**Report on Final project**

**Sireesha Maguluri**

**An assignment submitted in partial fulfillment of the requirements for MSDS-532**

**as part of the degree Masters**

**In**

**Data Science**

**School of Computing and Information Sciences**

**University of the Cumberlands**

**MSDS532 - A01 Data Science Programming with Python**

**DR. Arnett Campbell**

**October 18, 2023**

**Fall, 2023**

**Predicting Car Prices: A Machine Learning Approach**

**Abstract:**

This report delves into the intricate task of predicting used car prices in the present market. Here, I have used the provided dataset, which is car\_ads\_fp.RData, the study employs data preprocessing, exploratory data analysis, and machine learning models to make accurate price predictions. This research contributes valuable insights into the domain of used car pricing and emphasizes the importance of advanced machine learning techniques for enhancing prediction accuracy.

**Introduction**

The used car market is a significant sector in the automotive industry, where accurate price prediction is essential for buyers and sellers. This report presents a comprehensive analysis of predicting used car prices based on generation\_model, color and different aspects which are mentioned in the provided data i.e. car\_ads\_fp.RData. The research involves a multi-step process, starting with data collection and preprocessing and culminating in the implementation of machine learning models (Bukvić et al., 2022).

**Data Collection and Preprocessing**

The study begins by loading data from an RData file using the Pyreadr library and converting it into a pandas DataFrame. Later, checks for missing values, duplicates, and columns with missing data. The missing values in numeric columns are filled with their respective means to enhance data quality, while categorical columns are imputed with the mode. Furthermore, the 'Price’, ‘GenModel’, and ‘Color’ columns are converted to numeric, and outliers are identified and removed using Z-scores Joseph (2023). A correlation heatmap is created to explore relationships among numerical attributes.

**Advanced Machine Learning Models:**

The report discusses implementing advanced machine learning models for used car price prediction based on the required factors like year, model, and color. Specifically, a Random Forest Regressor is used, and its performance metrics are evaluated. The report mentions that the Random Forest Regressor is trained on selected features, and the results are compared using Mean Squared Error (MSE) and R-squared (R²). Additionally, the report introduces the concept of a Gradient Boosting Regressor as an alternative model, highlighting its MSE and R² scores.

**Results and Visualizations:**

It mentions the creation of scatter plots to visualize predicted prices against actual prices, offering a clear understanding of the model's performance. Furthermore, the report discusses the generation of a correlation matrix and provides statistical summaries of numeric attributes, which offer valuable insights into data patterns. The report also delves into the correlation matrix and statistical summaries of numeric attributes Joseph (2023). These visualizations provide insights into the relationships among variables and data distribution.

**Recommendations and Future Directions:**

The report encourages further research in the field of used car price prediction. It suggests considering the application of advanced machine learning techniques, such as neural networks or deep learning models, to enhance prediction accuracy (Bukvić et al., 2022). Moreover, the report highlights the potential for extending this research to other regions and markets to evaluate the generalizability of the developed models.

**Conclusion**

The report summarizes the essential findings and methodologies for predicting used car prices. It highlights the significance of advanced machine learning models, data preprocessing, and data exploration in achieving accurate predictions. The report underscores the importance of data-driven approaches in addressing the complexities of used car pricing, providing valuable insights for buyers and sellers in the automotive industry.

**References:**

1. Bukvić, L., Škrinjar, J. P., Fratrović, T., & Abramović, B. (2022). Price prediction and classification of Used-Vehicles using supervised Machine Learning. Sustainability, 14(24), 17034. <https://doi.org/10.3390/su142417034>
2. Joseph. (2023). Cars predictions. *Kaggle*. <https://www.kaggle.com/code/joseph64/cars-predictions>