may be somewhat higher than actual sales price and not reflective of the actual values of the cars. This error is consistent across the dataset (including the entries that will be used for testing) however and should not significantly affect measured model performance.

The dataset has 13 columns including, shown in the table below.

Table 1: Dataset features

FEATURE	EXPLANATION
ID	
PRICESOLD	The price at which the vehicle was listed at
YEARSOLD	The calendar year when the vehicle was sold
ZIPCODE	The zip code where the car was listed
MILEAGE	
MAKE	
MODEL	
YEAR	The production year of the vehicle
TRIM	The version/configuration of the model
ENGINE	The engine type/specification (including displacement in liters)
BODYTYPE	
NUMCYLINDERS	The number of cylinders of the engine
DRIVETYPE	The type of drivetrain (RWD, AWD, FWD, 4WD)

4.2 Data Cleaning and Normalization

The first step in cleaning the dataset provided from Kaggle was to identify variables which will not be useful for training the models. This includes features which are not correlated with price, have too many discrete values to draw inferences from, or have too many missing values. The features that were identified to be dropped from the dataset were: *ID*, *zipcode*, and *Trim*.

The next step is identifying and removing outliers for the ten remaining features. Keeping in mind the distribution of the data and the negative effect of removing too many values, appropriate minimum and maximum values were set for each feature to remove rows in the dataset which were extreme in any feature category. This was performed for the features *pricesold*, *Mileage*, and *Year*. (See Table 2) These were chosen somewhat arbitrarily but with the purpose of removing an appropriate percentage of uncommonly occurring extreme values in the dataset. This increases the performance of the models

Table 2: Removal of Outliers

OUTLIER CATEGORY	COUNT BEFORE REMOVAL	COUNT REMOVED	MINIMUM VALUE	MAXIMUM VALUE
PRICE	122,144	11,743	1000	50,000
MILEAGE	110,401	13,950	1000	200,000
YEAR (AGE OF VEHICLE)	96,451	15,902	0	50
OTHERS	80,549	38,562	NA	NA

The remaining features were categorical variables. Since all variables provided into regression models must be continuous, strategies for making these variables continuous must be employed. If the feature is numeric in nature, then it can be made continuous. The *Engine* feature was inconsistently in the dataset but included the displacement which is numeric. Some entries also included the number of cylinders, while omitting this from the *NumCylinders* feature. These could be extracted with regular expressions. The engine displacement was kept in the *Engine* feature rather than the engine type, which could not easily be made to be numeric.

Another strategy to convert categorical variables to numeric is creating dummy variables. If a feature assumes relatively few different values, the feature can be converted into several dummy variables where each unique categorical value becomes a different continuous variable. The values of each of these variables can only ever be zero or one. The creation of dummy variables was applied to the features *Make*, *Model*, *BodyType*, *DriveType*, and *NumCylinders*.

The dependent variable *pricesold* was log-normally distributed. In order to normalize it, the base-2 logarithm of each of the prices is taken as the price to train the model on (see Figures 2 and 3 below).

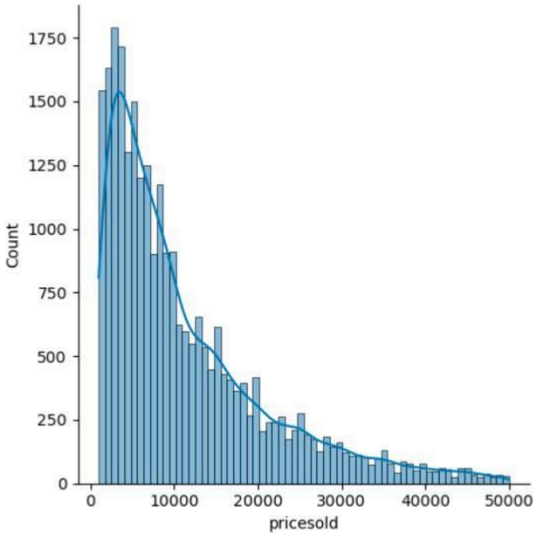


Figure 2: Density Histogram of Pricesold Variable

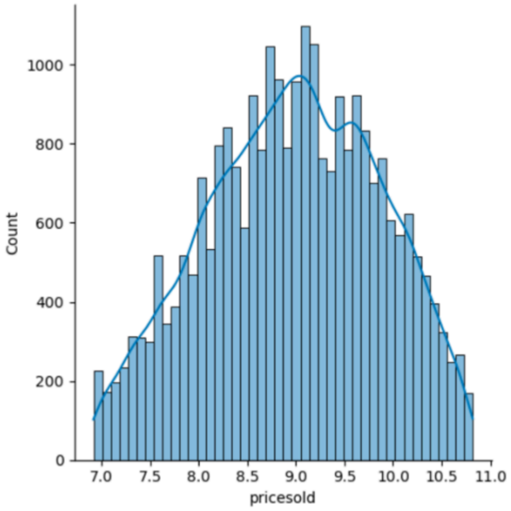


Figure 3: Density Histogram of Pricesold after Log Normalization

4.3 Machine Learning Algorithms

The data, after being cleaned and normalized, is split into training and test data using a randomized 80-20 split. This is to ensure that the data used for testing does not contain any of the data used for training. Thus 20% of the data is reserved for testing purposes (see 4.4 Inference). The training dataset was used to train the four price prediction ML models chosen: Multiple Linear Regression, Lasso Regression, Ridge Regression, and Random Forest Regression. All machine learning algorithms used in this report were imported from the sklearn library. Some models were provided input parameters to implement. The motivations for the choice of input parameters are explained in this section for the models that require them.

The Ridge Regression model was implemented with the argument alpha=0.01. An assortment of different alpha values were tried, and lower values performed slightly better. Values lower than 0.01 didn't noticeably perform improve model accuracy.

Lasso Regression was similarly implemented with an alpha=0.01 value after testing. Lower values gave better prediction results.

Random Forest was implemented with default parameters and random_state=0. The random state is necessary because it is a stochastic process that takes a seed value to begin. When testing different values, there was no noticeable performance increase. The random_state throughout training was therefore set consistently to 0, minimizing stochastic behavior resulting from varying the random_state.

4.4 Inference

Inference involves using the subset of the data that was reserved for testing (20%) to predict the price based on the features. This step was performed after the dataset was cleaned and normalized, and the models were optimized. The dataset was re-split, models were retrained, and inferences retaken a total of five times. This produced five separate inferences with the same parameters to be able to produce an average for the measurements. The inferences produced varies slightly each time as a result of the randomized 80-20 training-testing data split. Each model produced inferences from the same testing subset in every iteration. To judge overfitting, they were also tested on the training subset of the data.

Much better prediction results on the training data is an indication of overfitting.

For the Kaggle subset (cars sales from 2019), an inference was performed once for each model, on the entire dataset. This was done subsequent to simulating the aging of all the vehicles as described in Milestone 6 (see Project Method Description).

4.5 Measurements

The measurements taken for this study are described in this section. All of the measurements are taken from the same inference data for each model and using the formulas for the various metrics.

4.5.1 Training and Testing Accuracy Comparison

The three performance metrics that were taken for machine learning algorithm are R-squared, RMSE, and MAPE. These measurements were taken for both the training and testing inferences and averaged across all five iterations of inferences taken to produce a table of metrics.

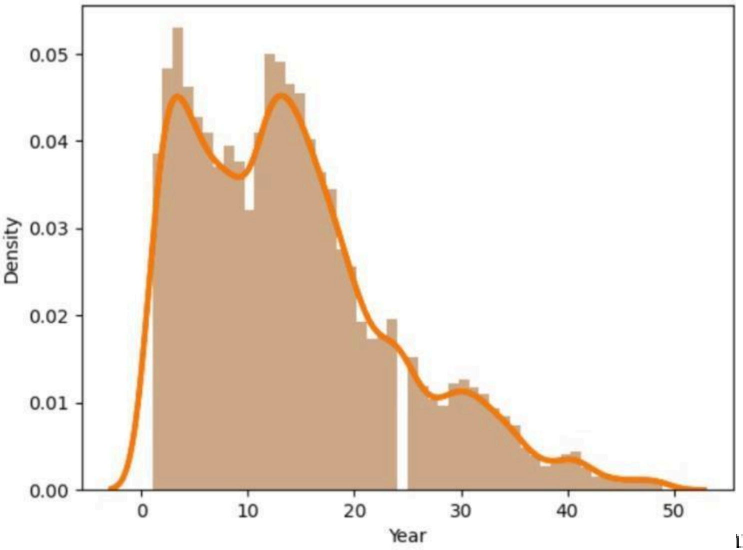
The MAPE is of special importance for evaluating the potential of the algorithms to be used for a consumer valuation tool and fulfilling Research Question 3. Therefore, the dataset was split into four price categories and MAPE measurements taken for the inferences each algorithm. The four price categories were each approximately 25% of the dataset each, with the first one being the lowest priced cars and the last being the highest priced cars. In other words, the MAPE was taken for cars belonging to the 0th-25th price percentile, 25th-50th price percentile, 50th-75th price percentile, and 75th-100th price percentile. This serves to demonstrate the performance of each algorithm across different price categories of cars.

4.5.2 Inferred Price Plots

For the first iteration of inferences (both training and testing), scatter plots were created for each algorithm where each data point is the actual price, plotted against the inferred price. A line of best was then calculated and drawn through these points as well as a line showing the actual values, y=x. If the algorithms do not demonstrate any systematic error, the line of best fit should match this line. See Appendix A for these scatter plots.

4.5.3 Measuring Depreciation

To measure the depreciation with respect to cars of different ages, samples from the inference of the Kaggle subset's two most common age spans were taken (see figure 4). Measurements of the percentage decline in predicted price was produced from this.



icle Age

5 Results

This chapter will present the results of all measurements performed. This includes tables demonstrating the training and testing accuracy of the models, the magnitude of coefficients for the models that utilize coefficients, depreciation measurements, an inference histogram and inference scatterplots. The results are presented with the use of tables and graphs.

5.1 Training and Testing Accuracy

Table 3 shows the prediction accuracy results for both training and testing for each of the four ML algorithms. The predictions accuracy is measured by taking the average value of five iterations.

Table 3: Performance Metrics Training and Testing Data

MODEL	RSME	R-SQUARED	MAPE
LR TRAINING	5799	0.6501	44.45
LR TESTING	5953	0.6448	45.12
RR TRAINING	5798	0.6501	44.45
RR TESTING	5953	0.6448	45.12
LASSO TRAINING	5796	0.6504	44.50
LASSO TESTING	5950	0.6452	45.15
RFR TRAINING	1975	0.9593	12.85
RFR TESTING	4799	0.7692	37.65

5.2 Magnitude of Coefficients

Table 4 shows a sample of coefficients from the Linear Regression model, with samples corresponding to the coefficients of the same features in the Ridge and Lasso models. The coefficients for all the models were of similar magnitude through all five iterations of inference.

Table 4: Sample of Coefficients

Model	Feature coefficient 4	Feature coefficient 114	Feature coefficient 125
Linear Regression	36960088	601928687	-22956190
Ridge Regression	-27	-48	32
Lasso Regression	-19	-41	45

5.3 MAPE by Price Percentile

Figure 5 shows the result of each algorithm for the testing dataset by the price percentile. The price boundaries for these percentiles (before log-normalization) are shown in Table 5.

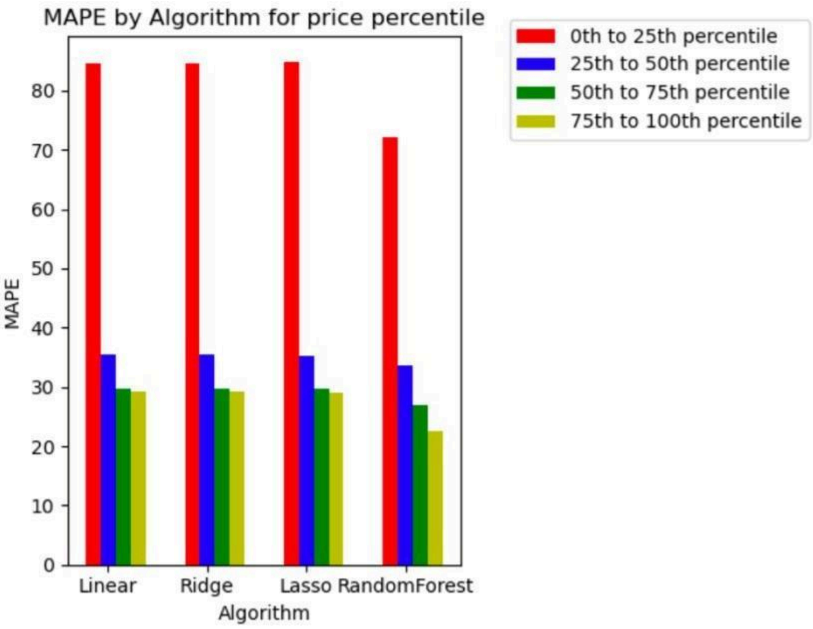


Figure 5: MAPE by Price percentile

Table 5: Price Boundaries for Price Percentiles

PERCENTILE	START PRICE	END PRICE
0TH-25TH	0	4,050
25TH-50TH	4,050	8,050
50TH-75TH	8,050	15,500
75TH-100TH	15,500	50,000

5.4 Measurement of Depreciation

The results of the measurement of depreciation are shown in the chart below. The chart below (Figure 6) displays the percentage decrease in price when simulating the aging of the vehicles in the testing dataset, for cars of two different age categories approximately 10 years apart. This information is plotted for each algorithm to compare.

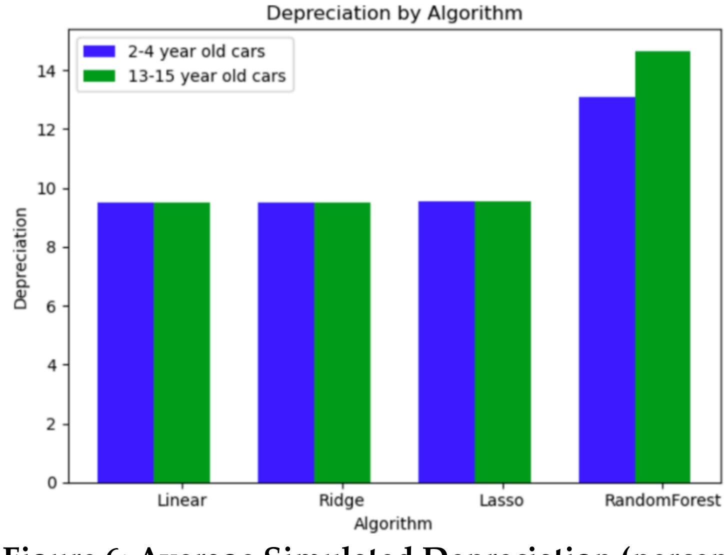


Figure 6: Average Simulated Depreciation (percent)

5.5 Inferred Price Plots

See Appendix A for the scatter plots of the actual and predicted values of the testing dataset for each algorithm.

5.5.1 Inference Histogram

Figure 7, shown below, is a density histogram of each algorithm's price predictions on the test data, along with the actual price distribution. The line depicting Linear Regression follows that of Ridge Regression very closely and is not easily visible.

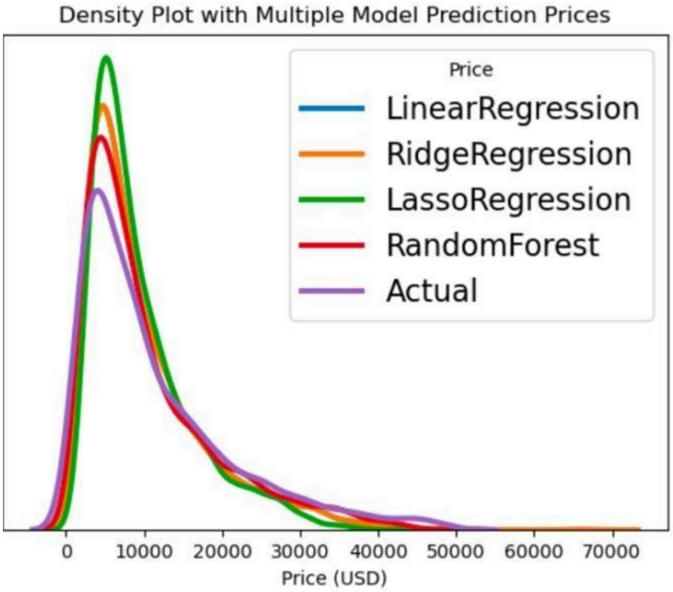


Figure 7: Prediction Density Histogram Test Data

6 Discussion

This chapter will discuss the implementation and results to relate them to the project milestones and the scientific goals.

6.1 Analysis and Discussion of Results

6.1.1 Training and Testing Accuracy

The measurements show that the RSME value for the testing dataset of Linear Regression was the same as for Ridge Regression at 5953, while Lasso Regression performed slightly better at 5950. Random Forest had a much lower value at 4799. Linear, Ridge, and Lasso had a very similar RSME for training data that was somewhat lower than their respective RSME's on the testing dataset. This is to be expected, since the models are trained to minimize the squared error on the data it is trained on. The Random Forest Regression algorithm had an RSME of 1975 for the training data.

The R-squared error for each similarly showed that Random Forest Regression had much better performance. The rest of the algorithms performed slightly worse on the testing data than the training data. Random forest had a value very close to 1, which is the highest possible value that is only reached when each price in the data set is predicted perfectly. This indicates overfitting. This occurs when the model is too complex for the data, and over-tunes the coefficients to predict the individual data points in the training set, while not generalizing well for unseen data (testing data).

Linear Regression, Ridge Regression, and Lasso Regression achieved similar performance to each other judged by the MAPE metric. Interestingly, they all demonstrated better performance on the testing data than the training data for one of the five iterations that were averaged to produce the table. The algorithms are all defined in the sklearn library to minimize the RSME on the training data rather than MAPE. Therefore, the algorithm is not optimized to give the lowest possible MAPE value for the training data, although this value is related to the RSME. The testing data is 25% the size of the training dataset and will therefore give more variance in the results. By chance the testing data can achieve a lower MAPE than the training data and did so in one of the five iterations.

The Random Forest Regression MAPE value for testing data shows that it performed better in its overall score, with a value of 37.65% on the testing data compared to the next best score of 44.45%. When examining this value by four different price percentiles, Random Forest showed higher performance in all four categories although the 0th-25th and 75th-100th percentile categories showed the largest increase in performance relative to the other algorithms (see figure 6). The histogram depicting the density of model predictions (see figure 7) shows that the distribution of the predictions are, for all of the models, very concentrated at roughly 5,000. Appendix A shows plots of the predicted values against the actual values for each algorithm. The line of best fit shows the center point of the line and demonstrates that Linear Regression, Ridge Regression, and Lasso Regression plot have a systematic error in the concentration of predictions for higher actual car prices. The algorithms are very likely to predict the price as lower than actual.

6.1.2 Magnitude of Coefficients

The magnitude of the coefficients for the Linear Regression model were very large as shown in Table 4. The coefficients of the Ridge Regression and Lasso Regression algorithms, which both penalize the use of large coefficients, were much smaller in comparison. This did not seem to have a significant impact on any of the performance metrics, as these three algorithms performed very similarly in all of them. A potential implementation-dependent issue is rounding errors stemming from the use of very large magnitude metrics which approach the upper and lower bounds of values of the datatype that is used to store the coefficients. Since no significant differences in performance was shown in the measurements, there didn’t seem to be any such issues.

6.1.3 Dataset Limitations

A limitation on any ML algorithm’s performance is the dataset. If the dataset does not include features that are strongly correlated to the price, the ML algorithm might not have access to enough information to accurately infer the price. Some strongly correlated features can be rendered redundant if another feature is included in the dataset and is strongly correlated to the redundant feature. If this is the case for a missing feature, it may be partially redundant and therefore unnecessary to include in the dataset. The dataset that was chosen for the training of the models in this work initially included a feature for the zip code for the sale. This feature was removed as part of the Data Cleaning and Normalization outlined in the method. A previous report by Sri Totakura and Harika Kosuru [12] comparing ML Regression model performance found that the "Region" feature in their dataset had the highest feature importance. The study concluded that this feature had the highest correlation to the price by comparing each feature's impact on the price compared to the rest, for their best performing algorithm (Light Gradient Boosted Machine). This suggests that deriving a "Region" feature of the car sales from the zip code available in the dataset could improve model performance. As it relates to the research questions to be answered in this work, the increase in performance could differ between models and affect the results of this study for comparing the performance of models.

6.1.4 Measurement of Average Simulated Depreciation

The average depreciation was shown to be the same for Linear Regression, Ridge Regression, and Lasso Regression, independent of the age of the vehicle. This value was approximately 9.7%. The Random Forest Regression algorithm showed approximately 13.2% depreciation for cars that were 2-4 years old, and approximately 14.6% for cars that were 13-15 years old.

6.2 Project method discussion

As described in the project method description, this work utilizes quantitative analysis with measurements of the implemented machine learning algorithms to achieve the purpose of this work and answer the research questions. To be able to take the performance measurements required as part of this, the machine learning algorithms to compare had to be implemented correctly and using a suitable dataset. The requirements for this dataset were for it to be large, be representative of the used car market which a consumer valuation tool is likely to target, span several years in its sales, and include many relevant features. The project milestones were chosen to accomplish this.

The project milestones for this research were all met. As part of Milestone 1, four machine learning algorithms were identified as the best candidates for comparison, based on previous research in similar areas and which are commonly utilized for regression analysis. Milestone 2 was to choose an appropriate dataset. The dataset chosen met the requirements, although it would potentially have increased the performance of the ML algorithms if it had exceeded the requirements to a greater degree. Because of missing information in some of the features, most of the cars in the dataset were dropped as part of the data cleaning process. Additionally, more features that could be retained after the data cleaning process would be beneficial. Particularly the "zip code" feature, which gives information on the region had too many discrete values to be made continuous. A dataset with more localized sales and therefore fewer discrete values would be more useful. It is possible that the dataset is not representative of the used car market, that is to say that cars with certain features are more likely to be included in this dataset and overrepresented. Milestone 3 was to make appropriate normalizations to the data. This was accomplished but at the cost of reducing the size of the dataset and number of features. Milestone 4 was to implement each of the models chosen as part of Milestone 1. The data cleaning and normalization process was refined based on the requirement for the implementations. The same data was used for all of the models, and the data cleaning and normalizations made should be made to suit all of the models, not to increase the performance of any one model. This was to ensure fair comparison in the performance measurements. Milestone 5 was to make the performance measurements. The metrics for this were chosen to best answer the research questions. To evaluate performance, two metrics commonly used in ML and related to the loss function that each model optimizes during training were used. Additionally, emphasis was put on using a third metric that is more useful for evaluating use for a consumer tool, for several price categories of cars. Milestone 6 was to compare the predicted depreciation for each model. This was to answer research question 2. There are several alternative approaches to doing this, but the one used in this research was to average the percentage decline in predicted price for all cars of a certain year in the testing dataset, and for different ages to see how well the models predict geometric depreciation. Since it is unlikely that the same car was sold multiple times in the dataset, the true depreciation could not be known to evaluate the performance. Instead, the measured average depreciation was compared to an average for all cars obtained from previous research in this area. For this comparison to be valid, the dataset needs to contain years where the depreciation followed this typical average depreciation, and the cars included in the dataset need to be representative of the total market. If these assumptions are true, then any deviation from the average can only be explained by inability of the models to detect depreciation.

6.3 Scientific discussion

The results of this study show that Random Forest Regression was able to achieve better performance for the prediction of price for used cars than the other algorithms tested. Further, the three others tested are variations of Linear Regression, and all performed very similarly despite large differences in the magnitude of coefficients. Random Forest Regression scored better on all three commonly used ML regression metrics and assessed depreciation much more accurately. This makes it more suited to developing a consumer tool for price prediction. However, even when broken down by price category, the model did not achieve a lower MAPE than 20%. Previous research into housing was able to achieve a MAPE of 6.37% for housing price prediction. The conclusions drawn are also limited by the weaknesses of the dataset and model implementation. The dataset was filtered to exclude cars with very low or high values in any feature category, as well as rare car brands. For cars with these excluded values, the model may not be able to predict their prices well and therefore conclusions for the performance of the models may not be applicable to them.

6.4 Ethical and societal discussion

This work utilized a public dataset published on Kaggle (see Appendix B). This dataset was webscraped and this webscraping must be handled ethically. Webscraping is the use of automated tools for collection and extraction of data from the Web for use of further analysis of this data. Web-crawling is one of these techniques that involves running a script that automatically browses a website and retrieves data. This was done by the creators of the dataset, who web-crawled Ebay.com to gather the data. The legality of webscraping depends on the terms of use of the website, infringement of copyright for commercial use, and if any damage occurred to the website. In addition, there are ethical concerns to consider. The privacy of data for individual users of the website must not be compromised. [13]

This research is intended to expand the knowledge needed to create of consumer tool for valuation of used cars using Machine Learning. Such a tool has the potential to change the market for used cars. The societal impact of this tool for consumers looking to buy and sell cars could, if handled responsibly, increase visibility and equality in the market for used cars, as far more individual factors could be considered for valuing a used car. This tool in the hands of an un-informed buyer, could ensure that they are receiving a fair price, and bypass the need for trusted "middle-men" to facilitate a sale.

The elimination of "middle-men" for a transaction means lower frictional costs in the market, and a potential for the seller to find a buyer more quickly and for a higher price. If the use of this tool is widespread and consumers base their buying and selling decision on its predictions, the price of cars could be influenced by the predictions. Errors and other problems with the predictions or tool could negatively affect some consumers. In addition to the impact for individual consumers, the car market and industry that facilitates the sale of used cars could change fundamentally. This could mean the loss of jobs or other negative effects.

7 Conclusions

This chapter will summarize the work as a whole by conclusively answering the research questions proposed in section 1.4, as well as giving examples of possible future work.

The first research question was to determine which of the models and parameters gives the best overall accuracy in making price predictions for used cars. The optimal parameters were determined in the process of implementing the models, and thus each model was implemented with the parameters that yielded the best performance by trial and error. The results show that out of the four models tested, Random Forest Regression provided the highest accuracy in all of the metrics used and highest overall accuracy.

The second research question was to determine which of the models can most accurately assess the depreciation of a car over time. All of the models approximated geometric appreciation, meaning that a constant percentage of value is lost every year independent of the age of the vehicle. Random Forest Regression had a significantly higher assessed average depreciation at approximately 13.8%, compared to the others with 9.7%. This is closer to the range of 15%-31% assessed by Karl Storchmann in his analysis of international depreciation rates [2].

The third research question is to determine which model demonstrates the best potential for development of a consumer tool for evaluating used cars or a particular subset of used cars. The results show that Random Forest Regression performed the best on all performance metrics and for all price percentile subsets of used cars. It was also much better able to approximate the depreciation.

7.1 Future Work

This section will explain some possible future research that can expand upon the knowledge gained through this research.

7.1.1 Applying the Method to Other ML Models

This work compared the performance of four ML Regression algorithms. A way to expand this work in the future is to apply the same method for comparing these algorithms to others that are suited to regression problems. Some example algorithms are Light Gradient Boosted

Machine (LGBM), Kth Nearest Neighbor Regression (KNN), Decision Tree Regression (DTR), and Artificial Neural Networks (ANN). The problem of price prediction deals with continuous variables which makes it suited to regression algorithms, but by creating discrete intervals for the continuous variables such as price, other algorithms could be applied.

7.1.2 Adding Additional Features Related to the Year

A potential improvement to the predictive power of all ML models, if they are able to take advantage of the information, is to add more correlated features. There are some features which are not related to the attributes of the car, such as the price of fuel. A car that uses more fuel will be worth less when fuel costs more. Other such features could include the economic conditions, or changes in the climate.

References

[1] Annina S, Mahima SD, Ramesh B, “An Overview of Machine Learning and its Applications”. *International Journal of Electrical Sciences & Engineering (IJESE)*, 2015. pp. 22-24.

[2] Karl Storchmann, “On the depreciation of automobiles: An international comparison”. *Department of Economics, Yale University*, 2004. pp. 372-373.

[3] Sakshi Gupta, “*Regression vs. Classification in Machine Learning*”, October 6, 2021 [Online] Available: https://www.springboard.com/blog/data-science/regression-vs-classification/ [Accessed March 6, 2022].

[4] Jim Frost, *“Overfitting Regression Models,”* Statistics by Jim, 2022. [Online], Available: https://statisticsbyjim.com/regression/overfitting-regression-models/ [Accessed April 18, 2022].

[5] Skikit-learn, supervised-learning. [Online] Available: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning [Accessed March 12, 2022]

[6] Aarshay Jain, “*A Complete Tutorial on Ridge and Lasso Regression in Python*”, January 28, 2016. [Online], Available: https://www.analyticsvidhya.com/blog/2016/01/ridge-lasso-regression-python-complete-tutorial/ [Accessed March 12, 2022]

[7] Leo Breiman, “Random Forests”. *University of California Berkeley,* 2001.

[8] c3.ai, Root Mean Squared Error. [Online] Available: https://c3.ai/glossary/data-science/root-mean-square-error-rmse/

[9] Ajitesh Kumar, “*Mean Squared Error or R-Squared – Which one to use?”.* [Online], Available: https://vitalflux.com/mean-square-error-r-squared-which-one-to-use/ [accessed March 12, 2022]

[10] Ahmad Abdulal, Nawar Aghi, "*House Price Prediction*". C.S. Bachelors Thesis, Kristianstad's University, 2020.

[11] Nathan Allard, Tobias Hagström, "*Modern Housing Valuation: A Machine Learning Approach*". C.S. Bachelors Thesis, KTH, 2021.

[12] Sri Totakura, Harika Kosuru, "*Comparison of Supervised Learning Models for predicting prices of Used Cars*". C.S. Bachelors Thesis, Blekinge Institute of Technology, Karlskrona, 2021.

[13] Vlad Krotov, "*Legality and Ethics of Web Scraping*", September 2018

Appendix A: Scatter Plots of Price Prediction Results

The scatterplots below show, for all of the datapoints in the set, the actual price of the vehicle on the x-axel and the predicted price on the y-axel. The pink line is the calculated line of best fit for all of these points, while the blue line shows the line $y = x$ (the resulting line of best fit if all predictions were without systematic error). All the graphs show some degree of underprediction, especially for higher actual values.

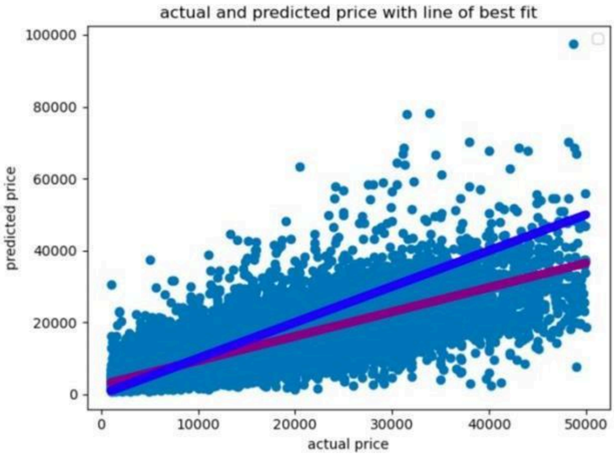


Figure 8: LR Training

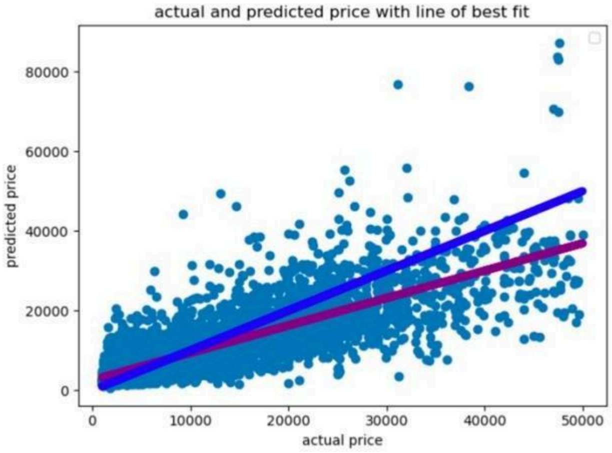


Figure 9: LR Testing

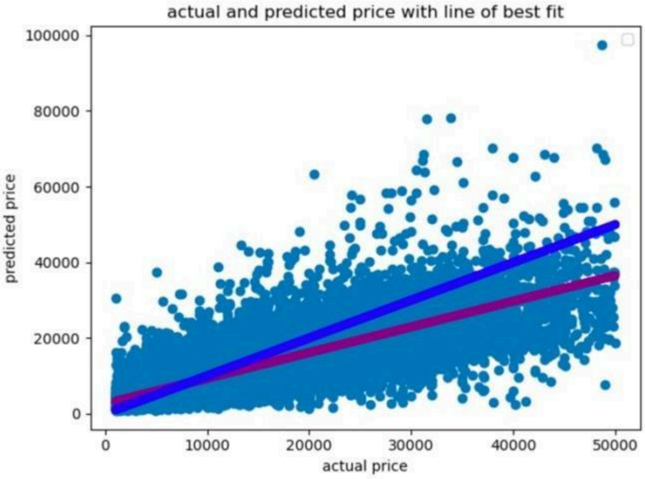


Figure 10: RR Training

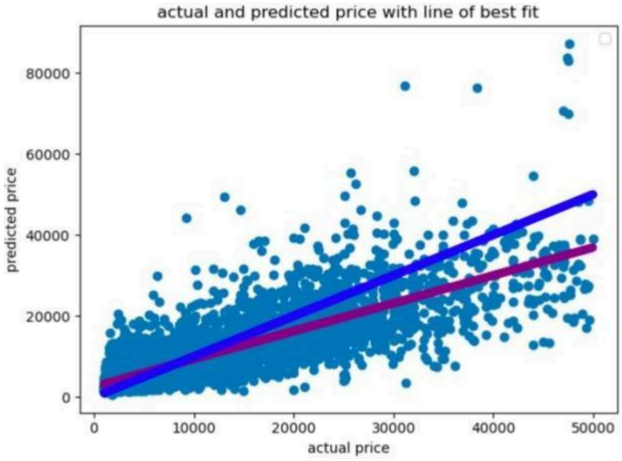


Figure 11: LR Testing

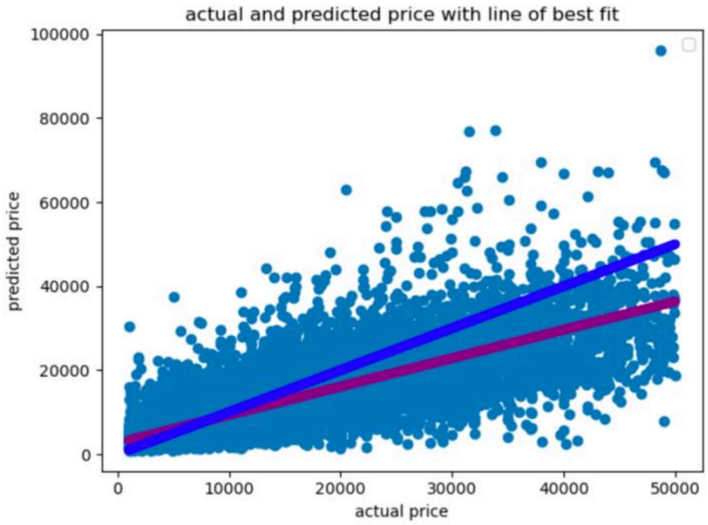


Figure 12: Lasso Training

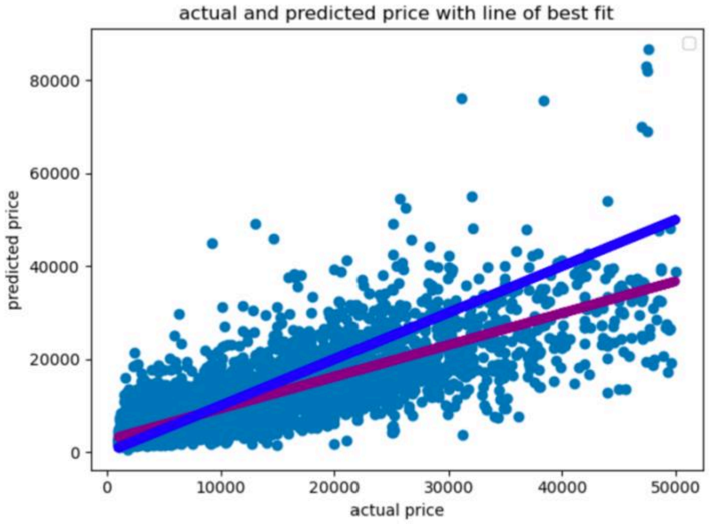


Figure 13: Lasso Testing

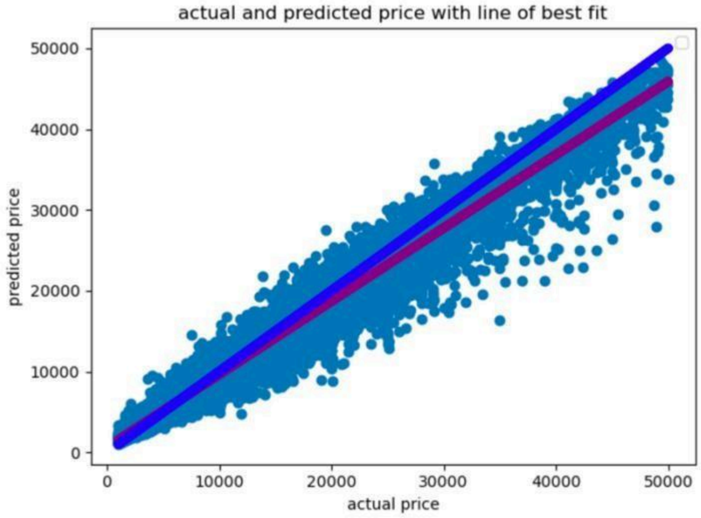


Figure 14: RF Training

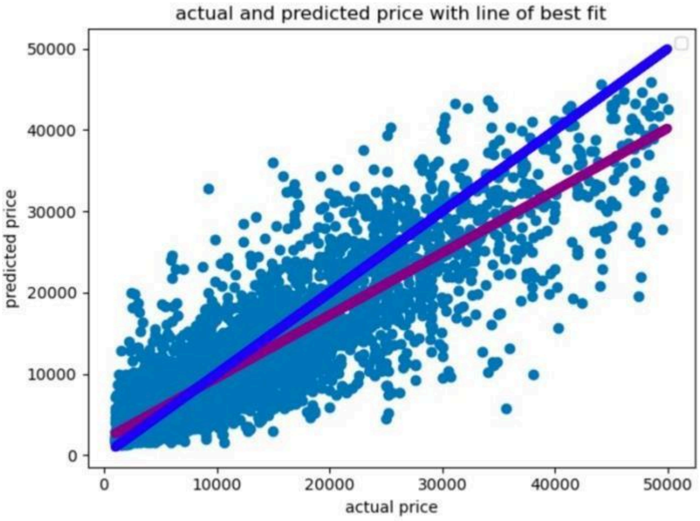


Figure 15: RF Testing