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.