Bridging the Gap: Introducing a Universal Data Monetization Method from Information and Game Theories

Domingos S. M. P. Monteiro, Felipe Silva Ferraz,
Silvio R. L. Meira
Center of Advanced Studies and Systems of Recife
Recife, Brazil
E-mail: {dsmpm, fsf, srlm}@cesar.school

Domingos S. P. Salazar
Distance Education and Technology Unit
Rural Federal University of Pernambuco
Recife, Brazil
domingos.salazar@ufrpe.br

Abstract— **Despite significant research on data monetization in recent years, the academic literature lacks universally applicable methods for this endeavor. This study seeks to introduce a versatile method suitable for various databases and prevalent challenges in both academic and commercial realms. Our methodology draws from information theory and game theory, leveraging the Return On Investment (ROI) metric as a value determinant. The derived method calculates the ROI contributed by distinct databases for binary decision-making, incorporating the Shapley Value concept from cooperative game theory. We tested this method on a practical dilemma—underwriting car insurance policies in Brazil. Our method adeptly pinpointed the financial contribution of each dataset to the assessed decisions. It can be adapted for other binary decision contexts where financial outcomes of decisions are either provided or quantifiable. Given the novelty of this research domain, we anticipate this study to spur further exploration into data valuation in the realm of data science and Big Data.**

Keywords-Big Data; Data Value; Big Data Monetization; Artificial Intelligence; Game Theory; Information Theory; Shapley Value; Digital Assets.

I. INTRODUCTION

Despite the advancements in Big Data applications, in research related to the subject, and in the widespread application of analytical and artificial intelligence solutions in recent decades, a method has not yet been established that can be widely disseminated to determine the intrinsic value of a specific piece of data (or specific database) within a Big Data context. Academic research, in this context, has provided little emphasis on the dimension of Data Value, especially when contrasted with investigations directed at the other three classic dimensions of Big Data, namely: Volume, Velocity, and Variety [1]. In the more specific context of financial or economic value, research is even scarcer, and available studies do not share a common view on methods for its measurement [2].

The aim of this research is to create a flexible method that can be applied to any database that becomes available to solve a specific problem, both in the academic and business environments. At this moment will center our proposal on the binary decision-making problem involving risk as explaned in section III. The exploration of this method aims to shed light on a gap in the domain of data analysis and valuation, providing a methodological structure that can be used effectively and relevantly in different contexts, regardless of the specifics inherent to the data.

Every day, new data is collected from various sources, formats, and domains. A lack of information can lead to inefficient decision-making, thus making the impacts of these decisions less predictable or riskier [3]. Given this scenario, our research questions were:

- Is it possible to develop a method to measure the financial impact (value) of new data available for a binary decicion making problem?
- Can the new information lead to more predictable and efficient decision-making?

The data monetization method proposed in this study is based on the concept of Return On Investment (ROI) [4] provided by different databases that become available for binary decision-making. In the Big Data context, this is a common and realistic scenario since new data is constantly arriving in larger volumes, with greater speed (velocity) and variety [1]. In the research phase, in the searching for suitables methods, we evaluated the application of both information theory [5] and game theory, and the final formulated method was based on the Shapley Value concept borrowed from cooperative game theory [6].

To validate our method, we have executed a series of controlled experiments applied to a real decision-making problem of underwriting car insurance policies in the Brazilian market. For our experiments, we used two databases provided by a partner company of the project, Neurotech SA (Neurotech) [7], and conducted eight (8) different experiments considering the results of each database individually and the combined databases for two (2) distinct real problems with two (2) distincts combination arrangements:

1. Claims: represents the occurrence of a covered risk during the insurance plan's validity period, and;
2. Theft: represents the occurrence of subtraction of the insured asset during the insurance plan's validity period.

The application of the proposed method was able to precisely isolate the financial value added by each of the databases used for the different decisions evaluated, and these results shed significant light on the research problem in focus.

The proposed method offers the possibility of replication in a multitude of scenarios characterized by problems of a similar nature [8]. Essentially, it is pertinent to those that constitute binary decisions, in which the assessment of the financial gain or loss of each decision, in itself, is provided by concrete information or can be duly inferred or measured. The reach of this method goes beyond its initial application, extending to various contexts that share similar characteristics.

We hope that this study can stimulate the development of other methodological approaches, whether derivations of the method proposed in this study or as new proposals themselves. We also hope that this stimulation can contribute to a more comprehensive and insightful understanding of data value in the universe of data science, when inserted into the complex environments that characterize Big Data.

This article is structured as follows: in Section 2, we explore the context that we have adopted in our study related to the topic of Big Data Monetization and detail our main objective; in Section 3, we present the data our experiments relyed on; in Section 4, we discuss how to apply information and games theorie to the problem; in Section 5, we proposed the method; in Section 6, we provide details on the related experiments; in Section 7, we present the method results to the available data; in Section 8, we present the conclusion and suggest further future studies.

II. Data: a Valuable Digital Asset

In a 2016 review comprising over a thousand and five hundred studies that mentioned the term “Big Data”, De Mauro et al. [9] proposed the following definition that would be able to bundle most of the assessed texts: “Big Data is the information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value”. This definition is the one taken into consideration in our study and what makes the proposed method even more relevant, once it highlights that the explicit objective of a Big Data environment is to turn digital assets into Value.

The debate on how “value” itself was formed has been going on for millennia, since pre-Christian era, when Aristotle argued that value is based on the need for exchange (Aristotle, 350 BC) [11]. This concept is in the core of economic adjustment and is the basis to define what will be produced, how it will be produced and who will produce it.

Another key discussion in economic theory includes questioning the reasons for a product or service to be priced the way it is, that is, how the value of a product or service is determined and how to calculate it correctly [12].

The theory was formulated and applied in a world where products and services were in their entirety represented by physical assets with well-defined characteristics: raw material, finished products and services provided by physical living beings (humans and animals) [13].

The advent of computers brought the world a new category of assets, digital ones, represented in a discrete numerical way and used in digital devices with computational processing. These digital assets are capable of delivering a new category of products and services: better decisions, increased performance, competitive advantages and they can even be sold directly as a product [10]. It is in this context of “data” as a digital asset and as a product itself that we will propose a way for its monetization in this study.

Our objective in this project was to apply a strategy that can accurately estimate the value of data in a real-world situation, using concepts from information theory and game theory in conjunction with machine learning. We will center our proposal on finding value of data to the binary decision-making problem involving risk. This decision was driven by two primary reasons:

1. The operational focus of our partner company, Neurotech, which has developed and implemented thousands of solutions for binary decision problems, impacting millions of decisions made daily by its clients related to credit risk analysis, underwriting insurance policies, among other areas (retail, finance, health plans, etc.);
2. Being a class of problems well-known and researched by the academic community [14];

Although our proposal concentrates on the binary decision-making problem, most used for classification purposes, these types of solutions can be grouped into decision trees that are applicable for both multiple classifications and regression [15]. Hence, the generalization of this method can encompass both classes of problems.

A. Value Search

The price of data can depend on various premises, such as its acquisition cost, its storage and update cost, its scarcity, etc. However, as Warren Buffett remarked regarding financial assets, "Price is what you pay, value is what you get" [16]. In other words, the value of an asset is an intrinsic characteristic that differs from its price. In the case of investments, Buffet evaluates a company, for instance, based on its ability to generate future profits.

An analogy with data assets would be that their value, and not their price, depends on their ability to inform better decision-making [17], which is often directly linked to a company's operational profit. In this way, the value of data can be mapped to the quantification of this data's capacity to enhance decision-making.:

$$Value\ of\ data = V(data, decision) \qquad (1)$$

It's interesting to note that the decision is part of the equation. In other words, even if the acquisition cost remains unchanged, data can have different values for different decision-making processes. Thus, it is expected that a rational agent would only purchase specific data if it were traded at a price equal to or less than its added value; for our method, this would be represented by a positive ROI upon the addition of the new data. However, quantifying this value can be a complex issue.

We evaluated information theory and game theory as potential paths for our method. We will delve deeper into these possibilities in the following sections using a subset of our experimental data, which we detail from this point forward.

III. AVALIABLE DATA SET

The data used in this study were provided by the partnering organization of this research. The company, Neurotech SA, is a leading data and analytics provider for the Brazilian market that serves over 200 major corporations, including large retailers, banks, financial institutions, and insurance companies in Brazil. We considered the initial database (*DB1*) as that containing Neurotech's proprietary and public collected data consisting of roughly 3.3 million vehicles, and their respective owners.

In our experiments, we have joined the former database (*DB1)* with a new database provided by a third-party company specialized in collecting data on automatic payments via radio-frequency identification (RFID) [18], often used in toll payments, commonly known as TAG. This database was considered to represent the new data available for the problem (*DB2*).

We investigated how the value of this new data *(DB2)* could be determined for monetization purposes, using the method detailed here. We selected an auto insurance company to test our method.

Currently, Neurotech has millions of car insurance quotes transacted monthly on its platforms that consider the initial database (*DB1*) in their analyses. In this project, we will supplement these quotes with the data from the new database to understand if this new data can assist in the risk decision-making of policy underwriting compared to decisions based solely on the original base.

For our project, a sample of 120 million transactions from *DB2* was provided. A transaction in this context means an event related to a TAG, such as passing through a particular toll. While toll usage is the most common application for a TAG, the market now allows TAG use in transactions at affiliated networks that involve not just tolls, but also use in parking lots, refueling at gas stations, and even purchases at some fast-food drive-thrus and restaurants. These 120 million transactions involve 2.2 million TAGs, roughly 2.2 million vehicles, and their respective owners.

The next step was to combine the consolidated new data (*DB2*) with the original database (*DB1*). Of the 2.2 million TAGs provided, we found similar keys (vehicle/owner) in the original base for 540,000 of these TAGs. The result leads us to an initial conclusion that approximately 25% of the TAG holders present in *DB2* sample also went through *DB1* of the partner company seeking insurance for their car.

As a result, we were able to enrich 540,000 policies from the original database *(DB1)* with the data from the new database *(DB2)*. This represented 16.5% of the partner company's original database, meaning 16.5% of the policies transacted in the original database involved individuals who had transaction data from their TAGs available for enrichment from the new database. We summarize these conclusions in the Venn diagram below.

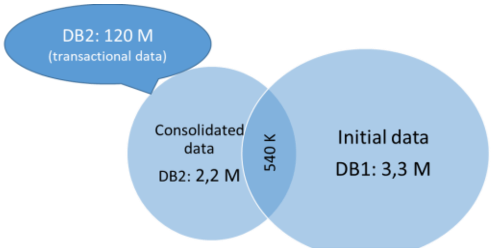


Figure 1. Data: Transformations and Enrichment.

IV. VALIDATING THE THEORIES

Our goal in this project was to apply a strategy that can accurately estimate the value of data in a real-world situation, by applying concepts from information theory and game theory in conjunction with machine learning.

A. Testing theories with a data samples

For the purpose of comparing theories and defining our strategy, we initially considered only one of the focal problems: identification of Theft or Robbery. In this example, we used a random sample from the available database consisting of 328,565 annual car insurance policies. Among them, 756 had an occurrence of a certain type of theft or robbery claim (0.23% of the cases), and 327,809 did not record the occurrence of this type of claim. The database consists of various explanatory variables that were available (or known) at the beginning of the underwriting process. The goal is to quantify the value that this information has with the aim of predicting the occurrence of the claim. For this, we will use the approach of information theory based on mutual information [5] and game theory considering the Shapley Value [6].

Consider the following variables and their respective descriptions.

TABLE I. VARIABLE DESCRIPTION

Variable	Descriptions
MEDIA_DISTANCIA_PARCEIROS	Average distance between the residential address and the nearest point of interest.
STD_VALOR_TRANSACAO	Standard deviation of the historical transactional values (in BRL) of the vehicle with partners..
QTD_NOITE	Number of vehicle transactions during the nighttime.
QTD_MADRUGADA	Number of vehicle transactions during the early morning hours.
MAIOR_DISTANCIA_PARCEIROS	Minimum distance between the residential address and the nearest point of interest.
QTD_TAGS	Number of tags registered for the vehicle in question.
QTD_TARDE	Number of vehicle transactions during the afternoon.
MIN_VALOR_TRANSACAO	Minimum among the historical transactional values (in BRL) of the vehicle with partners..
MEAN_VALOR_TRANSACAO	Average of the historical transactional values (in BRL) of the vehicle with partners.
QTD_PARCEIROS	Number of points of interest registered for the vehicle in question.

The selected variables are numerical, and depending on their value, one may observe a greater or lesser quantity of the target class (occurrence of Theft or Robbery). The dependency between each variable and the target class (Y=1) was represented in a bivariate analysis, and we will present below, for illustrative purposes, the result for the variable QTD_TAGS. The bivariate analyses for the other variables are available at the following Kaggle reference [19]. Those analysis supported the Entropy calculation when applying Information Theory.

TABLE II. FREQUENCY OF THEFT AND BURGLARY OCCURRENCES FOR THE VARIABLE QTD_TAGS

QTD_TAGS	Y=1	Y=0	Total
1.0 - 1.0	0,1%	22,9%	23%
2.0 - 3.0	0,1%	27,1%	27%
4.0 - 4.0	0,0%	12,3%	12%
5.0 - 7.0	0,0%	21,8%	22%
8.0 - 345.0	0,0%	15,6%	16%
TOTAL	0,2%	99,8%	100%
QTD_TAGS	Y=1	Y=0	Total

B. Apply Information Theory

Shannon's information theory is not directly applied to determine the specific value of a piece of data itself, but rather to quantify the information contained in a dataset or to understand how information is transmitted and processed. However, information theory can be used to address prediction and probability problems, and thus we will test an approach considering Shannon's information theory. Consider the Shannon entropy of a given information X [5],

$$H(X) = -\sum_i \quad p(x_i)\, log_2\, p(x_i) \qquad (2)$$

Where:

- $H(X)$ is the entropy of the information source X
- $p(x_i)$ is the probability of occurrence of the symbol xi na fonte de informação.
- The logarithm is in base 2, which measures information in bits.

Note that this magnitude depends solely on the given data X in question, but it does not depend on the decision for which the data X will be used. For this reason, it does not meet our value criteria in (1) and therefore would not be applicable to our method. One possible way to incorporate information theory into our method, to quantify the impact that a data point X has on a decision Y, would be to measure the mutual information:

$$I(X;Y) = -\sum_i \quad p(x_i,y_i)\, log_2\, [p(x_i,y_i)/p(x_i)q(y_i)], \qquad (3)$$

Where:

- $I(X;Y)$ is the mutual information between sources X and Y.
- $p(x_i,y_i)$ is the joint probability of xi and yj occurring in sources X and Y, respectively.
- p(xi) and q(yi) are the marginal probabilities of X and Y, respectively.
- Y is the target, we would like to predict.

In a binary decision problem, Y would be 0 or 1, suggesting two possible events (or decisions). Meanwhile, p(x,y) represents the probability of observing the data X=x and the target Y=y simultaneously, while p(x) and q(y) are the probabilities of observing X=x and Y=y, respectively. Intuitively, mutual information is the gain in information we have regarding the decision Y, given that X is known. In terms of entropy, it can be written as:

$$I(X;Y) = H(Y) - H(Y|X), \qquad (4)$$

Where H(Y|X) is the conditional entropy.

$$H(Y|X) = -\sum_i \quad p(x_i,y_i)\, log_2\, [p(x_i)/p(x_i,y_i)], \qquad (5)$$

Where:

- $H(Y|X)$ is the conditional entropy of Y given X
- $p(x_i,y_i)$ is the joint probability of xi and yi occurring in sources X and Y, respectively.

Note that, if X and Y are independent, we have $p(x_i,y_i) = p(x_i)q(y_i)$, which would result in $H(Y|X)=H(Y)$ and $(X;Y)=0$. That is, the information gain is null when knowing data X, since the decision Y does not depend on this data. In this specific case, it's expected that the data holds no value for this decision-making process. Alternatively, if X truly provides some information gain regarding the decision Y, we should have $I(X;Y)>0$, and the value of the information X can be given by:

$$Valor(X) = V(I(X;Y)). \qquad (6)$$

In summary, the value of data X depends on the decision Y through the information gain (or mutual information). For the value to be consistent, the function V(.) must be increasing and V(0)=0. In the following example, we will quantify the value of data X for the binary event Y that indicates the occurrence (Y=1) or non-occurrence (Y=0) of a claim on a car insurance policy over a one-year validity period, as per the bivariate analysis shown in Table II and others available on Kaggle [19] for the variables presented in Table III.

TABLE III. APPLYING ENTROPY DEFINITION, CONDITIONAL ENTROPY AND MUTUAL INFORMATION

Variable	H(X)	H(Y)	H(Y|X)	I(X;Y)
MEDIA_DISTANCIA_PARCEIROS	2,3212402	0,0234799	0,0234632	1,68E-05
STD_VALOR_TRANSACAO	2,3329737	0,0234799	0,0234459	3,40E-05
QTD_NOITE	2,3204087	0,0234799	0,0234735	6,50E-06
QTD_MADRUGADA	2,2215431	0,0234799	0,0234482	3,17E-05
MAIOR_DISTANCIA_PARCEIROS	2,3212402	0,0234799	0,0234564	2,36E-05
QTD_TAGS	2,2686251	0,0234799	0,0234591	2,08E-05
QTD_TARDE	2,3215146	0,0234799	0,0234591	2,08E-05
MIN_VALOR_TRANSACAO	1,6651711	0,0234799	0,0234486	3,14E-05
MEAN_VALOR_TRANSACAO	2,3329732	0,0234799	0,0234725	7,40E-06
QTD_PARCEIROS	2,3135011	0,0234799	0,023463	1,70E-05

One way to assess the value of the data from the Table III is as follows. Since a value function V(.) is increasing, it is expected, for instance, that the variable X1=MIN_VALOR_TRANSACAO holds more value for the business in question (which involves the decision Y) than the variable X2=QTD_NOITE, as I(X1;Y)=3.14E-05 > I(X2;Y)=6.50E-06. However, note that X1 has a lower entropy than X2, with H(X1)=1.665 and H(X2)=2.32. This result demonstrates that having higher entropy, and therefore more information, is not always synonymous with a better decision, as what matters in reality for our method is the mutual information I(X;Y) that data X1 and X2 have concerning decision Y. Thus, information theory does not determine specific data values but helps quantify the uncertainty or information contained in probabilistic events. To determine the specific value of data, we need to consider other methods, depending on the context and the data involved.

C. *Game Theory in Action*

The same variables from the previous section were assessed based on their SHAP values [20]. SHAP (SHapley Additive exPlanations) values are a model interpretability technique that quantifies the relative impact of individual features on the output of a machine learning model [21]. Within the realm of binary classification tasks, SHAP values are employed to discern the proportional contribution of each feature to a model's prediction.

This method becomes particularly valuable when dealing with highly complex models, such as Random Forests or Deep Neural Networks [22], where the functional relationship between the input features and the model's output can be highly non-linear and intertwined. SHAP values, grounded in cooperative game theory, provide a solution to the problem of fairly distributing "rewards" (in this context, the contribution to the model's prediction) to each "player" (feature).

Such distribution considers both the marginal contribution of each feature as well as all potential synergistic interactions among them. For this analysis, the variables were used to train an XGBoost binary classification model aiming to predict the target class (Y=1). Subsequently, the average absolute SHAP value of each feature was calculated and represented in Table IV.

TABLE IV. SHARP VALUE FOR EACH VARIABLE USED IN THE MODEL

Variável	mean(SHAP value)
MEDIA_DISTANCIA_PARCEIROS	0.09022027
STD_VALOR_TRANSACAO	0.07902433
QTD_NOITE	0.05610221
QTD_MADRUGADA	0.05555103
MAIOR_DISTANCIA_PARCEIROS	0.04823394
QTD_TAGS	0.04670959
QTD_TARDE	0.04214765
MIN_VALOR_TRANSACAO	0.02885941
MEAN_VALOR_TRANSACAO	0.0260881
QTD_PARCEIROS	0.0217269

In Table IV, the variable STD_VALOR_TRANSACAO had an absolute average SHAP value of 0.079 and ranks second among the variables, despite the same variable having shown the highest mutual information with the target class in the table from the previous section (Information Theory). Such behavior is expected since different methods were used, and although they quantify the variable's impact on the decision in some way, they won't necessarily agree on the importance of the variables and, consequently, the value of the data. Even considering the different methods, it's interesting that there's an agreement between them that STD_VALOR_TRANSACAO is a significant variable. Fig. 2 shows the relevance of each variable in the final constructed model.

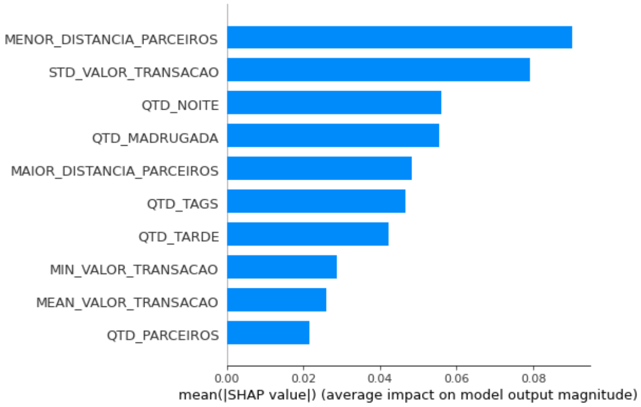


Figure 2. Variable importance according to Shapley Value.

D. *Theory Selection*

While we validated the potential use of both theories in the last two sessions, it became evident that we would face a much greater generalization challenge with Information Theory compared to Game Theory when applied to real-world problems.

In the case of the analysis presented based on the entropy formula and the mutual information of Information Theory, we simplified our problem to a bivariate analysis for each of the available attributes in a selected small sample. We would face a significant challenge in generalizing our method using this theory as the number of attributes or, equivalently, the number of databases to be combined and jointly evaluated increased.

When analyzing the Shapley Value (SHAP) calculation method from game theory, we discerned an objective very similar to what our method aims for. Analogously to SHAP, our method seeks to provide a solution to the problem of fairly distributing Monetary Value based on the contribution of each of the databases used for decision-making. To reach this conclusion, the financial gains obtained from decisions considering all possible combinations between the databases, both in isolation and combined, must be evaluated, precisely as proposed by the method based on game theory. By doing so, we will determine how the gains should be distributed. In our case, each "player" (resource) would be one of these available databases.

Such distribution considers both the marginal contribution of each resource and all possible synergistic interactions between them, and it's precisely for this reason that from this point forward, we'll be developing our method using Game Theory as the foundation.

V. THE METHOS BASED ON ROI

The data monetization method resulting from this study employs the concept of ROI added by a new database that becomes available for a binary decision-making when applied to a decision through a decision score (decision model) [23].

This is an anticipated scenario in the Big Data context as new data are continuously emerging.

ROI is a widely used financial indicator to assess the effectiveness and profitability of an investment. It compares the net gain achieved against the initial investment cost. Calculated as the difference between the net gain and the investment cost, divided by the investment cost, it's expressed as a percentage. ROI offers insights into an investment's efficiency, allowing organizations to evaluate the value produced relative to the invested resources [4].

We will illustrate the method using our experiment related to the decision-making process of analising insurance policies, in the underwriting process of a given quote. Specifically, car insurance underwriting may use data and/or an individual predictive score to determine whether an individual will be granted insurance (underwriting) or will be denied access following the quote process. An individual might be rejected, for instance, if there's any indication of potential fraud or risky behavior the insurer does not want to price.

In this case, the policy price is not shown, and the individual is said to have been declined during underwriting. Typically, this process does not reject a significant portion of the quotes. Underwriting rules tend to decline between 1% to 10% of the quotes, and this can vary among regions and profiles.

This process is beneficial for estimating data value since the operational outcome with an underwriting rule can be gauged on a historical policy base (backtest). This outcome is calculated based on the issued premium (in currency), which corresponds to the amount the insured pays to the insurer to access the insurance, and the observed indemnity during the policy period, which is the amount the insured received in the event of a claim. The insurer's operational gain equates to the difference between the premiums received from the insured (net of brokerage commission) and the compensations paid out resulting from occurred claims (indemnities).

$$Result_0 = \sum_{i=1}^{N} \quad Premium(i)\left(1 - \%Commission(i)\right) - Indemnity(i), \quad (7)$$

Where:

- $Result_0$ is the insurer operational Gain.
- $Premium(i)$ is the amount payed by insured i.
- $\%Commission(i)$ is the percentage commission payed by the insurer to the brokerage of the insured i.
- $Indemnity(i)$ is compensations paid out by the insurer resulting from occurred claims of insured i policy

A good underwriting rule can reject some of these policies that, in general, would have resulted in a loss. In this way, a new operational gain can be calculated without them, considering only the policies that have a risk score above a given cutoff point. This rule can be written as follows:

$$Result_1 = \sum_{i=1}^{N} \quad \theta(Score(i) - cutoff\_point) \, [Premium(i)(1 - \%Commission(i)) - Indemnity(i)] \, , \quad (8)$$

Where:

- θ(x) is the Heaviside step function [24].
- $Score(i)$ is is insurer i risk score.
- $cutoff\_point$ is minimum score for acceptance.
- The remaining itens are the same in (7).

Finally, the gain with an underwriting rule can be written as the difference between the results above (with and without the rule).

From this point on, we will build the artificial intelligence solutions (models) that will transform the raw data into the scores that will be used in the decision-making for the different scenarios of our experiment.

The models will transform the data into different scores according to the databases used (DB1, DB2, or both) and with the problem being addressed (target objective of the modeling). Table V summarizes all the models that were built in this research. At the end of each of the experiments, we apply (8), and the results are presented in section VIII.

VI. EXPERIMENTS

We conducted controlled experiments applied to real databases in a laboratory setting. The initial application assesses the ability to highlight the gain in understanding the risk of a new insurance quote when considering all the different scenarios including both available databases (DB1 and DB2) and how this gain can be measured in terms of ROI. We have used all the available data in our experiments, i.e., all the 540 thousands records (Fig. 1) where used for the modeling phase.

The experiments were repeated for different strategic decision-making, within the same domain (risk decision), allowing us to learn from the process and develop a generalizable method by the end of the study.

A. Model Construction (Risk Scores)

After the data was collected and processed, we moved on to the modeling phase, also known as knowledge discovery [25]. At this juncture, the data mining process begins. Algorithms will automatically sift through the data, trying to create the best possible representation of this data through scores. We split the dataset into two samples: one used for model training, with 75% of the available data, and the other for testing the model, where we assessed the performance of the built solutions, with the remaining 25%. For training, older insurance policies were used, while the newer ones were set aside for model testing.

With the datasets prepared, we could conduct experiments combining all possible scenarios. We had two databases (DB1 and DB2), two possible targets/objectives (claims and theft/robbery), and we also selected two different approaches to combine the resulting models: Stacking and linear combination.

All these possibilities resulted in eight different experiments as presented in Table V below.

TABLE V. 8 (EITGH) EXPERIMENTS

#Experiment	Description
1	DB2 x theft/robbery
2	DB2 x claims
3	DB1 x theft/robbery
4	DB1 x claims
5	Stacking x theft/robbery
6	Stacking x claims
7	Combinação Linear x theft/robbery
8	Combinação Linear x claims

B. The Choosen Technics

We chose two Machine Learning (ML) techniques to be applied in the model building for our project: Multilayer Perceptron (MLP) and eXtreme Gradient Boosting (XGBoost). The former was selected as it's a more traditional and widely used technique with a known performance track record for various problems. XGBoost has gained prominence in recent years for outperforming various ML algorithms [26]. For each of these techniques, different hyperparameter combinations were tested (For MLP: number of layers, number of neurons in each layer, etc. For XGBoost: depth, learning rate, lambda, etc.).

Each constructed model resulted in a risk scoring system (risk score) with the score indicating the risk level of that policy (claims or theft/robbery). A lower score indicates higher risk, while a higher score signifies lesser risk. After each model was built, to standardize our analyses, we divided the scored population into a uniform distribution with 10 bands (10 deciles), with the first decile (decile 1) representing the top 10% at highest risk, and the last decile (decile 10) representing the 10% at lowest risk.

We consider the model construction as a preliminary and necessary stage for our method and therefore will not delve deep into the details of each of the eight experiments. We will only detail the results of the third experiment here since it has the highest KS [27] among all (see summarized results in Table II).

C. Models Results

In the chart shown in Fig. 3 below, the average percentage of Theft/Robbery is depicted by the dashed line (0.45%). The chart displays the results of the model application to the test set of the third experiment conducted. The results for the other experiments can be found at the following Kaggle link [19]. The frequency of theft or robbery is represented by the red curve. We observed that the percentage of theft or robbery in the first score band (decile 1) is 283% (1.69% in the band) higher than the overall theft or robbery frequency of the base. In the last band (decile 10), this percentage is 91% lower (0.04% in the band).

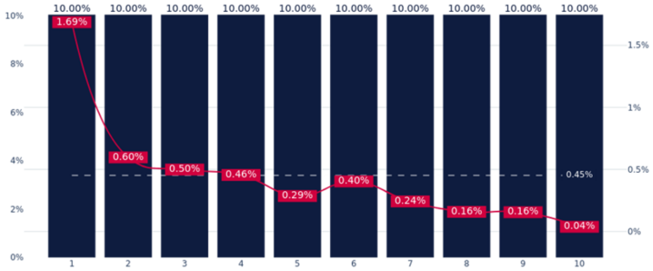


Figure 3. Histogram and risk curve by decile for the experiment 3.

The KS for this model was 33.4%, meaning we cannot assume the null hypothesis [28], and the model can distinguish the policies that will result in theft or robery from those that will not. We present in Table VI the results for the 8 (eight) experiments conducted:

TABLE VI. SUMMARY OF TECHNICAL RESULTS OF SOLUTIONS CONSIDERING THE ISOLATED AND COMBINED DATABATES

Model Database	Theft / Robery	Claims
DB2	28	4.2
DB1	33.4	5
Stacking Comb	33	5.6
Linear Comb	33	6.5

D. Experiment Conclusion

When evaluated individually, the solutions developed using the original database have a technically superior performance compared to those developed with the new database provided for both defined objectives. This was expected since the original database was developed and adjusted for this application by the partner company. The results from combining the databases were vastly different for the different objectives.

For the goal of predicting theft or robbery, none of the combinations achieved a technical result superior to the solution developed exclusively with the original database. We will investigate further to determine whether it's still possible to compute an added economic value to the addition of these new data using the proposed ROI method.

For the goal of predicting claims, both combinations achieved technically superior results to solutions developed with each of the databases exclusively. We will further investigate if this superior technical result can also be estimated in terms of ROI.

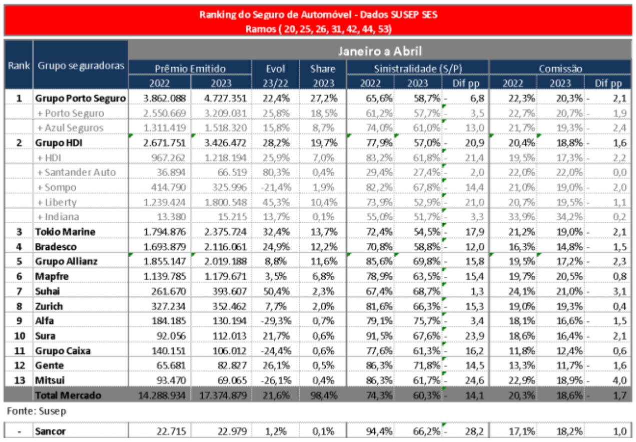
VII. RESULTS

To reach a conclusion about our method, we applied (8) to the 8 (eight) solutions developed during the experimental phase, considering a real database from a major Brazilian insurance company.

We are using a sample with R$ 100 million (one hundred million reais) in issued premium in the original portfolio (before the underwriting rule) and a 60.3% claims rate. We also considered an 18.2% commission. All these parameters were extracted from Table VII, which presents the comparison disclosed by SUSEP (Superintendency of Private Insurance) concerning the operational performance among all car insurance companies for the year 2023.

We used the parameters related to the month of April from this table in (8). Although the report provides data for each insurance company individually, we refrained from using the parameters of the insurance company under study due to confidentiality reasons.

TABLE VII. PUBLISHED DATA ABOUT OPERACIONAL RESULTS OF INSURANCE COMPANIES



Ranking do Seguro de Automóvel - Dados SUSEP SES
Ramos (20, 25, 26, 31, 42, 44, 53)

Rank	Grupo seguradoras	Janeiro a Abril									
		Prêmio Emitido		Evol	Share	Sinistralidade (S/P)			Comissão		
		2022	2023	23/22	2023	2022	2023	Dif pp	2022	2023	Dif pp
1	**Grupo Porto Seguro**	3.862.088	4.727.351	22,4%	27,2%	65,6%	58,7%	- 6,8	22,3%	20,3%	- 2,1
	+ Porto Seguro	2.550.669	3.209.031	25,8%	18,5%	61,2%	57,7%	- 3,5	22,7%	20,7%	- 1,9
	+ Azul Seguros	1.311.419	1.518.320	15,8%	8,7%	74,0%	61,0%	- 13,0	21,7%	19,3%	- 2,4
2	**Grupo HDI**	2.671.751	3.426.472	28,2%	19,7%	77,9%	57,0%	- 20,9	20,4%	18,8%	- 1,6
	+ HDI	967.262	1.218.194	25,9%	7,0%	83,2%	61,8%	- 21,4	19,5%	17,3%	- 2,2
	+ Santander Auto	36.894	66.519	80,3%	0,4%	29,4%	27,4%	- 2,0	22,0%	22,0%	0,0
	+ Sompo	414.790	325.996	-21,4%	1,9%	82,2%	67,8%	- 14,4	21,0%	19,0%	- 2,0
	+ Liberty	1.239.424	1.800.548	45,3%	10,4%	73,9%	52,9%	- 21,0	20,7%	19,5%	- 1,1
	+ Indiana	13.380	15.215	13,7%	0,1%	55,0%	51,7%	- 3,3	33,9%	34,2%	0,2
3	**Tokio Marine**	1.794.876	2.375.724	32,4%	13,7%	72,4%	54,5%	- 17,9	21,2%	19,0%	- 2,1
4	**Bradesco**	1.693.879	2.116.061	24,9%	12,2%	70,8%	58,8%	- 12,0	16,3%	14,8%	- 1,5
5	**Grupo Allianz**	1.855.147	2.019.188	8,8%	11,6%	85,6%	69,8%	- 15,8	19,5%	17,2%	- 2,3
6	**Mapfre**	1.139.785	1.179.671	3,5%	6,8%	78,9%	63,5%	- 15,4	19,7%	20,5%	0,8
7	**Suhai**	261.670	393.607	50,4%	2,3%	67,4%	68,7%	1,3	24,1%	21,0%	- 3,1
8	**Zurich**	327.234	352.462	7,7%	2,0%	81,6%	66,3%	- 15,3	19,0%	19,3%	0,4
9	**Alfa**	184.185	130.194	-29,3%	0,7%	79,1%	75,7%	- 3,4	18,1%	16,6%	- 1,5
10	**Sura**	92.056	112.013	21,7%	0,6%	91,5%	67,6%	- 23,9	18,6%	16,4%	- 2,1
11	**Grupo Caixa**	140.151	106.012	-24,4%	0,6%	77,6%	61,3%	- 16,2	11,8%	12,4%	0,6
12	**Gente**	65.681	82.827	26,1%	0,5%	86,3%	71,8%	- 14,5	13,3%	11,7%	- 1,6
13	**Mitsui**	93.470	69.065	-26,1%	0,4%	86,3%	61,7%	- 24,6	22,9%	18,9%	- 4,0
	Total Mercado	14.288.934	17.374.879	21,6%	98,4%	74,3%	60,3%	- 14,1	20,3%	18,6%	- 1,7

Fonte: Susep

-	**Sancor**	22.715	22.979	1,2%	0,1%	94,4%	66,2%	- 28,2	17,1%	18,2%	1,0

The two available databases (DB1 and DB2) on individuals were used to create separate underwriting rules. Subsequently, two new rules were created from the combination of the two databases using the different combination methods as explained in the section on Experiments. The operational gains from the four situations, extracted through backtesting, are listed in tables VIII and IX below. For this, an optimization of the cut-off point was considered, which would result in a refusal of 10% of the policies applying (8).

A. *Calculating ROI for Each Scenario*

To arrive at these results, we applied (8) for calculating the financial return for all the scenarios simulated in our experiment according to Lifts (applying an approval cut-off point at the 10th percentile) and their respective gains (ROI) presented in Tables VIII and IX.

TABLE VIII. RESULTS FOR THE THEFT AND ROBERY TARGET.

Modelo	Lift	Gain
DB2 isolated	170%	R$ 5.137.858,00
DB1 isolated	283%	R$ 10.711.628,20
Stacking Comb.	293%	R$ 11.204.882,20
Linear Comb.	296%	R$ 11.352.858,40

TABLE IX. RESULTS FOR THE CLAIM TARGET.

Model	Lift	Gain
DB2 isolated	19%	-R$ 2.310.277,40
DB1 isolated	40%	-R$ 1.274.444,00
Stacking Comb	33%	-R$ 1.619.721,80
Linear Comb	50%	-R$ 781.190,00

From the calculation of the result for all possible combinations, we can apply the Shapley Value to isolate the contributions of each of the databases to the final combined result.

B. *Calculating the Data Value*

Finally, by applying the values shown in Tables VIII and IX and using the SHAP formula, we can come to an exact conclusion about the value of each data for each of the selected problems and for each of the combinations made.

Just to recap, the formula for the data value we want to solve is presented below (as detailed in section II.A:

$$Value\ of\ data = V(data, decision)$$

As defined in our method, we will use the SHAP value to determine how much each of the Databases used contributes to the final value of the coalition (combination of the two databases). Recalling the Shapley Value formula that will be used in this calculation is presented below:

$$vi = \sum_{(S\subseteq N\setminus\{i\})} \frac{|S|!\,(|N|-|s|-1)!}{|N|!} * (v(S\cup\{i\}) - v(S)),$$

Where:

- N is a players set.
- is a coalition that does not include player i.
- v(S) is the coalition value S.
- v(S∪{i}) is the coalition that includes player i.

From now on we present the computation of the value of the both datbases calculated according to our method for each of the databases in each of the combinations carried out in our experiments.

1) *Value of DB1 and DB2 for Theft and Robery Using Stacking*

Table X shows all possible coalitions and their respective ROIs for the Theft and Robery problem using Stacking as a technique to combine DB1 and DB2 in their coalition.

TABLE X. RESULTS FOR THE THEFT TARGET WITH STACKING.

Coalisões	ROI
C({Ø})	R$ 0,00
C({DB1})	R$ 10.711.628,20
C({DB2})	R$ 5.137.858,00
C({DB1,DB2})	R$ 11.204.882,20

We can interpret the Result as follows: The coalition (combination) of the two databases yields a Result greater than any other coalition that has only one of the two databases.

By applying the SHAP formula, we arrive at the following division of Gains between DB1 and DB2:

TABLE XI. DATA VALUE OF DB1 AND DB2 FOR THEFT USING STACKING.

Player	Result (SHAP)
DB1	R$ 8.389.326,20
DB2	R$ 2.815.556,00

As we can see from Table XI, although the coalition with the two databases yields a higher result for the decision being made, the result for each of the databases is less than the ROI of each one individually. This outcome is expected because the new data added for decision-making is unlikely to be entirely independent of the data previously available. These characteristics and discussions regarding the results are also applicable to the other experiments, so we will present the results in a summarized form for the remaining experiments.

2) Value of DB1 and DB2 for Theft and Robery Using Linear Combination

TABLE XII. RESULTS FOR THE THEFT TARGET WITH LINEAR COMBINATION

Coalision	ROI
C({Ø})	R$ 0,00
C({DB1})	R$ 10.711.628,20
C({DB2})	R$ 5.137.858,00
C({DB1,DB2})	R$ 11.352.858,40

Applying SHAP formula, it possible to get to these gains between DB1 and DB2:

TABLE XIII. DATA VALUE FROM DB1 AND DB2, USING LINEAR COMBINATION

Player	Result (SHAP)
DB1	R$ 8.463.314,30
DB2	R$ 2.889.544,10

3) Value of DB1 and DB2 for Claims Using Stacking.

TABLE XIV. RESULTS FOR THE CLAIM WITH STACKING

Coalision	ROI
C({Ø})	R$ 0,00
C({DB1})	-R$ 1.274.444,00
C({DB2})	-R$ 2.310.277,40
C({DB1,DB2})	-R$ 1.619.721,80

Applying SHAP formula it is possible to get the following division between DB1 e DB2:

TABLE XV. VALUE OF DATA DB1 AND DB2 FOR CLAIMS USING STACKING.

Player	Result(SHAP)
DB1	-R$ 291.944,20
DB2	-R$ 1.327.777,60

4) Value of DB1 and DB2 for Claims Using Linear Combination.

TABLE XVI. RESULTS FOR CLAIMS USING LINEAR COMBINATION.

Coalision	ROI
C({Ø})	R$ 0,00
C({DB1})	-R$ 1.274.444,00
C({DB2})	-R$ 2.310.277,40
C({DB1,DB2})	-R$ 781.190,00

Applying the sharp method it was possible to calculate the following gaing between bases DB1 and DB2:

TABLE XVII. DATA VALUE OF DB1 AND DB2 FOR CLAIMS USING LINEAR COMBINATION.

Player	Result (SHAP)
DB1	R$ 127.321,70
DB2	-R$ 908.511,70

From the calculated data values from DB1 and DB2 for all the different combined experiments (Tables XI, XIII, XV and XVII) it became clear that a database value would depend cleary on the problem and the decision we are puirsuiting. We found situations where both data brought value do the decision (Tables XI, XIII), cases where none of the databases where able to add value do the decision (Tables XV) and situation where only one of then brought additional value (Table XVII).

VIII. CONCLUSIONS AND FUTURE WORK

In conclusion, this study addresses a significant gap in the field of data monetization, proposing a method that provides a systematic and adaptable approach to assess the value of data across various databases and problem domains. While data monetization has garnered substantial attention, the absence of widely applicable methods in academic literature has hampered the realization of its full potential [2]. By integrating information theory, game theory, and the ROI metric, this research introduces a new method rooted in the Shapley Value concept from cooperative game theory.

The successful application of the method to the real-world decision-making problem of underwriting car insurance policies in the Brazilian market exemplifies its efficacy in precisely quantifying the financial contribution of individual datasets to binary decision outcomes. By isolating the added ROI generated by each data set, this approach offers a comprehensive perspective on data's value in decision-making processes. Notably, the versatility of the proposed method extends to analogous scenarios featuring binary decisions with measurable financial implications.

Venturing into this emerging research domain, we anticipate this study will serve as a catalyst for the development of future methodologies. These methodologies might build on the foundations laid out in this work or introduce innovative frameworks further illuminating the multifaceted value of data in the dynamic landscape of data science in Big Data environments. As organizations continue to recognize data's strategic importance, the proposed method presents a promising avenue for maximizing the benefits derived from data monetization efforts.

Despite its contributions, the proposed method is not without limitations. Firstly, its application is confined to binary decision problems where financial gains and losses can be explicitly quantified. This constraint may hinder its direct applicability to scenarios with more complex decision structures or non-monetary objectives. Additionally, the method relies on the availability of accurate and reliable data to compute ROI, making it susceptible to inaccuracies stemming from data quality issues.

Moreover, while the Shapley Value provides a fair value attribution in cooperative games, its implementation may demand computational resources that could become burdensome for exceptionally large datasets or high-dimensional decision spaces [29].

Regarding future research directions, several paths merit exploration. A fundamental area is extending the proposed method to accommodate more complex decision structures, potentially involving multiple parties or sequential decisions. This could entail adapting concepts from cooperative game theory to capture such scenarios' dynamics.

Furthermore, as the data monetization field continues to evolve, exploring alternative value metrics beyond ROI could offer a more comprehensive understanding of data's value. This could involve incorporating qualitative factors, long-term strategic impact, or even societal implications.

Additional investigations into methods for dealing with noisy or incomplete data might enhance the proposed approach's robustness. Exploring machine learning techniques, data preprocessing, and statistical analysis could help mitigate data quality issues' impact.

Lastly, a broader application of the method across various sectors and contexts would provide empirical evidence of its versatility and limitations. Comparative studies involving different decision problems and data sets could shed light on the proposed approach's generalization and efficacy in diverse settings.

In conclusion, while the current method offers a valuable contribution to evaluating data's value in binary decision scenarios, it's imperative that future research overcome its limitations and broaden its scope to address the complexities of real-world decision-making environments. By embracing these challenges and pursuing innovative directions, researchers can propel the advancement of data monetization methodologies and pave the way for more informed, data-driven decisions in an increasingly data-rich world.

REFERENCES

[1] ORACLE, “What Is Big Data? Big Data Definition -,” Jul. 11, 2020. https://www.oracle.com/br/big-data/what-is-big-data.html (accessed Jul. 11, 2020).

[2] D. Monteiro, L. Monteiro, F. Ferraz, and S. Meira, “Big Data Monetization: Discoveries from a Systematic Literature Review,” Oct. 2020.

[3] A. McAfee and E. Brynjolfsson, “Big Data: The Management Revolution,” *Harv. Bus. Rev.*, 2012.

[4] R. S. Kaplan and D. P. Norton, “The Balanced Scorecard—Measures that Drive Performance,” *Harvard Business Review*, Jan. 01, 1992. Accessed: Aug. 31, 2023. [Online]. Available: https://hbr.org/1992/01/the-balanced-scorecard-measures-that-drive-performance-2

[5] C. E. Shannon, “A Mathematical Theory of Communication,” p. 55, Jan. 1948.

[6] L. S. Shapley, “A value for n-person games,” in *Contributions to the Theory of Games*, Princeton University Press, 1953, pp. 307–317.

[7] Neurotech SA, “Neurotech SA.” [Online]. Available: https://www.neurotech.com.br/

[8] D. S. Johnson, “The NP-completeness column: An ongoing guide,” *J. Algorithms*, vol. 11, no. 4, pp. 434–451, 1990.

[9] A. De Mauro, M. Greco, and M. Grimaldi, “A formal definition of Big Data based on its essential features,” *Libr. Rev.*, vol. 65, no. 3, pp. 122–135, 2016, doi: 10.1108/LR-06-2015-0061.

[10] K. Ruan, “Digital Assets as Economic Goods,” in *Digital Asset Valuation and Cyber Risk Management*, K. Ruan, Ed., Academic Press, 2019, pp. 1–28. doi: 10.1016/b978-0-12-812158-0.00001-6.

[11] Aristotle, *Nicomachean Ethics*. Publisher Not Specified, 350AD.

[12] A. Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations*. Publisher Not Specified, 1776.

[13] K. Marx, *Das Kapital*. Publisher Not Specified, 1867.

[14] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

[15] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. CRC Press, 1984.

[16] W. E. Buffett and L. A. Cunningham, *The Essays of Warren Buffett: Lessons for Corporate America*. Cunningham Group, 2013.

[17] F. Provost and T. Fawcett, *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O’Reilly Media, 2013.

[18] R. Journal, “What is RFID?” 2023. [Online]. Available: https://www.rfidjournal.com/what-is-rfid

[19] “Data Monetization - Auto Isurance data.” https://www.kaggle.com/datasets/domingosmonteiro/auto-insure-data (accessed Sep. 05, 2023).

[20] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[21] J. Li, B. Shao, J. Xu, H. Li, and Q. Wang, “A big data based product ranking solution,” in *2016 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, Jul. 2016, pp. 190–194. doi: 10.1109/SOLI.2016.7551685.

[22] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] S. Lohiya, *Decision Trees for Decision Making*. Springer, 2018.

[24] E. Kreyszig, “Advanced Engineering Mathematics, 10th Edition | Wiley,” *Wiley.com*, 2010. https://www.wiley.com/en-us/Advanced+Engineering+Mathematics%2C+10th+Edition-p-9781119455929 (accessed Aug. 31, 2023).

[25] G. Piatetsky-Shapiro, *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991.

[26] Q. Tang, G. Xia, X. Zhang, and F. Long, “A Customer Churn Prediction Model Based on XGBoost and MLP,” in *2020 International Conference on Computer Engineering and Application (ICCEA)*, Guangzhou, China: IEEE, Mar. 2020, pp. 608–612. doi: 10.1109/ICCEA50009.2020.00133.

[27] J. Berkson, “A Note on the Kolmogorov-Smirnov Test,” *J. Am. Stat. Assoc.*, vol. 40, no. 230, pp. 269–272, 1945.

[28] D. S. Moore, G. P. McCabe, and B. A. Craig, *Introduction to the Practice of Statistics*. W. H. Freeman, 2017.

[29] W. S. Jewell and C. H. Owen, *Cooperative Game Theory and Applications*. Oxford University Press, 1999.