Conference:

The Eighteenth International Conference on Software Engineering Advances (ICSEA 2023)

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

Domingos Monteiro
dsmpm@cesar.school

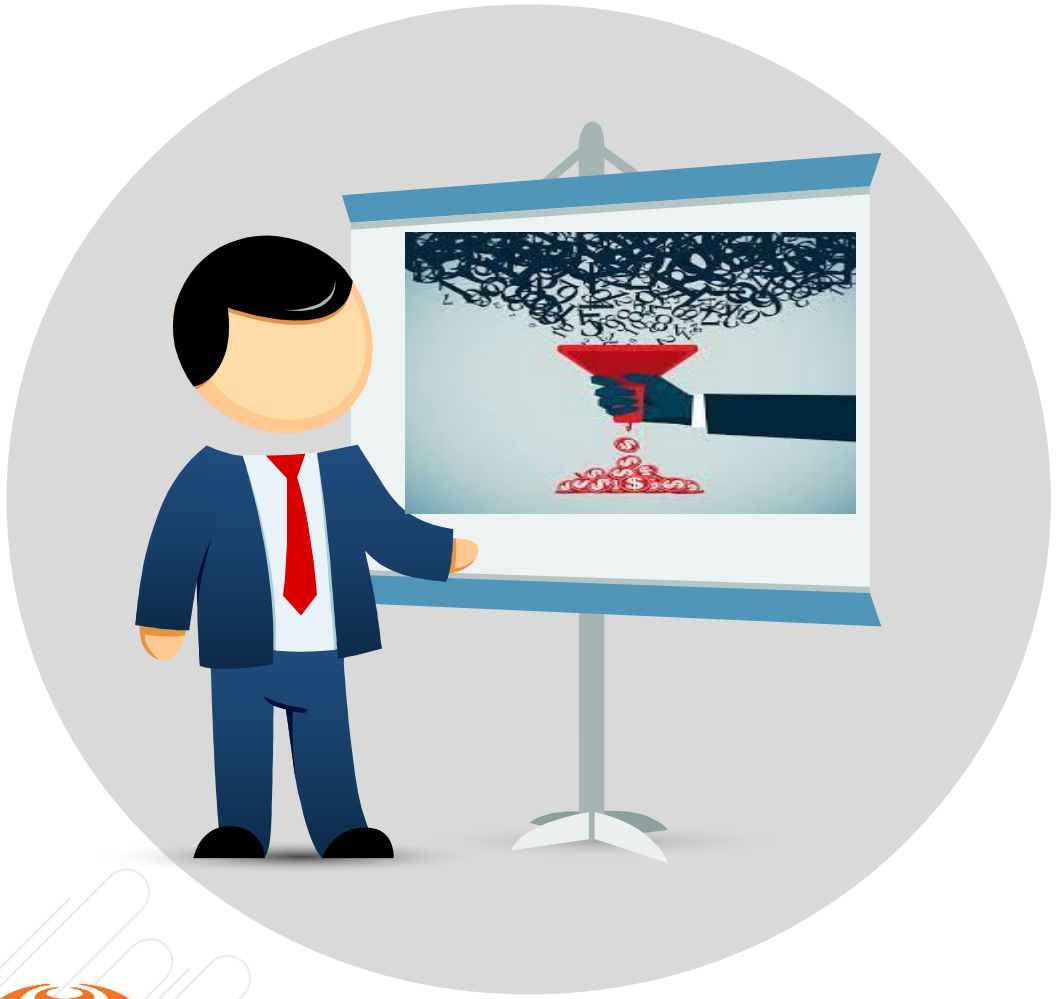
C.E.S.A.R school

About the Author

Domingos holds a Bachelor's degree in Computer Science from the Federal University of Pernambuco (UFPE) and earned a Master's degree in Artificial Intelligence from UFPE. He is an accomplished entrepreneur with a strong background in executive education, having completed an EPGC program focused on Executive Education at Stanford University Graduate School of Business. Presently, he is pursuing a Ph.D. at Cesar School.

Domingos boasts a wealth of experience in the field of Computer Science, specializing in Big Data and Artificial Intelligence. His professional journey has involved collaborating with diverse industries in Brazil, assisting them in the strategic utilization of data and intelligence to enhance predictability for future impact of their decisions.

Summary



1. Introduction & Research Question
2. Data: a Valuable Digital Asset
3. Available data set
4. Validating the theories
5. The method based on ROI
6. Experiments
7. Results
8. Conclusions and Future Work
9. References

C.e.S.A.R school

Introduction & Research Question

“Given the Lack of established methods to determine data's intrinsic value in Big Data [1], Would new information enhance decision predictability and efficiency? Can we measure the financial impact of it?”

- **Focus**: Flexible method for data valuation in binary decision-making with risk.
- **Gap**: Limited research on Data Value, especially in financial contexts.
- **Aim**: Improve predictability and efficiency by leveraging new data sources.
- **Methodology**:
  - Built on ROI concept for evolving Big Data.
  - Utilized Shapley Value from cooperative game theory.
- **Validation**:
  - Applied method to car insurance underwriting in Brazil.
  - Isolated financial value added by each database.
- **Scalability**:
  - Replicable across scenarios with similar binary decisions.
  - Applicable to decisions with concrete or measurable financial outcomes.

C.E.S.A.R school

Data: a Valuable Digital Asset

"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"[2].**

- Digital assets represent a new category capable of delivering better decisions, increased performance, competitive advantages, and can even be sold directly.
- Rationale for Binary Decisions:
  - Aligned with partner company operational focus.
  - Well-researched by the academic community, also valid to multiple classifications and regressions.

- Data Value Definition*:*

$$\textit{Value of data} = V(\textit{data, decision}). \quad (1)$$

  - Data's value depends on its ability to enhance decision-making
  - Rational agents purchase data if ROI is positive, i.e., value exceeds price.
  - Quantifying data's value is challenging.
- Further exploration of methods using experimental data.

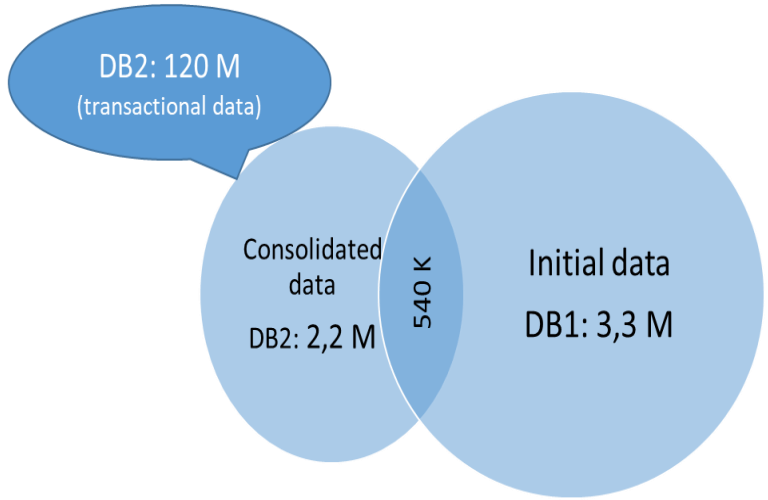
C.e.S.A.R school

AVAILABLE DATA SET

“Neurotech SA, a leading data and analytics provider for the Brazilian market, provided a sample of **120 million transactions,** involving 2.2 million TAGs, vehicles, and owners, for the project”

- **Avalilable Data:**
  - **Initial Database (DB1)**: Contained Neurotech's proprietary and public collected data of approximately 3.3 million vehicles and owners.
  - **New Database (DB2)**: Obtained from a third-party specializing in RFID data collection, representing new data for the study.

- **Objetive**: We want to determine the value of DB1 and DB2 for monetization, focusing on auto insurance.
  - DB2 was combined with DB1 to assess its impact on risk decision-making in policy underwriting.
  - We Merged DB2 with DB1, identifying 540,000 similar TAG holders in both databases, enriching policies.
  - Approximately 25% of TAG holders in DB2 also sought insurance from the partner company, enriching 16.5% of the original database.



C.E.S.A.R school

VALIDATING THE THEORIES (1)

"Our goal in this project was to apply a strategy that can estimate the value of data using machine learning. We approached information theory based on mutual information and game theory considering the Shapley Value."

- To test the theories we focused on identifying Theft or Robbery in a dataset of 328,565 annual car insurance policies.
  - 756 cases of Theft or Robbery (0.23%) and 327,809 cases without.
  - Explanatory variables provided insight.
  - We employed information theory (mutual information) and game theory (Shapley Value) for valuation.
- **Variables Analyzed** (Table I):
  - Examples:

  - MEDIA_DISTANCIA_PARCEIROS: average distance to nearest point of interest
  - QTD_TAGS: number of tags registered.

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.

VALIDATING THE THEORIES (2)

Information Theory vs. Game Theory

- **Information Theory**:
  - Quantifies information in data using entropy and mutual information.
  - Measures uncertainty or information contained in probabilistic events.
  - Does not determine specific data values for decision-making.
  - Requires additional methods to assign data value (e.g., mutual information).

- **Game Theory (Shapley Value)**:
  - Utilizes SHAP values from cooperative game theory.
  - Quantifies the impact of individual features on a model's prediction.
  - Suitable for complex models like Random Forests or Deep Neural Networks.
  - Provides a fair distribution of feature contributions.
  - Aligns with our goal of fairly distributing value among databases for decision-making.

- We selected Game Theory as the foundation for further method development.

C.e.S.A.R school

THE METHOD BASED ON ROI (1)

"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."

- **ROI as Data Value Indicator**:
  - ROI (Return On Investment) measures the profitability of an investment by comparing net gain to initial costs, expressed as a percentage.
  - Offers insights into resource efficiency and value produced.
- **Application in Car Insurance Underwriting**:
  - Examining the decision-making process in car insurance underwriting.
  - Assessing whether an individual will be granted insurance following the quote process.
  - Operational gain calculated based on issued premium and indemnity.
  - Evaluating data value by assessing the impact on underwriting rules.

- Artificial intelligence models were built to transform raw data into scores for decision-making in various scenarios (see side Table)

#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

THE METHOD BASED ON ROI (2)

"an optimization of the cut-off point was considered, which would result in a refusal of 10% of the policies applying (3)"

- **Calculation Framework**:
  - Operational Gain (Result_0) measured based on historical policy data.

$$Result_0 = \sum_{i=1}^{N} Premium(i)(1 - \%Commission(i)) - Indemnity(i)\ , \quad (2)$$

  - Operational Gain with Underwriting Rule (Result_1) calculated, rejecting policies that would typically result in losses.

$$Result_1 = \sum_{i=1}^{N} \theta(Score(i) - cutoff\_point)\,[Premium(i)(1 - \%Commission(i)) - Indemnity(i)]\ , \quad (3)$$

  - Data value determined as the difference between Result_1 and Result_0.

EXPERIMENTS

"Risk Scores: Data mining process started after data collection and processing. We split the Data into training (75%) and testing (25%) sets. Models built for different scenarios, resulting in eight experiments.

- **Controlled Experiments**:
  - Conducted controlled experiments on real databases in a lab setting.
  - Focus on evaluating data value in different scenarios involving DB1 and DB2, measured by ROI.
- **Chosen Techniques**:
  - Employed two Machine Learning (ML) techniques: Multilayer Perceptron (MLP) and eXtreme Gradient Boosting (XGBoost).
  - Various hyperparameter combinations tested for each technique.
- **Models Results**:
  - Risk scoring system created for each model.
  - Policies divided into deciles.
  - Highest KS observed in Experiment 3 (33.4%).
  - Results summarized in side Table.

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

- **Experiment Conclusion**:
  - Models developed using the original database outperform those using the new database individually.
  - Combinations of databases yield varying results depending on the objectives.
  - Further investigation needed to assess economic value using ROI.

C.E.S.A.R school

RESULTS (1)

"Data value depends on the problem and decision context. Scenarios vary in how databases contribute to decision-making. Careful consideration needed for data usage in specific contexts."

- **ROI Evaluation**:
  - Applied Rsult1 formula to the 8 experimental solutions.
  - Used real data from a major Brazilian insurance company.
  - Parameters extracted from SUSEP data for 2023n considering:
    - $ 100 million (one hundred million reais) in issued premium (before the underwriting rule)
    - 60.3% claims rate
    - 18.2% commission.
- **Calculating ROI for Each Scenario**:
  - Calculated financial return for all simulated scenarios.
  - Presented results for theft/robbery and claims targets.
- **Data Value Assessment**:
  - Applied SHAP formula to determine data value for each database in different combinations.
  - Highlighted scenarios where both databases add value, none add value, or only one adds value.

C.E.S.A.R school

RESULTS (2)

"…new data added for decision-making is unlikely to be entirely independent of the data previously available"

- **Data Value for all possible scenarios:**

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

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

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

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

- Utilizing the same dataset with different decision-making approaches results in varying data valuations.
- Data valuation depends on specific datasets and the choice of decision models.
- Context plays a critical role in determining the value of a database.

- **This outcome aligns entirely with our original postulation (1).**

C.E.S.A.R school

CONCLUSION

"Our research dives deep into the less-charted waters of data monetization, crafting a novel methodology that equips organizations to EVALUATE DATA'S VALUE SYSTEMATICALLY AND FLEXIBLY across diverse datasets and situations."

- Proposed method bridges a gap in data monetization by systematically assessing data value using information theory, game theory, and ROI.
- Successfully applied to underwriting car insurance policies, demonstrating its quantification of individual dataset contributions.
- Versatile method extends to similar binary decision scenarios with financial implications.
- Future research should explore complex decision structures, alternative value metrics, and methods for handling noisy or incomplete data.
- Broader application across diverse sectors and contexts can enhance method's generalization and effectiveness.
- Continued research can advance data monetization, enabling more informed, data-driven decisions in a data-rich world.

C.e.S.A.R school

References

[1] D. Monteiro, L. Monteiro, F. Ferraz, e S. Meira, “Big Data Monetization: Discoveries from a Systematic Literature Review”, out. 2020, vol. 9. The Ninth International Conference on Data Analytics. October, 2020

[2] A. De Mauro, M. Greco, e M. Grimaldi, “A formal definition of Big Data based on its essential features”, Library Review, vol. 65, no 3, p. 122–135, 2016, doi: 10.1108/LR-06-2015-0061. M. Gasparaite, K. Naudziunaite, and S. Ragaisis, *Systematic literature review of devops models*, vol. 1266 CCIS. Springer International Publishing, 2020.

Thank You

C.e.S.A.R school