Price prediction of used cars using machine learning

Miralireza Nabavi Bavil
School for Engineering of Matter, Transport & Energy
Arizona State University

1 Introduction

On average, used cars prices are almost 50% lower than new cars! You will be able to pay off a used car much faster, saving you financing fees. Consumers switch cars at an average of six years after purchase, and if you paid $10,000 for a used vehicle instead of $20,000 for a new one, you could opt into a nicer car for your next vehicle or buy another $10,000 vehicle, creating your very own two for one special!

Consumers complain about how quickly a new car depreciates—as soon as they drive it off the lot. The value of a new vehicle can drop 11% on the drive home meaning your $20,000 vehicle is worth only $17,800 once it leaves the lot. The vehicle continues depreciation as weeks, months, and years pass. With used vehicles, the bulk of the depreciation has already occurred. Some used vehicles may even gain value!

If you are in an accident with your new car, the insurance will pay for what the car is worth at that time, leaving a gap between the purchase price and what the vehicle is worth. That's where gap insurance comes in. Gap insurance will cover the difference between what you paid for a new vehicle and what its depreciated value is, but it will raise your insurance premium. Gap insurance is not necessary with a used car as the depreciation has already occurred.

Almost a quarter of the carbon dioxide a vehicle produces during its life-cycle occurs during manufacturing and initial shipment. Buying a used car reduces the carbon dioxide output into the environment. Used cars also impact the environment less than newer, hybrid vehicles. Hybrid vehicles use lithium-ion, lead-acid, or nickel-metal hybrid batteries that have a much larger environmental impact than a used car due to the toxic waste left behind by batteries and acid.

As mentioned above there are numerous benefits associated with purchasing a used car which have paved the way for its massive market. One of the challenges to purchase a used car is skepticism about its price. Predicting a correct price for used car requires expertise and deep understanding of both cars and the market. In this study, we will develop a model to facilitate the price prediction of used car using machine learning. This model will be useful for both families and business owners.

Machine learning is a field of computer science that aims to teach computers how to learn and act without being explicitly programmed. More specifically, machine learning is an approach to data analysis that involves building and adapting models, which allow programs to "learn" through experience. Machine learning involves the construction of algorithms that adapt their models to improve their ability to make predictions.

This paper is organized as follows. The dataset and features are introduced in §2. Section §3 describes the exploratory analysis of the dataset. The results of the machine learning model development is covered in §4 followed by conclusions and suggested future works in §5.

2 Dataset

2.1 Features

The used car dataset selected for this study consist of 20063 observation. For each observation there are 10 features which determines the price of the used car. The features of the dataset are as follow:

1. **trim**

   Trim levels are used by manufacturers to identify a vehicle's level of equipment or special features. The equipment/features fitted to a particular vehicle also depend on any options packages or individual options that the car was ordered with. This feature is categorical.

2. **isOneOwner**

   This feature is categorical and determines whether the used car had only one owner or more. Thus, it is a logical variable (True/False).

3. **mileage**

   This variable determine the mileage of the used car. The average mileage of the dataset is 73113 miles. This feature is numerical. The minimum and maximum mileage is 8 and 488000 miles respectively.

4. **year**

   The year the car was manufactured. The average value is 2007 and dataset contains cars from 1994 to 2014. This feature is numerical.

5. **color**

   This feature is categorical and determines the color of the used car. The colors are black, silver, white, gray, blue, unspecified and other.

6. **displacement**

   Engine displacement is the measure of the cylinder volume swept by all of the pistons of a piston engine, excluding the combustion chambers. This feature is categorical.

7. **fuel**

   This categorical feature determine the type of fuel that the used car consumes. The possible values are gasoline, hybrid and diesel.

8. **region**

   This categorical feature determines the location of the car in the US.

9. **soundSystem**

   This categorical feature specifies the type of sound system in the used car.

10. **wheelType**

    The type of the wheel is determined by this categorical feature.

11. **price**

    The target feature is the price of the used car which is a numerical feature. The average car price in the dataset is $30747.24. Dataset contains car with price from $599 to $80000.

2.2 Features engineering

As previously mentioned, the are several categorical variable in the dataset. The variable 'trim', 'isOneOwner', 'color', 'displacement', 'fuel', 'region', 'soundSystem' 'wheelType' are categorical. This variable are converted to numerical values using one-hot encoding. The variable 'mileage' and 'year' are treated as numerical variables. Thus, after conversion there are 37 features available which determine the target feature, price. The features are normalized using standard scalar method which ensures the standard deviation of each variable is 1 and the mean is zero. The target feature is normalized by $1000.

3 Exploratory data analysis

In this section we will gain some general knowledge about the trends governing the dataset using visualization and exploratory data analysis.

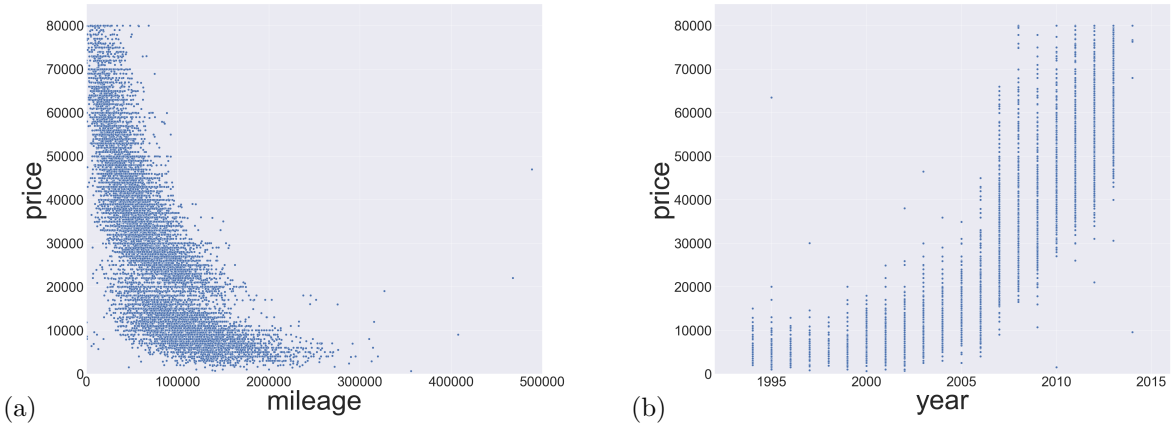


Figure 1: The scatter plot of price versus (a) mileage and (b) year

Figure 1 (a) shows the scatter plot of price versus mileage. We could infer from the plot that as the mileage increases the price of the used car decreases which makes sense. However, the anti-correlation is not linear for the whole range of the mileage. Similarly, in Figure 1 (b), we observe a positive correlation between the feature year and price. As the used car is manufactured more recently, it is more valuable which is logical.

To visualize the effects of categorical variables on the price, bar plots are used. Figure 2 (a) shows the average price of the used cars with a certain fuel consumption. We observe that cars using 'Diesel' fuel are the most expensive followed by 'Hybrid' and 'Gasoline'. Furthermore, from this figure, we could infer that the number of previous owner is a factor in predicting the used car price. Finally, the white used cars have the highest price on average.

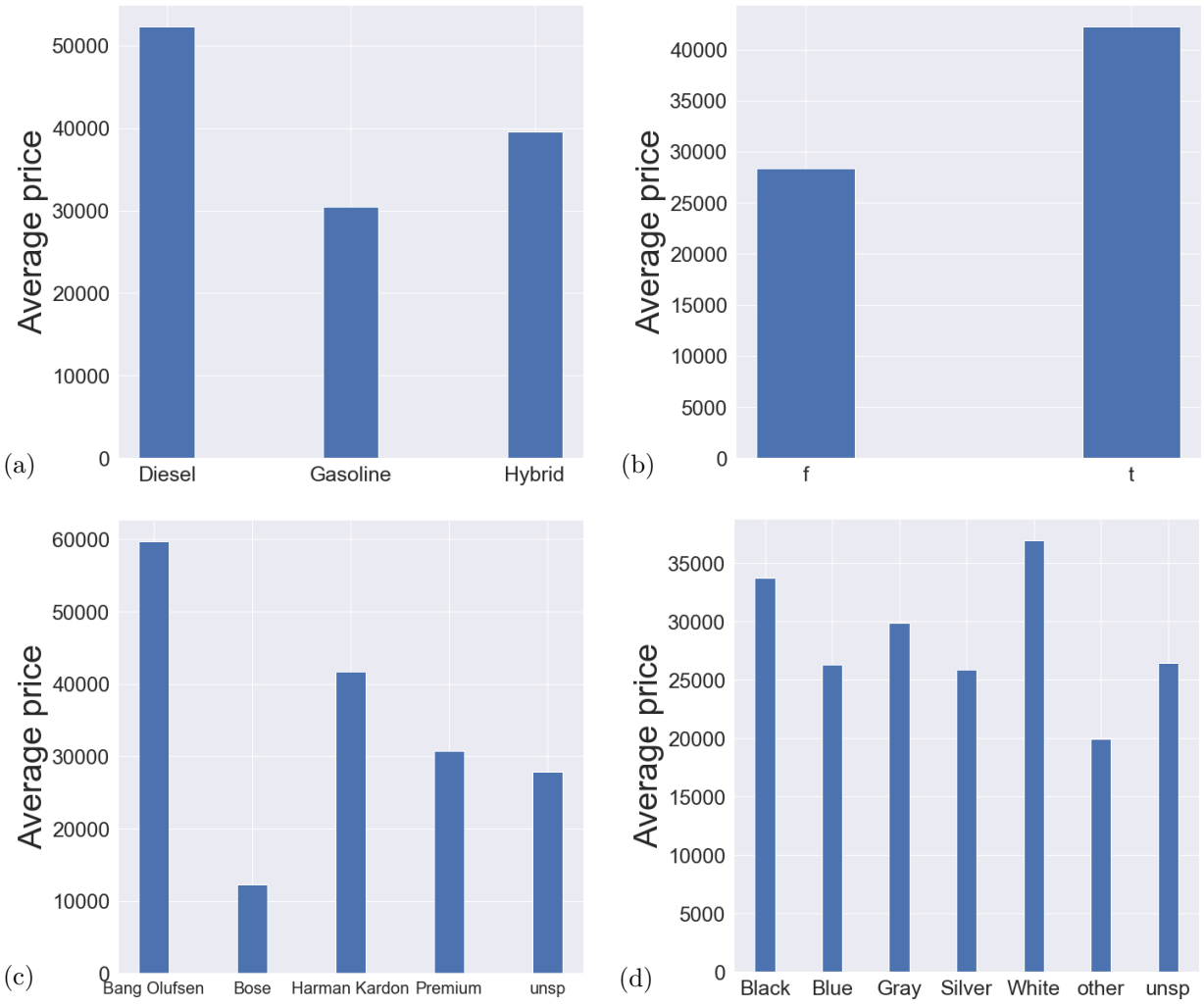
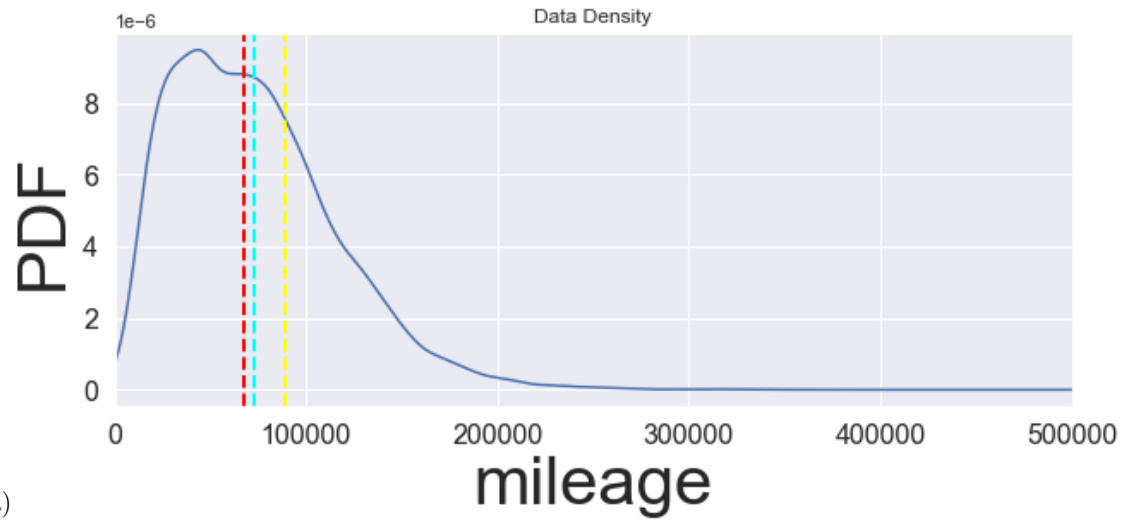


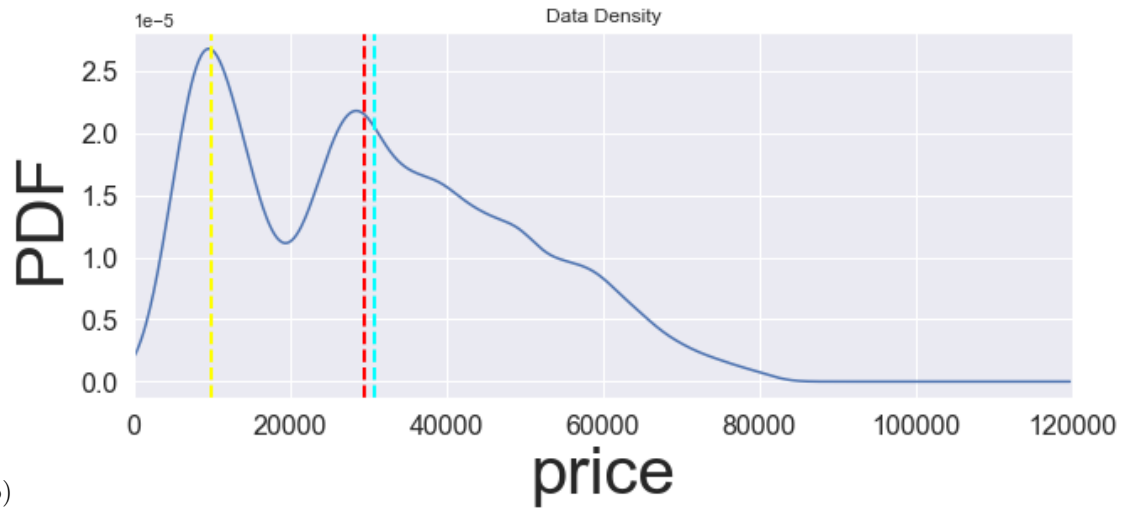
Figure 2: The bar plot of average price versus (a) fuel type, (b) isOneOwner, (c) soundSystem and (d) color.

Finally, Figure 3 shows the probability density function (PDF) of the price and mileage of the used cars. Mean, median and mode are shown with cyan, red and yellow dashed lines,

respectively.



(a)



(b)

Figure 3: The probability density function (PDF) of (a) mileage and (b) price. Mean, median and mode are shown with cyan, red and yellow dashed lines, respectively.

4 Machine learning prediction

In this section, several machine learning models will be developed and compared in terms of their ability to predict the price of used cars. The metrics used to evaluate the models are coefficient of determination (r2-score) and round mean square error (RMSE). The dataset is divided into two sections: 1) training set and 2) testing set. 20 percent of the dataset is reserved for final evaluation, which means that the size of the training dataset is (16050, 37), while the size of the test dataset is (4013, 37). Note that the test dataset was not used for model training. 10% of the training set was used for tuning the hyper-parameter of the models.

4.1 Linear regression with no regularization

Linear regression is a linear model that assumes a linear relationship between the input variables (x) and the single output variable (y). This model is the simplest model that we could use to fit the data. For each feature it provides an intercept and a slope. The linear model was able to achieve a r2-score of 0.894 and RMSE of 33.987. This suggest that price of the used cars is more or less linearly related to the features as the accuracy is quite high.

4.2 Linear regression with Lasso regularization

Regularization is an important concept that is used to avoid over-fitting of the data, especially when the train and test data are much varying. Regularization is implemented by adding a "penalty" term to the best fit derived from the trained data, to achieve a lesser variance with the tested data and also restricts the influence of predictor variables over the output variable by compressing their coefficients [1].

Lasso is one of the important regularization techniques. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. Lasso Regression uses L1 regularization technique.

In this study, a grid search study with 10 folds was performed to determine the best regularization constant. The number of constants used in cross-validation was 1000. Finally, the model was trained using the best regularization coefficient which achieved a r2-score of 0.895 and RMSE of 33.810 on the test set. The regularization slightly enhanced the accuracy of the model on the test set. Result using Ridge regularization was similar. Figure 4 shows the mean squared error versus regularization constant.

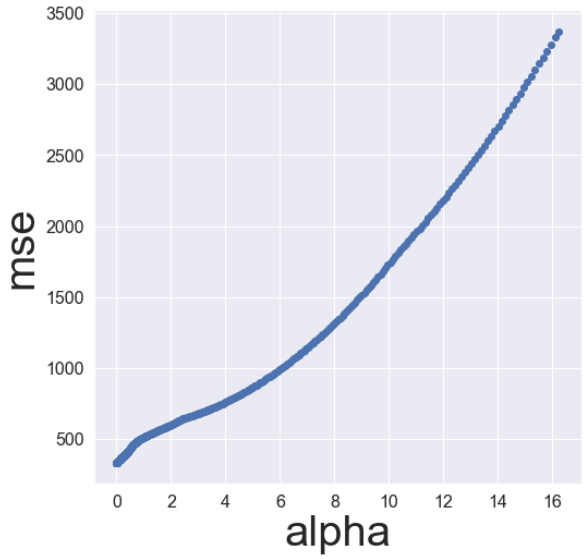


Figure 4: Regularization constant versus mean squared error.

4.3 k-nearest neighbors (KNN)

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. The KNN regressor could be trained with either uniform or distance-based weights. Furthermore, the number of neighbors is an influential parameter in training the model. The grid search analysis was conducted using these two parameters and 10 cross-validation folds. The results shows that with distance-based weights and 8 neighbors we have the best performance. The model was trained using these parameters and achieved a r2-score of 0.878 and RMSE of 39.102. Figure 5 shows the RMSE verus the k.

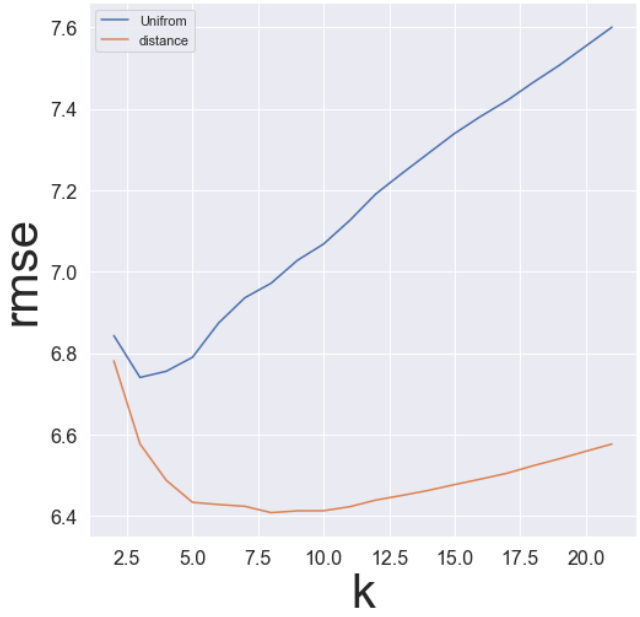


Figure 5: RMSE of the KNN model on the cross-validation test set versus the number of neighbors (k).

4.4 Support vector machine

The support vector machine was utilized the develop a regression model for the used car data. For this model an appropriate kernel and regularization parameter (C) should be selected for ideal training. The options for kernel function are radial-based function (rbf), sigmoid, polynomial and linear. The The grid search analysis showed that the rbf function with $C = 40$ provided the best fit on the cross-validation test set. Figure 6 shows the RMSE on the cross-validation test set versus the regularization parameter, trained with rbf kernel function. The final model was trainedwith ideal parameters and achieved a r2-score of 0.938 and RMSE of 20.105 on the test set.

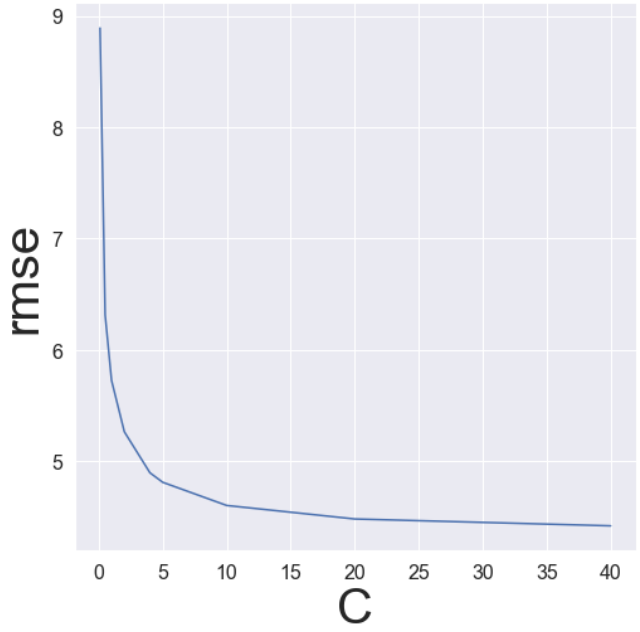


Figure 6: RMSE of the support vector machine model on the cross-validation test set versus the regularization parameter, trained with rbf kernel function.

4.5 Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning [2].

The hyper-parameters for this model are 1) 'criterion' which is the function to measure the quality of a split (available options: squared error, friedman mse, absolute error and poisson), the maximum leaf nodes of the model and maximum features which is the number of features to consider when looking for the best split (available options are auto, sqrt and log2). The best hyper-parameters, according to the grid search analysis, was friedman MSE criterion, 180 maximum leaf nodes and automatic maximum features selection. The model was able to achieve a r2-score of 0.934 and RMSE of 21.066 on the test set.

Using the decision tree model, the importance of each feature could be assessed. This provides priceless knowledge about the influential factors determining the price of the used cars. Furthermore, it could be used to simplify the model by not considering insignificant features. The decision tree model provides the importance factor for each feature, where the sum of these factors is one over the entire features. The features with the highest feature importance factor are 'year', 'mileage' and 'displacement' with factors of 0.817, 0.146 and 0.011 respectively. The feature importance factor was significantly smaller for other features, suggesting that training a model with only these three features would suffice.

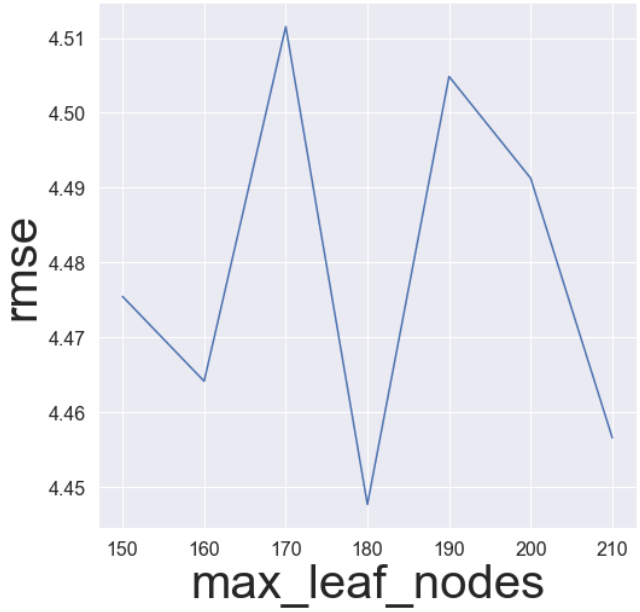


Figure 7: RMSE of the decision tree model on the cross-validation test set versus the maximum number of leaf nodes.

4.6 Random forest

The random forest is a classification/regression algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree

The hyper-parameters for the random forest models is number of estimators which is the number of trees in the forest. The grid search analysis was performed with 500, 600, 700, 800, 900 and 1000 trees. As shown in Figure 8, 600 trees provides the best fit on the cross-validation test set. The final model was trained with 600 trees and obtained a r2-score of 0.939 and RMSE of 19.793 on the test set.

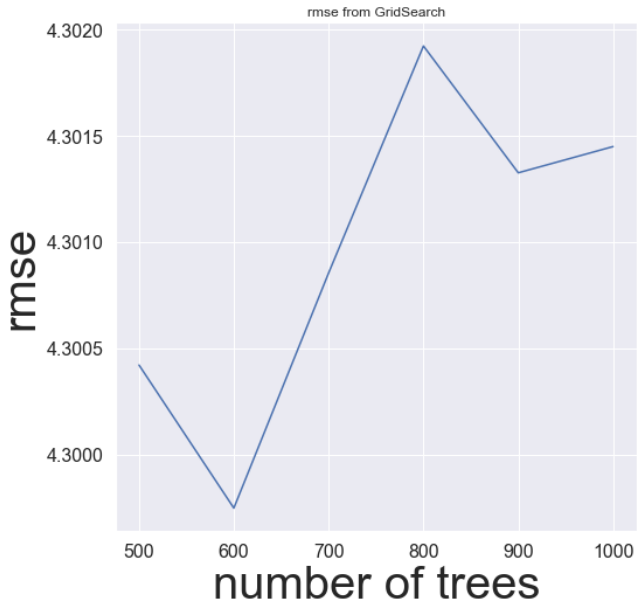


Figure 8: RMSE of the random forest model on the cross-validation test set versus the number of trees.

4.7 Neural Network

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature [1].

Several neural networks with different architecture were trained to fit the used car data. Both Relu and Tanh activation function were tested to fit the database. The performance of the Sigmoid activation function wasn't satisfactory at all. The Neural Network was trained using Adam optimizer on 2000 epochs and RMSE criteria as the metric. Figure 9 shows the architecture of the Neural Network used in this study. Furthermore, several dropout layers were added in the architecture. However, since the accuracy of the validation and training set is almost equal, dropout layers were unable to enhance the model's accuracy. Also, training for longer epochs seems unnecessary as it didn't improve the accuracy as shown in Figure 10. Finally, the results were summarized in Table 1. This table only contains architecture with one and two hidden layers as increasing the number of hidden layers didn't improve the results. The best structure were able to achieve a r2-score of 0.945 and RMSE of 17.72 on the test set

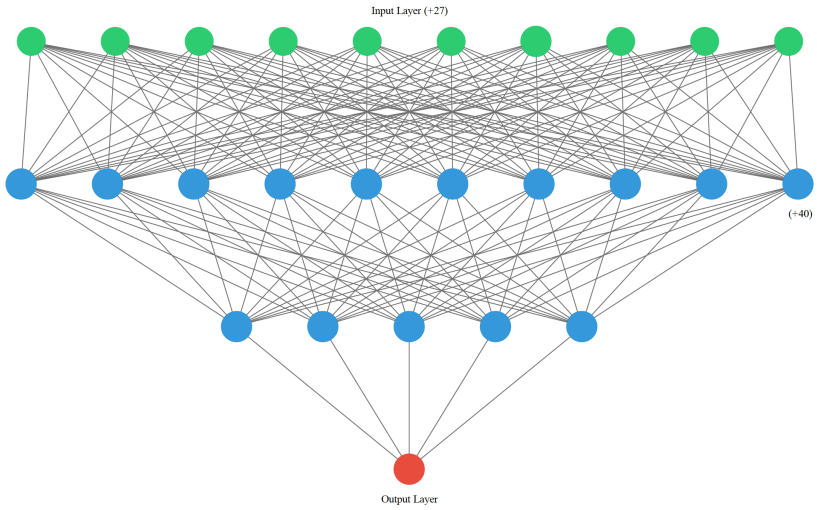


Figure 9: Architecture of the Neural Network applied in this study.

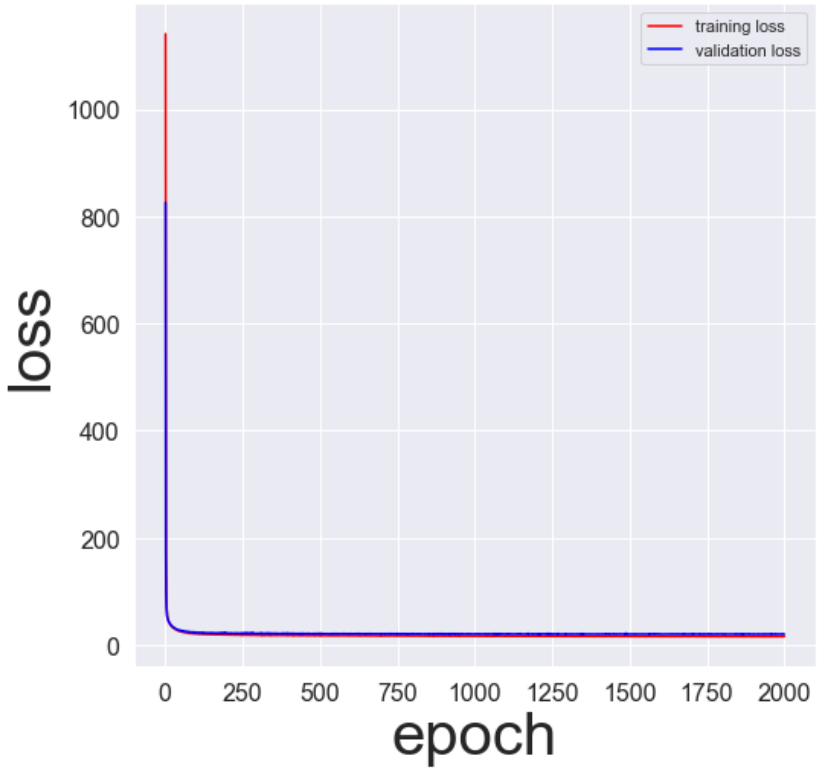


Figure 10: RMSE of the Neural Network on both training and validation sets.

No	Number of layers	L_1	Activation	r2	RMSE
1	$50 \times 5 \times 1$	0	Relu	0.942	18.576
2	$50 \times 5 \times 1$	0.1	Relu	0.943	18.161
3	$50 \times 5 \times 1$	1	Relu	0.940	19.071
4	$100 \times 20 \times 1$	0	Relu	0.940	19.289
5	$100 \times 20 \times 1$	0.1	Relu	0.944	18.097
6	$100 \times 20 \times 1$	0.5	Relu	0.942	18.516
7	$500 \times 100 \times 20 \times 1$	0	Relu	0.929	22.918
8	$500 \times 100 \times 20 \times 1$	0.1	Relu	0.942	18.521
9	$500 \times 100 \times 20 \times 1$	0.5	Relu	0.943	18.265
10	$500 \times 100 \times 20 \times 1$	1	Relu	0.940	19.403
11	$50 \times 5 \times 1$	0	Tanh	0.939	19.515
12	$50 \times 5 \times 1$	0.1	Tanh	0.944	17.855
13	$50 \times 5 \times 1$	1	Tanh	0.943	18.386
14	$100 \times 20 \times 1$	0	Tanh	0.923	24.704
15	$100 \times 20 \times 1$	0.1	Tanh	0.945	17.72
16	$100 \times 20 \times 1$	0.5	Tanh	0.944	18.017
17	$500 \times 100 \times 20 \times 1$	0	Tanh	0.929	22.918

Table 1: Architecture of the Neural Networks fitted to the used car data.

4.8 Gradient Boosting

The final machine learning model utilized in this study is gradient boosting. The hyper-parameters tuned for this model are the number of estimators and maximum depth. For the best hyper-parameters the model was able to achieve a r2-score of 0.944 and RMSE of 17.957 on the test set.

5 Conclusion

The results of the analysis shows that the price of the used cars is dependent on several variables. From those variables, 'year', 'mileage' and 'engine displacement' are the most influential variables determining the price of the used cars. The effects of features like color, fuel type and previous ownership provide insignificant impact on the price, compared to the aforementioned parameters.

Eight machine learning methods (with different levels of complexity) were applied to create a predictive model for the price of the used cars. Surprisingly, simple models like

linear regression were able to achieve an acceptable accuracy. This suggest that the used car dataset contains a strong linearity with respect to the influential feature. Complex models like Neural Networks, gradient boosting, random forest and SVM were able to achieve an equal accuracy on the test set. However, Neural Networks provides the best fit to this data. Table 2 provides a summary of the machine learning models and their accuracy.

No	model	r2	RMSE
1	Linear regression	0.894	33.987
2	Lasso, Rdige	0.895	33.810
3	KNN	0.87	39.10
4	SVM	0.938	20.105
5	Decision Tree	0.934	21.066
6	Random Forest	0.939	19.793
7	Neural Network	0.945	17.72
8	Gradient Boosting	0.944	17.957

Table 2: Summary of the machine learning models and their accuracy.

6 Future work

There are several other models that could be used to fit the used car data, for instance, XGboost. Furthermore, there has been a lot of the improvement in the structure of Neural Networks recently. The author suggest the application of these novel architectures like Cascade feed-forward neural networks, Multi-layer perceptron neural networks, Radial basis neural networks and Adaptive neuro-fuzzy inference systems.

Acknowledgement

The author is grateful to Dr. Robert McCulloch for fruitful discussions.

References

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[2] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.