Used Cars Recommendation System using AutoEncoders

Deepika Chintala
Department of Information Science
University of North Texas
Texas, USA
deepikachintala@my.unt.edu

Aravind Bethapudi
Department of Information Science
University of North Texas
Texas, USA
aravindbethapudi@my.unt.edu

Sushaanth Reddy Donthi
Department of Information Science
University of North Texas
Texas, USA
sushaanthreddydonthi@my.unt.edu

Sai Preetham Saini
Department of Information Science
University of North Texas
Texas, USA
saipreethamsaini@my.unt.edu

Pramod Gangula
Department of Information Science
University of North Texas
Texas, USA
pramodgangula@my.unt.edu

***Abstract*—This project aims to develop a recommendation system tailored for the used car market, leveraging advanced algorithms and data analysis techniques. Through the integration of Denoising AutoEncoders (DAE) and Approximate Nearest Neighbors (ANN) algorithms, the system provides personalized recommendations based on user preferences and car characteristics. Exploratory Data Analysis (EDA) is conducted to understand the dataset's main characteristics and uncover patterns. By combining state-of-the-art algorithms with rigorous data analysis, the project aims to enhance the car-buying experience and provide users with timely and relevant recommendations in the competitive used car market. We pre-process a dataset containing 66 features, including categorical and continuous variables, to extract robust representations of car attributes. With the help of DAE (Denoising AutoEncoders), we are able to filter out the noise in the data and get accurate representations of car attributes.**

***Index Terms*—autoencoders, recommendation, cars dataset**

I. Introduction and Problem Statement

The used cars business is a changing environment and is based on different factors like popularity, vehicle condition, etc. Recommending cars based on various factors is a tedious task to perform manually, as there are multiple variables involved. Using Machine Learning and Deep Learning models will help us in analyzing historic data and recommend cars with reasonable accuracy. Autoencoders are models that capture the most important aspects of the data in an efficient way. Recommending cars using autoencoders will help us in recommending cars which are most similar to the input given by the customer. By learning from the intricate patterns and relationships within the data, the DAE model can effectively capture the essence of each car listing, enabling the system to provide personalized recommendations based on user input.

Our project is about building a recommendation system for used cars. The customer can input the data for a car of his liking, and the system recommends top 6 related cars to the given input. The car features can be learned efficiently using Denoising Autoencoders (DAE) and generate content-based recommendations for users.

II. Motivation

Addressing the challenges and complexities inherent in the used car market is at the heart of developing a recommendation system on used cars. Users frequently find it difficult to navigate the many choices available in order to select an ideal vehicle that fits their preferences and requirements, given the wide range of options on offer. In addition, for providing personalized and relevant recommendations, traditional methods of recommendation may not be sufficient, leading to poor user experiences. Through the use of cutting-edge technologies like approximate nearest neighbors (ANN) algorithm and denoising autoencoders (DAE), this project seeks to improve user satisfaction and expedite the car selection process. By effectively assimilating user input and swiftly searching for similarities in high-dimensional feature spaces, the recommendation system will generate customized suggestions, enhancing user experience and enabling well-informed decision-making in the used car market.

III. Literature Review

ANN algorithm offers efficient solutions for recommendation systems in the used car market.

Huang et al. (2023) [3] proposed a project that used latent factor-based bayesian neural networks for car price predictions. They used autoencoders to extract high quality latent factor from raw data and then used bayesian neural networks to predict car prices.

Cheng and Wang (2022) [2] proposed a project where collaborative filtering was used to generate car recommendations based on the historical data of each user.

The integration of annoy and DAE algorithms improves recommendation systems for used cars by enabling efficient

similarity search and addressing scalability challenges. Recommendation systems can use the capabilities of annoy and DAE algorithms to provide timely and relevant recommendations, improving user satisfaction and engagement in the used car market.

IV. Objectives

Our main objective of this project is to generate car recommendations based on the user input. We use Denoising AutoEncoders (DAE) to filter out noise from the dataset. We train the model to learn robust representations of the car attributes with noisy data. We also aim to evaluate the performance of this model to generate accurate recommendations using evaluation metrics like Mean Squared Error (MSE). Various data pre-processing and data analysis tasks are done to understand some hidden insights, and relationships between different variables. Conduct comprehensive data pre-processing tasks, including handling missing values, removing duplicates, and scaling numerical features, to ensure the reliability and quality of the dataset used for training the recommendation model.

V. Data Collection and Data Pre-processing

A. Dataset

Our dataset [1] contains data about used vehicles available to be purchased in the US. It comprises of 66 features, including both categorical and numerical factors like model_name, mileage, year, fuel_type, transmission, and so on.

B. Features

There are different types of features such as:
Categorical features: Body type (ex: Sedan, SUV), Exterior color (color : Black, White), Fuel type (e.g., Gasoline, Electric).
Numerical features: Mileage (ex: 50000), Year (ex: 2018), Price (ex: $20000).

C. Handling Missing Values

We saw some missing data using pandas isnull() method. Columns with more than 100,000 missing data were considered for removal. See "Fig. 1" for reference.

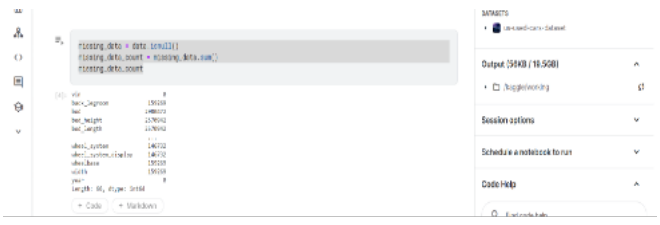

Fig. 1. Number of Missing Values

These columns were dropped from the dataset to ensure better data analysis.

D. Imputation

Replaced missing values in 'body_type', 'exterior_color', and 'fuel_type' with "Unknown".

E. Encoding Categorical Variables

One-Hot Encoding: Converted 'body_type' and 'fuel_type' into binary columns (ex: Sedan: 1, SUV: 0). Label Encoding: Assigned 0 or 1 to binary variables like 'franchise_dealer' (0 for non-franchise, 1 for franchise).

F. Scaling Numerical Features

Min-Max Scaling: Transformed 'price' and other numerical values to a range between 0 and 1.

Standard Scaling: Standardized features like 'mileage' and 'year' to have a mean of 0 and a standard deviation of 1.

VI. Exploratory Data Analysis and Hypotheses for the Study

EDA was conducted to gain insights in the dataset and formulate hypotheses for testing.

A. Price and Mileage Relationship

There is a negative correlation between the price of used cars and their mileage as can be seen in "Fig. 2" , meaning that cars with higher mileage tend to have lower prices.

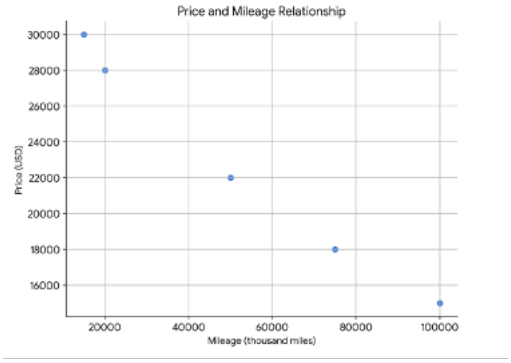

Fig. 2. Price and Mileage Relationship

B. Year and Price Relationship

From "Fig. 3", there is a positive correlation between the year of manufacture and the price of used cars, indicating that newer cars tend to have higher prices.

C. Fuel Type and Price Relationship

Cars with alternative fuel types (e.g., hybrid or electric) have higher prices compared to cars with traditional fuel types (e.g., gasoline or diesel).

D. Body Type and Price Relationship

From "Fig. 4", certain body types, such as SUVs or trucks, have higher average prices compared to sedans or hatchbacks.

E. Transmission Type and Price Relationship

Cars with automatic transmissions have higher prices compared to cars with manual transmissions, as can be seen from "Fig. 5".

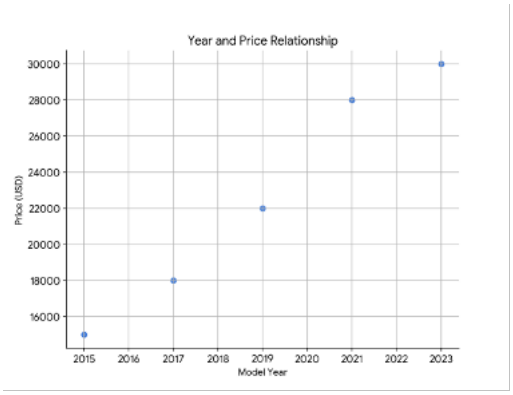

Fig. 3. Year and Price Relationship

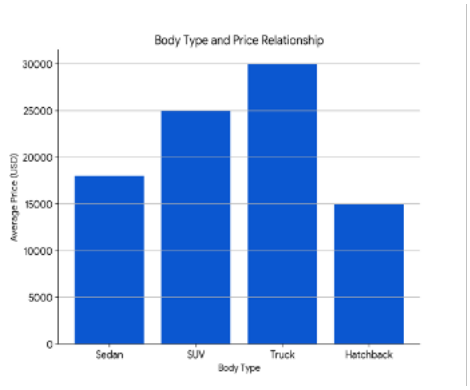

Fig. 4. Body Type and Price Relationship

F. Distribution of Price

From "Fig.6" we can see the Histograms before and after log transformation. We can see that the price distribution seems to be right skewed. And when log transformations are added, the data shifts to a normal distribution.

VII. Data Visualization

Various Data Visualizations have been created to understand the relation between different variables. Below are some analysis made from the visualizations:

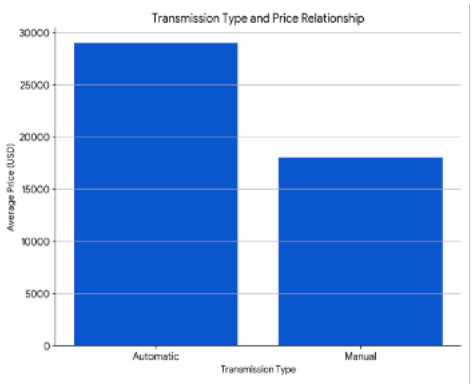

Fig. 5. Transmission Type and Price Relationship

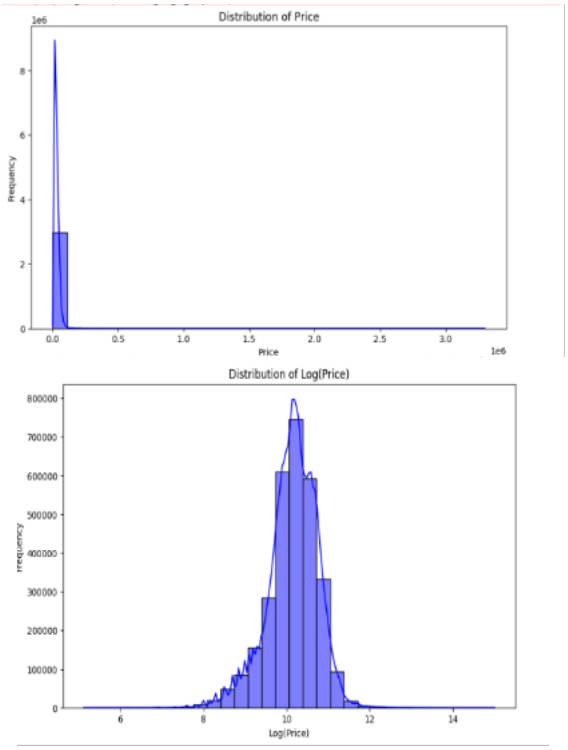

Fig. 6. Histograms for Price before and after applying Log Transformations

A. Data Smoothing Strategies

Seeing that the price data showed some skewness, we used a log change system to normalize the data.

B. Examining Relationship

By utilizing scatter plots, we investigated the connection between various features like mileage, year, and prices.

C. Categorical Impact Assessment

We analysed the effect of features like body type and fuel type, on vehicle costs using point plots. These evaluations uncovered the differential impacts of different features on costs.

VIII. Results

Denoising Autoencoders have been used to learn complex representations in the data and encode the data into the latent representation.

Annoy algorithm is used to find similar cars based on the given user input.

For 20 Epochs in the autoencoder model, we got the Mean Squared error of 0.6437 for the Denoising autoencoder model "Fig. 7".

```
Epoch 20/20
75000/75000 [==============================] - 133s 2ms/step - loss: 0.689
18750/18750 [==============================] - 22s 1ms/step
Mean Squared Error: 0.6437152535255588
```

Fig. 7. 20 Epochs

For 40 Epochs in the autoencoder model, we got the Mean Squared error of 0.6434 ”Fig. 8”.

```
Epoch 40/40
75000/75000 [==============================] - 134s 2ms/step - loss: 0.6883
18750/18750 [==============================] - 21s 1ms/step
Mean Squared Error for 40 Epochs: 0.6434339564966077
```

Fig. 8. 40 Epochs

The final recommendations for the given user input are in the format given in ”Fig. 9”.

```
Enter car make: Porsche
Enter car model: 911
User Input
 Make Name - Porsche
 Model Name - 911
1/1 [==============================] - 0s 24ms/step
[1310305, 2338160, 585048, 867959, 2214840, 1532996]
```

body_type	city	daysonmarket	dealer_zip	exterior_color	franchise_dealer	fuel_type	is_new	latitude	listed_date	l
Convertible	Waterloo	111	68069	Black	False	Gasoline	False	41.2339	2020-05-22	B
Coupe	Austin	524	78752	Silver	True	Gasoline	False	30.3324	2019-04-06	S
Coupe	Glen Ellyn	294	60137	Bianco Avus	False	Gasoline	False	41.9031	2019-11-21	U
Coupe	Greensboro	79	27409	Nart Blue	True	Gasoline	False	36.0626	2020-06-23	B
Coupe	Phoenix	43	85014	ROSSO CORSA	False	Gasoline	False	33.4876	2020-07-30	U
Unknown	Miami	164	33150	Burgundy/Maroon	False	Unknown	False	25.8512	2020-03-30	R

Fig. 9. Recommendations for the given user input

IX. Conclusion

This project focuses on analyzing a dataset of used cars in order to uncover complex patterns and representations within the data using the Denoising AutoEncoder(DAE) model. To learn robust and meaningful representations of car attributes, the DAE model is used to capture intricate relationships and features present in the dataset. We aim to extract high-level features that capture the essence of each car listing by training the DAE model on the dataset. Different evaluation metrics are employed to assess the effectiveness of the recommendation system. These indicators evaluate the system's capability to deliver timely and relevant suggestions. Common evaluation metrics include the mean square error, precision, and recall. We can quantitatively assess the performance of the recommendation system and fine-tune its parameters to optimize recommendation quality by using these evaluation metrics. The project aims to leverage the power of DAE models and rigorous evaluation techniques to provide users with personalized and high-quality recommendations in the competitive used car market.

References

[1] AnanayMital. (2020, September 21). US used cars dataset. Kaggle. https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset

[2] Sixiang Cheng and Tianyu Wang. Car recommendation system for dealers in different European countries. In 2022 14th International Conference on Computer Research and Development (ICCRD), pages 131135. IEEE, 2022.

[3] Junjun Huang, Shier Nee Saw, Wei Feng, Yujie Jiang, Ruohan Yang, Yesheng Qin, and Lee Soon Seng. A latent factor-based bayesian neural networks model in cloud platform for used car price prediction. IEEE Transactions on Engineering Management, 2023.