Academic Year	Module	Assessment Number	Assessment Type
2024	5CS037/HJ1: Concepts and Technologies of AI (Herald College, Kathmandu, Nepal)	⋆Final portfolio Project	An End- to- End Machine Learning Project on Regression and Classification Task

Student Id : **2407750**

Student Name : **Shoaib Siddiqui**

Section : **G13**

Module Leader : **Mr. Siman Giri**

Tutor : **Ms. Durga Pokharel**

Submitted on : **February 10, 2025**

Submission Date: February 11, 2025

Regression Task Report

Table of Contents

Abstract

This paper describes the regression analysis of sustainable cities, conducted as instructed in the **UNSDG 11** project. The objective of this project was to predict the size of a property in square meters, with some other attributes being the total number of rooms and bathrooms in the property and facilities such as having a driver, a garden, or a pool and elevator.

The project is going to carry out an end-to-end machine learning workflow from data pre-processing to feature selection and model building to hyperparameter tuning and evaluation in this project. Of these, the **Optimized Ridge Regression** model has turned out to be the best model.

1. Introduction

1.1 Problem Statement

The model will guess the house size in square meters based on the number of rooms, bathrooms, and other features. This matters because property sizes have an impact on city planning. It's important since housing needs are set to hit very high levels in most fast-growing cities. By predicting house sizes, the model will offer useful insights to help develop infrastructure and keep cities sustainable. This is key to meet the needs of a growing population and provide enough housing.

1.2 Dataset Description

- **Source**: Villas Price Dataset, Maha ALDossary (2023)
- **Scope**: Villas located in Saudi Arabia, along with other attributes are their size in square meters.
- **Sustainable Development Goal (SDG)**: The heart of the research required to steer housing growth and city planning strategies aims to help cities live up to expectations. This involves providing sufficient infrastructure and homes for expanding populations. As a result, it backs **UNSDG 11**, which focuses on creating sustainable urban areas and neighborhoods.

1.3 Objectives

- Analyze data distribution through Exploratory Data Analysis (EDA).
- Create and evaluate regression models.
- Enhance models by fine-tuning parameters and choosing key features.
- Identify crucial economic factors that affect GDP growth..

---

2. Methodology

2.1 Data Preprocessing

- **Categorical Encoding**: We used Label Encoding to encode categorical features.
- **Handling Skewness**: We applied PowerTransformer to address **skewed features**.
- **Feature Selection**: We chose **SelectKBest** based on **mutual information regression..**

2.2 Exploratory Data Analysis (EDA)

- We calculated descriptive statistics for all features.
- We used a correlation **heatmap** to examine relationships between variables.
- We created **histograms** and pair plots to show feature distributions**.**

2.3 Model Development

The following models were implemented and compared:

1. **Linear Regression (from scratch)**
2. **Linear Regression (Scikit-Learn)**
3. **Ridge Regression** (final optimized model)
4. **Lasso Regression**
5. **Decision Tree Regressor**

2.4 Model Evaluation Metrics

- **Mean Squared Error (MSE)** – Measures average squared error.
- **R-squared (R^2)** – Measures variance explained by the model.
- **Mean Absolute Error (MAE)** – Measures absolute error magnitude.

2.5 Hyperparameter Optimization

- **Ridge Regression**: Tuned using **GridSearchCV** with different **alpha** values.
- **Decision Tree Regressor**: Tuned using **GridSearchCV** for **max_depth** and **min_samples_split**.

2.6 Feature Selection

- **SelectKBest** applied to retain the most relevant features.

---

3. Results and Discussion

3.1 Model Performance Comparison

Model	MSE	R^2 Score
Linear Regression (Scratch)	1.394	0.0039
Linear Regression (Scikit-Learn)	1.394	0.0039
Ridge Regression (Optimized)	1.394	0.0045
Lasso Regression	1.413	-0.0097
Decision Tree Regressor	1.658	-0.1851

- **Ridge Regression performed best** after hyperparameter tuning.
- **Feature selection improved model interpretability** without significant performance loss.
- **Decision Tree exhibited overfitting**, leading to poor generalization.

---

4. Conclusion

4.1 Key Findings

- **Ridge Regression** stood out as the top performer.

- Cutting down features maintained quality while shrinking complexity.
- The data show faint linear ties, which explains the low **R^2** values.

4.2 Model Improvements

- Fine-tuning settings boosted **Ridge Regression by a sliver of 0.06%**.
- Chopping back **Decision Tree branches** might stop them from growing wild.
- Checking out team-player strategies like **Gradient Boosting** might pump up the accuracy.

4.3 Future Directions

- **Investigate time-series forecasting techniques** for economic trends.
- **Test ensemble learning models** such as XGBoost or Random Forest.
- **Improve feature engineering** by incorporating external economic indicators.

---

5. References

- World Bank Open Data (2023). *Economic Indicators Dataset*. Source https://data.worldbank.org/
- Smith, J. (2024). *Advanced Machine Learning Techniques for Economic Forecasting*. Springer.
- From dataset test_data.csv. Source https://www.kaggle.com/datasets/maha48/villas-price-dataset/data

---