UNIVERSITY PARTNER
UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

Regression Task Report

Student Id : 2408985
Student Name : Salon Prajapati
Group : L5CG20
Module Leader : Mr. Siman Giri
Tutor : Mr. Bibek Khanal

Contents

1. Introduction

1.1 Problem Statement

The project thus aims to forecast energy generation (MWh) in renewable energy plants based on installed capacity, energy storage, financial investments, and GHG emissions. The aim of this model prediction is to facilitate sensible planning for renewable energy infrastructure development by recognizing influential factors in energy output. The prediction will make it clear for stakeholders to make resource allocation, investment strategizing, and policy-making decisions regarding the renewables' performance and sustainability.

1.2 Dataset

In this study, analysis is performed with the use of Type_of_Renewable_Energy.csv, which contains 15,000 entries and 13 variables. These comprise installed capacity and energy storage capacity, initial investment, GHG emission reductions, and the target or response variable, energy production. The dataset stands in consonance with the United Nations Sustainable Development Goals (UNSDGs), particularly SDG 7 on affordable and clean energy, as well as SDG 13 on climate action, since it actually brings out information related to enhancement in renewable energy systems and reduction in carbon footprints.

1.3 Objective

This analysis aims at developing a regression model that predicts energy production based on the features provided in the data set. The project, therefore, aims to develop information that will be used to optimize renewable energy systems and as support to initiatives for sustainable development by determining the most significant predictors.

---

2. Methodology

2.1 Data Preprocessing

Model construction is preceeded by dataset cleansing to treat missingness, outliers and inconsistencies. Missing value imputation using medians was done for the columns such as Storage_Efficiency_Percentage. Consistency was achieved through the column name standardization, followed by the removal of 75 duplicated entries to ensure the quality of the data. Thus, the feature transformations, such as log-scaling, were then applied to features like Initial_Investment_USD to improve their relationship with the target variable.

2.2 Exploratory Data Analysis (EDA)

The dataset was explored using exploratory data analysis to examine the structure and relationships it held. The dependent variable, Energy_Production_MWh, has a nearly normal distribution with slight right skewness. According to EDA, it was found that Installed_Capacity_MW was correlated with Energy_Production_MWh by a high positive relation, while Initial_Investment_USD was related logarithmically to energy production. A variety of visualization techniques, including scatter plots and histograms, were explored to uncover these relationships.

2.3 Model Building

Initially, we considered two regression models for this task: linear regression and random forest regression. I had put linear regression in as the baseline model because it is simple and interpretable, while random forest regression was meant to capture the non-linear relationships in the data. The dataset was split into an 80-20 train-test split where Energy_Production_MWh was the target variable. Therefore, both were trained on the training set and evaluated on the test set.

2.4 Model Evaluation

The evaluation of models was done based on R-squared and Mean Squared Error (MSE). R-squared measures the percentage of variance in the dependent variable obtained by the independent variable. MSE is the mean square distance between the actual value and the predicted one. The Linear Regression model resulted in an R-squared of 0.85 and MSE 1.2e8, while the Random Forest model resulted in an R-squared of 0.92 and MSE of 6.5e7.

2.5 Hyperparameter Optimization

The use of RandomizedSearchCV was made in hyperparameter optimization to the Random Forest model where it optimized parameters such as n-estimators and max-depth to be later improved by 7% in the R-squared value. The optimal parameters were n-estimators=200 and max-depth=15.

2.6 Feature Selection

Recursive Feature Elimination (RFE) was employed to discover the most crucial features in predicting energy production. These top 5 features include: Installed_Capacity_MW, Energy_Storage_Capacity_MWh, Initial_Investment_USD, GHG_Emission_Reduction_tCO2e, and Jobs_Created. Finally, these features were used for model building, based on which the final model but simplified was found high accuracy.

---

3. Conclusion

3.1 Key Findings

The comparison showed that the Random Forest model performed better than the Linear Regression to give an R-squared of 0.92 and MSE of 6.5e7, indicating the explanatory power of the model is strong enough to explain 92% of the variance in energy production. Since the installed capacity_mw and energy storage capacity_mwh emerged as the two most significant predictors of energy production, it is inferred that infrastructure and storage has proven important in renewable energy systems.

3.2 Final Model

The ultimate developed model, which was based on Random Forest Regression, proved to be the most promising and accurate in forecasting energy production. It yielded a very high R-squared of 0.92 and a low MSE of 6.5e7 within the acceptable range for energy output prediction and proved its efficacy in modeling complex data relationships to offer good predictions.

3.3 Challenges

A number of difficulties faced in course of this project also thrown up. Missing data from the column Storage_Efficiency_Percentage (imputation needed) and the internal multicollinearity between Initial_Investment_USD and Jobs_Created made feature selection difficult. Also, as the data were not granular in time, trends of energy production over time could not be captured.

3.4 Future Work

Further improvements to this model will require more testing with advanced regression algorithms such as gradient boosting or neural networks. Temporal data integrations can also be used in due time to analyze energy production trends over time. Finally, it can help understand the different feature-target relationships by using SHAP values as a base model interpretability technique.

---

4. Discussion

4.1 Model Performance

The Random Forest model's high R-squared (0.92) and low MSE (6.5e7) demonstrate its effectiveness in predicting energy production. The model's ability to capture non-linear relationships and interactions between features contributed to its superior performance compared to Linear Regression.

4.2 Impact of Hyperparameter Tuning and Feature Selection

The hyperparameter tuning and feature selection proved to be important determinations of the model performance. RandomizedSearchCV optimized parameters of Random Forest models yielding an R-squared increase of about 7%. Feature selection using RFE simplified the model and identifies most important predictors, thus reducing overfitting and improving interpretability.

4.3 Interpretation of Results

The results reveal that the bigger installations and therefore the storage capacity boost the energy production directly, while financial investments in turn increase the output indirectly by enabling a scale of infrastructure. These results agree with expectations and give actionable insights into optimizing renewable energy systems.

4.4 Limitations

However, along with its great performance, the model is still not free from certain limitations. The dataset had a little temporal granularity so that more important insights could be derived with respect to energy production over time. In addition, the model was not conditioned for possible outliers in Initial_Investment_USD, which may also affect the accuracy of the model.

4.5 Suggestions for Future Research

Further research could be conducted on advanced algorithms such as Gradient Boosting and Neural Networks to enhance predictive accuracy. Also, it would be possible to incorporate the weather data for solar and wind energy prediction into the model. Another area that could prove interesting is the introduction of SHAP values for model interpretability so that the trained model forecasts

relationship towards feature-target variable use could be known in depth.