Project Proposal

Course: Reproducible Research
Major: Data Science and Business Analytics
Student Names: Maryam Abdulhuseynova 436856,
Parvin Badalov 456028,
Reikhan Gurbanova 468193
Date: 31.03.2025

Project Title

House Price Prediction: Reproducing and Improving a Model Using TFDF

Background and Motivation

We are choosing a project based on an existing Kaggle analysis. Our goal is to reproduce the analysis into a different programming language and improve the model. We want to work with the "House Prices Prediction using TFDF" by gusthema Kaggle Notebook, which uses TensorFlow Decision Forests (TFDF) in Python to predict housing prices. This dataset includes housing features from Ames, Iowa.

As Data Science students, we find this project valuable because it allows us to practice code translation, model reproduction, and improvement on a real-world problem. We'll be translating the notebook from Python to R to see how reproducibility holds across environments.

What We Plan to Do ?

1. **Reproduce the original analysis**: We will take the original Python code and translate it into R, using similar packages and functions.
2. **Compare results**: We'll evaluate if the translated version performs similarly (using metrics like RMSE or R^2) and discuss any differences.
3. **Improve the model**: We'll explore using different algorithms, hyperparameter tuning, and feature engineering techniques.
4. **Test robustness**: We'll run cross-validation and test how the model performs under different preprocessing methods.
5. **Summarize and reflect**: We'll write about what was easy or difficult in the translation process and how the improvements affected the model.

Tools and Languages

- **Original Language**: Python

- **New Language**: R
- R Packages: randomForest, xgboost, tidyverse, caret, mlr3
- Version Control: Git and GitHub for collaboration, version tracking, and project documentation

Dataset

We are using the "House Prices - Advanced Regression Techniques" dataset from Kaggle. It contains detailed information about homes sold in Ames, Iowa.

Final Deliverables

- A translated and well-documented R script or notebook replicating the original analysis
- A GitHub repository with all code, documentation, and version history
- A short report explaining the process, challenges in reproducibility, model improvements, and final outcomes

Why We Chose This Project ?

We wanted to choose a practical and well-known dataset that allows us to focus on the core elements of reproducible research: translating code, testing reproducibility, improving models, and clearly documenting the process. Using Git throughout the project also helps us apply proper version control and collaboration practices. This project lets us apply all of these while working on a real and structured prediction problem.