Industry 4.0: Predicting lead conversion opportunities with machine learning in small and medium sized enterprises

Luis Borges Gouveia and Oberdan Costa

Atlantic Business School / University Fernando Pessoa

Presenting author

Luis Borges Gouveia

lmbg@ufp.edu.pt, *http://homepage.ufp.pt/lmbg/*

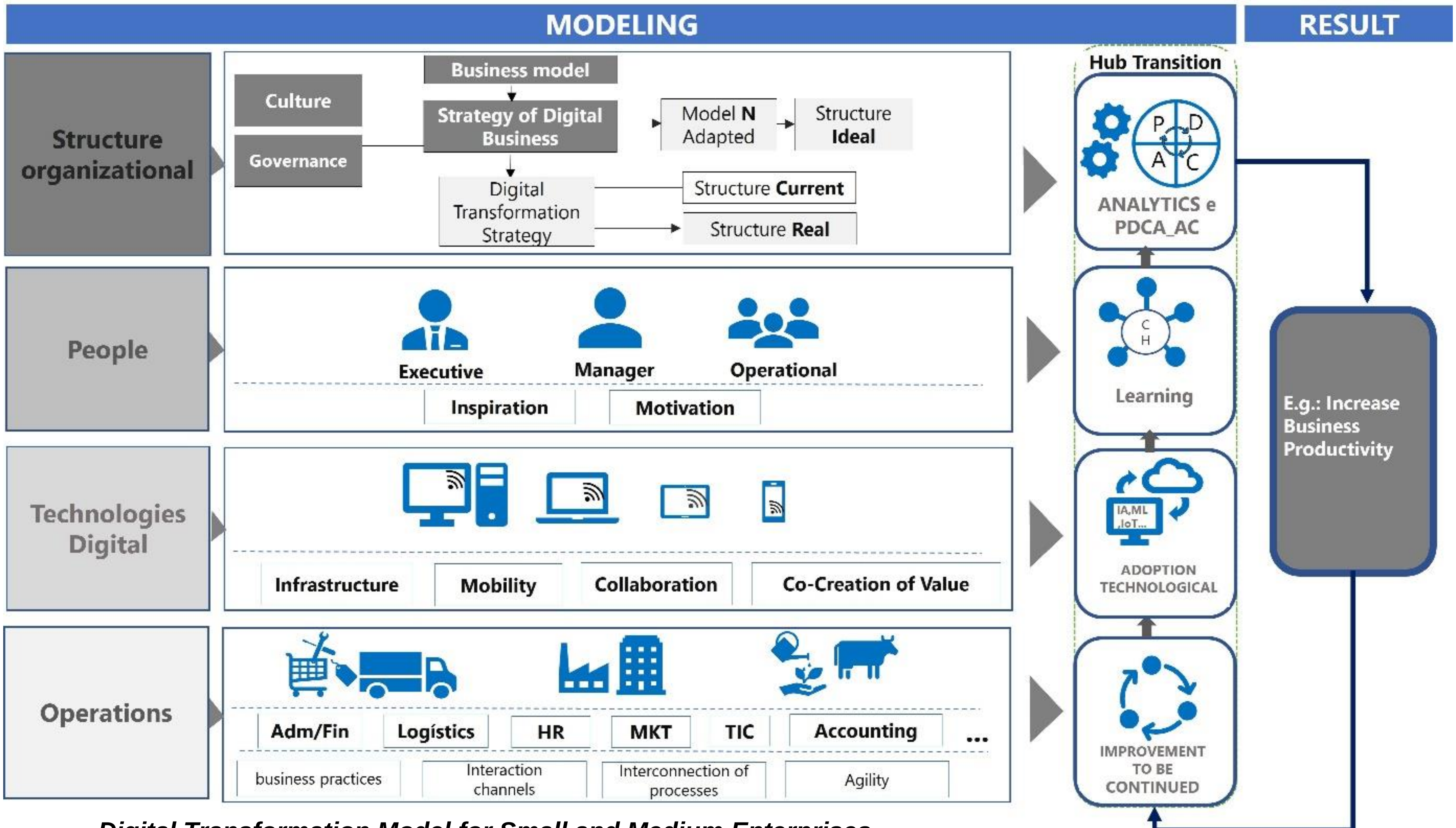
Movements of change

- **The context of after COVID-19 (that the current state of War reinforces)**
  - Accelerated processes of changes in the global economy (**time to respond**)
  - Changes in structures, business models and routines (**adaptation capacity**)
  - Small and Medium Enterprises (SMEs) faced challenges in finding paths for the journey of **digital transformation** and **adaptation to the industry 4.0**
  - Strong need to support, and **to integrate their transformations**
- **A movement of change in organizations, turn key to:**
  - **Triggering inflection points** for managers in organizations
  - **Acceleration of ways to address challenges**: multiple processes in human activity in structures, business models, and processes

Main goal and background

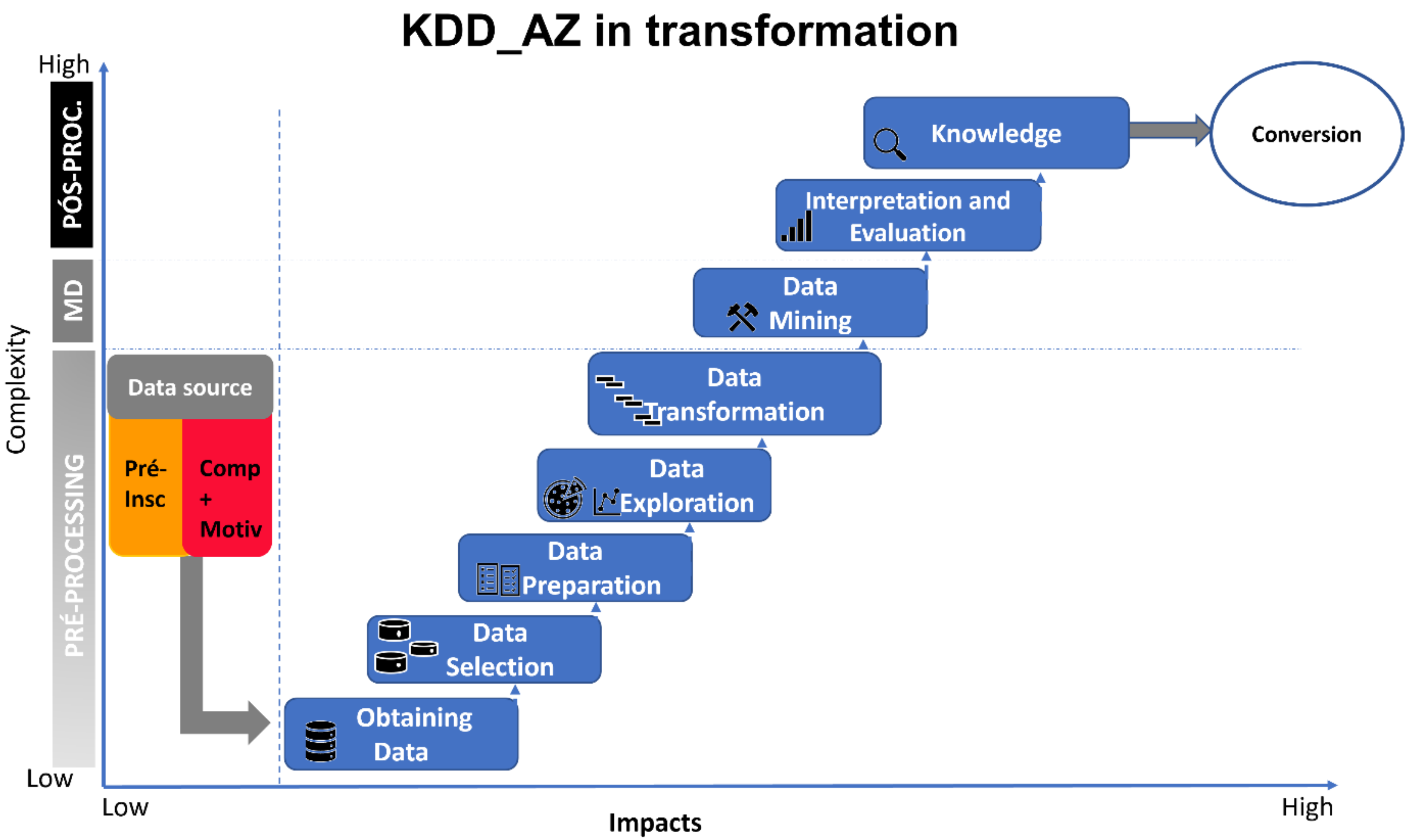
- **Work goal**
  - To predict the probability of converting leads using Machine Learning (ML) in order to improve the **process of enrollment opportunities** in small and medium-sized companies in the **education sector**
- **Background**
  - Digital Transformation Model for Small and Medium Enterprises (MTD_SMEs)
  - Specific approach in the technological resource supervised method of Machine Learning (ML)
  - Knowledge discovery model or KDD_AZ model in transformation

Digital Transformation Model for SMEs



Digital Transformation Model for Small and Medium Enterprises
Source : Costa, O. S. & Gouveia, L. B. (2021).

Knowledge discovery model or KDD_AZ model in transformation



Knowledge discovery model or KDD_AZ model in transformation
Source : Costa, O. S. & Gouveia, L. B. (2021).

Reported work

- Apply ML in the context of SMEs, with algorithms that perform different tasks
- Dataset from an **education hub of southern Brazilian university**, 2020/2021 enrollment period
  - **Dataset**: pre-registration data (name of the course), and data about the understanding of the composition and motivations at contact time.
- Procedure: a sequence of three stages of the KDD_AZ process:
  - (**Pre-processing**, Data mining (**Modeling**) and **Post-processing**)
- Use of ML techniques with simple and objective metrics to **predict the probability of closing the lead registration** as accurately as possible.
- **Development**: Python and tools available in the pandas and scikit-learn packages

Methodology used

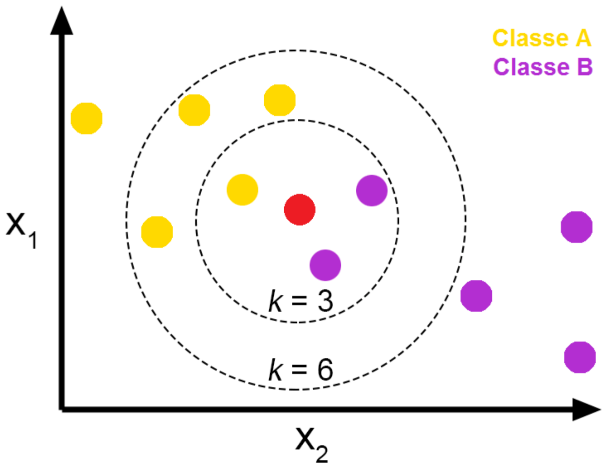
KDD_AZ process : Pre-processing

Data phases	Tasks
Obtaining	Imported from data from an education hub of a private university in southern Brazil, in the period 2020 and 2021, in **csv format**.
Selection	Selection of structured data consisting of **8 attributes / resources** and a sample size with **1596 leads**
Preparation	**Cleaning** (correcting/removing inconsistent data, deleting missing values or replacing them with **NA**, checking missing or incomplete data and identifying anomalies (**outliers**) and resource engineering (data integration and construction)
Exploration	Exploratory Data Analysis (**EDA**), which included the summarization of the data in a descriptive manner and the variation/dispersion of the data, frequency distribution and correlation analysis in a graphic way.
Transformation	**Normalization** and **Conversion/encoding** of data in appropriate ways to deliver the data mining step

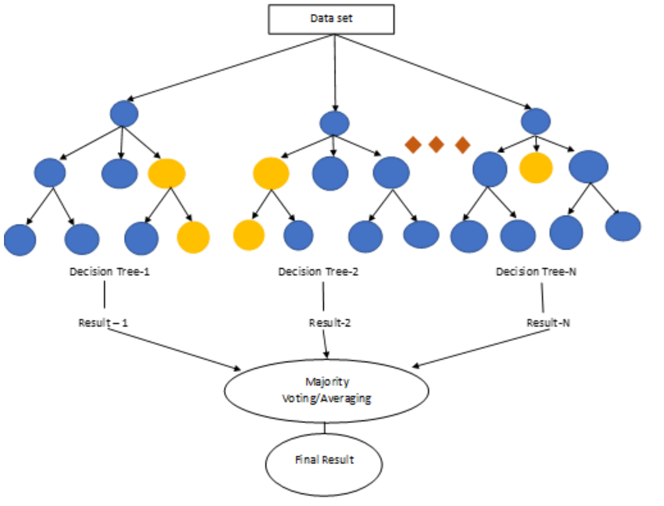
KDD_AZ process Step 1: Pre-processing
Source : Costa, O. S., & Gouveia, L. B. (2021).

Methodology used

KDD_AZ process : Data mining (Modeling)

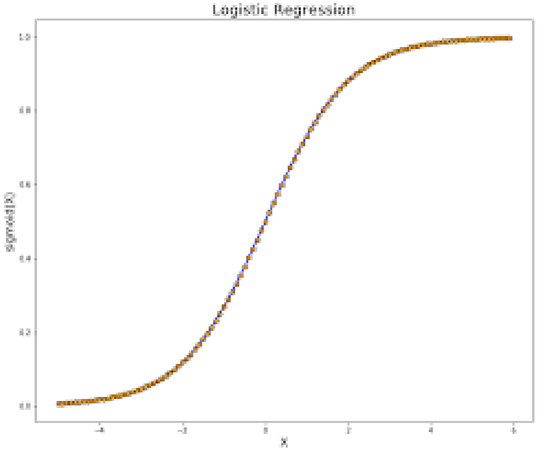


k-Nearest Neighbors (KNN)



Random Forest

Logistic Regression



KDD_AZ process Step 2: Data mining algorithm (Modeling)

Source: Elaborated by the authors

Methodology used

KDD_AZ process : Post-processing

Performance metrics

Performance Metric	Formula
Accuracy	(TP + TN) / (TP + TN + FP + FN)
Precision	TP/ (TP + FP)
Recall (Sensitivity)	TP / (TP + FN)
F1-Score	(2* Recall *Precision) / (Recall + Precision)

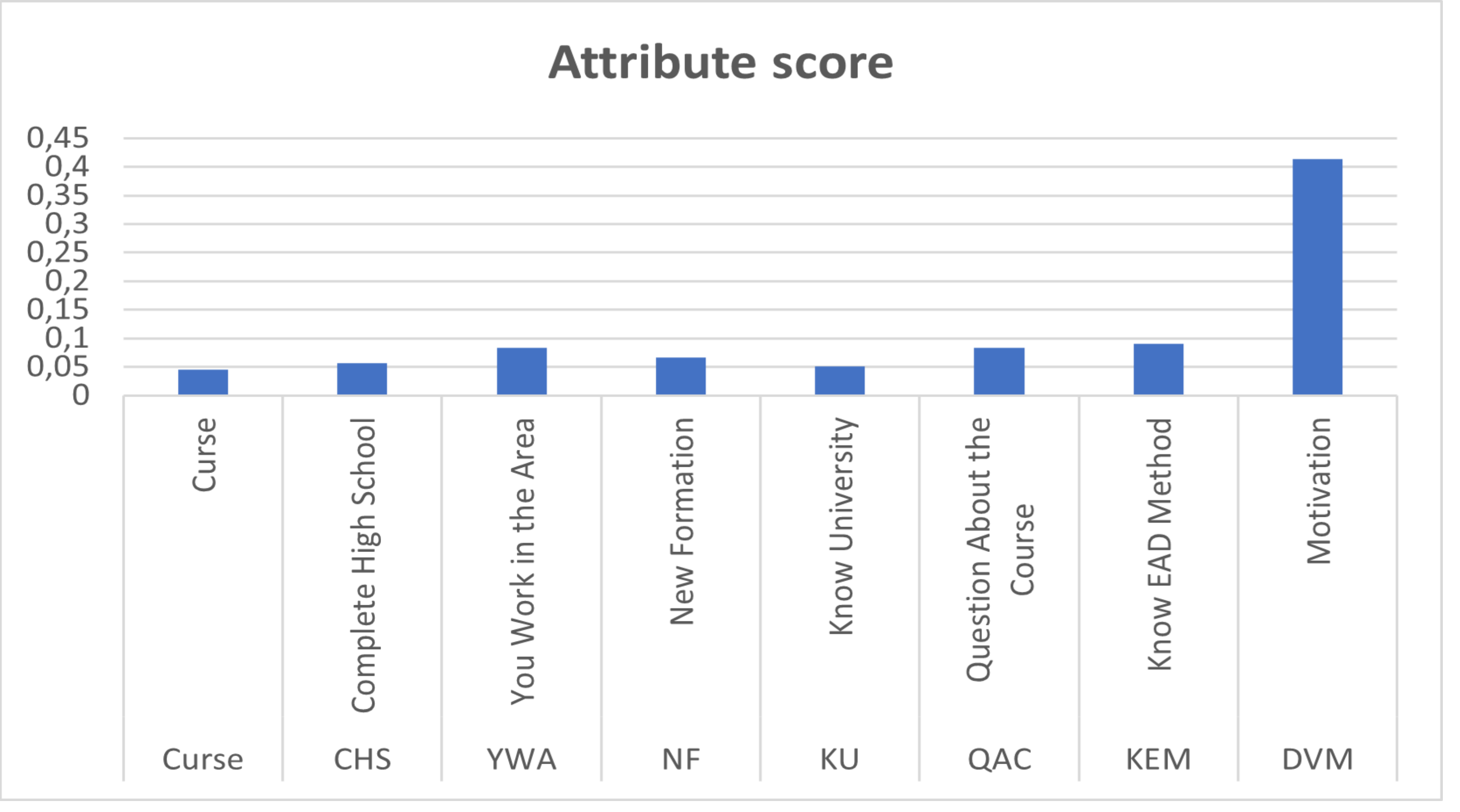
KDD_AZ process Step 3: Post-processing (Summary of performance metrics)
Source: Elaborated by the authors

RESULTS AND DISCUSSION: key issues

- Assessment of the significance of predictor attributes/variables for the **estimator**.

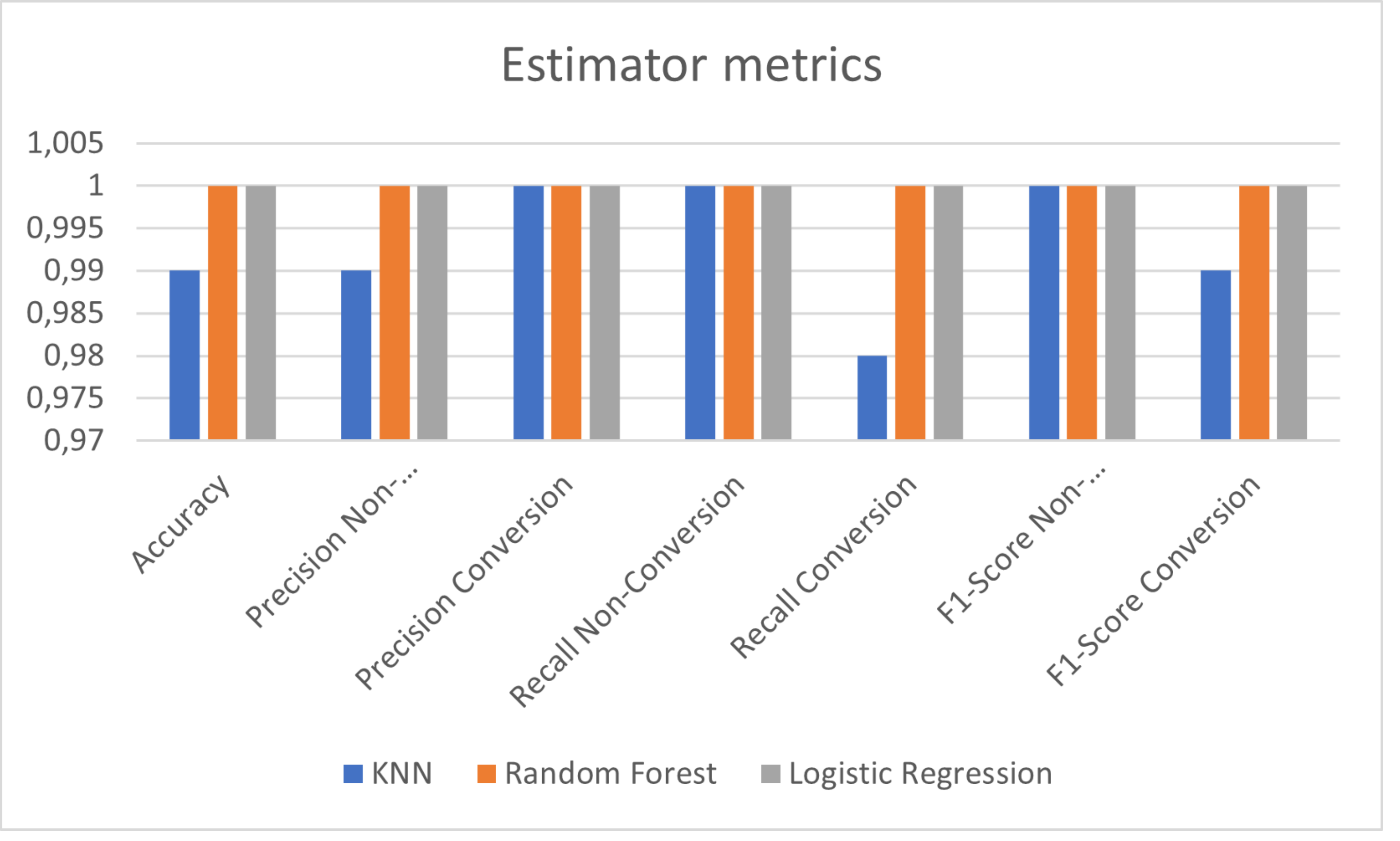
- **Evaluate the use** of different classification techniques in order to achieve the best possible result to predict the conversion probability of opportunities generated in lead capture

RESULTS AND DISCUSSION: key issues



Scores of attributes
Source: Elaborated by the authors

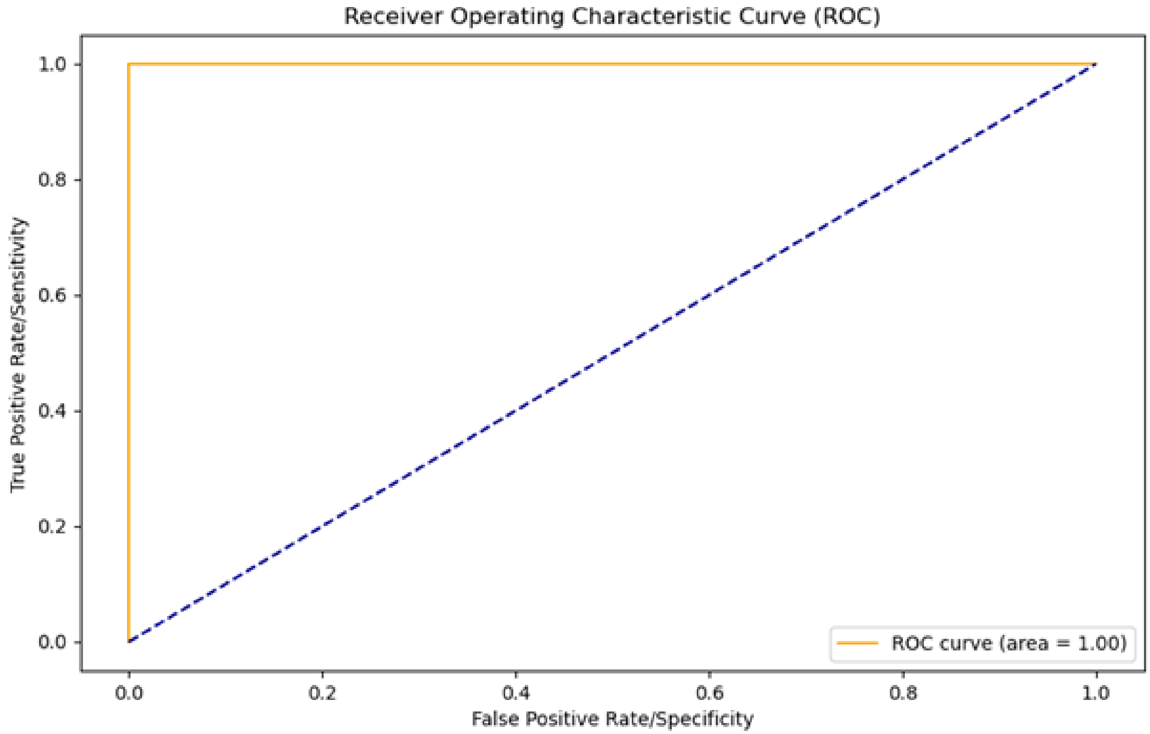
RESULTS AND DISCUSSION: key issues



Estimator performance evaluation metrics
Source: Elaborated by the authors

RESULTS AND DISCUSSION: key issues

Considering that the objective of analyzing the predictive power of a model is to ensure that it will detect as many true positives as possible, while minimizing false positives, we use the ROC and AUC curve tool to demonstrate the performance of the classification model, by means of ratio of the True Positive Rate (Sensitivity) and the False Positive Rate (1-Specificity), varying the threshold (cut-off point in the estimated probability).



high performance classifier
Source: Elaborated by the authors

RESULTS AND DISCUSSION: key issues

Now taking the RL estimator as the basis for the implementation operations, we executed the test data set, now parameterized for 16 examples and obtained the following results, as shown in Table

LEADS	
LEAD1=[1.38705442 -0.52917332],	Foreseen =[0.01096455 0.98903545]
LEAD2=[1.25761298 -0.39234234],	Foreseen =[0.02316462 0.97683538]
LEAD3=[-0.90816361 0.51411561],	Foreseen =[0.94106479 0.05893521]
LEAD4=[1.6327847 -0.15942508],	Foreseen =[0.04233817 0.95766183]
LEAD5=[1.03877541 0.03504071],	Foreseen =[0.17138649 0.82861351]
LEAD6=[1.03760658 -0.59207386],	Foreseen =[0.01224267 0.98775733]
LEAD7=[2.16199969 0.38161715],	Foreseen =[0.21653096 0.78346904]
LEAD8=[0.11076434 0.34126079],	Foreseen =[0.69963233 0.30036767]
LEAD9=[-0.77944662 0.60757554],	Foreseen =[0.95457127 0.04542873]
LEAD10=[0.91802345 0.57984128],	Foreseen =[0.73237132 0.26762868]
LEAD11=[-0.52820267 1.11515053],	Foreseen =[0.99357463 0.00642537]
LEAD12=[1.58148912 -0.08150395],	Foreseen =[0.06232312 0.93767688]
LEAD13=[0.31562106 1.01722554],	Foreseen =[0.97466543 0.02533457]
LEAD14=[0.61245334 0.98701273],	Foreseen =[0.96005853 0.03994147]
LEAD15=[0.03394077 0.19344227],	Foreseen =[0.5667005 0.4332995]
LEAD16=[0.84044667 0.5330707],	Foreseen =[0.7076786 0.2923214]

Parameterized RL estimator
Source: Elaborated by the authors

RESULTS AND DISCUSSION: key issues

When analyzing Table 4, we highlight 3 important points, they are:

Point 1: Leads number 1,2,4,5,6,7 and 12 have low conversion probability,

Point 2: Leads from numbers 3,8,9,10,11,13,14 and 16 have

Point 3: We can still identify situations in which we are not sure whether the lead is registered or not, as is the case for lead 15, where the probabilities of enrolling and not enrolling are very similar (56.67% not enrolling and 43, 32 % enrollment.

LEADS	
LEAD1=[1.38705442 -0.52917332],	Foreseen =[0.01096455 0.98903545]
LEAD2=[1.25761298 -0.39234234],	Foreseen =[0.02316462 0.97683538]
LEAD3=[-0.90816361 0.51411561],	Foreseen =[0.94106479 0.05893521]
LEAD4=[1.6327847 -0.15942508],	Foreseen =[0.04233817 0.95766183]
LEAD5=[1.03877541 0.03504071],	Foreseen =[0.17138649 0.82861351]
LEAD6=[1.03760658 -0.59207386],	Foreseen =[0.01224267 0.98775733]
LEAD7=[2.16199969 0.38161715],	Foreseen =[0.21653096 0.78346904]
LEAD8=[0.11076434 0.34126079],	Foreseen =[0.69963233 0.30036767]
LEAD9=[-0.77944662 0.60757554],	Foreseen =[0.95457127 0.04542873]
LEAD10=[0.91802345 0.57984128],	Foreseen =[0.73237132 0.26762868]
LEAD11=[-0.52820267 1.11515053],	Foreseen =[0.99357463 0.00642537]
LEAD12=[1.58148912 -0.08150395],	Foreseen =[0.06232312 0.93767688]
LEAD13=[0.31562106 1.01722554],	Foreseen =[0.97466543 0.02533457]
LEAD14=[0.61245334 0.98701273],	Foreseen =[0.96005853 0.03994147]
LEAD15=[0.03394077 0.19344227],	Foreseen =[0.5667005 0.4332995]
LEAD16=[0.84044667 0.5330707],	Foreseen =[0.7076786 0.2923214]

Parameterized RL estimator
Source: Elaborated by the authors

CONCLUSIONS

- To **support knowledge extraction** from the lead base (and increase the conversion rates)
  - Key to identify important pre-registration resources
- Issues highlighted in the results
  - Identification of the **most relevant attributes** for the correct classification of the conversion
    Results: Together, reach 87.43% of relevance
  - **Use of the different estimators** to find the best possible result of the precision of conversion forecast.
    Results: Logistic Regression allow for 100% correct classification of true positives and true negatives (maybe further research allow to evaluate overfitting)

CONCLUSIONS

- Main contribution
  - formation of a set of **significant variables to predict the probability of leads conversion**
    to support small and medium institutions in the area of education
- Benefits
  - It is expected that this will **reduce the time and effort of the conversion teams** and **help in the realignment of the prospection of qualified leads**
    to support the marketing area unit of the colleges

iSCSi 2022 official Sponsors

Platinum

Gold

Silver

Institutional