Portfolio 2 - Car Sales Price Prediction

Project Overview

This Jupyter Notebook, titled "Portfolio 2- Car Sales Price Prediction" is focused on predicting car selling prices using machine learning techniques, specifically linear regression models. The project explores the impact of feature selection and training data size on model performance. In addition to the predictive modeling, the notebook addresses ethical concerns in data science, particularly in data visualization and model transparency.

Table of Contents

1. **Introduction**
   The goal of the notebook is to train and evaluate different linear regression models for predicting the selling prices of cars. The notebook explores how feature selection and the size of training data affect the model's accuracy and error metrics, along with discussing ethical considerations in data handling.

2. **Libraries Used**

   - `pandas`
   - `numpy`
   - `matplotlib`
   - `scikit-learn`

3. **Data Loading and Preprocessing**
   The dataset used in this project is a cleaned version of a car sales dataset containing columns such as year, selling price, kilometers driven, fuel type, seller type, transmission, and number of previous owners. Categorical variables are converted to numerical values to be used in the regression models.

4. **Analysis / Modeling**
   Four linear regression models are trained to predict selling prices:

   - **Model A:** Uses two most correlated features (year, transmission) with 10% training data.
   - **Model B:** Uses two least correlated features (kilometers driven, owner) with 10% training data.
   - **Model C:** Uses two most correlated features with 90% training data.
   - **Model D:** Uses two least correlated features with 90% training data.

   The models are compared using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score to assess their performance.

5. **Data Science Ethics**
   This section addresses potential ethical concerns in data visualization and model transparency. The notebook includes an analysis of ethical issues, referencing two key resources:

   - **Example 1**: A case from Georgia where COVID-19 data visualization led to misinterpretation of trends.
   - **Example 2**: An article on ethical data visualization, emphasizing the importance of context and transparency in presenting data.

   Ethical issues discussed include:

   - Misrepresentation of data through improper visualization techniques.
   - The importance of clearly labeling axes and providing sufficient context to avoid misleading interpretations.
   - Transparency in machine learning models to ensure that predictions are not only accurate but also understandable and explainable.
6. **Results**

   - **Model C** (90% training data and most correlated features) had the best performance with the lowest MSE and RMSE.
   - **Model B** (10% training data and least correlated features) performed the worst, highlighting the importance of both feature selection and data size.
7. **Conclusion**
   Feature relevance and training data size significantly affect the accuracy of the regression models. The Year feature had the highest positive correlation with selling price, suggesting that newer cars command higher prices.

8. **Future Work / Improvements**
   Additional features could be explored to improve the model's predictive power, as the R^2 values indicate that current features do not fully explain the variability in selling prices.

How to Run the Notebook

1. Open the notebook using Jupyter:

   ```
   jupyter notebook Portfolio2.ipynb
   ```
2. Follow the code cells sequentially for a step-by-step execution of the analysis.

Files Included

- `48622613-Portfolio2.ipynb` - The Jupyter Notebook with all the code and analysis.
- `data` containing the dataset car_sells_clean_data.csv

Author

Isidora Gautier

In [ ]: