Class Data Challenge: Predicting Orange Juice Sales

As part of our machine learning course, students are presented with a real-world challenge that involves predicting the sales of orange juice brands at new grocery stores. This task requires applying machine learning techniques to analyze historical sales data and make predictions for future sales based on pricing and advertisement strategies.

Objective

The primary objective of this challenge is to predict the sales for three different orange juice brands (Tropicana, Minute Maid, and Dominicks) at new grocery stores. Students will utilize the provided datasets to train machine learning models and make predictions based on planned prices and advertisements.

Datasets

- `juice.csv`: Contains historical data, including sales, prices, advertisement flags (feat), store IDs, and orange juice brands.
- `new_grocery.csv`: Includes planned prices and advertisements for the orange juice brands at new grocery stores. There is no information on sales in this data set.

Instructions

1. **Data Inspection**: Begin by exploring both datasets to understand the features, target variable "sales", and the data's overall structure. Identify any missing values or anomalies that need to be addressed.

2. **Data Preparation**: Clean the data by handling missing values and outliers. Encode categorical variables appropriately. Split the `juice.csv` dataset into training and testing sets for model validation.

3. **Model Development**: Experiment with different machine learning models (e.g., Lasso, Ridge, Trees, Random Forests) to find the most powerful one for predicting orange juice sales. Evaluate each model's performance using suitable metrics and select the best model based on these evaluations.

4. **Model Application**: Apply the chosen model to the `new_grocery.csv` dataset to predict sales for the new grocery stores. Justify the choice of your model and discuss its advantages and limitations in this scenario.

5. **Report Writing**: Document your process, including data inspection, model choice and justification, evaluation metrics, and predictions. Include visual aids and a clear narrative to support your findings. This could be in the format of a JupyterNotebook or a PDF.

6. **Code and Predicted Output Submission**: Submit all the code used throughout the challenge. Provide a CSV file containing the predicted sales and corresponding store IDs for the new grocery stores.

Evaluation Criteria

Your work will be evaluated based on the accuracy of your predictions, the rationale behind your model selection, the clarity and completeness of your report, and the quality of your code. This challenge aims to simulate a real-world task, testing your ability to apply machine learning techniques effectively and communicate your findings.

Submission Deadline

Please submit all required documents and files by **01.06.2024**.