PR2: Regression

Overview and Assignment Goals:

The objectives of this assignment are the following:

- Explore how to best represent and transform features numerically, especially categorical and textual fields.
- Apply dimensionality reduction or feature selection techniques to reduce noise and improve performance.
- Compare different regression algorithms and evaluate their effectiveness.
- Address skewed distributions and outliers in the target variable.

Detailed Description:
Develop predictive models that can determine, given a large number of attributes describing a short term rental, the price that should be charged for that rental.

Your goal is to build regression models that predict the nightly rental price of a short term rental listing in the US, based on various listing and host attributes. Note that listings from different parts of the country are mixed and exact location is obscured. The dataset consists of listing metadata, such as room type, number of reviews, minimum nights, availability, amenities, and more.

The raw dataset includes over 100,000 listings and 60+ features, some numerical, others categorical, and some text-based features. Not all features will be useful in predicting price. Some listings have price outliers (e.g., $10,000+/night), so thoughtful handling of skewed distributions will be important.

Assignment Tasks:

1. **Data Understanding and Cleaning**
    - Handle missing values appropriately.
    - Cap extreme outliers in the `price` column if necessary.
    - Engineer new features if appropriate (e.g., neighborhood popularity, review scores per night, etc.).
2. **Feature Representation**
    - Convert categorical features using suitable techniques (e.g., one-hot encoding, frequency encoding).
    - Use text features like `name`, `reviews` or `description` only if you think they

can add value (not required).

3. **Dimensionality Reduction / Feature Selection**
    - Try at least one dimensionality reduction technique (e.g., PCA, LLE, random projections, etc.).
    - Discuss how it impacts model performance and training time.
4. **Modeling**
    - Train and compare at least three regression models (e.g., Linear Regression, Ridge/Lasso, Decision Tree, Random Forest, Gradient Boosting, etc.).
    - Evaluate using **Root Mean Squared Error (RMSE)**.
5. **Skewed Target Handling**
    - Visualize the distribution of the target (`price`).
    - Consider log-transforming `price` if the distribution is heavily skewed.

We will use root mean square error (RMSE) as the evaluation metric for this task.

Caveats:

+ Remember that not all features will be good for predicting the target. Think of feature selection, engineering, reduction (anything that works).
+ Use the data mining/machine learning knowledge you have gained until now, wisely, to optimize your results.

Rules:

- This is an individual assignment. Discussion of broad level strategies are allowed but any copying of prediction files and source codes will result in an honor code violation.
- You are allowed 5 submissions per day.
- After the submission deadline, only your chosen or last submission is considered for the leaderboard.

Deliverables:

- Valid submissions to the Leader Board website: https://clp.engr.scu.edu (username is your SCU username and your password is your SCU password).

Canvas Submission for the report:

- Include a 2-page, single-spaced report describing details regarding the steps you followed for feature extraction, feature selection, and classifier model development. The report should be in PDF format and the file should be called **<SCU_ID>.pdf**. Be sure to include the following in the report:
    1. Name and SCU ID.

2. Rank & RMSE for your submission (at the time of writing the report). If you chose not to see the leaderboard, state so.
3. Your approach.
4. Your methodology of choosing the approach and associated parameters.

- Ensure you submitted the correct code on CLP that matches your output. Code does not need to be submitted on Canvas.

Grading:

Grading for the Assignment will be split on your implementation (70%), report (30%).

Files: The datasets are on the HPC, under `/WAVE/projects/CSEN-140-Sp25/data/pr2`.