# Predicting used car prices in Northern America

**Abstract-** **This project investigates the use of Convolutional Neural Networks (CNNs) for predicting used car prices in North America. The study explores the integration of unsupervised autoencoders for feature extraction to assess their influence on supervised model performance. By evaluating CNN models under different configurations, the findings highlight the potential of advanced techniques like autoencoders in shaping predictive capabilities and demonstrate the impact of parameter tuning on model accuracy and efficiency. We will then use a pre-trained state-of-the-art model to evaluate how it addresses our problem.**

### 1. Introduction

Predicting car selling prices requires analyzing a range of factors such as a vehicle’s age, mileage, condition, and other attributes. This study leverages advanced machine learning techniques to develop accurate predictive models. Key tasks included:

1. **Developing and Evaluating CNN Models:** Implemented Convolutional Neural Networks (CNNs) to predict car prices based on numerical and categorical features.
2. **Exploring Unsupervised Autoencoders:** Utilized autoencoders for feature extraction to assess their influence on improving model efficiency and predictive accuracy.
3. **TabTransformer**

In this project, the state-of-the-art TabTransformer architecture was used to handle the tabular car price dataset. A TabTransformer-inspired model was implemented from scratch, while a pre-trained TabTransformer model was leveraged for transfer learning and fine-tuned to optimize performance for our specific dataset.

1. **Evaluating Results with Key Metrics:** Analyzed model performance using metrics Mean Absolute Percentage Error (MAPE), and R² Score.

The dataset, sourced from Kaggle [1], consists of 558,811 entries with a mix of categorical and numerical features, providing a robust foundation for model training and evaluation. This study aims to uncover insights into the predictive capabilities of CNNs and the potential role of autoencoders in shaping the feature space.

### 2. Dataset and Preprocessing

1. Dataset Characteristics:

**Size:** 558,811 entries, 16 columns.

**Features:** Numerical (e.g., year, odometer) and categorical (e.g., make, model, transmission).

**Challenges:** Missing values in critical columns such as make and model.

1. Preprocessing Steps:

**Handling Missing Values:** Dropped rows with missing critical attributes.

**Feature Engineering:** Created derived features (e.g., age) and label-encoded categorical data.

**Normalization:** Scaled numerical attributes using *MinMaxScaler*.

**Data Splitting:** 80% training, 20% testing.

### 3. Methodology

This study employs a systematic approach to predict car prices using Convolutional Neural Networks (CNNs) and unsupervised autoencoders. The methodologies are outlined below:

#### Supervised Learning: Convolutional Neural Network (CNN)

The CNN architecture for regression includes:

* **Convolutional Layers:** Extract patterns from numerical and categorical features.
* **Batch Normalization:** Stabilizes and accelerates training.
* **Dropout:** Reduces overfitting by deactivating neurons randomly during training.
* **Dense Layers:** Outputs the predicted selling price.

The models were configured with varying parameters:

* **Filters**: 64, 32, 128
* **Kernel Size**: 3, 5
* **Dropout Rate**: 0.1 to 0.3
* **Learning Rates**: 0.001, 0.0005

Training was conducted using the Adam optimizer with Mean Squared Error (MSE) as the loss function, for 50 epochs and a batch size of 32. Early stopping with a patience of 10 epochs was applied to prevent overfitting.

#### Unsupervised Learning: Autoencoders for Feature Extraction with CNN

* **Filters:** [32, 16, 8] across layers for hierarchical feature extraction.
* **Kernel Size:** 3, enabling fine-grained feature learning.
* **Encoding Dimension:** Reduced the feature space to one-third of the original input size.
* **Learning Rate:** 0.001 for stable training.
* **Epochs:** 50, with early stopping based on validation loss.
* **Batch Size:** 64 for efficient model updates.
* The encoded output from the autoencoder was reshaped and used as input for a CNN model. This transfer learning approach was compared to CNNs trained directly on the original dataset.

### 4. Results

#### A. Key Findings

1. A. CNN Configurations Evaluation

|  |  |  |
| --- | --- | --- |
| **Configuration** | **MAPE** | **R² Score** |
| Filters: 64, Kernel: 3, Dropout: 0.2, LR: 0.001 | 38.82% | 0.7961 |
| Filters: 32, Kernel: 5, Dropout: 0.3, LR: 0.0005 | 45.95% | 0.6951 |
| Filters: 128, Kernel: 3, Dropout: 0.1, LR: 0.001 | 36.24% | 0.7934 |

The configuration with 64 filters, a kernel size of 3, a dropout rate of 0.2, and a learning rate of 0.001 achieved the best results.

1. Impact of Feature Extraction:

**Model Performance**

The Transfer\_Learning\_Model achieved an R² score of 0.2322, indicating limited variance explained compared to models trained on the original dataset.

#### **Purpose of Autoencoder**

The autoencoder aimed to compress features and streamline the input for the CNN, enhancing model efficiency.

#### **Observations**

The compressed features retained less critical variance, leading to reduced predictive performance. This highlights the trade-off between dimensionality reduction and information retention.

#### **Future Directions**

Refining the autoencoder architecture or combining it with hybrid feature selection methods could improve feature retention and predictive outcomes.

### 5. Contribution

1. Tejinder (301232634): Conducted EDA and prepared the preprocessing pipeline. Designed and implemented CNN models for supervised learning.
2. Matheus (301236904): Developed and trained Autoencoder for feature extraction. Evaluated the impact of extracted features on model performance.
3. Bernice (301284811): Experimented with pre-trained models for transfer learning. Conducted performance comparisons between custom and pre-trained models.
4. Thejus (301301078): Visualized insights from EDA and model evaluations. Compiled results into the final report and coordinated group contributions.

### 6. Conclusion

This study highlights the significance of parameter tuning in CNNs for car price prediction. While unsupervised autoencoders offer potential for feature reduction, their application in this context worsened performance. Future work could explore advanced feature engineering or incorporate more sophisticated architectures like Transformers.

### References

[1] Kaggle, "Used Car Auction Prices Dataset," [Online]. Available: <https://www.kaggle.com/datasets/tunguz/used-car-auction-prices>. [Accessed: Dec. 8, 2024].