Machine Learning lifecycle – GreenScape project

Authors - Kacper Janczyk, Simona Dimitrova, Matyas Tatar, Yuliia Bobrovytska, Collin Limoncelli-Buyskes

Agenda

1. Introduction

In this paper the Machine Learning lifecycle will be explained in the context of the project. The machine learning lifecycle is a cyclic process that can be done repeatedly and does not need any programming. The ML project life cycle can generally be divided into three main stages: data preparation, model creation, and deployment. All three of these components are essential for creating quality models that will bring added value to your business. It is called a cycle because, when properly executed, the insights gained from the existing model will direct and define the next model to be deployed. With regard to the project, the Cross-Industry Standard Process cycle is being followed and utilized.

2. Data Collection

In this section data collection and gathering will be explained in the context of the project, specifically gathering data and analyzing feasibility from a data perspective. In the matter of the project Data Quality report was documented that provided an assessment of the quality and reliability of the dataset. It is typically prepared as part of the understanding phase in the machine learning lifecycle. The purpose of the data quality report is to evaluate the fitness of the data for the intended analysis and decision-making process. Data collection is crucial when it comes to the machine learning lifecycle as it involves gathering the data that will be used to train and build the machine learning models. The quality, relevance, and representativeness of the collected data significantly impact the performance and effectiveness of the models.

2.1 Data requirements

In the context of the project, it was essential first to determine the machine learning task in order to collect the data. Structured data (tabular data with rows and columns) was the type of data we used. When it comes to data volume, there was no defined desired size of the dataset. During data collection, our main objective was to define the quality standards and expectations for the collected data, such as accuracy, completeness, consistency, and reliability. To be suitable for the intended ML analysis, we ensured that the data is trustworthy and error-free. Data Labeling – since supervised machine learning was implemented, we had to specify the required target output value which in our case is the green score index that needed to be associated with the input data (quality of life, public nuisance, income, environment statistics, housing, segregation, etc.). In order to determine the relevance and applicability of the data, we had to define specific features, and attributes (neighborhoods, regions, years) that need to be captured to achieve desired outcomes. We tried to avoid collecting unnecessary and redundant data as much as possible so that we do not deviate from the project's research questions. When collecting the data, we excluded all kinds of personal and sensitive information, to ensure that appropriate measures are in place to protect data privacy and confidentiality.

2.2 Data Sources

The data sources refer to the location from which the required data is obtained. For collecting relevant data, it is important to choose appropriate data sources. All datasets are gathered from Publicly Available websites 'Breda in Numbers', and 'Allcharts.info' except for two datasets (green score index and livability score) which are obtained from an internal database – BUasVPN.

2.3 Data Labelling

This section refers to data labeling which involves assigning the correct output values which in our case is the green score index to the input values which are the factors that have a correlation with the green score index, enabling the training of ML models to make accurate predictions.

3. Data Preparation

The crucial phase in the ML lifecycle is Data Preparation since it involves preprocessing raw data into a suitable format for training the Machine Learning model. Key aspects of this process are data cleaning, feature selection, data transformation, data encoding, train-test-split, and data integration. Most machine learning algorithms require data to be formatted in a specific way.

3.1 Data Cleaning

- Duplicates
- Missing values
- Outliers

Once the data has been cleaned, it is transformed to suit the analytics and ML modeling. The structure was changed, as the datasets were merged and aggregated along important dimensions. The results were stored in a new main dataset which was utilized for machine learning modeling.

4. Model Training

The core component of the machine learning lifecycle is model training, where the ML model learns patterns and relationships in the data to make predictions. It involves machine learning algorithm selection, defining the model architecture, and parameter optimization.

4.1 Algorithm selection

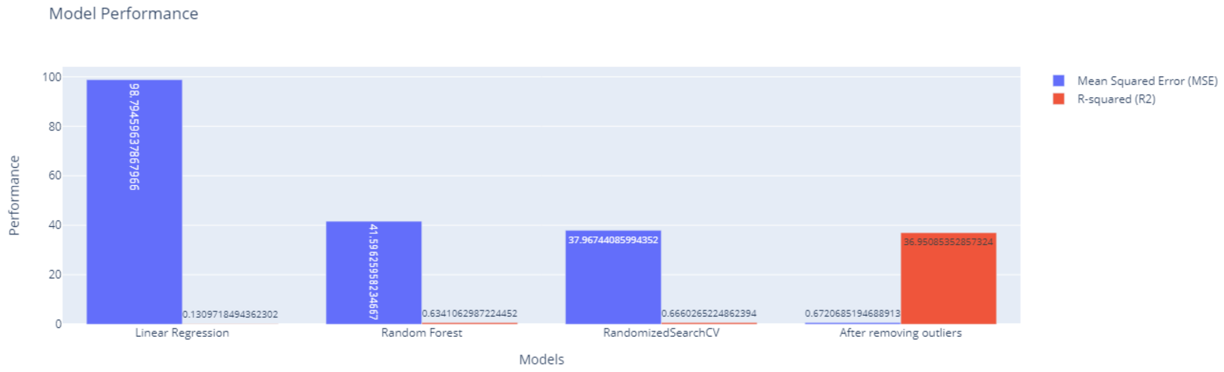
The machine learning algorithm that is being utilized in the context of the project is Random Forest. Since we were working with tabular data, in this matter regression task was most suitable for making machine learning predictions on the green score index. We came to the conclusion that random forest has the best performance after assessing the performance of different machine learning algorithms. By using evaluation regression metrics such as mean square error (MSE), mean absolute error (MAE), and r-squared coefficient (R2) we determined which metrics were the closest to our target values. Lower the MSE and MAE indicated better accuracy in terms of indicating smaller differences between the predicted and target values.

5. Model Evaluation

This section assesses the performance and quality of the trained machine learning model. In the context of the project, this involves measuring how well the model is generalized to unseen data, making informed decisions about model deployment, and assessing the limitations. Finally, we compared the results with the success metrics and decided whether to deploy

Mean Squared Error (MSE): Assess the model's accuracy in regression problems by measuring the average squared or square root of the differences between predicted and actual values.

R-squared (R2): Measures the proportion of the variance in the dependent variable (target) that is predictable from the independent variables (features) used in the model. The R2 score provides an indication of how well the model fits the data and explains the variability in the target variable.



6. Model Deployment

A key phase in the ML lifecycle, model deployment calls for careful consideration of infrastructure, packaging, serving, monitoring, and security issues. Organizations may assure a seamless and efficient deployment of their ML models, enabling real-world applications and value creation, by following best practices and integrating deployment into a CI/CD pipeline. The model will be deployed on the Streamlit dashboard as part of our main objective for this project.

7. Model Monitoring

Prediction monitoring involves analyzing the actual predictions made by the model in real-world scenarios. This can include comparing the predicted outcomes to the ground truth values and evaluating prediction errors. Monitoring prediction outputs helps identify any drift or deviations from expected behavior, enabling proactive intervention to maintain model accuracy. In the context of the project, we have focused on prediction monitoring since our main goal is to assess how the green score index changes over time and in neighborhoods. Prediction monitoring entails examining the model's actual predictions in actual situations. This might involve assessing prediction errors and contrasting anticipated results with actual values. Monitoring prediction outputs enables proactive action to preserve model accuracy by seeing any drift or departures from predicted behavior.

8. Conclusion

The main objectives are met and the research questions are answered by the final Streamlit dashboard machine learning analysis and EDAs. In the context of the project, interpretability is more important than accuracy meaning the model is simpler and more transparent since it depends on concerns to what degree the model allows for human understanding.