IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Development of Intelligent Robotic Process Automation: A Utility Case Study in Brazil

BRUNA VAJGEL[1], PEDRO L. P. CORRÊA[2,3], THAIS TOSSOLI[1], ROSA V. ENCINAS QUILLE[2,3], JOHN A. R. BEDOYA.[4], GUSTAVO MATHEUS DE ALMEIDA[5], LUCIA V. L. FILGUEIRAS[2], VANESSA R. S. DEMUNER[4], AND DENIS MOLLICA[4]

[1]EY, Praia de Botafogo 370 - 6 Andar, Botafogo, Rio de Janeiro, RJ 22250-040, Brazil (e-mails: Bruna.Vajgel@br.ey.com, Thais.Tossoli@br.ey.com)
[2]Escola Politécnica da Universidade de São Paulo, São Paulo, SP, 05508-010, Brazil (e-mail: pedro.correa@usp.br, lfilguei@usp.br)
[3]Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, São Paulo, SP, 03828-000, Brazil (e-mail: encinas@usp.br)
[4]EdP Brazil, Avenida Presidente Juscelino Kubitschek, 04543-011, São Paulo, São Paulo, Brazil (e-mails: john.bedoya@edpbr.com.br, vanessa.souza@edpbr.com.br, denis.mollica@edpbr.com.br)
[5]Departamento de Engenharia Química da UFMG, Universidade Federal de Minas Gerais, Pampulha, Belo Horizonte - MG, 31270-901, Brazil (e-mail: galmeida@deq.ufmg.br)

This work was supported by the Brazilian National Energy Power Agency (ANEEL)'s R&D program.

ABSTRACT Robotic Process Automation (RPA) refers to process automation applications of traditional Information Technologies based on robot software with the ability to capture and interpret the specific processes of organizations. Studies show that RPAs are able to reduce resources and optimize processes effectively in relation to customers. Some of these call center business processes deal with customers most likely to complain; therefore, a "Proactive Notification" robot was developed to classify these types of customers to be prioritized. This robot defines the creation of an RPA architecture for proactive notifications applied to an electric company in Brazil. The methodology used for the development of this project consisted of data management, predictive models, and peripheral components for sending SMS and making calls. It was tested against all customers in 40 cities (two states) and the model considers the historical basis of 3 years of occurrences to predict customers with a high probability of filing a complaint due to power failure. The results show that customers who were called for this type of problem did not call the call center again to complain, suggesting positive acceptance of the robot. In conclusion, the robot presented herein is capable of making proactive notifications with high precision to customers with the highest probability of complaints, predicting possible problems.

INDEX TERMS Proactive notification, RPA, robotic process automation, predictive models, disruptive technology, electric power sector, technological innovation, digital transformation.

I. INTRODUCTION

DUE to the Internet revolution, companies are increasingly using applications and systems to help provide better services to their customers [1], [2]. These services are usually operations (processes) that need to be automated with recent trend tools and technologies applied to robotic processes [3]–[7]. A new generation of automation systems has evolved and is called Robotic Process Automation (RPA), which offers faster, more accurate performance and doubles investment returns [8]–[10]. As the provider of mobile telecommunications in the United Kingdom, Telefónica O2 is one of the first companies that succeeded in automating its processes to deal with its 24 million customers. In 2015, they deployed 160 "robots", generating a return on investment in three years between 650 and 800%, improving response time ∼1500-fold (from days to just a few minutes), and reducing customer "chase" calls by more than 80%. This resulted in significant cost reductions making them exponentially more competitive in the mobile market [11].

Initially, first generation (G1) RPAs were developed to transform "Back-office" activities with service automation in order to save time and keep people for more intellectual activities. In the "G1" generation, the term RPA is synonymous with automation of service tasks that were previously performed by human beings [12]–[15]. In business processes, RPAs come as an automation solution using software (or

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3075693, IEEE Access

robots) configured to connect to ERP (Enterprise Resource Planning), CRM (Customer Relationship Management) systems, other systems through APIs and other standard integration methods. Over time, automation via RPAs has evolved towards second generation (G2) and third generation (G3) robots that are able to deliver greater value via the application of Artificial Intelligence, Internet of Things and Big Data Analytics, providing learning means and methods, in addition to analyzing certain business process contexts [16]–[19]. As automation generates value, it leaves the "Back-office" area and begins to be used in the companies' areas of operation. However, discussions related to risk control [20] and best practices for project management that allow business operations through RPAs have been studied [21]. These studies include research on choosing the right automation approach, selecting the right implementation provider, and redesigning processes to maximize the benefits and to minimize automation risks [22]–[24].

Serious research into RPA development is conducted by both Fortune 500 companies as well as new startups [25]; as are the cases of Shop Direct, Co-operative Banking Group, Fidelity Investments, RWEnpower, the NHS and O2 respond quickly to changes in business through agile operations [26], [27]. In addition, corporations are adopting an emerging technology RPA to streamline company operations and to reduce costs [28]–[33]. Along this line, we have developed an RPA to handle the large call volumes to an electrical utility company's call center due to temporary loss of service (e.g., resulting from a brownout or blackout).

Therefore, we here report the construction of the "Proactive Notification" robot. This robot is capable of monitoring the system responsible for mapping the power outage occurrences, estimating the duration for each occurrence, subsequently prioritizing and communicating. The communication is via SMS and telephone, with customers who have a high probability of filing a complaint with power utility companies. Proactive notification tracks the occurrences to ensure when they are closed and if it is necessary to rectify the estimation of power outage duration. In order to forecast the duration and predict the clients that must be contacted, both the areas responsible for the operations of restoring power supply and for communicating with clients, IOC (Integrated Operation Center) and Call Center, respectively, must be accessed. As such, additional functions have been added to the robot to access the relevant information in the internal systems of the Electricity Utility Company (EUC).

This paper presents the results of this study, organized as follows: Section II presents a review for Intelligent Robotic Process Automation; Section III presents the methodology; Section IV details our architecture of the solutions applied to intelligence automation; Section V details the statistical model for intelligent robot function and Section VI is dedicated to presenting the results and discussion about the methodology proposed.

II. INTELLIGENT ROBOTIC PROCESS AUTOMATION

One of the challenges in the era of digital transformation is the application of disruptive technologies on a large scale, the analysis of benefits and socio-economic and cultural impacts that may arise. One of these disruptive technologies is the Intelligent Robotic Process Automation [34], a subject that has been developing in artificial intelligence, digital technologies, software robots, and software development [35]. Agostine et al. [36] present four research challenges to include intelligence in current RPA technology. They explain that these can be automated with AI techniques, however, only a sample of RPA tools were analyzed, considering it as a first step towards intelligent solutions for RPA.

More recent work, by Syed et al. [37] present a review of the literature identifying contemporary issues and challenges in RPA. Through more than 100 papers, they identify their benefits, capabilities and challenges of RPAs:

- Benefits
  - Operational efficiency;
  - Quality of service;
  - Implementation and integration;
  - Risk management and compliance.
- Capabilities
  - Employee level capabilities;
  - Organization and process related capabilities;
  - Process transparency, standardisation and compliance;
  - Process intelligence for decision-making;
  - Flexibility, scalability and control.
- Challenges
  - Support for benefit accruement;
  - Comprehensive metrics for benefits;
  - Models for organizational readiness assessments;
  - Mechanisms for infrastructure assessments;
  - Models for organizational capabilities assessments;
  - Maximizing analytical capabilities;
  - Methodological support for adoption and implementation;
  - Socio-technical implications;
  - Techniques for task selection;
  - Techniques for managing scalability;
  - Others.

  Note that challenges may arise according to specific areas, such as accounting and auditing [38]. It is clear that for the realization of the points mentioned above, take into account the organizational readiness for RPA, capabilities of the RPA technology to be adopted, and implementation and delivery of an RPA solution. A limitation found in current RPAs is the mandatory requirement of structured data; with AI and complementary technologies we can use unstructured/semi-structured data when it comes to larger scales (big data).
  In 2020, Rizk et al. [39] presented a prototype of a Conversational Digital Assistant for Intelligent Process Automation based on natural language processing . Other

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3075693, IEEE Access

works were presented in relation to intelligent RPAs through Advance Process Analysis Model [40]. The difference in our work is that we present the development of an intelligent RPA for G1, G2 and G3+ robots, by which not only simple repetitive tasks are automated, but also processes already automated by robots of the first generations. The project involves techniques and methods of Analytics, Visual Analytics, as well as related market products, techniques and methods of metrics, two business processes, benchmarking and performance analysis, techniques and methods of robotization, software automation, artificial intelligence, decision-making systems, computational logics and work related to social transformation and human work forces.

III. METHODOLOGY

Figure 1 presents a proactive communication about the lack of power in a simple way.

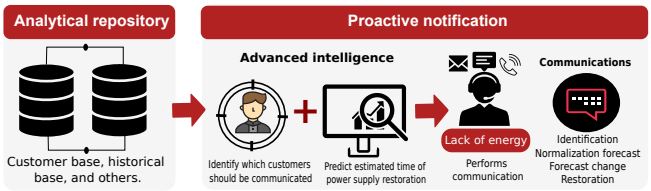

FIGURE 1. Proactive communication: prioritizing affected customers to be informed about the lack of electricity and its expected restoration

The aim of the solution is to decrease by 20% the number of calls for energy outage complaints received by the call center. The challenge is to build a solution which, within 15 minutes, is able to identify the occurrences in the company's system, classify their state (new, ongoing and closed), prioritize the clients that should be informed about the energy outage, estimate time the power supply restoration, send the communication and register the communication history in the internal company's systems. Moreover, operational teams should have autonomy to customize and monitor the solution, while not requiring IT (Information Technology) support. In order to achieve those goals, it is necessary to develop a solution with three main components:

- **Data Management Component:** data structure designed to identify all the occurrences of power outages, classify them (new, ongoing and closed occurrences), identify all the clients affected in each occurrence, collect the output of the predictive and classifying models, organize the list of messages and calls to be sent, create the data base to be consumed by the dashboards, and register the generated forecast and communications in the internal systems.
- **Intelligent Component:** predictive and classifying models to predict the time for restoring power and to classify the clients more prone to make a complaint once they are affected by a next energy outage episode, respectively.
- **Peripheral components:** those components are responsible for sending SMS and placing calls (Twilio and Messenger API), inserting the information communicated to the clients in the internal systems (Webservice and API) and monitoring the robot activity (PBi dashboard).

A. DATA MANAGEMENT

Organizations seek competitive advantages in order to stand out, improve internal processes, and adapt to the current needs of customers [41]. One of the main steps of any business is to have an adequate data management, which guarantees an assertive decision making process for meeting the objectives. However, many companies have a serious problem with the quality of their data. These problems start with the process of capturing customer-related data to be processed as it can be difficult to guarantee the reliability and quality of the data. Over time the data become a huge database because of daily operations, and these can lead to critical problems, when no attention is paid to data management. This leads to subsequent integration problems, examples of which may be the capture of incorrect customer data, duplicate information, and others. Effective data management requires a data strategy employing reliable methods to access, integrate, clean, govern, store, and prepare data for analysis [42], [43].

B. STATISTICAL MODEL

Statistical model is a simplified and mathematically formalized way of approaching and approximating reality (that is, the one that generates the data) and, optionally, making predictions based on the approximation. These statistical models are used as algorithms in Machine Learning, a method of data analysis in the field of Artificial Intelligence that automates the creation of analytical models. Through algorithms that learn from different databases and accumulated experiences, Machine Learning allows predicting and learning certain patterns and behaviors automatically, without human intervention [44]–[47] [48]. Recent research shows that the use of machine learning in RPA is capable of real-time detections, classifying them with greater precision and taking dynamic actions [48].

For the project, two statistical models were developed, namely the forecasting model and the prioritization model. The forecasting model consists in generating the prediction (or forecast) for the closure of each occurrence, predicted based on 3 years of occurrences. The prioritization model consists in inferring the probability of each customer filing a complaint with EUC and prioritizing the communication of customers with a greater propensity to complain.

C. PERIPHERAL COMPONENTS

Communication with the expected power supply restoration will only occur for customers with a high probability of filing a complaint from EUC. If the first forecast of the model is not feasible, the aforementioned customers will receive a new

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3075693, IEEE Access

message with a new estimated forecast for restoring power. The channels used by the solution are SMS and phone call.

IV. ARCHITECTURE OF THE SOLUTION APPLIED TO INTELLIGENT AUTOMATION

Throughout this chapter, the solution architecture built to support the data structure and its components will be presented. The information flow, tool, and tool's purpose will be presented for each solution. A macroscopic view of the architecture is presented in Figure 2. This solution was implemented using the cloud computing service, Microsoft Azure [38], [49], [50].

- Azure Data Factory
- Azure Data Lake
- Azure SQL DB
- PowerBI
- Power Apps

A. HOW THE PROCESS SOLUTION WORKS

The processing of the solution is divided into 3 stages:

1) Processing of historical load (cold data) for:
   a) Updating the technical customer base;
   b) Updating the equipment registration base;
   c) Updating historical events;
   d) Training models.
2) Continuous load processing (hot data) for:
   a) Application of the return forecast;
   b) Customer prioritization model for sending communications.
3) Auxiliary processing for:
   a) Backup of the solution;
   b) Visualization dashboards.

B. HISTORICAL LOAD DATA PROCESSING

The processing of the historical load data begins with the ingestion of both the customer's technical registration base and the registration of equipment data, which are both logged monthly. The tables involving individual occurrences are updated daily. This periodicity is necessary because the energy return forecast model requires the calculation of *features* based on historical variables.

1) Updating the customer and equipment registration base

This process aims to update the technical registration base of EUC customers and equipment on a monthly basis.

2) Update of occurrence history

This process aims to update the occurrence history and recalculate the historical *features*. The sequence of steps is required to complete this process is described here:

1) Import historical data:
   a) Persistence of historical bases such as **Occurrence**, **complaint** and, **Key occurrence** - joining the Occurrence tables, Events, and Interrupted Customers from the Data Lake to DBFS;
   b) Cleaning tables in which data from another state exists;
   c) Assignment of the **COMPANY** field to tables where this type of identification does not occur;
   d) Construction of *features* in the **Complaint** table.
2) Data cleaning:
   a) Formation of the historical dataset divided by **Occurrence**, **Client**, and **Complaint**;
   b) Construction of *features* for the historical dataset.
3) Construction of historical features: construction of historical features (calculation of moving averages) of the forecast model;
4) Backup daily data from the *Historical Output Model* and the *Historical Robot Consumption*.

3) Model training

The training process for these models includes the energy return forecasting model and the construction of historical features for the affected clients' model. A block diagram is provided in Figure 3, which shows the ingestion, execution, and processing steps used for training these models.

C. PARAMETERIZATION - POWER APPS

The main reason for developing of Power Apps is to provide security, control, and stability to our solution. The Power Apps are incorporated to ensure (1) that the process remains active consistently (11 hours a day on weekdays and 4 hours on Saturdays), which requires constant care from the IT team, (2) direct communication with the customer (messages and calls can be triggered to any of the captive customers who have a valid phone) , which requires attention from the Call Center team, and (3) the energy return predictions are reported to the customer, which requires monitoring by the IOC team. For this, a series of parameterizations can be made that have a direct impact on the function of the flow of the process, among them are:

- Robot shutdown (suspension of all communications);
- Choice of cities on which data should be reported;
- Minimum and maximum forecast reporting for the client;
- Changing the text of communication messages.

Note that these parameterizations can be performed independently for São Paulo and Espírito Santo and directly impact the flow of the Proactive Notification solution.

D. COMMUNICATION - MESSAGE SYSTEM

After applying the models to open events and affected clients, a list of customers is generated from those most likely to contact EUC in the event of a power outage. Two different approaches are used to communicate with these costumers. Calls are made to landlines and text messages are sent to cellular phones. The list of customers is sent to the APIs responsible for communications, following these parameters:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3075693, IEEE Access

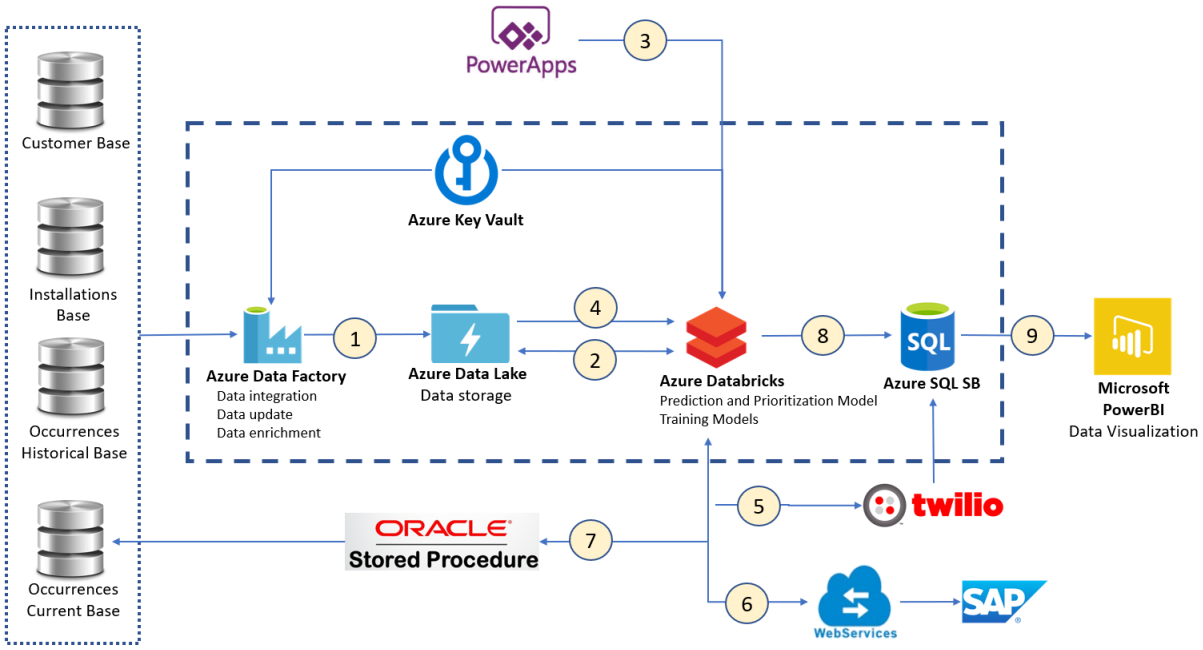

FIGURE 2. Macro-architecture of the solution applied to intelligent RPA

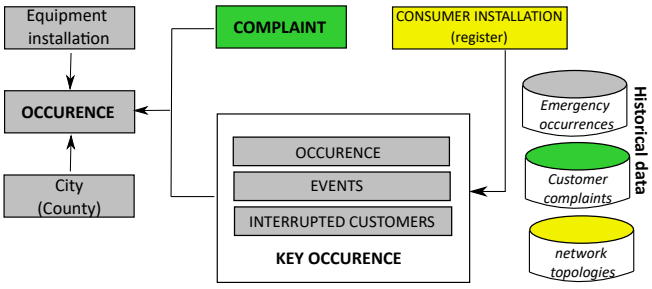

FIGURE 3. Block diagram responsible for the process

- For communications via Short Message Service (SMS) (EUC messenger) it is necessary to fill in the URL, company, user, customer, contact, and message fields.
- For communication via call (Twilio) it is necessary to fill in the fields of account_sid and auth_token (contained in this section), url, customer contact, and *Twilio* phone number.

E. RELATIONAL MODEL

In order to obtain all the necessary information for our solution, it is necessary relationships between the tables for constructing the *dataset*. In the specific case of Proactive Notification, the interaction produces both the models for predicting the energy return and prioritizing the affected customers.

F. SQL DATABASE

The Standard Query Language (SQL) is used for three different functions within the Proactive Notification solution:

- To serve as a beacon for all ingestions made within the Data Lake through Azure Data Factory;
- Serve as a backend of the PowerApps parameterization spreadsheet;
- Serve as Data Warehouse for Power BI views.

G. DASHBOARD - POWER BI

The dashboard was designed to provide the user with important information about the robot. This information includes the number of occurrences, affected customers identified, customer communications, evaluation of the forecasting and prioritization models, information on the flow performance, and flow error log.

V. STATISTICAL MODELS FOR INTELLIGENT ROBOT FUNCTIONS

The intelligent functions of this robot are composed of two predictive models:
1) Statistical model for predicting the energy return time; and
2) Machine learning based Customer prioritization model.
The following sections describe each model in detail.

A. MODEL FOR PREDICTING THE ENERGY RETURN TIME

The main objective of this Model is to estimate how long it will take for power to be reestablished, based on complaint history of the installation times for each outage occurrence.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3075693, IEEE Access

To study the modeled phenomenon, it was necessary to understand and match the stages of the outage occurrence cycle (opening of the occurrence, dispatch of the repair team, displacement and arrival of the team in the field, maintenance, end of service, registration of completion in the system) with the time required for each stage of the cycle. (example: average service time of the installation, average service time for the type of equipment, average service time of the plant etc.). Based on these historical variables, our statistical model estimates how long it will take for the energy to be reconnected, so that the robot can communicate this information to the affected customers and prioritize needs using the Customer Prioritization Model.

To develop this model we used various classification and regression techniques. The classification techniques used were Logistic Regression, Extra Trees, Random Forest, Ada Boost, and Gradient Boosting Tree. The regression techniques used were Random Forest, Negative Binomial, and Gamma.

The Model involves two predictions:
1) The first estimate is the light return time and, for events in which the first forecast expires,
2) The second estimate is the power restoration time. The training of the models responsible for the first and the second energy return time predictions are performed monthly. Once trained, the models are saved in the Data Lake to be used to guide the services provided by the robot. In the following sections, all stages of the predictive modeling are described in detail.

1) Basic modeling settings

In this first stage of modeling, basic parameters of the training, validation and test bases used in the model are configured. The first configuration is the years used to generate the training, validation, and test bases. Most of the variables that make up the Forecast Model are moving averages recording power failure events and installation times for each of these occurrences. The chronological order of the events matters. Therefore, the training base uses data that is defined as being three years behind the current date, the validation base two years behind, and the test base one year behind. For the first and second prediction models currently in use, the training, validation, and test bases are built with data from 2017, 2018 and 2019.

Two important filters that need to be defined for this robot are the operating hours and the times that customers will be contacted. These filters limit the training of the models to events that occurred between 8:00 am to 10:00 pm and lasted up to 12 hours. These parameters were defined by the IOC (EUC Integrated Operation Center) and can be modified at any time. These filters were created to prevent customers from being disturbed late at night and to prevent customers from worrying about wait times that exceed 12 hours. These decisions are motivated entirely by the professional experience employees at EUC.

The next model configuration block is related to the selection of variables. The lower correlation limit will be set to 0.01, which means that all the variables that have a correlation coefficient of less than 0.01 are excluded from the model.

2) Characteristics of the Training, Validation, and Test Bases

In this second stage of modeling, the bases for training, validation and testing are loaded. The training, validation and test basis were filtered by year to compose the training (2017), validation (2018) and test (2019) dataframes. The training, validation and test bases have 139,590, 188,068 and 183,339 rows and 103 columns. The columns considered for constructing of the model were 59, related to information about the number of the occurrence, code of the occurrence of the occurrence, call center, municipality of the occurrence, date of creation of the occurrence, time the occurrence was resolved in PowerOn, Total time for resolving the occurrence - from the opening to the end of the maintenance, Total time of the occurrence - from the opening to the end of the occurrence in the system, Forecast of the end of the occurrence given by the operator in the field, Type of equipment, Season of the year (summer, autumn, winter, spring) in which the occurrence occurred (model variable), Period of the day (morning, afternoon, night and dawn) in which the occurrence occurred, Zone (rural or urban) of the occurrence (model variable), Region of occurrence, Date when the team was sent to resolve the occurrence etc.

The last treatment performed on the training, validation and test bases are the robot's operating time and time limit filters to be informed. The first filter limits training, validation and testing of the model to events that were opened between 8:00 am and 10:00 pm, while the second filter restricts training, validation and testing of the model to occurrences with the resolution time equal to or less than 12 hours or 720 minutes. No sanitation is performed on the basis for missing information. Those occurrences that have missing information are excluded from the model.

3) Selection of variables

In this third stage of modeling, the selection of the most important variables for the model is performed. The first step is to analyze the Pearson correlation coefficients between the explanatory variables and the target variable of the model. Pearson's coefficient (ρ) measures the linear correlation, whether positive or negative, between two variables:

- $\rho = 1$: It means a perfect positive correlation between the two variables, that is, if one increases, the other also increases;
- $\rho = -1$: It means a perfect negative correlation between the two variables, that is, if one increases, the other always decreases;
- $\rho = 0$: It means that the two variables do not depend linearly on each other. However, there may be a non-linear dependency.

Table 1 shows all the variables with Pearson's coefficients of less than 0.01 are removed from the model. Thus, the variables selected for modeling are 30.

Variable	Description
WeekOccurrence	Week of the month (first, second, third or fourth) of the occurrence
FlagHoliday	Flag that informs you whether or not it is a holiday period
CnaeEssencial	Flag that says if CNAE is classified as essential affected by the occurrence
CnaeCritic	Flag that says if CNAE has been classified as critical affected by the occurrence
MeanDispInstall	Moving average of installation dispatch time
StdDispInstall	Standard deviation within the moving average of the installation dispatch time
MeanDispEquip	Moving average of equipment dispatch time
StdDispEquip	Standard deviation within the moving average of the dispatch time of the equipment
MeanDispCentral	Moving average of the dispatch time of the call center
StdDispCentral	Standard deviation within the moving average of the dispatch time of the call center
MeanArriInstall	Moving average of the installation arrival time
StdArriInstall	Standard deviation within the moving average of the installation arrival time
MeanArriEquip	Moving average time of arrival of the equipment
StdArriEquip	Standard deviation within the moving average of the arrival time of the equipment
MeanArriCentral	Moving average of the call center arrival time
StdArriCentral	Standard deviation within the moving average of the call center arrival time
MeanMaintInstall	Moving average of installation maintenance time
StdMaintEquip	Standard deviation within the moving average of the equipment maintenance time
StdMaintCentral	Standard deviation within the moving average of the call center maintenance time
MeanDisplaDisEquip	Moving average of the travel time and dispatch of the equipment
MeanDisplaInstall	Moving average of the displacement and dispatch time of the installation
MeanDisplaArriEquip	Moving average of travel time and equipment arrival
MeanDisplaDispCentral	Moving average of the travel and dispatch time of the call center
MeanDisplaArriInstall	Moving average of the time of travel and arrival of the installation
MeanDisplaArriCentral	Moving average of displacement and arrival of the call center
MeanDisplaMaintEquip	Moving average of travel time and equipment maintenance
MeanDisplaMaintCentral	Moving average of the travel and maintenance time of the call center
TotalOccurOpenCentral6h	total occurrences opened in the last six hours in the call center
TotalOccurOpenCentral3h	total occurrences opened in the last three hours in the call center
TotalOccurOpenCentral1h	total occurrences opened in the last hour at the call center

TABLE 1. Variables selected from Pearson's correlation coefficient.

4) Model Training

At this stage, the two training sessions of the model are described: first forecast and second forecast.

First forecast: At this fourth stage, the first training of the model is carried out, considering only the list of variables selected in the previous stage and standardized by the respective minimum and maximum of their distributions (Standard Scaler). The training uses the adjustment of generalized linear models of the python library called statsmodel, choosing the negative binomial distribution and the link to identity function with alpha = 0.1. After the first training, the variables with the twenty largest z-scores are selected.

With these twenty new variables selected, a second training of the model is performed, again using a negative binomial regression with the link equal to the identity function and the alpha equal to 0.1. Table 2 shows the main metrics of the best adjusted model; with MAPE (Mean Absolute Percentual Error), MAE (Mean Absolute Error) and R^2 respectively, for Training and Validating the 1st State and 2nd State.

	MAPE (%)	MAE (Minutes)	R^2
Training	3.2	94.9	0.04
Validation (Total)	52.4	98.5	0.05
Validation (1st State)	51.4	104.1	0.02
Validation (2nd State)	53.2	94.5	0.06

TABLE 2. Metrics of the best adjusted model.

This is the final version of the model for the first forecast. The model is saved to Data Lake. The first model will be consumed in production for the first forecast of the occurrences and the second will be a backup of this model. Hence, if there is any problem, it is possible to recover the older versions of the model.

Second forecast: The next step of the modeling is to separate in the validation sample those occurrences that were underestimated by the model of the first forecast. A quick analysis was performed to understand when the underestimated moment of occurrence has its having its first predictions expired. The result of the analysis of the underestimated forecasts is as follows:

- Number of forecasts that expired before dispatch time: 25,771 (48%);
- Number of forecasts that expired during the travel time: 8,933 (18%);
- Number of forecasts that expired between arrival and before the end of maintenance: 13,100 (24%);
- Number of forecasts that expired before the closing time: 2,869 (5%);
- Number of forecasts that expired after the closing time: 2,807 (5%).

It can be seen that the worst bottleneck of occurrences is the time it takes to dispatch the service teams, followed by the time to carry out maintenance.

Once the population of underestimated occurrences has been selected in the validation and test bases, a second training is performed using the underestimated validation base, again using a negative binomial regression with the link equal to the identity function and the alpha equal to 0.1.

Table 3 shows the metric of the second forecast model:

	MAPE (%)	MAE (Minutes)	R^2
Validation	26.3	91.8	0.11
Teste	27.1	91.7	0.09

TABLE 3. Metric of the second forecast model

Since it is important for the relationship with the client that the forecasts are not underestimated, an analysis was also performed to quantify how many occurrences are underestimated and how many are overestimated for each state:

- First State
  - Underestimated (Actual> Expected): 9850 (40%);
  - Overestimated (Actual $\leq$ Expected): 14625 (60%).
- Second State
  - Underestimated (Actual> Expected): 10661 (37%);
  - Overestimated (Actual $\leq$ Expected): 18339 (63%).

Comparing the real time at the end of the event with the first and second predictions in the validation sample, it can be seen that:

- The number of second forecasts greater than or equal to real time is 32,964 (62%) occurrences;
- The number of second forecasts greater than or equal to the first forecasts is 53,475 (100%) occurrences;
- The number of second forecasts greater than or equal to the first doubled forecasts is equal to 0 (0%) occurrence;

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3075693, IEEE Access

- The number of second forecasts greater than or equal to real time and less than the first doubled forecasts is 32,964 (62%) occurrences;
- The average ratio between the second forecast and the first doubled forecast is 79%;
- The average difference between the second forecast and the real time: 0.09 minutes;
- The average difference between the second forecast and the first forecast: 129 minutes.

Making the same comparison now for the test sample, it appears that both results are consistent, as shown in the numbers below:

- The number of second forecasts greater than or equal to real time is 32,650 (64%) occurrences;
- The number of second forecasts greater than or equal to the first forecasts is 50,865 (100%) occurrences;
- The number of second forecasts greater than or equal to the first doubled forecasts is equal to 0 (0%) occurrence;
- The number of second forecasts greater than or equal to real time and less than the first doubled forecasts is 32,650 (64%) occurrences;
- The average ratio between the second forecast and the first doubled forecast is 79%;
- The average difference between the second forecast and the real time: 5.6 minutes;
- The average difference between the second forecast and the first forecast: 130 minutes.

We can guarantee that the second forecast is never less than the first forecast and that the second forecast is not simply the first forecast multiplied by a factor of two. The second light return time forecast model is saved in the same way as the first forecast on Data Lake.

5) Evaluation of Results

Evaluating the model from the perspective of the EUC business, that is, considering that the correct predicted values are greater than or equal to the actual resolution time of the occurrences, we have the following numbers:

- First State 57,661 occurrences were evaluated in test, of which:
  - An average assertiveness of 59% is obtained, only considering the first forecast;
  - 34,036 forecasts are correct in the first forecast (1st forecast real time);
  - 23,625 forecasts are errors in the first forecast (1st forecast <real time);
  - An average assertiveness of 83% is obtained also considering the second forecast;
  - 48,071 forecasts are correct in the first or second forecast (1st forecast real time or 2nd forecast real time);
  - 14,035 wrong forecasts in the first forecast are converted into hits by the second forecast (1st forecast <real time and 2nd forecast real time).
- Second State 79,886 occurrences were evaluated in test, of which:
  - An average assertiveness of 63% is obtained, only considering the first forecast;
  - 50,023 forecasts are correct in the first forecast (1st forecast real time);
  - 29,863 forecasts are errors in the first forecast (1st forecast <real time);
  - An average assertiveness of 86% is obtained considering also the second forecast;
  - 3 68,398 forecasts are correct in the first or second forecast (1st forecast real time or 2nd forecast real time);
  - 18,375 wrong forecasts in the first forecast are converted into hits by the second forecast (1st forecast <real time and 2nd forecast real time).

Figure 4 dissects the figures presented above for the first state by time range foreseen for the first forecast.

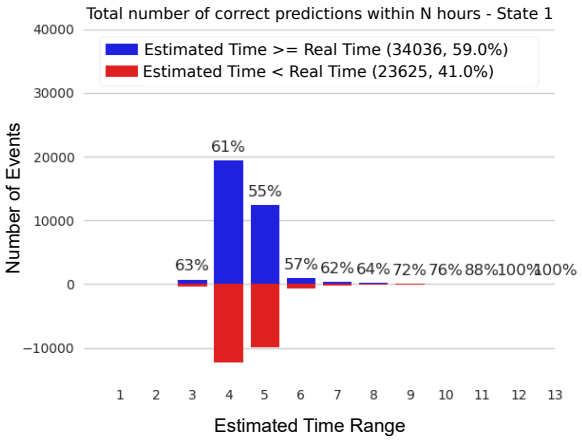

FIGURE 4. First forecast for the first State

The blue bars represent the number of occurrences correctly predicted in the first forecast (1st forecast $\geq$ Real Time) with their respective percentages per predicted time range, while the red bars represent the number of occurrences incorrectly predicted in the first forecast (1st forecast $<$ Real Time) with their respective percentages by estimated time range. It can be seen that the two bands with the highest number of predictions made are the 4 and 5 o'clock bands, with 61% and 55% of assertiveness, respectively.

When applying the second forecast only for the occurrences of the red bars and analyzing them by predicted time range, the result in Figure 5 is obtained.

In this graph, the blue bars represent the number of occurrences correctly predicted in the second forecast that are incorrectly predicted in the first (2nd Forecast $\geq$ Real Time and 1st Forecast $<$ Real Time) with their respective percentages by predicted time range, while the red bars represent the number of erroneously predicted occurrences also in the second forecast (2nd Forecast $<$ Real Time and 1st Forecast $<$ Real Time) with their respective percentages by predicted

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3075693, IEEE Access

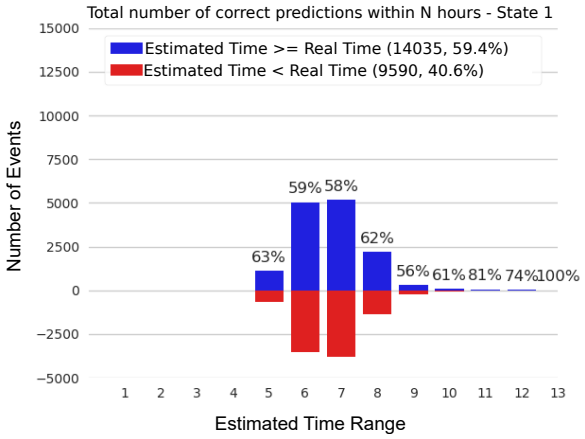

FIGURE 5. Second forecast for the first State

time range. With the second forecast, it is possible to reverse the error of 59.4% of the 41% of the predicted errors in the first revision, thus raising the assertiveness of the model to 83% in the first state, as shown in the graph as follows with the consolidated view of the correctly predicted occurrences in at least minus one of the forecasts (1st Forecast Real Time or 2nd Forecast $\leq$ Real Time).

6) Forecast application

In this section, the application of the models of the first and second forecast of the energy return time in the forecast of occurrences in operation will be explained. The application of the forecasts is performed the Return during each 15 minute cycle of the robot. After loading the models and the forecast bases and applying the necessary transformations, the models are applied to generate the forecasts. The distribution of the expected value for the analyzed events is estimated and compared with the distributions observed during the training phase. Very different distributions from that observed during training may indicate a change in the condition of the problem between training and operation, which may be an indication of the need for retraining before the monthly period. Subsequently, these occurrences can be re-analyzed when the real service time is known.

B. PRIORITIZATION MODEL FOR CLIENTS TO BE COMMUNICATED

The main objective of the Customer Priority Model to be Communicated is to estimate the propensity of affected customers to call the call center in a power outage event and, based on the call center's complaints history, to prioritize customers most likely to complain to be communicated. In order to study the phenomenon modeled, it was necessary to understand and to map the factors most influence a customer so that he/she decides to call the call center (ex: weekday, school holidays and holidays, if he is a customer who suffers greatly with power failure events etc.).

Based on these historical variables, the model estimates the probability of a customer calling the call center, when affected by a power failure event, so that the robot prioritizes the communication of affected customers with a greater propensity to complain. As will be explained in the following sections, the Priority Model for Customers to be Communicated is a Random Forest model.

The Random Forest model is a versatile supervised machine learning algorithm that can be used for both regression and classification. The algorithm builds a set of decision trees based on random samples of the data; from each tree obtained, it makes a prediction of the target variable and chooses the final prediction as being the most frequent. It is considered highly accurate and robust because it includes a large set of decision trees in the process and is not at risk of over fitting problems, since it uses the average of predictions; It can also be applied as a variable selection method, selecting the characteristics chosen as the most important for the classifying model. The main disadvantage of this model is the computational cost and the time to make the predictions, since the model needs to generate multiple decision trees for each prediction; also for this reason, the model can be difficult to interpret if compared to just a single tree decision-making.

For developing of the model, the training of the model, responsible for prioritizing the communication with customers affected by a power failure event that is more likely to make a complaint, is carried out monthly. Once trained, the model is saved in the Data Lake to be consumed in production by the robot. In the following sections, all the modeling steps are described in detail.

1) Basic modeling settings

In this first step of the modeling, general parameters of the algorithm are configured. The first configuration is to list the features available in the training and test bases, and the columns with qualitative data. It is necessary to apply a treatment to the data to be used in the model. Next, the parameters of five different classifier models used during the development phase are configured; however, only the Random Forest Classifier model is active, as this model presented the best results. Other important settings for the model are the definition of the target prediction variable and the number of subsets into which the training base will be partitioned to perform the cross-validation process (qty = 5). The main component analysis parameters (PCA) are also configured. This technique is used to provide a visualization in lower dimensions of the same data set, transforming the variables available into a new ordered set of orthogonal variables, known as main components, being that the variation present in the main components decreases from the first to the last. The technique is applied in order to reduce the dimensionality of a data set with many correlated variables, while the variation of the data set is conserved.

2) Characteristics of the Training, Validation and Test Bases

At this second stage of modeling, the bases for training, validation and testing are loaded. For the construction of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3075693, IEEE Access

model, 48 columns are used, which are information similar to the previous model (Prediction model). The choice of these columns was previously treated; however, in order to prevent unwanted data from being passed to the model, events with dates older than 2017 are filtered. To reduce the number of customers to be used in the training base, a sampling technique was applied, query was set up to bring a Simple Random Sample of 500,000 customers. The technique is based on the randomness of the sample and ensures that each individual in the population has the same chance of being included in the sample.

3) Model Training

At this step, the categorical variables are converted into dummy columns, which can be used by the Random Forest model. For the machine learning models, it is common to divide data into training and testing. To evaluate the results from the model, the technique of cross-validation of data, or K-fold Cross-Validation, was applied. The method initially consists in randomly dividing the data into K mutually exclusive subsets, previously defined as 5 subsets. The model will be trained K times and at each iteration of the process, a different subset will be adopted as the test subset, while the other subsets are the training data. When using this technique to evaluate the performance of the model, we use the entire dataset for both training and testing, in different process iterations. The method permits to evaluate the model's generalization capacity from a data set. The results can be aggregated in a single average model that fits the entire data set. After training the classifier model, we went through the entire forest and collected all the information specified in the previous section. This information allows us to describe which thresholds dominate the separations –see Figure 6 and Figure 7.

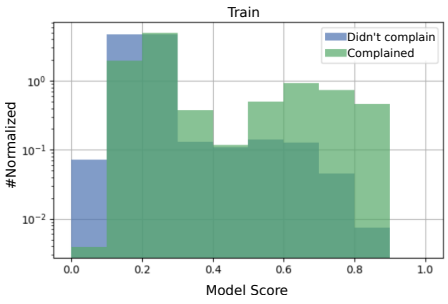

FIGURE 6. Train

In the chart, we see that there are two distributions. Green corresponds to the nodes where customers complained and blue corresponds to customers who did not complain. These weighted distributions for the model score indicate that whenever the models score resource is used to decide whether there was a greater possibility of complaints, the description is dominant for higher model scores, which is illustrated by the difference box between the distributions spikes for limits greater than 0.5.

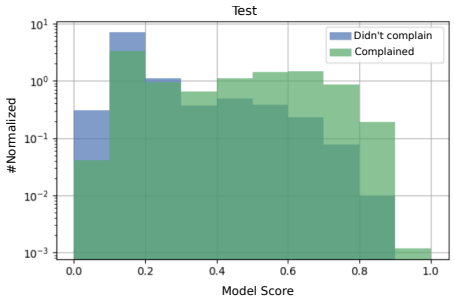

FIGURE 7. Test

4) Model evaluation

Some metrics are essentially defined for binary classification tasks, such as f1_score and ROC curve (roc_auc_score). When extending a binary metric for problems of multiple classes or multiple labels, the data is treated as a collection of binary problems, one for each class. There are several ways to average binary metrics across the set of classes, each of which can be useful in some scenarios.

- "macro" simply averages binary metrics, assigning equal weight to each class. In problems in which infrequent classes are important, macro-media can be a means of highlighting their performance. On the other hand, the assumption that all classes are equally important is often false, and the macro-medium overemphasizes the typically low performance in an infrequent class.
- "weighted" is responsible for class imbalance, averaging the binary metrics in which the score of each class is weighted by its presence in the true data sample.
- "micro" provides each sample class pair with a contribution equal to the overall metric (except as a result of the sample weight). Instead of adding the metric by class, it adds the dividends and dividers that make up the metrics by class to calculate a general quotient. Micro-averaging may be preferred in configurations of multiple labels, including classification into various classes, whereas the majority class should be ignored.

The following Tables 4 and 5 show the result for the model metrics.

TRAIN	Precision	recall	f1-score	support
Did not complain	0.99	0.97	0.98	138715
Complained	0.13	0.26	0.17	2546
Micro avg	0.96	0.96	0.96	141261
Complained	0.56	0.61	0.58	141261
Complained	0.97	0.96	0.96	141261

TABLE 4. Training results

Statistical analysis of the binary classification, the F1-score is a measure of the accuracy of a test. It considers accuracy p and recall r of the test to calculate the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier and

TEST	Precision	recall	f1-score	support
Did not complain	0.98	0.93	0.96	350322
Complained	0.12	0.40	0.18	8417
Micro avg	0.92	0.92	0.92	358739
Complained	0.55	0.66	0.57	358739
Complained	0.96	0.92	0.94	358739

TABLE 5. Test results

r is the number of correct positive results divided by the number of all the relevant samples, that is, all the samples that should have been identified as positive. The F1 score is the harmonic average of precision and recovery, whereby an F1 score reaches its best value at 1 and worst at 0.

The ROC curve, calculates the receiver operating characteristic curve, a receive operating characteristic (ROC), or simply ROC curve, is a graph that illustrates the performance of a binary classifier system, as its discrimination threshold varies. It is created by plotting the fraction of true positives from positives (TPR = true positive rate) versus the fraction of false positives from negatives (FPR = false positive rate), in various limit settings. TPR is also known as sensitivity, and FPR is a minus specificity or true negative rate. This function requires true binary value and target scores, which can be estimates of positive class probability, confidence values or binary decisions.

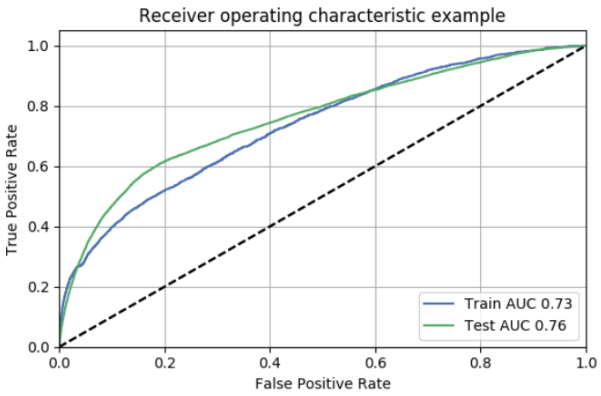

FIGURE 8. ROC curve

5) Forecast application

This section will explain the application of the customer prioritization model of occurrences in operation, with a greater propensity to call the call center to complain about the power outage. The application of the customer prioritization model is performed each 15 minute cycle of the robot.

After loading the models and the forecast base and applying the necessary transformations, the model and the threshold cut of the model, defined in the detection threshold parameter, are applied to generate the prioritizations.

VI. RESULTS AND DISCUSSIONS

The evaluation of the models can be discussed in the points described in the following subsections.

A. STATISTICAL MODELS FOR INTELLIGENT ROBOT FUNCTIONS

For State-1 All the occurrences from 40 cities - Real time less than 720min, results in Figure 9.

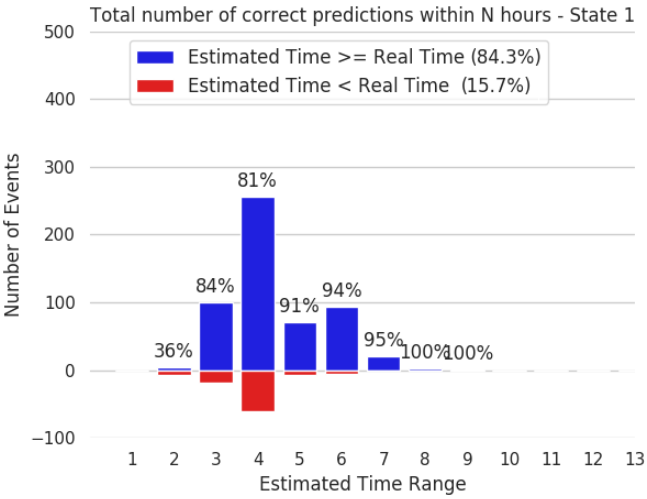

FIGURE 9. All the occurrences from 40 cities - Real time of less than 720min

In this figure, the blue bars represent the number of occurrences correctly predicted in the first forecast (when 1st Forecast $\geq$ Real Time) or in the second forecast (when 1st Forecast $<$ Real Time and 2nd Forecast $\leq$ Real Time) with their respective percentages by the time range predicted, while the red bars represent the number of occurrences erroneously predicted in the first forecast that the second forecast was unable to correct (1st Forecast $<$ Real Time and 2nd Forecast $<$ Real Time) with their respective percentages by the predicted time range. Thus, with the second forecast, the model goes from 59% to 83.4% of assertiveness.

For State-2 All the occurrences from 40 cities - Real time less than 720min, results in Figure 10.

B. STRATEGIC FOR MANAGING AND DEPLOYING INTELLIGENT ROBOTS

Lessons Learned:

- Considerations about tools and architecture: During the execution of this project, the use of a traditional RPA tool to send communications to customers selected by the prioritization model was analyzed. In this case, its use would be restricted to the consumption of an API responsible for triggering communications for each item registered in a work queue. When evaluating its performance for carrying out this activity, due to the effort of orchestration time of the solution and the cost associated with this tool, it was decided to use Databricks was chosen to use this communication API instead of a traditional RPA tool. The choice of Databricks

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3075693, IEEE Access

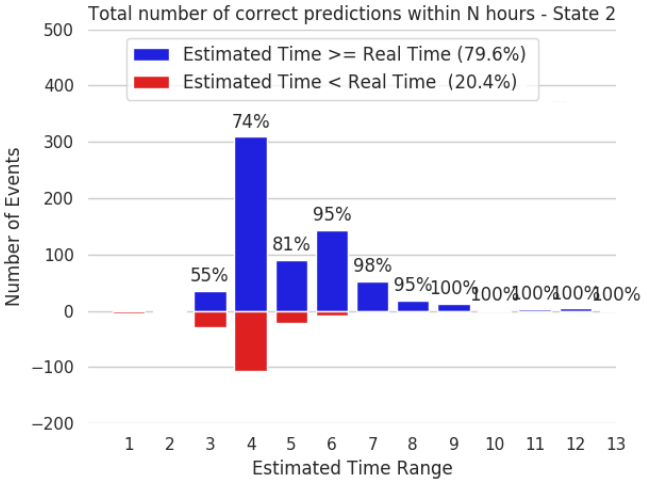

FIGURE 10. All occurrences from 40 cities - Real time less than 720min

was also stimulated by the fact that we had already built the entire data structure of the robot on it and that in terms of time and cost this solution proved to be more efficient.

- Solution Cost: The cost of the Cloud solution is calculated based on different information related to the services that will be used and the time they will consume. The information used to calculate the costs of each service varies according to its type. The following items describe the information taken into account when pricing our solution, for each of the services used:

- Azure Data Factory: Region, Type, Service Type (Data Pipeline or SQL Server Integration Services). For the Data Pipeline service type, the information are activity runs (in thousands), data integration unit hours, number of pipeline activity execution hours, number of external pipeline activity execution hours, Data Flow type, quantity of instances and number of hours running (Compute Optimized vCores, Compute Optimized vCores, Memory Optimized vCores), and number of Entity units (50,000 entities) for Read/Write operations and Monitoring operations. For SQL Server Integration Services service type, the information includes Tier, Instance, number of Virtual machines, and number of hours running;
- Azure Databricks: Region, Workload (Data Analytics, Data Engineering or Data Engineering Light), Tier, Instance, number of Virtual machines, number of hours running, and number of hours Databricks Unit running. Note that this product allows a discount for booking for 1 year or 3 years.
- Azure Data Lake: This product allows the payment by commitment which permits to pre-pay the storage size. For the Pay-as-you-go pricing type, the information is Region, Storage used, and number of Transaction units (10,000 transactions) for Write and Read transactions.

- Scale the solution: This solution was built with the possibility of parameterizing the cities to be served through the correct configuration of the application developed in Power Apps. Note that all the development of the models and architecture of the solution was carried out considering the data of all the municipalities present in the EUC concession area. Therefore, the solution can be perfectly scaled for the entire concession area without the need for any further development effort. If the concession area is expanded to new cities or states, the solution can also include these new municipalities by retraining the models and making some adjustments in data intake and Power Apps considering the new data.

The models present the technical assessment of the model, in which MAPE and MAE are presented in minutes of the forecasts. It is explicit in numerical form the general values of the day obtained for these metrics considering the first forecast and the second forecast, while in the graphs (Dashboard - Power BI) the metrics are presented considering each iteration of the flow.

It is important to note that events with a closing time of less than 60 minutes are filtered, since this is the minimum time reported by the robot, it is also important to understand how the first and second prediction classification occurs, as this classification changes the values through MAPE. Once the first prediction of an occurrence expires, the occurrence is classified as a second prediction, so the MAPE of the first prediction presents the error percentage only of occurrences overestimated by the first prediction, therefore, the occurrences underestimated by the first prediction are classified as a second forecast.

For those interested in exploring dataset in detail, it has been made available in zenodo.org [1].

VII. HUMAN FACTORS

Clients' interaction with the RPA is understood as a relevant factor for the acceptance of G3 technology. Following the Brazilian Business Pact for Humanized Work Digitalization, the company carried study on users after 18 months of RPA G1 adoption. The study intended to understand the workers' experience with the RPA. The study revealed that RPA has provided an overall positive user experience mainly due to the perceived utility of the spared time, the effective upgrade in career opportunities and the pride for actively participating in the innovation adoption [authors' paper omitted for blind review].

The G3 RPA described in this paper, unlike the previous G1 experience, is intended to interact with the company's client. In that previous study, workers had mentioned their concern with the automated communication with their clients: Do you imagine a robot calling you? This interaction has to be the most humanized possible. The person being served should have the impression of a not-so-cold service and that he/she is actually having a similar interaction than she would have with a person. Both clients' and workers'

[1]http://doi.org/10.5281/zenodo.3995046

experience with the G3 technology are interesting research themes that remain open for investigation.

VIII. CONCLUSIONS

In this work an intelligent "Proactive Notification" RPA was developed for the electric power sector (Electricity Utility Company). Currently, RPA is capable of providing highly accurate proactive notifications to customers with the highest probability of complaints. This proposed RPA is capable of monitoring the system responsible for mapping power interruptions, estimating for each occurrence. By SMS and phone calls, it thus prioritizes and communicates, clients with a high probability of filing a complaint with the public service company.The acceptance of the robot was good, people who were called did not call to complain. The models show that the forecasts proposed one after the other (first and second forecast) are increasingly accurate, going from approximately 60% to 85% of accuracy.

The proposed abandons the traditional RPA concept of automating "Back-office" tasks to be used in companies' areas of operation. Although the challenges of the benefits and capabilities that an intelligent RPA can offer us have been overcome, the test time is still short to be able to observe all the advantages to the maximum. This work is a strong foundation for the creation of G1, G2, and G3+ robot RPAs. The era of digital transformation requires it, applying disruptive technologies on a large scale, analysis of benefits and socio-economic and cultural impacts.

IX. ACKNOWLEDGE

This research has been supported by the ANEEL's R&D program.

REFERENCES

[1] Santiago Aguirre and Alejandro Rodriguez. Automation of a business process using robotic process automation (rpa): A case study. In Workshop on engineering applications, pages 65–71. Springer, 2017.

[2] R. Uskenbayeva, Z. Kalpeyeva, R. Satybaldiyeva, A. Moldagulova, and A. Kassymova. Applying of rpa in administrative processes of public administration. In 2019 IEEE 21st Conference on Business Informatics (CBI), volume 02, pages 9–12, 2019.

[3] S. Gupta, S. Rani, and A. Dixit. Recent trends in automation-a study of rpa development tools. In 2019 3rd International Conference on Recent Developments in Control, Automation Power Engineering (RDCAPE), pages 159–163, 2019.

[4] P. Desai. Robotic process automation: Rpa pre-requisite and pivotal points : Special issue: Special issue:iaisct(ss4). In 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), pages 446–451, 2020.

[5] T. Kobayashi, K. Arai, T. Imai, and T. Watanabe. Rpa constitution model for consumer service system based on iot. In 2019 IEEE 23rd International Symposium on Consumer Technologies (ISCT), pages 82–86, 2019.

[6] Y. Ma, D. Lin, S. Chen, H. Chu, and J. Chen. System design and development for robotic process automation. In 2019 IEEE International Conference on Smart Cloud (SmartCloud), pages 187–189, 2019.

[7] C. Xue. A task parallel processing technology for robot process automation. In 2019 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), pages 543–5432, 2019.

[8] Han Ping Fung. Criteria, use cases and effects of information technology process automation (itpa). Advances in Robotics & Automation, 3, 2014.

[9] James R Slaby. Robotic automation emerges as a threat to traditional low-cost outsourcing. HfS Research Ltd, 1(1):3–3, 2012.

[10] Mary C Lacity and Leslie P Willcocks. A new approach to automating services. MIT Sloan Management Review, 58(1):41–49, 2016.

[11] Mary Lacity, Leslie P Willcocks, and Andrew Craig. Robotic process automation at telefonica o2. 2015.

[12] Esko Penttinen, Henje Kasslin, and Aleksandre Asatiani. How to choose between robotic process automation and back-end system automation? 2018.

[13] Wil MP Van der Aalst, Martin Bichler, and Armin Heinzl. Robotic process automation, 2018.

[14] Aleksandre Asatiani and Esko Penttinen. Turning robotic process automation into commercial success–case opuscapita. Journal of Information Technology Teaching Cases, 6(2):67–74, 2016.

[15] Jerome Geyer-Klingeberg, Janina Nakladal, Fabian Baldauf, and Fabian Veit. Process mining and robotic process automation: A perfect match. In BPM (Dissertation/Demos/Industry), pages 124–131, 2018.

[16] M Gotthardt, D Koivulaakso, OKYANUS Paksoy, CORNELİUS Saramo, MINNA Martikainen, and OM Lehner. Current state and challenges in the implementation of robotic process automation and artificial intelligence in accounting and auditing. ACRN Oxford J. Finance Risk Perspectives, 8:31–46, 2019.

[17] Jan Mendling, Gero Decker, Richard Hull, Hajo A Reijers, and Ingo Weber. How do machine learning, robotic process automation, and blockchains affect the human factor in business process management? Communications of the Association for Information Systems, 43(1):19, 2018.

[18] Somayya Madakam, Rajesh M Holmukhe, and Durgesh Kumar Jaiswal. The future digital work force: robotic process automation (rpa). JISTEM-Journal of Information Systems and Technology Management, 16, 2019.

[19] X. Ling, M. Gao, and D. Wang. Intelligent document processing based on rpa and machine learning. In 2020 Chinese Automation Congress (CAC), pages 1349–1353, 2020.

[20] Sascha Chandler, Clare Power, Morven Fulton, and Nathalie Van Nueten. Who minds the bots? why organisations need to consider risks related to robotic process automation, 2017.

[21] José Gonzalez Enríquez, A Jiménez-Ramírez, FJ Domínguez-Mayo, and JA García-García. Robotic process automation: a scientific and industrial systematic mapping study. IEEE Access, 8:39113–39129, 2020.

[22] Mary Lacity, LP Willcocks, and A Craig. Robotizing global financial shared services at royal dsm. The outsourcing unit working research paper series, 2016.

[23] Abderrahmane Leshob, Audrey Bourgouin, and Laurent Renard. Towards a process analysis approach to adopt robotic process automation. In 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE), pages 46–53. IEEE, 2018.

[24] Sandeep Vishnu, Vipul Agochiya, Ranjit Palkar, et al. Data-centered dependencies and opportunities for robotics process automation in banking. Journal of Financial Transformation, 45(1):68–76, 2017.

[25] Bo Liu and Ning Zhang. Decision-making for rpa-business alignment. In LISS2019, pages 741–756. Springer, 2020.

[26] Lambert Rutaganda, Rudolf Bergstrom, Avijeet Jayashekhar, Danushka Jayasinghe, Jibran Ahmed, et al. Avoiding pitfalls and unlocking real business value with rpa. Journal of Financial Transformation, 46:104–115, 2017.

[27] Leslie Willcocks, John Hindle, and Mary Lacity. Keys to rpa success. Technical report, Executive Research Report. Knowledge Capital Partners, 2018.

[28] Bendik Bygstad. Generative innovation: a comparison of lightweight and heavyweight it. Journal of Information Technology, 32(2):180–193, 2017.

[29] Craig Le Clair, A Cullen, and M King. The forrester wave™: Robotic process automation, q1 2017. Forrester Research, 2017.

[30] Pavle Mijović, Evanthia Giagloglou, Petar Todorović, Ivan Mačužić, Branislav Jeremić, and Ivan Gligorijević. A tool for neuroergonomic study of repetitive operational tasks. In Proceedings of the 2014 European Conference on Cognitive Ergonomics, pages 1–2, 2014.

[31] Leslie P Willcocks, Mary Lacity, and Andrew Craig. Robotic process automation at xchanging. 2015.

[32] Leslie Willcocks, Mary Lacity, and Andrew Craig. Robotic process automation: strategic transformation lever for global business services? Journal of Information Technology Teaching Cases, 7(1):17–28, 2017.

[33] J. A. E. Arias, J. A. B. Beltrán, and S. Bedoya. Rpa implementation for automation of managenment process of personal in compañía nacional de empaques s.a. In 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), pages 1–5, 2020.

[34] Peter Hofmann, Caroline Samp, and Nils Urbach. Robotic process automation. Electronic Markets, 30(1):99–106, 2020.
[35] Julia Siderska. Robotic process automation — a driver of digital transformation? Engineering Management in Production and Services, 12(2):21–31, 2020.
[36] Simone Agostinelli, Andrea Marrella, and Massimo Mecella. Research challenges for intelligent robotic process automation. In Chiara Di Francescomarino, Remco Dijkman, and Uwe Zdun, editors, Business Process Management Workshops, pages 12–18, Cham, 2019. Springer International Publishing.
[37] Rehan Syed, Suriadi Suriadi, Michael Adams, Wasana Bandara, Sander J.J. Leemans, Chun Ouyang, Arthur H.M. ter Hofstede, Inge van de Weerd, Moe Thandar Wynn, and Hajo A. Reijers. Robotic process automation: Contemporary themes and challenges. Computers in Industry, 115:103162, 2020.
[38] Max Gotthardt, Dan Koivulaakso, Okyanus Paksoy, Cornelius Saramo, Minna Martikainen, Othmar Lehner, et al. Current state and challenges in the implementation of smart robotic process automation in accounting and auditing. ACRN Journal of Finance and Risk Perspectives, 2020.
[39] Yara Rizk, Vatche Isahagian, Scott Boag, Yasaman Khazaeni, Merve Unuvar, Vinod Muthusamy, and Rania Khalaf. A conversational digital assistant for intelligent process automation. In International Conference on Business Process Management, pages 85–100. Springer, 2020.
[40] Devansh Hiren Timbadia, Parin Jigishu Shah, Sughosh Sudhanvan, and Supriya Agrawal. Robotic process automation through advance process analysis model. In 2020 International Conference on Inventive Computation Technologies (ICICT), pages 953–959, 2020.
[41] John R Talburt. Entity resolution and information quality. Elsevier, 2011.
[42] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. Scientific data, 3(1):1–9, 2016.
[43] Jinchuan Chen, Yueguo Chen, Xiaoyong Du, Cuiping Li, Jiaheng Lu, Suyun Zhao, and Xuan Zhou. Big data challenge: a data management perspective. Frontiers of Computer Science, 7(2):157–164, 2013.
[44] Sebastian Thrun and Lorien Pratt. Learning to learn. Springer Science & Business Media, 2012.
[45] Taiwo Oladipupo Ayodele. Types of machine learning algorithms. New advances in machine learning, 3:19–48, 2010.
[46] Mark R Segal. Machine learning benchmarks and random forest regression. 2004.
[47] Thomas G Dietterich. Ensemble methods in machine learning. In International workshop on multiple classifier systems, pages 1–15. Springer, 2000.
[48] P. Martins, F. Sá, F. Morgado, and C. Cunha. Using machine learning for cognitive robotic process automation (rpa). In 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), pages 1–6, 2020.
[49] Marshall Copeland, Julian Soh, Anthony Puca, Mike Manning, and David Gollob. Microsoft azure. Apress: New York, NY, USA, 2015.
[50] Rawan Aljamal, Ali El-Mousa, and Fahed Jubair. Benchmarking microsoft azure virtual machines for the use of hpc applications. In 2020 11th International Conference on Information and Communication Systems (ICICS), pages 382–387. IEEE, 2020.

BRUNA VAJGEL holds Ph.D. degree in Astrophysics at Federal University of Rio de Janeiro and at Harvard University and Master degree at National Observatory. Nowadays she works as a data scientist, specialist in developing solutions based on data mining, data extraction, data analysis and statistical modeling. She has more than ten years of experience in the research field, both in the national and international environment, planning and coordinating scientific projects of data science. Recently she worked in advanced data analytic projects, developing predictive and descriptive models using machine learning techniques in a health insurance company, power utilities company and in a bank. She also participated of a strategic analytic project, using data mining and data analysis techinques over public databases to elaborate a study about the brazilian Market scenario to a OilGas Institute.

PEDRO LUIZ PIZZIGATTI CORRÊA undergraduated in computer science at Universidade de São Paulo in 1987, master in computer science at Universidade de São Paulo in 1992, PhD in electrical engineering from the Escola Politécnica of the University of São Paulo (2002), Post-doc in Data Science at the University of Tennessee (2015) and Associate Professor at the University of São Paulo (2017). He is currently associate professor at the Computer Engineering and Digital Systems Department (PCS) of Escola Politécnica of University of São Paulo, SP, Brazil, working mainly on distributed databases, data science, modeling of computer systems, architecture of distributed systems, computing and biodiversity, agricultural automation and electronic government.

THAIS TÓSSOLI DE SOUSA is graduated in Electrical Engineering at Universidade Estadual de Campinas in 2013 and Master in Automation at Universidade Estadual de Campinas in 2015. She is currently working on robotic process automation projects. Her research interests are automation and Robotic Process Automation.

ROSA VIRGINIA ENCINAS QUILLE is graduated in Systems Engineering from the National University of the Altiplano (2007) - Peru, Master in Computer Science and Computational Mathematics from the University of São Paulo (2014) - Brazil. Currently, she is a PhD student at the Information System of EACH USP, and the EPUSP Big Data research and extension group at Computer Engineering and Digital Systems department (PCS) of Escola Politécnica da Universidade de São Paulo, SP, Brazil. Her academic and professional activities focus on Data Science, Big Data analytics, IoT, Artificial Intelligence, data base, Environmental data analysis and Data mining.

JOHN A. R. BEDOYA is graduated in Physics from the National University of Colombia in 2007, holds its M.S. and Ph.D. in Theoretical Physics from the Institute for Theoretical Physics from the São Paulo State University in 2009 and 2013. Post-doc in Mathematical Physics And Quantum Dynamical Systems at Bahia Federal University. He is specialized in Dynamical Systems, Quantum and Classical Symmetries, Mathematical Physics, and Computational Modeling of Physical Systems. He is a Senior Data Scientist acting in the Construction of predictive models, Machine Learning, and IA applications in the Electrical Sector, he is also acting in the academical sector in the research of applications of dynamical systems and symmetries. At the moment John is a referee of the IOP's Machine Learning: Science and Technology Journal and the Journal of Physics A: Mathematical and Theoretical.

LUCIA VILELA LEITE FILGUEIRAS received the B.S., M.S., and Ph.D. degrees in Electrical Engineering from the Escola Politécnica, Universidade de São Paulo, in 1983, 1989, and 1996, respectively. She has been an Assistant Professor with the Computer Engineering Department, Escola Politécnica, Universidade de São Paulo, since 1990. She was a member of IEEE Reliability Society and IEEE RS Human Interface Technology Committee from 1998 to 1994. Her research interests are in the areas of human interaction with automation systems, including human reliability, human-agent interaction, human-data interaction, and information visualization.

VANESSA R. S. DEMUNER graduated in Electrical Engineering from Faculdade Novo Milênio (2016). She holds a postgraduate degree from Faculdade Estácio de Sá/ES in the MBA course in Project Management. She currently works as an Engineer in Technological Development/R&D at EDP, a multinational company with operations in 14 countries, in the Electricity Distribution, Generation, Transmission and Commercialization segment. She has articles published in journals (congresses, seminars). She has technical training in Electromechanics from the Federal Institute of Espírito Santo. Specialization in Power Systems and Master in Engineering in progress. Among the projects in which she worked, the themes of Intelligent Electric Grids, Robotization, Customer-focused Solutions, Autonomous-Cooperative Inspection System for Electric Energy Assets using VANTS, Distributed Solar Generation, Evaluation of the R&D Program stand out ANEEL and Electric Mobility.

DENIS MOLLICA is graduated in Physics at Universidade Federal Fluminense, holds a M.S. in nuclear engineering and energetic planning from Universidade Federal do Rio de Janeiro and Ph.D. in Electric Engineering from Universidade de São Paulo. He has been an Assistant Professor with the Computer Engineering Department, Escola Politécnica, Universidade de São Paulo, since 2003. His research interests are Data Governance and Data Quality, BI, Big Data, Social networks in Recommender System and human-data interaction.

GUSTAVO MATHEUS DE ALMEIDA Graduation (2000) and Master (2003) in Chemical Engineering from the Federal University of Minas Gerais, and specialization in Paper and Cellulose (2004) and PhD in Chemical Engineering (2006) from the University of São Paulo, with an internship at IDIAP Research Institute, Switzerland (2005). Post-doctorate at the Federal University of Minas Gerais (2007-2008), and visiting professor at the University of Coimbra, Portugal (2011). Professor at the Department. of Chemical and Statistics Engineering at the Federal University of São João del-Rei from 2009 to 2014. Since 2015, professor at the Department. of Chemical Engineering at the Federal University of Minas Gerais. Permanent professor of its Graduate Program in Chemical Engineering, and founder and coordinator of the Data Analysis and Visualization Research Laboratory.