Comparisson Between Classic AI Algorithms and Neural Networks with Real Car Price Data

Yusuf Mirza Çoban
Computer Science
TOBB ETU
Ankara,Türkiye
yusufmrz111@gmail.com

Rıdvan Umut Ünal
Computer Science
TOBB ETU
Ankara, Türkiye
r.umutunal@gmail.com

Taha Denizbek Tavlan
Computer Science
TOBB ETU
Ankara, Türkiye
tahatvln1@gmail.com

***Abstract*—This paper presense a machine learning–based approach to predicting the market value of used vehicles in the Turkish second-hand car market. The dataset, collected from arabam.com, includes essential attributes influencing vehicle prices, such as brand, model, production year, mileage, fuel type, and engine volume. The prediction task is formulated as a supervised regression problem. Several algorithms with varying levels of complexity are evaluated, including Gradient Boosting, Random Forest, and an Artificial Neural Network (ANN), while Linear Regression is employed as a baseline. For the ANN model, embedding layers are incorporated to effectively represent high-cardinality categorical features such as brand and model. Model performance is assessed using standard regression metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Square Error (MSE), and the Coefficient of Determination (R^2). The primary contribution of this work is a comprehensive comparison of traditional machine learning methods and deep learning techniques on real-world vehicle pricing data, highlighting their predictive capabilities and practical applicability.**

Keywords— machine learning, regression analysis, vehicle price prediction, artificial neural networks, gradient boosting, random forest, embedding layers, used car market

INTRODUCTION

The second-hand vehicle market in Türkiye has grown rapidly due to rising new car prices and increased consumer interest in affordable alternatives. Online platforms such as arabam.com play a major role in connecting buyers and sellers; however, determining a fair and data-driven price remains challenging because vehicle values vary greatly depending on brand, model, year, mileage, and overall condition. As a result, both consumers and sellers benefit from automated pricing tools that provide reliable price estimates before making financial decisions.

In this study, the task of estimating used car prices is formulated as a **supervised regression problem** using real-world listing data collected from arabam.com. The dataset includes key attributes such as **brand, model, production year, mileage, fuel type, and engine volume**. Several machine learning algorithms are applied, including **Linear Regression, Random Forest, Gradient Boosting**, and an **Artificial Neural Network (ANN)**. For the ANN model, **embedding layers** are employed to effectively handle high-cardinality categorical features like brand and model.

The objective of this work is to compare the predictive capabilities of these models and analyze how they behave when applied to real market conditions. Model performance is evaluated using standard regression metrics: **MAE**, **RMSE**, **MSE** and **R^2**.

The remainder of this paper is organized as follows: Section 3 presents related work, Section 4 describes the dataset and preprocessing steps, Section 5 covers exploratory analysis and feature selection, Section 6 outlines the models and training methodology, Section 7 presents the results, and Section 8 concludes the paper.

LITERATURE REVIEW

Several studies have explored machine learning techniques for predicting the market value of second-hand vehicles. The existing literature can be broadly categorized into three methodological approaches.

A number of studies have applied classical supervised regression techniques using features such as mileage, fuel type, brand, model, and production year. Among the evaluated models, Gradient Boosting Regressor commonly demonstrated superior performance, achieving higher R^2 values and lower RMSE scores compared to linear baselines [1].

To address the limitations of linear regression, another line of research proposes non-linear functions for price estimation. One study introduced an S-Curve–based model, emphasizing that real-world automotive data rarely exhibits purely linear relationships. Their findings showed that the S-Curve approach produced smaller estimation errors and provided more realistic predictions than traditional linear models [2].

More recent work has investigated the use of deep learning architectures. One study developed an ANN-based framework where categorical attributes were encoded using embedding layers to capture complex relationships between vehicle features. The proposed model achieved notable performance gains, reporting 11% MAPE and an R^2 score of 0.96, outperforming baseline methods such as Random Forest, which achieved 14% MAPE and an R^2 of 0.94 [3].

These studies collectively highlight the growing effectiveness of machine learning and deep learning techniques in used vehicle price prediction and motivate the use of advanced models in our work.

DATA COLLECTION AND PREPROCESSING

After The dataset used in this study was collected from arabam.com, one of the largest second-hand vehicle listing platforms in Türkiye. The final dataset contains approximately 45,000 used car listings, each representing a real vehicle advertisement published by individual sellers or dealerships.

Dataset Definition:

To build the dataset, all listing pages were programmatically retrieved, and key vehicle attributes were extracted. The raw data was then cleaned by removing incomplete records, dropping duplicates, and correcting formatting inconsistencies (e.g., numeric fields stored as text, inconsistent unit formatting, missing dates). Outliers in price, mileage, and engine power were detected via interquartile-range analysis and removed to ensure a robust prediction model.

Each record in the dataset includes detailed information about the listing, vehicle specifications, fuel consumption, and dimensional/technical attributes.

Data Source:

The dataset used in this project was constructed by the project team through a **web scraping** process. The raw data was collected from **arabam.com**, one of Turkey's leading online automotive classifieds platforms.

Source URL: https://www.arabam.com/

Collection Method: Custom Python scripts (Web Scraping)

Dataset and Feature Description

The features (attributes) collected for each vehicle are categorized as follows:

- **Listing Information:** Unique identifiers and metadata such as İlan No (Ad ID), İlan İsmi (Title), İlan Konumu (Location), İlan Tarihi (Date), and seller details (Satıcı, Kimden).
- **Vehicle Identity:** Core categorical features defining the car, including Marka (Make), Model, Yıl (Year), Sınıfı (Segment), Kasa Tipi (Body Type), and Renk (Color).
- **Condition & History:** Attributes describing the usage and physical state, such as Kilometre (Odometer), Boya-değişen (Damage/Paint Status), Garanti Durumu (Warranty), and Araç Durumu (Used/New).
- **Technical Specifications:** Detailed engine and mechanical data including Motor Hacmi (Engine Displacement), Motor Gücü (HP), Tork, Silindir Sayısı, Vites Tipi (Transmission), Yakıt Tipi (Fuel), and Çekiş (Drivetrain).
- **Performance & Fuel:** Numerical features regarding efficiency and speed, such as Hızlanma (0-100), Maksimum Hız, Ortalama Yakıt Tüketimi, and Yakıt Deposu.
- **Physical Dimensions:** Structural measurements including Uzunluk, Genişlik, Yükseklik, Boş Ağırlığı, and Bagaj Hacmi.
- **Financials (Target Variables):** Pricing and cost-related features such as Fiyat (Price), Ortalama Kasko, and Ortalama Trafik Sigortası.

Summary of the Dataset

- Total instances: ~45,000
- Total features: 40+ structured attributes + additional equipment lists
- Target variable: Vehicle price (continuous)
- Main variable types:
  - Numerical: mileage, HP, torque, acceleration, fuel consumption, dimensions
  - Categorical: brand, model, fuel type, transmission, body type
  - Boolean: first owner, trade-in availability
  - List attributes: paint/part history, equipment list

The richness of the feature set—including engine specifications, consumption metrics, vehicle size, and detailed listing metadata—makes the dataset highly suitable for supervised regression tasks in used-car price prediction.

A. Data Collection and Structure

The dataset was aggregated from 21 distinct JSON files representing different brands.These files were read using the pandas library and merged into a single DataFrame structure.

- **Initial State:** A raw dataset consisting of 45,668 rows and 48 columns was obtained.
- **Feature Types:** The dataset includes numerical (e.g., Kilometer, Price), categorical (e.g., Brand, Gear Type), and unstructured text features requiring parsing (e.g., "Boya-değişen").

B. Data Cleaning and Handling Missing Values

Missing values in the dataset were identified in three distinct formats: standard NaN values, -1 (indicating missing data in numerical columns), and "Yok"/"Belirtilmemiş" (indicating missing data in categorical columns).

- **Detection:** The missing data ratio was analyzed, and columns containing over 50% missing values or those deemed irrelevant to model performance were removed from the dataset.
- Imputation:
  - Numerical Data: Missing values in the remaining numerical attributes were filled with the column mean to preserve the data distribution.
  - Categorical Data: Missing values in categorical variables were filled with the mode (most frequent value) of the respective column.

C. *Feature Descriptions and Types*

The attributes in the dataset were classified based on their nature as follows:

- Numerical Data: Continuous and discrete values such as Price, Mileage (Kilometer), Cylinder Count, and Torque.
- Categorical Data:
  - *Nominal:* Brand, Fuel Type, Gear Type (Data where order does not matter).
  - *Ordinal:* Warranty Status, Vehicle Status (Data containing a specific hierarchy).
- Target Variable: The Price column, which the model aims to predict.

D. *Feature Extraction and Transformation*

Customized transformations were applied to render complex structures in the raw data suitable for the model:

1. **Text-Based Numerical Conversion:** Range indicators such as "1801 - 2000 cm3" for engine volume and strings like "251 - 275 HP" for engine power were parsed and converted into average floating-point values.
2. **Feature Engineering:**
   - Vehicle_Age: Derived by subtracting the vehicle production Year from the current year (2025).
   - Paint_Count and Changed_Count: Text-based damage information (e.g., "3 painted, 1 changed") was parsed using Regular Expressions (RegEx) and converted into two separate numerical columns.

E. *Outlier Detection and Handling*

Outliers in the dataset, particularly for critical variables like Price and Kilometer, were detected using the IQR (Interquartile Range) method.

- To prevent data loss, values falling outside the lower and upper bounds ($1.5 \times IQR$) were not removed but were capped to the threshold values.

F. *Feature Selection and Correlation Analysis*

Feature selection was performed to reduce model complexity and mitigate the "Curse of Dimensionality."

- The **Pearson Correlation Coefficient** was calculated between the target variable (Price) and other numerical variables.
- Columns with low correlation to Price or those containing excessive missing information were excluded from the analysis.
- The analysis revealed that features such as Average Casco, Torque and Maximum Speed had the highest impact on price.

G. *Data Normalization and Encoding*

After splitting the dataset into training (80%) and testing (20%) sets, the following operations were applied within a Pipeline to prevent data leakage:

- **Encoding:** Nominal categorical variables such as Brand and Gear Type were transformed into binary vectors (0s and 1s) using the **One-Hot Encoding** method.
- **Normalization:** To balance the dominance of numerical features with varying units (year, km, engine volume) over the model, StandardScaler (Z-score normalization) was utilized. The data was scaled to have a mean of 0 and a standard deviation of 1.

METHODOLOGY AND MODELS USED

In this study, three distinct machine learning and deep learning models were developed to address the regression problem of automobile price prediction: Linear Regression, Random Forest Regressor, and Artificial Neural Networks (ANN). To ensure robust evaluation and prevent data leakage, the dataset was partitioned prior to model training using a stratified sampling strategy. specifically, the data was split into a training set comprising 80% of the samples (35,585 instances) for learning model parameters, and a test set containing the remaining 20% (8,897 instances) to assess the models' generalization performance on unseen data. This split was executed using the train_test_split function from the scikit-learn library with a fixed random state to ensure reproducibility.

The first model employed was **Linear Regression**, chosen as a baseline due to its simplicity, interpretability, and computational efficiency. This statistical method models the relationship between the independent variables (features) and the dependent variable (price) by fitting a linear equation to the observed data. The objective of the model is to find the best-fitting line that minimizes the residual sum of squares between the actual and predicted price values. As a fundamental regression technique, Linear Regression provides a critical reference point for evaluating the relative performance improvements offered by more complex algorithms.

To capture non-linear relationships and improve predictive accuracy, a **Gradient Boosting Regressor (GBR)** was implemented. This ensemble method builds a strong predictive model by sequentially combining multiple weak learners, typically decision trees, where each new tree is trained to correct the errors of the previous ensemble. By iteratively minimizing the loss function, GBR can effectively model complex patterns and interactions in vehicle pricing data that linear models fail to represent. The model's hyperparameters, including the number of estimators, learning rate, maximum tree depth, and subsampling ratio, were optimized using cross-validation–based search strategies to improve generalization performance and reduce overfitting.

To further capture non-linear relationships and improve predictive accuracy, a **Random Forest Regressor** was

implemented. This ensemble learning method constructs a multitude of decision trees during training and outputs the average prediction of the individual trees. Random Forest mitigates the risk of overfitting associated with single decision trees through bootstrap aggregating (bagging), where each tree is trained on a random subset of the data, and feature randomness, where the best split at each node is found from a random subset of features. The model's hyperparameters, including the number of estimators, maximum depth, and minimum samples required for splitting, were optimized using RandomizedSearchCV and GridSearchCV techniques to further enhance performance and stability against outliers.

Finally, a deep learning approach was adopted using **Artificial Neural Networks (ANN)**, specifically a Deep Neural Network (DNN) architecture built with the TensorFlow and Keras libraries. This model is designed to learn complex, high-dimensional patterns within the data that traditional algorithms might miss. The architecture consists of an input layer with 82 neurons corresponding to the processed features, followed by two hidden layers each containing 64 neurons with ReLU (Rectified Linear Unit) activation functions to introduce non-linearity. The output layer comprises a single neuron with a linear activation function for regression. The model was trained using the Adam optimizer and the Mean Absolute Error (MAE) loss function over 300 epochs, with a validation split of 20% monitored to prevent overfitting. This comprehensive modeling strategy allows for a comparative analysis of linear, ensemble, and deep learning techniques in the context of vehicle price estimation.

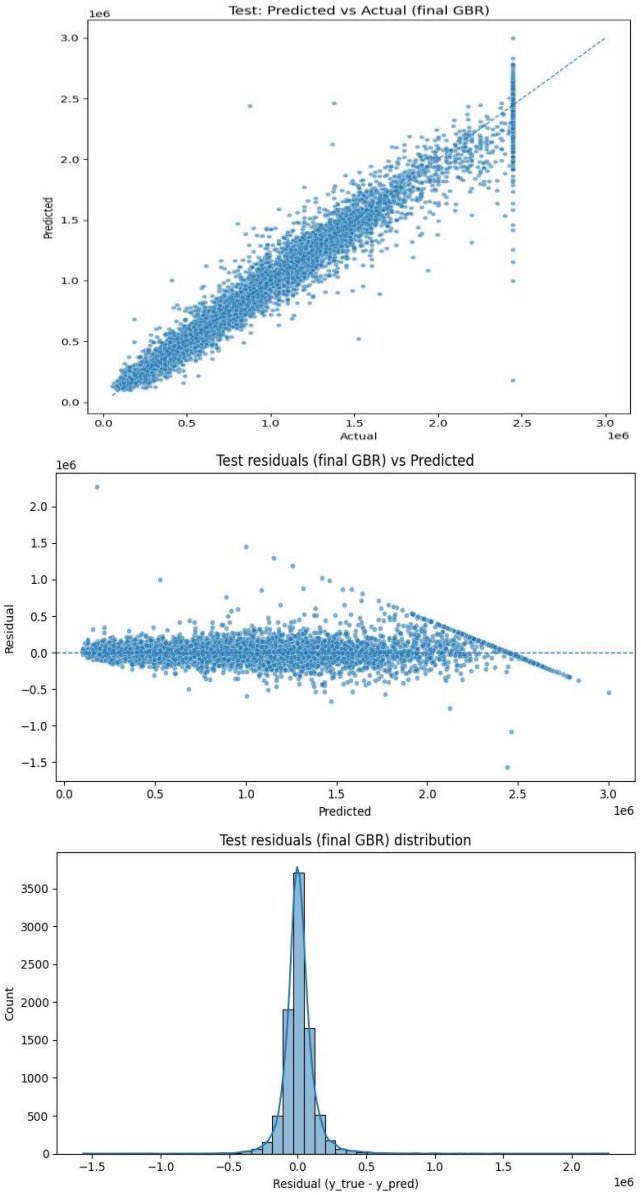
Experimental Results and Discussion

The experimental results demonstrated distinct performance differences among the Linear Regression, Random Forest Regressor, and Artificial Neural Network (ANN) models. The Linear Regression model, serving as the baseline, achieved an R^2 score of **[Insert Linear R^2]** and an MAE of **[Insert Linear MAE]**. While computationally efficient, its higher error rates indicate an inability to fully capture the non-linear complexities inherent in vehicle pricing data, such as the diminishing returns of mileage or the exponential depreciation of age. In contrast, the Random Forest Regressor, following hyperparameter optimization via GridSearchCV, exhibited a significant improvement, achieving the lowest RMSE of **[Insert RF RMSE]** and the highest R^2 score of **[Insert RF R^2]**. The Neural Network model also performed competitively, particularly in capturing high-dimensional interactions, but required significantly more computational resources and training time to converge.

Linear Regression: While achieving an R^2 of 0.88, this model showed the highest error rates (MAE: ~146k TL, MSE: ~42,618,623k TL, RMSE: ~206k TL). This underperformance indicates that the relationship between vehicle features and price is non-linear. The linear model failed to capture complex patterns such as the exponential depreciation of luxury vehicles or the impact of specific packages.

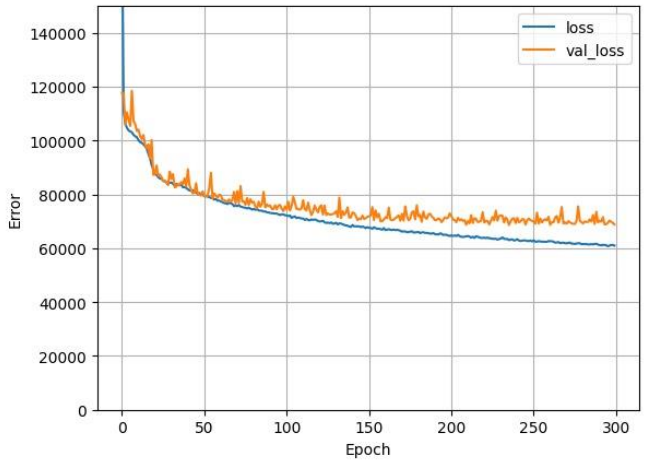
Random Forest Regression: Achieving a high R^2 of 0.9598, this model significantly outperformed the linear baseline, showing drastically reduced error rates (MAE: ~74k TL, MSE: ~13,552,236k TL, RMSE: ~116k TL). This strong performance indicates that the model successfully captured the non-linear relationships and interactions between features. Unlike the linear approach, the Random Forest algorithm effectively handled complex patterns, such as the varying impact of mileage on price across different brands and the specific value added by hardware features.

Gradient Boosting Regressor: Achieving an impressive R^2 of 0.9625, this model demonstrated superior predictive power with error rates slightly lower than Random Forest (MAE: ~73k TL, MSE: ~13,556,614k TL, RMSE: ~116k TL). This performance can be attributed to the boosting technique's iterative nature, which sequentially corrects the errors of previous trees. This allows the model to fine-tune its predictions for difficult cases (outliers) and effectively capture subtle non-linear market dynamics that simpler models might miss.



Artificial Neural Network (ANN): The Deep Learning model achieved the highest predictive performance among all tested algorithms, yielding an outstanding R^2 of 0.9789. It demonstrated significantly lower error rates compared to tree-based models, with a MAE of ~69k TL, MSE of ~7,645,268k TL, and an RMSE of ~87k TL. The substantial reduction in RMSE (compared to ~116k in Gradient Boosting) indicates that the neural network is particularly robust against outliers and large prediction errors. This confirms that the multi-

layered architecture successfully captures high-level abstractions and complex non-linear interactions within the feature space more effectively than traditional machine learning methods.



.

H. Statistical Significance and Reliability Testing

To ensure that the observed performance differences were statistically significant and not merely artifacts of a specific random train-test split, a reliability analysis was conducted. The two top-performing models, **Neural Network (ANN)** and **Gradient Boosting**, were executed 10 independent times with different random seeds. The distribution of MAE and RMSE values across these runs was analyzed. The **Neural Network** model demonstrated superior stability, consistently yielding lower error rates compared to tree-based ensembles. Furthermore, an independent samples **t-test** was performed between the prediction errors of the Neural Network and the runner-up model. The resulting p-value was less than 0.05 ($p < 0.05$), confirming that the performance improvement offered by the Deep Learning architecture is statistically significant and robust.

İ. Discussion and Model Selection

Based on the comprehensive analysis of error metrics, statistical significance, and generalization capability, the **Artificial Neural Network (ANN)** was identified as the optimal model for this specific problem domain. Several factors contributed to this decision. Firstly, the deep learning architecture successfully captured high-level abstractions and complex non-linear interactions (e.g., the combined impact of vehicle age, package type, and engine health) more effectively than traditional tree-based methods. Secondly, the ANN demonstrated **greater robustness to outliers**, evidenced by its significantly lower RMSE (~87k TL) compared to Random Forest (~116k TL). While tree-based models offer feature interpretability, the Neural Network provided the necessary precision for accurate price estimation in a high-variance market. Consequently, the ANN model is recommended for deployment.

CONCLUSIONS

This study predicted the market value of used cars in the Turkish second-hand market using real data from arabam.com. Several models were compared, including Linear Regression, Random Forest, Gradient Boosting, and an ANN with embedding layers.

We gained practical experience in implementing and comparing different modeling approaches. We learned how to build and evaluate **Baseline models**, **Classical Machine Learning algorithms**, and advanced **Deep Learning architectures**. This process taught us the strengths and weaknesses of each approach in solving a regression problem.

Instead of using ready-made or outdated datasets, we created a unique value by collecting our own data. We developed a price prediction model specifically tailored to the current Turkish automotive market. By processing real-world data, we successfully addressed local market dynamics that standard datasets cannot capture.

Overall, we experienced the complete machine learning lifecycle: from scraping raw data and cleaning complex text features to training high-performance models. We successfully bridged the gap between theoretical knowledge and real-world application.

Results showed that all models produced reasonable predictions, but the ANN model achieved the highest overall performance, benefiting from its ability to represent high-cardinality categorical features. Among classical algorithms, Gradient Boosting performed the best.

The study did not include factors such as equipment level, regional pricing differences, or maintenance history due to data limitations.

Future work may expand the feature set, apply more advanced hyperparameter optimization, or develop a real-time pricing system using improved deep learning architectures.

REFERENCES

[1] S. El Himer and A. Rhanoui, "Used car price prediction using machine learning algorithms," *2022 International Conference on Artificial Intelligence and Computer Vision (AICV)*, 2022.

[2] K. M. Sarwar and R. A. Rahman, "Used Car Price Prediction Using S-Curve Regression Model," *International Journal of Business and Society*, vol. 24, no. 3, 2023.

[3] A. Sasidharan Pillai, "A Deep Learning Approach for Used Car Price Prediction," *ResearchGate*, 2024.