International Journal of Business Analytics
Volume 9 • Issue 1

The Impact of Data-Complexity and Team Characteristics on Performance in the Classification Model:
Findings From a Collaborative Platform

Vitara Pungpapong, Chulalongkorn Business School, Chulalongkorn University, Thailand
https://orcid.org/0000-0001-6386-6651

Prasert Kanawattanachai, Chulalongkorn Business School, Chulalongkorn University, Thailand

ABSTRACT

This article investigates the impact of data-complexity and team-specific characteristics on machine learning competition scores. Data from five real-world binary classification competitions hosted on Kaggle.com were analyzed. The data-complexity characteristics were measured in four aspects including standard measures, sparsity measures, class imbalance measures, and feature-based measures. The results showed that the higher the level of the data-complexity characteristics was, the lower the predictive ability of the machine learning model was as well. The authors' empirical evidence revealed that the imbalance ratio of the target variable was the most important factor and exhibited a nonlinear relationship with the model's predictive abilities. The imbalance ratio adversely affected the predictive performance when it reached a certain level. However, mixed results were found for the impact of team-specific characteristics measured by team size, team expertise, and the number of submissions on team performance. For high-performing teams, these factors had no impact on team score.

KEYWORDS

Big Data, Binary Classification, Data Characteristics, Data Complexity, Imbalanced Data, Machine Learning, Predictive Accuracy, Team

INTRODUCTION

In the digital age, more and more businesses are relying on machine learning algorithms to analyze an ocean of data. Undoubtedly, the field of data science has received wide attention from both practitioners and researchers over the past decade (Davenport & Patil, 2012; Jordan & Mitchell, 2015; Krittanawong, 2018). Organizations have been trying hard to maximize the usage of data to gain competitive advantage in big data era. Data analytics and machine learning have now become essential in the business world. When planning a machine learning project, firms have often focused on investing in technological infrastructure and human capital. Recognizing what determines the difficulty level of a machine learning project will also help organizations evaluate both the technical and financial feasibility of such an undertaking.

DOI: 10.4018/IJBAN.288517

Although machine learning algorithms have been rapidly enhanced and developed to improve their predictive abilities (Gilliland, 2020), little attention has been given to understanding the characteristics of data thus far. For example, imbalanced data sets are data sets where one class outnumbers the other class which are commonly found in practice. Several studies have been shown that imbalanced class leads to biased classification and thus impact the accuracy of the model. Specifically, a classifier tends to favor the majority class that can lead to a high number misclassifications of the minority class (Bekkar, Djemaa, & Alitouche, 2013; Chicco & Jurman, 2020). Prior studies have shown that dataset characteristics are important factors in determining the classification algorithm's performance (Kwon & Sim, 2013; Luengo & Herrera, 2015; Oreški & Begičević Ređep, 2018; Sánchez, Mollineda, & Sotoca, 2007). However, the measurements of data complexity used in most previous studies have been relatively simple and limited. Ho and Basu (2002) and Lorena, Garcia, Lehmann, Souto, and Ho (2019) introduced a comprehensive set of geometric and statistical measurements extracted from a training dataset to determine the difficulty in a classification problem. Although the work of Ho and Basu (2002) and Lorena et al. (2019) have great theoretical value, several of these proposed geometric measurements rely on the results from specific classifiers or algorithms, which impose a practical limitation when dealing with big and complex data. In particular, it adds another layer of analysis, with some suggesting that these algorithms involve the process of hyperparameter tuning. Although several theories have been proposed to capture the complexity of data, few theories have been empirically tested with real-world datasets.

With the advances in information technology and the high demand for data scientists, crowdsourced machine learning competitions have served as playgrounds for data scientists to get their hands dirty with real-world problems while the companies sponsoring the competitions can benefit from their solutions (Bojer & Meldgaard, 2020; Stallkamp, Schlipsing, Salmen, & Igel, 2011). As today's machine learning competitions become more complex, they encourage talents from various areas of expertise worldwide to collaborate in self-organized virtual teams to compete for rewards and build their professional profiles. Given the nature of self-organized teams, team members' backgrounds and the size of the team can be very different. Furthermore, the use of a team has proven to increase task effectiveness and efficiency (Mysirlaki & Paraskeva, 2019; Sundstrom, De Meuse, & Futrell, 1990). Two common team characteristics, 1) team intellectual capital and 2) team size (Haeussler & Sauermann, 2020; Mao, Mason, Suri, & Watts, 2016; Rasch & Tosi, 1992; Rodríguez, Sicilia, García, & Harrison, 2012), have been conceptually identified and empirically tested as important in problem-solving tasks. However, studies on the influence of virtual teams on performance in crowdsourcing contests have been limited.

Hence, this research aims to examine how the complexity characteristics of datasets affect classification predictive performance using real business-related datasets from crowdsourcing machine learning competitions. Additionally, the impact of team-specific characteristics on performance is also explored in this article. To do so, five complete binary classification competitions from Kaggle.com were selected, and the data-complexity measures were then computed. Leaderboards from these five competitions were then obtained to get team-specific characteristics as well as team scores, which represent the predictive performance. The results provided new empirical evidence on the theory of data complexity and virtual team characteristics in online, crowdsourced machine learning platforms. A better understanding of how the data complexity and team-specific characteristic impact the predictive performance can facilitate organizations in planning predictive analytics resources and personnel capabilities more appropriately.

LITERATURE REVIEW

Data Characteristics

Real-world datasets are typically complex and messy, thereby posing difficulties in a classification problem. Like in other supervised learning tasks, there is no clear winning algorithm for every possible classification dataset. Data complexity has affected the classification algorithm's performance in several studies (Kwon & Sim, 2013; Oreški & Begičević Ređep, 2018). Ho and Basu (2002) and Lorena et al. (2019) each introduced an extensive set of geometric descriptors, of which many relied on the results from specific classifiers or algorithms. For example, linearity separability can be measured using the results from a Support Vector Machine (SVM) classifier. This may lead to high computation costs and may not be feasible in practice, especially for a big and complex dataset. Table 1 shows the data-complexity characteristics proposed in previous studies.

All the measurements listed in Table 1 can be computed directly from the dataset without needing an extra data modeling step. Here, the measurements are divided into four categories, as follows:

(1) **Standard measures:** the standard measurements regarding the dimensionality of the dataset
(2) **Sparsity measures:** the measurements based on the number of features relative to the sample size and ratio of missing data
(3) **Class imbalance measures:** the measurements regarding the ratio of samples between classes
(4) **Feature-based measures:** the measurements which quantify how effective the available features are to separate classes

Standard measurements included the sample size (n) and the number of features (p) in the original dataset. The ratio of nominal features (m) was evaluated using the number of nominal features divided by the total number of features. Since the nominal data is less informative than the continuous one, a higher value of m thus indicates a higher data complexity. Typically, a set of (k - 1) indicator variables are employed to recode a nominal feature with k categories. The number of features may increase dramatically after the recoding process. Hence, it is fairer to use the number of features after pre-processing to quantify the complexity of a dataset. Other measurements evaluated using the number of features should use the number of features after recoding accordingly.

For feature-based measures, several studies have proposed the use of maximum Fisher's discriminant ratio (F1) and maximum individual feature efficiency (F3), which illustrate the best scenario in which there is at least one discriminative feature (Lorena et al., 2019). However, this study considered the average F1 and F3 across all features to evaluate the overall discriminative power of all features in a dataset.

Team Characteristics

Complex problems are rarely solved by the individual. When working in teams, team members can learn from each other to complement and accelerate productivity (Wuchty, Jones, & Uzzi, 2007). The relationship between team characteristics and team performance has been studied extensively in a broad range of fields, including psychology, economics, and management science (Mathieu, Hollenbeck, van Knippenberg, & Ilgen, 2017; Salas et al., 2008). However, teams in online crowdsourced competitions differ in nature (McComb & Maier, 2018b).

Crowdsourced competitions typically encourage participants to work in teams, which online digital technology makes seamless. Based on the definition by Kanawattanachai and Yoo (2007), a virtual team refers to "a temporary, geographically dispersed, and electronically communicating work group". In addition, crowdsourcing teams are self-forming teams. As such, team size and team members' backgrounds can vary significantly from team to team.

For predictive analytics contests, such as Netflix and Kaggle competitions, individuals or small teams are allowed to participate in zero-sum games that are highly competitive. Dissanayake, Zhang,

Table 1. Data complexity characteristics

	Measurement	Abbreviation	Min	Max	References
Standard Measures	Sample size	*n*			(Oreški & Begičević Rođep, 2018); (Li & Hung, 2009); (Kwon & Sim, 2013)
	Number of features	*p*			(Oreški & Begičević Rođep, 2018); (Li & Hung, 2009); (Kwon & Sim, 2013)
	Ratio of nominal features	*m*	0 *(simple)*	1 *(complex)*	(Kwon & Sim, 2013)
Sparsity Measures	Average number of features per dimension	S1	≈ 0 *(simple)*	*p* *(complex)*	(Basu & Ho, 2006); (Lorena et al., 2019)
	Ratio of instances which have missing values	S2	0 *(simple)*	1 *(complex)*	(Dempster, Laird, & Rubin, 1977); (Merlin, Sorjamaa, Maillet, & Lendasse, 2010); (Kwon & Sim, 2013)
Class Imbalance Measures	Entropy of Classes proportions	C1	0 *(complex)*	1 *(simple)*	(Lorena, Costa, Spolaôr, & de Souto, 2012); (Lorena et al., 2019)
	Imbalance Ratio	C2	0 *(simple)*	1 *(complex)*	(Tanwani & Farooq, 2010); (Lorena et al., 2019)
Feature-based Measures	Fisher's discriminant ratio	F1	≈ 0 *(simple)*	1 *(complex)*	(Mollineda, Sánchez, & Sotoca, 2005); (Caliński & Harabasz, 1974); (Lorena et al., 2019)
	Volume of overlapping region	F2	0 *(simple)*	1 *(complex)*	(Lorena et al., 2019)
	Individual feature efficiency	F3	0 *(simple)*	1 *(complex)*	(Lorena et al., 2019)

and Gu (2015) found that both intellectual and social capital in a virtual team had a significantly positive impact on team performance and effectiveness. The combined technical skills of all team members can be viewed as the team's intellectual capital, while their ability to work as a team effectively is regarded as their social capital. In this context, the authors primarily focused on the impact of a team's

intellectual capital on its performance because crowdsourced predictive analytics competitions are very intense and typically require collective knowledge of expertise to win. Furthermore, Dissanayake et al. (2015) also showed that the number of submissions was also significantly positively related to the rank of the team.

The relationship between team size and performance is also often of interest (Mao et al., 2016). Although several works have demonstrated the benefits of working in teams rather than as individuals, the empirical evidence in favor of larger or smaller teams is still inconclusive (Dissanayake et al., 2015). Interestingly, McComb and Maier (2018a) found that smaller teams are more likely to win crowdsourced competitions.

METHOD

Data Collection

After exploring Kaggle Competitions, the authors selected five complete binary classifications. The inclusion criteria were:

(1) competitions must be in a featured category, which typically involves full-scale machine learning and contain commercially-related competitions
(2) competitions must be from 2015 onwards
(3) competition rewards must be a monetary prize
(4) datasets must be tabular without visual or textual data
(5) the evaluation metric was the area under the receiver operating characteristic curve (ROC) Curve (AUC)
(6) real-world business problem datasets

The description of these five Kaggle featured competitions is shown in Table 2. Only teams who submitted before the competition deadline were considered in this study.

For these five competitions, both the private and public leaderboard scores and the associated data, including team name, number of team members, member rank, highest member tier, number of submissions, and submission date, were obtained via the Kaggle API and website. The data for these were then pre-processed, transformed, and merged to suit further analysis.

The training datasets from the chosen competitions were also downloaded to assess their complexity characteristics. The authors performed preliminary data pre-processing, including removing features with no variation and recoding nominal features into sets of dummy variables. To make the data more manageable, nominal classes with very low-frequency classes (less than 0.1% of samples) were grouped into a single group. The complexity characteristics were computed based on the features after data pre-processing, as shown in Table 3.

According to Kaggle, the public leaderboard score was computed based on approximately 30% of all test data, while the private leaderboard score was based on the remaining 70%. Each competition sets a limit on the number of kernel submissions differently. Whenever a team submits a kernel, Kaggle will automatically compute the public score shortly after submission and display it on the leaderboard. Teams, therefore, can benefit from the feedback on team performance right away. Furthermore, before the competition deadline, Kaggle allows each team to designate a maximum of two kernels as "Use for Final Score". The public and private scores based on the selected kernels are then computed at the end of the competition. If a team selects more than one kernel to use for the final score, the leaderboard only shows the scores computed from the kernel that received the best private score. The number of submissions from each team is also reported on the leaderboard.

Table 2. Description of five chosen competitions

Competition	# Teams	Prize	Start date	Duration (days)	Sample size	# Features	Data type
Homesite Quote Conversion	1,754	20,000	2015-11-10	92	261,000	298	numeric/ categorical data
Microsoft Malware Prediction	2,405	25,000	2018-12-14	91	8,921,483	83	numeric/ categorical data
Santander Customer Satisfaction	5,115	60,000	2016-03-03	62	76,000	370	numeric
Santander Customer Transaction Prediction	8,749	65,000	2019-02-14	57	200,000	201	numeric
Springleaf Marketing Response	2,220	100,000	2015-08-14	68	145,000	1,933	numeric/ categorical data

Table 3. Complexity characteristics of five chosen competitions (bold fonts indicate the most complex data for each dimension)

	Standard			Sparsity		Class Imbalance		Feature-based Measures		
Competition	*n*	*p*	*m*	**S1**	**S2**	**C1**	**C2**	**F1**	**F2**	**F3**
Homesite Quote Conversion	260,753	605	0.5554	0.0023	0.78579	0.6962	0.5618	0.9953	0	0.2711
Microsoft Malware Prediction	8,921,483	254	0.7953	3×10^{-3}	0.9987	1	0	0.9991	0	0.1220
Santander Customer Satisfaction	76,020	335	0	0.0044	0	**0.2403**	**0.9177**	0.9993	0	0.0986
Santander Customer Transaction Prediction	200,000	200	0	0.0010	0	0.4705	0.7793	0.9989	0	**0.9999**
Springleaf Marketing Response	145,231	37,704	**0.9502**	**0.2596**	**0.9999**	0.7824	0.4449	**0.9997**	0	0.2928

ANALYSIS

Given that several data complexity measures were computed from the same variable, of which some share a similar aspect of data complexity, multicollinearity could pose a problem in fitting the regression model. The authors thus considered ridge regression models with standardized covariates and response variables in this research. Furthermore, some of the complexity measures were omitted in the model to eliminate redundancy. For example, C1 was closely related to C2. In addition, F2 was omitted as it was found to be zero for all competitions. Based on the exploratory data analysis, the second-order form of C2 and F3 were included in the model to capture the curvilinear relationship.

For team-specific characteristics, the teams' members and the number of submissions were obtained from Kaggle's leaderboards. For each team, the Kaggle badge system progress for each team member was collected and weight-averaged to measure the team's intellectual capital. Team member composite score was calculated by assigning a weight to each badge (grandmaster=5; master=4; expert=3; contributor=2; novice=1). Higher team composite scores thus indicated a higher intellectual capital for the team. Since the team scores were nested in specific competitions, the ridge regression model can be written as

$$Score_{ij} = \beta_0 + \beta_1 m_i + \beta_2 S1_i + \beta_3 S2_i + \beta_4 C2_i + \beta_5 C2_i^2 + \beta_6 F1_i + \beta_7 F3_i + \beta_8 F3_i^2 + \beta_9 CompositeScore_{ij} + \beta_{10} TeamSize_{ij} + \beta_{11} Submissions_{ij} + \varepsilon_{ij}$$

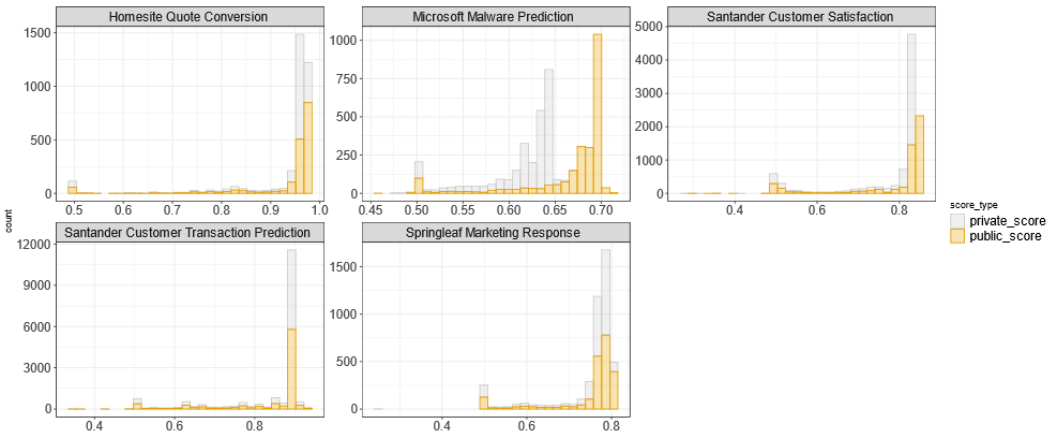
where β s represents regression coefficients and $Score_{ij}$ is the score of the j -th team in the i -th competition. The significance of each independent variable was then evaluated.

RESULTS

Exploratory Analysis

Initially, the authors explored the distribution of public and private scores obtained from Kaggle leaderboards. As shown in Figure 1, the shape of the distribution of the public and private scores was quite similar except for the Microsoft Malware Prediction competition. Unlike in other competitions, none of the teams achieved scores higher than 0.8. Furthermore, the distributions of the scores were

Figure 1. Histogram (all teams)



heavily left-skewed, making it difficult to see what happened on the right tail of the distributions. To get a clearer picture, Figure 2 showed the histograms of scores for teams that earned a medal, thereby revealing a better distinction between the public and private scores. Overall, private scores were lower than public ones. As expected, the difference between the two scores was apparent in the Microsoft Malware Prediction competition.

To further investigate the relationship between public and private scores, Pearson's correlations were calculated to measure the association. Since the winners of competitions were determined

Figure 2. Histogram (only teams earning medal)

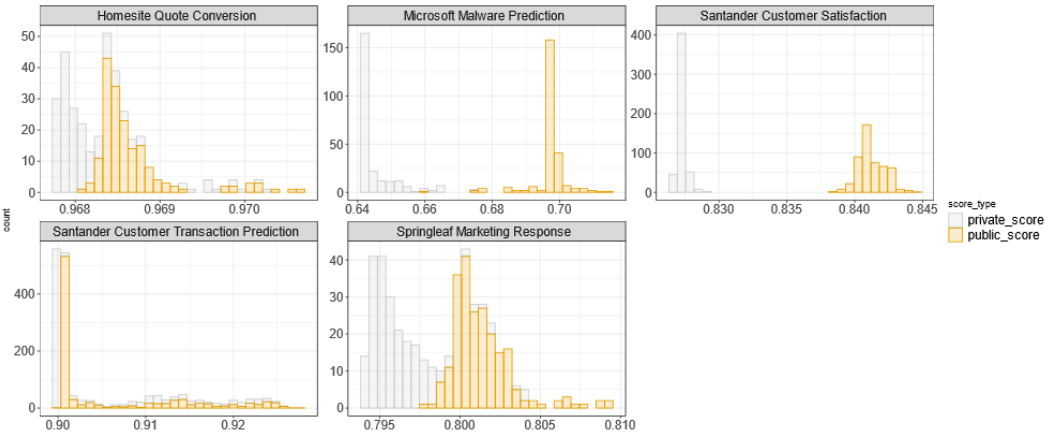


Table 4. Correlation between public and private scores

	Pearson's Correlation		Spearman's Correlation	
Competition	All Teams	Teams awarded medals	All Teams	Teams awarded medals
Homesite Quote Conversion	0.9992	0.9751	0.9851	0.8560
Microsoft Malware Prediction	0.9586	-0.1876	0.9161	0.3483
Santander Customer Satisfaction	0.9943	0.2660	0.8243	0.0999
Santander Customer Transaction Prediction	0.9986	0.9980	0.9804	0.8001
Springleaf Marketing Response	0.9910	0.9716	0.9978	0.9363

by rank, Spearman's rank correlation was also reported here. As shown in Table 4, the correlation between public and private scores was very strong, considering all of the teams in the leaderboards. However, the correlations remained high in only three competitions when considering on teams that were awarded medals alone. The development of machine learning algorithms that can generalize well was underway, indicating that predictive performance may vary wildly for different datasets given the same algorithm.

Team-specific characteristics were also summarized for each competition, as shown in Table 5. Overall, all team-specific measurements were comparable across the five competitions. Although Kaggle allows users to form a team on their own, most teams had only 1-2 members. In fact, the authors found that only a few teams had more than five members, and the maximum number was 11. The average number of submissions and team composite scores were also higher for teams who achieved a higher medal class. However, this was not true for the Microsoft Malware Prediction competition, in which the highest average number of submissions and team composite scores were among the teams that were awarded silver medals. The authors then examined the impact of data-complexity and team-specific characteristics on predictive performance, as reported in the next section.

Table 5. Descriptive statistics of team-specific characteristics (standard deviations are shown in parentheses)

Competitions	# Teams	Avg. # Submissions	Avg. team size	Avg. team composite score
Homesite Quote Conversion	1,754	20.67 (40.73)	1.09 (0.45)	2.21 (0.88)
gold	13	266.46 (142.08)	3.69 (2.39)	4.33 (0.49)
silver	74	90.68 (66.69)	1.24 (0.54)	3.42 (0.83)
bronze	88	59.84 (72.79)	1.26 (0.75)	3.15 (0.93)
no medal	1,579	13.19 (17.94)	1.05 (0.28)	2.08 (0.78)
Microsoft Malware Prediction	2,405	18.02 (38.00)	1.19 (0.66)	1.92 (1.02)
gold	14	49.57 (63.85)	1.21 (0.43)	2.29 (1.09)
silver	107	77.96 (96.58)	1.45 (1.04)	2.83 (1.23)
bronze	121	18.14 (36.55)	1.27 (0.84)	2.41 (0.99)
no medal	2,163	14.84 (29.24)	1.17 (0.63)	1.84 (0.98)
Santander Customer Satisfaction	5,115	18.24 (28.52)	1.11 (0.50)	2.04 (0.76)
gold	20	111.25 (99.99)	2.05 (2.31)	3.84 (1.02)
silver	235	32.39 (33.51)	1.16 (0.52)	2.53 (1.04)
bronze	256	29.32 (38.19)	1.12 (0.54)	2.45 (0.92)
no medal	4,604	16.50 (25.83)	1.11 (0.47)	1.99 (0.70)
Santander Customer Transaction Prediction	8,749	11.89 (16.59)	1.12 (0.46)	1.62 (0.85)
gold	27	93.81 (44.77)	2.93 (1.47)	4.18 (0.79)
silver	412	35.90 (31.32)	1.34 (0.85)	2.83 (1.02)
bronze	440	20.40 (19.88)	1.18 (0.50)	2.21 (0.96)
no medal	7,870	9.88 (12.83)	1.10 (0.41)	1.51 (0.75)
Springleaf Marketing Response	2,220	17.67 (30.15)	1.12 (0.50)	2.20 (0.86)
gold	14	146.07 (91.11)	2.64 (2.02)	4.40 (0.55)
silver	97	76.14 (59.55)	1.41 (0.99)	3.25 (0.90)
bronze	111	50.05 (43.74)	1.19 (0.61)	2.84 (0.95)
no medal	1,998	12.14 (17.60)	1.09 (0.40)	2.10 (0.79)
Overall	20,243	15.62 (27.39)	1.12 (0.50)	1.87 (0.89)

The Impact of Data-complexity and Team Characteristics on Scores

Four ridge regression models were fitted to investigate the impact of team and data-complexity characteristics on both public and private scores. Model 1 and Model 2 used all entries in the leaderboards from the five competitions to fit the models, while Models 3 and 4 focused only on teams that were awarded medals. The authors considered both public and private scores as response variables. The model coefficients along with their corresponding standard error and p-values are shown in Table 6 and Table 7. The R^2 and adjusted R^2 for each model are also reported.

Data-Complexity Characteristics and Performance

In terms of the data complexity variables, all measurements, regardless of the ratio of missing instances (S2), were found to statistically associate with the competition scores at the 5% level of significance

Table 6. Ridge regression results: all teams in leaderboards (n=20,243)

	Model 1 (public score)					Model 2 (private score)			
	Beta	**SE**	**p-value**			**Beta**	**SE**	**p-value**	
Data Characteristics									
m	-0.02259	0.00255	2×10^{-5}	***		-0.01961	0.00255	6×10^{-5}	***
S1	-0.08402	0.00550	3×10^{-7}	***		-0.06148	0.00550	4×10^{-6}	***
S2	-0.00001	0.00211	0.996			-0.00068	0.00211	0.7552	
C2	0.02162	0.00390	5×10^{-4}	***		0.03126	0.00390	4×10^{-5}	***
$C2^2$	-0.19752	0.00307	4×10^{-12}	***		-0.24415	0.00307	7×10^{-13}	***
F1	-0.25912	0.00645	$2. \times 10^{-10}$	***		-0.26181	0.00645	1×10^{-10}	***
F3	0.14079	0.00755	7×10^{-8}	***		0.15838	0.00755	3×10^{-8}	***
$F3^2$	-0.00809	0.00656	0.252			-0.01194	0.00656	0.1061	
Team Characteristics									
CompositeScore	0.12567	0.00731	2×10^{-7}	***		0.11635	0.00768	4×10^{-7}	***
TeamSize	0.00784	0.00768	0.315			0.00703	0.00731	0.3645	
Submissions	0. 13419	0.00773	1×10^{-7}	***		0.12171	0.00773	3×10^{-7}	***
R2		0.35976					0.44677		
Adjusted-R2		0.35945					0.44649		
* Sig. at $\alpha = 0.1$									
** Sig. at $\alpha = 0.05$									
*** Sig. at $\alpha = 0.01$									

across all four models. Specifically, the ratio of nominal variables (*m*), the average number of features per dimension (S1), and Fisher's discriminant ratio (F1) were negatively related to both public and private scores, indicating that the scores became lower for more complex data. For C2, which measured the level of imbalance, the authors found that the coefficient of the first-order term was positive, while the coefficient of the second-order term was negative. In addition, both the first- and second-order terms of C2 were statistically significant. The results, therefore, suggested that the predictive accuracy was not hurt by the small value of C2 until it reached a certain level, at which point it had a negative impact on the performance. Individual feature efficiency (F3), which measured the ratio of overlapping individuals, also behaved in a similar way as C2 visually. Particularly, empirical results showed that accuracy dropped when C2 surpassed 0.6 and F3 surpassed 0.3. The estimated coefficients of the first-order and second-order terms of F3 were positive and negative, respectively; thus, the model was able to capture the curvature pattern. However, only the first-order term was found to be statistically significant. In sum, all of the data-complexity characteristics except S1, which represented the ratio of missing data, were associated with scores. The findings thus demonstrated that more complex data resulted in lower competition scores.

Table 7. Ridge regression results: only teams awarded medals (n=2,029)

	Model 3 (public score)					Model 4 (private score)			
	Beta	**SE**	**p-value**			**Beta**	**SE**	**p-value**	
Data Characteristics									
m	-0.04636	0.00823	5×10^{-4}	***		-0.03590	0.00823	0.002	***
S1	-0.11103	0.01770	2×10^{-4}	***		-0.06656	0.01770	0.006	***
S2	-0.01679	0.00802	0.070	*		-0.01413	0.00802	0.116	
C2	0.05690	0.01336	0.003	***		0.06279	0.01336	0.002	***
$C2^2$	-0.36185	0.01007	4×10^{-10}	***		-0.38494	0.01007	2×10^{-10}	***
F1	-0.39215	0.02103	7×10^{-8}	***		-0.35313	0.02103	2×10^{-7}	***
F3	0.22750	0.02358	1×10^{-5}	***		0.23023	0.02358	1×10^{-5}	***
$F3^2$	-0.01488	0.02088	0.496			-0.01838	0.02088	0.404	
Team Characteristics									
CompositeScore	0.02301	0.02375	0.361			0.01676	0.02375	0.500	
TeamSize	0.01338	0.02444	0.599			0.01040	0.02444	0.682	
Submissions	0.01215	0.02666	0.661			0.00843	0.02666	0.760	
R^2		0.99534					0.99665		
Adjusted-R^2		0.99532					0.99663		
* Sig. at $\alpha = 0.1$									
** Sig. at $\alpha = 0.05$									
*** Sig. at $\alpha = 0.01$									

Team Characteristics and Performance

Based on the results of all entries in the leaderboards, Model 1 and Model 2 revealed that the team's composite score and the number of submissions were significantly associated with team scores. Hence, the authors concluded that the intellectual capital of each individual in a team as well as their engagement were important and had a significant positive impact on team performance. However, when focusing on only teams that were awarded medals, none of the team characteristics were significant, while better fit models were obtained ($R^2 > 0.99$). This implied that among top-performing teams, the effect of the data characteristics dominated the team ones.

DISCUSSION AND CONCLUSION

Findings

As an exploratory study to better understand the impact of data-complexity and team characteristics on the accuracy of the machine learning model, the current research provides several insights into how each aspect of both data complexity and team's characteristics impact the model evaluation metric score of five real-world binary classification machine learning competitions.

In particular, results from the ridge regression model indicated that the higher the complexity level of data, as measured by m, S1, C2, F1, and F3, was, the lower the evaluation metric score was as well. In fact, C2, which measured the level of imbalance of the target variable, was the most important

variable as it had the smallest, and therefore, most significant, p-value. Specifically, C2 had a nonlinear relationship with model performance. When C2 was low, it did not hurt model performance. However, once C2 passed a certain point, it had a negative impact on the evaluation metric score. This suggests that a well-crafted model can withstand the level of imbalance in the target variable up to a certain level. On the other hand, the ratio of missing instances (S2) had no impact on model performance. Due to the data limitations, however, the authors did not know which strategy each team used in dealing with missing data. Further investigation is thus needed in future research.

In addition to the data-complexity characteristics, team-specific characteristics, as measured by team composition (based on Kaggle's badge system), and performance feedback, as measured by the number of submissions, also positively affected the evaluation metric score. In line with previous studies, team size had no impact on model performance (Dissanayake et al., 2015). Unexpectedly, however, all team characteristics for top-performing teams were not associated with competition scores. In summary, the effect of data-complexity characteristics dominated performance score.

Limitations

Although this study has several contributions to the field, there are still several limitations that the reader should consider in evaluating the results. Although the results were based on a real-world business problem, only five binary classification problems were considered in this study. In addition, most of the top-performing teams in leaderboards did not share their codes either on Kaggle.com or GitHub.com. Hence, the relationship between the complexity of datasets on a specific algorithm could not be broadly examined.

Because the findings on the impact of team characteristics on performance are insignificant, it will be important in future work to examine the other aspects of team such as team interdependence (Gully, Incalcaterra, Joshi, & Beaubien, 2002), and shared mental model (Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000) in a highly cognitive demanding task.

Implications

Theoretical Implications

Despite limitations, a meaningful and interpretable pattern of relationships among the data-complexity characteristics, team's characteristics, and predictive ability of the machine learning model was attained. As an initial empirical study exploring the data-complexity characteristics of real-world binary classification problems, this study provides several important implications.

Although the past literature introduced how to measure the data-complexity measures in various dimensions, very little is known about the form of relationship between these measures and model performance. This study discovered that all of the data-complexity measures except S2 (ratio of instances which have missing values) were significantly associated with competition score. Apparently, this study went a step further by empirically proving that there is a curvilinear relationship between class imbalance ratio (C2) and the accuracy of machine learning model.

Practical Implications

Implementing a successful machine learning project has been a new challenge for all organizations in digital era. When planning any machine learning project, an organization must be able to justify the cost and benefit of it. Estimating the cost for a machine learning project in advance is not an easy task (Wamba et al., 2017). Several studies have shown that predictive ability of classification models highly depend on data complexity (Sánchez et al., 2007). Hence, organizations that plan to implement and deploy machine learning need to pay attention not only to technological infrastructure and human capital but also data complexity. With a better understanding of how data-complexity characteristics directly impact the model evaluation score, the project manager will have a more objective way to measure the difficulty of such a project to properly allocate the needed resources.

Among data complexity measurements, the results from this study showed that the class imbalance measure played the most important role in determining performance of classification model. Although a well-crafted model can handle small degree of imbalance, severely imbalanced class has an impact on predictive accuracy. The imbalance ratio is also very easy to compute. Thus, the authors suggest that a machine learning project manager should put the imbalance ratio at the top of the list when evaluate data complexity. Currently, there have been many approaches to tackle imbalanced data. The most popular methods include sampling methods and cost-sensitive learning methods (He & Garcia, 2009). The sampling methods attempt to get balance distributions in a training dataset while cost-sensitive learning methods take the costs associated with misclassifying examples into account. Organizations can no longer ignore the fact that data preprocessing stage is the most time-consuming activity in machine learning project (Munson, 2012). The authors agree that it is worth to invest time and effort to handle severely imbalanced class properly to avoid a biased classifier. In addition, choosing an appropriate metric to assess the performance of a classifier is also important when dealing with imbalanced data. In addition to its popularity, AUC has been proved to be a consistent and robust metric to measure on imbalanced condition (Halimu, Kasem, & Newaz, 2019; Wardhani, Rochayani, Iriany, Sulistyono, & Lestantyo, 2019).

REFERENCES

Basu, M., & Ho, T. K. (2006). *Data Complexity in Pattern Recognition* (M. Basu & T. K. Ho, Eds.). Springer. doi:10.1007/978-1-84628-172-3

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*, *3*(10).

Bojer, C. S., & Meldgaard, J. P. (2020). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27. doi:10.1080/03610927408827101

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13. doi:10.1186/s12864-019-6413-7 PMID:31898477

Davenport, T. H., & Patil, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, *90*(10), 70–76. PMID:23074866

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society. Series B. Methodological*, *39*(1), 1–22. doi:10.1111/j.2517-6161.1977.tb01600.x

Dissanayake, I., Zhang, J., & Gu, B. (2015). Task Division for Team Success in Crowdsourcing Contests: Resource Allocation and Alignment Effects. *Journal of Management Information Systems*, *32*(2), 8–39. doi:10.1080/07421222.2015.1068604

Gilliland, M. (2020). The value added by machine learning approaches in forecasting. *International Journal of Forecasting*, *36*(1), 161–166. doi:10.1016/j.ijforecast.2019.04.016

Gully, S. M., Incalcaterra, K. A., Joshi, A., & Beaubien, J. M. (2002). A meta-analysis of team-efficacy, potency, and performance: Interdependence and level of analysis as moderators of observed relationships. *The Journal of Applied Psychology*, *87*(5), 819–832. doi:10.1037/0021-9010.87.5.819 PMID:12395807

Haeussler, C., & Sauermann, H. (2020). Division of labor in collaborative knowledge production: The role of team size and interdisciplinarity. *Research Policy*, *49*(6), 103987. Advance online publication. doi:10.1016/j.respol.2020.103987

Halimu, C., Kasem, A., & Newaz, S. S. (2019). Empirical comparison of area under ROC curve (AUC) and Mathew correlation coefficient (MCC) for evaluating machine learning algorithms on imbalanced datasets for binary classification. *Proceedings of the 3rd international conference on machine learning and soft computing*. doi:10.1145/3310986.3311023

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. doi:10.1109/TKDE.2008.239

Ho, T. K., & Basu, M. (2002). Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(3), 289–300. doi:10.1109/34.990132

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. doi:10.1126/science.aaa8415 PMID:26185243

Kanawattanachai, P., & Yoo, Y. (2007). The impact of knowledge coordination on virtual team performance over time. *Management Information Systems Quarterly*, *31*(4), 783–808. doi:10.2307/25148820

Krittanawong, C. (2018). The rise of artificial intelligence and the uncertain future for physicians. *European Journal of Internal Medicine*, *48*, e13–e14. doi:10.1016/j.ejim.2017.06.017 PMID:28651747

Kwon, O., & Sim, J. M. (2013). Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications*, *40*(5), 1847–1857. doi:10.1016/j.eswa.2012.09.017

Li, Y., & Hung, E. (2009). *Building a Decision Cluster Forest Model to Classify High Dimensional Data with Multi-classes*. doi:10.1007/978-3-642-05224-8_21

Lorena, A. C., Costa, I. G., Spolaôr, N., & de Souto, M. C. P. (2012). Analysis of complexity indices for classification problems: Cancer gene expression data. *Neurocomputing*, *75*(1), 33–42. doi:10.1016/j.neucom.2011.03.054

Lorena, A. C., Garcia, L. P. F., Lehmann, J., Souto, M. C. P., & Ho, T. K. (2019). How Complex Is Your Classification Problem? A Survey on Measuring Classification Complexity. *ACM Computing Surveys*, *52*(5), 107. Advance online publication. doi:10.1145/3347711

Luengo, J., & Herrera, F. (2015). An automatic extraction method of the domains of competence for learning classifiers using data complexity measures. *Knowledge and Information Systems*, *42*(1), 147–180. doi:10.1007/s10115-013-0700-4

Mao, A., Mason, W., Suri, S., & Watts, D. J. (2016). An Experimental Study of Team Size and Performance on a Complex Task. *PLoS One*, *11*(4), e0153048. doi:10.1371/journal.pone.0153048 PMID:27082239

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *The Journal of Applied Psychology*, *85*(2), 273–283. doi:10.1037/0021-9010.85.2.273 PMID:10783543

Mathieu, J. E., Hollenbeck, J. R., van Knippenberg, D., & Ilgen, D. R. (2017). A century of work teams in the Journal of Applied Psychology. *The Journal of Applied Psychology*, *102*(3), 452–467. doi:10.1037/apl0000128 PMID:28150984

McComb, C., & Maier, T. (2018a). Designing Improved Teams for Crowdsourced Competitions. *Proceedings of the ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference 30th International Conference on Design Theory and Methodology*.

McComb, C., & Maier, T. (2018b). *Designing improved teams for crowdsourced competitions.* Paper presented at the International Design Engineering Technical Conferences and Computers and Information in Engineering Conference.

Merlin, P., Sorjamaa, A., Maillet, B., & Lendasse, A. (2010). X-SOM and L-SOM: A double classification approach for missing value imputation. *Neurocomputing*, *73*(7), 1103–1108. doi:10.1016/j.neucom.2009.11.019

Mollineda, R. A., Sánchez, J. S., & Sotoca, J. M. (2005). *Data Characterization for Effective Prototype Selection*. doi:10.1007/11492542_4

Munson, M. A. (2012). A study on the importance of and time spent on different modeling steps. *SIGKDD Explorations*, *13*(2), 65–71. doi:10.1145/2207243.2207253

Mysirlaki, S., & Paraskeva, F. (2019). Virtual Team Effectiveness: Insights from the Virtual World Teams of Massively Multiplayer Online Games. *The Journal of Leadership Studies*, *13*(1), 36–55. doi:10.1002/jls.21608

Oreški, D., & Begičević Redep, N. (2018). Data-driven decision-making in classification algorithm selection. *Journal of Decision Systems, 27*(sup1), 248-255. 10.1080/12460125.2018.1468168

Rasch, R. H., & Tosi, H. L. (1992). Factors affecting software developers' performance: An integrated approach. *Management Information Systems Quarterly*, *16*(3), 395–413. doi:10.2307/249535

Rodríguez, D., Sicilia, M. A., García, E., & Harrison, R. (2012). Empirical findings on team size and productivity in software development. *Journal of Systems and Software*, *85*(3), 562–570. doi:10.1016/j.jss.2011.09.009

Salas, E., DiazGranados, D., Klein, C., Burke, C. S., Stagl, K. C., Goodwin, G. F., & Halpin, S. M. (2008). Does team training improve team performance? A meta-analysis. *Human Factors*, *50*(6), 903–933. doi:10.1518/001872008X375009 PMID:19292013

Sánchez, J. S., Mollineda, R. A., & Sotoca, J. M. (2007). An analysis of how training data complexity affects the nearest neighbor classifiers. *Pattern Analysis & Applications*, *10*(3), 189–201. doi:10.1007/s10044-007-0061-2

Stallkamp, J., Schlipsing, M., Salmen, J., & Igel, C. (2011). *The German Traffic Sign Recognition Benchmark: A multi-class classification competition.* Paper presented at the 2011 International Joint Conference on Neural Networks.

Sundstrom, E., De Meuse, K. P., & Futrell, D. (1990). Work teams: Applications and effectiveness. *The American Psychologist*, *45*(2), 120–133. doi:10.1037/0003-066X.45.2.120

Tanwani, A. K., & Farooq, M. (2010). Classification Potential vs. Classification Accuracy: A Comprehensive Study of Evolutionary Algorithms with Biomedical Datasets. Academic Press.

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, *70*, 356–365. doi:10.1016/j.jbusres.2016.08.009

Wardhani, N. W. S., Rochayani, M. Y., Iriany, A., Sulistyono, A. D., & Lestantyo, P. (2019). *Cross-validation metrics for evaluating classification performance on imbalanced data.* Paper presented at the 2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA). doi:10.1109/IC3INA48034.2019.8949568

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. *Science*, *316*(5827), 1036–1039. doi:10.1126/science.1136099 PMID:17431139

Vitara Pungpapong (Ph.D. Purdue University) is an Assistant Professor in the Chulalongkorn Business School at Chulalongkorn University, Thailand. She received the B.S. in Statistics with the concentration in Business Information Technology from Chulalongkorn University, Bangkok, Thailand, A.M. degree in Statistics from Harvard University, Cambridge, MA, USA and Ph.D. degree in Statistics from Purdue University, West Lafayette, IN, USA. Her research interests include data modeling for massive data, machine learning, and computational statistics.

Prasert Kanawattanachai (Ph.D. Case Western Reserve University) is an Associate Professor in the Chulalongkorn Business School at Chulalongkorn University, Thailand. His research interests include Virtual Team, Business Simulation Game, Knowledge Representation and Trust. His work has published in journals such as MIS Quarterly, Academy of Management Journal, The Journal of Strategic Information Systems, Journal of Management Education, and International Journal of Organization Analysis.