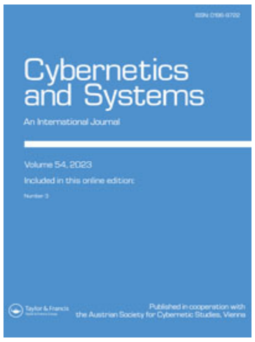
Cybernetics and Systems
An International Journal
Volume 54, 2023
Included in this online edition:
Number 3
Published in cooperation with the Austrian Society for Cybernetic Studies, Vienna

Cybernetics and Systems
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucbs20

Improving the Prediction of Project Success in the Telecom Sector by Means of Advanced Data Balancing

Nuño Basurto, Alfredo Jiménez, Secil Bayraktar & Álvaro Herrero

To cite this article: Nuño Basurto, Alfredo Jiménez, Secil Bayraktar & Álvaro Herrero (2023) Improving the Prediction of Project Success in the Telecom Sector by Means of Advanced Data Balancing, Cybernetics and Systems, 54:3, 306-320, DOI: 10.1080/01969722.2022.2030007

To link to this article: https://doi.org/10.1080/01969722.2022.2030007

Published online: 01 Feb 2022.

CYBERNETICS AND SYSTEMS: AN INTERNATIONAL JOURNAL: AN INTERNATIONAL JOURNAL
2023, VOL. 54, NO. 3, 306–320
https://doi.org/10.1080/01969722.2022.2030007

Taylor & Francis
Taylor & Francis Group

Improving the Prediction of Project Success in the Telecom Sector by Means of Advanced Data Balancing

Nuño Basurto[a], Alfredo Jiménez[b], Secil Bayraktar[c], and Álvaro Herrero[a]

[a]Departamento de Ingeniería Informática, Grupo de Inteligencia Computacional Aplicada (GICAP), Universidad de Burgos, Burgos, Spain; [b]Department of Management, KEDGE Business School, Bordeaux, France; [c]Department of Human Resources Management and Business Law, TBS Business School, Toulouse, France

ABSTRACT

As governments access capital, technology, and managerial expertise from private investors, there is an increasing trend of privatizations in infrastructure projects worldwide. Given the large size of these investments, notably in the sector of telecommunications, investors typically create a consortium with other interested firms, an investment vehicle known as private participation project (PPP). Given the critical repercussions on the rest of the economy, and its large financial losses in case of failure, the prediction of the success of PPPs in telecommunications is of utmost importance. The success of PPPs can be predicted by Machine Learning and it is probed in this article. Hence, widely acknowledged classifiers (*k*-nearest neighbors [k-NNs], support vector machines [SVMs], and random forest [RF]) are applied to PPPs publicly available data from the World Bank. The results on this highly imbalanced dataset are greatly improved by the application of data balancing techniques. It includes some standard ones (random oversampling, random undersampling, and synthetic minority oversampling technique [SMOTE]), together with some other advanced ones (density-based SMOTE and borderline SMOTE). The satisfactory results validate the proposed application of classifiers on the dataset improved by data-balancing techniques.

KEYWORDS

Classification; data balancing; internationalization; private-participation projects; telecommunications

Introduction and Previous Work

In recent decades, governments worldwide have increasingly relied on privatizations in order to attract foreign investors who can transfer to the home country the skills and resources that large infrastructure projects need. Investors, due to the large size of these projects, typically create a consortium of “sponsors” with other firms also interested in the project, reducing the risk and achieving complementarities (Ramamurti and Doh 2004; García-Canal and Guillén 2008). Among the sectors where private

CONTACT Nuño Basurto nbasurto@ubu.es Departamento de Ingeniería Informática, Grupo de Inteligencia Computacional Aplicada (GICAP), Universidad de Burgos, Av. Cantabria s/n, 09006 Burgos, Spain.

participation projects (PPPs) are more common, given their large size and complexity, the telecommunication sector is one of the flagships.

Telecommunications play a central role in any economy, and due to its strategic value and repercussions in the rest of the sectors, governments usually try to regulate it and keep it under strong scrutiny. In fact, the entrance of foreign investors was not always allowed, and even nowadays restrictions are sometimes applied. Yet, the trend in both developed and developing countries is now to be more open to foreign capital, in order to increase competition and access cutting-edge technology and expertise (Ramamurti and Doh 2004; García-Canal and Guillén 2008). Consequently, the composition of the ownership of privatized projects in this sector has greatly diversified in terms of nationalities.

Given their popularity and ubiquity, a growing stream of scientific literature has developed in order to understand the antecedents of project performance and the key variables for success (Dorobantu, Lindner, and Müllner 2020). Thus, previous studies have analyzed the impact of diverse factors, both at the project (Djankov 1999) and at the country levels (Jiménez et al. 2017; Jiang et al. 2015; Ramamurti 2003). Although many projects succeed because investors carefully choose the projects in which they bid and, similarly, governments select the most suitable and capable consortium of sponsors, the economic repercussions of failed projects are of such magnitude that the prediction of success rates of telecommunication projects is of utmost importance. We precisely aim to contribute to this field by analyzing a sample of 9176 PPPs, located in 32 host markets, covering a wide time span (2004–2013), on which we test and compare various techniques to predict the success of projects.

Several different intelligent techniques have previously contributed to the enterprise management field (Herrero and Jiménez 2019; Jiménez and Herrero 2019; Simić et al. 2019; Herrero, Jiménez, and Bayraktar 2019). However, scant attention has been devoted so far to the previously described problem from the supervised-learning research community. One of the very few articles on this topic is (Huang et al. 2004), where corporate credit rating analysis is conducted based on support vector machine (SVM) and artificial neural networks (ANN). These classifiers are applied to data from the United States and Taiwan markets trying not only to forecast but also to get a model with better explanatory power. More recently (Chou, Cheng, and Wu 2013), proposed the combination of SVM and fuzzy logic to address a real case study in construction management. This hybrid system tried to predict project dispute resolution outcomes (i.e., mediation, arbitration, litigation, negotiation, and administrative appeals) when the dispute category and phase in which a dispute occurs are known during project execution.

In (Malhotra, Sharma, and Nair 1999) *k*-nearest neighbor (*k*-NN) is compared to ANN, discriminant analysis, quadratic discriminant analysis, and multinomial logistic regression analysis to provide input to managers who make business decisions. These models were applied to retail department store data, showing that they are most useful when uncertainty is high and a priori classification cannot be made with a high degree of reliability. Additionally (Yu 2011), proposed the application of k-NN to multi-criteria inventory classification in order to manage inventory more efficiently. *k*-NNs are compared to SVM, ANN, and multiple discriminant analysis when applied to four benchmark datasets. SVM was identified as the most accurate among all of them due to its high generalization capability, as well as its use of kernel functions to increase the learning efficiency.

The initial steps in this research line include (Herrero and Jiménez 2020), who presented the application of different classifiers (SVM, k-NN, and random forest [RF]) for predicting the final success of several PPPs that involved infrastructures. Based on these initial results, standard data balancing techniques were validated (Basurto et al. 2021) for improving the SVM and RF classifiers performance when applied to such imbalanced PPP datasets. Advancing from this previous work, in this article, advanced data balancing techniques are studied and its effect is comprehensively validated when applied in conjunction with several different classifiers.

The rest of this article is organized as follows: the applied techniques are described in Soft-Computing Techniques section. The addressed problem (including the analyzed dataset) is described in International PPPs section while the obtained results are presented and discussed in Experiments and Results section. Finally, the conclusions of this study and future work are stated in Conclusions and Future Work section.

Soft-Computing Techniques

Different balancing techniques, as detailed in Data-Balancing section, have been applied to improve the success prediction. The base results and the effect of data balancing are validated by considering the classifiers described in Classifiers section.

Data-Balancing

Data balancing is one of the existing and widely applied pre-processing techniques whose main objective is to obtain a data set with a balancing between existing classes. These techniques are used when the classes of a dataset are unbalanced, i.e., in case of a one-class problem, one of the two classes has more instances than the other. This usually causes a biased learning unsupervised method and hence, classification results are negatively affected. For data balancing, there are three main approaches: oversampling, undersampling, and hybrid methods.

In the oversampling approach, the main objective is to generate new instances of the minority class until the number of instances belonging to the majority and minority classes becomes similar. A well-known method is called random over sampling (ROS) (Hoens and Chawla 2013) that generates completely random new instances by duplicating existing instances from the minority class. On the other hand, the synthetic minority oversampling technique (SMOTE) (Chawla et al. 2002) has been proposed, which generates new artificial instances. These are generated through the linear interpolation of different instances (randomly chosen) belonging to the minority class and their neighbors belonging to the same class. Based on the SMOTE approach, new variants have been generated, such as density-based SMOTE (DBSMOTE) (Bunkhumpornpat, Sinapiromsaran, and Lursinsap 2012). This new approach combines the benefits of SMOTE with the DBSCAN clustering method. In this way, it generates new instances on the shortest path from the instance to a pseudocentroid generated in the cluster. Another variant is borderline-SMOTE (BLSMOTE) (Han, Wang, and Mao 2005), in this case, an oversampling is performed only on the minority class instances found on the border.

The undersampling methods have a completely opposite approach to the above, their approach is to eliminate the instances of the majority class in order to achieve a balance between classes. One of the most popular techniques is random under sampling (RUS) (Hoens and Chawla 2013), which removes instances completely randomly from the majority class.

Finally, there are some techniques whose approach is the union of both approaches (Shamsudin et al. 2020) by performing first one and then the other, thus reducing the impact of a single method. In this article, the combination of SMOTE + RUS and ROS + RUS has been used.

Classifiers

In order to predict the success of the projects, supervised machine learning models are applied. Based on previous recent results (Herrero and Jiménez 2020; Herrero, Bayraktar, and Jiménez 2020; Moscoso-López et al. 2021; Gonzalez-Cava et al. 2021; Tang et al. 2021; Jia and Sun 2021), the classifiers SVMs (Boser, Guyon, and Vapnik 1992; Cortes and Vapnik 1995), RFs (Breiman 2001) and k-NNs (Cunningham and Delany 2022) have been selected for this research. In order to train such models, the success of the projects in the dataset is used as the class label (success or failure).

One of the best-known classifiers is the SVM's, these have been used on a great number of occasions, being able to solve several real-life problems (Cortes and Vapnik 1995). Its use has been very focused on one-class problems (Shin, Eom, and Kim 2005). The objective of this classifier is to identify a hyperplane that maximizes the separation of instances of different classes, trying to universalize the archetype. One of the characteristics of SVM is the high sensitivity of its parameters such as gamma and cost. In this research, several values for each of these parameters have been used (Results Obtained by RF section).

SVMs can be seen as classifiers where the loss function is the Hinge function, defined as:

$$L\left[y, f(x)\right] = \max\left[0, 1 - yf(x)\right] \tag{1}$$

Being x an observation from input features, y the class x belongs to, and $f(x)$ the output of the classifier. Additionally, there is the gamma parameter that states the influence of a single training example, i.e., a low value means a far influence while a high value means closes.

The classification trees (Safavian and Landgrebe 1991) are inductive learning methods, in their structure, there are two types of nodes, leaf nodes, which are those designed for making a final decision or prediction, and inner nodes that perform the association of different branches given a question at the expense of the value of the feature of the training data set, each of them has two child nodes.

Labels are assigned to the archs connecting a node to one of its child nodes (their content is related to the responses to the node question) and leaf nodes (their content is one of the classes in the training dataset).

The RF algorithm can be interpreted as an aggregation of several classification trees. Each example is sampled independently but with the same distribution in all trees. One of the characteristics that make RF stand out is that thanks to the assignment of the new data to the class that had already been predicted by the different trees, there is less variance.

Finally, the approach offered by k-NN is completely different from that seen with the other two classifiers. An instance belongs to the same class to which its k-NNs belongs. In this research, we have worked with different values of k; 5, 10, and 15. And we have also worked with the well-known Euclidean distance.

The metric used to fairly differentiate between classes has been the area under the curve (AUC). It has been selected due to its ability to define how good a model is, according to the distinction it makes between classes. This metric is widely applied in unbalanced data sets.

International Private Participation Projects

PPPs can be conceptualized, according to the World Bank, as privatized infrastructure projects in which local or foreign firms play a significant role as private investors with equity in the ownership of the project. Given their size and complexity, firms participating in PPPs can rarely do it alone, and instead, they usually create a consortium of sponsors who much engage in close collaboration in order to ensure the success of the project throughout all its phases, including the bidding process, the capital raise, the construction, and the execution.

As PPPs are legally independent entities from the sponsors, the success of the project is vital in order to achieve the goals of the project as well as to repay the project funding requested. Otherwise, investors will have to abandon the project when this one is unable to generate enough revenue or the government requests its termination because the project is not generating the expected outcome. The overall process of these privatizations has many repercussions for multiple stakeholders (users, suppliers, banks, society, …), attracting a lot of media attention, and it is highly politicized (Jiménez et al. 2019).

Following previous research on PPPs (Jiang et al. 2015), we draw our sample from the World Bank's private participation in infrastructure (PPI) dataset. We selected from the data source all the privatizations in the telecommunication sector, which is precisely one of the most numerous in the total dataset. Specifically, we collected information on 9,176 projects, located in 32 host countries and covering a wide time span (2004–2013).

Also building on previous studies employing the PPI dataset, we conceptualize the success of projects based on the completion of the bidding process, the successful fulfillment of the binding agreements, as well as the access to the needed capital. We thus construct a dichotomous variable based on the project status provided in the dataset, following (Jiménez et al. 2017; Jiang et al. 2015; Jiménez et al. 2019). It is important to note that the particular characteristics of these projects, in which the government privatize partially or fully a large infrastructure, and sponsors find the capital to carry out the project *via* non-recourse loans which are secured only by the future cash flows generated by the infrastructure (Dorobantu, Lindner, and Müllner 2020), make traditional firm performance measures unsuitable for this context. In contrast, their performance is usually measured in terms of whether the consortium of sponsors has successfully fulfilled the legally binding agreement to invest funds, developed the facilities, and provided the service (Jiménez et al. 2019). We coded projects as 1 = successful (98.55% of the data samples) when the project status is shown "operational," "merged," or "concluded." In contrast, we coded projects as 0 = failed (1.45% of the data samples), when the project

Table 1. Ranges and statistical features of continuous variables.

Continuous	Min	Mean	Max	SD
AGE_INVESTMENT_2014	1.000	3.794	10.000	2.405
DELAY	0.000	14.190	23.000	5.396
LOG_TOTAL_INVESTMENT	−0.693	3.412	8.321	2.818
GDP_HOST	9.140	11.440	12.420	0.432
GDP_GROWTH_HOST	−7.820	4.461	15.210	2.987
UNEMPLOYMENT_HOST	0.000	4.007	27.100	6.026
HOST_POLCONV	0.000	0.393	0.787	0.201
HOST_CORRUPTION_WGI	−1.423	−0.791	1.562	0.383

status is shown “cancelled” (when the private sponsor(s) have abandoned the project) or “distressed” (when the government or the sponsor has either requested termination or is in international arbitration).

As explanatory factors, we also included in the model several factors that were previously employed in the literature (Jiménez et al. 2017). At the country level, we include in the estimations the logarithm of the GDP in the host country, the % of GDP growth, the logarithm of the level of unemployment, the score of the POLCONV index (representing the degree of political stability), and the score of the control of corruption variable in the World Bank Worldwide Governance Indicators dataset. This latter control variable is reversed, as it is commonly done in empirical studies, in order to facilitate its interpretation, i.e., higher levels of the variable representing more corruption and vice-versa.

At the project level, we included the logarithm of the total amount of investment in the privatization, the age of the project, the time lag since the time of project commitment and the one of project closure, and whether the project is Greenfield (started from scratch) or brownfield (pre-existing). Furthermore, we also control for whether the main investor is a foreign company, whether one investor or more is a local company from the host country, whether the government kept some ownership in the project, and whether the project is publicly traded or not. Thus, we included a total of 13 features in each of the datasets analyzed for all the project instances.

Our choice of features is therefore consistent with previous literature in the field which has identified that the success of projects is contingent on the combined effect of country-level factors, both macroeconomic and institutional, and project-level ones related to the characteristics of the privatization and the composition of the consortium of sponsors (Djankov 1999; Jiang et al. 2015; Ramamurti 2003; Doh, Teegen, and Mudambi 2004; Inoue, Lazzarini, and Musacchio 2013).

For a better understanding of the problem addressed in this research, the statistical values of the variables are shown in Table 1.

On the other hand, Table 2 shows the different statistical values obtained in the discrete variables.

Table 2. Ranges and statistical features of discrete variables.

Discrete	Min	Mean	Max	SD
EAST_ASIA_PACIFIC	0.000	0.166	1.000	0.372
EUROPE_CENTRAL_ASIA	0.000	0.101	1.000	0.302
LATIN_AMERICA_CARIBBEAN	0.000	0.084	1.000	0.277
MIDDLE_EAST_NORTH_AFRICA	0.000	0.017	1.000	0.130
SOUTH_ASIA	0.000	0.398	1.000	0.489
SUBSAHARAN_AFRICA	0.000	0.235	1.000	0.424
FOREIGN_SPONSOR_MAIN_INVESTOR	0.000	0.556	1.000	0.497
LOCAL_SPONSOR_PRESENCE	0.000	0.478	1.000	0.500
GREENFIELD	0.000	0.786	1.000	0.410
PUBLICLY_TRADED	0.000	0.398	1.000	0.489
HOST_GOV_OWNERSHIP	0.000	0.343	1.000	0.475
Class	0.000	0.986	1.000	0.120

Experiments and Results

The dataset previously presented in International Private Participation Projects section has been analyzed with the techniques introduced in Soft-Computing Techniques section. The obtained results are presented and discussed in this section, grouped by the applied classifier. The results obtained without applying any data-balancing technique are referred as "None," that are included for a comprehensive comparison comprising all scenarios. Due to the high number of experiments, only the AUC metric is shown in this section, being one of the most appropriate ones for unbalanced datasets. To aid replicability, a seed of 59 has been established and 75% has been used for training and 25% for testing. For the validation of the experimental results, a cross-validation with 5-fold has been performed. For a better understanding of the data set, a sample of the size of the data set has been provided in Table 3. Prior to processing, the target or class column is omitted.

As a consequence of the use of the balancing methods, the dimensions of the data set are as shown in Table 4. This shows the big difference between performing classification tasks with undersampling and oversampling techniques.

Results Obtained by SVM

Results obtained when applying SVM after balancing the dataset with the techniques described in Data-Balancing section are shown in Figure 1.

For the purpose of direct comparison, values obtained with the same gamma value but with a different cost value have been compared. The obtained results show that the use of balancing techniques improves the results obtained by none in a fairly general way, with the ROS and SMOTE oversampling techniques, together with their hybridization with RUS, outperforming all the other ones. It can also be said that there is a great difference between the results achieved by the different gamma values, with the

Table 3. Dimensions of the data set before processing.

	Data	Testing	Training
Rows from Class 0	133	29	104
Rows from Class 1	9,043	2,265	6,778
Total rows	9,176	2,294	6,882
Total columns	20	19	19

Table 4. Dimensions of the data set after the application of different balancing methods in the training set.

	None	ROS	SMOTE	DBSMOTE	BLSMOTE	RUS	ROS + RUS	SMOTE + RUS
Rows from Class 0	104	6,870	6,760	6,760	6,770	104	3,461	3,440
Rows from Class 1	6,778	6,778	6,778	6,778	6,778	104	3,421	3,441
Total rows	6,882	13,648	13,538	13,538	13,548	208	6,882	6,881

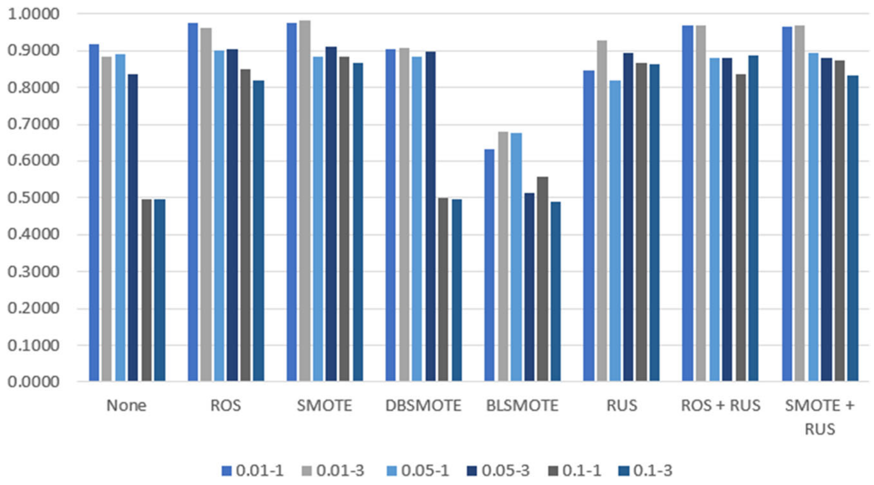

Figure 1. AUC results by SVM per data-balancing algorithm and different values of gamma (0.01, 0.05, and 0.1) and cost (1 and 3) parameters.

0.01 value being generally better than the rest. In terms of cost, no major differences are being observed, with smaller variations than in the case of gamma. The low AUC scores obtained with the SMOTE variations (DBSMOTE and BLSMOTE) are worth mentioning, being these values similar to those of NONE when the gamma value is 0.1.

Finally, it must be highlighted that the best value for the SVM classifier has been obtained when applying SMOTE with a gamma value of 0.01 and cost 3, getting an AUC value of 0.9815.

Results Obtained by RF

Results obtained when applying the RF classifier, after balancing the dataset with the techniques described in Data-Balancing section are shown in Figure 2.

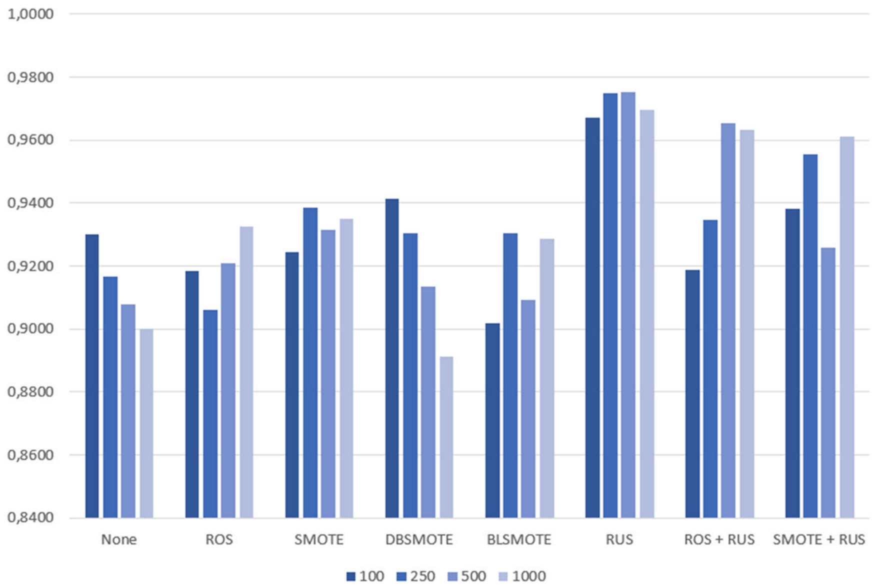

Figure 2. AUC results by RF per data-balancing algorithm and different numbers of trees (100, 200, 500, and 1000).

In general terms, it can be said that the prediction results are improved by the use of balancing techniques, as in the case of the SVM classifier (see previous subsection). For the RF classifier, it can be seen that RUS achieves the highest AUC values for all the numbers of trees included in the experiments (100, 200, 500, and 1,000). This differs from previous results as in the case of SVM, RUS performance was modest. SMOTE combined with RUS in the case of the smallest numbers of trees (100 and 250) and ROS combined with RUS in the case of the highest numbers of trees (500 and 1,000) are those that have obtained second-best results. It can be concluded that undersampling techniques are the best approach, outperforming the other ones when predicting the project success by RF. The best AUC value achieved by this classifier is 0.9754, when applying RUS and using 500 trees.

The oversampling techniques in this case turn out to obtain the worst classification performance. It is worth mentioning the case of DBSMOTE (with $n = 1,000$), that has obtained values even worse than none (raw data without any balancing).

Results Obtained by k-NN

As for the previous classifiers, results by *k*-NN are presented in this section. Three values for the *k* parameter have been used, in order to find a greater variability in the results and covering a wider range: 5, 10, and 15. Obtained AUC values are shown in Figure 3.

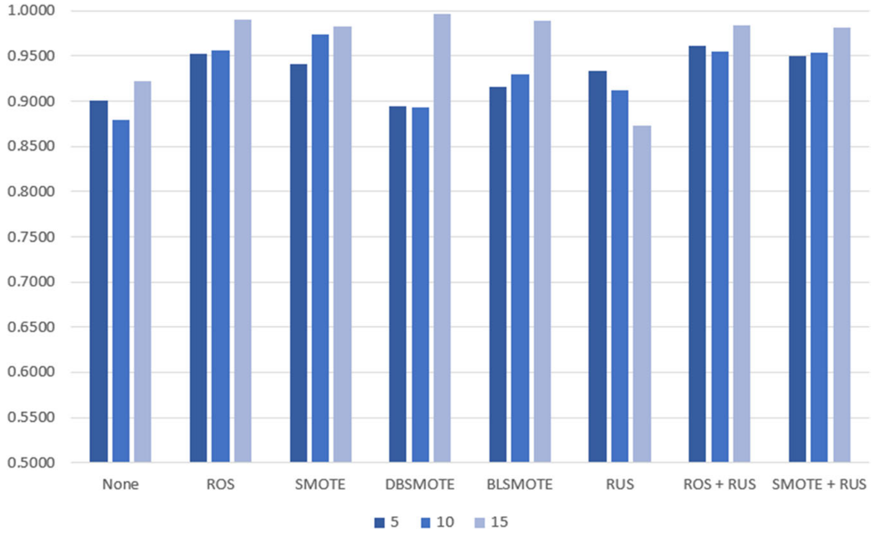

Figure 3. AUC results by *k*-NN per data-balancing algorithm and different *k* values 5, 10, and 15.

The obtained results show that in most cases the value of $k = 15$ outperforms the other ones, except in the case of RUS. In general, it can be observed that the oversampling algorithms obtain better results than the other balancing techniques. The SMOTE, DBSMOTE, and BLSMOTE variants have obtained very high-performance values for $k = 15$. On the other hand, the performance values of these techniques for $k = 5$ and $k = 10$ are quite low compared to the other ones, being similar to those of none. Additionally, it should be noted that this classifier is the one obtaining the highest AUC value of all the run experiments: 0.9969. This has been obtained by balancing the data with DBSMOTE and using a number of neighbors (k) equal to 15. The second best AUC value is 0.9904, obtained by balancing with the ROS technique and the same number of neighbors.

Comparison of Classifier Results

Finally, the results obtained by the different classifiers and balancing methods are visually compared in Figure 4.

This figure shows a similarity in the vast majority of classification methods except the poor results by SVM after applying BLSMOTE. On the other hand, it can be clearly seen that *k*-NN outperforms the other classifiers in a total of six balancing techniques, out of eight (seven valancing combinations and "None"). Additionally, the best classification ratio is obtained by this classifier after applying DBSMOTE. It is worth noting that *k*-NN is not the best classifier when combining it with RUS and none, showing its great ability to deal with oversampled data.

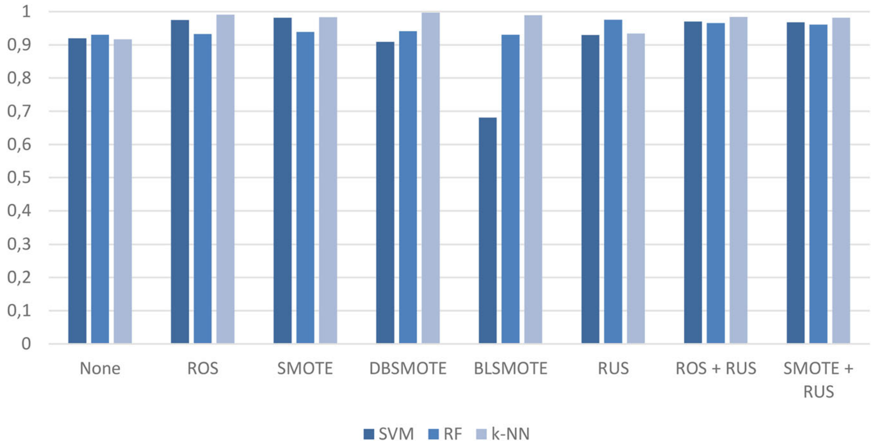

Figure 4. Best AUC values by the three classifiers per data-balancing algorithm.

Conclusions and Future Work

An analysis of the results obtained shows that the balancing techniques help to improve the prediction results. When comprehensively looking at the results, it can be said that the performance of the balancing techniques greatly varies from one classifier to another one. In general terms, RUS is the one that best performs for RF and, on the other hand, the oversampling techniques are the ones that have obtained better results for SVM and *k*-NN classifiers. It can also be noted the relevance of the hybrid balancing techniques, which reduce the impact of using a single technique and thus lead to more stable values. As mentioned in the previous section, the best result has been obtained when combining DBSMOTE and the *k*-NN classifier. This research advances previous results (Herrero and Jiménez 2020) by proposing new data-balancing and classification methods.

A more accurate prediction of the success of privatization projects in telecommunications bears important repercussions both for investors and governments. First, the higher predictability of the outcome of the project, the more likely projects will be able to obtain access to the funding required for the construction of the project. Furthermore, it will also be easier to attract sponsors to the consortium, given the lower levels of risk, increasing synergies, and complementarities. Managers are therefore particularly interested in being able to employ the most convenient estimation techniques in order to reduce uncertainty and obtain tangible advantages when designing and bidding for the project. Second, better predictability also provides benefits for governments, as they will be able to increase the attractiveness of privatizations and, consequently, the number and quality of proposals. As governments receive not only capital, but also technology and knowledge from investors in privatization projects, increasing the number of proposals can allow them to select, among a larger pool, the one that is most beneficial given the government and country needs. Overall, the improvement in the methods to predict the success of PPPs improves efficiency and brings advantages to multiple stakeholders. We therefore encourage scholars to pursue this fascinating area of research.

Notwithstanding its contributions, this article is subject to some limitations. First, the dataset does not provide information regarding whether the goal of the projects is related to IT or to infrastructure construction. Second, our geographical reach is restricted to the countries covered in the PPI database. Third, the PPI dataset contains data of projects and sponsors, but not on the lenders.

This research could be extended in the future by considering some other sectors and comparing the obtained results with those from the telecommunications one. Additionally, some other classifiers and combinations of them (ensembles) could be also applied. Finally, features selection techniques could be applied in order to identify the most relevant features involving the success of PPP.

Funding

The work was conducted during the research stay of Nuño Basurto at KEDGE Business School in Bordeaux (France).

References

Basurto, N., A. Jiménez, S. Bayraktar, and Á. Herrero. 2021. Data balancing to improve prediction of project success in the telecom sector. *15th international conference on soft computing models in industrial and environmental applications.* New York, NY: Springer International Publishing.

Boser, B. E., I. M y Guyon, and V. N. Vapnik. 1992. A training algorithm for optimal margin classifiers. 5th Annual Workshop on Computational Learning Theory, Pittsburgh, PA, July 27–29, 144–52. doi:10.1145/130385.130401.

Breiman, L. 2001. Random forests. *Machine Learning* 45 (1):5–32. doi:10.1023/A:1010933404324.

Bunkhumpornpat, C., K. y Sinapiromsaran, and C. Lursinsap. 2012. DBSMOTE: Density-based synthetic minority over-sampling technique. *Applied Intelligence* 36 (3):664–84. doi:10.1007/s10489-011-0287-y.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16:321–57. doi:10.1613/jair.953.

Chou, J. S., M. Y. Cheng, and Y. W. Wu. 2013. Improving classification accuracy of project dispute resolution using hybrid artificial intelligence and support vector machine models. *Expert Systems with Applications* 40 (6):2263–74. doi:10.1016/j.eswa.2012.10.036.

Cortes, C. y., and V. Vapnik. 1995. Support-vector networks. *Machine Learning* 20 (3): 273–97. doi:10.1023/a:1022627411411.

Cunningham, P. Y., and S. J. Delany. 2022. k-Nearest neighbour classifiers. *ACM Computing Surveys* 54 (6):128. doi:10.1145/3459665.

Djankov, S. 1999. Ownership structure and enterprise restructuring in six newly independent states. *Comparative Economic Studies* 41 (1):75–95. doi:10.1057/ces.1999.3.

Doh, J. P., H. y Teegen, and R. Mudambi. 2004. Balancing private and state ownership in emerging markets' telecommunications infrastructure: Country, industry, and firm influences. *Journal of International Business Studies* 35 (3):233–50. doi:10.1057/palgrave.jibs.8400082.

Dorobantu, S., T. Y. Lindner, and J. Müllner. 2020. Political risk and alliance diversity: A two-stage model of partner selection in multipartner alliances. *Academy of Management Journal* 63 (6):1775–806. doi:10.5465/amj.2017.0265.

García-Canal, E. Y., and M. F. Guillén. 2008. Risk and the strategy of foreign location choice in regulated industries. *Strategic Management Journal* 29 (10):1097–115. doi:10.1002/smj.692.

Gonzalez-Cava, J. M., R. Arnay, J. A. Mendez-Perez, A. León, M. Martín, J. A. Reboso, E. Jove-Perez, and J. L. Calvo-Rolle. 2021. Machine learning techniques for computer-based decision systems in the operating theatre: Application to analgesia delivery. *Logic Journal of the IGPL* 29 (2):236–50. doi:10.1093/jigpal/jzaa049.

Han, H., W. Y. Wang, and B. H. Mao. 2005. *Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning* Berlin, Heidelberg, Germany: Springer Berlin Heidelberg.

Herrero, Á. Y., and A. Jiménez. 2019. Improving the management of industrial and environmental enterprises by means of soft computing. *Cybernetics and Systems* 50 (1):1–2. doi:10.1080/01969722.2019.1560961.

Herrero, Á. Y., and A. Jiménez. 2020. One-class classification to predict the success of private-participation infrastructure projects in Europe. *14th international conference on soft computing models in industrial and environmental applications*. New York, NY: Springer International Publishing.

Herrero, Á., A. Y. Jiménez, and S. Bayraktar. 2019. Hybrid unsupervised exploratory plots: A case study of analysing foreign direct investment. *Complexity* 2019:1–14. doi:10.1155/2019/6271017.

Herrero, Á., S. Y. Bayraktar, and A. Jiménez. 2020. Machine learning to forecast the success of infrastructure projects worldwide. *Cybernetics and Systems* 51 (7):714–31. doi:10.1080/01969722.2020.1798645.

Hoens, T. R. Y., and N. V. Chawla. 2013. Imbalanced datasets: From sampling to classifiers. In *Imbalanced learning: Foundations, algorithms, and applications*, 43–59. Hoboken, NJ: John Wiley & Sons. doi:10.1002/9781118646106.ch3.

Huang, Z., H. Chen, C. J. Hsu, W. H. Chen, and S. Wu. 2004. Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems* 37 (4):543–58. doi:10.1016/S0167-9236(03)00086-1.

Inoue, C. F. K. V., S. G. Y. Lazzarini, and A. Musacchio. 2013. Leviathan as a minority shareholder: Firm-level implications of state equity purchases. *Academy of Management Journal* 56 (6):1775–801. doi:10.5465/amj.2012.0406.

Jia, H. y., and K. Sun. 2021. Improved barnacles mating optimizer algorithm for feature selection and support vector machine optimization. *Pattern Analysis and Applications* 24 (3):1249–74. doi:10.1007/s10044-021-00985-x.

Jiang, Y., M. W. Peng, X. Yang, and C. C. Mutlu. 2015. Privatization, governance, and survival: MNE investments in private participation projects in emerging economies. *Journal of World Business* 50 (2):294–301. doi:10.1016/j.jwb.2014.10.006.

Jiménez, A. y., and Á. Herrero. 2019. Selecting features that drive internationalization of Spanish firms. *Cybernetics and Systems* 50 (1):25–39. doi:10.1080/01969722.2018.1558012.

Jiménez, A., G. F. Jiang, B. Petersen, and J. Gammelgaard. 2019. Within-country religious diversity and the performance of private participation infrastructure projects. *Journal of Business Research* 95:13–25. doi:10.1016/j.jbusres.2018.08.027.

Jiménez, A., M. Russo, J. M. Kraak, and G. F. Jiang. 2017. Corruption and private participation projects in central and Eastern Europe. *Management International Review* 57 (5): 775–92. doi:10.1007/s11575-017-0312-4.

Malhotra, M. K., S. Y. Sharma, and S. S. Nair. 1999. Decision making using multiple models. *European Journal of Operational Research* 114 (1):1–14. doi:10.1016/S0377-2217(98)00037-X.

Moscoso-López, J. A., D. Urda, J. J. Ruiz-Aguilar, J. González-Enrique, and I. J. Turias. 2021. A machine learning-based forecasting system of perishable cargo flow in maritime transport. *Neurocomputing* 452:487–97. doi:10.1016/j.neucom.2019.10.121.

Ramamurti, R. 2003. Can governments make credible promises? Insights from infrastructure projects in emerging economies. *Journal of International Management* 9 (3):253–69. doi:10.1016/S1075-4253(03)00036-X.

Ramamurti, R. Y., and J. P. Doh. 2004. Rethinking foreign infrastructure investment in developing countries. *Journal of World Business* 39 (2):151–67. doi:10.1016/j.jwb.2003.08.010.

Safavian, S. R. Y., and D. Landgrebe. 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics* 21 (3):660–74. doi:10.1109/21.97458.

Shamsudin, H., U. K. Yusof, A. Jayalakshmi, M. N. A. Khalid. 2020. Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent transaction dataset. 2020 IEEE 16th International Conference on Control & Automation (ICCA), Singapore, October 9–11. doi:10.1109/ICCA51439.2020.9264517.

Shin, H. J., D. H. Y. Eom, and S. S. Kim. 2005. One-class support vector machines—an application in machine fault detection and classification. *Computers & Industrial Engineering* 48 (2):395–408. doi:10.1016/j.cie.2005.01.009.

Simić, D., V. Svirčević, V. Ilin, S. D. Simić, and S. Simić. 2019. Particle swarm optimization and pure adaptive search in finish goods' inventory management. *Cybernetics and Systems* 50 (1):58–77. doi:10.1080/01969722.2018.1558014.

Tang, X., X. Gu, L. Rao, and J. Lu. 2021. A single fault detection method of gearbox based on random forest hybrid classifier and improved Dempster-Shafer information fusion. *Computers & Electrical Engineering* 92:107101. doi:10.1016/j.compeleceng.2021.107101.

Yu, M. C. 2011. Multi-criteria ABC analysis using artificial-intelligence-based classification techniques. *Expert Systems with Applications* 38 (4):3416–21. doi:10.1016/j.eswa.2010.08.127.