Title Of Project: Benchmarking Regression Models

Team Members: Jala Daniel, Eugene Lowe, Anirudh Vannemreddy, Jahnavi Chintala, Victor Mgbeafulu

Goal of the study: The goal of the study is to deepen our understanding of regression by applying various regression methods to multiple different datasets. We will choose data sets of various topics in different areas of study to ensure a thorough analysis of how different regression models perform under different conditions. Through data analysis, model building, and evaluation of metrics, we will compare the performance of each model. This study will hone our skills in model selection and interpretation of model results, preparing us to make data-driven decisions.

Preliminary Work: We have decided to use regression methods for our project, this includes Linear regression, Ridge regression, Lasso Regression, Principal Component Regression (PCR), Regression Splines and Smoothing Splines. We are using datasets from different areas of study so that we can have diversity in our study. Some of these datasets will have non-linear relationships, multicollinearity, some datasets will be large, and some will be small.

These datasets are nontrivial and interesting because they represent real world scenarios, meaning that they are complex enough to give us relevant experience handling and processing complex data. We specifically chose datasets that needed cleaning and preprocessing so that we could get the most out of this study.

Expected Outcome and Measure of Success: The expected output of this study is a report containing comparisons of different regression methods and insights into which models perform best with which datasets. We will gain insights based on the summaries of performance metrics for each model. These metrics will allow us to identify what conditions a model performs it's best and worse in. Another expected output is our Jupyter notebook, which will contain all of the code used during the study with markdown annotations explaining each step.

We will measure our success in multiple ways, one way being our report. The main goal is to deepen our understanding of regression, so a comprehensive report that accurately demonstrates different regressions methods with statistical evidence that supports our conclusions would show success. Another way that we will measure success is by evaluating and comparing our metrics using methods such as R-squared to measure how

well the model is dealing with variance in the data. We will also use cross-validation to ensure that our model isn't underfitting or overfitting the data. We will also use metrics such as MSE, RSS and RSE to compare method performances.

Dataset	Number of features	Number of Instances
Song Popularity Dataset	15	13070
Seoul Bike Sharing Demand	13	8760
Health Insurance charge	7	1338
Student Performance Factors	20	6607
Gym Membership Dataset	17	1000
Diabetes Dataset	34	70000
Communities and Crime	127	1994
Productivity Prediction of Garment Employees	15	1197
Porter Delivery Time Estimation	14	197428
Appliances Energy Prediction	28	19735

Datasets:

1. Song Popularity Dataset: https://www.kaggle.com/datasets/yasserh/song-popularity-dataset
2. Seoul Bike Sharing Demand: https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand
3. Health Insurance charges: https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset/data
4. Student Performance Factors: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors
5. Gym Membership Dataset: https://www.kaggle.com/datasets/ka66ledata/gym-membership-dataset
6. Diabetes Dataset: https://www.kaggle.com/datasets/ankitbatra1210/diabetes-dataset
7. Communities and Crime: https://archive.ics.uci.edu/dataset/183/communities+and+crime
8. Productivity Prediction of Garment Employees: https://www.kaggle.com/datasets/ishadss/productivity-prediction-of-garment-employees
9. Porter Delivery Time Estimation: https://www.kaggle.com/datasets/ranitsarkar01/porter-delivery-time-estimation
10. Appliances Energy Prediction https://archive.ics.uci.edu/dataset/374/appliances+energy+prediction