Computer Science and Information Systems 18(3):1001–1021 https://doi.org/10.2298/CSIS200422016T

Patient Length of Stay Analysis with Machine Learning Algorithms

Savo Tomović

Faculty of Mathematics and Natural Sciences,
Univeristy of Montenegro, 81000 Podgorica, Montenegro
savot@ucg.ac.me

Abstract. In this paper the problem of measuring factor importance on patient length of stay in an emergency department is discussed. Historical dataset contains average patient length of stay per day. Factors are agreed with domain expert. The task is to provide factors' impact measure on specific day that does not belong to the historical dataset (new observation) and average length of stay for that day is higher than specified threshold. Observations are represented as multidimensional numeric vectors. Each dimension represents factor. The basic idea consists of identifying appropriate neighbourhood and measure distances between the new observation and its neighbourhood in the historical dataset with respect to each factor. Impact measure of a factor is derived from the Error Sum of Squares. Factor impact is proportional to distance between the observation and its neighbourhood with respect to the dimension representing that factor. Nearest neighbour and clustering methods for neighbourhood determination are considered.

Keywords: length of stay analysis, nearest neighbours, clustering, SSE

1. Introduction

In this paper the problem of explaining patient length of stay (LOS) elevation in an emergency department is discussed. In the rest of the paper the length of stay explanation problem is referred to as LOSEP.

The task is to identify the most significant factors and objectively measure their impact on patient length of stay. Historical dataset is available. It contains average patient length of stay per day along with a set of features - factors. User is interested in investigating new observations - dates that do not belong to the historical dataset and for which registered average length of stay is higher than a threshold σ. The aim is to identify which factors are the most significant in contributing to longer patient stay in an emergency department, providing the leadership to improve concrete aspects in the organization.

More precisely, let the historical dataset H is given. It contains average length of stay per day along with available features - factors. Features are representing organizational or other aspects that potentially can cause longer patient stay than it is desired or expected. Available features are referred to as *factors* in the rest of the paper. Every record from H is represented with multidimensional numeric vector of the following form $(factor_1, factor_2, ..., factor_n, LOS)$. Let $q \notin H$ represents object $q = ((factor_1(q), factor_2(q), ..., factor_n(q), LOS(q))$ such that $LOS(q) > \sigma$. In the rest of the paper objects like q are referred to as *new observations*. The task is to create a

methodology to objectively estimate impact of each factor on the length of stay augmentation registered on the new observation q. Factors can be sorted with respect to the estimated impacts. Such ordering determines the most significant factors causing the situation $LOS(q) > \sigma$.

Length of stay is considered as one of the most important indicators for any hospital department efficiency. The compulsion is to keep length of stay below some value. Such value is not simply average or median, but it is estimated by specific methodology. In large hospitals it is challenging for management to manually analyse every organizational aspect that can affect length of stay. The results of such analysis should assist leaders of hospitals and its staff in understanding what is happening on daily basis and what actions should be taken.

The LOSEP problem is different from problems of length of stay prediction and determining factors influencing patient length of stay where the overall impact of the contributing factors is derived from correlations existing in the training dataset. Such approaches usually create model based on available dataset and use the model to estimate the impact of each factor to the target variable - length of stay in this case. Such estimation is ”static” meaning that for every new observation the model will produce the same factor ranking without considering any specificities related to concrete object. For example, the model can detect the highest importance of the number of emergency department visits based on the training dataset. So, for any new observation representing days when average length of stay is higher than σ the answer will be the same: length of stay is elevated due to higher number of visits. Of course, length of stay can be high because number of visits is elevated, but only this factor may not be an issue especially if the patients were of low triage level. Investigating the number of the sickest patients can indicate that this factor is also of significant importance.

The method presented in this paper is able to independently analyse every new observation and provide factors ranking related to specificities of the considered observation. In general, it is possible to obtain different explanations and factors ranking for each new observation. In such manner the solution can cover dynamical aspects of hospital behaviour meaning that on different days different aspects can be of different importance.

The motivation for this study originates from the challenges in designing and implementing system devoted to hospital management in American healthcare system. In the case study that is exposed in the fifth section an emergency department is considered. Data about average length of stay - *AVG_LENGTH_OF_STAY* per day expressed in hours is stored in historical dataset. Factors of interest are defined as follows: number of visits - *ED_VISITS*, number of ambulance visits - *AMBULANCE_VISITS*, average triage level - *AVG_TRIAGE_LEVEL*, number of patients with triage level 1 and 2 (most sick patients) - *TRIAGE_LEVEL_1_2_COUNT* and diversion hours - *DIVERSION_HOURS*. Granularity of the historical dataset is on a day level.

The proposed solution consists of two components. The first component is to find appropriate neighbourhood of a new observation. Nearest neighbour and clustering methods for neighbourhood determination are considered. With the nearest neighbour method the algorithm finds k closest objects from historical dataset to construct neighbouring cluster. In clustering method observations from the historical dataset are clustered in preprocessing step and the cluster with the closest centroid to a new observation is considered

as neighbouring cluster. Standard Euclidean distance is used to determine the distance between objects.

The second component of the solution is for objective impact estimation of each factor. The proposed procedure calculates increment to the Error Sum of Squares if the new observation was added to the neighbouring cluster and distributes the increment value among factors proportionally.

The paper is decomposed into several sections as follows. The next section is devoted to a number of studies available in the literature, where we try to compare with and signify our contribution. Section 3 presents a motivating example for this research study. Two procedures for a new observation explanation are exposed in the third section. The same section introduces the function for objective impact measurement of each factor. The applicability of the proposed method is demonstrated in section 5 on the case study related to an emergency department. The last section contains concluding remarks and possible extensions to the proposed method.

2. Related work

There is a huge amount of scientific papers dealing with computer applications in medical research. For example, results from almost 300 papers appeared in numerous journals and conferences between 1999 and 2013 were presented in [12]. Specifically, analysis of healthcare services quality occupies significant effort in research community. Length of stay is considered as the most important indicator for any department efficiency, especially from the patient's perspective.

Many studies with objective to predict length of stay and identify and quantify impact of different factors were presented. Number of examined factors is huge and some of them were even more carefully interpreted regrading department specificity.

In [2] authors provide detailed review of length of stay applications and methods to calculate and predict length of stay. Authors classified algorithms into four categories: arithmetic methods, statistical methods, data-driven approaches and multi-stage models.

Arithmetic methods are the simplest. They assume that length of stay is normally distributed [2] and usually calculate average length of stay or median [7]. These measures can be misleading [26] because typically length of stay has an exponential distribution [2].

Statistical methods can be categorized into two subgroups [2]: survival analysis and regression analysis.

Survival analysis [11] uses length of stay as surrogate to estimate the impact of patient data on survival time.

Regression analysis can be considered as a statistical method that identifies factors which possibly predict length of stay. There is a huge number of analysed factors, internal and external, including organizational factors (patient arrival time, physicians and nurse characteristics, physicians and nurse shift changes, admission to specific hospital wards, laboratory performance, imaging, consultation, etc.), demographic data (age, gender, martial status, occupation, place of residence, etc.), information related to hospitalization (diagnosis related group - DRG, specialty of physician, history of admission to hospital, triage acuity level, type of admission, type of treatment, patient condition, method of payment for hospital costs) [8], [1], [21], [9], [20], [4]. Within this type, among others, linear

regression and logistic regression and regression trees are found [32], [23], [10]. Percent of length of stay variation that could be explained with this approach vary from about 35% to almost 70%. As it is known from the literature and stated in some of these studies, for regression analysis there are several assumptions that must be satisfied: linearity, normality, homoscedasticity, data must not show multicollinearity, etc. [29].

In [18] authors claim that models based on regression analysis are heavily dependant on available data and even minor change in the data can generate completely different models from which different patterns or rules are extracted. Authors propose the procedure to create diverse regression models through re-sampling of the training data and achieve more stable and accurate models. Other examples of combining and averaging models to reduce prediction error include bagging [5], boosting [13] and random forest [31].

Data driven approaches usually refer to data mining techniques that are used to predict length of stay above or bellow a certain threshold that is for example specific for diagnosis related groups. Authors in [6] apply classification techniques to categorize length of stay in intensive care unit with respect to recommended seven-day norm; authors in [27] try to predict length of stay for post-coronary patients longer than 120 days; authors in [14] consider patients suffering from burns and create a model to predict whether their length of stay will be less than one-week; authors in [19] present research devoted to appendectomy patients and develop a model that recognizes those patients whose length of stay will exceed recommended five-day period. Authors in [28] apart from, testing several classification algorithms, reported results about finding the most significant input variables to predict the target variable - length of stay. Among thirty six input variables, the most significant variables affecting length of stay according to generated classification model were drug categories, co-morbidity (that is the presence of one or more diagnosis co-occurring with the primary disease), gender (men had longer length of stay than women) and age (patient younger than 50 years and older than 80 years had longer length of stay).

Interesting approach is presented in [3]. Authors categorized length of stay into three groups as short, medium and long. Training dataset is constructed by clustering similar claims after which classification is performed using ten different classifiers. For each classifier, using clustering as a preprocessing step gives better accuracy as compared to non-clustering based training dataset.

Wide variety of classification algorithms were applied in previously mentioned studies: decision trees, support vector machines - SVM, artificial neural networks - ANN, Naive Bayesian classifier etc. According to [30], decision tree implementation C4.5 is a classifier that has the best combination in terms of error rate and speed. Authors in [33] found that decision tree R-C4.5s (successor of C4.5) algorithm creates more robust and smaller trees. In [22] authors reported that Naive Bayesian classifier is robust to missing data.

Clustering can be considered as data driven approach, too. In [16], [15] clustering is used to create clinically meaningful groups with respect to length of stay and covariates: gender, age, primary diagnosis, etc.

Multi-stage models are based on modelling patient flow in hospital with Markov and semi-Markov chains. Length of stay is considered as an array of successive stages which the patients go through until they leave the hospital completely. It is possible to include additional variables that may influence patient length of stay as it is suggested in [25],

[24], [17]. The final model represents length of stay based on five of the most important variables, namely age, gender, admission method, Barthel grade and destination on departure from hospital.

As it is stated in the introductory section, all studies treat length of stay and its influencing factors statically in terms of that discovered correlations are interpreted as universal truth. For example, factor *X* predicts *Y%* of the length of stay augmentation. It is not possible to give different explanation for length of stay prolongation on two different observations.

In this study we do not predict length of stay neither do we use training dataset to develop a model from which the most significant factors are detected. The LOSEP problem discussed in this paper consists of defining procedure that is able to objectively estimate impact of available factors on length of stay elevation registered on new observations. Historical dataset H and threshold σ are given. New observations represent dates outside the historical dataset for which registered length of stay is $> \sigma$. Using the set H, the procedure must objectively (mathematically) measure impact of each factor causing higher length of stay than it is expected or desired. It is necessary to consider every new observation independently and provide potentially different explanations for different observations although the same set H is always used.

For example, consider two observations $a = (X_1(a), X_2(a), ..., X_n(a), LOS(a))$ and $b = (X_1(b), X_2(b), ..., X_n(b), LOS(b))$ for which $LOS(a) \geq \sigma$ and $LOS(b) \geq \sigma$ and factors $X_1, X_2, ..., X_n$. Explanations why $LOS(a) \geq \sigma$ and $LOS(b) \geq \sigma$ hold can be represented in the following forms $a : \{X_{i1} : impact_{i1}, X_{i2} : impact_{i2}, ..., X_{in} : impact_{in}\}$ and $b : \{X_{j1} : impact_{j1}, X_{j2} : impact_{j2}, ..., X_{jn} : impact_{jn}\}$ where $(i_1, i_2, ..., i_n) \neq (j_1, j_2, ..., j_n)$ are different permutations of the set $(1, 2, ..., n)$. The previous means that the most significant factor for the object a is X_{i1} while the most significant factor for the object b is X_{j1}.

3. A Motivating Example

In this section, an example from the case study presented in section 5, is briefly discussed.

Consider that historical dataset covers the period between January and August 2017 and let $\sigma = 4$ hours. The first two days of September 2017 are depicted on figure Fig. 1. The task is to explain elevation in average length of stay with respect to available factors: *AMBULANCE_VISITS, ED_VISITS, AVG_TRIAGE_LEVEL, TRIAGE_LEVEL_1_2_COUNT* and *DIVERSION_HOURS*. As an example of traditional approach, the Random Forest Regression is fitted with the historical data. Generated model ranks mentioned factors according to their importance to length of stay as follows *AVG_TRIAGE_LEVEL* $\succ$ *AMBULANCE_VISITS* $\succ$ *ED_VISITS* $\succ$ *TRIAGE_LEVEL_1_2_COUNT* $\succ$ *DIVERSION_HOURS*. It is illustrated on figure Fig. 2. This ranking is learned from the provided dataset and all future observations must be explained in such manner.

The procedure presented in this paper is able to estimate factors independently for each new observation. Results are presented on figure Fig. 3. It can be seen that on the 1st September the most important factors for length of stay elevation are *ED_VISITS* and *TRIAGE_LEVEL_1_2_COUNT*, while for the 2nd September the combination *AMBULANCE_VISITS*, *DIVERSION_HOURS* and *TRIAGE_LEVEL_1_2_COUNT* causes higher length of stay.

SEPTEMBER 2017 Dates	1	2	...
AVG_LENGTH_OF_STAY	4:54	5:08	
ED_VISITS	114	144	
AMBULANCE_VISITS	16	20	
AVG_TRIAGE_LEVEL	3.4	3.29	
TRIAGE_LEVEL_1_2_COUNT	24	32	
DIVERSION_HOURS	0	4	

Fig. 1. Observations that have to be explained

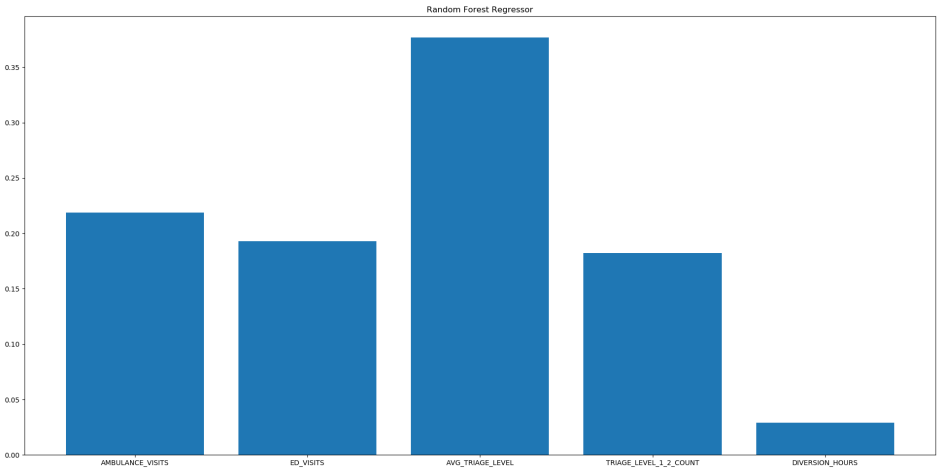

Fig. 2. Contributing factors ranked by Random Forest Regression

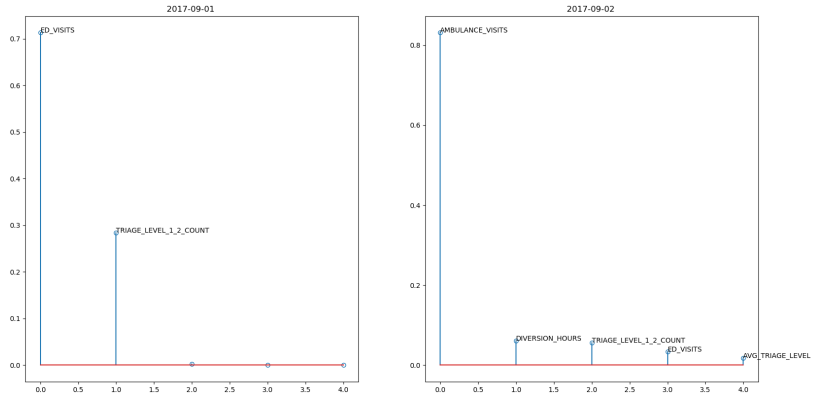

Fig. 3. Contributing factors ranked in LOSEP problem

To achieve clear distinction between the LOSEP problem and problems of determining factors influencing patient length of stay and length of stay prediction in this paragraph more general definition of the LOSEP problem is given. Consider that data matrix $H_{m\times n+1}$ is given. It contains historical data represented in the form of m multidimensional vectors with the following dimensions: $var_1, var_2, ..., var_n, TARGET\_VALUE$. Without loss of generality, we can take that the last column is target variable that should be explained in terms of other columns representing factors of interest. Additionally, threshold value σ is provided. Consider that stream of new observations Q is provided. The task is to objectively measure impact of the defined factors on the target value elevation registered on the new observation $q \in Q$. The result can be represented in the following form $q : \{var_{i1} : impact_{i1}, var_{i2} : impact_{i2}, ..., var_{in} : impact_{in}\}$, where $impact_{i1} > impact_{i2} > ... > impact_{in}$ and $impact_{i1} + impact_{i2} + ... + impact_{in} = 1$. Of course, for different observations $q_1 \neq q$ the explanation is generally different $q_1 : \{var_{j1} : impact_{j1}, var_{j2} : impact_{j2}, ..., var_{jn} : impact_{jn}\}$.

It is obvious that the LOSEP problem is an instance of the previous more general problem. Average length of stay, *AVG_LENGTH_OF_STAY*, on a specific day in an emergency department is *TARGET_VALUE* and factors are *ED_VISITS, AMBULANCE_VISITS, AVG_TRIAGE_LEVEL, TRIAGE_LEVEL_1_2_COUNT, DIVERSION_HOUR*. It can be seen that such selection of factors mostly reflects operational aspects of an emergency department rather than single patient characteristics. The threshold value σ is set to 4 hours. Patient length of stay expressed in hours and averaged on daily basis for each day from the future period forms the stream Q. The task is to estimate impact of each factor on days $q \in Q$ where $AVG\_LENGTH\_OF\_STAY(q) > 4$ and indicate the most important aspect that causes elevation in length of stay.

Also, the LOSEP problem is different from the outlier analysis. The new observations introduced in the above example may or may not be outliers. Instead of finding outlier objects in a dataset, the LOSEP problem consists of finding the factors best explaining why length of stay for the new observation that is not present in the given dataset, is higher than σ. Additionally, in the LOSEP problem it is necessary to introduce the function eligible to objectively measure the impact of the factors on the length of stay value registered on a new observation.

4. Explanation of New Observations

Historical dataset in the LOSEP problem contains data over specific past period of time. The historical dataset H can be considered as a collection of records or data objects. Each object consists of fixed set of variables. So, objects from the historical dataset can be thought of as vectors (points) of the following form $(factor_1, factor_2, ..., factor_n, LOS)$ in a multidimensional space, where each dimension corresponds to exactly one factor. In addition, the last dimension corresponds to the average patient length of stay on a specific day.

In other words, dataset H can be interpreted as $m \times n + 1$ matrix, where there are m rows, one for each object, and $n + 1$ columns, one for each dimension.

Granularity of historical dataset considered in the case study from the fifth section is on a day level. For each $i, 1 \leq i \leq n$, $factor_i(o)$ is a result of some aggregate function (sum, average, count) of variable $factor_i$ registered on every patient processed on a

specific day that is represented with o. Accordingly, for each record $o \in H$, $LOS(o)$ is average patient length of stay on a specific day that is represented with o. To emphasize this situation, in the rest of this section LOS is replaced with *AVG_LENGTH_OF_STAY*.

The historical dataset is partitioned on two parts based on *AVG_LENGTH_OF_STAY* dimension and user defined threshold σ. The value for σ can be a combination of looking at historical data and a benchmark leadership wants to hit. Objects $o \in H$ for which $AVG\_LENGTH\_OF\_STAY(o) < \sigma$ holds, are part of a *good partition*. Objects belonging to the good partition are referred to as *good objects*. The other partition contains objects $o \in H$ for which $AVG\_LENGTH\_OF\_STAY(o) \geq \sigma$ is true.

More precisely, good partition H_{GP} is a subset of the given historical dataset H that is determined by the threshold σ as follows:

$$H_{GP} = \{o | o \in H \wedge AVG\_LENGTH\_OF\_STAY(o) < \sigma\}. \quad (1)$$

The LOSEP problem consists of providing explanation of a new observation. New observations belong to the same multidimensional space as the objects from the set H, but they are not known in advance. Actually, new observations represent future observations, dates outside the H, for which *AVG_LENGTH_OF_STAY* is $\geq \sigma$. Explanation of an observation q consists of objective measurement of the factors' impact on the average length of stay elevation and can be represented in the following form $q : \{factor_{i1} : impact_{i1}, factor_{i2} : impact_{i2}, ..., factor_{in} : impact_{in}\}$, where $impact_{i1} > impact_{i2} > ... > impact_{in}$ and $impact_{i1} + impact_{i2} + ... + impact_{in} = 1$ hold. It means that the greatest impact on the value $AVG\_LENGTH\_OF\_STAY(q) \geq \sigma$ has $factor_{i1}$ followed by $factor_{i2}$ and the smallest importance is estimated for $factor_{in}$. Here $(i_1, i_2, ..., i_n)$ is a permutation of the set $\{1, 2, ..., n\}$.

The new observation is explained with respect to the appropriate neighbourhood. The neighbourhood is determined among available good objects from H_{GP}. Two procedures, namely *nearest neighbour based method* and *clustering method* are considered.

Nearest neighbour method finds *k* closest good objects to the given observation $q = (factor_1(q), factor_2(q), ..., factor_n(q), AVG\_LENGTH\_OF\_STAY(q))$. The number of good objects constituting the neighbourhood is a user defined constant. Standard Euclidean distance is used to determine distance between objects.

When *k*-neighbourhood of the new observation q is determined, the algorithm calculates its centroid. The centroid c_q is the mean of all objects in the *k*-neighbourhood. Notice that all neighbouring objects are good objects.

The impact for every factor is estimated based on the formula for calculating the sum of squared errors (SSE). SSE is usually used as an objective function to estimate the quality of clustering. The SSE takes the sum of the squared distances between every object and the closest centroid. Set of clusters with the smallest SSE is considered as the best clustering solution.

Consider that *k*-neighbourhood of the new observation q constitutes the neighbouring cluster $C_q, |C_q| = k$. The SSE of the C_q is given by the following formula:

$$SSE = \sum_{x \in C_q} dist^2(c_q, x). \quad (2)$$

If the observation q is added to the cluster C_q, SSE of the cluster is increased by the amount $dist^2(c_q, q)$. As it is mentioned earlier, $dist$ is a standard Euclidean distance, so impact of i^{th} factor can be estimated by the formula:

$$impact(factor_i) = (c_q[factor_i] - q[factor_i])^2 / dist^2(c_q, q). \quad (3)$$

The algorithm based on the nearest neighbour approach is presented in the following listing. The algorithm is implemented in Python3 using sklearn library.

```
procedure kNN_impact_estimation(X, q, k)
{X is representing good objects}
{q is the new observation}
{k is the user defined size of the new observation neighbourhood}
begin
    {available good objects are fitted to the NearestNeighbors class}
    nnbrs = NearestNeighbors(n_neighbors=k)
    nnbrs.fit(X)

    {cluster and centroid of the new observation neighbourhood}
    C_q = nnbrs.kneighbors(q)
    c_q = (1/k) * sum_{x in C_q} x

    {impact[i] is impact of the factor_i}
    {q = (factor_1, factor_2, ..., factor_n)}
    impact = []
    for i in range(0, dim(q)):
        impact[i] = (c_q[i] - q[i])^2 / dist^2(c_q, q)

    return impact
```

With the clustering method neighbourhood of a new observation is determined by clustering of all good objects and determining the closest centroid. Theoretically, clusters can be created with any clustering algorithm, but exhaustive experiments that were part of the case study presented in the next chapter indicated that the best choice is Affinity propagation method. Partition-based methods usually require specifying number of clusters in advance, which was impossible to properly estimate in the available dataset. Density based methods during model building declare significant number of good objects as outliers. Consequently, such objects are eliminated from factor estimation that was unacceptable, bearing in mind that the number of good objects is generally limited.

When clusters are created the algorithm determines the closest centroid to the observation q. Let C_q be the cluster with the centroid c_q that is closest to the new observation q. Adding the q to the cluster C_q increases SSE of the cluster by the amount $dist^2(c_q, q)$, where $dist$ is a standard Euclidean distance. As before, the impact of i^{th} factor can be estimated by the formula (3).

The algorithm based on clustering approach is presented in the following listing. The algorithm is implemented in Python3 using sklearn library.

```
procedure clustering_impact_estimation(X, q)
{X is representing good objects}
{q is the new observation}
begin
    {available good objects are fitted to the Affinity Propagation class}
    clustering = AffinityPropagation(X)
```

```
{determine the closest centroid}
c_q = min_{c∈clustering.cluster_centres} dist(c, q)

{impact[i] is impact of the factor_i}
{q = (factor_1, factor_2, ..., factor_n)}
impact = []
for i in range(0, dim(q)):
    impact[i] = (c_q[i] − q[i])^2 / dist^2(c_q, q)

return impact
```

Nearest neighbour based method and clustering method are not equivalent and in general generate different results as it will be experimentally confirmed (section 5.2). The main difference is due to determining different neighbouring cluster of a new observation. Nearest neighbour based method simply selects the closest *k* good objects from historical dataset. Notice that these objects can be very diverse from each other. On the other hand, clustering method uses clustering as pre-processing step to create clusters of good objects. After that, factor ranking for a new observation is estimated considered good objects from only one cluster, the cluster that is the closest to that observation.

5. Case study - Emergency Department

In this section real life problem, length of stay explanation in an emergency department, is presented and usability of discussed algorithms is demonstrated.

5.1. Modelling the data

The raw data was exported by the author directly from information system of one hospital acquisition and management company. Specific data preprocessing procedures were necessary to be designed and implemented on the raw data to obtain historical dataset H of the form introduced in the previous section.

Eventually, the historical dataset is represented as data matrix that contains columns: *ED_VISITS* - total emergency department visits, *AMBULANCE_VISITS* - number of patients brought in by ambulance, *AVG_TRIAGE_LEVEL* - average triage level of all patients on specific day, *TRIAGE_LEVEL_1_2_COUNT* - number of patients with triage level 1 or 2 on specific day, and *DIVERSION_HOURS* - diversion hours number on specific day. Granularity of the historical dataset is on a day level.

The previous columns represent factors under consideration. The factors considered in this case study are suggested by domain expert.

In addition, there are three more columns: *AVG_LENGTH_OF_STAY*, representing average length of stay in hours in emergency department on a specific day, *DATE*, representing calendar date, and *FACILITY*, representing hospital name. Notice that *DATE* and *FACILITY* constitute primary key.

To conclude, every record from the historical dataset $H_{m\times n}$ is multidimensional vector of the following form *(DATE, FACILITY, LENGTH_OF_STAY, ED_VISITS, AMBULANCE_VISITS, AVG_TRIAGE_LEVEL, TRIAGE_LEVEL_1_2_COUNT, DIVERSION_HOURS)*. It means that $n = 8$. Total number of records after data preprocessing is $m = 602646$.

The raw data was separated among several tables in relational database. All records originated from emergency departments from four different hospitals in California. The covered period was from January, 2013 to April 2018. All hospitals belong to the same hospital management company, so transactional data from every emergency department are stored together in the same database.

Source tables are: Emergency department (ED), Triage level (TL), and Diversion (DV). Details are presented in Table 1. Types and attributes of each table are shown in Table 2, Table 3, Table 4.

Table 1. Source datasets characteristics

Dataset name	Number of records	Starting date	Ending date
ED	603006	2013-01-01	2018-04-14
TL	7474	2013-03-01	2018-04-12
DV	541936	2013-01-01	2018-04-14

Table 2. Types and attributes of ED datasets

Attribute	Type	Explanation
PATIENT_ACCOUNT	string	Unique patient account number
MRN	string	Medical record number of patient
ARRIVAL_TIMESTAMP	datetime	arrival timestamp of patient into the emergency department
DISCHARGE_TIMESTAMP	datetime	
PATIENT_TREATED	boolean	False=Patient registered then left; True=patient was actually treated
ICU_ADMIT	boolean	True=Patient was admitted from ED to the intensive care unit (ICU)
ADMIT	boolean	True=Patient was admitted to the hospital
MEDICARE_ICU_ADMIT	boolean	True=Patient was admitted from ED to the ICU and had Medicare insurance
MEDICARE_TREATED	boolean	True=Patient had Medicare insurance and was treated
UN_INSURED_TREATED	boolean	True=Patient had no insurance and was treated
LEFT_AFTER_TRIAGE	boolean	True=Patient left after being triaged
LEFT_BEFORE_TRIAGE	boolean	True=Patient left before being triaged
LEFT_WITHOUT_BEING_SEEN	boolean	True=Patient left without being seen (LWBS)
ELOPED	boolean	True=Patient left and being assessed by a nurse
AMA	boolean	True=Patient left against doctor's orders
TRANSFER	boolean	True=Patient was transferred to another hospital within the same system
AN_OTHER_HOSPITALS	boolean	True=Patient was transferred to a hospital outside system
EMS	boolean	True=Patient brought in by ambulance
EMS_ADMIT	boolean	True=Patient brought in by ambulance and was admitted
UN_INSURED_ADMIT	boolean	True=Uninsured patient was admitted to the hospital
TRIAGE_START_TS	timestamp	initial triage started
BED_ASIGN_TS	timestamp	bed was assigned to the patient
NURSE_TS	timestamp	the nurse saw and triaged the patient
PHYSICIAN_TS	timestamp	the physician has come in and done their assessment
ARRIVAL_MODE	string	the method of arrival: walk-in, ambulance, police, etc.
PAYER_CODE	string	initial payer code description
FACILITY_NAME	string	

Emergency department dataset contains high level data such as medical record number, patient account, insurance type, payer code, arrival mode etc. Most importantly, ED dataset contains all timestamps identifying starting and ending points of treatment procedure: arrival (*ARRIVAL_TIMESTAMP*), arrival to triage (*TRIAGE_START_TS*), triage to bed (*BED_ASIGN_TS*), bed to nurse (*NURSE_TS*), nurse to doctor (*PHYSICIAN_TS*), doc-

Table 3. Types and attributes of TL datasets

Attribute	Type	Explanation
CPT4_CODE	integer	
PATIENT_ACCOUNT	string	
ARRIVAL_TIMESTAMP	datetime	
HOSPITAL_NAME	string	

Table 4. Types and attributes of DV datasets

Attribute	Type	Explanation
HOURS_OF_DIVERSION	integer	number of hours the diversion occurred
DATE	date	
HOSPITAL_NAME	string	

tor to disposition (*DISCHARGE_TIMESTAMP*). Overall length of stay is calculated as the difference between arrival time-stamp and departure time-stamp $LENGTH\_OF\_STAY = DISCHARGE\_TIMESTAMP - ARRIVAL\_TIMESTAMP$. Of course, $LENGTH\_OF\_STAY = arrivaltotriage + triagetobed + bedtonurse + nursetodoc + doctodisposition$. Also, records from ED dataset contain information if patient was treated in ambulance, $EMS = True$. Based on it, number of ambulance visits per day is calculated.

To conclude, for the purpose of the analysis ED dataset is projected on a schema of the form *ED(FACILITY, ARRIVAL_TIMESTAMP, DISCHARGE_TIMESTAMP, EMS)*.

Triage level dataset, among others, contains information about CPT4 codes from which triage level for each patient visit can be extracted. Triage level is an integer value that describes degree of patient sickness. Value 1 for triage level means *very sick*. Value 5 means *not sick*.

To conclude, for the purpose of the analysis TL dataset is projected on a schema of the form *TL(FACILITY, ARRIVAL_TIMESTAMP, TRIAGE_LEVEL)*. The *TRIAGE_LEVEL* column is derived from the original *CPT4_CODES* column with the mapping provided in Table 5.

Table 5. TRIAGE_LEVEL mapping

CPT4_CODE	TRIAGE_LEVEL
99281	1
99282	2
99283	3
99284	4
99285	5
99291	1

Diversion dataset contains records about how many hours the emergency department had to shut down because it was full. Each record summarizes number of hours the diversion occurred for every day and every hospital.

For the purpose of the analysis DV dataset is projected on a schema of the form *DV(FACILITY, DATE, HOURS_OF_DIVERSION)*.

Finally, data matrix $H_{602646\times 8}$ with columns *(DATE, FACILITY, (AVG_LENGTH_OF_STAY, ED_VISITS, AMBULANCE_VISITS, AVG_TRIAGE_LEVEL, TRIAGE_LEVEL_1_2_COUNT, DIVERSION_HOURS)* can be obtained. The aim of defining the unique schema was to enclose necessary data from the three before mentioned datasets. To achieve the previous schema some specific transformations must be done on the original tables.

Length of stay for one visit is calculated as the difference between arrival time-stamp and departure time-stamp, $LENGTH\_OF\_STAY = DISCHARGE\_TIMESTAMP - ARRIVAL\_TIMESTAMP$. Both time-stamps are present in the ED dataset. Average length of stay, *AVG_LENGTH_OF_STAY*, is obtained by grouping records with the same date and facility name. Similarly, with grouping by day and facility and counting $DISCHARGE\_TIMESTAMP - ARRIVAL\_TIMESTAMP <= 24$ hours *ED_VISITS* - number of patients in ED per day and facility is found. It means that this study takes into account patients who spent less than one day in an emergency department. From available data (column *ADMIT* in ED dataset) it is possible to separate patients who were admitted to the hospital for further treatment - INPATIENT, from those who were only treated in emergency department - OUTPATIENT. But, in this case study only patients who spent less that 24 hours in an emergency department are covered, regardless of their admission to the hospital for further treatment (INPATIENT or OUTPATIENT).

Finally, by counting $COUNTIF(EMS = True)$ the number of ambulance visits per day is obtained.

The previous can be concisely expressed with the following pseudo SQL query:

```
SELECT date_part(ARRIVAL_TIMESTAMP)
  , FACILITY
  , COUNT(DISCHARGE_TIMESTAMP - ARRIVAL_TIMESTAMP <= 24 hours)
     AS ED_VISITS
  , COUNTIF(EMS = True) AS AMBULANCE_VISITS
  , AVG(DISCHARGE_TIMESTAMP - ARRIVAL_TIMESTAMP) AS AVG_LENGTH_OF_STAY
FROM ED
GROUP BY date_part(ARRIVAL_TIMESTAMP), FACILITY
```

In the previous query *date_part* stands for a function that can extract date part from *ARRIVAL_TIMESTAMP* time-stamp. In that way roll-up is performed by climbing up to the day level in the time dimension.

From the TL dataset *AVG_TRIAGE_LEVEL* and *TRIAGE_LEVEL_1_2_COUNT* are calculated with the following pseudo SQL query:

```
 SELECT date_part(ARRIVAL_TIMESTAMP)
   , FACILITY
   , AVG(TRIAGE_LEVEL) AS AVG_TRIAGE_LEVEL
   , COUNTIF(TRIAGE_LEVEL in (1, 2)) AS TRIAGE_LEVEL_1_2_COUNT
FROM TL
GROUP BY date_part(ARRIVAL_TIMESTAMP), FACILITY
```

On the DV dataset the following pseudo SQL query is run:

```
 SELECT DATE
   , FACILITY
   , SUM(HOURS_OF_DIVERSION) AS DIVERSION_HOURS
FROM TL
GROUP BY DATE, FACILITY
```

At the end, the data matrix $H_{602646\times 8}$ is generated by calculating natural join on the results of the previous three SQL queries.

5.2. Factor impact estimation

In the following experiments historical dataset H is filtered to contain records between January and August 2017, about 115K objects. The set Q contains observations between September and December 2017 for which *AVG_LENGTH_OF_STAY* is $\geq \sigma$. The threshold is set to $\sigma = 4$ hours. For this case study the value for σ was suggested by domain expert. In general, σ is a combination of looking at historical data and a benchmark leadership wants to hit.

Two methods for new observation explanation presented in the fourth section were applied. Because of the lack of space, in this section results of explaining only subset of Q is reported. The subset $Q_{exp} \subset Q$ contains the top 5 observations having the highest values for patient length of stay, 5 objects with the smallest patient length of stay that is $> \sigma$ and 5 objects around the median value of the objects from Q.

The results obtained from the method $kNN\_impact\_estimation$ are presented in figures Fig. 4, Fig. 5, Fig. 6. The new observation is identified with date that is shown as chart title. It can be seen that the method is able to clearly identify the most significant factor for every observation from the set Q_{exp}.

The results obtained from the method $clustering\_impact\_estimation$ are presented in figures Fig. 7, Fig. 8, Fig. 9. Titles of the charts are dates of the observations. It can be seen that the method is able to clearly identify the most significant factor for every new observation from the set Q_{exp}.

The results from the previous experiments indicate that the $clustering\_impact\_estimation$ provides clearer distinction between the most significant factor and the others. For example, the difference between the impacts of the most significant factor and the following factor is in average 0.43 for $clustering\_impact\_estimation$. The method $kNN\_impact\_estimation$ achieves 0.38 as the average difference between impact of the two most important factors.

Execution time of the proposed methods is measured on machine with Intel(R) Core(TM) i7-55000U CPU at 2.40GHz and 8GG of RAM memory. The

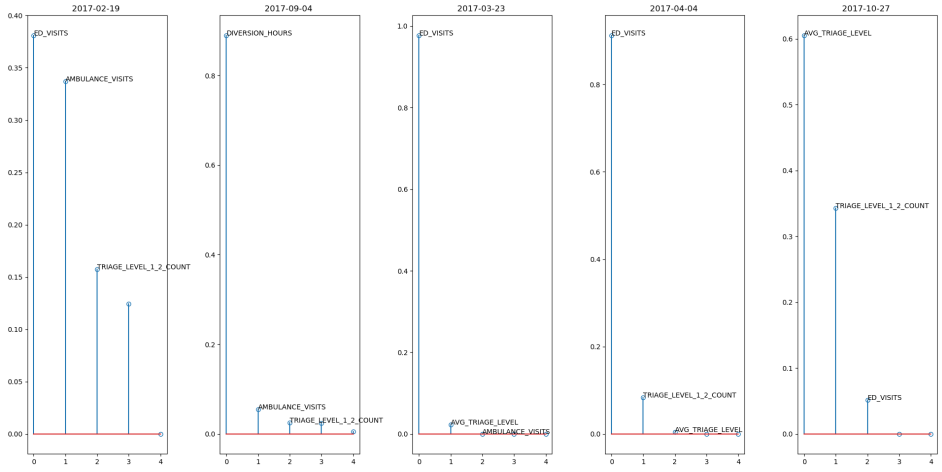

Fig. 4. The method $kNN\_impact\_estimation$ for the top 5 new observations

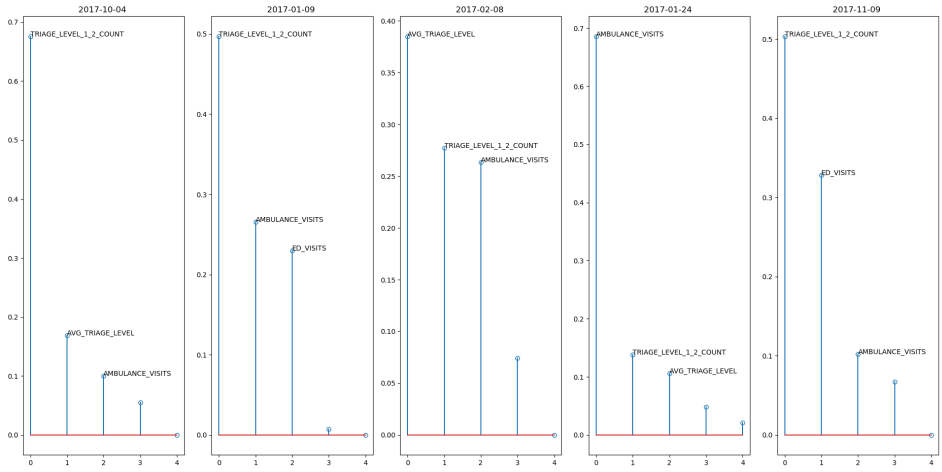

Fig. 5. The method $kNN\_impact\_estimation$ for 5 objects around the median

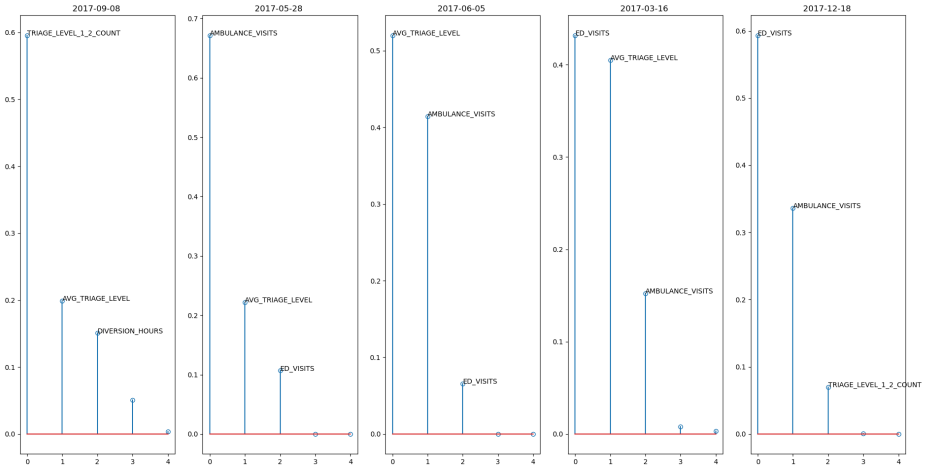

Fig. 6. The method $kNN\_impact\_estimation$ for the smallest 5 new observations

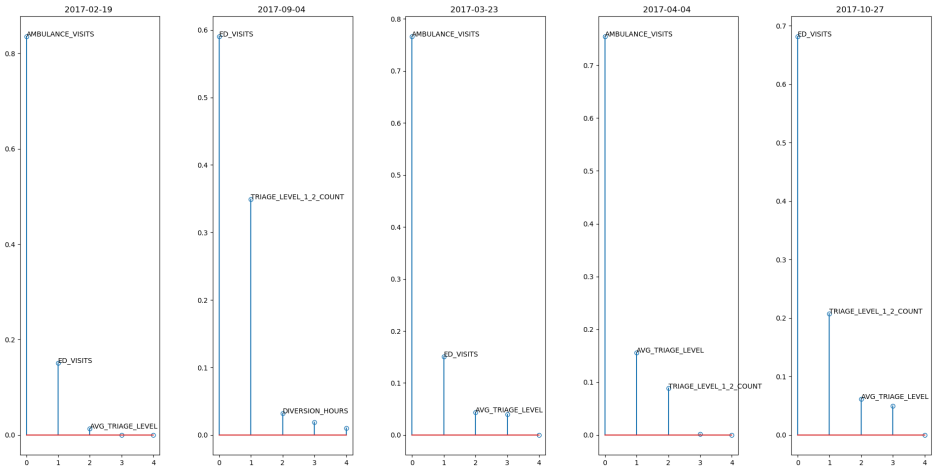

Fig. 7. The method $clustering\_impact\_estimation$ for the top 5 new observations

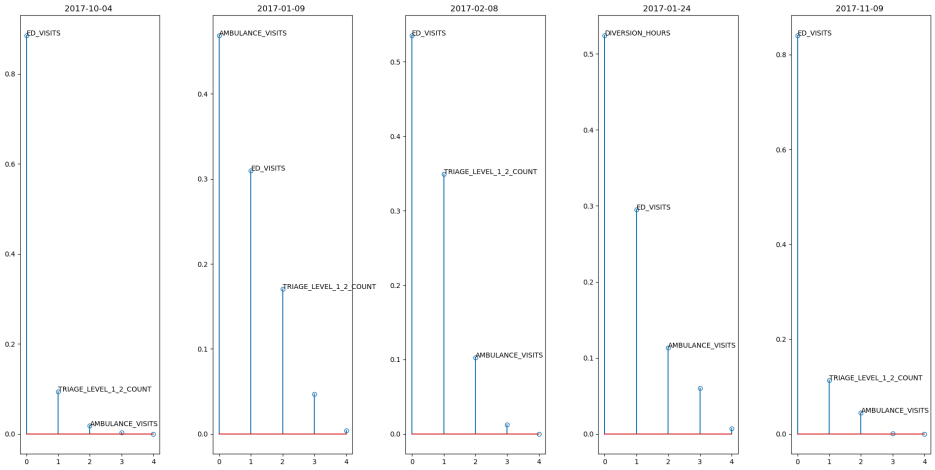

Fig. 8. The method $clustering\_impact\_estimation$ for 5 objects around the median

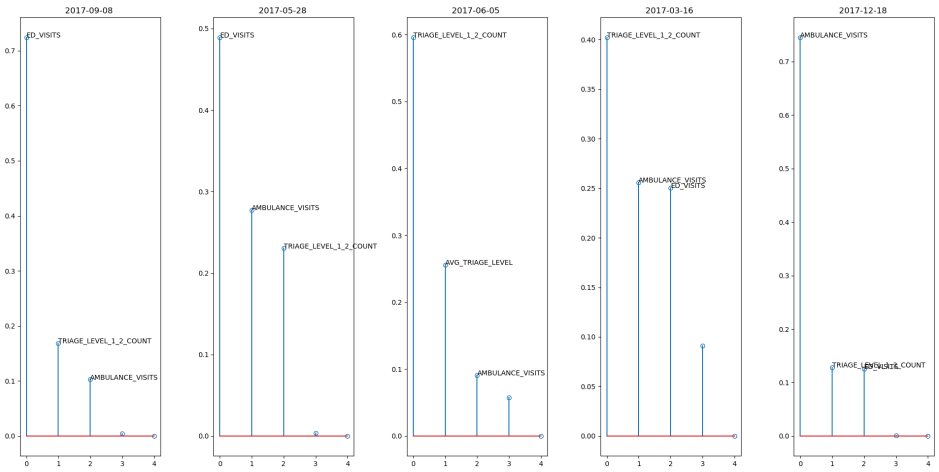

Fig. 9. The method $clustering\_impact\_estimation$ for the smallest 5 new observations

$clustering\_impact\_estimation$ requires 2.0005 seconds comparing to 6.8509 seconds that are necessary for $kNN\_impact\_estimation$ to finish.

Numerical and objective estimation of factors' impact allows user to determine major factors to specific anomaly (elevation in length of stay) and to express how much in % of total elevation this factor is contributing to the problem. For example, consider the observation representing 8th September from the figure Fig. 9. It can be seen that the most important contributing factor in this case is *ED_VISITS* contributing more than 70% to length of stay elevation registered on 8th September. Similarly, explanations for all other cases can be generated.

Length of stay is especially important parameter because it can be interpreted as efficiency. The results of such analysis should assist leaders of hospitals and its staff in understanding the "story" or "narrative" of their organization.

6. Conclusions

In this paper length of stay explanation problem - LOSEP is introduced. The problem consists of estimating impact of available factors on length of stay values that are higher than the threshold σ. Historical dataset is given. Objects from the historical dataset representing cases when registered length of stay is $\leq \sigma$ are referred to as good objects. The set of good objects can be considered as the knowledge database for the proposed methods. Observations of interest are new observations, possibly coming from a stream, for which length of stay is higher than σ. The system is queried to provide explanation about length of stay elevation on a new observation in a form of estimated importance of each factor.

The paper presents two methods: $kNN\_impact\_explanation$ and $clustering\_impact\_explanation$. In both approaches a new observation o is explained based on its neighbourhood.

The essential difference between $kNN\_impact\_explanation$ and $clustering\_impact\_explanation$ is in the process of finding the appropriate neighbourhood. The neighbourhood consists of good objects. In the method $kNN\_impact\_explanation$ the algorithm finds k-neighbourhood consisting of the closest k good objects. The algorithm from $clustering\_impact\_estimation$ finds neighbourhood of the q by clustering good object and determining the cluster with the closest centroid. Standard Euclidean distance is used to determine the distance between objects.

When the neighbourhood of the new observation q is determined, the impact for every factor is estimated. New observation q can be represented as $q = (factor_1(q), ..., factor_n(q))$. The procedure for objective impact estimation of each factor calculates increment to the SSE if the observation was added to the neighbouring cluster and distributes the increment value among factors proportionally.

Also, results of the case study in which proposed methods were applied on length of stay explanation in an emergency room are discussed. The historical dataset contains above 600K data objects. The following factors are considered: *ED_VISITS* - total emergency department visits, *AMBULANCE_VISITS* - number of patients brought in by ambulance, *AVG_TRIAGE_LEVEL* - average triage level of all patients on specific day, *TRIAGE_LEVEL_1_2_COUNT* - number of patients with triage level 1 or 2 on specific

day, and *DIVERSION_HOURS* - diversion hours number on specific day. Granularity of the historical dataset is on a day level.

Results of the analysis show that proposed methods are capable to recognize the most important factor and additionally to express how much in % of total elevation every factor is contributing to the specific observation.

Experiments show that the proposed two methods are not equivalent. In general, they assign different factors' impacts for the same observation. As future work, it can be interesting to implement voting schema and combine these two methods. Additionally, these methods potentially can be extended towards expert system that will be able to independently construct narrative explanation of the problem and propose possible actions regarding the situation.

References

1. Aghajani1, S., Kargari, M.: Determining factors influencing length of stay and predicting length of stay using data mining in the general surgery department. Hospital Practices and Research 1, 53–58 (2016)
2. Awad, A., Bader-El-Den, M., McNicholas, J.: Patient length of stay and mortality prediction: a survey. Health Services Management Research 30, 105–120 (2017)
3. Azari, A., Janeja, V., Mohseni, A.: Healthcare data mining: Predicting hospital length of stay (phlos). International Journal of Knowledge Discovery in Bioinformatics (IJKDB) 3, 44–66 (2012)
4. Bashkin1, O., Caspi, S., Haligoa, R., Mizrahi, S., Stalnikowicz, R.: Organizational factors affecting length of stay in the emergency department: initial observational study. Israel Journal of Health Policy Research 4, 1–7 (2015)
5. Breiman, L.: Bagging predictors. Machine learning 24, 123–140 (1996)
6. Buchman, T.G., Kubos, K.L., Seidler, A.J.: A comparison of statistical and connectionist models for the eprediction of chronicity in a surgical intensive care unit. Crit Care Med. 22, 750–762 (1994)
7. Castillo, M.G.: Modelling patient length of stay in public hospitals in mexico. Thesis (Doctoral), University of Southampton, Southampton Business School 1, 318pp (2012)
8. Chaou, C.H., Chiu, T.F., Yen, A.M.F., Chip-Jin, Chen, H.H.: Analyzing factors affecting emergency department length of stay—using a competing risk-accelerated failure time model. Medicine 95, 1–7 (2016)
9. Chua, J.M.: Factors associated with prolonged length of stay in patients admitted with severe hypoglycaemia to a tertiary care. Endocrinology, Diabetes Metabolism 1, 1–5 (2019)
10. Combe, C., Kadri, F., Chaabane, S.: Predicting hospital length of stay using regression models: Application to emergency department. In: MOSIM'14. vol. 124, pp. 672–674 (2014)
11. DO, D.R.C.: Analysis of Survival Data. Chapman HAll (1984)
12. Esfandiari, N., Babavalian, M.R., Moghadam, A.M.E., Tabar, V.K.: Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications 41, 4434–4463 (2014)
13. Friedman, J.: Greedy function approximation: a gradient boosting machine. Annals of statistics 29, 1189–1232 (2001)
14. Frye, K.E., Izenberg, S.D., Williams, M.D.: Simulated biologic intelligence used ot predict length of stay and survival of bums. J Burn Care Rehabil 17, 540–546 (1996)
15. Garg, L., Mcclean, S., BJ, B.M., Millard, P.: Phase-type survival trees and mixed distribution survival trees for clustering patients hospital length of stay. Informatica 22, 57–72 (2011)

16. Garg, L., McClean, S., Barton, M., BJ, B.M., Fullerton, K.: Intelligent patient management and resource planning for complex, heterogeneous, and stochastic healthcare systems. Systems, Man and Cybernetics, Part A: Systems and Humans 42, 1332–1345 (2012)
17. Golouke, N., Huibers, C., Stalpers, S., Taekema, D., Vermeer, S., Jansen, P.: An observational, retrospective study of the length of stay, and its influencing factors, among elderly patients at the emergency department. European Geriatric Medicine 6, 331–335 (2015)
18. Grubinger, T., Kobel, C., Pfeiffer, K.: Regression tree construction by bootstrap: Model search for drg-systems applied to austrian healthdata. BMC Medical Informatics and Decision Making 10, – (2010)
19. Hu, P.: A data-driven approach to manage the length of stay for appendectomy patients. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 39, 1339–1347 (2009)
20. Jus, E.: Factors influencing length of stay in the emergency department in a private hospital in north jakarta. Endocrinology, Diabetes Metabolism 27, 165–173 (2008)
21. Khosravizadeh, O., Vatankhah, S., Bastani, P., Kalhor, R., Alirezaei, S., Doost, F.: Factors affecting length of stay in teaching hospitals of a middle-income country. Electronic Physician 8, 3042–3047 (2016)
22. Liu, P., Lei, L., Yin, J., Zhang, W., Naijun, W., El-Darzi, E.: Healthcare data mining: predicting inpatient length of stay. Proceedings of the 3rd International IEEE Conference on Intelligent Systems Los Alamitos 1, 832–837 (2006)
23. Liu, Y., Phillips, M., Codde, J.: Factors influencing patients' length of stay. Australian Health Review 24, 63–70 (2001)
24. Marshal, A.H.: Conditional phase-type distributions for modelling patient length of stay in hospital. International Transactions in Operational Research 10, 567–576 (2003)
25. Marshal, A.H., McClean, S.I., Shapcott, C.M., Millard, P.: Modeling patient duration of stay to facilitate resource management of geriatric hospitals. Helath care management science 5, 313–319 (2002)
26. Marshall, A., Vasilakis, C., El-Darzi, E.: Length of stay-based patient flow models: recent developments and future directions. Health Care Management Science 8, 213–220 (2005)
27. Mobley, B.A., Leasure, R.: Artificial nerual network predictions of lengths of stay in a post-coronary care unit. Heart Lung 24, 251–256 (1995)
28. PR, P.H., Ahmadi, M., Alizadeh, S., Sadoughi, F.: Use of data mining techniques to determine and predict length of stay of cardiac patients. Healthcare informatics research 19, 121–129 (2013)
29. SixSigma: Box plot diagram to identify outliers (2019), available from: `https://www.whatissixsigma.net/box-plot-diagram-to-identify-outliers/`
30. TS, T.L., Loh, W., Shih, Y.: A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine Learning 40, 203–228 (2000)
31. Vapnik, V.: The Nature of Statistical Learning Theory. Springer (2000)
32. Yoon, P., Steiner, I., Reinhardt, G.: Analysis of factors influencing length of stay in the emergency department. Can J Emerg M 5, 155–161 (2003)
33. Z, Z.Y., Liu, P., Lei, L., Yin, J.: R-c4. 5 decision tree model and its applications to health care dataset. Proceedings of ICSSSM 2005 2, 1099–1103 (2005)

Savo Tomovic received his PhD in computer science from University of Montenegro in 2011. During his PhD studies he was involved in the project Linear Collider Flavour Identification (LCFI) with the aim to compare different data mining and classification algorithms as well as to understand the relative importance of the various input variables for the resulting tagging performance. He is currently an associated professor in the Faculty

of Science at University of Montenegro and Head of the Computer Science Department. He teaches a wide variety of undergraduate and graduate courses in several computer science disciplines, especially data mining, machine learning and data warehousing. In addition, he is currently engaged as consultant in several software companies on projects for design and implementation of cognitive systems and data warehouse models.

Received: April 22, 2020; Accepted: January 27, 2021.