Data Science Use Case: Applying CRISP-DM Methodology for Predicting Length of Stay in a Portuguese Hospital

by Salma Kaisan Syauqi

Data science has been used widely in many fields, including the healthcare field. Hospitals as a part of the healthcare field use one of the core disciplines of data science, which is medical data mining (Stedman & Hughes, 2021) to get insight for supporting clinical decision-making. One of the hospitals in Portuguese used data mining to predict the Length of stay (LOS) (Caetano et al., 2015). Length of stay (LOS) is necessary for the management of a hospital which directly impacts on hospital costs and patient satisfaction. Besides that, LOS has relation with the severity of a disease and also mortality (Chrusciel et al., 2021). This hospital used the CRISP-DM methodology that consisted of five phases which are business understanding, data understanding, data preparation, modeling, and evaluation (Caetano et al., 2015). In this article, we are going to understand deeper each of those phases by using a real case.

1. *Business Understanding*

   One of the problems that hospital has is that it is pressured to provide more beds for new patients which impacts on hospital admission scheduling. The solution that can be offered to this problem is the prediction of Length of stay (LOS) which made by using regression model in order to make the prediction more accurate. The software that are used are SQL for extracting data from hospital database and R tool for analyzing the data.

2. *Data Understanding*

   In the period of data collecting, from October 2000 to March 2013, the data that are successfully stored consisted of 26,462 inpatient and 15,253 patients. An expert medical panel that consists of 9 physicians from different medical specialities is responsible for selecting the relevant data for LOS prediction which can be seen in the table below.

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. *Data Preparation*

In data preparation phase, there are some processes of the data that adopt the R tool which are performing an exploratory data analysis and preprocessing the original dataset. From that process, the physicians detect a few outliers and discard some attributes. After cleaning, there are 26,431 records in the database and only 14 attributes that are used as input variables of the regression models which are shown in the table below.

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]

4. *Modeling*

For selecting the appropriate model, six regression models are tested. Those models are Average Prediction (AP), Multiple Regression (MR), Decision Tree (DT), Articial Neural Network (ANN) ensemble, Support Vector Machine (SVM) and Random Forest (RF). And the model that is selected is retrained by using all training data.

5. *Evaluation*

Regression metrics are used for evaluating the models. Those regression metrics are the coecient of determination (R2), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Then, for comparing some regression methods in a graph, the Regression Error Characteristic (REC) curse can be used. The evaluation result determine that RF model is the best predictive model.

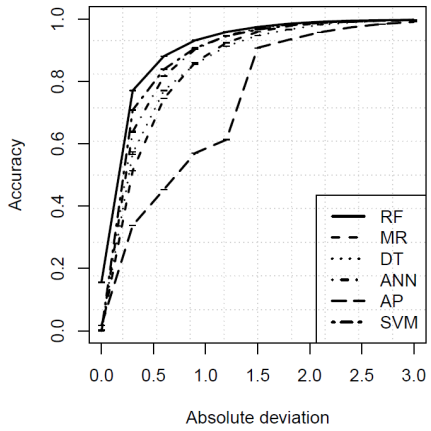

Fig. 1. REC curves for all tested models.

Reference:

Caetano, N., Cortez, P., & Laureano, R. M. S. (2015). Using data mining for prediction of hospital length of stay: An application of the CRISP-DM methodology. *Lecture Notes in Business Information Processing*, *227*. https://doi.org/10.1007/978-3-319-22348-3_9

Stedman, C., & Hughes, A. (2021, September 7). data mining. SearchBusinessAnalytics. Retrieved September 19, 2022, from https://www.techtarget.com/searchbusinessanalytics/definition/data-mining

#:%7E:text=Data%20mining%20is%20a%20key,useful%20information%20in%20data%20sets.++

Chrusciel, J., Girardon, F., Roquette, L., Laplanche, D., Duclos, A., & Sanchez, S. (2021, December). The prediction of hospital length of stay using unstructured data. *BMC Medical Informatics and Decision Making*, *21*(1). https://doi.org/10.1186/s12911-021-01722-4

GitHub Repository link: https://github.com/salmakaisan/amlp-pzsib