NOVA IMS Information Management School

2021/22

Predictive Methods of Data Mining

The Spies Among Us

Group Project

Supervised Models

1. DESCRIPTION

The president of the USA is getting increasingly worried that foreign forces might be infiltrating spies to steal its most precious asset: information.

This has happened in the past, causing major setbacks to the economy and loss of trust among the citizens. Many attempts have been made to contain the problem, but the population started to protest against the additional measures because people started to feel invaded in their privacy. In order to attend to general privacy concerns, the president asked his ministers for a strategy that would protect the country's information while preserving the privacy of ordinal citizens. As the minister of defence noted, based on previous data, it seems that there are some groups of people more likely to be spies than others. That is why your team of data scientists has been hired.

Previously detected cases of espionage have been meticulously registered in the general citizen database, where all citizens are anonymously registered for statistical purposes, and which is updated yearly. Your goal is to build a predictive model that, for this year's additions to the database, answers the question **"Which citizens should be placed under close surveillance?"** using a set of pre-selected variables accessible from the last year's citizen database, explained bellow.

As data scientists, your team is asked to analyse and transform as needed the data available and apply different models in order to answer the defined question in the most accurate way.

2. OBJECTIVES

Your goal is to build a predictive model that answers the question "Which citizens should be placed under close surveillance?" using the small quantity of data accessible from the USA database that contains general information about the citizens.

3. DATASETS

The data has been split into two groups:

- **Train set** - The training set should be used to build your machine learning models and assess their performance if needed. In this set, you also have the ground truth associated with each citizen, i.e., if the citizen is a spy or not.
- **Test set** - The test set should be used to see how well your model performs on unseen data. In this set, you don't have access to the ground truth, and the goal of your team is to predict that value (0 or 1) by using the model you created based on the training set. The predicted values in the test set should be submitted on Kaggle. The score of your predictions will be evaluated using the **F1 Score.**

Variable	Description
ID	ID
ID_ORIGINAL	The unique identifier of the citizen
Gender	The gender of the citizen
Age	The age (in years) of the citizen
Area_Residence	Area of residence
Household_Income	The household income
Household_Size	Number of people in the citizen's household including himself
Foreign_Citizenship	If the citizen has foreign citizenship
Frequent_Traveler	If the citizen is a frequent traveler (travels more than twice a year)
Social_Person	If the citizen is seen as social
Cellphone_Usage	Cell phone usage level of the citizen
Occupation	The type of occupation the citizen has (if employed the type of employer is mentioned – self-employed, government, private or public company)
Military_Service	If the citizen has completed military service
Political_Participation	Political participation level (considering activities such as voting in elections, starting and participating in petitions, etc.)
Satisfaction_Level	The satisfaction level with the standard of living in the country
Spy	If the citizen has been identified as a spy

4. DELIVERABLES

- A Jupiter notebook with all the code implemented to obtain the results presented in the report. The file naming format should be ”202122 PMDM GroupXX Notebook.ipynb”, where ”XX” should be your group number.
- A report that describes the analytical processes and the conclusions obtained, with at most 5000 words. The file naming format should be ”202122 PMDM GroupXX Report.pdf”, where ”XX” should be your group number.

4.1. NOTES

- All topics mentioned will be evaluated based on the report - a well-structured and succinct report will have a big weight on the evaluation.
- The jupyter notebook will be analysed only if some doubt arises during the report evaluation. If some steps are done in the Jupyter notebook but not described in the report, those will not be evaluated. As an example, imagine that you check the outliers, and at the end of your project, you decide to keep them. In the report, you should mention how you checked for outliers, what the steps were to remove them and why did you decide to keep them at the end, among other insights that can be relevant. The jupyter notebook should be delivered with all the cells already ran.
- Both the report and the code will go through a process of plagiarism checking.

5. EVALUATION CRITERIA

The following table quantifies the major evaluation criteria.

Criteria	Percentage	Maximum Grade
Introduction and Methodology	5%	1
Exploration and Understanding	12.5%	2.5
Pre-processing	15%	3
Modelling	15%	3
Performance Assessment	10%	2
Conclusions	5%	1
Visualizations	7.5%	1.5
Other predictive models (not given during classes)	5%	1
Creativity & Other Self-Study	10%	2
Model performance (kaggle)	15%	3
TOTAL	100%	20

A project that focuses only on the techniques and methodologies learned during the practical classes will have at most 17 values. The remaining 3 values are possible to achieve if contributions based on self-study and creativity are applied and clearly explained in the report (Criteria "Other predictive models" and "Creativity & Other Self-Study").

This bullet list provides some details about each aspect:

- **Model performace**: The performance obtained on the test set on Kaggle platform.
- **Other predictive models:** A theoretical explanation of the algorithm should be provided in the annex. Includes the depth and the quality of the comparative analysis provided by the different algorithms, the theoretical explanation of the algorithm itself, and the justification of the chosen parameters.
- **Creativity and Other Self-Study**: If other techniques not given during practical classes are applied, a theoretical explanation of the algorithm/technique should be provided in the annex.

All topics are evaluated through a comparison of the work provided by different groups.