Using Data Mining for Prediction of Hospital Length of Stay: An Application of the CRISP-DM Methodology

Nuno Caetano[1], Paulo Cortez[2], and Raul M. S. Laureano[3]

[1] HFAR - Hospital das Foras Armadas, Azinhaga Ulmeiros,
1620-060 Lisboa, Portugal
nmcaetano@gmail.com
[2] ALGORITMI Research Centre, Department of Information Systems,
University of Minho,
4800-058 Guimarães, Portugal
pcortez@dsi.uminho.pt
WWW home page: http://www3.dsi.uminho.pt/pcortez
[3] Business Research Unit (UNIDE-IUL),
Instituto Universitário de Lisboa (ISCTE-IUL),
Av. das Forças Armadas, 1629-026 Lisboa, Portugal
raul.laureano@iscte.pt

Abstract Hospitals are nowadays collecting vast amounts of data related with patient records. All this data hold valuable knowledge that can be used to improve hospital decision making. Data mining techniques aim precisely at the extraction of useful knowledge from raw data. This work describes an implementation of a medical data mining project approach based on the CRISP-DM methodology. Recent real-world data, from 2000 to 2013, were collected from a Portuguese hospital and related with inpatient hospitalization. The goal was to predict generic hospital Length Of Stay based on indicators that are commonly available at the hospitalization process (e.g., gender, age, episode type, medical specialty). At the data preparation stage, the data were cleaned and variables were selected and transformed, leading to 14 inputs. Next, at the modeling stage, a regression approach was adopted, where six learning methods were compared: Average Prediction, Multiple Regression, Decision Tree, Artificial Neural Network ensemble, Support Vector Machine and Random Forest. The best learning model was obtained by the Random Forest method, which presents a high quality coefficient of determination value (0.81). This model was then opened by using a sensitivity analysis procedure that revealed three influential input attributes: the hospital episode type, the physical service where the patient is hospitalized and the associated medical specialty. Such extracted knowledge confirmed that the obtained predictive model is credible and with potential value for supporting decisions of hospital managers.

Key words: medical data mining, hospitalization process, length of stay, CRISP-DM, regression, random forest.

1 Introduction

In recent decades, hospitals have been collecting large amounts of data into their clinical information systems. All this data hold valuable knowledge and therefore there is an increasing potential of the use of Data Mining (DM) [1], to facilitate the extraction of useful knowledge and support clinical decision making, in what is known as medical data mining [2]. There are several successful medical data mining applications, such as the prediction of mortality [3] and degree of organ failure [4] at Intensive Care Units, and the segmentation of tissue from magnetic resonance imaging [5], among others.

This work focuses on the prediction of the Length Of Stay (LOS), defined in terms of the inpatient days, which are computed by subtracting the day of admission from the day of discharge. Extreme LOS values are known as prolonged LOS and are responsible for a major share in the hospitalization total days and costs. The use of data-driven models for predicting LOS is of value for hospital management. For example, with an accurate estimate of the patients LOS, the hospital can better plan the management of available beds, leading to a more efficient use of resources by providing a higher average occupancy and less waste of hospital resources [6, 7].

DM aims at the extraction of useful knowledge from raw data [1]. With the growth of the field of DM, several DM methodologies were proposed to systematize the discovery of knowledge from data, including the tool neutral and popular Cross-Industry Standard Process for Data Mining (CRISP-DM) [8], which is adopted in this work. The methodology is composed of six stages: business understanding, data understanding, data preparation, modeling, evaluation and implementation.

This study describes the adopted DM approach under the first five stages of CRISP-DM, given that implementation is left for future work. The main goal was to predict generic LOS (for all hospital services) under a regression approach using past patterns existing in the hospitalization process, based on a DM techniques. The data is related with a Portuguese hospital, based on recent data collected from the hospitalization process between 2000 and 2013, including a total of 26462 records from 15253 patients. At the preprocessing stage, the data were cleaned and attributes were selected, leading to 14 inputs and the LOS target. During the modeling stage, six regression methods were tested and compared: Average Prediction (AP), Multiple Regression (MR), Decision Tree (DT) and state-of-the-art regression methods [9], including an Artificial Neural Network (ANN) ensemble, Support Vector Machine (SVM) and Random Forest (RF). The predictive models were compared using a cross-validation procedure with three popular regression metrics: coefficient of determination (R^2), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Moreover, the best predictive model (RF) was opened using a sensitivity analysis procedure [10] that allows ranking the input attributes and also measuring the average effect of a particular input in the predictive response.

This work is organized as follows. Firstly, the relevant related work is presented (Section 2). Then, the adopted DM approach is detailed in terms of the

CRISP-DM methodology first five phases (Section 3). Finally, closing conclusions are drawn (Section 4).

2 Related Work

Nowadays, hospital managers are pressured to accomplish several goals, such as providing better health care, increasing the number of available beds for new admissions and reduce surgical waiting lists. Under this context, LOS is used worldwide as a highly relevant measure to analyze the hospital resources consumption and to monitor hospital performance [7]. Given the importance of LOS, a large number of studies have adopted a data-driven approach for modeling LOS. In the next few paragraphs, we present some examples of related studies.

In 1998, Merom et al. [11] estimated the rate of inappropriate hospital days (failure of established criteria for admission) and the identification of the variables associated with this impropriety. During such study, 1369 patients from 24 hospitals were analyzed under a multiple regression model. Several attributes were used in their analysis: occupation, group age, gender, inappropriate days, government, another hospital entity, another diagnosis, origin, admission diagnosis and period of stay.

In 2007, Abelha et al. [12] evaluated LOS of patients submitted to a noncardiac surgery and admitted to a surgical Intensive Care Unit (ICU) between October 2004 and July 2005. The attributes used to categorize patients were: age, gender, body mass index, physical status, type and magnitude of surgery, type and duration of anesthesia, temperature on admission, LOS in the ICU and in hospital mortality in the ICU and hospital. A simple linear regression model was adopted and from the results it was found that the average LOS was 4.22±8.76 days.

In 2010, Oliveira et al. [13], proposed to evaluate factors associated with higher mortality and prolonged LOS in ICUs. The study included 401 patients consecutively admitted to the ICU, within a six-month period. The collected attributes were: gender, age, diagnosis, personal history, APACHE II score, mechanical ventilation days, endotracheal reintubation, tracheostomy, LOS in the ICU, ICU discharge or death. In terms of results, the average LOS in the ICU was 8.2±10.8 days. The study concluded that factors such as APACHE II, reintubation and tracheostomy were associated with higher mortality and prolonged LOS in the ICU.

Also in 2010, Karla et al. [14] studied the temporal trends of the workflow in the internal medicine service of an University Hospital. The data analyzed were obtained in that service for three different time periods spanning through 13 years. The most relevant data features data were: date of admission, date of departure or death, gender, age, residence code, financial entity and primary diagnosis. Their results have confirmed several changes in LOS behavior through time (e.g., the number of admissions in the internal medicine service statistically increased from 1991 to 2004).

More recently, Freitas et al. [15] analyzed in 2012 LOS outliers based on inpatient episodes of Portuguese public hospitals belonging to the national health system, with data collected between 2000 and 2009. The variables used for analysis were: age, distance from residence to hospital, year of discharge, comorbidities,A-DRG complexity, readmission, admission and DRG type, discharge status and hospital type. In the analysis they used logistic regression models to examine the association of each variable with the time of admissions outliers, and model with all variables to calculate the adjusted odds ratios and respective confidence intervals (95%). In terms of results, nine million inpatient episodes were analyzed, of which 3.9% were considered high LOS outliers. They concluded that age, type of admission and hospital type attributes were significantly associated with high LOS outliers.

In the same year (2012), Azari et al. [6] explored a classification approach to predict LOS. The main attributes of their analysis were: specialty services, days elapsed since the first act of the year, primary condition group (generalized code for the principal diagnosis) and Charlson index (diagnostic code) and LOS. The LOS was divided in three different classification groups: one to two days, greater than two and less than seven days, and longer or equal to seven days. The study concluded that the performance of classification techniques could be improved by incorporating a clustering step during the training stage.

Also in 2012, Castillo [7] developed a statistical model to predict the LOS in Mexican public hospitals. The following attributes were used: age, gender, occupation, education level, previous visits, origin, surgical first diagnosis, diagnosis, surgical procedure, number of surgical procedures and ward. The best predictive model was given by a probabilistic model based on a cluster analysis.

Finally, Sheikh-Nia in 2012 [16] used a sequential ensemble of classification algorithms to predict LOS of patients in the next year, based on the patient previous medical history. The main attributes considered were: age at first claim, gender, provider, year, medical specialty, number of days from the first record, primary condition group, Charlson index and LOS. The results showed that all of the independent classifiers exceeded the baseline by a factor of 1.78 for the ANN, 1.20 for K-Nearest Neighbor and 1.17 for DT.

Instead of predicting LOS in specialized medical services, as in UCI [12, 13] or internal medicine [14], in this work we predict generic LOS, for all hospital services (e.g., internal medicine, general surgery, pneumology), which is a more challenging task. Also, as a case study, we only analyze data from one hospital. Nevertheless, we approach a much larger dataset (with 26462 records collected from 2000 to 2013), when compared with the datasets used by some of the mentioned works (e.g., Merom et al. [11] included data from 1369 patients and Oliveira et al. [13] analyzed only 401 records). In addition, the attributes that we adopt (described in Section 3) were defined by a hospital expert's medical panel and are commonly available at the hospitalization process. Most of the proposed attributes (e.g., gender, age, episode type, medical specialty) were also adopted by the literature (as shown in Section 3.3). Moreover, in contrast with several literature works [6, 7, 16], we do not perform a classification task, which requires defining *a priori* which are the interesting LOS class intervals. Instead, we adopt the more informative pure regression approach, which predicts the actual number of LOS days and not classes.

3 CRISP-DM Methodology

In this section, we describe the main procedures and decisions performed when following the first five phases of the CRISP-DM methodology for LOS prediction of a Portuguese hospital, namely: business understanding (Section 3.1), data understanding (Section 3.2), data preparation (Section 3.3), modeling (Section 3.4) and evaluation (Section 3.5).

3.1 Business Understanding

The prediction of LOS is inserted within the wider problem of hospital admission scheduling, where there is a pressure to increase the availability of beds for new patients. In this particular Hospital, most patients come from the emergency department and from the region of Lisbon. The goal was set in terms of predicting LOS using regression models, thus favoring predictions that are closer to the target values. As a baseline business objective (to determine if there is success), we defined a coefficient of determination with a minimum value of R^2=0.6, which often corresponds to a reasonable regression.

In terms of software, we adopted open source tools, using structured query language (SQL) to extract data from the hospital database and the **R** tool for the data analysis (`http://www.r-project-org`). In particular, we adopt the **rminer** package [17], for applying the DM regression models (i.e., AP, MR, DT, ANN, SVM and RF) and sensitive analysis methods.

3.2 Data Understanding

The data was collected between October 2000 and March 2013. During this period, a total of 26462 inpatient episodes were stored, related with 15253 patients and associated with the distinct hospital medical specialties.

The selection of relevant data attributes for LOS prediction was performed by an expert medical panel. The panel was composed with 9 physicians from different medical specialties (e.g., internal medicine, general surgery, gynecology). The panel presented a total of 28 attributes that were considered related with LOS and that were analyzed in the data preparation phase (Table 1). The first seven rows of Table 1 are related with the patient's characteristics while the remaining rows are related with the inpatient clinical process. The description column of the table contains in brackets the attribute type (date, nominal, ordinal or numeric), as found in the original hospital database.

Table 1. List of attributes related with LOS prediction (attributes used by the regression models are in **bold**).

Name	**Description (attribute type)**
	Patient Characteristics:
Sex	Patient gender (nominal)
Date of Birth	Date of birth (date)
Age	Age at the time of admission (numeric)
Country	Residence country (nominal)
Residence	Place of residence (nominal)
Education	Educational attainment (ordinal)
Marital Status	Marital status (nominal)
	Inpatient clinical process:
Initial Diagnosis	Initial diagnosis description (ordinal)
Episode Type	Patient type of episode (nominal)
Inpatient Service	Physical inpatient service (nominal)
Medical Specialty	Patient medical specialty (nominal)
Origin Episode Type	Origin episode type of hospitalization (nominal)
Admission Request Date	Date for hospitalization admission request (date)
Admission Date	Hospital admission date (date)
Admission Year	Hospital admission year (ordinal)
Admission Month	Hospital admission month (ordinal)
Admission Day	Hospital admission day of week (ordinal)
Admission Hour	Hospital admission hour (date)
Main Procedure	Main procedure description (nominal)
Main Diagnosis	Main diagnosis description (ordinal)
Physician ID	Identification of the physician responsible for the internment (nominal)
Discharge Destination	Patient destination after hospital discharge (nominal)
Discharge Date	Hospital discharge date (date)
Discharge Hour	Hospital discharge hour (date)
GDH	Homogeneous group diagnosis code (numeric)
Treatment	Clinic codification for procedures, treatments and diseases (ordinal)
GCD	Great diagnostic category (ordinal)
Previous Admissions	Number of previous patient admissions (numeric)
	Target attribute:
LOS	Length Of Stay (numeric)

3.3 Data Preparation

In this phase, a substantial effort was performed using a semi-automated approach to preprocess the data. In particular, the **R** tool was adopted to perform an exploratory data analysis (e.g., histograms and box plots) and preprocess the original dataset. The processing involved the operations of cleaning, discarding redundant attributes, handling missing values and attribute transformations.

During the exploratory data analysis step, a few outliers were first detected and then confirmed by the Physicians. The respective records were cleaned: one LOS with 2294, an age of 207 and 29 entries related with a virtual medical specialty, used only for testing the functionalities of the hospital database. After cleaning, the database contained 26431 records.

Then, fourteen attributes from Table 1 were discarded in the variable selection analysis step: Date of Birth (reason: reflected in Age); Country (99% patients were from Portugal); Residence (30% of missing values, very large number of nominal levels); Admission Request Date (48% of missing values, reflected in Admission Date); Admission Date (reflected in Admission Month, Day, Hour and LOS); admission year (not considered relevant); Physician ID (19% of missing values and large number of 156 nominal levels); Initial Diagnosis (63% of missing values); and attributes not known at the patient's hospital admission process (i.e., GDH, GCD, Treatment, Discharge Destination, Date and Hour). The remaining 14 attributes (**bold** in Table 1) were used as input variables of the regression models (Section 3.4). As shown in Table 2, all input attributes proposed in this study (except for Marital Status) were also used in previous works, which is a clear indication that the selected attributes (**bold** in Table 1) can have a potential predictive LOS value. In particular, there are three input variables (gender, age and main diagnosis) that were used in five or more studies.

Table 2. List of input attributes proposed in this work and that were also used in the literature.

Attribute Name	**Previous LOS studies that adopted this attribute**
Sex	[11] [12] [14] [13] [7] [16]
Age	[12] [14] [13] [15] [7] [16]
Education	[7]
Episode Type	[15] [7]
Inpatient Service	[7]
Medical Specialty	[6] [16]
Origin Episode Type	[11]
Admission Month	[14]
Admission Day	[14]
Admission Hour	[14]
Main Procedure	[12] [7]
Main Diagnosis	[11] [14] [13] [6] [7] [16]
Previous Admissions	[7]

Next, missing values were replaced by using the hotdeck method [18], which substitutes a missing value by the value found in the most similar case. In particular, the **rminer** package uses a 1-nearest neighbor applied over all attributes with full values to find the closest example [17]. The following attributes were affected by this operation: Education (11771 missing values), Marital Status

(10046 values), Main Procedure (19407 values) and Main Diagnosis (19268 values).

Finally, several attributes were transformed, to facilitate the modeling stage. To reduce skewness and improve symmetry of the underlying variable distribution, the logarithm transform y=ln(x+1) was applied to the Previous Admissions and LOS variables. This is a popular transformation that often improves regression results for right-skewed variables [19]. Also, the Admission Hour variable was standardized to include only 24 levels. Moreover, the values of nominal attributes with a large number of levels were recoded/standardized to reduce the number of levels: Education (transformed from 14 to 6 levels), Main Procedure (from hundreds of values to 16 levels) and Main Diagnosis (from hundreds to 19 levels). Finally, using medical knowledge, we transformed the Age numeric attribute into 5 ordinal classes: A - lower than 15 years; B - between 15 and 44; C - between 45 and 64; D - between 65 and 84; and E - equal or higher than 85.

3.4 Modeling

Due to its importance, in the last decades, several methods have been proposed for regression, such as DT, ANN, SVM and RF [9]. In this phase, we tested six regression methods, as implemented in the **rminer** package [17]: AP, MR, DT, ANN, SVM and RF.

The AP is a naive model that consists in predicting the same average LOS ($\overline{y}$, as found in the training set) and is used as baseline method for the comparison.

The DT is a branching structure that represents a set of rules, distinguishing values in a hierarchical form.

The MR is a classical statistical model defined by the equation:

$$\hat{y} = \beta_0 + \sum_{i=1}^{I} \beta_i x_i \tag{1}$$

where $\beta_0, \ldots, \beta_i$ are the set of parameters to be adjusted, usually by applying an ordinary least squares (OLS) algorithm.

ANN is based in the popular multilayer perceptron, with one hidden layer of H hidden nodes and logistic activation functions, while the output node uses the linear function. Since ANN training is not optimal, the final solution is dependent of the choice of starting weights. To solve this issue, **rminer** first trains N_r different networks and then uses an ensemble of these networks such that the final output is set in terms of the average of the distinct N_r individual predictions.

The SVM model performs a nonlinear transformation to the input space by adopting the popular Gaussian kernel. SVM regression is achieved under the commonly used ϵ-insensitive loss function. Under this setup, the SVM performance is affected by three parameters: γ – Gaussian kernel parameter; ϵ and C – a trade-off between fitting the errors and the flatness of the mapping. Finally, RF is an ensemble of T unpruned DT, where each tree is based on a random feature selection with up to m features from bootstrap training samples. The RF predictions are built by averaging the outputs of T trees. RF is a substantial modification of bagging (fit of several models to bootstrap samples of training data) and on many problems RF performance is similar to boosting, while being more simpler to train and tune [9].

The **rminer** package full implementation details can be found in [17]. Under this package, before fitting the MR, ANN and SVM models, the input data is first standardized to a zero mean and one standard deviation [9]. Except for the hyperparameters of the most complex methods (ANN, SVM and RF), **rminer** adopts the default parameters of the learning algorithms, such as: MR and ANN – BFGS algorithm, as implemented in **nnet** package; DT - CART algorithm, as implemented in the **rpart** package; SVM - sequential minimal optimization algorithm, as implemented in the **kernlab** package; and RF - Breiman's random forest algorithm, as implemented in the **randomForest** package.

In this work, we set $N_r = 3$ for the ANN ensemble. Also, heuristics were adopted to set two of the three SVM hyperparameters [17]: $C = 3$ (for standardized data) and $\epsilon = 3\sigma_y\sqrt{\log(N)/N}$, where σ_y denotes the standard deviation of the predictions given by a 3-nearest neighbor and N is the dataset size. For RF, we adopted the default $T = 500$ value. For the most complex methods, **rminer** uses grid search to select the best hyperparameter values: H for ANN, γ for SVM and m for RF. In this work, the grid method searches ten values for each hyperparameter ($H \in${0,1,...,9}; $\gamma \in \{2^{-15}, 2^{-13}, ..., 2^{3}\}$; and $m \in \{1, 2, ..., 10\}$). During the grid search, the absolute error is measured over a validation set (with 33% of the training data). The configuration that corresponds to the lowest validation error is selected. Finally, the selected model is retrained with all training data.

The method used for estimating the predictive performance of a model was a 5-fold cross-validation, which divides the data into 5 partitions of equal size. In each 5-fold iteration, a given subset is used as test set (to measure predictive capability) and the remaining data is used for training (to fit the model). To assure statistical robustness, 20 runs of this 5-fold procedure were applied to all methods. For demonstration purposes, we present here a portion of the R/rminer code used to test the RF model:

```
library(rminer) # load the library
# read the data:
d=read.table("data.csv",header=T,sep=",")
# execute 20 runs of 5-fold using RF:
M=mining(LOS~.,data=d,Runs=20, method=c("kfold",5),
         model="randomforest", search="heuristic10")
# save the results into a file:
savemining(M,"rf.results")
```

3.5 Evaluation

To evaluate the predictions, three regression metrics were selected, the coefficient of determination (R^2), Mean Absolute Error (MAE) and Root Mean Squared

Error (RMSE), which can be computed as [20]:

$$\begin{array}{l} R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \overline{y}_i)^2} \\ MAE = 1/N \times \sum_{i=1}^{N} |y_i - \hat{y}_i| \\ RMSE = \sqrt{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 / N} \end{array} \quad (2)$$

where N denotes the number of predictions (test set size), y_i is the target value for example i, $\overline{y}_i$ is the average of the target values in the test set and $\hat{y}_i$ is the predicted value for example i.

R^2 is a popular regression metric that is scale independent, the higher the better, with the ideal model presenting a value of 1.0. The lower the RMSE and MAE values, the better the predictions. When compared with MAE, RMSE is more sensitive to extreme errors. The Regression Error Characteristic (REC) curve is useful to compare several regression methods in a single graph [21]. The REC curve plots the error tolerance on the x-axis versus the percentage of points predicted within the tolerance on the y-axis.

Table 3 presents the regression predictive results, in terms of the average of the 20 runs of the 5-fold cross-validation evaluation scheme. From Table 3, it is clear that the best results were obtained by the RF model, which outperforms other DM models for all three error metrics. A pairwise t-student statistical test, with a 95% confidence level, was applied, confirming that the differences are significant (i.e., p-value<0.05) when comparing RF with other methods. We emphasize that a very good R^2 value was achieved (0.813), much higher than the minimum success value of 0.6 set in Section 3.1.

Table 3. Predictive results (average of 20 runs, as measured over test data; best values in **bold**).

	Metrics		
Method	$\mathbf{R^2}$	**MAE**	**RMSE**
AP	0.000	0.861	1.085
MR	0.641	0.446	0.650
DT	0.622	0.415	0.667
ANN	0.736	0.340	0.558
SVM	0.745	0.296	0.547
RF	**0.813***	**0.224***	**0.469***

* statistically significant under a pairwise comparison with other methods.

The REC analysis, shown in Figure 1, also confirms the RF as the best predictive model, presenting always a higher accuracy (y-axis) for any admitted absolute tolerance value (x-axis). For instance, for a tolerance of 0.5 (at the logarithm transform scale), the RF correctly predicts 85.4% of the test set examples. The REC results are further complemented in Table 4, which compares the accuracy of the best two models (RF and SVM) for eleven absolute deviation values within the range [0,1]. The table confirms the superiority of the RF model, which always presents higher accuracy values, with a difference that ranges from 2.1 percentage points (for a tolerance of 1.0) to 15.6 (for a tolerance of 0.0).

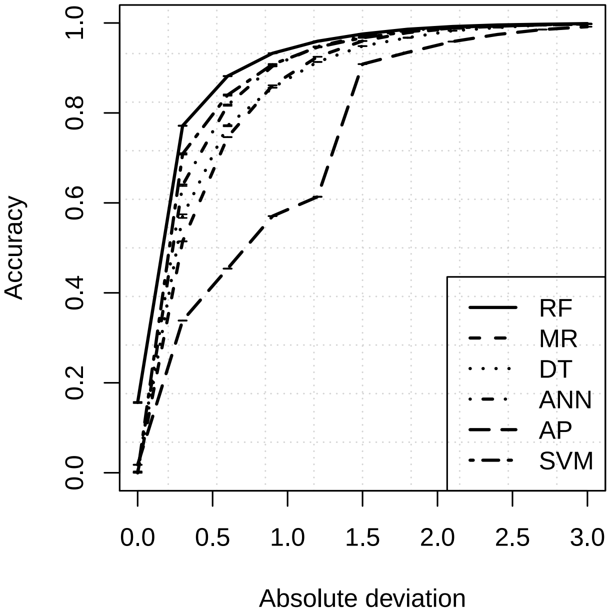

Fig. 1. REC curves for all tested models.

The quality of the predictions for the RF model can also be seen on Figure 2, which plots the observed (x-axis) versus the predicted values (y-axis). In the plot, values within the 0.5 tolerance are shown with solid circles (85.4% of the examples), values outside the tolerance range are plotted with the + symbol and the diagonal dashed line denotes the performance of the ideal prediction method. It should be noted that the observed (target) values do not cover the full space of LOS values, as shown in Figure 2. This is an interesting property of this problem domain that probably explains the improved performance of RF when compared with other methods, since ensemble methods (such as RF) tend to be useful when the sample data does not cover the tuple space properly. The large diversity of learners (i.e., T=500 unpruned trees) can minimize this issue, since each learner can specialize into a distinct region of the input space.

It should be noted that the presented predicted results were computed over the logarithm transform scale (see Section 3.3). In Figure 2 and within a 0.5 tolerance (solid circles), the predictions are above the origin point (point A, x=0) and below the right upper observed values (point B, x=4.2). This means that at the normal scale (x', using the inverse of the logarithm transform), the RF model error is capable of correctly predicting 85.4% of the examples with

Table 4. RF vs SVM accuracy for some absolute deviation values (average of 20 runs, best values in **bold**).

Absolute Deviation	SVM Accuracy	RF Accuracy
0.0	0.0%	**15.6**%
0.1	50.1%	**61.3**%
0.2	63.3%	**70.9**%
0.3	70.9%	**77.2**%
0.4	76.3%	**81.8**%
0.5	80.6%	**85.4**%
0.6	84.0%	**88.2**%
0.7	86.7%	**90.3**%
0.8	89.0%	**91.9**%
0.9	80.8%	**93.3**%
1.0	92.3%	**94.4**%

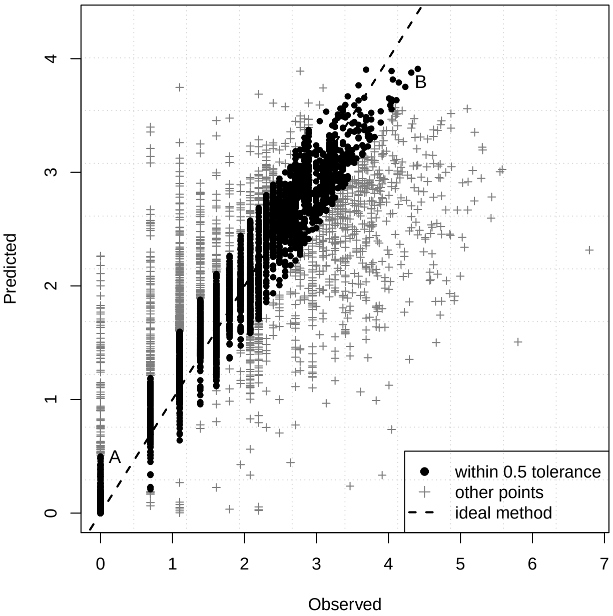

Fig. 2. Observed versus predicted RF values.

a real maximum error that ranges from 0.7 days (point A, x'=0) to 26.0 days (point B, x'=65.7 days).

When compared with DT and MR, the ANN, SVM and RF data-driven models are difficult to be interpreted by humans. Yet, sensitivity analysis and visualization techniques can be used to open these complex models [10]. The procedure works by analyzing the responses of a model when a given input is varied though its domain. By analyzing the sensitivity response changes, it is possible to measure input relevance (higher changes denote a more relevant input) and average impact of an input in the model. The former can be shown using an input importance bar plot and the latter by plotting a Variable Effect Characteristic (VEC) line curve or segments.

To extract explanatory knowledge from the RF model and open the black-box, we applied the Data-Based Sensitivity Analysis (DSA) method, as implemented in the *Importance* function of the **rminer** package. DSA has the advantage of being a fast method that can measure the overall influence of a particular input, including its iterations with other inputs (Cortez and Embrechts, 2013). The DSA algorithm was executed over the RF model fit with all data. The obtained sensitivity responses were first used to rank the RF inputs, according to their relevancy in the predictive model (Figure 3). Then, the average effects of the most relevant inputs were analyzed using VEC line segments (Figures 4, 5 and 6).

The input importance bar plot (Figure 3) ranks the Episode Type (30.1% impact) as the most relevant attribute, followed by Inpatient Service (12.3%) and Medical Specialty (10.1%). Overall, the bar plot shows a much greater influence of the inpatient clinical process attributes (e.g., Episode Type, Medical Specialty, Previous Admissions) when compared with the patients' characteristics (e.g., Education, Sex). This is an interesting outcome for hospital managers. In the next paragraphs, we detail the particular influence of the top three inputs by analyzing their VEC line segments.

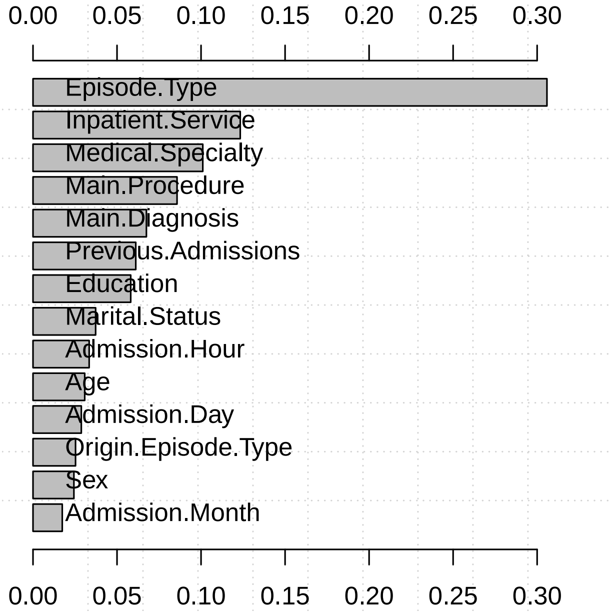

Fig. 3. Input importance bar plot for the RF model.

Figure 4 shows the global influence of the most relevant input (Episode Type), which is a nominal attribute with two classes. The VEC line segments clearly confirm that the ambulatory type (scheduled admission, typically involving a 1 day LOS) is related with an average lower LOS (0.1 in the logarithm transform scale, 0.1 days in the normal scale) when compared with the internment type (1.58 in the logarithm scale, 3.9 days).

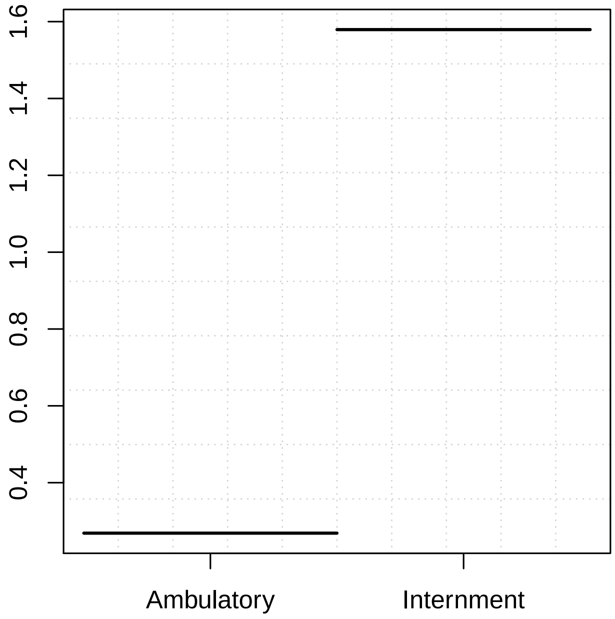

Fig. 4. VEC line segments, showing the average influence of the Episode Type (x-axis) on the RF model output (y-axis).

Next, we analyze the average influence of the Inpatient service (Figure 5). The greatest LOS is associated with five services: medicine, average LOS of 1.45, corresponding to 3.3 days at the normal scale; orthopedics, average of 1.39, corresponding to 3.0 days; specialties, average of 1.37, corresponding to 2.9 days; surgery, average of 1.36, corresponding to 2.9 days; and pneumology, average of 1.32, corresponding to 2.7 days.

Finally, we analyze the third most relevant attribute, the Medical Specialty (Figure 6). The internal medicine is related with the highest average LOS (1.64, corresponding to 4.2 days). The second highest average LOS (1.50, corresponding to 3.5 days) is related with orthopedics. Two Medical Specialty values are ranked third in terms of their average effect on LOS: general surgery and urology, both related with an average LOS of 1.40, corresponding to 3.1 days.

These results were shown to hospital specialists and a positive feedback was obtained, confirming meaningful and interesting effects between these attributes and the average expected LOS. Moreover, we would like to stress that the top

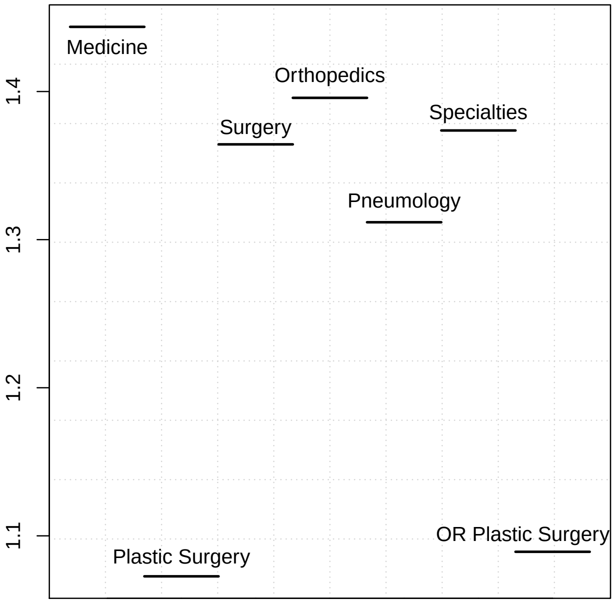

Fig. 5. VEC line segments, showing the average influence of the Inpatient service (x-axis) on the RF model output (y-axis).

four relevant attributes were also in agreement with several literature works (Table 2). For instance, the Episode Type was proposed by [15][7], the Inpatient Service was adopted by [7], the Medical Specialty was used in [6, 16], and the Main Procedure was approached in [12, 7]. We also highlight that Education and Marital Status are two of the proposed attributes that are scarcely adopted by the literature. Yet, these attributes were ranked at 7th and 8th place, with a total contribution of around 10% of the input importance (Figure 3), thus confirming their added value for the LOS prediction model.

4 Conclusions

Due to advances in Information Technology, hospitals are collecting vast amounts of data related with their clinical information systems. All this data can hold valuable knowledge. The development of the Data Mining (DM) field has created new exciting possibilities for extracting such clinical knowledge, in what is known as medical data mining. This work describes an implementation of a medical data mining project approach based on the CRISP-DM methodology. In particular, a DM approach was applied to estimate the Length Of Stay (LOS) of patients at their hospital admission process. We analyzed recent real-world data from a Portuguese hospital, involving a large dataset that included 26462 records (from 15253 patients) and an initial set of 28 attributes (as defined by a medical panel).

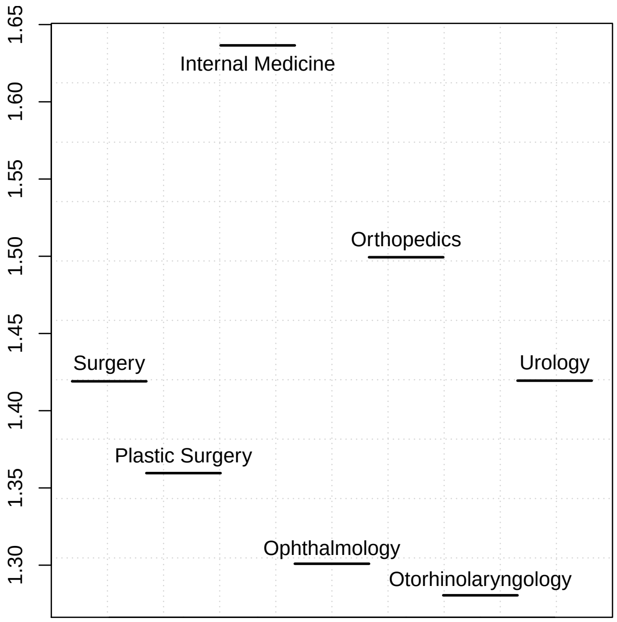

Fig. 6. VEC line segments, showing the average influence of the Medical specialty (x-axis) on the RF model output (y-axis).

The DM approach was guided by the popular CRISP-DM methodology, under a regression approach. After the data preparation phase of CRISP-DM, a cleaned dataset (without outliers and missing data) was achieved, with a total of 26431 records, 14 input attributes and the LOS target. During the modeling phase, six distinct regression models were explored: Average Prediction (AP), Multiple Regression (MR), Decision Tree (DT), Artificial Neural Network (ANN) ensemble, Support Vector Machine (SVM) and Random Forest (RF). These models were compared and tested under a robust evaluation scheme that used 20 runs of a 5-fold cross-validation. Finally, at the evaluation phase of CRISP-DM, the obtained results were analyzed.

The best prediction performance was achieved by the RF model, which presents a very good coefficient of determination value of R^2=0.81 and that is 21 percentage points higher than the minimum threshold of R^2=0.60 set in the business understanding phase. A Regression Error Characteristic (REC) curve analysis revealed that the RF model can correctly predict 85.4% of the examples under a tolerance deviation that ranges from 0.7 (for observed LOS of 0 days) to 26 days (for observed LOS of 66 days). At the same evaluation phase of CRISP-DM, sensitivity analysis and visualization techniques were used to extract explanatory knowledge from the best predictive model (RF). The sensitivity analysis revealed a high impact of inpatient clinical process attributes, instead of the patient's characteristics. In effect, the top three influential input attributes were: the hospital Episode Type, the Inpatient Service where the patient is hospitalized and the associated Medical Specialty. Moreover, the average influence of each of these input attributes in the prediction model has been detailed by using a Variable Effect Characteristic (VEC) analysis. Such analysis has confirmed that several input values associated with high LOS, such as: 'internment" (for Episode Type), "medicine" (for Inpatient Service) and "internal medicine" (for Medical Specialty).

The obtained DM predictive and explanatory knowledge results were considered credible by the hospital specialists and are valuable for hospital managers. By having access to better estimates of the LOS that is more likely to occur in the future and which factors affect such estimates, hospital managers can make more informed decisions. Such informed decisions can lead to a better planning of the hospital resources, resulting in a better hospital management performance, with an increase in the number of available beds for new admissions and reduction of surgical waiting lists.

In the future, we intend to address the implementation phase of CRISP-DM by testing the obtained data-driven model in a real-environment (e.g., by designing a friendly interface to query the RF model). After some time, this would allow us to obtain additional feedback from the hospital managers and also enrich the datasets by gathering more examples. The proposed approach has also the potential to predict well LOS using data from other hospitals, since we address generic LOS and use 14 variables that are easily available at the hospitalization process.

Acknowledgments

We wish to thank the physicians that participated in this study for their valuable feedback. Also, we would like to thank the anonymous reviewers for their helpful suggestions. The work of P. Cortez has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: PEst-OE/EEI/UI0319/2014.

References

1. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. *Advances in Knowledge Discovery and Data Mining*. MIT Press, 1996.
2. K. Cios and G. Moore. Uniqueness of Medical Data Mining. *Artificial Intelligence in Medicine*, 26(1-2):1–24, September-October 2002.
3. Á. Silva, P. Cortez, M. F. Santos, L. Gomes, and J. Neves. Mortality assessment in intensive care units via adverse events using artificial neural networks. *Artificial Intelligence in Medicine*, 36(3):223–234, 2006.
4. Á. Silva, P. Cortez, M. F. Santos, L. Gomes, and J. Neves. Rating organ failure via adverse events using data mining in the intensive care unit. *Artificial Intelligence in Medicine*, 43(3):179–193, 2008.
5. Gabriele Chiusano, Alessandra Staglianò, Curzio Basso, and Alessandro Verri. Unsupervised tissue segmentation from dynamic contrast-enhanced magnetic resonance imaging. *Artificial Intelligence in Medicine*, 61(1):53–61, 2014.

6. Ali Azari, Vandana P Janeja, and Alex Mohseni. Predicting hospital length of stay (phlos): A multi-tiered data mining approach. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pages 17–24. IEEE, 2012.
7. M.G. Castillo. *Modelling patient length of stay in public hospitals in Mexico*. PhD thesis, University of Southampton, 2012.
8. Chris Clifton and Bhavani Thuraisingham. Emerging standards for data mining. *Computer Standards & Interfaces*, 23(3):187–193, 2001.
9. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag, NY, USA, 2nd edition, 2008.
10. Paulo Cortez and Mark J. Embrechts. Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, 225:1–17, March 2013.
11. Dafna Merom, Tamy Shohat, Gil Harari, Meir Oren, and Manfred S Green. Factors associated with inappropriate hospitalization days in internal medicine wards in israel: a cross-national survey. *International Journal for Quality in Health Care*, 10(2):155–162, 1998.
12. Fernando Abelha, Paula Maia, Nuno Landeiro, Aida Neves, and Henrique Barros. Determinants of outcome in patients admitted to a surgical intensive care unit. *Arquivos de Medicina*, 21(5-6):135–43, 2007.
13. A. Oliveira, O. Dias, M. Mello, S. Arajo, D. Dragosavac, A. Nucci, and A. Falcão. Fatores associados à maior mortalidade e tempo de internação prolongado em uma unidade de terapia intensiva de adultos. *Revista Brasileira de Terapia Intensiva*, 22(3):250–256, 2010.
14. A.D. Kalra, R.S. Fisher, and P. Axelrod. Decreased length of stay and cumulative hospitalized days despite increased patient admissions and readmissions in an area of urban poverty. *Journal of general internal medicine*, 25(9):930–935, 2010.
15. Alberto Freitas, Tiago Silva-Costa, Fernando Lopes, Isabel Garcia-Lema, Armando Teixeira-Pinto, Pavel Brazdil, and Altamiro Costa-Pereira. Factors influencing hospital high length of stay outliers. *BMC Health Services Research*, 12(265):1–10, 2012.
16. Samaneh Sheikh-Nia. An Investigation of Standard and Ensemble Based Classification Techniques for the Prediction of Hospitalization Duration. Thesis for Master Science Degree, University of Guelph, Ontario, Canada, 2012.
17. P. Cortez. Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool. In P. Perner, editor, *Advances in Data Mining – Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining*, pages 572–583, Berlin, Germany, July 2010. LNAI 6171, Springer.
18. M. Brown and J. Kros. Data mining and the impact of missing data. *Industrial Management & Data Systems*, 103(8):611–621, 2003.
19. Scott Menard. *Applied logistic regression analysis*. Number 106. Sage, 2002.
20. I.H. Witten, E. Frank, and M. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Franscico, USA, San Francisco, CA, 3rd edition, 2011.
21. J. Bi and K. Bennett. Regression Error Characteristic curves. In T. Fawcett and N. Mishra, editors, *Proceedings of 20th Int. Conf. on Machine Learning (ICML)*, Washington DC, USA, AAAI Press, 2003.