CS235 Fall'23 Project Final Report: Use Deep Learning to Predict Car Sales Price

Liam Hsieh
University of California, Riverside
Data Science, MSOL
862395843
lhsie013@ucr.edu

Abstract

This project leverages Artificial Neural Networks (ANNs) to predict car sales prices, comparing their performance against linear regression. The selected ANN model, with three hidden layers (5-10-5) and a linear activation function, outperforms linear regression in Mean Absolute Error. The project follows a systematic flow, encompassing data preprocessing, model exploration, and final evaluation. Visualizations support the model selection, contributing valuable insights to the automotive industry

Keywords: Car sales price prediction, Neural Networks, Mean Absolute Error (MAE)

ACM Reference Format:
Liam Hsieh. 2023. CS235 Fall'23 Project Final Report: Use Deep Learning to Predict Car Sales Price. In *Proceedings of Data Mining Techniques (CS 235)*. ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The proposed project aims to leverage Deep Learning to predict car sales prices based on a dataset obtained from Kaggle [1]. The dataset consists of 500 records with 9 columns, encompassing customer details such as name, email, country, gender, age, annual salary, credit card debt, net worth, and car purchase amount. The objective is to explore the potential of ANN models in predicting car sales prices compared to a baseline model using linear regression. Linear regression will serve as the benchmark model, and various ANN structures will be evaluated to identify the most suitable and accurate predictive model. By harnessing the capabilities of ANN, we aim to enhance the accuracy and precision of car sales price predictions, providing valuable insights for the automotive industry and aiding customers in making informed purchasing decisions.

[1] https://www.kaggle.com/datasets/yashpaloswal/ann-car-sales-price-prediction

CS 235, Fall 2023, University of California, Riverside

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

2 Related Work

The prediction of car prices using machine learning has attracted considerable attention. Cars, being both ubiquitous and essential commodities in the modern world, possess well-known features familiar to most people. The pricing of cars shows a clear correlation with several features, contributing to the popularity of applying machine learning for predicting car prices. One prevalent application involves forecasting the retail or invoice price of cars based on features and dealer information, especially in the context of used cars [1, 3].

During the COVID-19 pandemic, with a shortage of vehicles in the market, the prices of used cars experienced a significant boost. This surge has led many individuals to seek scientific references to aid in decision-making. Consequently, there is a growing demand for machine learning-based methods to address this need.

Given the strong linear relationship between features and price, linear regression has proven to be effective in predicting car prices [2, 4].

3 Proposed Approach

The proposed approach involves initially splitting the dataset into a training set (80%) and a testing set (20%). The core of the project revolves around leveraging Artificial Neural Networks (ANNs) to predict car sales prices. The ANN architecture comprises an input layer representing selected features, multiple hidden layers activated using linear functions, and an output layer for regression, employing the mean absolute error loss. The Keras library, built on TensorFlow, will be utilized to implement the ANN model. Specifically, the output layer will employ a linear activation function to predict numeric car purchase amounts. Different activation functions, e.g., *relu*,*sigmoid*, and *linear*, number of hidden

layers, and optimizers, e.g., *adam*,*sgd*, and *nadam*, including will be explored and compared to determine the most effective model configuration. The ANN model will be compiled using a defined optimizer and loss function, and subsequently trained on the training set. Evaluation of the ANN's performance will be conducted using the testing set. After a screening experiment for different optimizers, number of neurons, and activate functions, we further applied cross validation to determine the potential of adding additional hidden layers from just two to five and to select the most accurate and efficient model for car sales price prediction.

The model we eventually used for the prediction is described as follows:

- three hidden layers.
- number of neurons for each hidden layers are 5, 10, and 5, sequentially.
- the activate function is linear for each layer.
- optimizer is adam.
- number of epochs for model training is 80 and batch size is 64.

Fig. 1 is the lost plot for the proposed ANN model.

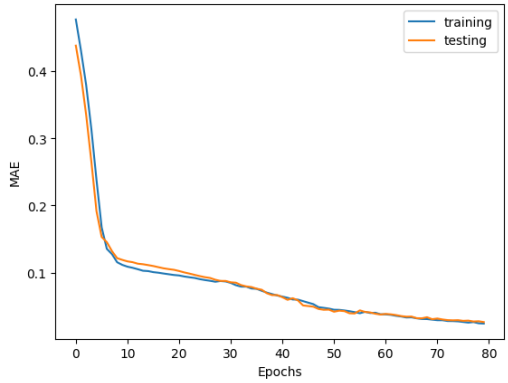


Figure 1. ANN training loss vs testing loss plot

The high-level flow for proposed approach is shown as Fig. 2. The presented flowchart outlines the sequential steps in a data-driven predictive modeling process. It begins with raw data, progresses through data cleaning and preprocessing, including MinMax normalization. The dataset is then split into training and testing sets with an 80:20 ratio. Two main branches of modeling are pursued: one involves Linear Regression as baseline, and the other conducts a pre-screening with 81 distinct Artificial Neural Network (ANN) models to narrow down the exploration then determine the number of hidden layers later. The final step involves a comprehensive evaluation of model performance using Mean Absolute Error, providing a comparative analysis between Linear Regression and various ANN models.

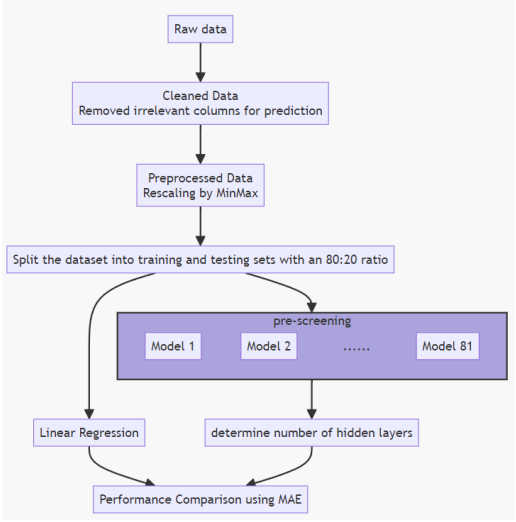


Figure 2. Project Overview.

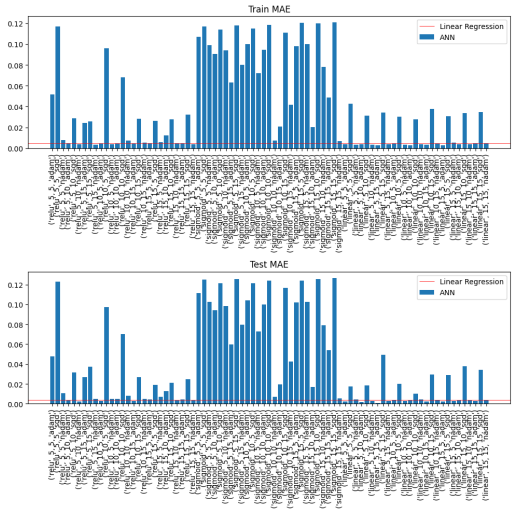


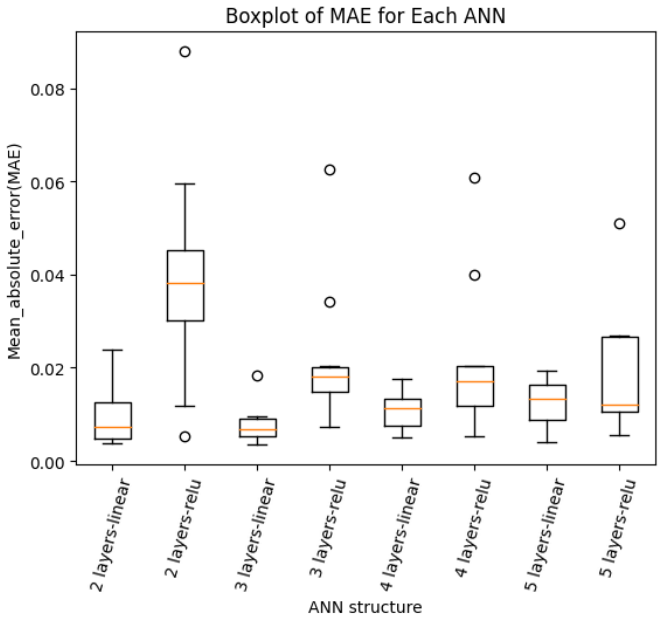
Figure 3. Output Overview

The results of pre-screening have been illustrated by Fig. 3 it helps narrow down those low performance model configuration rapidly before we moved to determine the number of hidden layers.

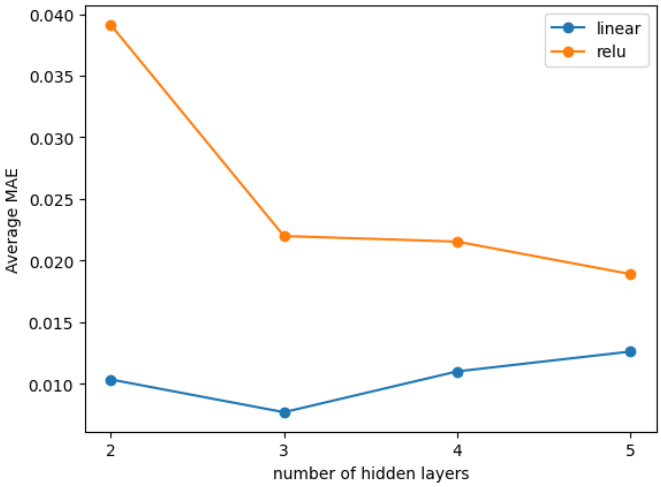
Moreover, Fig. 4a and Fig. 4b support our decision of selecting the 3-layers-linear as our prediction model because it not only has the lowest average MAE but also has smallest standard deviation of MAE.

4 DISCUSSION and CONCLUSIOINS

This paper proposed a multi-layers artificial neural network model to predict car sales price and implement from scratch using Keras. Comparing to the performance of linear regression, the most common prediction method



(a) Boxplot of different ANN models



(b) Mean MAE of ANN Models with Varying Number of Hidden Layers

Figure 4. Comparison of ANN models.

in this kind of problem due to strong linearity, the proposed model can outperform linear regression by mean absolute error.

References

[1] Muhammad Asghar, Khalid Mehmood, Samina Yasin, and Zimal Mehboob Khan. 2021. Used cars price prediction using machine learning with optimal features. *Pakistan Journal of Engineering and Technology* 4, 2 (2021), 113–119.

[2] Lucija Bukvić, Jasmina Pašagić Škrinjar, Tomislav Fratrović, and Borna Abramović. 2022. Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning. *Sustainability* 14, 24 (2022), 17034.

[3] Chejarla Venkata Narayana, Nukathoti Ooha Gnana Madhuri, Atmakuri NagaSindhu, Mulupuri Aksha, and Chalavadi Naveen. 2022. Second Sale Car Price Prediction using Machine Learning Algorithm. In *2022 7th International Conference on Communication and Electronics Systems (ICCES)*. IEEE, 1171–1177.

[4] MUTİ Sumeyra and Kazım YILDIZ. 2023. Using Linear Regression For Used Car Price Prediction. *International Journal of Computational and Experimental Science and Engineering* 9, 1 (2023), 11–16.