Project Report

Taravat Ashabi, Mukesh Mohan, Shahin Rajabi, and Tanmaya Rodda

April 7th, 2024

AIT 736 - Applied Machine Learning

**Linear Regression Model: Predict Used Car Prices**

1. **Introduction and Problem Formulation (Motivation, goals, challenges)**

The "Used Car Price Predictions" dataset provides valuable information for predicting the prices of 58 different used cars models based on various features. The target variable, Price, represents the price of the used car being sold. There are various factors that contribute to the value of a used car, which are the variables provided in this dataset. Those key attributes that can influence the pricing, such as the Year of purchase, Mileage (the number of kilometers driven), and the City and State where the car was sold. Each car is identified by its unique Vehicle Identification Number), and the dataset includes information about the Manufacturer (Make) and the specific Model of the vehicle.

By analyzing this dataset, machine learning models can be trained to learn the complex relationships between these features and the used car prices. Such models could be valuable for various stakeholders, including car dealers, buyers, and sellers, to estimate fair market prices for used vehicles accurately. The objective is to develop a linear regression model that can accurately predict the selling price (Price) of used cars based on the provided features, such as Year, Mileage, City, State, Make, and Model. The linear regression model should establish a linear relationship between the target variable (Price) and the predictor variables (Year, Mileage, City, State, Make, Model).

The linear regression model aims to identify the significant predictor variables that have a substantial impact on the used car prices. Evaluate the model's performance using appropriate metrics, such as mean squared error (MSE), root mean squared error (RMSE), or R-squared, to assess its predictive accuracy. Identify any outliers or influential observations that may be affecting the model's performance. Interpret the model coefficients and provide insights into how changes in the predictor variables (e.g., Year, Mileage, Make, Model) influence the predicted used car prices.

1. **Dataset Description**

Below is a sample of our dataset:

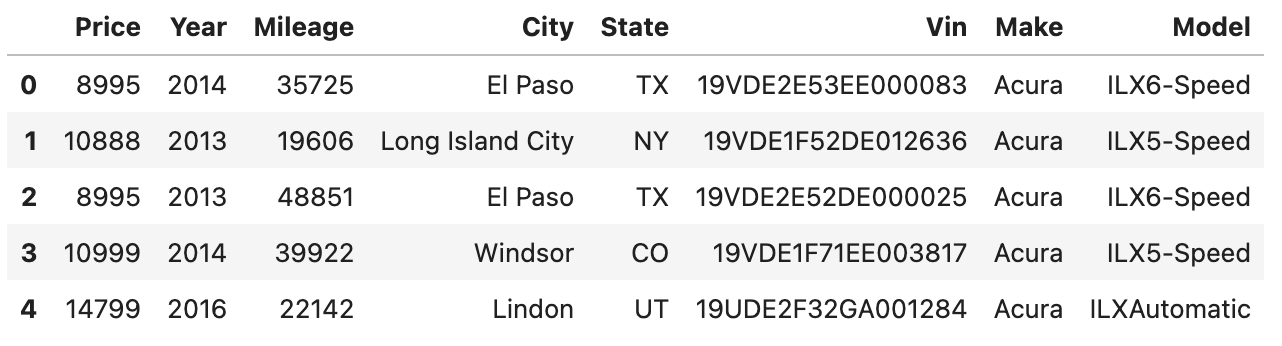


Figure 1. First five rows of the true\_car\_listings dataset

Our dataset, which is sourced from Kaggle, comprises listings of used cars, featuring approximately 850,000 entries. It covers a comprehensive range of information for each vehicle, including the price, year of manufacture, mileage, and location (city and state). Additionally, details such as the car's make, model, and unique identifier (VIN) are provided [3].

The data was collected from numerous sources, including automobile dealership lists, internet car sales platforms, and direct inputs from car vendors. The variety of brands, models, years, and geographic areas indicates an in-depth collection effort that spans a large market sector.

Given the lack of apparent missing values and placeholders, the dataset indicates that it has undergone cleaning processes. Duplicate listings have been removed, the Make, Model, City, and State fields' formats were standardized, and the accuracy of Vehicle Identification Numbers (VINs) has been verified [3].

Missing data was not an issue in our dataset, and the model was reverted after we removed the outliers from the numerical columns, as this did not affect the model's evaluation metrics.

A screenshot of a computer

Description automatically generated

Figure 2. true\_car\_listings dataset does not have null values

We partitioned our dataset with an 80:20 ratio (80% training and 20% testing sets) by using the train\_test\_split function. This function split our dataset into two parts: one for training our linear regression model and the other for testing its performance. After performing the split, 80% of the data is used for training the model, ensuring it learns to predict the target variable accurately. The remaining 20% serves as the testing set to evaluate the model's performance on unseen data.

1. **Develop your Model (Clearly explain what you did here completely and include your results)**

We chose to run a linear regression algorithm on the true\_car\_listings.csv dataset to build a model that predicts the price of cars. To build our model we used Python [6] via Google Colab and Jupyter Notebook [5]. Jupyter Notebook was accessed through Anaconda Navigator [1].

The data were read using the pandas library [7]. During the exploratory and preprocessing stages, the data were separated into numerical columns and categorical columns. Since the ‘Vin’ column is the unique identifier for each record, it was dropped. The category\_encoders library [9] was used to target and encode the categorical features and the sklearn library [2] was used to scale the numerical features. The interquartile range method was used to remove outliers from the numerical columns. This step did not affect the model’s evaluation metrics, so it was removed. Cross-validation was used to explore various feature combinations and select the best combination that yielded the highest average R-squared score. The best feature combination was found to be X = ['Year', 'Mileage', 'Make', 'Model'], with an average R-squared score of 0.7398.

After selecting the best feature combination, the sklearn library was used to conduct the following analysis. The data was split into 80% training and 20% testing sets. The linear regression model was created and trained with the training data and selected features. It was then used to make predictions on the testing data. The model was evaluated with the mean squared error (MSE) and R-squared metrics. Our evaluation results are an MSE of 46205175.65 and an R-squared value of 0.749.

1. **Evaluation Results (Run your model on a test set and use evaluation metrics explained in class to measure the correctness of your classifier—e.g., use confusion matrix, ROC, F1-score so forth).**

Our research has led us to the identification of a number of important predictor variables that have a big impact on used automobile prices. These factors—year, mileage, make, and model—are crucial in figuring out how much a pre-owned car is worth on the open market. We want to shed light on the intricate dynamics behind used automobile market pricing by including these variables in our predictive model.

Our model's performance indicators highlight how well it captures the volatility of used car values while reducing prediction mistakes. Interestingly, the R-squared, or coefficient of determination, is a respectable 0.749. This suggests that the predictor factors in our model account for about 74.9% of the variation in used automobile pricing. Our model's accuracy and reliability are further supported by the fact that its mean squared error (MSE), of 46205175.65, indicates a comparatively low degree of prediction error.

Our linear regression model uses important characteristics like the make, model, mileage, and year of purchase to accurately forecast used car pricing. Our model's excellent R-squared value and low mean square error (MSE) demonstrate its durability, which gives us confidence that it may offer insightful information to both customers and industry professionals involved in the used automobile market.

The following visual was created with the matplotlib and seaborn libraries [4] [8].

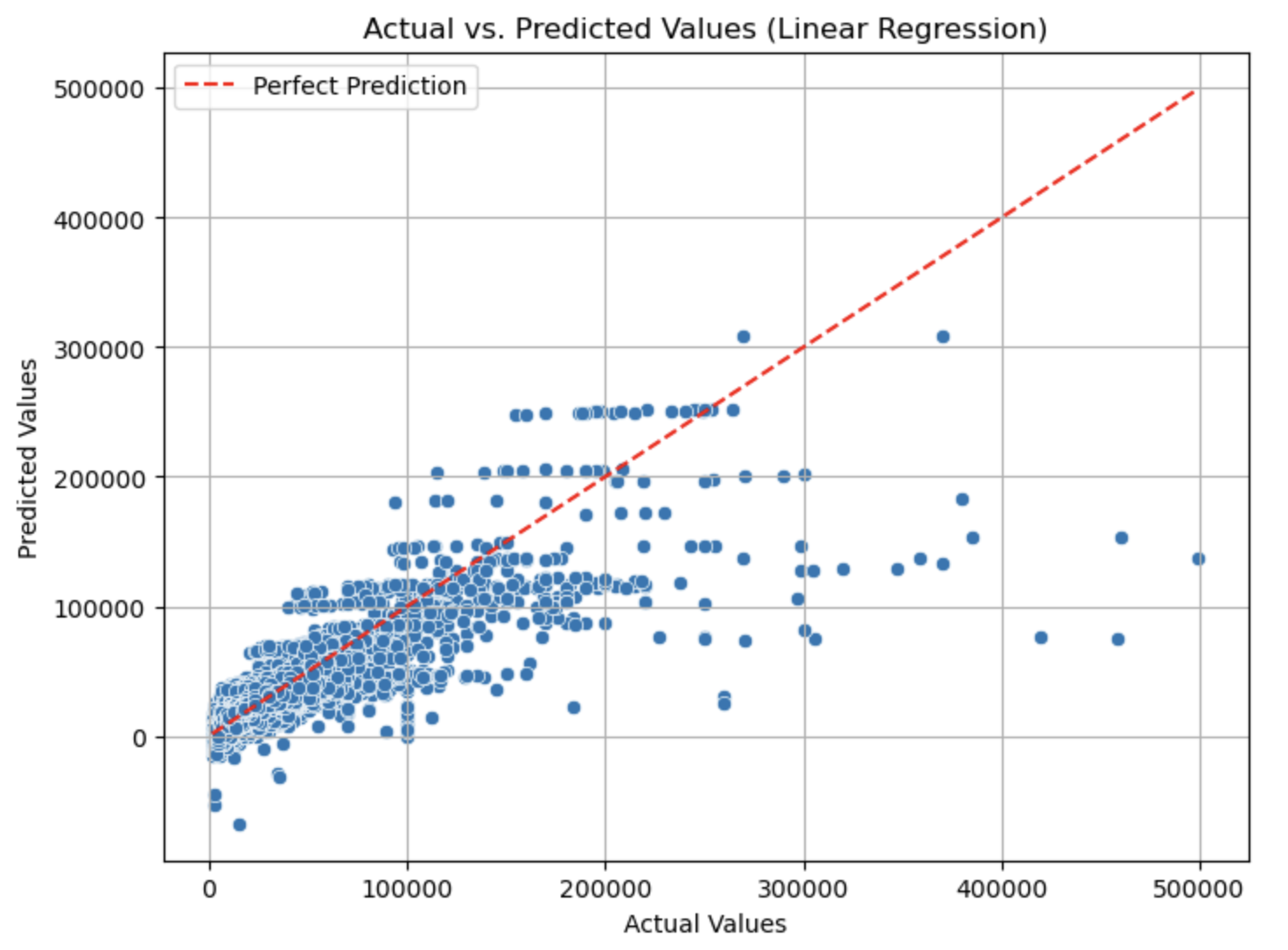


Figure 3. Visual of the results of the linear regression model's predicted values vs. the actual price values

In conclusion, Our linear regression model is a solid foundation that makes it easier to estimate fair market values for used cars, giving both buyers and sellers accurate information. Our approach deftly captures the complex correlations between different parameters and used car pricing by using machine learning techniques. This makes it possible for decision-making processes involving stakeholders to be more informed.

Ensuring the model's resistance against any outliers and influential observations has been made possible by meticulous study and preprocessing of the dataset. We have strengthened the model's trustworthiness by protecting its performance from unwarranted impacts by taking care of these data points.

Out of all the features taken into account, the combination of year, mileage, make, and model turns out to be especially useful for pricing prediction. This emphasizes how important these elements are in the ever-changing used car industry.

Furthermore, analyzing the insights obtained from interpreting the model coefficients provides important details on how changes in predictor variables affect the projected prices of used cars. For vehicle dealers, buyers, and sellers alike, this information can be quite helpful in helping them make well-informed judgments about negotiations and pricing.

To sum up, our linear regression model provides a practical and effective way to forecast used car prices, which benefits different parties involved in the automotive sector. Its successful implementation highlights the long-lasting effectiveness of this conventional but potent method in handling the intricacies involved in used car price prediction.

Moreover, our model has the potential to be a useful tool that is used on internet platforms, car dealerships, and online markets with additional integration and refining. It stands to expedite transactions and promote fair dealing within the market by offering quick and precise pricing estimations for second-hand autos.

**References**

[1] *Anaconda Software Distribution* (Version 2-2.4.0). (2016). Anaconda.<https://www.anaconda.com>

[2] Avijeet Biswal. (2023, April 3). *Sklearn Linear Regression*. Simplilearn.Com.<https://www.simplilearn.com/tutorials/scikit-learn-tutorial/sklearn-linear-regression-with-examples>

[3] HARIKRISHNAREDDYB. (2022, April 20). *Used car price predictions*. Kaggle. <https://www.kaggle.com/datasets/harikrishnareddyb/used-car-price-predictions/data>

[4] J. D. Hunter. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.<https://doi.org/10.1109/MCSE.2007.55>

[5] Project Jupyter. (2022). Project Jupyter/JupyterLab (Version 3.4.4). Retrieved from<https://jupyter.org/>

[6] Python Software Foundation. (2024). Python Language Reference (Version 3.12.2). Retrieved from<https://www.python.org/>

[7] The pandas development team. (2020). *pandas-dev/pandas: Pandas* (2.0.2). Zenodo.<https://doi.org/10.5281/zenodo.3509134>

[8] Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021, <https://doi.org/10.21105/joss.03021>.

[9] Will McGinnis, Scott Hendrickson, & Eric Hambro. (2018). category\_encoders: A collection of scikit-learn compatible transformers for encoding categorical variables into numeric. Retrieved from<https://github.com/scikit-learn-contrib/category_encoders>