Date: 29.11.2024

Consulting Report for Module 5

Prepared by: Lukasz Macias

For: Hazel Bunning, Marcel Grossjean

Project: Predicting Crime Patterns in Switzerland 2030, Demographic and Offence trends

****

Project Overview

The primary objective of this project is to forecast the total number of offences in Switzerland for the years 2024–2030. The authors leverage various predictive machine learning techniques on a dataset obtained from the Swiss Federal Statistics Office, specifically focusing on offences and accused individuals under the Swiss Criminal Code (SCC).

Dataset Characteristics

The dataset comprises records from 2009 to 2023, with the following key features:

- **Size:** 1,488 rows.
- **Demographic Information:** Includes variables such as sex, age group, and nationality.
- **Offence Counts:** Total number of offences committed by legal or natural persons.

Analysis Overview

The analysis is structured into several stages:

1. **Exploratory Data Analysis (EDA):**

The authors conduct a comprehensive EDA to gain insights into the dataset's structure and trends.

2. **Model Development:**

Five machine learning models are employed to predict the number of offences:

a. Multiple Linear Regression
b. Decision Tree Regressor
c. Random Forest Regressor
d. XGBoost
e. Artificial Neural Networks

3. **Model Evaluation:**

The models are assessed using the following evaluation metrics:

a. Mean Squared Error (MSE)
b. Mean Absolute Error (MAE)
c. R-squared

4. **Results Presentation:**
   a. Predictions from the Random Forest Regressor are visualized using a plot.
   b. Feature importance analysis highlights the key drivers of offences.
   c. A dendrogram is presented to illustrate hierarchical clustering of features.
   d. Dimensionality reduction is performed, followed by K-Means clustering of the results to identify potential groupings.

5. **Discussion:** The authors conclude by discussing the outcomes, including the strengths and limitations of their approach and potential implications for predicting offence trends.

Positive Aspects and Strengths

The authors selected a highly engaging and relevant topic, addressing an important societal issue. The dataset, sourced from the Swiss Federal Office for Statistics, is likely to be highly accurate and reliable, providing a strong foundation for the analysis.

A notable strength of the project is the use of a diverse range of machine learning techniques. The authors applied and compared five predictive models, ranging from fundamental approaches like Linear Regression to more advanced methods such as Artificial Neural Networks.

Additionally, the analysis goes beyond basic predictions by employing advanced machine learning techniques. The use of dendrograms for hierarchical clustering and K-Means clustering on a reduced-dimensionality dataset provides deeper insights into the structure and relationships within the data.

Suggestions for Improvement

1) The research question addressed in the project is too simplistic and general. Predicting the total number of offences in upcoming years can be done visually by examining a line plot, making the application of machine learning techniques appear excessive for this purpose. To enhance the analysis, I recommend focusing on a more sophisticated and less obvious topic. An exapmle would be analyzing demographic associations: are certain demographic groups (e.g., age, sex, nationality) more strongly associated with fluctuations in offence counts over time?

2) The presentation is too brief, lacks structure and sufficient explanation on the slides. A well-structured presentation typically begins with an agenda, which is missing here. Including an agenda helps set the stage for the audience, outlining the key sections of the analysis. To improve clarity and flow, it is recommended to separate the different parts of the analysis into distinct sections:
    a. Exploratory Data Analysis (EDA): Clearly outline the insights gained during the initial exploration of the data.
    b. Model Training: Present the details of the machine learning models used, their configurations, and the evaluation process.
    c. Conclusions and Discussion: Summarize the key findings, discuss their implications, and address any limitations or future work.

3) Exploratory Data Analysis - The purpose of Exploratory Data Analysis (EDA) is to familiarize the audience with the dataset used in the analysis. The authors employed pie charts to explore and visualize the data, illustrating aspects such as the number of offences by type, sex, or age group.

    However, several issues with the presentation of EDA should be addressed:

1. **Sorting of Values:** The values in the pie charts were not sorted, making it difficult to quickly identify the most and least dominant offence types. Sorting the values would improve readability and allow for a clearer interpretation of the data.
2. **Lack of Numerical Context:** The charts do not include the actual number of offences for each category. Without this information, it is challenging to gauge the relative weight or significance of each group.
3. **Choice of Visualization:** Pie charts are generally less effective for comparing data across categories. A simple bar chart with the number of offences on the Y-axis would be a better alternative. Percentages could optionally be added as labels to provide additional context without sacrificing clarity.

4) The Model Evaluation slide is lacking critical information regarding the training process and the steps taken to prepare the models. Specifically, there is no mention of how the models were trained, which features were selected, or the data preprocessing steps that were undertaken. Additionally, the hyperparameters used in each model are not provided. Instead, the slide only presents the evaluation metrics for each model, without the necessary context to understand how the models were built and optimized.

5) The K-Means clustering slide lacks crucial information on how the number of clusters was determined and the method used for this selection. It is important to clarify the approach taken, such as using the elbow method or silhouette score, to justify the choice of the optimal number of clusters. Additionally, in K-Means clustering, it is a best practice to assign meaningful and descriptive names to the clusters based on their characteristics. This was not included in the slide. Furthermore, there is no interpretation of the clusters, which would help contextualize the results and provide insights into the patterns identified by the model.

Conclusions

The project demonstrated a solid approach in applying machine learning techniques, with strengths including the use of diverse models and advanced methods for feature analysis. The dataset, sourced from a reputable institution, provided a strong foundation for the analysis, and the use of feature importance analysis and clustering techniques revealed deeper insights into the structure of the data.

However, there are several areas for improvement. The primary research question, predicting the total number of offences, is relatively simplistic and could be enhanced by addressing more complex topics. The presentation could benefit from a clearer structure, with distinct sections. Additionally, more context is needed around the model evaluation process, such as details on data preprocessing, feature selection, and hyperparameter tuning.

In conclusion, while the project provides valuable insights and a comprehensive use of machine learning techniques, there is room for further refinement in both the analysis and presentation. Future work could focus on more sophisticated predictive questions and enhanced clarity in the explanation and visualization of the results.