Collaboration Study Case between the Industry and the University: Use of Lightgbm Algorithm for Data Analytics in the Execution History of Scheduled Routines (Job)

Juan J. Arenas[1], Cesar Soto[2], and Freddy Paz[1]
[1] Department of Engineering Pontifical Catholic University of Peru, Lima, Lima 32, Peru
[2] Graduate School Pontifical Catholic University of Peru, Lima, Lima 32, Peru
Email: {jjarenas; fpaz}@pucp.edu.pe; cesar.soto@tcs.com

Abstract —At present, the link between the University and the industry for the generation of innovation is becoming more frequent. This link is achieved through cooperation projects, where a company presents a challenge to the university. In the case of computer engineering, the challenges are in the development of software, systems auditing or data analytics, among others. In this paper, we will present the work done by the university for a company. The objective of this project was to analyze a set of more than 5 million data to predict whether a Job (routine program to execute an executable) will be executed correctly or not. For the project, CRISP-DM was used as a methodology, and the activities carried out during the execution of the project range from the understanding of the business to the validation of the selected model. The algorithm presented for the proposed model was LightGBM, which has been widely used due to the speed of training with large amounts of data.

Index Terms—Information system, data analytics, case study, data methodology, machine learning.

I. Introduction

At present, the link between the University and the industry is getting stronger [1]. This link occurs through joint projects where the University seeks to help the industry generate innovation. In the case of information technology, projects are much more common, due to the fact that the industry, many times, does not have computer resources [2]. For example dedicated servers, software licensing, human resources dedicated to research. The industry has problems with personnel trained in specific topics, time for research and technological tools [3].

The industry is going through a digital revolution, where the company is in need of competitive advantage through information obtained in large amounts of data. This data, many times, stored by companies but not analyzed to obtain strategic information [4] and this is because companies do not have knowledge in data analytics. Although companies could contract the service of a specialist, currently there is the alternative of finding an ally to do innovation projects. This ally can be the university, because it has been generating knowledge in the area of information systems, more specifically, in data analytics.

Data analytics is an area of knowledge corresponding to the science of data and results in providing information to a user or machine [5]. The analysis of the data is divided into two activities [6]: The first, a data preparation activity, in which different techniques are performed to produce a table with data ready to be analyzed. The second activity has to do with the techniques used to obtain the expected results; from an exploratory analysis to a behavior prediction. These techniques are used by means of algorithms created in different programming languages.

In this article, we are going to present the data analytics project. The project was carried out between the company Tata Consulting Services Peru and the Pontifical Catholic University of Peru. This project aimed to analyze more than 5 million data, corresponding to the planning and execution of Jobs of all 2017. These data had to be prepared to then enter a behavior that allows predicting if a Job is going to run correctly or will have some error during its execution. It is necessary to highlight that the variables, delivered by the company, were very limited and, at the stage of data preparation, it was necessary to supplement these variables with external information. For example, the date was linked to information on non-working days (holidays) and the name of the day of the week. The information of the predecessor job also had to be added.

The project used CRISP-DM as a data analysis methodology [7], which has the facility to perform a set of activities in a cyclical way (once the last activity is finished, the first activity can be performed again information of the previous cycle). The cycles facilitated the analysis of the data because the behavior of the data was not known and more than one cycle had to be carried out to determine the correct model. For this project, 3 cycles had to be carried out. In the first cycle a model with the linear regression algorithm was used, in the second cycle a logistic regression algorithm was used, and then in the third cycle, using the LightGDM algorithm. Among the activities considered, of each cycle, are:

Manuscript received May 6, 2019; revised December 5, 2019.
Corresponding author email: jjarenas@pucp.edu.pe.
doi:10.12720/jcm.15.1.101-106

- Understanding the business, in this case, the company had meetings every two weeks with the work team of the university to explain the behavior of the data. The team defined the objectives that were required and resolved the doubts about the types of variables presented. The objective of the project was to determine if a Job is going to run correctly or if it was going to have an error during its execution.
- Preparation of the data, in this stage the data had to be complemented with external information. First, the name of the day of the week was added to the variables delivered by TCS. Second, a new variable was added that determined whether the day of execution was a working day or not (holiday). The objective of the new variables was to determine if one day was related to the execution situation of the Job. On the other hand, the variables corresponding to the execution of the predecessor Job were added, because in the understanding of the business, it was explained that on several occasions the delay of a Job could influence whether the next Job is executed correctly or not.
- Modeling, the last algorithm used was LightGBM, which is a gradient improvement framework that uses a tree-based learning algorithm. This algorithm was chosen due to its high speed in handling large amounts of data. One of the problems of traditional algorithms is that their times increase while the amount of data to be processed increases. For the creation of the model, 5 million records were used, which correspond to the total of jobs executed during 2017. Within the model training technique, the data had to be separated into 70% to train the model and 30% to verify if the trained model is correct. We also had to configure the input parameters of the LightGBM algorithm with values that allow rapidity in the result, and there is no overfitting.
- Evaluation, after finishing executing the model it is necessary to validate the prediction accuracy. For this, 2 million data corresponding to 2018 will be used. As previously explained in the model, it was created with 5 million records, which correspond to the history of job executions in 2017. The validation aims to compare the result of the model with the real. This comparison is made by means of three techniques: precision, recall, and accuracy. It will be satisfied if the model predicts more than 70% of the total of the jobs with erroneous execution.

This article will be divided into three sections. The first is the explanation of the case study, which will explain the context of work that has been done between the company and the university. The second section will present the activities carried out, according to the CRISP-DM methodology. We will show the table of the variables that were determined for the use of the algorithm. It will also present the execution of the algorithm, which will detail the lines of code used for the execution of the algorithm. Finally, the technique will be shown to validate the results obtained. We will describe the values of the prediction accuracy given by the LightGBM model. In the third and last section, we will present the conclusions of the case study, as well as future work to be done.

II. Description of the Case

In 2018 Tata Consulting Services Company (TCS) signed an agreement to exchange information with the Pontifical Catholic University of Peru (PUCP). In this agreement, each of its parties is committed to providing information with the aim of generating new knowledge. Within the agreement, the data analytics project was generated, whose objective was to predict the performance of Jobs' execution through the execution history. The objective of this project was to obtain information through data mining techniques.

It is necessary to highlight that the PUCP has a computer engineering section, in which there are teachers and students dedicated to the management, preparation, and management of data [8]. This group of people has been developing various academic projects related to the study of data analytics since 2016. The group consists of 2 full-time teachers, one with the profile of information systems and the other with the profile of Machine learning techniques. Additionally, the group has the participation of 8 students in the computer engineering career, which had been doing extracurricular activities with respect to the use of data mining techniques.

And in the case of the company, TCS provides technology services [9] and, within its services to clients, is the monitoring and control of the execution of routines, also called a job. This service has activities carried out manually. Therefore, a person must be available 24 hours a day, 365 days a year to verify the correct functioning of the systems. One of the difficulties that this service has is that the person in charge must be available at non-conventional times, such as dawns or holidays. The response times, before an event, could often be affected by the lack of personnel available to address the problem.

Table I: Variables of the Jobs

Variable	Description
Mesh	It is the set to which a Job belongs.
Scheduled date	It is the date on which a Job should be executed.
Initial execution date	Start date and time of the execution.
Final execution date	Date and time of the end of the execution.
Duration	Time, in seconds, of the duration of the Job.
Average	Duration usually of Job. This duration is the sum of time between the amounts of the same Job.
Agent	Name of the server where the Job was executed.
State	The value that determines if the Job had an error or not.

In this context, TCS generated a collaborative project to identify some performance of the jobs through data analytics. TCS delivered to the PUCP a quantity of 7,536,162 of data on Jobs executed. These records contained variables about the programming, execution, and status (correct execution or execution error) of the Jobs. Table I shows the description of each of the fields received by TCS.

About the way of work, it can be described that the group determined weekly meetings. At each meeting, tasks were assigned on data preparation, preparation of development environments, implementation of algorithms, among other activities. It was also determined to have a meeting every two weeks with the TCS team, where the objective was to review the progress, verify if the predictive models were correct, and prepare the information by means of verification with data not considered for the training of the algorithm.

III. Methodology

Regarding the methodology, it should be mentioned that all the activities presented by CRISP-DM were used. Each activity had a final result. In the case of understanding the business, where we obtained the final result was the objective of the data analysis. For the preparation of the data, the variables necessary for the analysis were obtained as a result. For the modeling, the final model obtained from the execution of the LightGBM algorithm was obtained. Finally, the evaluation of the model, whose result was the values obtained in the accuracy, precision, and Recall.

A. Business Understanding

This activity aims to know the situation of the company, know the reasons why you need to predict behavior. In the case of TCS, he has described the need to modernize the traditional software factories towards digital software factories, using technology used in Industry 4.0, guaranteeing efficiency in the use of available resources: FTE, infrastructure, among others. In order to that and to make an effective follow-up and control the execution of Jobs, they propose the adoption of new tools and technology to solve the problem of follow up and control the execution of Jobs in the mainframe. It is known that this execution can be carried out during all 24 hours during the 365 days, so if it is desired to carry out an efficient control, manually, it is impossible. Among the alternatives to solve this problem, the case was presented to the PUCP, with the objective of analyzing the situation and defining the objective of the data analysis. After meetings it was determined that the objective of the data analysis was to predict if a Job was running correctly or an error was going to arise (the variable to be predicted was considered in a binary way, that is, YES in the case it was executed correctly and NOT in the case that there was an error). It was also possible to know the description of each variable and the possible values found in the more than 7 million records.

B. Preparation of the Data

Once the analysis objective was determined, the variables and types of data presented by TCS had to be reviewed. These data contained 8 variables, of which 7 were independent variables, and we had a dependent variable. As a first step in the preparation of the data, the dependent variable (State) had to be modified. Changed the data of Yes or No by binary values 0 (in the case there was no error) and 1 (in the case there was an error).

After this change, the average time of execution of the Job had to be corrected. This correction was due to the fact that the current value was being calculated by the total number of Jobs. This was an error that had to be corrected because the average should be determined by that Job's history and not by a total of Jobs, where future information is being considered. Fig. 1 show the error committed.

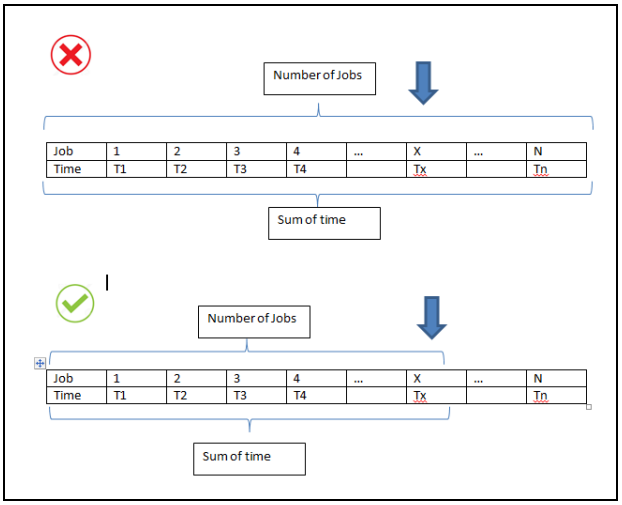

Fig. 1. Average calculation for Job X

The next step was to add information related to the behavior of the days. First a column was added where the first letter of the day of the week was entered. The objective was to know if the behavior of the day can affect the execution of a Job. As in the case of the name of the day of the week, the execution time was classified in four shifts: morning (6:00 am – 12:00 pm), afternoon (12:01 pm - 6:00 pm), night (6:01 pm - 12:00 am) and early morning (12:01 am - 5:59 am). This change was made for the same reason to the new name column of the day, which was to know if any range of hours affected the execution of the Job.

Finally, the day and time of the predecessor Job was added. This information was obtained thanks to the information of the Jobs mesh given by the TCS during the second cycle of the methodology. In general, the model would have 8 independent variables and a dependent binary variable. Table II shows the list of variables considered for the model.

TABLE II: VARIABLES OF THE JOBS

Variable	Description
Weekday scheduled	Field in text format that gives the name of the week. (Monday, Tuesday, Wednesday Thursday, Friday, Saturday, Sunday)
Weekdays start of execution	Field in text format that gives the name of the week. (Monday, Tuesday, Wednesday Thursday, Friday, Saturday, Sunday)
Weekdays final of execution	Field in text format that gives the name of the week. (Monday, Tuesday, Wednesday Thursday, Friday, Saturday, Sunday)
Scheduled time range	Field in text format that provides the range of hours. Early morning, morning, afternoon or evening.
Start time range of execution	Field in text format that provides the range of hours. Early morning, morning, afternoon or evening.
Average duration	Field in integer format that provides the average time in seconds.
Scheduled weekday of the previous Job	Field in text format that gives the name of the week. (Monday, Tuesday, Wednesday Thursday, Friday, Saturday, Sunday)
Scheduled time range of the previous Job	Field in text format that provides the range of hours. Early morning, morning, afternoon or evening.
State	Field in integer format. 0 (correct execution) and 1 (execution with error)

C. Modeling

The processing of the algorithm was done in Python. For this, different libraries were used, which facilitated the different steps necessary to execute the algorithm. The library that was used for the execution of LightGBM was LGBMCClassifier.

For the process of modeling the data, the job history executed in 2017 was used, and its quantity exceeds 5 million records. As part of the process, the 5 million were separated into two parts for training and verification. According to the theory, the training of the data can be done in several ways; however, for the project, the technique was used where 70% of the data were used to train the model, while the other 30% was used to validate the model [10]. Fig. 2 shows the source code that separated the data.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=10)
```

Fig. 2. Division of records

Once the distribution of the data has been determined, it is necessary to configure the parameters used in the LightBGM algorithm. The objective of these values was to perform a training quickly and avoid over-adjustment [11]. Within the parameters it was defined that the 'boosting-type' would be by default the traditional Gradient Boosting Decision Tree 'dbdt', that the number of sheets would also be by default, the learning rate, which determines the impact of each tree in the final result [12], was placed with a value of 0.05 (the range of values goes from 0.1 to 0), `bagging_fraction`, which determines the training speed and its values go from 0 to 1, a value of 0.8 was placed, `feature_fraction`, which determines the percentage of characteristics that will be used to train, in this case of placed 0.9,` bagging-frec`, which determines the number of iterations to carry out the bagging, in this case the value of 5 was placed. Figure 3 shows the parameters placed.

```
params = {
    'task': 'train',
    'boosting_type': 'gbdt',
    'objective': 'binary',
    'metric': { 'auc'},
    'num_leaves': 31,
    'learning_rate': 0.05,
    'feature_fraction': 0.9,
    'bagging_fraction': 0.8,
    'bagging_freq': 5,
    'verbose': 2
}
```

Fig. 3. Parameters of LightBGM

Finally, the LightGBM algorithm had to be executed. The 'Train' method of the LGBMCClassifier library was used to execute the algorithm. Fig. 4 shows the execution of the algorithm, where the execution of 100 trees was placed as a parameter.

```
lgb_train = lightgbm.Dataset(X_train, y_train)
lgb_test = lightgbm.Dataset(X_test, y_test, reference=lgb_train)
lgbm = lightgbm.train(params,
                lgb_train,
                num_boost_round=100,
                valid_sets=[lgb_train, lgb_test],
                     feature_name =cols,
                verbose_eval=2)
```

Fig. 4. Training execution

The execution gave as a result that the area under the ROC curve (AUC) is greater than 0.9 in all the trees. In Fig. 5 you can see some of the results.

```
[2]     training's auc: 0.998596        valid_1's auc: 0.998559
[4]     training's auc: 0.998633        valid_1's auc: 0.998593
[6]     training's auc: 0.998654        valid_1's auc: 0.998615
[8]     training's auc: 0.998662        valid_1's auc: 0.99862
[10]    training's auc: 0.998664        valid_1's auc: 0.998623
[12]    training's auc: 0.998674        valid_1's auc: 0.998633
[14]    training's auc: 0.998676        valid_1's auc: 0.998637
[16]    training's auc: 0.998687        valid_1's auc: 0.998648
[18]    training's auc: 0.998694        valid_1's auc: 0.998655
[20]    training's auc: 0.9987  valid_1's auc: 0.99866
[88]    training's auc: 0.998872        valid_1's auc: 0.998808
[90]    training's auc: 0.998874        valid_1's auc: 0.998811
[92]    training's auc: 0.998878        valid_1's auc: 0.998814
[94]    training's auc: 0.998879        valid_1's auc: 0.998815
[96]    training's auc: 0.998883        valid_1's auc: 0.998818
[98]    training's auc: 0.998885        valid_1's auc: 0.99882
[100]   training's auc: 0.998889        valid_1's auc: 0.998824
```

Fig. 5. Results of the execution of the trees

D. Validation of the Model

As a final step in the methodology, the model created by the LightGBM algorithm had to be validated and the data corresponding to 2018 was used. This means that the modeling was carried out with data from 2017 and the evaluation was made with data from the 2018 (records for 2018 correspond to more than 2 million). The evaluation

strategy was done by verifying the accuracy, recall, precision and the area under the ROC curve [13]. The accuracy determines the percentage of the predicted with respect to the total of the data, recall determines the amount of Job correctly predicted that it was going to be executed with error with respect to the total of Jobs that was executed with error and the precision determines the amount of Job predicted correctly that it was going to run with error on the total amount of predicted values that were going to be a Job with error. So the first thing was to obtain the values predicted by the algorithm and the real values for this new data set. It is necessary to highlight that the predicted values go between 0 and 1. After obtaining the predicted values and the real values, the evaluation must be determined by different cut-off values. For this case 10 different cut values were made. Fig. 6 shows the algorithm used to obtain the verification results for each of the cut-off values.

```
y_predict_lgbm=lgbm.predict(np.array(df_val[cols]))
y_val_true=np.array(df_val[target]).ravel()
thresh=[0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9]
for t in thresh:
    y_t=[(1 if y_predict_lgbm[i]>t else 0) for i in range(len(y_predict_lgbm))]
    print(t,accuracy_score(y_val_true, y_t),recall_score(y_val_true, y_t),precision_score(y_val_true, y_t),roc_auc_score(y_val_t
rue, y_t))
```

Fig. 6. Verification algorithm

Execution of the algorithm resulted in a list of accuracy, recall and precision values. These values can be seen in Fig. 7. The first column being the cut-off value (the cut-off value is the value that will determine whether there was an error or not.) For example, if the cut-off value is 0.4, it means that the predicted values by the algorithm that are less than 0.4 are determined as there was no error while values greater than 0.4 if there was an error in the execution), the second column represents the value of the accuracy, the third column the value of the recall, fourth column the value of the precision and finally, the AUC value.

```
0 0.009503879144991718 1.0 0.009503879144991718 0.5
0.1 0.9975942936380771 0.9990485252140818 0.7984557339650785 0.9983144326995399
0.2 0.9971512012707082 0.9394715655419746 0.7970443043869726 0.9685881034638718
0.3 0.9967943623626074 0.8694283832247677 0.8079028802666032 0.9337224155062347
0.4 0.9965408189279042 0.8181951255214813 0.8178958150424349 0.9082235918886554
0.5 0.9963968312983198 0.7701090536485399 0.8376721598598837 0.8843385659531643
0.6 0.9958633119172022 0.6639098294664422 0.8700364473431805 0.8314791290636045
0.7 0.9953711802846851 0.581973212325258 0.893979425487661 0.7906554874272427
0.8 0.9948039246715151 0.48034838615238235 0.94663204961777 0.7400442737420524
0.9 0.9925255117212538 0.2194979140745078 0.9735432559649407 0.6097203396704097
```

Fig. 7. List of accuracy, recall and precision values.

Then, after obtaining the results of accuracy, recall, and precision, it was determined that the ideal cut-off value by means of an analysis of the results, where it was decided to have a recall value greater than 0.65 and a precision greater than 0.85.

IV. Conclusions

We have made a practical case of data analytics using the CRISP-DM methodology in a real case of the company TCS. This methodology was very helpful because it allows ordering the work throughout the process of data analytics. It is also necessary to describe that the data received was very limited and that the understanding of the business helped to determine new variables that were not described at a glance. On the other hand, the analysis of the variables helped to determine that the average placed by TCS was poorly calculated and an algorithm had to be used to calculate the real average. It should also be concluded that the work was not done by a single technique, but that the linear regression and logistic regression techniques were used. However, this technique did not help in the accuracy of the predicted values. Finally, the parameters placed in the LightGBM algorithm were determined by trial and error, knowing that there was little documentation on ideal values in this regard.

This project was a success that at present TCS is requesting the encapsulation of the algorithm to create a Web service, which will be integrated into an early detection system of errors. This system will allow having a visionary strategy on what can happen during days where there is a risk of little personnel and even a system that can correct errors automatically is being developed.

In this way, it is possible to respond with greater precision and effectiveness to the alerts generated in the 24x7 equipment, minimizing the uncertainty inherent in the agility of the business. Subsequently, and for future research linked to Analytics, a window opens in which the analysis of the collected data is pertinent in order to quantify the relationship between the real events and the predictive models generated in order to adjust appropriately to the model in search of its scalability to other industries.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Juan J. Arenas, Freddy Paz conducted the research; Juan J. Arenas, Cesar Soto execute the case study; Juan J. Arenas, Freddy Paz, Cesar Soto wrote the paper; all authors had approved the final version.

References

[1] R. Grimaldi, M. Kenney, D. S. Siegel, and M. Wright, "30 years after Bayh–Dole: Reassessing academic entrepreneurship," *Research Policy*, vol. 40, no. 8, pp. 1045–1057, 2011

[2] P. P. Khakbaz, "The role of research and development in growth of small and medium enterprise in technological cluster of regions," *Information Management and Business Review*, vol. 4, no. 5, pp. 234–241, 2012.

[3] S. Mayer and W. Blaas, "Technology transfer: An opportunity for small open economies," *Journal of Technology Transfer*, vol. 27, no. 3, pp. 275–289, 2002.

[4] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Proc. 24th Annual International Conference on the Theory and Applications of*

Cryptographic Techniques, Berlin, Heidelberg, 2006, pp. 486–503.
[5] R. Baeza–Yates and B. Ribeiro–Neto, *Modern Information Retrieval*, ACM Press, 1999.
[6] National Academies Press, Strengthening Data Science Methods for Department of Defense Personnel and Readiness Missions, National Academies Press, 2017.
[7] IBM. (2012). Manual CRISP-DM de IBM SPSS Modeler. IBM Corporation. [Online]. Available: ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf
[8] 2018. IA Research Group PUCP. [Online]. Available: http://ia.inf.pucp.edu.pe/
[9] 2018. Tata Consulting Group Peru. [Online]. Available: https://www.tcs.com/worldwide
[10] K. Dobbin and R. Simon. (2011). Optimally splitting cases for training and testing high dimensional classifiers. Dobbin and Simon BMC Medical Genomics. [Online]. Available:https://bmcmedgenomics.biomedcentral.com/articles/10.1186/1755-8794-4-31
[11] 2018. LightGBM. [Online]. Available: https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html
[12] G. Ke, Q. Meng, T. Finley, and T. Wang, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. 31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017.
[13] T. Yuan, H. Wu, J. Zhu, L. Shen, and G. Qian, "MS-UCF: A reliable recommendation method based on mood-sensitivity identification and user credit," in *Proc. International Conference on Information Management and Processing*, London, 2018, pp. 16-20.

Juan J. Arenas Master in Management and Policies of Innovation and Technology. Master in Management of business innovation. Currently a university professor with knowledge in the creation of innovation systems for companies. Research on technology transfer from research centers (universities, laboratories, etc. to companies). Activities in support of university entrepreneurship through workshops and advice to PUCP students. Also consultancies to companies for innovation issues through computer technologies created in the university. Activities in support of companies for innovation management issues through workshops and consultancies. More than 10 years of work experience in the public and private sector in issues of innovation management for IT.

Freddy A. Paz. Full-time professor at the Pontifical Catholic University of Peru. PhD in Engineering from the Pontifical Catholic University of Peru. Master in Computer Science with a minor in Software Engineering from the Pontifical Catholic University of Peru. Master in Computer Engineering from the Pontifical Catholic University of Valparaíso. Systems Engineer CIP 123414 with Professional Degree from the National University Pedro Ruiz Gallo. Research areas: Human-Computer Interaction, Programming Languages and Experimentation in Software Engineering.

Cesar A. Soto Extensive skills in Business 4.0 and over 10 years of experience on IT, specializing in software design, engineering, consulting and management of strategic and organizational digital projects of high complexity companies in the financial services and insurance, retail and telecommunications sector. Strong knowledge and experience in development of high quality software projects; and analytical skills and assertiveness in negotiating with clients, suppliers and the project team to achieve common objectives aligned to digital business strategy. Expert on C/C++, Machine Learning for music innovation and apasionated for code at all related technologies.