

**Institute of Business Administration**  
**CSE-473: Introduction to Machine Learning**  
**Assignment 02**  
**Due Date: 2<sup>nd</sup> December 2024 (11:55 PM)**

---

**Instructions**

- You are required to participate in the challenge posted on the Kaggle website individually. The link to the Kaggle challenge is copied below:  
<https://www.kaggle.com/t/2ab76d0cde5f43758cbcd9af46b01279>
- The challenge must be implemented using Python.
- You are expected to use your proper full name. This is important because many people have similar first names. Any entry with an indistinguishable name/username (like warrior, beast, etc.) would be removed.
- You are allowed to post 20 entries per day. This is to ensure you start participating from the very first day. Being on the top of the leaderboard is important but more important is the breadth of your experiments and making a consistent effort.
- You are required to submit/fill a small Google form after every entry to ensure that we have a detailed record of your submissions, like which model you used, what data transformations you applied, how you handled the missing data, what model parameters were used, etc. The link to the form submission is copied below.  
[https://docs.google.com/forms/d/e/1FAIpQLSefXWYwQMN0O9ljwiziJtDDtxxZRHL\\_uu2sUtDEm91\\_O16ilw/viewform](https://docs.google.com/forms/d/e/1FAIpQLSefXWYwQMN0O9ljwiziJtDDtxxZRHL_uu2sUtDEm91_O16ilw/viewform)
- Your task is to primarily try following regression algorithms along with the data preprocessing, cleaning, and transformation techniques to increase the score.
  - Linear Regression
  - Polynomial Regression
  - KNN Regression
  - Regression Trees
  - Random Forest Regressor
  - AdaBoost Regressor
  - Gradient Boosting Regressor
  - XGBoost Regressor
  - Stacking
  - Neural Networks
- You must also submit the Python code notebook (on LMS), demonstrating your best entry on Kaggle. Please note that the submitted file must show the same score as the one visible on Kaggle upon execution. The Python notebook must also include your overall findings and insights in a single markdown cell. The analysis should encompass, but should not be limited to the following aspects:
  - Cover the breadth of your experiments by providing the details about each method you employed (e.g., KNN, Random Forest, etc.), such as,
    - How your KNN entries' score gradually improved including the combinations of techniques that proved most effective.
    - The impact of various data transformation techniques (e.g., column filtering based on correlation) on your performance scores.
    - The role of specific factors such as feature importance, maximum depth, and learning rate, and how these influenced the outcomes across different methods.
  - An analysis of which algorithm performed best for the given dataset and the reasons behind its effectiveness.
  - The key challenges you encountered while striving to improve your score and how you addressed them.