Quantitative Analysis of the Romanian Private Security Market. A Machine Learning Approach

Alexandru-Costin Băroiu

Abstract Market segmentation and analysis have benefited greatly from advancements in Machine Learning. Supervised and unsupervised learning techniques have been applied with great success in market analysis. This paper proposes such an approach that aims to first introduce a new dataset and identify the groups in which the market is segmented, by applying k-means++ clustering, and then to develop a well performing classifier that would correctly identify future companies and place them in the previously identified clusters for a disregarded industry, the Romanian private security market. First, a clustering algorithm is applied to group the companies into clusters. Then, the results are analyzed and findings are discussed about the market segmentation. 6 distinct groups are identified and the main factors that differentiate the companies are number of employees and turnover. Second, multiple classification algorithms are trained and benchmarked in order to find the best performing model. Due to the nature of the data, which is heavily imbalanced, sampling algorithms and weights adjustments are applied in order to improve model performance. Lastly, the best combination of classification algorithm and sampling technique is presented for the given data. The best performing model is a Multi-Layer Perceptron network, with an average F-score of 71.8. The best performing technique used to counter imbalanced data is ADASYN, with an average F-score of 69.9. The best performing combination and the best result overall is achieved by the Multi-Layer Perceptron in tandem with Random Oversampling, with an F-score of 83.3.

Keywords Market segmentation · Market analysis · Machine learning

1 Introduction

Even though it isn't a novel approach, Machine Learning (ML) has gained increased popularity for the past decade when it comes to applications in business related fields. Regressions models have long been used to make predictions, especially in financial

A.-C. Băroiu (✉)
Bucharest University of Economic Studies, 010552 Bucharest, Romania
e-mail: baroiualexandru12@stud.ase.ro

M. Busu (ed.), *Digital Economy and the Green Revolution*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-031-19886-1_1

markets, and classification algorithms have been applied to a plethora of fields, such as business intelligence.

Another application of ML widely used is clustering, applied when there is latent knowledge in data that needs to be extracted. One such example is market segmentation. This field has been approached at large, with research raging from tourism to real estate. The prospect of grouping businesses or consumers and gain a better understanding of what motivates them has proven to be a very lucrative area of interest.

This paper proposes such an approach for the Romanian private security market. Various press articles have been published on this subject, often as a critique to its highly controversial nature. The industry has often associated with corruption, political affiliation and tax evasion. As such, this market could prove to be a very lucrative one when it comes to conducting business analysis and to developing a better understanding of how Romanian companies operate in a still emerging capitalist market. However, no research has been conducted for this market and its potential for knowledge has not yet been tapped. A first endeavor will be presented in this paper, which will cluster the Romanian private security companies and then will build a classifier that can quickly identify future companies and place them in their respective class.

2 Literature Review

2.1 Clustering

Clustering is the grouping of heterogeneous data into more homogenous groups. There are different types of algorithms: based on the process of cluster forming, agglomerative or partitioning clustering, and based on the organization of the data, centroid-based, density-based, distribution-based and hierarchical-clustering. An exhaustive list of clustering algorithms was formed by Xu and Tian [1].

Clustering is of interest for this paper due to its large adoption in related research, analyzing the business dynamics through the means of grouping. One example of clustering is when it is used to gain insights from companies in the hospitality and tourism sector. Tourism is an active research field for clustering, where it is often used to group customers and get a better understanding of their behavior [2, 3].

Another research field that leans on clustering is real estate, where market segmentation and analysis are essential. It has been used to analyze the impact of policies on the real estate sector or to group listings to gain latent knowledge [4]. Research in financial markets has also used clustering to improve prediction quality [5].

As previously stated, clustering is widely used in research. Mainly, it is applied to market segmentation. Still, comprehensive guidelines for developing robust Business to Business (B2B) market segments are sparse, with few frameworks being present [6]. Hurdles in B2B market segmentation are the vast majority of the market

segmentation literature deals with consumer markets [7], industrial marketers have been slower to adopt market segmentation beyond mere traditional industry segments [8] and information searching and purchasing in B2B contexts are more formalized than in consumer (B2C) contexts [9].

The clustering algorithm that will be implemented is k-means, with the k-means++ variation for clusters initialization [10]. K-means aims to partition n observations into k clusters, with observation belonging to the cluster with the nearest. K-means clustering minimizes within-cluster variances, or squared Euclidean distances [11]. K-means++ was shown to offer excellent time complexity and good cluster purity [12]. It is an algorithm for choosing the initial values for the *k*-means clustering algorithm.

2.2 Classification

Classification is a ML technique that enables algorithms to observe instances and correctly place them in a certain class. Classification is part of the supervised learning branch of ML and it has been widely applied in research. For example, sentiment analysis is a popular research field where classification is the main task of the algorithms used. Labeled data, texts accompanied by a certain sentiment, is processed by a model that enables it to later recognize the sentiment of new texts.

One interesting approach of using ML classification is the framework for analyzing financial behavior using machine learning classification of personality through handwriting analysis. The paper details the evaluation of handwriting features for one of the Big Five Personality Traits. The 7 features that the results yielded were evaluated with machine learning techniques. The author found that individuals who scored high on extraversion were likely to overspend and to invest in risky financial products [13].

Some of the most popular classification algorithms are Support Vector Machine (SVM), DecisionTrees (DT) and K Nearest Neighbors (kNN). A SVM seeks to minimize the upper threshold of its classifications [14]. It achieves this by transforming training examples from their original dimension space to another space, with a greater number of dimensions, where a linear separation is approximated by a hyper plane [15]. The SVM seeks to minimize the margin of the classification hyper plane during the training stage. Kernel functions enable the transformation from the space of original dimensions to the space in which classifications are done [16].

DT involve subdivision of the data into subsets separated by the values of the input variables until the basic classification unit is obtained. One of the most popular DT is the Random Forest (RF) algorithm. The algorithm is based on the combination of the consensual classifications of the most accurate trees into a single one [17]. The combination of decision trees in the RF technique can be used in regressions or classifications.

For kNN, the k in the k-nearest neighbor is a positive integer. The input will consist of the nearest k training examples in classification within a field. In this form, the

result is class membership. This would recognize a new entity among its k closest to the class most familiar neighbors. If k = 1 then the object is assigned to the class of the single nearest neighbor.

Another classification approach that has gained popularity for the past half-decade is the use of neural networks. Neural networks model biological processes [18]. The basic unit of these networks, the neuron or perceptron, emulates the human equivalent, with dendrites for receiving input variables to emit an output value [19], which can serve as input for other neurons. The layers of basic processing units of the neural networks are interconnected, attributing weights for each connection [20], which are adjusted in the learning process of the network [21]. The first training phase optimizes not only the interconnections between the layers of neurons but also the parameters of the transfer functions between one layer and another, thus minimizing the errors. Finally, the last layer of the neural network is responsible for summing all the signals from the previous layer into just one output signal, the network's response to certain input data [5].

2.3 Imbalanced Data

One hurdle that classification algorithms may encounter is the imbalanced class distribution in a dataset. This means that, for example, for a binary classification algorithm the positive class has 90% of the observations while the negative class has only 10% of the observations. Such a model could have a classification accuracy of 92%, which seems to be an excellent performance. But, for this case, a random classifier would net an accuracy of 90%, the minimum achievable. As such, accuracy is no longer a great assessment of model performance. The preferred metric for benchmarking in this case is f-score, the harmonic mean between precision and recall. This metric pays more attention to mislabeled examples and, as such, offers a better evaluation for competing models.

In the case of imbalanced data, there are a few techniques that have been developed and studied that can improve model performance. The ones implemented in this paper are: RandomOverSampler (ROS), RandomUnderSampler (RUS), Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic (ADASYN) sampling approach and class and sample weights.

ROS oversamples the minority class or classes by randomly picking samples with replacement. RUS undersample the majority class or classes by randomly picking samples with or without replacement.

SMOTE is an oversampling approach in which the minority class is oversampled by creating synthetic examples rather than by oversampling with replacement. Therefore, the minority class is oversampled by taking each minority class sample and introducing synthetic examples along the line segments, by joining any or all of the k minority class nearest neighbors. Neighbors from the k nearest neighbors are randomly chosen, depending on the amount of oversampling required [22].

ADASYN uses a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for minority class examples that are harder to learn compared to those minority examples that are easier to learn [23].

Class weights are used to make the classifier pay more attention to certain classes, while sample weights are used to make the classifier pay more attention to certain samples.

Therefore, from the literature previously presented and the knowledge in the field, two research questions are postulated:

1. *Is the Romanian private security market mature?*
2. *What is the best solution to apply on the resulting market grouping to achieve the best classification results?*

3 Methodology

The research consists of a quantitative analysis, conducted on data from secondary sources. The pipeline is as follows: data preprocessing, unsupervised learning, clustering, to label the data and then applying supervised learning, classification, on the newly labeled data.

3.1 Data

The dataset is obtained from external sources, courtesy of www.securitateinromania.ro. It is an aggregate of the short balance sheet from the 2019 reported fiscal year (researched period is 01.01.2019–31.12.2019) of all the Romanian private security companies, registered under CAEN code 8082. The total number of companies and entries in the dataset is 1.504. From the attributes present in the dataset for each company, only the following are used to conduct the research: no. of employees, turnover, net profit, profit margin and turnover/employee. Feature engineering was performed on the dataset, adding a new column: net profit/employee.

3.2 Preprocessing

Before applying the clustering algorithm, data preprocessing is required. Firstly, rows containing Nan values will be removed. This leaves the dataset with 1.076 companies, removing 428 in the process. An important second step is normalizing the data. Z-normalization is applied, resulting in attributes with mean 0 and standard deviation 1.

3.3 *Clustering*

Before clustering the data, a preliminary analysis can be made. The correlation between the different attributes of the dataset can be observed in Fig. 1.

It can be noted that no strong correlations are present in the data, the strongest one being of 0.4. Nevertheless, relations are present, especially between turnover, no. of employees and net profit, which all are positively correlated. A negative correlation occurs between the net profit/employee and turnover/employee. Because there is no high correlation value present in the data, above 0.8, all attributes will be kept in the analysis.

K-means is a partitioning algorithm that requires the number of clusters as an input. The elbow method will be used to determine the optimal number of clusters.

From analyzing Fig. 2, 6 was chosen as the optimal number, because that is the place where the elbow point occurs. Following this step, k-means clustering is performed. In order to plot the clusters, PCA analysis is required to reduce the dimensionality of the data to only two dimensions. The result is presented in Fig. 3.

Multiple observations can be made. Firstly, the presence of an outlier cluster can be noted, cluster 3, which consists of a single data point. Secondly, it can be noted that cluster 2 contains a lot of highly dense points. Cluster 1 is the most dispersed, while clusters 4, 5 and 6 are well defined but in close proximity to majority class, cluster 2, especially cluster 6.

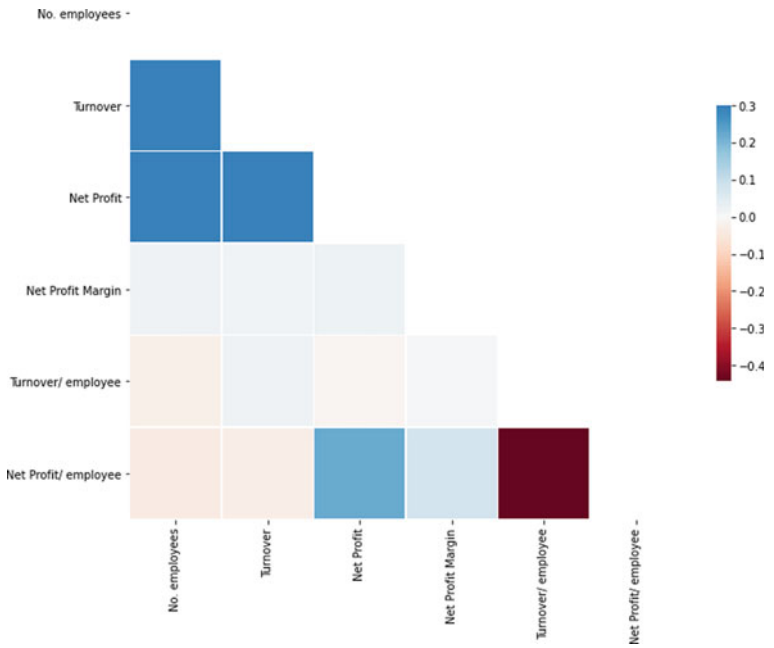

Fig. 1 Dataset correlation

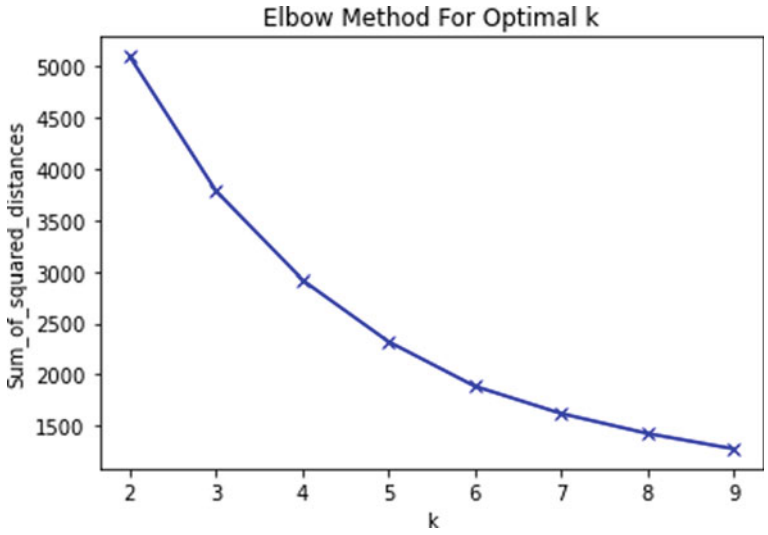

Fig. 2 Elbow method

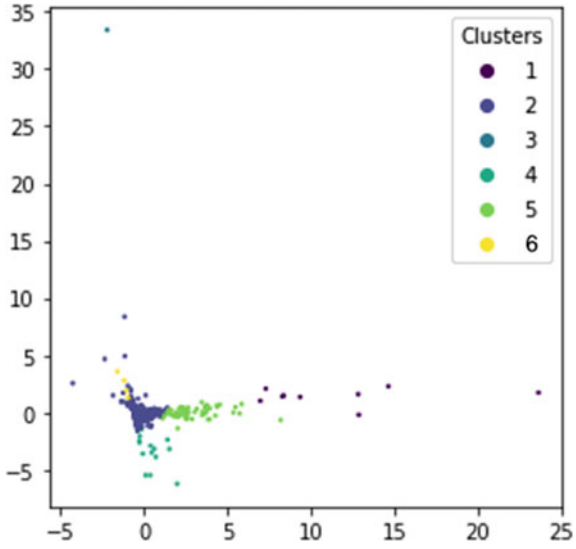

Fig. 3 K-means clustering plot (PCA = 2)

A more informative, 3D, representation of the clustering result can be observed in Fig. 4.

Additional clustering metrics that can be presented are maximum distance between two clusters, 38.67 between clusters 3 and 6, minimum distance, 3.14 between clusters 2 and 5, and average distance between clusters, 20.23.

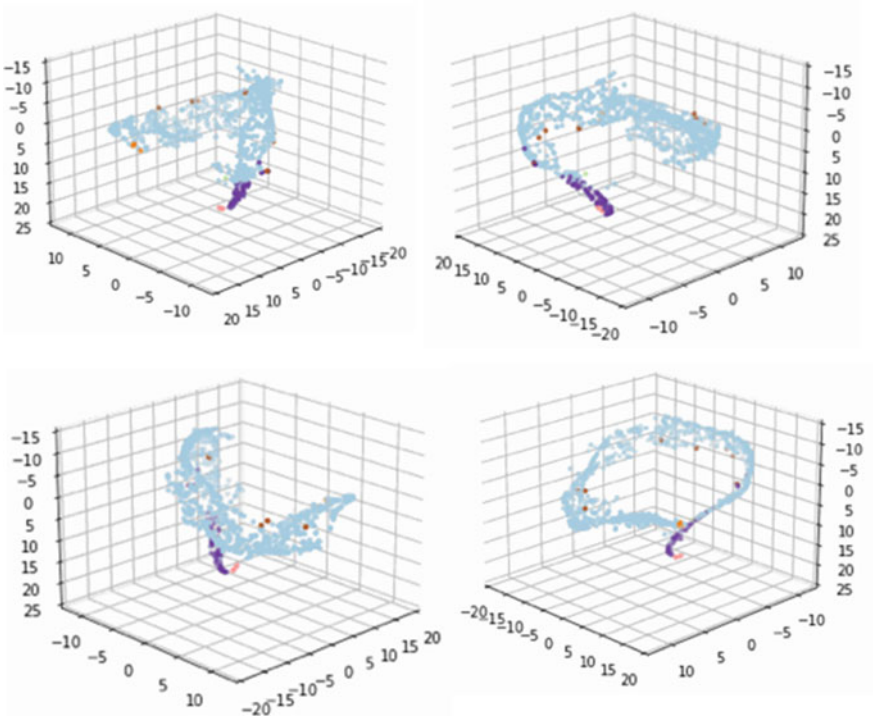

Fig. 4. 3D representation of the k-means clustering

Next, feature importance will be analyzed. Feature importance identifies the importance of the attributes when the clustering algorithm partitions the data. Values on the X axis represent the feature importance score.

It can be observed from Fig. 5 that the most important features when partitioning the data are turnover and no. of employees, followed by net profit.

Analyzing the characteristics of the clusters will offer better insights into the profile of the companies. Figure 6 offers a plethora of information regarding the clusters. First, companies in cluster 2, the majority cluster, have the highest turnover/employee while having the lowest net profit/employee. These are relative small companies with few employees, given the high turnover/employee ratio that struggle to net a profit or even operate at a small loss. This being the majority class could be a sign of concern for the market. Cluster 3 is composed of the biggest players on the market, market leaders, with the greatest number of employees, turnover and net profit. Overachievers can be found in cluster 6, companies that manage to net a large profit with a small number of employees. These companies could be innovators or the "apartment" businesses often found in Romania. Further analysis is required. Companies in cluster 4 struggle to stay in business, highlighted by their high negative net profit margin. Market followers can be found in clusters 5 and 1, companies that manage a balanced business that assures their continued existence on the market.

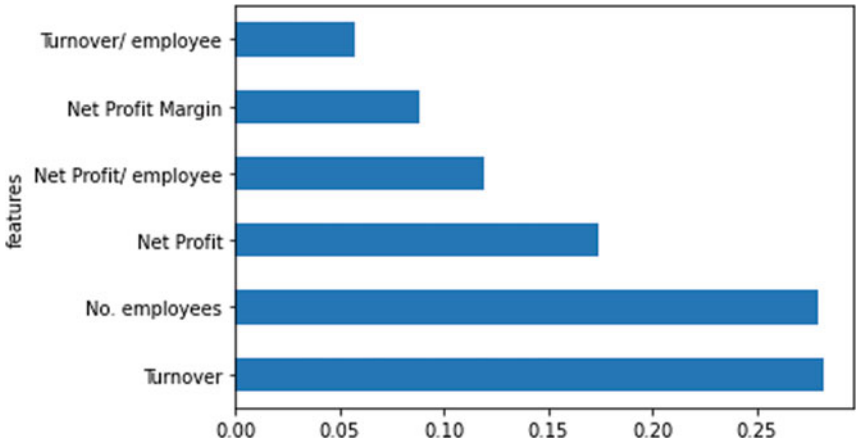

Fig. 5 Feature importance

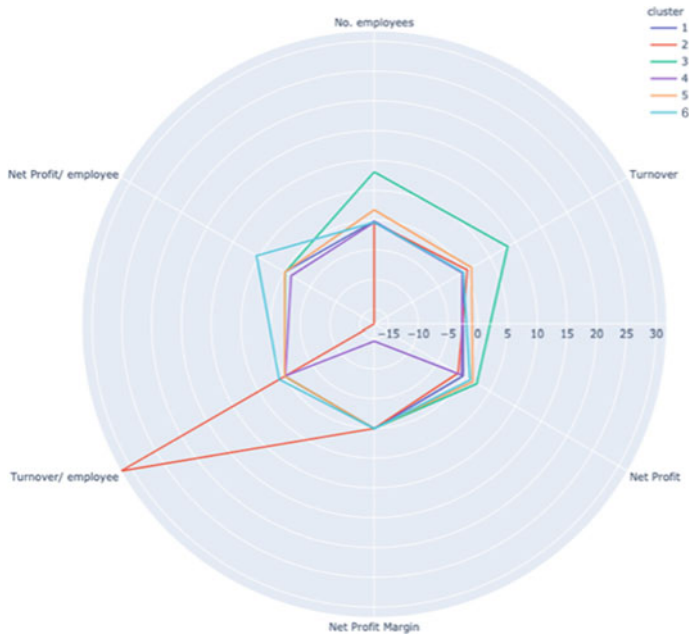

Fig. 6. Clusters line plot

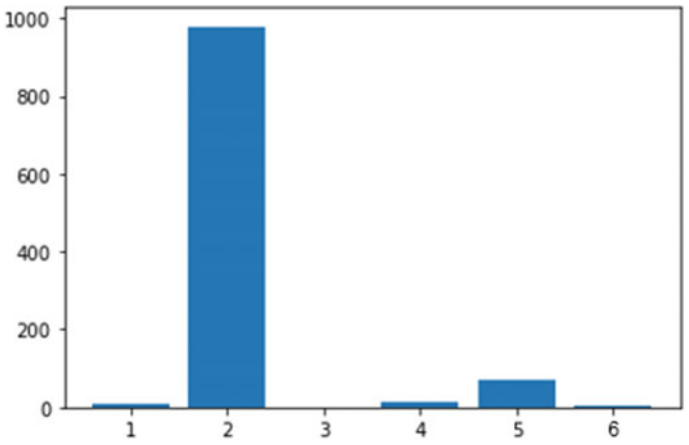

Fig. 7 Data class distribution

3.4 Classification

As previously stated, the labeled data obtained from clustering will be used to train classifiers that will be able to correctly identify new companies and place them in the according class.

Multiple classifiers will be trained and benchmarked. The classification algorithms will be machine learning based, SVM, DecisionTrees and KNN, and deep learning based, neural networks (NN).

For the classification task the dataset is split into a training and a test set, with a 70/30 split. Before training, a representation of the data is required in order to understand the distribution of the classes. Class 2 is the majority class, with 979 observations (98% of total). Class 5 is the second most represented, with 69 observations (41%), followed by class 4 (14, 1.3%), class 1 (9, 0.83%), class 6 (4, 0.37%) and, lastly, class 3 (1, 0.09%). Figure 7 illustrates these results.

It is clear that the dataset is imbalanced, with 90% of the observations in a single class. Therefore, imbalance tackling techniques, such as oversampling, undersampling and weights, will be required. First, classifiers will be trained without taking into the account the imbalance, then the data will be processed and the classifiers will be trained again. The results will then be presented and compared, in order to determine the best solution for this task.

4 Results and Discussion

As previously stated, the classifiers will first be trained on the imbalanced dataset, without applying any correction. This will also establish a baseline of the performance. The macro-scores will be reported for each model (accuracy, precision, recall, f-score). Results achieved at this step are presented in Table 1.

The models used to classify the data are Decision Tree, Support Vector Machine, K Nearest Neighbors and a Neural Network, which can be observed in Fig. 8.

Analyzing performance results from Table 1, it can be noted why accuracy isn't a great measure for imbalanced datasets. The smallest value achieved is 96% (DecisionTree), which is a great value for a balanced dataset. This is not the case here. The minimum accuracy score achievable is 90%, the representation of the majority class. A better metric to evaluate the performance of the models is f-score. F-score is the harmonic mean of precision and recall and shows the performance of the model when it comes to identifying all the members of a certain class and correctly classifying the identified members. Going forward, only f-score will be reported for the performance of the models.

It can be seen that the neural network achieves the best results (f-score of 82.6) followed by the SVM (66), DT (63.1) and KNN (45.1). The neural network is able to better understand the relationships present in the imbalanced data and to correctly classify the members of each class, a task that proves too complex for classic,

Table 1 Model performance for the imbalanced dataset

Model	Accuracy	Precision	Recall	F-score
Decision tree	96	62.6	63.7	63.1
SVM classifier	99.1	65.4	66.6	66
KNN	97.2	48.8	42.4	45.1
NN	**99.1**	**83.2**	**82**	**82.6**

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_53 (Dense)             (None, 500)               3500
_________________________________________________________________
dense_54 (Dense)             (None, 100)               50100
_________________________________________________________________
dense_55 (Dense)             (None, 50)                5050
_________________________________________________________________
dense_56 (Dense)             (None, 6)                 306
=================================================================
Total params: 58,956
Trainable params: 58,956
Non-trainable params: 0
```

Fig. 8 Neural network model summary

simpler models. In order to better understand the performance and results of the NN, the confusion matrix will be presented. It provides information in regards to the classification output and it helps in identifying mislabeled data.

By analyzing Fig. 9, it can be seen that the NN is unable to correctly label the single member of class 3, which was obviously an impossible task for this model, seeing how class 3 wasn't present in training. Regarding classes seen in training, the NN mislabels two members of class 4 in class 2, the only misfire of the model. As such, it can be concluded that the NN achieves great results for the task at hand.

Following the results on the imbalanced dataset, the results on the balanced datasets will be presented. Multiple techniques were used for balancing the dataset: RandomOverSampler (ROS), SMOTE, ADASYN, RandomUnderSampler (RUS), balanced class weights and sample weights. The RandomOverSampler, when used alone, was implemented with sampling_strategy = 'auto' in order to oversample all the under-represented classes. For SMOTE, ADASYN and RandomUnderSampler, a pipeline was established. For SMOTE and ADASYN, ROS was first applied for the minority class, and then the other techniques. For RUS, ROS was applied for the minority classes followed by RUS for the majority class. Weights were used for class and sample, separately. Results are presented in Table 2.

Firstly, it can be noted that the NN achieves the best performance across the board, with an average f-score of 71.8. Regardless, this is an evaluation of the imbalance battling techniques and not so much of the classification models. It can be seen that ADASYN nets the best result, with an average f-score for the models of 69.9. DecisionTrees is the worst performing model, with an average f-score of 37.7, and

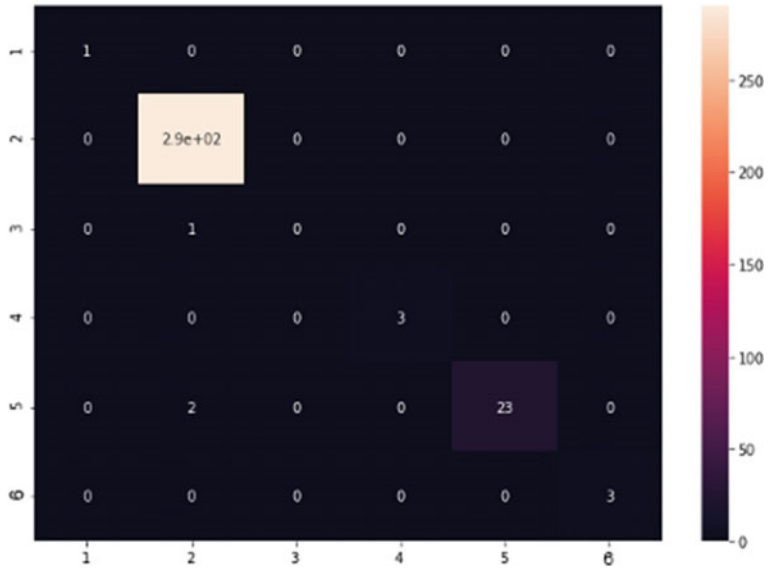

Fig. 9 Neural network classifier confusion matrix

Table 2 Classification models performance for the balanced dataset

Model	ROS	SMOTE	ADASYN	ROS + RUS	Class	Sample	Average
DecisionTrees	33.4	33.4	49.2	0.1	63.1	46.9	37.7
SVM classifier	65.0	65.6	65.3	16.5	66.0	17.6	49.3
KNN	65.3	65.6	82.3	16.5	45.1	45.1	53.3
NN	**83.3**	**82.9**	**82.9**	**33.2**	**82.6**	**65.9**	**71.8**
Average	61.8	61.9	**69.9**	16.6	64.2	43.9	53.0

ROS + RUS is the worst performing balancing implementation, with an average f-score of 16.6. The lowest performance is achieved by DecisionTree when applying ROS + RUS balancing, f-score of 0.1, while the best performance is achieved by the NN when applying ROS, f-score of 83.3.

Comparing the results of the balanced dataset with those of the imbalanced dataset, it can be seen that improvements are achieved by the NN and KNN, with a f-score difference between the best performing instances of 0.7 (ROS) for the NN and 37.2, greatest increase, for the KNN (ADASYN). No improvements have been observed for the other models, with DT and SVM achieving their imbalanced performance when applying Class Weights (f-score of 63.1 and 66, respectively).

The only balancing technique that generated results for this dataset is ADASYN, improving the average original, imbalanced, f-score of 64.2 by 5.7. The other techniques netted better results for KNN and the NN, while decreasing the performance of DT.

Firstly, the segmentation of the Romanian security market can be discussed. The data was partitioned into 6 clusters, each defining a type of company activating on the market: leaders, followers, struggling companies, innovators/shell companies and a plethora of small companies that barely manage to turn a profit.

Secondly, the performance of the classification models can be discussed. The NN managed to achieve the best results, for both the imbalanced and balanced datasets. The NN manages to better understand the latent knowledge in the dataset, enabling it to make better classification predictions. The best balancing technique of the ones implemented was ADASYN, the only one that improved performance over the imbalanced set.

5 Conclusion

In closing, both research questions have been answered. First, the Romanian private security market presents the characteristics of a mature market. There are a small number of large, well established, and profitable companies, a characteristic of a mature market. Additionally, most companies present tend to be smaller, struggling ones, pointing to the difficulty of penetrating the market. However, some question marks remain, especially with the overachieving companies identified, which could

point to a plethora of directions given the past that this market has in Romania. For the second research question, the best solution for this new dataset has been identified, namely the MLP in tandem with ROS, which achieved the best results. This classification task is essential to track the evolution of the market, of both existing companies or new entries.

For future work, more sophisticated classification models can be implemented, in tandem with different sampling algorithms, to net better classification results. In regards with the market analysis, more in-depth data could be collected, to gain a better understating of each company, or a qualitative study could be conducted among the management of these companies, to go beyond the data presently available.

It can be concluded that a greater understating of the Romanian security market was achieved, along with the development of a competent company classifier for this industry. This paper has introduced a new dataset, by clustering publicly available data, and has analyzed the maturity of a neglected market, the Romanian private security industry. Then, multiple classification algorithms were trained and benchmarked to find the best solution that could serve in the future to attribute new companies to the identified groups or to regroup the already identified companies in new fiscal years. This works serves as a starting point for future research that can be conducted on the selected market and that could help find new insights in a under researched but economically impactful segment of the Romanian economy.

References

1. Xu, D., Tian, Y.: A comprehensive survey of clustering algorithms. Ann. Data Sci. 165–193 (2015)
2. García-Villaverde, P.M., Elche, D., Martinez-Perez, A. and Ruiz-Ortega, M.J.: Determinants of radical innovation in clustered firms of the hospitality and tourism industry. Int. J. Hospitality Manage. 45–58 (2017)
3. Perkins, R., Khoo-Lattimore, C., Arcodia, C.: Collaboration in marketing regional tourism destinations: constructing a business cluster formation framework through participatory action research. J. Hospitality Tourism Manag. 347–359 (2021)
4. Guo, K., Wang, J., Shi, G., Cao, X.: Cluster analysis on city real estate market of China: based on a new integrated method for time series clustering. Procedia Comput. Sci. 1299–1305 (2012)
5. Henrique, B.M., Sobreiro, V.A., Kimura, H.: Literature review: machine learning techniques applied to financial market prediction. Expert Syst. Appl. 226–251 (2019)
6. Cortez, R.M., Clarke, A.H., Freytag, P.V.: B2B market segmentation: a systematic review and research agenda. J. Bus. Res. 415–428 (2021)
7. Hutt, M.D., Speh, T.W.: "Business Marketing Management: B2B," Cengage Learning (2016)
8. Clarke, A.H., Freytag, P.V.: An intra- and inter-organisational perspective on industrial segmentation: a segmentation classification framework. Eur. J. Mark. 1023–1038 (2008)
9. Müller, J.M., Pommeranz, B., Weisser, J., Voigt, K.I.: Digital, social media, and mobile marketing in industrial buying: still in need of customer segmentation? Empirical evidence from Poland and Germany. Ind. Mark. Manage. 70–83 (2018)
10. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (2007)
11. Jain, A.: Data clustering: 50 years beyond K-means. Pattern Recognit. Lett. 651–666 (2010)

12. Rodriguez, M.Z., Comin, C.H., Casanova, D., Bruno, O.M., Amancio, D.R., Costa, L.D.F., Rodrigues, F.A.: Clustering algorithms: a comparative approach. Plos One (2019)
13. Thomas, S., Goel, M., Agrawal, D.: A framework for analyzing financial behavior using machine learning classification of personality through handwriting analysis. J. Behav. Exp. Finance (2020)
14. Huang, W., Nakamori, Y., Wang, S.Y.: Forecasting stock market movement direction with support vector machine. Comput. Oper. Res. 2513–2522 (2005)
15. Kara, Y., Boyacioglu, M.A., Baykan, Ö.K.: Predicting direction of stock price index movement using artificial neural networks and support vector machines: the sample of the Istanbul Stock Exchange. Expert Syst. Appl. 5311–5319 (2011)
16. Pai, P.F., Lin, C.S.: A hybrid ARIMA and support vector machines model in stock price forecasting. Omega 497–505 (2005)
17. Breiman, L.: Random forests. Mach. Learn. 5–32 (2001)
18. Adya, M., Collopy, F.: How effective are neural networks at forecasting and prediction? A review and evaluation. J. Forecast. 481–495 (1998)
19. Laboissiere, L.A., Fernandes, R.A., Lage, G.G.: Maximum and minimum stock price forecasting of Brazilian power distribution companies based on artificial neural networks. Appl. Soft Comput. 66–74 (2015)
20. Lahmiri, S.: Improving forecasting accuracy of the S&P500 intra-day price direction using both wavelet low and high frequency coefficients. Fluctuation Noise Lett. 1450008 (2014)
21. Kumar, M., Thenmozhi, M.: Forecasting stock index returns using ARIMA-SVM, ARIMA-ANN, and ARIMA-random forest hybrid models. Int. J. Bank. Account. Finance 284–308 (2014)
22. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority oversampling technique. J. Artif. Intell. Res. 321–357 (2002)
23. He, H., Bai, Y., Garcia, E.A., Li, S.A.: ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: IEEE International Joint Conference on Neural Networks, pp. 1322–1328 (2008)