Prioritising Environmental Complaints (DSSGx UK 2021)

PROJECT PARTNER

The Superintendencia del Medio Ambiente (SMA) is a Chilean government agency reporting to the presidency through the Ministry of the Environment. Established in 2010, the mission of the SMA is to protect the environment and people's health, ensuring compliance with environmental regulations. It is responsible for the monitoring and enforcement of compliance with environmental regulations in Chile. As part of this remit, the SMA has powers to investigate suspected violations of environmental regulations by individuals or businesses, and to impose sanctions on those found to be in breach of their obligations.

As part of its enforcement activities, the SMA receives complaints from individuals and organisations about possible violations. These complaints vary widely in their nature and potential severity. To allocate resources effectively, SMA staff apply a triage process that assigns incoming complaints to three classes:

1. Redirect: complaints that are not relevant to the SMA remit. These are redirected to the relevant authority.
2. Archive: complaints that are relevant to the SMA remit but are not actionable, for example because key identifying information is missing or they lack a legal basis. These are archived.
3. Relevant: complaints that are relevant and actionable. These are analysed to determine whether the alleged breach of regulations is serious enough to warrant an inspection. If an inspection is carried out and violations are found, a sanctions procedure is initiated which can lead to monetary fines or other penalties.

Due to its limited resources, SMA cannot afford to inspect all the complaints that end up in the third group. It must prioritize where to allocate its efforts and has an interest in inspecting the most serious violations.

THE PROBLEM

The process of filing complaints was a paper-based process until 2018 when SMA launched an online platform for submitting complaints. The online system offered some clear advantages. Firstly, it eliminated the need to transcribe information from paper forms into SMA systems which was a slow

and error-prone process. Secondly it made the complaints procedure accessible to more people in more remote locations, enabling SMA to get more reliable information to help them protect ecosystems and public health. This accessibility created a new problem, however. Even before the introduction of the online system, SMA received thousands of complaints per year to process. Within a year of the introduction of the online system, this number had quadrupled and threatened to overwhelm SMA's ability to process the information.

The challenge SMA brought to the 2021 DSSGxUK summer programme was to build a machine learning model that could assist SMA staff in sorting and prioritising the large number of complaints that were being gathered by the online complaints system.

DATA SCIENCE FOR SOCIAL GOOD (DSSG) PROJECT

Data and modelling approach

The data available for the project consisted of the registry of past complaints that includes some structured data (facility type and sector, geographical location, previous inspection history) and free text from the complaints themselves. It was decided to build two models:

- a relevance model to classify incoming complaints into the 3 categories redirect, archive or relevant described above.
- A sanctions gravity model to predict the severity of the sanction that would be applied to relevant complaints if an inspection finds a violation of regulations.

Relevance model

Using the structured data alone, a random forest classifier reached an overall accuracy of 70%, but missed 34% of the complaints that should have been classed as relevant. This was taken as the baseline model. To improve on this, additional features were engineered from the free text data using some standard Natural Language Processing techniques. Latent Dirichlet Allocation (LDA) was first used for topic modelling. Term Frequency-Inverse Document Frequency (TF-IDF) was then used to extract the most important unigrams, bigrams and trigrams in each class. Finally Rapid Automatic Keyword Extraction (RAKE) was used to detect relevant phrases in the data with the presence or absence of RAKE keywords then providing additional information for the classifier.

Most of the common classification algorithms were tested – logistic regression, SVM, XGBoost, random forest and naïve Bayes. Starting with the original structured data field 'Complaint Type', features were added iteratively: the text features extracted from the NLP pipeline, the seasonality, facility-related features, etc. For each new combination that was fed into the model, the overall accuracy, F1-score, the confusion matrix and the discrepancy of performance between the validation set (average of 10 cross-validation folds) and the test set – to watch out for overfitting – were observed and recorded. The best model with the best performance, least overfitting issues and simplest predictors was chosen and used for future model evaluation and optimisation. In the end, the best model was a random forest model which made heavy use of NLP features in conjunction with a subset of the structured fields. After hyperparameter tuning, this model was able to attain an accuracy of 80.3%, with 12.5% of Relevant complaints missed. The model struggled to accurately predict the minority class (archive), partly due to the class imbalance but also due to the heterogeneity of the complaints in this class – there were lots of reasons why a complaint might end up archived.

Sanctions gravity model

The process for building the sanctions gravity model was almost identical with a similar NLP pipeline used to extract features from the free text. This model was designed to perform a simple binary-classification into low and high gravity sanctions. The best performing model was again a random forest model which, after hyperparameter tuning, attained 90% accuracy for the prediction of potential sanction gravity level.

Based on the performance of these prototype models, SMA estimated that, if deployed to assist their analysts in processing complaints, the use of these models could speed up redirection of complaints by 80%, the archiving of complaints by 85% and the identification of complaints that require inspection by 65%.

The code from the summer project can be found here:

- https://github.com/DSSGxUK/sma

A presentation of the results of the project from the 2021 (online) DataFest event can be viewed at:

- https://youtu.be/QO1cLMXL8Fo

FOLLOW UP WORK AND IMPACT

Following the completion of the summer fellowship project, the technical mentor of the SMA team, Amit Kohli, spent some additional months working with SMA to deploy the models within their cloud. After deployment in November 2021, the relevance model was run in observation mode, meaning that it was run in parallel with the work of the team of SMA analysts responsible for classifying incoming complaints. This enabled SMA to compare the decisions of the model to the decisions of human experts over a period of time to see how the model would perform on completely unseen data. Since the original model was trained on data obtained primarily from the period prior to the introduction of the online complaints system, the team had identified a strong likelihood of concept drift, data drift or a combination of the two. This was due to the fact that the online system was accessible to more citizens and had much wider geographical coverage.

After 9 months of observation, it was noted that the model retained a stable accuracy of approximately 80% for the “redirect” class. However, the accuracy of the “relevant” class, degraded over the same period from an initial 80% to only 50%. Surprisingly, the accuracy for the minority “archive” class actually *increased* over the same period from 16% to 40%. This lack of model stability for the “archive” and “relevant” classes showed that the expected drift effects are too strong to use the model reliably for these classes, especially since the “relevant” is the most important one for SMA. Furthermore, the improvement of model performance on the minority class during the observation period points to something technically interesting which is currently not understood. A likely resolution of this problem is to retrain the model using only data obtained from the online complaints system. At this point, a sufficient volume of such data has been accumulated and processed from the new system to make such an exercise feasible.

As of 2023, the model remains deployed in the SMA cloud but remains under observation and is not embedded in the analysts’ pipeline. SMA have indicated that they are considering performing a model retraining exercise towards the end of 2023. At this point it will be worth checking in with them again to see if comparable and stable model performance can be attained on the new data.

Amit Kohli’s presentation on the development, deployment and validation of the SMA model from the 2022 Datafest event in the Shard is available here:

- https://youtu.be/6dD_QK41Qr0?t=2818

"As Superintendent of Environment of Chile, we are pleased to have participated in the DSSGx program with the project of environmental complaints prioritization. Soon after we began working together, we noticed the compromise of the team which not only ended up translating into a high-quality product but also gave us many insights about the particular problem and the complaint process in general. It has been a wonderful experience for us and we highly recommend being part of this initiative to other organizations who want to get the most of their data and engage with a team of bright professionals."

Pablo Aguirre Hörmann, SMA