Data Mining in the Contractual Management of the Brazilian Ministry of Health: A Case Study

27

Alexandre Vinhadelli Papadópolis and Edna Dias Canedo

Abstract

Data mining is a process of analyzing data from different perspectives and summarizing it into useful information that can be used to classify data samples. Basically data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Machine learning is a branch of artificial intelligence which works with construction and study of systems that can learn from data. The core of machine learning deals with representation and generalization. The use of Data Mining is an important activity to control the development process of a software factory. Outsourcing software development by public and private organizations requires continuous monitoring to ensure compliance with service levels and small inspection teams coexist with work overload. This study seeks to apply data mining techniques and machine learning algorithms to create models to predict the delay in the delivery of demands. Following the CRISP-DM reference model, this study presents the process of creation and application of a predictive model for the contractual management of the Brazilian Ministry of Health.

Keywords

Data mining · Contractual management · Machine learning · CRISP-DM

A. V. Papadópolis · E. D. Canedo (✉)
Computer Science Department, University of Brasília (UnB), Brasília, DF, Brazil
e-mail: alex@sbpi.com.br; ednacanedo@unb.br

27.1 Introduction

Software development in most of Brazil's Federal Public Administration (APF), as well as in public and private organizations around the world, is carried out by outsourcing Information Technology (IT) services and can range from design to application execution [10, 16]. In APF's context, Normative Instructions (IN) of the Ministry of Planning (MP) govern software factory services contracting and dictates the rules for the preliminary study of the contracting, term of reference, contractual management and termination actions [12, 13].

According to Normative Instruction Number 04/2014-MP [6], the purpose of contractual supervision is to ensure adherence between convocational instruments (bidding notice, term of reference, contract) and its execution. Among other inspections is the compliance to service levels related to demands delivery deadlines, checked by a inspection team composed by a contract manager and one or more technical fiscals [1]. However, there is often an overload of the inspection team due to the large number of demands without the corresponding availability of people, which impairs the effectiveness of the control and, therefore, the objectives foreseen in the hiring. Excessive time is spent being dedicated to demands that do not delay, and also little attention to demands that delay. The specialists agree that one way to reverse this situation imposes the increase the inspection team efficiency, that is, to a greater assertiveness in the choice of demands that deserve more attention from the fiscals [4].

This study seeks to contribute to the topic through the application of data mining techniques in the logs of deliveries made by the software factory responsible for the Brazilian National Health Bus [3], a Service-Oriented Architecture (SOA) [22] platform that integrates citizens health data of dozens of applications provided by Ministry of Health of Brazil, in the search for a model to predict delays by the con-

S. Latifi (ed.), *17th International Conference on Information Technology–New Generations (ITNG 2020)*, Advances in Intelligent Systems and Computing 1134,
https://doi.org/10.1007/978-3-030-43020-7_27

tracted company and better guide the actions of the inspection team [3]. The conduction of the study will be guided by data mining methodology aimed at the discovery of knowledge [21] from databases, applying the necessary steps for selection, transformation and use in linear regression algorithms. Thus, the objective of this work is to answer the following research questions (RQ):

RQ.1: Is it possible to create a predictive model to indicate the possibility of delay in the execution of software construction demands for the Brazilian National Health Bus?
RQ.1: How can such a model be incorporated into the contractual management to positively impact on the monitoring of contract delivery times?

To answer the main objective of this work, a practical case study was conducted in a contract firmed by Brazilian Ministry of Health and a software factory to develop and maintain the Brazilian National Health Bus based on SOA architecture. Contract management is supported by Redmine, used as a demand management system.

The main contribution of this work was to build and apply a model to predict delays in the execution of software construction demands. In addition to improving the monitoring of deadlines, the study promoted improvement in contractual management and proved adequate to be extended to other contracts as well as to related domains, providing a better control of public spending with software development.

27.2 Background

Information Technology (IT) services are almost ubiquitous in all types of organizations, whether to define business strategies or to assist with operational processes [18]. Rather than maintaining its own staff, it is a common practice to outsource these services to specialized companies, not only to devote themselves to their mission, but also because the constant evolution of technologies [29]. This trend has also been observed since last century in APF. It is a practice so common that, in order to guarantee its transparency and control, Brazilian legislation provides several legal instruments specifically in this area. In addition to the Normative Instructions issued by the Ministry of Planning, there are also laws, decisions of the Brazilian Federal Court of Accounts, recommendations of the Office of the Union's General Controller, and ordinances of the Presidency of the Republic's Civil Cabinet [4].

The main advantage of outsourcing lies in the delegation of responsibility for recruiting, selecting, hiring and maintaining technical team of multidisciplinary IT professionals. Consequently, financial benefits (reduction and cost control, budget predictability), technical (professional specialization, agility in technological updating) and strategic benefits (focus on the core business) [4, 5, 13]. There are also significant risks: loss of knowledge, if the transition and contractual closure are not well planned; and lack of activities control, when the management team is not well trained and equipped. There are several legal recommendations and determinations of Brazilian external control bodies to mitigate such risks [7, 8].

In order to guarantee the desired results with the outsourcing, it is necessary to establish service levels that consider the length and complexity of the demands, service deadlines (levels of service) and intuitive indicators to verify compliance with these requirements. In addition, it is necessary the full understanding of rights and duties by stakeholders, creating a common language among those involved. From this minimum structure, the inspection teams of the contractor can monitor the execution of the demands and verify that the parameters are being respected.

The contracted company can also proactively identify bottlenecks, promote improvements in their work process, qualify their teams. In order to guarantee the desired results with outsourcing, it is necessary to establish service levels that consider the extent and complexity of the demands, service deadlines (levels of service) and intuitive indicators to verify compliance with these requirements. In addition, it is necessary the full understanding of rights and duties by stakeholders, creating a common language among those involved. From this minimum structure, the inspection teams of the contractor can monitor the execution of the demands and verify that the parameters are being respected. The contracted company can also proactively identify bottlenecks, promote improvements in their work process, qualify their teams.

27.2.1 Data Mining in the Context of Information Technology (IT)

Software development organizations coexist with the constant challenge of minimizing the time of creation and maintenance of products of a specialized nature. The study conducted by Goby and Brandt [27] proposed a Predictive Analytics-based model [15] to better estimate software delivery times by incorporating it into the product lifetime cycle (PLM). The CRISP-DM (Cross Industry Standard Process for Data Mining) [25] data model served as a roadmap for handling product-related data throughout its product life cycle and creating reusable models for a more accurate and continuously updated estimate of the lead time. Another way to identify useful metrics to predict success or failure in software construction is to make use of software repositories to gain a deeper understanding of the development process from the perspective of the data flow

associated with its steps [9]. Demand management systems used to track software-building activities can provide large amounts of data from event logs. The mining of these data enables them to understand and gain insight to improve business processes [24].

Predictive modeling also applies to IT incident resolution to help implement infrastructure changes [11] and to support the help-desk teams [23]. Random Forests and Gradient Boosting Machine classifiers proved to be efficient in relating specific components to the root cause of incidents. Other machine learning algorithms [31] have also been able to enable alerting mechanisms to anticipate incidents and improve knowledge sharing among users. In both cases, the input data was the tickets generated by the demand management applications.

27.2.2 Cross Industry Standard Process for Data Mining—CRISP-DM

CRISP-DM [26] is a well-known and widely adopted reference model for data mining, which offers flexibility to fit the needs of each project, allowing the creation of customized models. It suggests a set of progressive steps composed of tasks that can be chosen according to their applicability in the project.) Due to these characteristics, it was chosen to conduct this study.

Its six major phases are: (1) Business Understanding aims to understand business objectives and project requirements, culminating in the definition of a data mining problem; (2) Data Understanding encompasses collecting and gaining familiarity with data, identifying quality problems, increasing understanding, and formulating hypotheses; (3) Data Preparation aims at obtaining a final version of the database from the raw data, by means of various methods of transformation, selection and sanitization; (4) Modeling applies techniques of modeling and calibration of parameters to optimize the results; (5) Evaluation of the models obtained in relation to the predefined objectives; and (6) Deployment of the models in production [2, 19, 26].

The advantages of this model are: independence of business model, can be applied to analyze commercial data, financial, human resources, industrial production, service provision, among others; existence of several tools for its implementation; and close relationship with the KDD (Knowledge Discovery in Databases) process models, a process of extracting information from the database which creates relations of interest not directly perceived by experts in the subject, and assists the validation of extracted knowledge.

27.3 Methodology

This section presents the methodology used in this study, which was segmented in stages according to the phases proposed by CRISP-DM [17, 26]. The result of each phase is described. The preparation of this work was initiated through a bibliographical review and interviews with the managers and fiscals of the software factory contract, responsible for overseeing activities and assessing service levels. The purpose of these interviews was to identify the guidelines adopted for the process, as well as the elements considered as priorities. Thus, we sought to understand the objectives to be achieved in this study and the criteria for success.

During the activities the databases that could contribute to this article were identified, seeking the understanding of its semantics and verifying the levels of its quality. The insights obtained during this phase allowed the choice of data and the definition of the necessary selections, joins and transformations. The final database was generated and treated with data mining tools to create alternatives of predictive models that indicate the possibility of delay in the execution of the contractual demands. The obtained models were discussed together with the experts to choose the best option, identification of their limitations, possibilities of evolution and form of application in the routine of contract's inspection. As foreseen in CRISP-DM [26], the conduct of the work continuously provided a review of previously taken steps, in order to correct any errors. Finally, the results obtained were presented to the higher administrative hierarchy, aiming to ensure its applicability and continuous review, as well as evaluating the possibility of applying related studies in other contracts. The results of the CRISP-DM phases are detailed: Business Understanding, Data Understanding, Data preparation, Modeling and Evaluation.

27.3.1 Business Understanding

27.3.1.1 Scenario

Software construction activities are performed through contract with software factory and the contractual metric is the function point (PF). The tool to support contractual management is Redmine, used to record the demands from the initial request of the business area to the delivery of the version to be published in the production environment. The steps taken in this process are recorded in the form of phases and stages, each with a record of the start and finish times. By contractual requirement, the maximum execution times of the demands are defined based on the number of function points, which, in turn, is estimated based on the analysis of the impact of the demand and the use cases elaborated together with the requesting area. Failure to comply with the time limit

defined for each demand implies admonition and mulct for the service provider.

27.3.1.2 Project Goal

Elaboration of predictive models to indicate the possibility of delay in the execution of demands, based on the limits defined in the contract and in the times practiced (which will allow future refinement of the contractual instruments).

27.3.1.3 Success Criteria

In order to respond to the objectives of this study, data mining will be performed in the records of the contractual management system adopted by Datasus to discover existing behaviors and standards, and the success criterion is the creation of a predictive model that indicates the possibility of delay in implementation of a new demand, with a success rate of more than 75%.

27.3.2 Data Understanding

This section statistically describes the quality of the data, analyzes its behavior and describes the source of the data related to the business objectives. Initial data gathering:

- Demands database: when an area of the Ministry of Health requests new functionality in an information system, the information is recorded from the initial request to the delivery in production. The attributes of the process can be useful to identify the factors that determine the occurrence of delays.
- Holidays database: it informs dates considered without expedient by the Ministry of Health. The records in this base cover the year of beginning of the contractual execution until the year following the current one. At each beginning of the year, records are inserted to guarantee this rule.

27.3.3 Description of the Data

(a) Quantity of data: The database of the processing of demands is composed of **582 records**, relating to 130 different demands. This base is small because the contract under review was started just 2 months before this study.
(b) Quality of data
  - The database includes characteristics relevant to the analysis of the research problem, since it contains the periods of compliance of each process and the indication of the responsible person (software factory or Ministry of Health).
  - The prioritization of the relevant attributes was done with the help of business specialists, in this case the contract manager and the technical fiscals.
  - Considering that the problem to be addressed involves the analysis of the demands, and the steps are only elements that will contribute to the response, it will be necessary to add to the level of demand the time spent by each person in working days. Only after this aggregation will the other activities of the research be performed.

    The database contains no typographical errors, neither on mensurations. There are no coding inconsistencies or invalid metadata. The missing data is not represented by codes (nulls,?, 999).
  - The attributes of the initial database, extracted from the demands management system of the Ministry of Health, are the following:
    – *Demand*: Numerical sequential that uniquely identifies demand (primary key).
    – *Type*: Text containing information about the service catalog item that was demanded. From the catalog item it is possible to determine whether the demand deals with new project or existing project support; whether to produce new artifact or alter existing artifact; and whether the artifact to be worked on is a middleware or a service in the Bus. In addition, the text also informs whether the production start of the artifact resulting from the demand will require stopping the production server.
    – `Original Status`: numeric code that indicates the previous step to the current one.
    – `Status`: numerical code that informs the demand step to which the execution period refers. Through the code it is possible to determine if the person in charge is the software factory or the Ministry of Health.
    – `Description Status`: description of the `Status` code.
    – `Start`: timestamp of the moment that the step indicated by `Status` has been started. When the record refers to the first step of the demand, the value is equal to the moment of demand creation.
    – `End`: timestamp of the moment the step indicated by `Status` has been completed. When the record refers to the last processing of the demand, this attribute has no value.

27.3.4 Data Preparation

27.3.4.1 Selection of Data

The databases used in this study were obtained by extracting the records of the demand management system and the records related to holidays, applying SQL commands, restricted to the contract of evolution and maintenance of services based on the SOA (Service-Oriented Architecture)

of the Brazilian National Health Bus. The results of the SQL queries were saved in CSV (comma-separated values) and converted to Microsoft Excel XLSX format with three sheets: (1) `Holidays`: List of holidays used to calculate business days, extracted from the demand management system; (2) `Redmine Data`: base worksheet for the application of the transformation rules on the records; (3) `Times and SLA`: Worksheet generated after information is worked in Microsoft Excel Power Query.

27.3.4.2 Preparation of Data

In order to generate the final database to be used in the study, it was necessary to execute some aggregations and transformations in the data originating from the Redmine database. To do this, the following tasks were performed with the help of the Microsoft Excel Power Query tool:

- Creation of the function `FunctionWeekEndHolidayDays` to calculate the number of non-working days (weekends and holidays) between two dates.
- Execution of the procedure `CS-Data-Refined`: works the extracted data of Redmine to facilitate the handling of the information.
- Execution of the procedure `CS-DATA-HEADER-PROCEDURES`: prepares the data of the common steps in all phases.
- Execution of the query `CS-TIMES-FOR-RESPONSIBLE`: calculates the time spent by the software factory and by the Ministry of Health in each step.
- Execution of procedure `CS-CALCULATION-SLA`: uses the result of the previous procedures to assign data to each process and calculates SLAs, payment adjustments and glosses.

27.3.4.3 Final Database

The import of the database to the R required the conversion of attributes understood as numerical for categorical attributes. The final database contains the following attributes:

- `category_demand`: categorical variable that informs the demand category.There are two possible values:
  - `Development`: the demand will result in the creation of a new feature in existing functionality or a new functionality.
  - `Maintenance`: the demand will provide behavior adjustment in existing functionality.
- `artifact_developed`: categorical variable that informs the involved artifact, being able to be:
  - `Middleware`: includes the (a) Installation, administration, maintenance, updating, evolution, configuration, and migration of versions of both the middleware and the additional components associated with the SOA Bus; (b) Perform active monitoring of the middlewares and other components of the SOA Bus; and execute the necessary actions according to a plan of action negotiated with the contracting party; (c) Maintain solutions that use the middlewares and other components of the Ministry of Health Bus; and (d) Maintain documentation of the middleware and other components of the SOA Bus, as well as follow standards of service support documentation in force in MS.
  - `Service Units`: The term `Service Units` (service units) should be understood as being each service operation exposed in the National Health Bus, that is, to each function or capacity exposed through an SOA service contract on the web.
- `development_new`: categorical variable that indicates whether it is the development of a new artifact (value 0) or maintenance in an existing artifact (value 1). This variable should not be confused with `category_demand`, as there are new developments that can be made in existing artifacts. In the same way, there may be maintenance demands that generate new artifacts.
- `item_catalog`: categorical variable that informs the service catalog item used as the basis for the creation of demands.
- `points_function`: categorical variable that informs the measured quantity of function points.
- `interruption_server`: categorical variable that indicates the need to interrupt (value 1) or not (value 0) the production server to deploy the new artifact.
- `month_open`: categorical variable that informs the month of demand opening.
- `days_sla`: categorical variable that informs the limit in working days for the conclusion of the demand by the software factory.
- `days_factory`: continuous variable that informs the amount of business days that the software factory took to have the delivery accepted definitively by the Ministry of Health. This data does not exist in the Redmine database and needs to be calculated through a specific function, which considers weekends and the holidays database to set the number of business days between the start and end dates of each activity.
- `days_ms`: continuous variable that informs the amount of working days that the Ministry of Health took to approve the delivery. The procedure for obtaining this data is the same as described in `days_factory`.
- `days_execution`: continuous variable that informs the total amount of working days spent for the completion of the demand, from the initial request to the delivery in production. Basically, it is the sum of `days_factory` and `days_ms`.
- `delay_factory`: continuous variable that informs the number of business days beyond the SLA that the software factory took to have its delivery accepted definitively by the Ministry of Health. This is the variable we will use to train the predictive model.

- `perc_glosa`: categorical variable that informs the percentage of the value of the demand that should be glossed for non-compliance with the SLA. It can assume 3 values:
  - 0: the SLA was met or the delay was less than 10% of the deadline.
  - 1: there was a delay between 10 and 20% of the time frame defined in the SLA.
  - 10: delay was greater than 20% of the time limit defined in the SLA.

27.3.4.4 Data Correlation Analysis: Preparation of Numerical Data to Obtain the Pearson Coefficient

We used Pearson's correlation coefficients to analyze the relationships between each of the variables and the delay in the delivery of demand. The purpose of this step was to identify variables significantly related to the delay. This gives us indications of which variables will be useful predictors for the possibility of delay. In order to analyze the correlation between the variables using this method, it was necessary to create numerical dummy variables from the following categorical variables: category_demand, artifact_involved, development_new, interrupt_server, days_sla, days_factory, days_ms and delay.

27.3.4.5 Correlation Analysis Using the Pearson Coefficient

The correlation graph using the Pearson Coefficient was constructed using R [30] and the result is seen in Fig. 27.1. The color of each number indicates whether the correlation is positive (blue) or negative (red). The value indicates the degree of correlation, with values ranging from −1 (no correlation) to 1 (full correlation).

According to the diagram, the variables that are most related to the delay in the execution of the demands are: `artifact_involved`, `days_sla`, `days_factory` and `days_ms`. These correlations were discussed with the experts, who confirmed the results found. The definition should be the best possible, to ensure greater accuracy of the model.

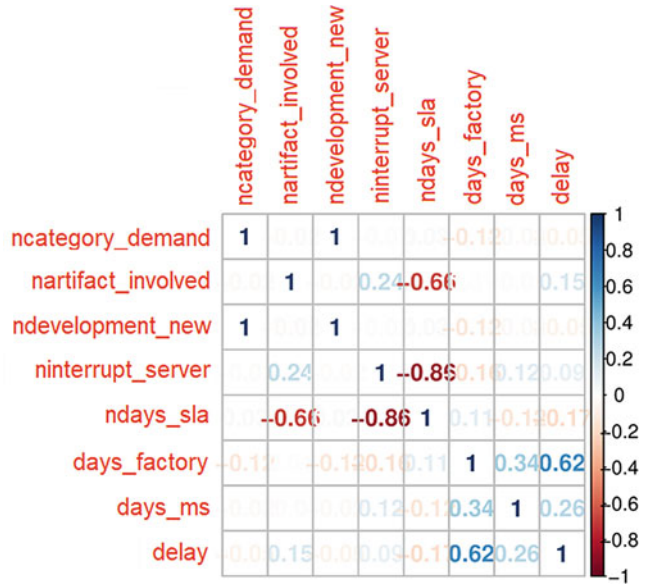

Fig. 27.1 Correlation diagram

27.3.5 Modeling

To answer the proposed problem, the GLM (Generalized Linear Modeling) and GBM (Gradient Boosting Method) regression algorithms were used to construct the models. The database was divided into two parts, used for model training (70%) and validation (30%). The construction of the models was made using the platform H2O.ai [14] integrated to R [28]. The parameters used for creating the models was:

- GLM: this algorithm was implemented with tenfold cross validation. The use of this validation technique was reproduced in other algorithms used. The GLM algorithm obtained AUC = 0.9772727 with tenfold cross-validation.

```
glm.model <- h2o.glm(myX, myY,
                     training_frame = dados.train,
                     validation_frame = dados.valid,
                     family = "binomial",
                     nfolds = 10,
                     alpha = 0.005,
                     model_id = "glm_alex")
```

- GBM: the first results showed that this algorithm had lower performance than the previous one. Then, a grid of parameters was chosen and applied to it. The parameters used were:
  - maximum trees: 100, 500 and 1000.
  - maximum depth: 5, 7 and 10.
  - stopping tolerance: 0.001

  The best model was built with tenfold cross validation, maximum 500 trees and 7 maximum depth of them. The GBM algorithm obtained AUC = 0.9924242 with tenfold cross-validation.

```
gbm.model <- h2o.glm(myX, myY,
                     training_frame = dados.train,
                     validation_frame = dados.valid,
                     family = "binomial",
                     model_id = "gbm_alex")
```

27.3.6 Evaluation

This section shows the evaluation of the models created and discusses the alternatives that best meet the defined objectives. To evaluate the models with the specialists, the following indicators were used:

- Confusion Matrix: table that allows visualization of the performance of an unsupervised learning algorithm. The rows represents instances in a predicted class while each column represents the instances in an actual class (or vice versa). This representation makes it easy to see if the system is confusing two classes (Fig. 27.2).
- MSE (Mean Square Error) is a measure of the quality of an estimator, always non-negative, and values closer to zero are better.
- RMSE (Root Mean Square Error): like MSE, RMSE is always non-negative, and a value of 0 (almost never achieved in practice) would indicate a perfect fit to the data. In general, a lower RMSD is better than a higher one.
- AUC (Area under the ROC curve): invariant metric scale used to compare predictive models, regardless of the classification threshold. A model whose predictions are 100% wrong has an AUC of 0, while a model whose predictions are 100% correct has an AUC of 1.
- R^2: value directly proportional to the number of predictors of a model, being more useful in comparing models of the same size. The higher the value of R^2 (always between 0 and 100%) the better the model adjusts its data.
- LogLoss: indicates the percentage of reliability of a classification and evaluates predictions of the probability of an entry belonging to a particular class, ranging from 0 to 1. Because it is a measure of loss, smaller values are better and 0 indicates a perfect error value.

The measures of accuracy of the models prove that the model based on the GBM algorithm surpassed GLM in all aspects. So, based on these metrics and the obtained results

GLM

	0	1	Error	Rate	Recall
0	109	3	0.0268	3 / 112	0.98
1	2	16	0.1111	2 / 18	0.84
Total	111	19	0.0385	5 / 130	
Precision	0.97	0.89			

GBM

	0	1	Error	Rate	Recall
0	108	4	0.0357	4 / 112	0.99
1	1	17	0.0556	1 / 18	0.81
Total	109	21	0.0385	5 / 130	
Precision	0.96	0.94			

Fig. 27.2 Confusion matrix

Model	MSE	RMSE	AUC	R^2	LogLoss
GLM	0.0930291	0.3050067	0.9772727	0.0351757	0.3305058
GBM	0.0491163	0.2216221	0.9924242	0.490604	0.1756327

Fig. 27.3 Accuracy indicators

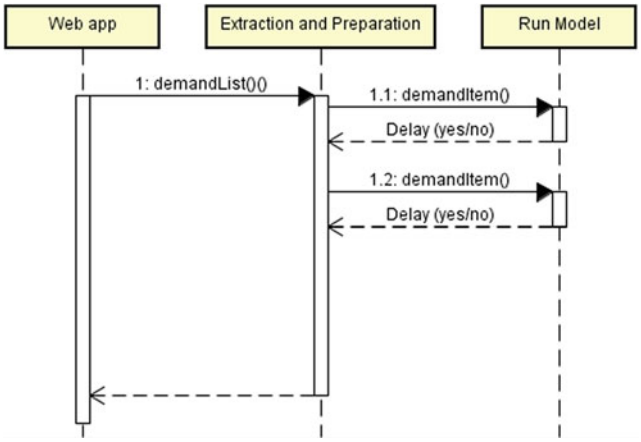

Fig. 27.4 Model incorporated to Redmine

(shown in Fig. 27.3), those involved in this study agreed that the GBM model best predicts the possibility of delay in the execution of contractual demands.

27.4 Deployment

In this section we present the form of introduction of the model obtained in the routine work of the contractual inspection teams and the strategies to guarantee its continuous updating. The implementation of the model will be done through the incorporation of Java classes to Redmine. These classes will extract new demands data and submit them to the predictor model. Figure 27.4 displays this planning in the form of a UML sequence diagram. The class responsible for model execution will be constructed based on the POJO [20] class generated by H2O.ai [14].

Considering that the model was built with a small amount of data, as already explained due to the short contractual execution time, monthly revisions are planned in the first 3 months of use. Then, four more bimonthly reviews will be made and, at the end, two more semi-annual reviews. In this way, it is expected to guarantee the continuous improvement of the model and a greater assertiveness in the inspection activities. Functioning of the built-in model to Redmine:

- The web app will be Redmine, the software used by Datasus for demand management. Its behavior will be adjusted to incorporate the necessary steps to use the proposed model and display them to the users in an organic way.
- The preparation and extraction steps will be triggered after events of demands inclusion or update. Instead of the mechanisms used in this study, the procedures will be incorporated into Java routines to ensure code and environment homogeneity.

- The final step will be to submit the data to the model, which will predict the delay and return a binary response.
- The result will be stored in the Redmine database and displayed in queries, listings, and view screens for each demand.

27.5 Results

The results allowed the inspection team to direct their actions to the demands predicted by the model as prone to delay. The overall assessment is that the oversight will be optimized. By incorporating the model to Redmine, as presented in Sect. 27.4, there will be more fluidity to the inspection process, in addition to dynamic updating and visualization of real-time predictions, minimizing the possibility of delays.

27.5.1 RQ.1. Is It Possible to Create a Predictive Model to Indicate the Possibility of Delay in the Execution of Software Construction Demands for the Brazilian National Health Bus?

The main objective of this work is to create a predictive model for the possibility of delay in the execution of demands, with a success rate greater than 75%. In this sense, the data available for analysis were extracted from Redmine and treated to generate a database to be submitted to GLM and GBM based algorithms.

The generated models were compared from six different metrics (confusion matrix, MSE, RMSE, AUC, R2 and LogLoss) and the model generated by the GBM algorithm proved to be the most suitable for the purposes of this study. However, it is important to point out that the initial choice of the algorithm was made from the results obtained with a small set of data. As more demands are available for analysis, the results may confirm or change the choice. Other variables may interfere in the choice of the algorithm, such as cost and complexity of the demands, as well as experience of developers team, which will be evaluated for the possibility of incorporation into the model, but the involved agreed that the priority is the team's perception about the possibility of using predictive models to increase their tasks efficiency. Thus, regarding research question **RQ.1** this study proved that it is feasible to create a model to predict the possibility of delay in the execution of the demands. In spite of the inherent fragility due to the small amount of data, the initiative was considered positive by the Administration of the Ministry of Health and new studies are being planned.

It is well understood that the use of data mining can bring greater objectivity to contractual management activities, increasing the security of the contracting parties. Another understanding is that automated prediction through machine learning algorithms can minimize the effects of the scarcity of human resources to manage high cost contracts with great relevance to Brazilian health.

27.5.2 RQ.2. How Can Such a Model Be Incorporated into the Contractual Management to Positively Impact on the Monitoring of Contract Delivery Times?

Another objective of this study was to establish a way to integrate the model generated here in the work process of contractual inspection teams, giving greater effectiveness to their tasks. Alternatives of integration to the demand management system in use in the organization were evaluated, maintaining the premise of generating little or no impact on stakeholders. It is hoped that this approach will minimize possible resistance and facilitate the incorporation of the model into daily activities, thus favoring other initiatives similar to this in other outsourcing contracts of the Brazilian Ministry of Health.

Thus, the answer to this question **RQ.2** considered the need to incorporate this work's results into the contractual demand management system, integrating the prediction of delays to the interfaces used to record and monitor the demands. This concern aims to reduce the natural resistance in the use of new technologies, even among software development teams. The implementation of the model could be done in different ways, for example an independent web application, but this would require the inspection teams to access another environment, different from the one used for recording and monitoring the demands. The choice of how to use the model should consider its accessibility by the people, being preferred those that are more integrated to the usual processes of work. For this reason, Java classes will be built and incorporated into Redmine, which will be seen by users only as additional features to the existing ones, respecting the patterns of visual behavior and navigation in use.

27.6 Conclusion

In this document we focus on the challenge of seeking mechanisms to expedite the control of IT contracts in the Ministry of Health of Brazil. To solve this problem, we proposed a regression-based model to predict delays in the execution of demands related to the National Health Bus. The proposal is applicable in similar situations and serves to give greater efficiency to the process of supervision of outsourcing contracts, using data collected from the demand

execution flow and creating reusable data analysis models. The methodology used is based on the CRISP-DM and uses a dataset obtained from a repository of the Datasus. The CRISP-DM reference model for data mining provides an overview of the life cycle of a data mining project. It contains the phases of a project, their respective tasks, and their outputs. The research questions that were defined were answered with the execution of this work.

Following CRISP-DM, domain data was collected, transformed and evaluated on an open source platform. Afterwards, the optimization problem was formulated and the data analysis plan was executed and implemented. The application (Fig. 27.4) can be internal or external to Redmine. The present work has its limitations. For example, the number of demands available for the preparation of the study and the institutional experience in the definition of aspects relevant to the analysis. The results of optimization can be improved by more accurate estimates of these factors, but efforts of both the research group and the audit team are needed.

For future work, it is desired that data analysis processes become routine in the organization. Efforts in the development and use of prescriptive analysis models, as described in this document, together with the description of predictive analytical models will invariably improve the future application of data analysis.

The adoption of data mining tools for contractual management in the Ministry of Health is in the embryonic stage. Its use can be expanded to other contracts (help-desk, mobile development, functional size counting) and domains (benchmarking of contractual execution, lead time of development processes, consumption of computing resources), increasing the objectivity of IT contracting. This study proved that it is possible to create and apply predictive models for the management of software factory contracts. More than that, it provoked a reflection on the use of data mining as a tool to support strategic IT management in the Brazilian Ministry of Health.

Acknowledgments This research work has the support of the Research Support Foundation of the Federal District (FAPDF) research grant 05/2018.

References

1. BRASIL: Instrucao Normativa MP/SLTI N° 4/2014. Ministério do Planejamento (2014)
2. Caetano, N., Cortez, P., Laureano, R.M.S.: Using data mining for prediction of hospital length of stay: an application of the CRISP-DM methodology. In: ICEIS (Revised Selected Papers). Lecture Notes in Business Information Processing, vol. 227, pp. 149–166. Springer, Berlin, (2014). https://doi.org/10.1007/978-3-319-22348-3_9
3. Chaim, R.M., Oliveira, E.C., Araujo, A.P.F.: Technical specifications of a service-oriented architecture for semantic interoperability of eh—electronic health records. In: 2017 12th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–6. IEEE, Piscataway (2017). https://doi.org/10.23919/CISTI.2017.7975923
4. Clara, A.M.C., Canedo, E.D., de Sousa Júnior, R.T.: Elements that orient the regulatory compliance verification audits on ICT governance. In: Proceedings of the 18th Annual International Conference on Digital Government Research, pp. 177–184. ACM, New York (2017) https://doi.org/10.3233/IP-170059
5. Clara, A.M.C., Canedo, E.D., de Sousa Júnior, R.T.: A synthesis of common guidelines for regulatory compliance verification in the context of ICT governance audits. Inf. Polity **23**(2), 221–237 (2018)
6. da Cruz, C.S., de Andrade, E.L.P., Figueiredo, R.M.D.C.: Processo de contratação de software e serviços correlatos para entes governamentais. Rev. Program. Bras. Qual. Prod. Softw. **1**, 103–110 (2010)
7. de Freitas, S.A.A., Canedo, E.D., Felisdório, R.C.S., Leão, H.A.T.: Analysis of the risk management process on the development of the public sector information technology master plan. Information **9**(10), 248 (2018)
8. Federal Court of Accounts of Brazil (TCU): Get.it: governance evaluation techniques for information technology: a WGITA guide for supreme audit institutions. In: International Organization of Supreme Audit Institutions (INTOSAI). Working Group of Information Technology (WGITA) vol. 1, pp. 1–136 (2016)
9. Finlay, J., Pears, R., Connor, A.M.: Data stream mining for predicting software build outcomes using source code metrics. Inf. Softw. Technol. **56**(2), 183–198 (2014)
10. França, A., da C. Figueiredo, R. M., Venson, E., and Silva, W.: Storytelling on the implementation of a decentralized model for software development in a Brazilian government body. In: Proceedings of the Seventh International Digital Government Research Conference, pp. 388–396 ACM, New York (2016). https://doi.org/10.1145/2912160.2912201
11. Goby, N., Brandt, T., Feuerriegel, S., Neumann, D., Research Goby, C.: Business Intelligence for Business Processes: The Case of IT Incident Management. In: Proceedings of the European Conference on Information Systems (ECIS), pp. 1–15 (2016) https://doi.org/10.13140/RG.2.1.2033.9604
12. Granja, T.H.M., da Costa Figueiredo, R.M., Canedo, E.D.: Management tool for software factory contracts for a Brazilian public agency. In: AMCIS, pp. 1–8. Association for Information Systems, Atlanta (2017). https://aisel.aisnet.org/amcis2017/SystemsAnalysis/Presentations/9/.
13. Guarda, G.F., Oliveira, E.C., Sousa Júnior, R.T.D.: Analisys of IT outsourcing contracts at the TCU (Federal Court of Accounts) and of the legislation that governs these contracts in the Brazilian federal public administration. J. Inf. Syst. Technol. Manag. **12**(1), 81–106 (2015)
14. H2O: Welcome to H2O3. H2O (2019)
15. Hayn, D., Veeranki, S., Kropf, M., Eggerth, A., Kreiner, K., Kramer, D., Schreier, G.: Predictive analytics for data driven decision support in health and care. Inf. Technol. **60**(4), 183–194 (2018)
16. Kamei, F., Pinto, G., Cartaxo, B., Vasconcelos, A.: On the benefits/limitations of agile software development: an interview study with Brazilian companies. In: Proceedings of the 21st Evaluation and Assessment in Software Engineering Conference (EASE), pp. 154–159. ACM, New York (2017). https://doi.org/10.1145/3084226.3084278
17. Leão, H.A.T., Canedo, E.D., Ladeira, M., Fagundes, F.: Mining enade data from the ulbra network institution. In: Information Technology-New Generations, pp. 287–294. Springer, Berlin (2018). https://doi.org/10.1007/978-3-319-77028-4_39
18. Linden, R., Schmidt, N., Rosenkranz, C.: The changing role of advisory services in information technology outsourcing. In: ICIS, pp. 1–8. Association for Information Systems (2018). https://aisel.aisnet.org/icis2018/management/Presentations/1/
19. Nabati, E.G., Thoben, K.: On applicability of big data analytics in the closed-loop product lifecycle: Integration of CRISP-DM stan-

dard. In: PLM. IFIP Advances in Information and Communication Technology, vol. 492, pp. 457–467. Springer, Berlin (2016). https://doi.org/10.1007/978-3-319-54660-5_41
20. Oracle: IBM SPSS Modeler CRISP-DM Guide. Oracle (2011)
21. Piad-Morffis, A., Gutiérrez, Y., Muñoz, R.: A corpus to support ehealth knowledge discovery technologies. J. Biomed. Inf. **94**, 1–12 (2019)
22. Pulparambil, S., Baghdadi, Y.: Service oriented architecture maturity models: a systematic literature review. Comput. Stand. Interfaces **61**, 65–76 (2019)
23. Sarnovsky, M., Surma, J.: Predictive models for support of incident management process in IT management. Acta Electrotech. Inf. **18**(1), 57–62 (2018)
24. Sastry, S.H.: Implementation of CRISP methodology for ERP systems. Int. J. Comput. Sci. Eng. **2**(5), 203–217 (2013)
25. Sharma, V., Stranieri, A., Ugon, J., Vamplew, P., Martin, L.: An agile group aware process beyond CRISP-DM: a hospital data mining case study. In: ICCDA, pp. 109–113. ACM, New York (2017). https://doi.org/10.1145/3093241.3093273
26. Spring: Understanding POJOs. Spring (2018)
27. Sun, K., Li, Y., Roy, U.: A PLM-based data analytics approach for improving product development lead time in an engineer-to-order manufacturing firm. Math. Model. Eng. Probl. **4**(2), 69–74 (2017)
28. Team, R.C., et al.: R: a language and environment for statistical computing. Citeseer **1**, 1–114 (2013)
29. Trinkenreich, B., Santos, G., Barcellos, M.P.: SINIS: a GQM+strategies-based approach for identifying goals, strategies and indicators for IT services. Inf. Softw. Technol. **100**, 147–164 (2018)
30. Verzani, J.: Using R for Introductory Statistics. Chapman and Hall/CRC (2014). https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf
31. Yuvaraj, N., SriPreethaa, K.R.: Diabetes prediction in healthcare systems using machine learning algorithms on hadoop cluster. Cluster Comput. **22**(Suppl 1), 1–9 (2019)