The Panic Room

Product lifecycle

The Crime Crystal Ball

Introduction

Introducing The Crime Crystal Ball, a cutting-edge machine learning model made to Improve the crime-fighting field by prediction and prevention. In today's society and economy, ensuring safety and security on the street is more important and challenging than ever. By using machine learning and data analytics, The Crime Crystal Ball works to provide crime prevention services and make them stay on top of their game. We do this by giving them more insights into how we try to predict the ever-changing future. Our aim in this project is to prioritize fairness and transparency, seeing as these are essential when working with our ethical considerations. We want to keep our model's biases mitigated and make sure that there is equity for all the communities that might be affected by our product.

Machine learning lifecycle

This file will document everything related to the product, its lifecycle, and the product iterations.

Data

Data used in the project is open-sourced and available to the public from the municipality of Breda and the Central Bureau of Statistics.

Data has been cleaned by splitting, combining, and renaming columns as well as transforming data into a more machine-friendly format. After Exploratory Data Analysis was performed to better understand available information, findings proved that the original idea was not feasible and a new one needed to be drafted.

The new project is supposed to utilise machine learning to analyse the influence of different factors on crime rates in neighbourhoods on a yearly scale. The year 2019 has been chosen for training the model as this year provides the most data and the Covid outbreak happened at the end of the year so it is a non-factor.

Data used:

Total amount of crime committed in neighbourhoods in a year;

Number of citizens with a level of education in neighbourhoods *('Laag', 'Midden', 'Hoog');*

Percentage of households in a low and high income level *('Lowincome', 'Highincome');*

Population of neighbourhood divided into 5 Age groups *('<30 Jaar, '30-44 Jaar', '45-64 Jaar', '>= 75 Jaar');*

Quality of life in neighbourhoods, available data was only for years 2018 and 2020 as such interpolation has been used to ... data for the year 2019;

Matey has performed different operations-

In preparation for machine learning data has been put into two dataframes, one with crime rate, the second is a result of merging dataframes: ‘population’, ‘income’, ‘education’ and later added ‘QoL’. The resulting data frames have information about 48 neighbourhoods.

ML model

First, the crime rate has been clustered into three classes using the KMeans algorithm.

Then, Clustered data used as a prediction target hereafter referred to as ‘y’, and dataframe of merged data frames used as features hereafter referred to as ‘X’ was split into Training and test split of 80/20%.

After that, the clustering algorithm SVC was used to perform classification based on X. To improve performance, the library Optuna has been used to find the best hyperparameters. 9 out of 10 performed tests return correctly predicted y. For transparency, feature importance in the model has been plotted.

Changes not present in the ML model

Binary classification instead of 3 class classification resulted in overfitting as with two classes there is a big class population difference. With 3 classes for one class, there is one representative. However, it filters out the outlier and makes the remaining 2 classes more balanced;

We are taking into account genders in age groups of the population dataset.

The intention behind the project;

The intention behind our ML model was originally to predict the number of crimes that will be committed in a certain area. It ended up being a bit unfeasible because it required a lot more work than we anticipated. This made us rethink our approach and the thing we want to achieve.

The datasets behind the project;

As stated before:

Data used:

The total amount of crime committed in neighbourhoods in a year;

Number of citizens with a level of education in neighbourhoods (‘Laag’, ‘Midden’, ‘Hoog’);

Percentage of households in a low and high-income level (‘Lowincome’, ‘Highincome’);

The population of the neighbourhood is divided into 5 Age groups (‘<30 Jaar, ‘30-44 Jaar’, ‘45-64 Jaar’, ‘>= 75 Jaar’);

These datasets have mainly been gathered from the municipality’s website. These are metrics that they have provided themselves and why we had to make sure that there was no personal information, or sensitive information that might be harmful to others.

The metrics behind the project;

The original idea for our metrics was things like map charts, bar charts, and tree charts. We wanted to use these metrics to show our findings easily seeing as the data was already divided into neighbourhoods. We want to use bar charts to show the differences between the neighbourhoods. These are mainly used to clarify the data and show what areas need attention.