**Predicting the Manufacturer’s Suggested Retail Price (MSRP) of Cars Using Various Regression Models**

Student’s Name

Professor’s Name

Institutional Affiliation

Course Code

Due Date

**Introduction**

The manufacturer’s suggested retail price (MSRP) is a crucial element in the automotive industry. Accurate MSRP predicting models provide manufacturers, dealers, and customers foundational decision support. In modeling, this research uses purged Linear Regression, K-nearest Neighbors (KNN), or Decision Tree regression to estimate future car MSRPs. It assesses the effectiveness and accuracy of those models to improve them for the improvement of accurate MSRP predictions. This will help enhance effective pricing and, hence, the purchase decisions of the different organizations across the industry.

**Linear Regression**

Linear regression is a method of analysis that predicts the value of a dependent variable based on the value of one or more independent variables using a straight line (James etal, 2023). The main aim is to utilize information about the table to try and predict or estimate the value of a dependent variable using independent variables.

**K-Nearest Neighbors (KNN)**

KNN is a data-driven and model-free learning technique for categorizing classes (classification) or predicting continuous values (Regression). It is about detecting patterns in the training set and then predicting based on the similarity of such patterns in new, unseen data. KNN calculates the value of the new data’s point by using universal averaging for ’k’ closest neighbors in the feature space (Zhang & Li, 2021). The selection of ‘k’ and the distance function form part of the model’s design, determining its performance standards.

**Decision Trees**

Decision Trees are supervised learning algorithms used for classification and regression tasks (Lee etal, 2022). They divide data into partitions according to the input variables' value, forming a tree-like structure. A node represents an attribute, a branch represents a decision, and a terminal node represents an outcome. Cross-validation is a process in which the data are split into a training and testing set, used to check against the performance of the model.

**Methodology**

**Data Preprocessing**

This study used a dataset involving numerous car attributes, including the kind of engine, type of transmission, driven wheels, number of doors, market class, size, style, highway mileage per gallon, city mileage per gallon, and level of popularity. Data cleaning involved handling missing values through data imputation techniques for numerical features, mean imputation for numerical variables, and feature engineering to break down categorical variables through one hot encoding method. The missing values were managed by replacing them using the mean of each feature, while outliers were defined and filtered out as data points three or more standard deviations away from the mean.

**Feature Selection**

Backward elimination was applied to select the most significant features. We started with an initial set of features and iteratively removed the least considerable feature based on p-values until all remaining features had p-values less than 0.05.

**Model Training and Tuning**

Three regression models were trained, and the models included Linear Regression K-Nearest Neighbors and Decision Trees. Therefore, to compare the performance of each of the created models, Mean Absolute Error (MAE), Mean Squared Error (MSE), and coefficient of determination or R-squared (R²) were calculated. These algorithms had their hyperparameters tuned using cross-validation of GridSearchCV.

**Results and Discussion**

**Model Performance Before and After Tuning**

**Linear Regression Results:**

* Before Tuning: MAE = 4626.13, MSE = 36606305.48, R-squared = 0.7486
* After Tuning (Ridge Regression): MAE = 4621.40, MSE = 36618834.88, R-squared = 0.7485

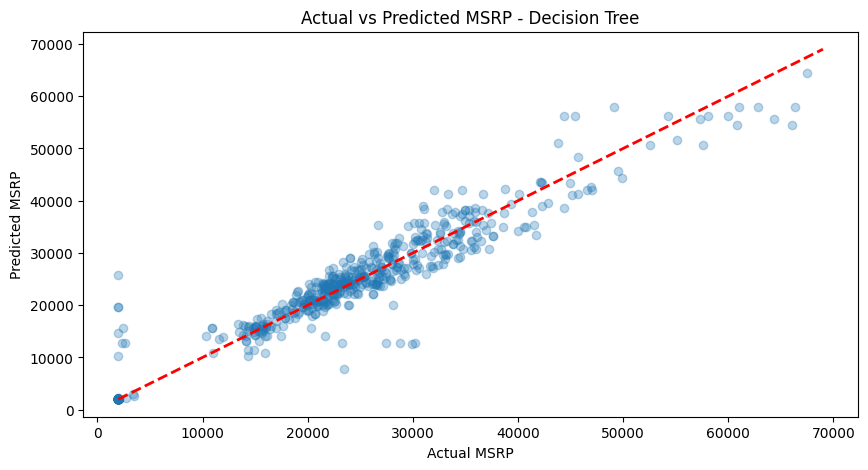
**K-Nearest Neighbors Results:**

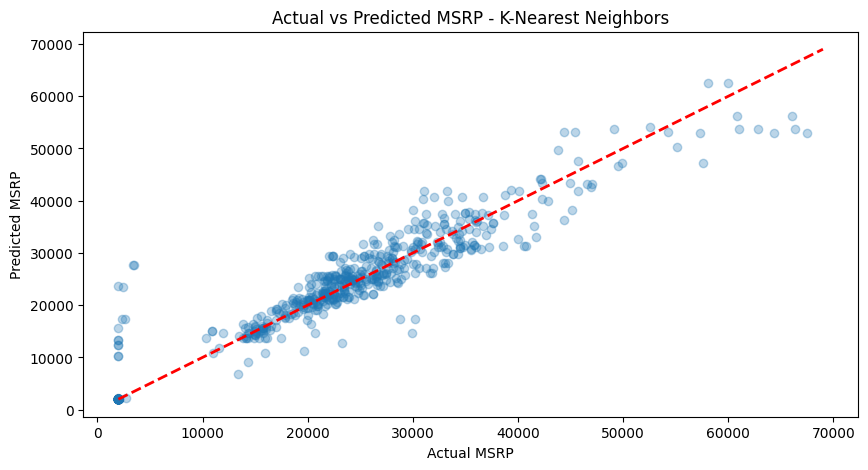
* Before Tuning: MAE = 4626.13, MSE = 36606305.48, R-squared = 0.7486
* After Tuning: MAE = 2556.50, MSE = 16699252.73, R-squared = 0.8853

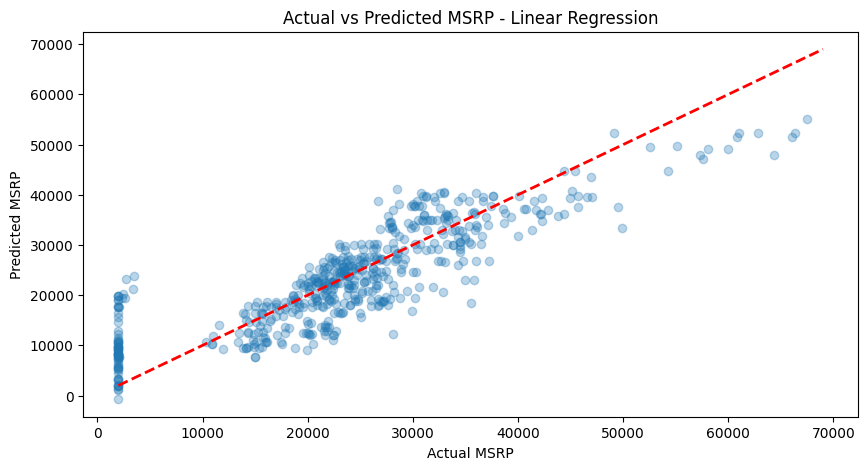
**Decision Tree Results:**

* Before Tuning: MAE = 4626.13, MSE = 36606305.48, R-squared = 0.7486
* After Tuning: MAE = 2293.20, MSE = 13319938.81, R-squared = 0.9085

The following graphs illustrate the performance of each model:







**Discussion**

The results show that with the help of fine-tuning, the predictive analysis enhanced the overall accuracy of K-Nearest Neighbors and Decision Tree final models, supported by a further increase in the proper R-squared value. For KNN, it changed from 0.7486 to 0.8853, while for a decision tree, the improvement changed from 0.7486 to 0. 9085.These enhancements were also supported by the decrease in the MAE and MSE used from the previous iteration. The linear regression model R squared remained the same (0.7485) both before and after fine tuning. This indicated that fine tuning the linear regression model does not affect the accuracy of predication.

**Conclusion**

The experimental results indicate that all fine-tuning approaches considered have the potential to contribute significantly to the effectiveness of predicting car prices. When the results of the models before and after fine-tuning were compared, it was found that the decision tree is the only model that could be pointed to as the best in price prediction of cars. For a study test, the model used is the decision tree model. Such coarse variation makes it possible for the model to have a higher R-square value as the fine-tuning showed that the model R-square value had improved from 0. 7486 to 0. 9085, which indicates that there is a high correlation in the coefficient to find the price estimate of cars. MAE and MSE values were also relatively reduced from the results obtained in the first split evaluation. The KNN model was also fine-tuned and gave good results, but here, the Decision Tree model outperformed KNN. In general, the performance of the linear regression model was statistically insignificant and much lower than that of a few non-linear models. Based on these results, non-linear classification techniques such as KNN and Decision Trees are better suited for this dataset. It could make more accurate models by using more complicated adjusting processes on other models.

**References**

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Linear regression. In *An introduction to statistical learning: With applications in python* (pp. 69-134). Cham: Springer International Publishing.

Lee, C. S., Cheang, P. Y. S., &. Moslehpour, M. (2022). Predictive analytics in business analytics: decision tree. *Advances in Decision Sciences*, *26*(1), 1-29

Zhang, S., & Li, J. (2021). KNN classification with one-step computation. IEEE Transactions on Knowledge and Data Engineering, 35(3), 2711-2723.