

**The Spies Among Us**

Predictive Methods of Data Mining

2021/2022

2nd Semester

Ana Francisca Dias **m20211085**

Ana Rita Silva **m20200899**

André Oliveira **m20211253**

Raquel Jessen Machado **m20211092**

**Professors:**

Roberto Henriques

Carina Albuquerque

Lara Oliveira

Carlotta Lehman

Ana St. Aubin

Index

[Introduction and Methodology 3](#_Toc105892750)

[Exploration and Understanding 4](#_Toc105892751)

[**Variable: Gender** 5](#_Toc105892752)

[**Variable: Occupation** 5](#_Toc105892753)

[**Variable Political\_Participation** 6](#_Toc105892754)

[**Variable: Area\_Residence** 6](#_Toc105892755)

[**Variable: Frequent traveller** 7](#_Toc105892756)

[**Variable: Cellphone usage** 7](#_Toc105892757)

[**Variable: Household size** 7](#_Toc105892758)

[**Variable: Foreign citizen** 8](#_Toc105892759)

[**Variable: Age** 8](#_Toc105892760)

[**Variable: Household\_Income** 9](#_Toc105892761)

[**Variable: Satisfaction Level** 9](#_Toc105892762)

[Verify data quality 11](#_Toc105892763)

[Missing values 12](#_Toc105892764)

[Pre-processing 12](#_Toc105892765)

[Data Cleaning 12](#_Toc105892766)

[Missing values 12](#_Toc105892767)

[Duplicates 13](#_Toc105892768)

[Outliers 13](#_Toc105892769)

[Data Transformation 15](#_Toc105892770)

[Binning 15](#_Toc105892771)

[Feature Selection 17](#_Toc105892772)

[SelectKBest 17](#_Toc105892773)

[Scaling 17](#_Toc105892774)

[Modelling 17](#_Toc105892775)

[Gradient Boosting 17](#_Toc105892776)

[Hyperparameter Tuning 18](#_Toc105892777)

[Performance Assessment 18](#_Toc105892778)

[Conclusions 20](#_Toc105892779)

[References 21](#_Toc105892780)

[Figure 1 - Bar chart Foreign Gender Data test 5](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892781)

[Figure 2 - Bar chart Foreign Gender Data train 5](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892782)

[Figure 3 - Bar chart Occupation data test 5](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892783)

[Figure 4 - Bar chart Occupation data train 5](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892784)

[Figure 5 - Bar chart Political Participation data test 6](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892785)

[Figure 6 - Bar chart Political Participation data train 6](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892786)

[Figure 7 - Bar chart Area Residence Data test 6](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892787)

[Figure 8 - Bar chart Area Residence Data train 6](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892788)

[Figure 9 - Bar chart Frequent Traveler data test 7](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892789)

[Figure 10 - Bar chart Frequent Traveler data train 7](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892790)

[Figure 11 - Bar chart Cellphone Usage data test 7](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892791)

[Figure 12 - Bar chart Cellphone Usage data train 7](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892792)

[Figure 13 - Bar chart Household Size data train 7](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892793)

[Figure 14 - Bar chart Household Size data test 7](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892794)

[Figure 16 - Bar chart Foreign Citizenship Data train 8](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892795)

[Figure 15 - Bar chart Foreign Citizenship Data test 8](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892796)

[Figure 17 - Bar chart age Data test 8](#_Toc105892797)

[Figure 18 - Bar chart age Data train 8](#_Toc105892798)

[Figure 19 - Bar chart Household Income data test 9](#_Toc105892799)

[Figure 20 - Bar chart Household Income data train 9](#_Toc105892800)

[Figure 21 - Bar chart Satisfaction\_Level data test 9](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892801)

[Figure 22 - Bar chart Satisfaction\_Level data train 9](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892802)

[Figure 23 - correlation matrix data test 11](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892803)

[Figure 24 - Correlation matrix data train 11](https://liveeduisegiunl.sharepoint.com/sites/MPDM648/Shared%20Documents/General/202122%20PMDM%20Group16%20Report.docx#_Toc105892804)

[Figure 25 - Household\_Size Outliers in Test Dataset 13](#_Toc105892805)

[Figure 26 - Household\_Size Outliers in Train Dataset 14](#_Toc105892806)

[Figure 27 - Household\_Income Outliers in Test Dataset 14](#_Toc105892807)

[Figure 28 - Household\_Income Outliers in Test Dataset 14](#_Toc105892808)

[Figure 29 - Satisfaction\_Level Outliers Train Dataset. 14](#_Toc105892809)

[Figure 30 - Satisfaction\_Level Outliers Test Dataset. 15](#_Toc105892810)

[Figure 31 - Age Outliers Train Dataset. 15](#_Toc105892811)

[Figure 32 - Age Outliers Test Dataset. 15](#_Toc105892812)

[Figure 33 – Age Boxplot 16](#_Toc105892813)

[Figure 34 - Household\_Size Boxplot 16](#_Toc105892814)

[Figure 35 - Satisfaction\_Level Boxplot 16](#_Toc105892815)

[Figure 36 - Household\_Income Boxplot 16](#_Toc105892816)

[Figure 37 - Best model parameters 19](#_Toc105892817)

[Figure 38 - Features included in the best model. 19](#_Toc105892818)

# Introduction and Methodology

Within the scope of the Predictive Methods of Data Mining course, we were proposed to carry out a project taking on the role of data scientists.

The basis of the project is set on the assumption that the USA president needs to obtain a solution for finding the spies responsible for steeling information without compromising the privacy of USA citizens. As the minister of defence already, based on previous data, discovered that there are some groups of people that are more likely to be spies than others, our team was hired to predict which citizens should be placed under close surveillance using a set of pre-selected variables accessible from the last year’s citizen database. 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. As data scientists, your team is asked to analyse and transform as needed the data available and apply different models to answer the defined question in the most accurate way.

The dataset given to the team contains the following variables:

Table 1 - Description of each variable.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| ID | - |
| ID\_ORIGINAL | The unique identifier of the citizen |
| Gender | The gender of the citizen |
| Age | The age 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\_Transfer | If the citizen is a frequent traveller (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 |

# 

# Exploration and Understanding

The purpose of this chapter is to apply a combination of numerical and visualization techniques to summarize the data and provide understanding of the dataset. This section includes the types of variables attributes or variables that makeup the dataset, treatment of missing values and outliers and statistical analysis.

According to the information received, there are 8 000 citizens registered in the train dataset and 493 citizens in the teste dataset. The type of variable and its quantity are shown in table 2:

Table 2 - Type and quantity of each variable in test and train dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type of variable** | **Train Dataset Quantity** | **Test Dataset Quantity** |
| ID | ID | 8 000 | 493 |
| ID\_ORIGINAL | ID | 8 000 | 493 |
| Gender | Nominal | 8 000 | 493 |
| Age | Nominal | 8 000 | 493 |
| Area\_Residence | Nominal | 7 924 | 485 |
| Household\_Income | Continuous | 8 000 | 457 |
| Household\_Size | Nominal | 7 670 | 476 |
| Foreign\_Citizenship | Binary Symmetric | 7 862 | 486 |
| Frequent\_Transfer | Binary Symmetric | 7 924 | 490 |
| Social\_Person | Binary Symmetric | 7 924 | 485 |
| Cellphone\_Usage | Nominal | 8 000 | 493 |
| Occupation | Nominal | 7 876 | 489 |
| Military\_Service | Binary Symmetric | 7 924 | 485 |
| Political\_Participation | Nominal | 7 876 | 489 |
| Satisfaction\_Level | Ordinal | 7670 | 476 |

Exploring the test and train dataset variables, it was possible to conclude some aspects in each dataset.

### **Variable: Gender**

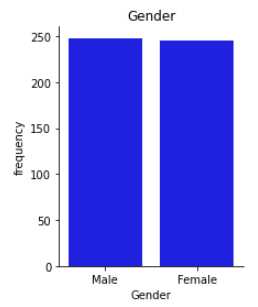
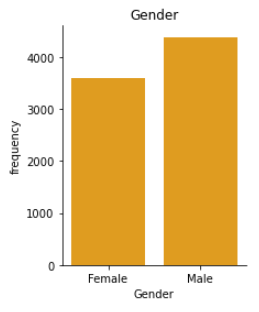


Figure 1 - Bar chart Foreign Gender Data test



In the test and train dataset the percentage of male and female citizens is about 50% each (figure1 and figure 2), which reveals that the variable “gender” won’t be a good variable for the predictive model.

Figure 2 - Bar chart Foreign Gender Data train

### **Variable: Occupation**

The variable occupation characterizes the citizen employment according to the following categories: “Private company”, “Government”, “self-employment”, “Public company, “Student”” or “Nothing”

As figure 3 and 4 show that both test and train datasets present that 70% of the citizens occupation is in a private company as the remaining citizens are divided into the remaining categories with hardly any difference.

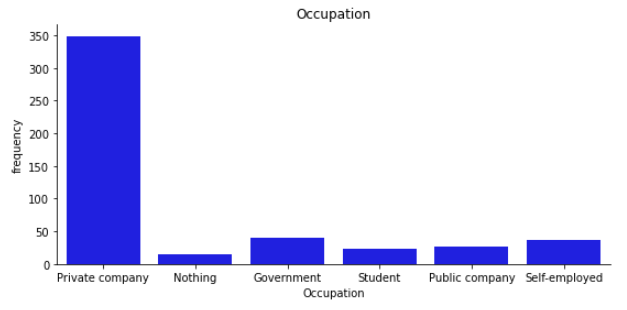




Figure 3 - Bar chart Occupation data test

Figure 4 - Bar chart Occupation data train

### **Variable Political\_Participation**

The variable Political\_Participation characterizes the citizens level of political participation in 4 categories: “No involvement”, “Some involvement”, “Strong involvement” and “Unknown” considering activities such as voting in elections, starting, and participating in petitions, and other political activities.

As the figures 5 and 6 show both in the test and train dataset, have around 70% of the citizens characterized as having no political involvement (36% and 39% in the test and train dataset respectively) or the political participation unknown (34% and 31% in the test and train dataset respectively).

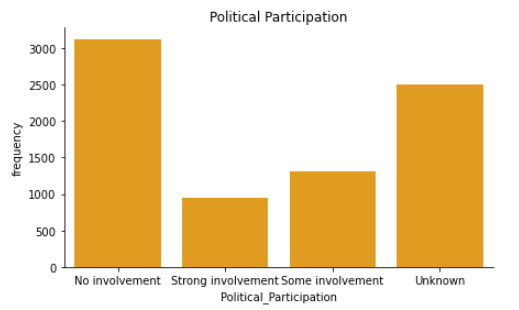
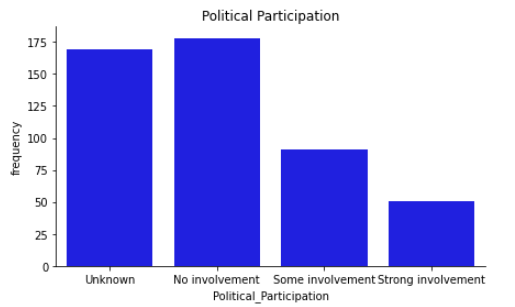
****

Figure 5 - Bar chart Political Participation data test

Figure 6 - Bar chart Political Participation data train

### **Variable: Area\_Residence**

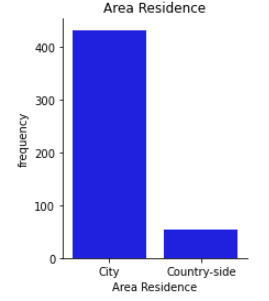


Figure 7 - Bar chart Area Residence Data test

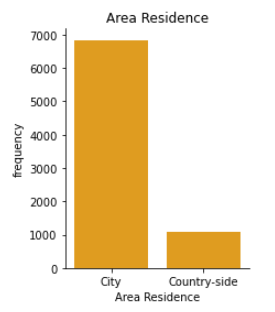
 The variable Area Residence characterizes the citizens by their home address in 2 categories: “City” or “Country-Side”. As the figures 7 and 8 show both in the test and train dataset show that around 87% of the citizens registered have their home address in the city.

Figure 8 - Bar chart Area Residence Data train

### **Variable: Frequent traveler**

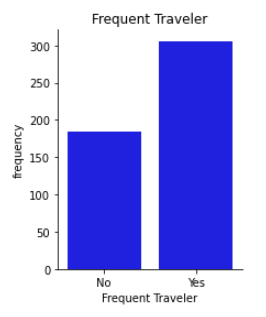
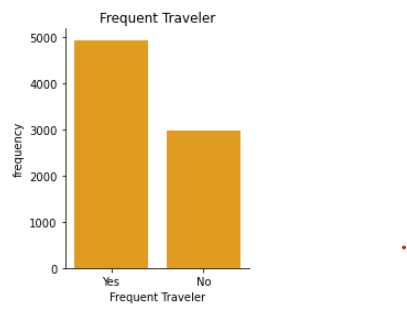
****

Figure 9 - Bar chart Frequent Traveler data test



The variable Frequent\_Traveler identifies if the citizen is a frequent traveller or not. As the figures 9 and 10 show, 55% to 58% of citizens, in both test and train datasets, are frequent travellers.

Figure 10 - Bar chart Frequent Traveler data train

### **Variable: Cellphone usage**

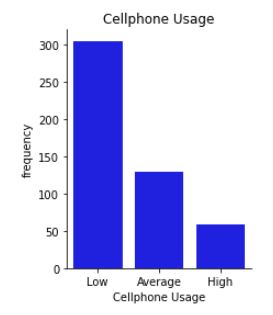


Figure 11 - Bar chart Cellphone Usage data test

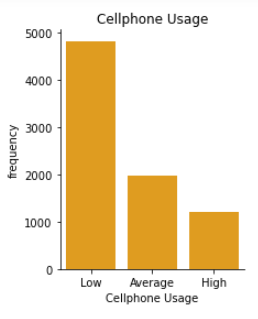
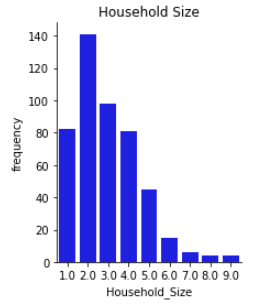


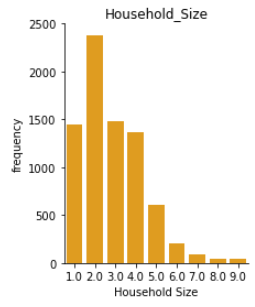
Figure 12 - Bar chart Cellphone Usage data train

The variable cellphone\_usage identifies how often the citizen uses their cell phone. This variable is categorized in 3 levels: Low, Average and High.

As the figures 13 and 14 show, the majority of citizens has a low level of cell phone usage (~60%), followed by average (~25%).

### **Variable: Household size**



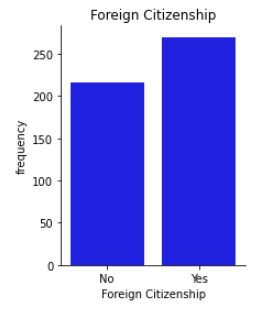


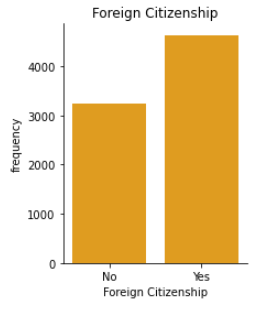
The variable household\_size identifies the number of people in the citizen’s household including himself. As the figures 13 and 14 show, majority of the registered citizens has a between a 1 and 4 household size (~86%).

Figure 13 - Bar chart Household Size data test

Figure 14 - Bar chart Household Size data train

### **Variable: Foreign citizen**





As figure 15 and 16 show, there are ~55% of citizens with a foreign citizenship registered in both train and datasets.

Figure 15 - Bar chart Foreign Citizenship Data train

Figure 16 - Bar chart Foreign Citizenship Data test

### **Variable: Age**

Both test and train datasets contain registrations of citizens between 18 and 88 years old. Analyzing the age variable, the figures 17 and 18, it possible to identify that there is a higher count of citizens with ages between 25 and 50 years old.

Chart, bar chart, histogram

Description automatically generated

Figure 17 - Bar chart age Data test

Train

Chart, bar chart

Description automatically generated

Figure 18 - Bar chart age Data train

### **Variable: Household\_Income**

The train and test data show a different in the values of the household\_Income variable. As in the train dataset the majority of the registered citizens as a household income between 7.000€ and 8. 261€, in the test dataset the majority of the citizens has a household income between 0€ and 1.261€. This difference values between the datasets can lead to a faulty prediction model.

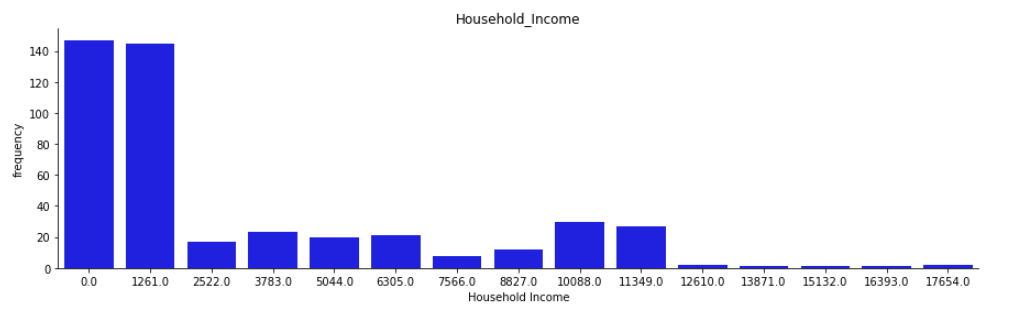


Figure 19 - Bar chart Household Income data test

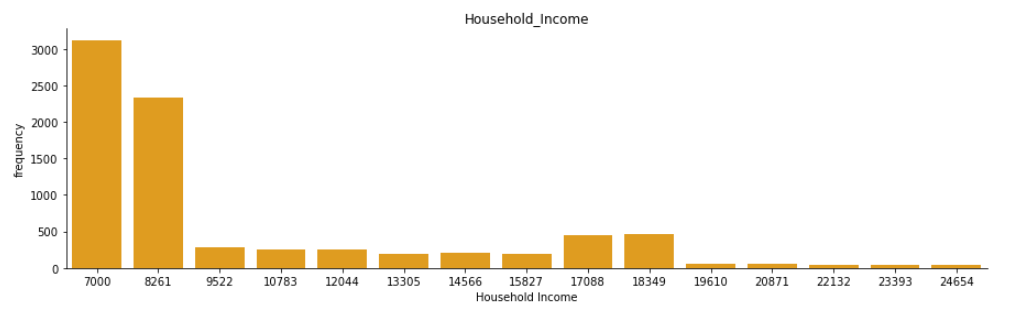
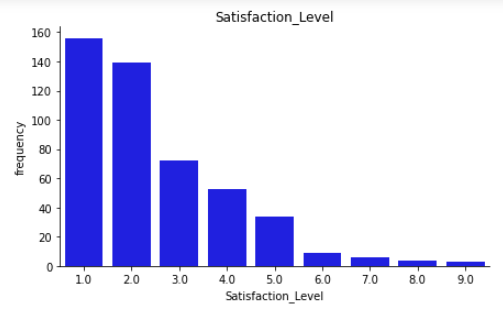


Figure 20 - Bar chart Household Income data train

### **Variable: Satisfaction Level**

The variable satisfaction level represents the level of satisfaction of the citizens with the standard of living in the country. This variable is categorised between 9 levels of satisfaction.

As figures 21 and 22 demonstrate there is a slight difference between the datasets.

In the train set the majority of citizens has a satisfaction level of 1 as in the test dataset the majority of the citizens has a satisfaction level of 2. However, both datasets show that in average, the citizens have a low level of satisfaction of standard living conditions.

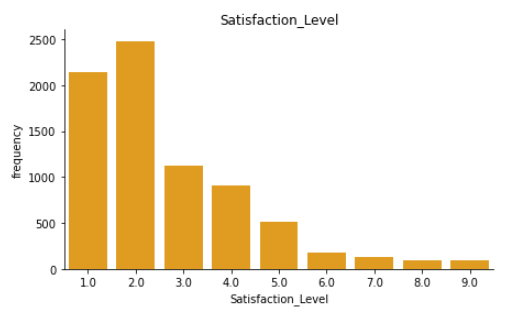


Figure 21 - Bar chart Satisfaction\_Level data test

Figure 22 - Bar chart Satisfaction\_Level data train

Initially, the basic descriptive statistics were plotted to get an overall view of the dataset, measuring the central tendency and the distribution of each attribute.

Table 3 - Summary statistics for all variables of the train dataset

Graphical user interface, application, table

Description automatically generated

Table

Description automatically generated

Table 4 - Summary statistics for all variables of the test dataset

Both household\_size and age variables are highly correlated with satisfaction\_level variable (figure 23 and 24). Due to this high correlation, producing ratios and combined features of this variables can lead to improvement of the model performance.

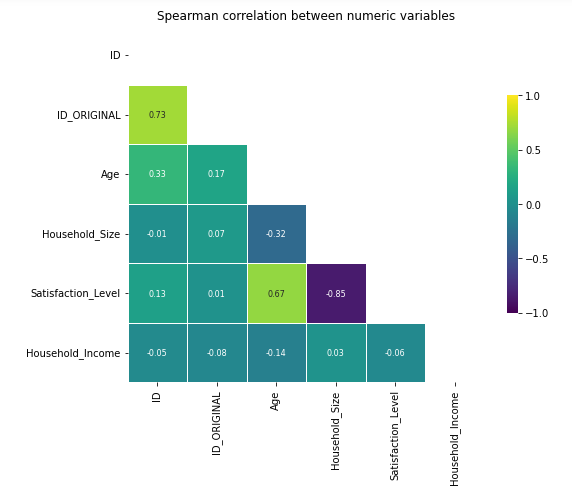


Figure 23 - correlation matrix data test

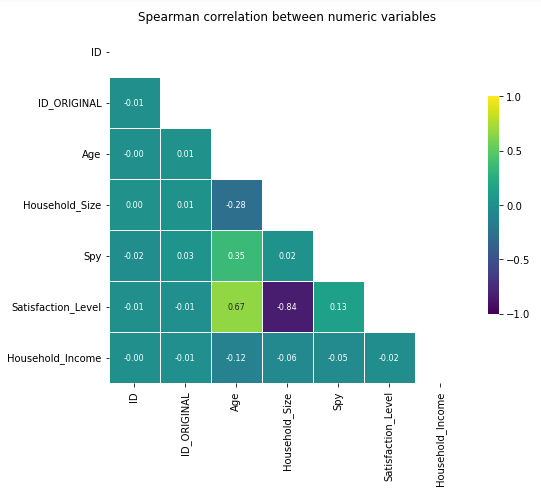


Figure 24 - Correlation matrix data train

## Verify data quality

Data quality is a crucial part of the process, as data quality directly impacts the final output. It involves cleaning, filtering, and integrating data, to produce a high-quality dataset to be used on the modelling stage.

### Missing values

The preliminary assessment has shown that there are missing values on some of the features of the test and train dataset. Table 5 below, represents the number of missing values in Test and Train Dataset.

Table 5 – Missing Values

|  |  |  |
| --- | --- | --- |
| **Variable** | **Test Dataset Missing values** | **Train Dataset Missing Values** |
| ID | 0 | 0 |
| ID\_ORIGINAL | 0 | 0 |
| Gender | 0 | 0 |
| Age | 0 | 0 |
| Area\_Residence | 8 | 76 |
| Household\_Income | 36 | 0 |
| Household\_Size | 17 | 330 |
| Foreign\_Citizenship | 7 | 138 |
| Frequent\_Traveler | 3 | 77 |
| Social\_Person | 8 | 76 |
| Cellphone\_Usage | 0 | 0 |
| Occupation | 4 | 124 |
| Military\_Service | 8 | 76 |
| Political\_Participation | 8 | 124 |
| Satisfaction\_Level | 17 | 330 |

# Pre-processing

The pre-processing phase is responsible for preparing the data for modelling, and to achieve that we must clean and transform the raw data. We proceeded to clean the data by checking for missing values, duplicates and outliers. Afterwards we performed transformations in the data such as binning, feature selection and scaling.

## Data Cleaning

### Missing values

In the data understanding and data exploration phase it was possible to detect the existence of missing values on some of the features of the test and train dataset. When analysing in detail the variables with missing values, the full picture and treatment applied can be summarized as present in table 6:

Table 6 - Test and Train Dataset Missing Values : number, percentage and treatment.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Test Dataset**  **Missing values** | | **Train Dataset**  **Missing Values** | | **Missing Values treatment** |
|  | **Number** | **Percentage** | **Number** | **Percentage** |  |
| ID | 0 |  | 0 | - | Not applicable |
| ID\_ORIGINAL | 0 |  | 0 | - | Not applicable |
| Gender | 0 | - | 0 | - | Not applicable |
| Age | 0 | - | 0 | - | Not applicable |
| Area\_Residence | 8 | 1,62% | 76 | 0,95% | As the variable is categorical, the missing values were replaced by “unknown” |
| Household\_Income | 36 | 7,3% | 0 | - | The missing values were replaced by “0” being the minimum income salary possible. |
| Household\_Size | 17 | 3,45% | 330 | 4,13% | The missing values were replaced by “1”, being the minimum number of people in the household including himself |
| Foreign\_Citizenship | 7 | 1,42% | 138 | 1,73% | As the variable is categorical, the missing values were replaced by “unknown” |
| Frequent\_Traveler | 3 | 0,61% | 77 | 0,96% | As the variable is categorical, the missing values were replaced by “unknown” |
| Social\_Person | 8 | 1,62% | 76 | 0,95% | As the variable is categorical, the missing values were replaced by “unknown” |
| Cellphone\_Usage | 0 | - | 0 | - | Not applicable |
| Occupation | 4 | 0,81% | 124 | 1,55% | As the variable is categorical, the missing values were replaced by “unknown” |
| Military\_Service | 8 | 1,62% | 76 | 0,95% | As the variable is categorical, the missing values were replaced by “unknown” |
| Political\_Participation | 4 | 0,81% | 124 | 1,55% | As the variable is categorical, the missing values were replaced by “unknown” |
| Satisfaction\_Level | 17 | 3,45% | 330 | 4,13% | As the variable is numerical, the missing values were replaced by 0 |

There are different approaches to handle missing values, such as replacing them with the mean or median/mode for a random value (numerical/categorical variables) or using kNN Imputer (match a point with its closest k neighbours). After testing these options, we concluded that replacing for a constant like “unknown”, “0” and “1”, was better for the performance of the model.

### Duplicates

We checked for duplicate values, but none were found.

### Outliers

We proceeded to check the distribution of outliers through boxplots, however, we decided to not remove any outliers because the Kaggle submission required the same number of rows as the original train dataset. Also, comparing the outlier’s ratio we concluded that the number of outliers wouldn’t affect the performance of the model.

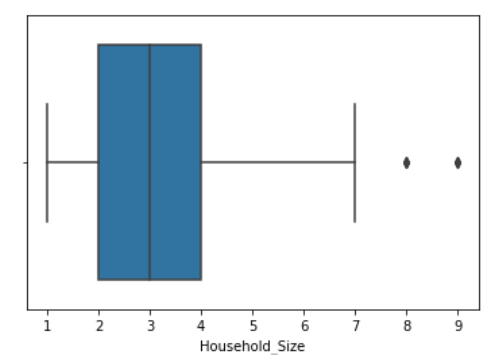


Figure 25 - Household\_Size Outliers in Test Dataset

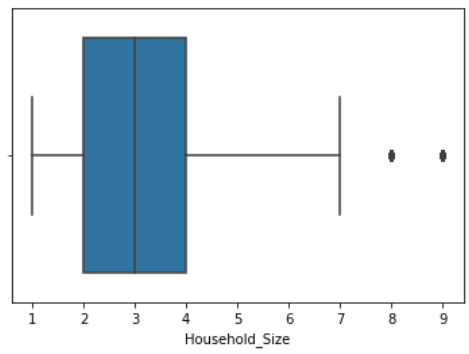


Figure 26 - Household\_Size Outliers in Train Dataset

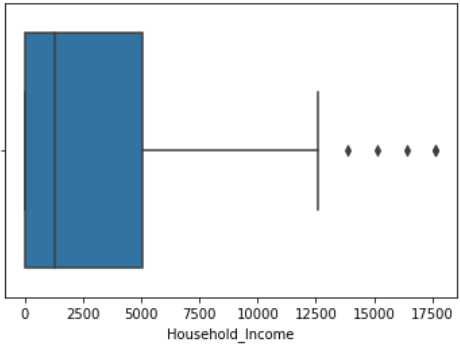


Figure 27 - Household\_Income Outliers in Test Dataset

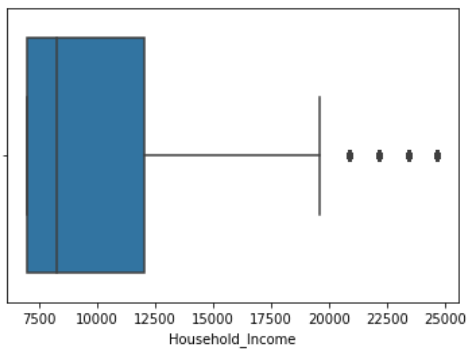


Figure 28 - Household\_Income Outliers in Test Dataset

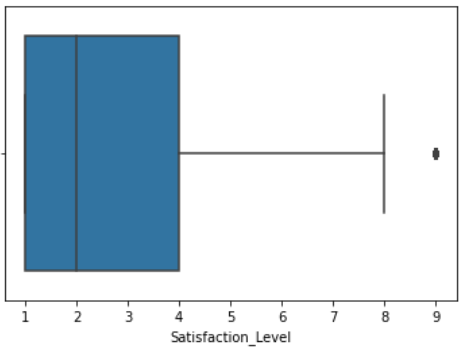


Figure 29 - Satisfaction\_Level Outliers Train Dataset.

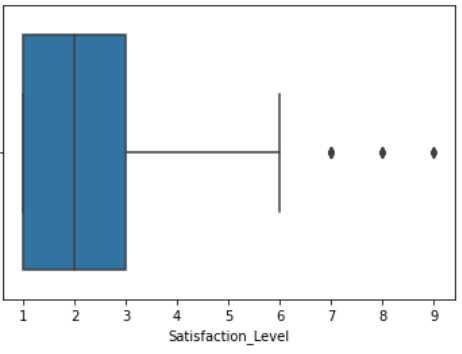


Figure 30 - Satisfaction\_Level Outliers Test Dataset.

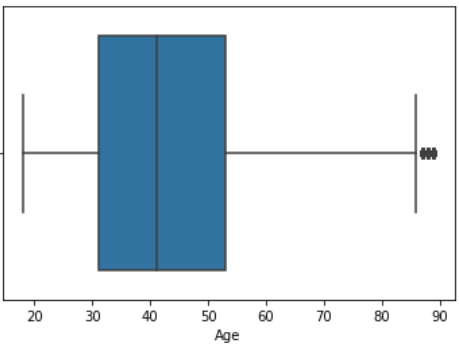


Figure 31 - Age Outliers Train Dataset.

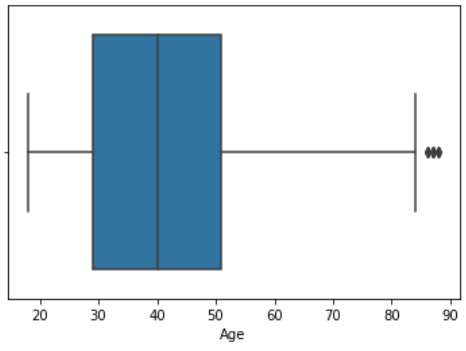


Figure 32 - Age Outliers Test Dataset.

## Data Transformation

To prepare the data for the modelling, we opted to do label encoding, a good technique to handle categorical variables. Label encoding replaces categorical values with numerical values.

### Binning

We opted to bin the quantitative variables (Age, Household\_Income, Household\_Size and Satisfaction\_Level) in order to partition the numerical variables into bins.

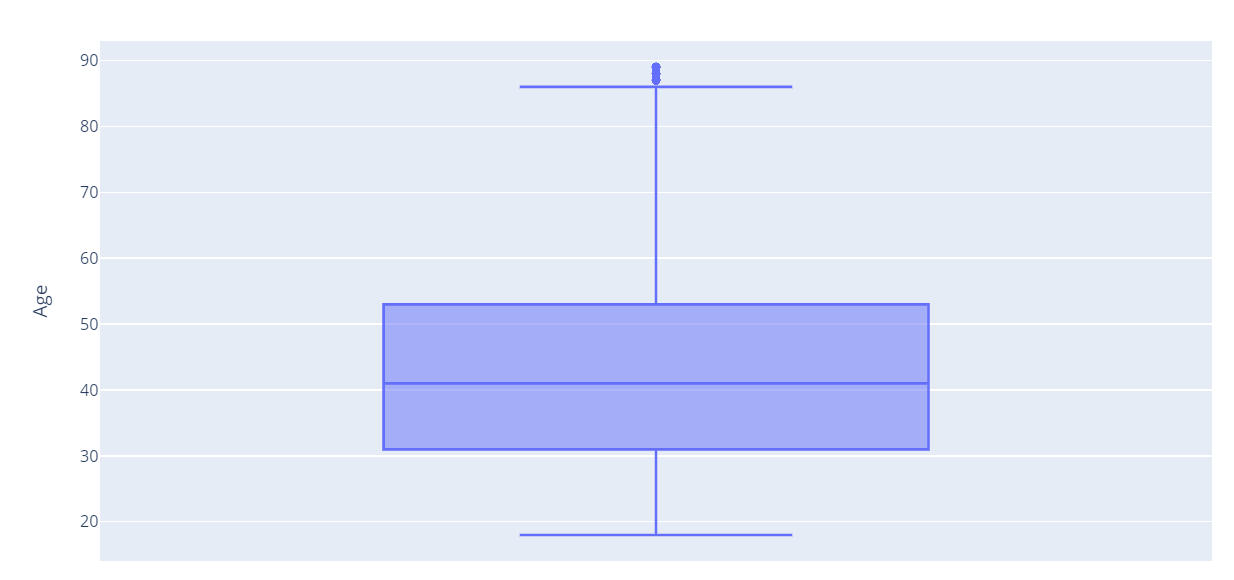


Figure 33 – Age Boxplot

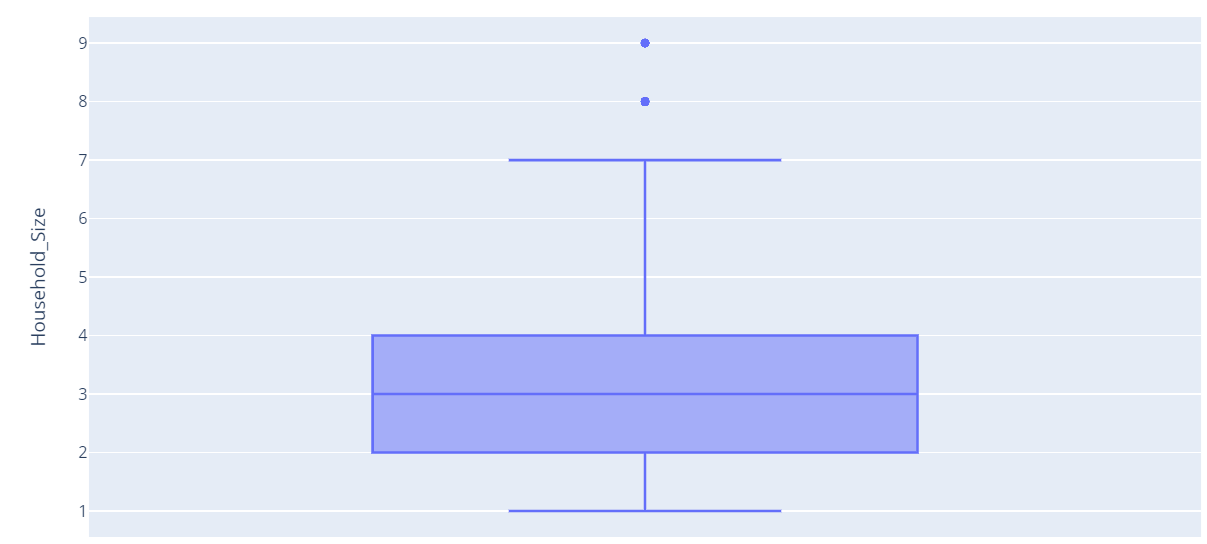


Figure 34 - Household\_Size Boxplot

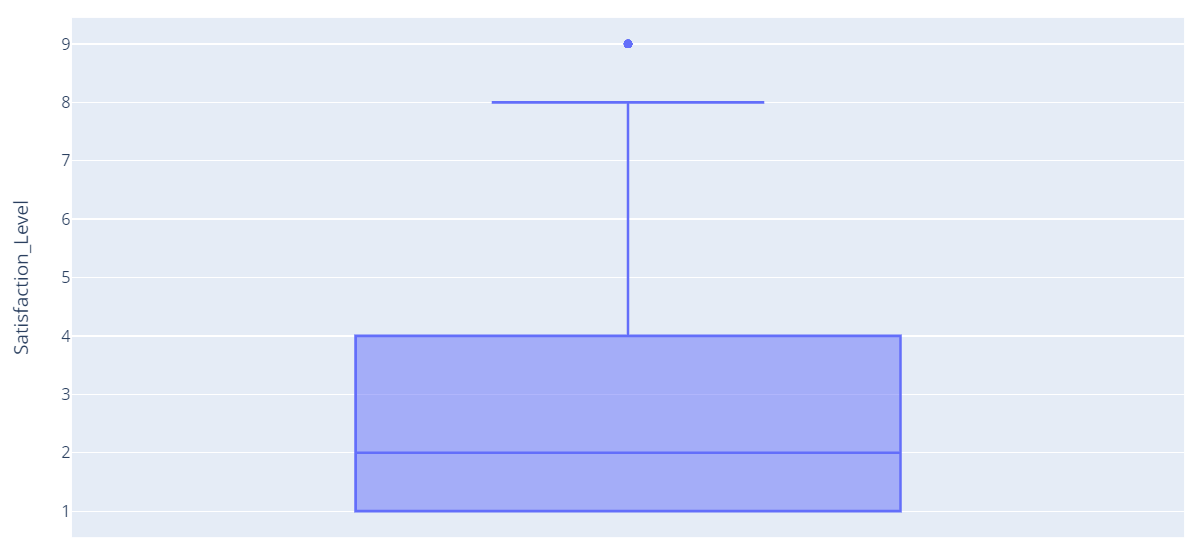


Figure 35 - Satisfaction\_Level Boxplot

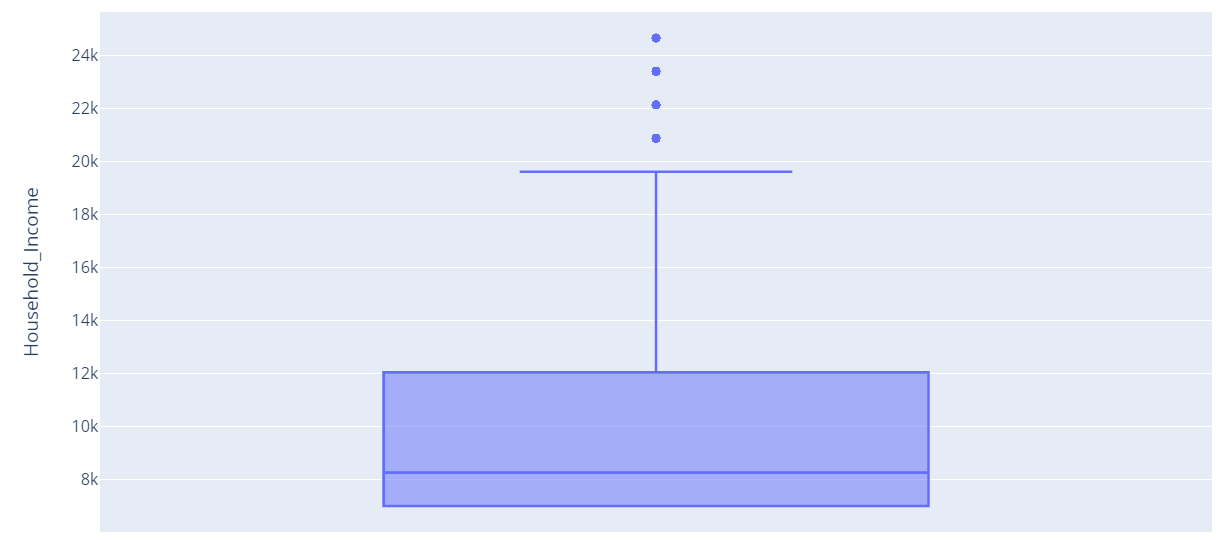


Figure 36 - Household\_Income Boxplot

### Feature Selection

In supervised learning methods, when preparing a large dataset for training, it is important to select the best features to include, and it is possible to that resorting to feature selection.

Feature selection is a pre-processing tool that is effective and efficient in improving the performance of a machine learning model, especially in cases of high dimensionality.

When dealing with large amounts of features (high dimensionality), learning models tend to overfit and that can reduce the performance of a model. Therefore, feature selection directly selects a subset of relevant features whilst building a model.

Resorting to this has a lot of advantages, such as: improvement of learning performance, increasing computational efficiency, decreasing memory storage, and constructing better models. [1]

The purpose of this step is to reduce the number of input variables to the most useful to the model, in order to predict the target variable, in this case, Spy.

Support Vector-Machine (SVM) and Neural Networks can be affected by not relevant predictors, therefore using irrelevant attributes can decrease the performance of the model.

Linear and logistic regressions can be sensible to correlated predictors, thus dealing with redundancy will help reduce multicollinearity and result in a better predictive model. [2]

### SelectKBest

To implement feature selection, we opted for SelectKBest. SelectKBest is a method from feature selection, that will help select the features according to k highest score.

We used SelectKBest to score the features against the target variable using the function f\_classify. Since we are dealing with quantitative attributes, f\_classify computes the ANOVA F-value between each variable and the target vector. [3]

Analysis of Variance (ANOVA) is a statistical method used to check the means of two or more groups that are significantly different from each other.

### Scaling

Due to existing some data mining algorithms affected by differences in the attribute’s ranges, it is important to perform scaling before modelling. Differences in ranges can contribute differently to the model and affect its results, for example, variables with a greater variation in their range, can lead to a higher impact on the results. [4]

Since in this dataset we have variables measured at different scales, we performed normalization before modelling. The method we chose is Min-max Normalization and it is one of the most commonly used.

# Modelling

### Gradient Boosting

The selected model for this project was Light GBM – a fast, distributed, high-performance gradient boosting framework based on decision tree algorithm. [5] Gradient boosting is a type of machine learning boosting, a method for creating an ensemble that is a combination of simple individual models that together create a more powerful new model.

Boosting works by fitting an initial model (e.g. decision tree) to the data. Then a second model is built that focuses on accurately predicting the cases where the first model performs poorly. The combination of these two models is expected to be better than either model alone. The boosting process is repeated many times, as each successive model attempts to correct the shortcomings of the combined boosted ensemble of all previous models. [6]

Gradient boosting relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error. The name gradient boosting arises because target outcomes for each case are set based on the gradient of the error with respect to the prediction. Each new model takes a step in the direction that minimizes prediction error, in the space of possible predictions for each training case. [6]

### Hyperparameter Tuning

Hyperparameter tuning is the task of choosing a set of optimal hyperparameters for a machine learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins, as a result, different models have different hyperparameters to be set. In practice, the aim of hyperparameter tuning is to find the hyperparameters of a given machine learning algorithm that return the best performance as measured on a validation set. [7]

The selected method of hyperparameter tuning for this project was Bayesian Optimization. It is a sequential model-based optimization algorithm that uses the results from the previous iteration to decide the next hyperparameter value candidates. It works by creating a probabilistic model, mapping hyperparameters to a probability of a score on the objective function . [8] This model is called a surrogate for the objective function and is represented as . The surrogate is much easier to optimize than the objective function, therefore, Bayesian methods work by finding the next set of hyperparameters to evaluate on the actual objective function by selecting hyperparameters that perform best on the surrogate function. [8]

This method is efficient because it selects hyperparameters in an informed manner. By prioritizing hyperparameters that appear more promising from past results, it can find the best hyperparameters in less time (fewer iterations) than both grid search and random search methods. [8]

# Performance Assessment

In order to avoid data leakage, the test set has only been used to assess the model, while the train set has been split into 5 stratified cross validation folds (as the target feature is unbalanced). Each fold is divided into a train and validation subsets – allowing the gradient boosting model to be trained, while the validation subset is used to allow the Bayesian Optimization algorithm to search for the best hyperparameters, and the SelectKBest algorithm to select the right set of features. The final set of hyperparameters and features is based on the fold that presents better evaluation metric scores. Finally, the model is retrained automatically with all training data.

The evaluation metric selected to assess the model’s performance was the F1-score, as it combines the precision and recall of a classifier into a single metric by taking their harmonic mean:

In binary classification, precision attempts to assess the proportion of positive identifications that were actually correct, while recall estimates the proportion of actual positives that were identified correctly. The F1-score reaches its best value at 1 and worst value at 0. A low F1 score is an indication of both poor precision and poor recall and the other way around.

When analyzing Figure 37, we can see that the Bayesian Optimization algorithm selected the hyperparameters carefully to avoid overfitting – pruning the maximum depth of the decision trees to reduce complexity.

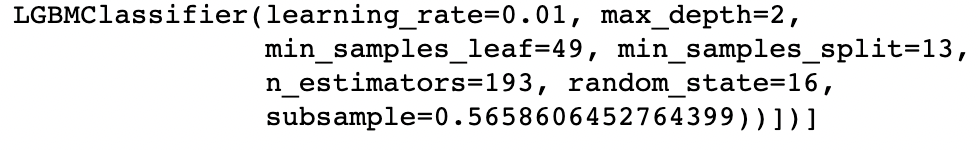


Figure 37 - Best model parameters

The final set of features to model – provided by the SelectKBest algorithm in Figure 38 – has selected 10 columns to model (support = True), including binned variables. The algorithm has filtered variables related with Gender, Household Size, Occupation and Military Service. The inclusion of binned variables has actually increased the model’s performance significantly.

Uma imagem com mesa

Descrição gerada automaticamente

Figure 38 - Features included in the best model.

The model evaluation results in Table 7 show that the model is not overfitting, as the f-score on the train set is lower than on the test set, by a small gap. The model could potentially be underfitting, but when its complexity was increased, it didn’t result on a better performance.

Table 7 - Model evaluation f1-score results.

|  |  |
| --- | --- |
| **Set** | **F1-score** |
| Average F1-score Cross-validation | 0.73862 |
| Average F1-score Train Set | 0.75010 |
| F1-score Train Set by Class ([0 1]) | [0.76278 0.73742] |
| Average F1-score Test Set | 0.74956 |
| F1-score Test Set by Class ([0 1]) | [0.76388 0.73524] |
| F1-score Kaggle | 0.72164 |

Table 8 - Model evaluation precision and recall results.

|  |  |
| --- | --- |
| **Set** |  |
| Precision Train Set by Class ([0 1]) | [0.75847458 0.74208861] |
| Recall Train Set by Class ([0 1]) | [0.76714286 0.7328125] |
| Precision Test Set by Class ([0 1]) | [0.75495751 0.74511401] |
| Recall Test Set by Class ([0 1]) | [0.77302393 0.72561459] |

The performance on the Kaggle competition set has been worse than on the test set by a wide margin, indicating that the model cannot capture every pattern perfectly. As presented in Table 5, the f1 score of spies (class 1) is always by far lower than in the case of non-spies (class 0), showing that the model cannot define a spy as well as it can distinguish a non-spy.

The root cause of this model behavior can be derived from Table 8, we the values of precision and recall by are presented. The precision values are quite balanced by class, indicating that the model is correctly categorizing the positive identifications, in practice, it is correct around 74% of the time when it predicts a person is a spy. However, the recall values are far lower categorizing spies than non-spies, indicating that the model struggles to correctly identify the actual positives, in other words, it correctly identifies 77% of all non-spies but only 72% of all spies.

# Conclusions

During the modelling phase, we we’re able to determine, through various attempts (trial and error), that some processes helped the model to be more effective and others didn’t.

We concluded that using the Min-max feature scaling was more effective than standard scaling or the use of no scaling at all.

To avoid overfitting, the use of Bayesian Optimization hyperparameter tuning turned out to be effective. Cross validation allowed to improve the model’s final score effectively.

We also decided to bin the features in the model and that has proven to be more effective, allowing the gradient boosted decision trees to have more data.

After trying both, we concluded that SelectKBest has revealed to be more effective than Recursive Feature Elimination (RFE) to perform feature selection.

The potential use of correlations to produce ratios and combined features, as we mentioned in the data understanding phase, turned out to not be effective in improving the model’s performance, so we discarded it.

Resorting to sampling techniques, such as undersampling or oversampling, to balance uneven datasets has also proven not to be very effective. We considered this happened, mostly because there are many unbalanced features in the dataset – and on the current project this was only tested to balance the target feature (Spy).

Throughout the modelling phase, different models were tested such as kNN, SVC, Naive Bayes, Logistic Regression and Ada Boost, but gradient boosting models were more effective due to class imbalance in the data.

On a final note, we consider that implementing deep learning techniques, such as Neural Networks, could have been a powerful way to improve the final score.

Also, it would be beneficial to include more data from external sources, to classify a spy and therefore improving the model’s results.

# References

|  |  |
| --- | --- |
| [1] | J. Li, K. Cheng and S. Wang, "Feature Selection: A Data Perspective," p. 45, 2017. |
| [2] | "Sklearn.feature\_selection.SelectKBest.," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature\_selection.SelectKBest.html. [Accessed 12 06 2022]. |
| [3] | "ANOVA F-value For Feature Selection," [Online]. Available: https://chrisalbon.com/code/machine\_learning/feature\_selection/anova\_f-value\_for\_feature\_selection/. [Accessed 12 06 2022]. |
| [4] | "Everything you need to know about Min-Max normalization: A Python tutorial," [Online]. Available: https://towardsdatascience.com/everything-you-need-to-know-about-min-max-normalization-in-python-b79592732b79. [Accessed 12 06 2022]. |
| [5] | "LightGBM’s Documentation," June 2022. [Online]. Available: https://lightgbm.readthedocs.io. |
| [6] | "Gradient Boosting Explained – The Coolest Kid on The Machine Learning Block," [Online]. Available: https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/. [Accessed June 2022]. |
| [7] | "Hyperparameter Optimization," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Hyperparameter\_optimization. [Accessed June 2022]. |
| [8] | W. Koehrsen, "A Conceptual Explanation of Bayesian Hyperparameter Optimization for Machine Learning," Towards Data Science, [Online]. Available: https://towardsdatascience.com/a-conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f. [Accessed June 2022]. |