SEAS 6414 - Python Applications in Data Analytics
Homework 6

Due Date: February 24, 2024 (10:00am EST)

Instructions: To complete the following task using Python, please download an Integrated Development Environment (IDE) of your choice. Ensure that your solution includes both the written code (input) and its corresponding output. Once completed, upload your solution in PDF format or any other format you prefer. **The questions are worth 50 points each**.

Question 1: Financial Sentiment Analysis

Background:
The utilization of statistical techniques in financial sentiment analysis is often limited due to complexities in practical applications and a shortage of high-quality training data. In finance and economics texts, where annotated datasets are scarce and largely proprietary, this challenge becomes more pronounced. To address this, a collection of approximately 5000 sentences has been compiled to establish benchmarks for various modeling techniques.

Dataset:
The original Financial Phrase Bank data contains 4840 sentences, annotated by 16 individuals with robust backgrounds in financial markets, including researchers from Aalto University School of Business. However, this assignment focuses on approximately 47% of the dataset, which corresponds to sentences with 100% agreement among the annotators. The analysis will center on sentences with unanimous annotator agreement, labeled as 'positive', 'negative', and 'neutral'.

Tasks:

1. **Data Preparation:**
    - Load and convert the text file into a DataFrame with columns: 'Text' and 'Sentiment'.
2. **Exploratory Data Analysis (EDA):**
    - Perform EDA and plot bar charts for the frequency of the top 20 words in each sentiment category.
3. **Class Imbalance Analysis:**

- Compute and visualize the frequency of each sentiment label with a bar graph. Discuss class imbalance.

4. **Word Count Analysis:**

   - Create box plots for word/token counts per sentiment label. Discuss discrepancies.

5. **Data Splitting:**

   - Split data into training (80%) and testing (20%) sets using stratified splitting with a random seed of 64.

6. **Model Development and Evaluation:**

   - **Vectorization:** Use `CountVectorizer` and `Tf-Idf`.
   - **Model Training:** Employ Naive Bayes
   - **Model Training - Bonus Question (5pts):** Employ Random Forest and Support Vector Machines.
   - **Metrics Tracking:** Present Accuracy, Precision, Recall, F1-Score for each class, and an overall Confusion Matrix.
   - **Analysis:** Discuss the effectiveness of each model based on the tracked metrics.

Objective:
Apply and evaluate supervised learning techniques for sentiment analysis in the financial sector, enhancing skills in data preprocessing, exploratory analysis, and evaluating machine learning models based on multiple performance metrics.

Question 2: Predicting Building Energy Efficiency

Objective:
Apply regression techniques using Scikit-learn to analyze and predict the energy efficiency of buildings, focusing on heating and cooling load requirements. This involves the use of various regression models, feature engineering, and model evaluation.

Dataset:
The dataset for this assignment, *Energy Efficiency Dataset*, can be found at the UCI Machine Learning Repository. It includes architectural features and energy efficiency metrics of buildings. The dataset columns are renamed for clarity as follows:

```
column_names = {'X1':'Relative_Compactness', 'X2': 'Surface_Area',
                'X3': 'Wall_Area', 'X4': 'Roof_Area',
                'X5': 'Overall_Height', 'X6': 'Orientation',
                'X7': 'Glazing_Area', 'X8': 'Glazing_Area_Distribution',
                'Y1': 'Heating_Load', 'Y2': 'Cooling_Load'}
```

Tasks:

1. **Data Preprocessing:**
   - Perform exploratory data analysis (EDA) after loading the dataset.
   - Conduct feature engineering if necessary.
2. **Model Development:**
   - Implement various regression models (Linear Regression, Ridge, Lasso, and Elastic Net).
   - Implement Random Forest Regression (Bonus Question - 5pts)
   - Perform hyperparameter tuning for optimization.
3. **Model Evaluation:**
   - Evaluate models using RMSE, MAE, and R^2 score.
   - Assess performance on training and testing datasets.
4. **Target Variable Analysis:**
   - Develop separate models for *Heating Load* and *Cooling Load* as target variables.
   - Compare the effectiveness of models for each target.

Deliverables:

- All code, analysis, and outputs.

- A report summarizing findings and model comparisons.

This assignment focuses on applying and evaluating regression analysis in a real-world context, emphasizing feature selection, model optimization, and interpretation of results in the context of building energy efficiency.