****

**Course: DATA522 - Solving Big Data Problems-Data Analytics - Spring 2025 (Online)**

**Report: Final Project Abstract**

**Date: April 15, 2025**

**Name: Meagan O’Briant Student ID: 770047676**

**Predicting House Prices Using Advanced Regression Techniques**

**Team Members:** Meagan O’Briant; Osly Ariel Cabrera Fletes; Orellana Barroso Javier Andres; Akhil Abburi

**Kaggle Competition Chosen:** House Prices - Advanced Regression Techniques (<https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques>)

**GitHub URL:** <https://github.com/MeagOBriant/House-Prices-Regression-Final-Project.git>

**Abstract**

For this project, we are building a model that can predict the sale prices of houses using the Ames Housing dataset featured in the Kaggle competition *House Prices – Advanced Regression Techniques*. The dataset contains 79 variables that describe various aspects of residential homes, including lot size, year built, quality ratings, neighborhood, and more. Since pricing homes accurately is such an important part of real estate, this dataset gives us a great chance to dig into regression modeling and get hands-on experience using RStudio. The objective is to apply statistical methods to accurately predict house prices and evaluate model performance using Root Mean Square Error (RMSE), which serves as the competition’s scoring metric. This project also emphasizes the practical application of data cleaning, transformation, and feature selection to build an effective, data-driven solution.

**Approach/Strategy**

The project will be conducted in RStudio using a set of packages tailored for data analysis and machine learning workflows. The initial steps involve data exploration and cleaning using dplyr and tidyverse. This step will help us in identifying and managing missing values, outliers, and skewed distributions. Then feature engineering will be applied, including both numerical transformations and categorical variable encoding to optimize model input. Visualizations will be created using ggplot2 to uncover patterns and relationships in the dataset.  
  
 Next, for modeling, we will begin with basic linear regression to establish a performance baseline, followed by more advanced ensemble models such as random forest and gradient boosting. These will be implemented through packages like randomForest, xgboost, and caret. Model performance will be assessed using cross-validation and RMSE to ensure reliability and generalizability. All work will be documented and version-controlled via GitHub, with the final deliverable structured in PowerPoint format per course guidelines.

**References:**

Balbaa, M. E., Tashtemirovich, A. O., Ismailova, N., & Batirova, N. (2023). Real-time analytics in financial market forecasting: A big data approach. In The International Conference on Future Networks and Distributed Systems (ICFNDS ’23), December 21–22, Dubai, United Arab Emirates. ACM. (UoNA e-library).

Liu, Y., & Tang, A. (2022).

Loshin, D. (2013). *Big data analytics : from strategic planning to enterprise integration with tools, techniques, NoSQL, and graph*. Elsevier/Morgan Kaufmann.

‌  
Prediction method of government economic situation based on big data analysis. Digital Government: Research and Practice, 3(4), Article 27. (UoNA e-library).