

COMP 8157

Advanced Database Topics

Used Car Price Prediction

Project Milestone 3

Submitted to: Dr. Shafaq Khan

March 28, 2023

*Submitted By:*

*Team Turing*

*Harsimran Singh - 110090200  
Hardik Sharma - 110090028  
Talha Mohammed Shamoon Choudhary - 110087321  
Abhishek Singh - 110089745*

Table of Contents

[Abstract 4](#_Toc130940214)

[Introduction & Motivation 4](#_Toc130940215)

[Related Work 5](#_Toc130940216)

[Proposed Methodology 6](#_Toc130940217)

[Result 8](#_Toc130940218)

[Limitations or Challenges 11](#_Toc130940219)

[Conclusion and Future Work 12](#_Toc130940220)

[References 12](#_Toc130940221)

Table of Figures

Figure 1: Proposed Model 6

Figure 2: Process Architecture 8

Figure 3: Before Outliers removal...…………………………………………………………..….…………9

Figure 4: After Outliers removal……………………………………………………………………………9

Figure 5: User Interface for Input Values…………………………………………………………………10

Figure 6: Use Interface for Price Prediction Result……………………………………………………….10

Figure 7: Actual Listing from Website………………...……………………….…………………………11

# Abstract

The goal of this project is to apply artificial neural network machine learning techniques to anticipate used automobile prices. Many auto buyers and sellers struggle with identifying the appropriate market value of a used car, which can result in either losing money or paying more than necessary. The purpose of this project is to develop a reliable and accurate car price prediction model to aid in the decision-making process of purchasing and selling old cars.

The utilization of a comprehensive dataset that contains a wide range of features such as automobile make, model, year, mileage, fuel type, engine size, and many more distinguishes this work. The dataset was gathered from many sources, and a thorough preprocessing phase was carried out to assure the data's quality and consistency. Data of thousands of cars currently listed online on websites such as autotrader.ca and autotemp.com was used to gather data which makes it relevant to analyze the current prices of the used cars.

This project employs Artificial Neural network machine learning methods to estimate car pricing. The performance of the model was analyzed using several metrics such as Mean Squared Error (MSE), and R-squared (R2) score. The correlation technique was used for selecting the most critical elements that influence automobile costs.

The results show that the Neural network model beats the other models, with an R2 score of 0.77. The most essential factors influencing car pricing are the year, mileage, engine capacity, and fuel type.

In conclusion, our study offers a dependable and accurate model for predicting automobile prices that can help sellers and buyers of cars make wise choices. By offering a fair and transparent pricing mechanism, the results of this effort may have substantial effects on the used car market.

**Keywords**: Car Price Prediction, supervised learning, neural networks, tensorflow

# Introduction & Motivation

Millions of automobiles are bought and sold each year in the complicated and dynamic used car market. Used automobiles worth $1.21 billion were sold in Canada in December 2022 [1]. Finding the fair market value of a used car is one of the most difficult tasks for both buyers and sellers of cars. Several elements, including the brand, model, year, mileage, condition, and many others, have an impact on a used car's pricing. To make sure that both buyers and sellers are making informed decisions and not losing money, it is essential to determine the exact worth of a used car.

This project aims to develop a neural network machine learning model that can efficiently forecast used car values based on a variety of factors. The fair market value of a used car can be determined using this methodology, which will help buyers and sellers of cars make educated choices.

A vital industry that significantly affects the economy is the used automobile market. To assess the fair market value of a used car, however, is challenging for both buyers and sellers due to the market's lack of transparency and information asymmetry. As a result, you can overpay for a car or lose money when you sell it.

The creation of a trustworthy and precise automobile price prediction model might have a number of advantages. First off, offering an estimate of the fair market value of a used car based on a variety of qualities, it will help car purchasers make informed selections. Second, it will help auto dealers appropriately price their vehicles so they don't lose money while selling them. To overcome the difficulties experienced by both car buyers and sellers in the used automobile market, it is crucial to construct a machine learning-based car price prediction model.

# Related Work

With the increase in car production, every year proportionally creates a market for old car usage. Which leads to the selling of cars at an unreasonable price to customers. The main goal is to help customers which can be any institution (car rental companies), end consumers to buy cars at the current market price rather than getting duped.

Multiple research papers in recent years have been published describing the models and techniques to predict the price in which they are using different supervised learning techniques (Lasso Regression, Multiple Regression, and Regression trees) for price evaluation of used cars. Some of which are using use 5-fold cross-validation to show the effects of linear regression and random forest under different circumstances. This project considers current market listing from websites such as autotraders, autotempest and inflation factor which is missing already proposed papers.

**Ganesh, Mukkesh & Venkatasubbu, Pattabiraman, Dec 2019** **[2]** tried to explain price evaluation using supervised learning techniques. They proposed the model which consists of four main steps firstly the **null hypothesis** has been used to overcome the problem of overfitting in regression trees by using multiple and lasso regression tree after that the **lasso regression** model used on training dataset to find optimal subset of attributes. Once the attributes are selected, they used **multiple regression** to predict price and find co-relation between attributes. For some seeds the mean error was high in case of regression tree an multiple regression leads to less accuracy. This model predicts with the accuracy 95%.

**Chen, C., Hao, L., & Xu, C., May 2017 [3]** built price prediction model by using 5-fold cross validation to show the effects using linear regression and random forest with three factors algorithm, number of samples and explanatory variables. In this paper they proposed comparative analyses of multiple models. Model based on car make is not suitable as it used linear regression which is only suitable for those car manufacturing year which trades in large volume(more sample size). Model based on specific type of car series used 13 explanatory variable with 4 type of car series which founds out having good enough sample size the random forest performs better than linear regression. Another model called universal which used 19 explanatory variable predicts price with better accuracy in case on random forest. The challenge this paper faced is with hardware which takes hours to establish apart from that sample size used in this paper is small so price accuracy can be increased. This paper only limited its implementation to two algorithms in every model.

**Samruddhi, K., & Ashok Kumar, R., Aug 2020 [4]** proposed car price prediction model using K-Nearest neighbour. In this paper they proposed model using K-Nearest Neighbour algorithm and its get cross validated using k fold method to overcome overfit issue. Once the data preprocessing part is done, they used KNN algorithm to be train on different k values. For instance, with 3 different ratios of k values leads to 85% accuracy Root-Mean Squared Error and Mean Absolute Error rate changes with value of k. Due to inconsistency of price accuracy while selecting train and test dataset cross-validation is used to inspect overfitting. After cross-validation on 5 and 10 folds they found better accuracy of 82% for 10 folds with k value of 4. Overall accuracy of 85% is far better than linear regression. This paper only compared its accuracy with one algorithm it lacks multiple comparison with other algorithms.

**L. D'costa, A. W. D’Souza, A. k, and D. M. Varghese, Jan 2020 [5]** proposed Multiple Linear Regression model to predict true value used car. Once the data collection from multiple resources with sample size of 1870 part is done the procedure and data analyses takes place using square root transformation to find the correlation between mileage and price. In order to fitting relationship between more than 2 variables they used multiple linear regression and used R-square and adjusted R-square are main reasons of accuracy checking. There is residual standard error because of the attributes which are not considered. This model found linear relationship among variables. Additionally, they used NCV test to check any heteroscedasticity among residuals. This model claims the accuracy of 89.33%. This paper have not considered other attributes for less residuals and not considering combination of other algorithms.

All of them are showing different level of accuracy of prices the main thing was the features they were considering while building models due to which the outputs were not satisfactory.

# Proposed Methodology

The project deals with used cars in Canada. Using Python Scraper, the dataset from autotrader.ca and autotempest.com was scraped in order to build an effective intelligent model. The proposed solution also includes a user interface where users can enter input and can get desired results. Figure 1 below explains the different steps involved in developing the solution.

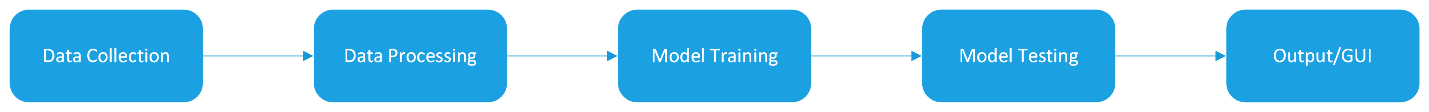


Figure 1: Proposed Model

**Data Collection:** A Python scrapper is written using BeuatifulSoup to scrape car listing URLs and then Selenium to scrape the actual listing on the website to get the following features: Mileage, Condition, Manufacturer, Model year, Car type, Engine and drive type. The collected data then is stored in MongoDB.

**Data preprocessing**: The input data is cleaned where the name and the brand on of the car is extracted from the title. The outliers are removed. The input features are then generalized for example: the transmission values ‘4-speed-automatic’ and ‘8-speed-automatic’ are both converted to ‘automatic’. The categorical values were then converted to numerical values by replacing each unique value in every columns with a number and storing those numbers in ‘column.json’ file.

After the conversion to numerical values correlation was used to check the effect of different feature on the Price and irrelevant features were dropped namely: 'MTW Status', 'Stock Number', 'Status', 'Exterior Colour'.

**Model Training:** Using ‘train\_test\_split’ from sklearn the data is split in 90-10 train-test split.

The data is then normalized using ‘StandardScaler’. The model used for training is a sequential neural network, the first layer is a dense layer with 128 neurons. The activation function used is ReLU (Rectified Linear Unit), which returns the maximum between 0 and the input value. The second and third layer consists of 64 and 32 neurons respectively with the same activation function of ReLU. The output layer consists of a single neuron with a linear activation function. The model is then compiled with the ‘mean squared error(mse)’ loss function and adam optimizer to minimize the loss. The model is then trained on the training set for 100 epochs, with a batch size of 50.

**Model Test:** The trained model is used to predict the values for the test set which is compared with the actual values of test and r2 is calculated.

**Output/GUI:** The webpage contains the user interface for users to interact. The Webpage is developed using ReactJs. It has a single webpage where a user can add the year, model, mileage, and other factor and click on the search button. After searching the database and evaluating the model, the value is predicted by the model which will be displayed to the user.

Figure 2 shows the architecture of the prosed solution using MVC Structure.

Graphical user interface, diagram

Description automatically generated

Figure 2: Process Architecture

The architecture process uses MVC structure. The View includes the user Application or the user interfaces where user can add input the values and see the result. The request to fetch the data goes to the server controller and is processed to check valid request. The inputs are then passed to Neural Network model whioch predicts the price and return the value. The Python scrappers runs to fetch the data from websites and scraped data is stored in MongoDB which later is used to train the Neural Network model.

Overall, the proposed methodology involves collecting and preprocessing data, selecting relevant features, developing a neural network model, training and evaluating the model, fine-tuning the model's hyperparameters, and deploying the model for practical use.

# Result

Theoretically, all the other models were already tried out in the research papers that we reviewed. We then planned to go with the Artificial neural network model for our project as outcomes from all the other models were already listed in the research paper.

By cleaning and preprocessing our data and calibrating our model, Figure 3 below shows the data before outlier removal and Figure 4 shows the data after outlier removal which highlights the importance of data preprocessing.

Chart, scatter chart

Description automatically generated

Figure 3: Before Outliers removal

Chart, scatter chart

Description automatically generated

Figure 4: After Outliers removal

The model achieved an overall accuracy of 78% with the limited data.

The aim for the future is to increase the data and train the model with a larger dataset which would increase accuracy further.

Figure 5 shows the webpage for entering the input value. User can enter up to 9 factors (Model Name, Brand Name, Model Year, Fuel, Drive Type, Transmission, Car Type, Engine Cylinder, Mileage) which will predict the price. After Enter the input in dropdown and Number inputs user can click on Predict Price Button which will fetch data from server.

Graphical user interface

Description automatically generated

Figure 5: User Interface for Input Values

A screenshot of a car

Description automatically generated with medium confidence

Figure 6: Use Interface for Price Prediction Result

Figure 6 shows the retrieved predicted price from the server and neural Network Model.

Figure 7 shows the comparison from the actual listing from the website. It can be seen that the predicted price was around $75000 and the actual price of different listing varies from $70,000 - $80,000. This shows the accuracy of the model.

Graphical user interface

Description automatically generated

Figure 7: Actual Listing from Website

# Limitations or Challenges

Initially, while connecting to the server there were handshake problems with the server. There were also latency problems since multiple machines were trying to scrape the data, also causing a delay in server's response.

The script ran to scrape around 200k rows for several days on several machines. But the data got corrupted while ingesting it into the database. After rerunning it, only 28 thousand rows were retrieved. Due to the ingestion error, the range of car models was vastly reduced. The efficiency of the model could be improved in the future with the availability of more data.

While neural networks are a powerful tool for many machine learning tasks, they can have limitations when it comes to predicting used car prices. Some potential limitations include:

* **Lack of data**: Neural networks require large amounts of data to train effectively. However, when it comes to predicting used car prices, this project has used smaller datasets.
* **Complex features**: Used car prices can be influenced by a wide range of factors, including the car's age, mileage, condition, location, and features. It is challenging to capture all of these factors accurately in a neural network model. We have used some of these factors and some can be included in future scope.
* **Non-linear relationships**: Used car prices can be influenced by non-linear relationships between features. For example, a car's value may decrease rapidly as it reaches a certain age or mileage threshold. Capturing these non-linear relationships accurately is challenging for neural networks.

Overall, while neural networks can be effective for predicting used car prices, they may not always be the best choice depending on the available data, complexity of features, and need for interpretability. The above challenges make it harder to predict the accurate value. Certain limitations can be resolved such as smaller datasets and factors correlation by using accurate measures.

# Conclusion and Future Work

Based on the results of the neural network model for predicting used car prices, it can be concluded that the model is effective in predicting car prices with a reasonable level of accuracy. The model was trained in a dataset of car features and their corresponding prices and was able to learn patterns and relationships in the data to make accurate predictions on new, unseen data. An efficient neural network machine learning model is built by training, testing, and evaluating the dataset. As a result of pre-processing and transformation, an accuracy of 78% was achieved.

However, it is important to note that the accuracy of the model may be impacted by factors such as the quality and completeness of the data used to train the model, as well as any limitations or biases in the model architecture. Therefore, it is recommended to continue to monitor and evaluate the performance of the model over time, and to adjust and refine the model as needed to ensure its continued effectiveness in predicting used car prices.

In the future, more data will be collected using efficient web scrapping techniques and deep classifiers will be tested. More models like SVM, ANN, and Quantile Regression will be tested to analyze accuracy. Future work will also include using the model prediction for the suggesting best car for the user.

# References

1. “Canada Used Car Dealers Retail Sales,” ycharts.com. https://ycharts.com/indicators/canada\_used\_car\_dealers\_retail\_sales#:~:text=Canada%20Used%20Car%20Dealers%20Retail%20Sales%20is%20at%20a%20current
2. Ganesh, Mukkesh & Venkatasubbu, Pattabiraman. (2019). Used Cars Price Prediction using Supervised Learning Techniques. International Journal of Engineering and Advanced Technology. 9. 216-223. 10.35940/ijeat.A1042.1291S319.
3. Chen, C., Hao, L., & Xu, C. (n.d.). *Comparative analysis of used car price evaluation models*. Retrieved February 28, 2023, from https://aip.scitation.org/doi/pdf/10.1063/1.4982530
4. Samruddhi, K., & Ashok Kumar, R. (2020). Used car price prediction using K-nearest neighbor based model. *International Journal of Innovative Research in Applied Sciences and Engineering*, *4*(2), 629–632. https://doi.org/10.29027/ijirase.v4.i2.2020.629-632
5. L. D'costa, A. W. D’Souza, A. k, and D. M. Varghese, “Predicting true value of used car using multiple linear regression model,” *International Journal of Recent Technology and Engineering*, vol. 8, no. 5S, pp. 42–45, 2020.