ICEIS 2014

Proceedings of the
16th International Conference on
Enterprise Information Systems

Volume 1

Lisbon, Portugal

27 - 30 April, 2014

Sponsored by
INSTICC – Institute for Systems and Technologies of Information, Control and Communication

In Cooperation with
AAAI – Association for the Advancement of Artificial Intelligence
IEICE / SWIM – IEICE Special Interest Group on Software Interprise Modelling
ACM SIGART – ACM Special Interest Group on Artificial Intelligence
ACM SIGMIS – ACM Special Interest Group on Management Information Systems
ACM SIGCHI – ACM Special Interest Group on Computer Human Interaction

In Collaboration with
IRC – Informatics Research Center

Industrial Partner
Taylor & Francis

Edited by Slimane Hammoudi, Leszek Maciaszek and José Cordeiro

Printed in Portugal
ISBN: 978-989-758-027-7
Depósito Legal: 373558/14

http://www.iceis.org
iceis.secretariat@insticc.org

FOREWORD

This book contains the proceedings of the 16th International Conference on Enterprise Information Systems (ICEIS 2014), which was sponsored by the Institute for Systems and Technologies of Information, Control and Communication (INSTICC), held in cooperation with the Association for the Advancement of Artificial Intelligence (AAAI), IEICE Special Interest Group on Software Interprise Modelling (SWIM), ACM SIGART - ACM Special Interest Group on Artificial Intelligence, ACM SIGMIS - ACM Special Interest Group on Management Information Systems, ACM SIGCHI - ACM Special Interest Group on Computer Human Interaction and in collaboration with the Informatics Research Center (IRC). This year ICEIS was held in Lisbon, Portugal.

The purpose of the 16th International Conference on Enterprise Information Systems is to bring together researchers, engineers and practitioners from the areas of "Databases and Information Systems Integration", "Artificial Intelligence and Decision Support Systems", "Information Systems Analysis and Specification", "Software Agents and Internet Computing", "Human-Computer Interaction" and "Enterprise Architecture", interested in the advances and business applications of information systems.

ICEIS 2014 received 313 paper submissions from 50 countries in all continents, which demonstrates the success and global dimension of this conference. From these, 47 papers were published and presented as full papers (30min oral presentation), 82 papers reflecting work-in-progress were accepted for short presentation and another 82 papers were presented in a poster session. These numbers, leading to a full-paper acceptance ratio of 15% and an oral paper acceptance ratio of 41%, show the intention of preserving a high quality forum for the next editions of this conference.

The high number and high quality of the received papers imposed difficult choices in the selection process. To evaluate each submission, a double blind paper review was performed by the Program Committee, whose members are highly qualified researchers in ICEIS topic areas.

All presented papers will be available at the SCITEPRESS Digital Library and will be submitted for indexation by Thomson Reuters Conference Proceedings Citation Index (ISI), INSPEC, DBLP, EI (Elsevier Index) and Scopus. Additionally, a short list of presented papers will be selected to be expanded into a forthcoming book of ICEIS 2014 Selected Papers to be published by Springer in the LNBIP Series.

The technical program of the conference included a panel and 5 invited talks delivered by internationally distinguished speakers, namely: Kecheng Liu (University of Reading, United Kingdom), Jan Dietz (Delft University of Technology, The Netherlands), Antoni Olivé (Universitat Politècnica de Catalunya, Spain), José Tribolet (INESC-ID/Instituto Superior Técnico, Portugal) and Hans-J. Lenz (Freie Universitat Berlin, Germany). Their participation positively contributes to reinforce the overall quality of the Conference and to

provide a deeper understanding of the fields addressed by the conference.

Moreover, ICEIS 2014 had a special session on Information Systems Security, a satellite workshop on Security in Information Systems and a Doctoral Consortium on Enterprise Information Systems. We are thankful to the chairs for their dedication and hard work in organizing these events.

We sincerely thank all the authors for their submissions and participation in ICEIS 2014. Furthermore, we would like to thank all the members of the program committee and reviewers, who helped us with their expertise, dedication and time. We would also like to thank the invited speakers for their excellent contribution in sharing their knowledge and vision and the workshop/special session chairs whose collaboration with ICEIS 2014 was much appreciated. Finally, we gratefully acknowledge the professional support of the ICEIS 2014 team for all organizational processes.

We hope that all colleagues find this a fruitful and inspiring conference. We hope to contribute to the development of the Enterprise Information Systems community and look forward to having additional research results presented at the next ICEIS, to be held in Barcelona.

Slimane Hammoudi
ESEO, MODESTE, France

José Cordeiro
Polytechnic Institute of Setúbal / INSTICC, Portugal

Leszek Maciaszek
Wroclaw University of Economics, Poland and Macquarie University, Sydney, Australia

ARTIFICIAL INTELLIGENCE AND DECISION SUPPORT SYSTEMS

Full Papers

Short Papers

A Data-driven Approach to Predict Hospital Length of Stay
A Portuguese Case Study

Nuno Caetano[1], Raul M. S. Laureano[1] and Paulo Cortez[2]
[1]*Instituto Universitário de Lisboa (ISCTE-IUL), Av. das Forças Armadas, 1629-026 Lisboa, Portugal*
[2]*ALGORITMI Research Centre, Department of Information Systems, University of Minho, 4800-058 Guimarães, Portugal*
nmcaetano@gmail.com, raul.laureano@iscte.pt, pcortez@dsi.uminho.pt

Keywords: Medical Data Mining, Length of Stay, CRISP-DM, Regression, Random Forest.

Abstract: Data Mining (DM) aims at the extraction of useful knowledge from raw data. In the last decades, hospitals have collected large amounts of data through new methods of electronic data storage, thus increasing the potential value of DM in this domain area, in what is known as medical data mining. This work focuses on the case study of a Portuguese hospital, based on recent and large dataset that was collected from 2000 to 2013. A data-driven predictive model was obtained for the length of stay (LOS), using as inputs indicators commonly available at the hospitalization process. Based on a regression approach, several state-of-the-art DM models were compared. The best result was obtained by a Random Forest (RF), which presents a high quality coefficient of determination value (0.81). Moreover, a sensitivity analysis approach was used to extract human understandable knowledge from the RF model, revealing top three influential input attributes: hospital episode type, the physical service where the patient is hospitalized and the associated medical specialty. Such predictive and explanatory knowledge is valuable for supporting decisions of hospital managers.

1 INTRODUCTION

In the last few decades, hospitals have been storing data regarding electronic clinical information systems. Thus, there is an increasing potential of the use of Data Mining (DM) (Fayyad et al., 1996), to facilitate the creation of knowledge and support clinical decision making, in what is known as medical data mining (Cios and Moore, 2002; Silva et al., 2006; Silva et al., 2008).

In this work we target 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 (Azari et al., 2012; Guzman Castillo, 2012): 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.

Given the importance of LOS prediction, a large number of studies have approached DM techniques in this area. Instead of predicting LOS in specialized medical services, as in UCI (Abelha et al., 2007; Oliveira et al., 2010; Pena et al., 2010) or internal medicine (Kalra et al., 2010), in this study we predict generic LOS, for all hospital services, which is more challenging task. Also, as a case study, only one Portuguese hospital is analyzed. Nevertheless, a large dataset is considered (data collected from 2000 to 2013 with 26462 records from 15253 patients) when compared with some of the mentioned works (e.g., (Pena et al., 2010) only considered 110 patients and (Oliveira et al., 2010) analyzed records from 401 patients). In addition, the attributes that we adopt (described in Section 2) were defined by a hospital expert's medical panel and are commonly available at the hospitalization process. Most of these attributes (e.g., sex, age, episode type, medical specialty) are also adopted by the literature. For instance, the episode type is proposed in (Guzman Castillo, 2012), while the medical specialty was used in (Azari et al., 2012). Moreover, in contrast with several literature works, such as (Pena et al., 2010; Azari et al., 2012; Guzman Castillo, 2012; Sheikh-Nia, 2012), we do not perform a classification task, which requires defining *apriori* which are the interesting LOS class intervals. Instead, we adopt the more informative pure regression approach, which predicts the actual num-

ber of LOS days and not classes.

DM aims at the extraction of useful knowledge from raw data (Fayyad et al., 1996). 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) (Clifton and Thuraisingham, 2001), 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. 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 techniques were tested and compared: Average Prediction (AP), Multiple Regression (MR), Decision Trees (DT) and state-of-the-art regression methods (Hastie et al., 2008), including an Artificial Neural Network (ANN) ensemble, Support Vector Machines (SVM) and Random Forests (RF). The predictive models were compared using a cross-validation procedure with three regression metrics, including the popular coefficient of determination. Moreover, the best predictive model (RF) was opened using a sensitivity analysis procedure (Cortez and Embrechts, 2013) that allows ranking the input attributes and also measuring the average effect of a particular input in the predictive response.

This paper is organized as follows. Firstly, the adopted DM approach is detailed in terms of the CRISP-DM methodology first five phases (Section 2). Then, closing conclusions are drawn (Section 3).

2 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.

2.1 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 value of 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 (Cortez, 2010), for applying the DM regression models (i.e., AP, MR, DT, ANN, SVM and RF) and sensitive analysis methods.

2.2 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 7 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.

2.3 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 boxplots) 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.

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

Name	Description (attribute type)
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)
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)

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, GDC, Treatment, Discharge Destination, Date and Hour). The remaining 14 attributes (**bold** in Table 1) were used as input variables of the regression models (Section 2.4).

Next, missing values were replaced by using the hotdeck method (Brown and Kros, 2003), 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 (Cortez, 2010). 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 (Menard, 2002). 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.

2.4 Modeling

In this phase, we tested six regression methods, as implemented in the **rminer** package (Cortez, 2010): AP, MR, DT, ANN, SVM and RF. The AP is a naive model that consists in predicting the same average LOS ($\bar{y}$, 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:

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

where $\beta_0, \ldots, \beta_i$ are the set of parameters to be adjusted, usually by applying a least squares 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 (Hastie et al., 2008).

The **rminer** package full implementation details can be found in (Cortez, 2010). 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 (Hastie et al., 2008). 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 (Cortez, 2010): $C = 3$ (for standardized data) and $\varepsilon = 3\sigma_y\sqrt{\log(N)/N}$, 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 paper, 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 valiation 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")
```

2.5 Evaluation

To evaluate the predictions, three regression metrics were selected (Witten et al., 2011): coefficient of determination (R2), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). R2 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 (Bi and Bennett, 2003). 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 2 presents the regression predictive results, in terms of the average of the 20 runs of the 5-fold cross-validation evaluation scheme. From Table 2, 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 R2 value was achieved (0.813), much higher than the minimum success value of 0.6 set in Section 2.1.

Table 2: Predictive results (average of 20 runs, as measured over test data; best values in **bold**).

	Metrics		
Method	**R2**	**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 quality of the predictions for the RF model can also be seen on Figure 2, which plots the observed (x-axis) versus de 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.

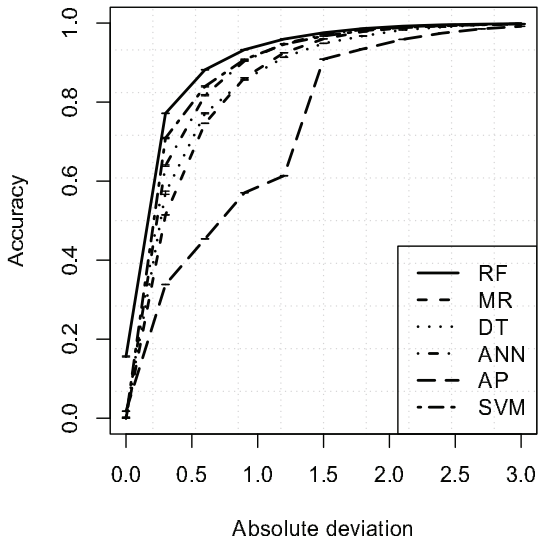

Figure 1: REC curves for all tested models.

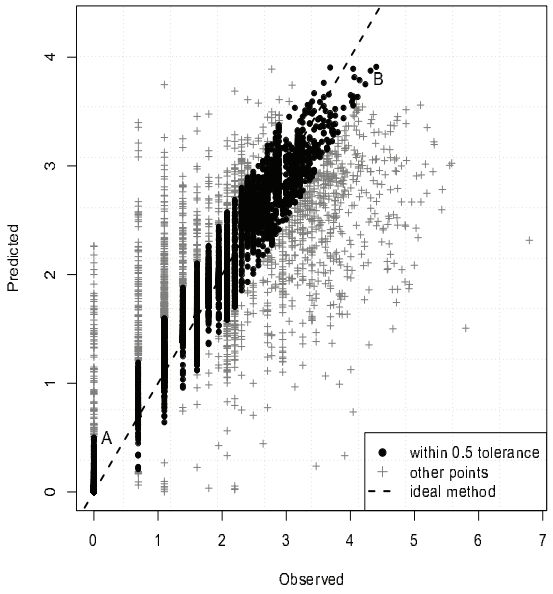

Figure 2: Observed versus predicted RF values.

It should be noted that the presented predicted results were computed over the logarithm transform scale (see Section 2.3). 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 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 (Cortez and Embrechts, 2013). 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 the Variable Effect Characteristic (VEC) curve.

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 curves (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) 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 curves.

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).

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

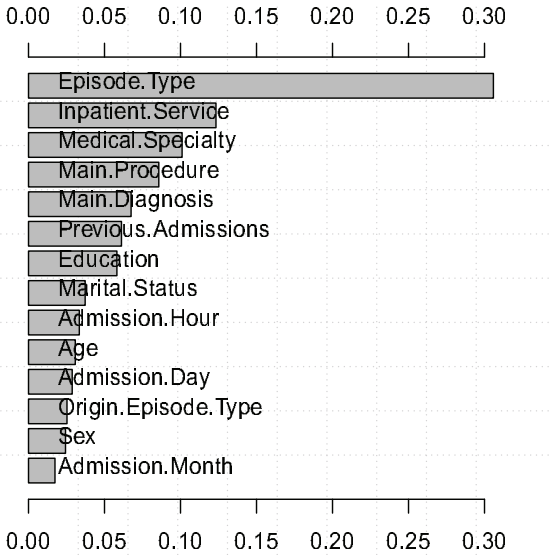

Figure 3: Input importance bar plot for the RF model.

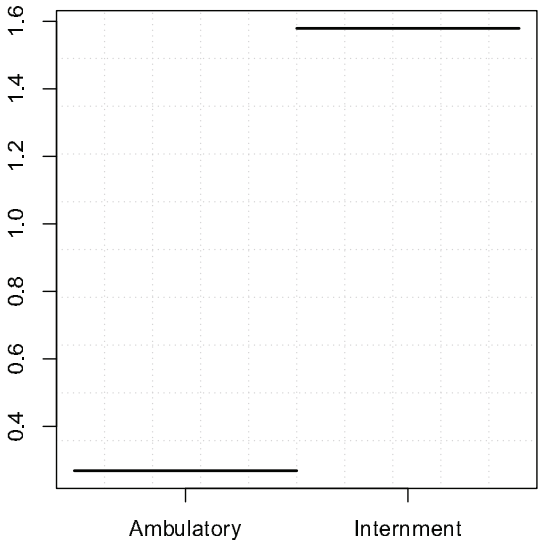

Figure 4: VEC line segments, showing the average influence of the Episode type (x-axis) on the RF model output (y-axis).

2.9 days; and pulmonology, 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

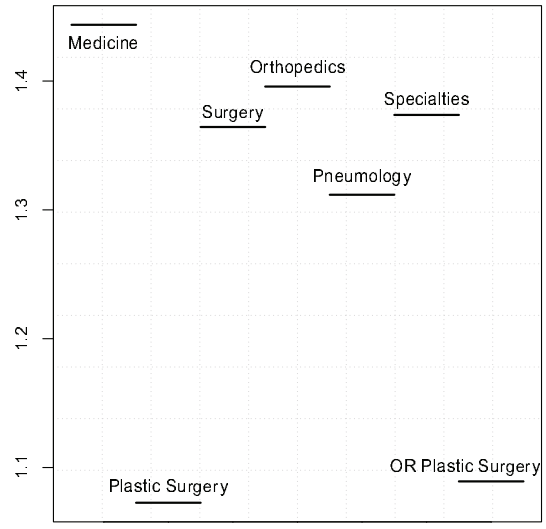

Figure 5: VEC line segments, showing the average influence of the Inpatient service (*x*-axis) on the RF model output (*y*-axis).

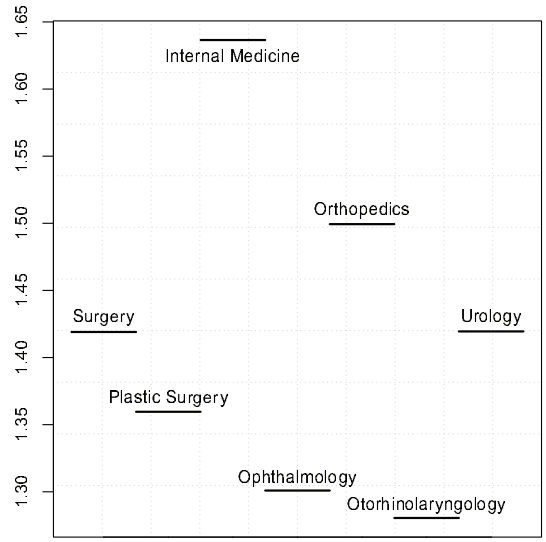

Figure 6: VEC line segments, showing the average influence of the Medical specialty (*x*-axis) on the RF model output (*y*-axis).

meaningful and interesting effects between these attributes and the average expected LOS. Moreover, we would like to stress that the top four relevant attributes were also in agreement with several literature works. For instance, the Episode Type was proposed by (Guzman Castillo, 2012; Freitas et al., 2012), the Inpatient Service was adopted by (Guzman Castillo, 2012), the Medical Specialty was used in (Azari et al., 2012; Sheikh-Nia, 2012), and the Main Procedure was approached in (Abelha et al., 2007; Guzman Castillo, 2012).

3 CONCLUSIONS

The development of the Data Mining (DM) field has created new exciting possibilities for the field of medical data mining. In this paper, a DM approach was applied to estimate the length of stay (LOS) of patients at their hospital admission process. As a case study, 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).

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 compared and tested, under a robust evaluation scheme (20 runs of a 5-fold cross-validation). Finally, at the Evaluation phase of CRISP-DM, the best results were obtained by the Random Forest (RF) model, which presents a very good coefficient of determination value (R2=0.81, 0.21 pp higher than the minimum threshold of 0.6 set in the Business Understanding phase). Such model can correctly predict 85.4% of the examples under a tolerance that ranges from 0.7 (for observed LOS of 0 days) to 26 days (for observed LOS of 66 days). Ensemble methods methods, such as RF, are usually usefull when the sample data does not cover the tuple space properly and the diversity of learners can minimize this problem

Moreover, sensitivity analysis and visualization techniques were used to extract explanatory knowledge from the best predictive model (RF). This analysis revealed a high impact of inpatient clinical process attributes, instead of the patient's characteristics. In particular, the top three influential input attributes were: the hospital episode type, the physical service where the patient is hospitalized and the associated medical specialty.

The obtained DM predictive and explanatory knowledge results are valuable for hospital managers. By having access to better estimates of what is more likely to occur in the future and which factors affect such estimates, hospital managers can make more informed decisions (e.g., better planning of the hospital resources), in order to accomplish their goals (e.g., increase the number of available beds for new admissions and reduce surgical waiting lists).

In future work, we intend to explore more ensemble methods, such as Adaptive Boosting (Freund and Schapire, 1995). We will also address the Implemen-

tation 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.

ACKNOWLEDGEMENTS

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

Abelha, F., Maia, P., Landeiro, N., Neves, A., and Barros, H. (2007). Determinants of outcome in patients admitted to a surgical intensive care unit. *Arquivos de Medicina*, 21(5-6):135–43.

Azari, A., Janeja, V. P., and Mohseni, A. (2012). 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.

Bi, J. and Bennett, K. (2003). Regression Error Characteristic curves. In Fawcett, T. and Mishra, N., editors, *Proceedings of 20th Int. Conf. on Machine Learning (ICML)*, Washington DC, USA, AAAI Press.

Brown, M. and Kros, J. (2003). Data mining and the impact of missing data. *Industrial Management & Data Systems*, 103(8):611–621.

Cios, K. and Moore, G. (2002). Uniqueness of Medical Data Mining. *Artificial Intelligence in Medicine*, 26(1-2):1–24.

Clifton, C. and Thuraisingham, B. (2001). Emerging standards for data mining. *Computer Standards & Interfaces*, 23(3):187–193.

Cortez, P. (2010). Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool. In Perner, P., editor, *Advances in Data Mining – Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining*, pages 572–583, Berlin, Germany. LNAI 6171, Springer.

Cortez, P. and Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, 225:1–17.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). *Advances in Knowledge Discovery and Data Mining*. MIT Press.

Freitas, A., Silva-Costa, T., Lopes, F., Garcia-Lema, I., Teixeira-Pinto, A., Brazdil, P., and Costa-Pereira, A. (2012). Factors influencing hospital high length of stay outliers. *BMC Health Services Research*, 12(1):265.

Freund, Y. and Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer.

Guzman Castillo, M. (2012). *Modelling patient length of stay in public hospitals in Mexico*. PhD thesis, University of Southampton.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, NY, USA, 2nd edition.

Kalra, A. D., Fisher, R. S., and Axelrod, P. (2010). 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.

Menard, S. (2002). *Applied logistic regression analysis*. Number 106. Sage.

Oliveira, A., Dias, O., Mello, M., Arajo, S., Dragosavac, D., Nucci, A., and Falcão, A. (2010). 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.

Pena, F., Soares, J., Peixoto, R., Jnior, H., Paiva, B., Moraes, F., Engel, P., Gomes, N., and Pena, G. (2010). Análise de um modelo de risco pré-operatrio especfico para cirurgia valvar e a relação com o tempo de internação em unidade de terapia intensiva. *Revista Brasileira de Terapia Intensiva*, 22(4):339–345.

Sheikh-Nia, S. (2012). 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.

Silva, A., Cortez, P., Santos, M. F., Gomes, L., and Neves, J. (2006). Mortality assessment in intensive care units via adverse events using artificial neural networks. *Artificial Intelligence in Medicine*, 36(3):223–234.

Silva, A., Cortez, P., Santos, M. F., Gomes, L., and Neves, J. (2008). Rating organ failure via adverse events using data mining in the intensive care unit. *Artificial Intelligence in Medicine*, 43(3):179–193.

Witten, I., Frank, E., and Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Franscico, USA, San Francisco, CA, 3rd edition.