Data-driven models for mortality assessment at the Intensive Care Unit

Research Paper Business Analytics

Ali el Hassouni
Vrije Universiteit Amsterdam
Faculty of Sciences
The Netherlands
a.el.hassouni@student.vu.nl

Supervisor: Dr. Mark Hoogendoorn
Vrije Universiteit Amsterdam
Faculty of Sciences
The Netherlands
m.hoogendoorn@vu.nl

ABSTRACT

The Intensive Care Unit (ICU) is known as the department with the highest mortality numbers in any hospital. Patients at the ICU require extensive assessment by hospital staff. This is partially achieved through the use of state-of-theart monitoring devices that provide measurements and trends about each patient. This means that large amounts of useful data is available for research. With the increasing number of patients, maintaining a high standard of care at the ICU becomes time consuming and brings high operational costs with it. In this paper we develop data-driven models that can be used by hospital staff to assess the physical state of patients at the ICU. To be more precise, these models provide hospital staff with predictions about the risk of mortality of patients at the ICU. Two models were developed and compared based on accuracy and scalability. The first approach follows a pipeline that prepares the data for predictive modelling with logistic regression. The second approach allows the use of the instance-based learning model K-nearest neighbor (KNN) with Dynamic Time Warping (DTW). We show that logistic regression (AUC of 0.84) significantly outperforms KNN (AUC of 0.68) by conventional criteria. This paper provides insights on the performance of both algorithms and could be used as inspiration for further research.

1. INTRODUCTION

Successfully applying machine learning techniques to health data can make a big difference in the medical domain [6]. Machine learning allows the use of techniques that can help caregivers make more informed assessments about the state of their patients. The fields of application vary from diagnosing heart disease, cancer prognosis and detection to mortality assessment at the ICU [6]. On the one hand, the massive amount of health data within Electronic Medical Records (EMRs) offers tremendous value through its high variety if used wisely. On the other hand, the variety of the data also means that the task of extracting meaningful features can be very complicated. For instance, an EMR of a patient can contain structured data such as waveform data as well as unstructured data such as notes written by caregivers. Furthermore, the data of patients in EMRs can be multidimensional and incomplete which makes the task of modelling even more complicated.

To address the complexity and disparity of data in EMRs, two feature extraction and data aggregation pipelines have been designed. The first pipeline describes the extraction of features in a way that allows the use of predictive modeling techniques such as logistic regression. Hereby, features are extracted and data is aggregated over a time dimension. Furthermore, other machine learning techniques such as feature engineering and data imputation are applied. The second pipeline prepares the data for instance-based learning. Here, we focus on the k-nearest neighbor algorithm. This algorithm requires the computation of patient similarities. This is achieved by using two different ways of computation: the Euclidean distance and the Dynamic Time Warping algorithm [4]. Herewith, we aim at handling the disparity of the data caused by for example waveforms that are out of phase or exist over non-equal time periods.

Both approaches considered in this paper have proven to perform well on medical data [3],[6]. However, literature shows that the number of studies that compare the two approaches is negligible [3],[10]. This has to do with the fact that most research conduct on this topic mainly focused on predictive modelling. This paper seeks to compare predictive modeling with instance-based learning in two ways. The models are compared in terms of their accuracy at predicting mortality of patients at the ICU and their scalability to the number of patients taken into consideration. For the first approach inspiration was drawn from [3]. This decision enables us to compare the results obtained with our logistic regression model with the ones obtained by [3]. For the second approach a customized implementation of the k-nearest neighbor algorithm with Dynamic Time Warping has been developed.

The structure of this paper is as follows. First, we give a brief literature review in Section 2. Section 3 gives a description of the data that has been used in this paper. Thereafter, we describe the methods used to develop our models in Section 4. Section 5 shows our experimental setup. Finally, we discuss the results and draw conclusions in Section 6.

2. RELATED WORK

Much work has been done on the field of machine learning for medicine. Several applications of machine learning in the

medical field are organ localization, tissue classification, disease prognosis and detection, and computer-aided mortality assessment. The applicability of these techniques in real life situations has been thoroughly discussed. A machine learning model is labeled as useful in medical applications if it meets the following criteria: good performance, ability to deal with incomplete and noisy data, transparency, capacity to explain its decisions, and the ability to perform in an acceptable amount of time [6].

The application of machine learning for mortality assessment in the ICU in particular has been widely investigated. Most of existing models rely on logistic regression for making predictions [1],[3]. Logistic regression is known for its simplicity and transparency [6]. Other algorithms that have been researched are survival models [3], Bayesian Networks [3], Neural networks [1] and the k-nearest neighbors [7] algorithm. Bayesian Networks and survival models are nearly as good as logistic regression, but these models are less stable at predicting the outcome [3]. The Neural networks and k-nearest neighbors algorithms have been researched using different datasets and proved to perform nearly as good as logistic regression [1],[7].

The k-nearest neighbor algorithm is known for its simplicity and effectiveness in handling numerical values [6]. Defining an appropriate distance measure is a key component in the implementation of this algorithm. The most commonly used techniques are Dynamic Time Warping and the Euclidean distance. Research also shows that little work has been done to compare a predictive modelling approach (e.g. logistic regression) with an instance based model (e.g. k-nearest neighbor) [3],[10].

3. DATA DESCRIPTION

In this section a brief description of the dataset is given. Data selection and pre-processing are discussed in detail. Hereto, inspiration is drawn from [3]. The MIMIC-II V2.6 database [2],[8] contains a wide range of detailed measurements of a big number of patients. This data is collected between 2001 and 2008 at a teaching hospital in Boston. The data comes from workstations at the ICU as well as hospital archives. The database contains monitoring information such as patient demographics, physiological metrics, waveforms, and trends. Furthermore, chart data such as fluid balance, medications, and reports are available. Demographics and background information are the only type of data we use from the hospital archives. This choice is made to limit the scope of the research to patients at the ICU. The database contains the following types of data:

1. **Continuous and ordinal measurements** consist of 6 categories, namely: cardiovascular, chemistries, hematology, arterial blood gases, ventilation, and miscellaneous. The total number of features across all categories is 64. This group of measurements consists of numerical values (continuous and ordinal) that result from data observed from patients. For a full description see [3]. For each feature, statistically rigorous methods were used to define valid ranges and remove outliers [3]. Additionally, hold limits are defined and used for applying the hold approach. This method describes the amount of time a measurement is retained and reused for the imputation of missing values in future time points until a new measurement is available.
2. **Categorical measurements** consist of 35 features that are binary or ordinal in nature. For a full description see [3]. Similar to the previous group of measurements, the hold method is applied.
3. **Medication measurements** contains 51 medications along with the doses medicines a patient administered. Some medicines are accompanied with per-kilogram units dose values while others have absolute dose values or both. For each medicine it was made sure that both types are available. Furthermore, for each medicine a binary feature was added that describes the presence (1) or absence (0) of the medication. For a full description see [3].
4. **Input/output measurements** contain features that describe input measurements related to blood and output measurements related to urine production. For a full description see[3].
5. **Demographics** are constant in nature and do not vary in time. The features that were included from this category are: age, ethnicity, and sex. For a full description see[3].

To further increase the quality of the data patient filtering is applied. Patients who do not meet the following criteria were removed from the dataset [3]:

- At least one Blood Urea Nitrogen observation
- At least one Glasgow Coma Scale observation
- At least one Hematocrit observation
- At least one Heart rate observation
- At least one Intravenous medication recorded
- Receive adult care (are not neonates)

Further preprocessing steps include the hold approach and choosing the granularity of the data [3]. Applying the hold approach means imputing missing values in future time points if these entries are within the provided hold ranges. The granularity of the data was set to one entry for every 15 minutes.

The original dataset contained information from 26647 patients. This amassed to a raw dataset of 120 features and 29550651 entries. After excluding patients with incomplete data (e.g. patients who left or died within a couple of hours from admission to the ICU) the dataset was reduced to 13923 patients.

4. METHODOLOGY

This section describes the methods used throughout the paper. Section 4.1 gives an overview of the machine learning approaches. Sections 4.2 and 4.3 describe the feature extraction for respectively the logistic regression and the k-nearest neighbors approaches. Sections 4.4 and 4.5 describe the setup to respectively the logistic regression and k-nearest neighbor models.

4.1 Machine learning approach

To enrich the set of features that are used by the logistic regression model and the k-nearest neighbor algorithm, feature extraction and selection were applied. These techniques usually help obtain higher quality results in models that use medical data [5]. Basically, one can abstract features from temporal data in various ways. Here we use two different approaches. These two approaches are described in detail in Sections 4.2 and 4.3.

4.2 Feature extraction for logistic regression

The first approach abstracts features on the time level. Hereto, inspiration was draw from [3]. Depending on the type of data, feature extraction is performed in various ways:

- **Continuous and ordinal variables**
  For this type of variables the minimum, maximum, mean value and standard deviation are calculated over a certain time period. Furthermore, a linear regression model is fit on observed values and time. The best-fit line obtained is used to derive the slope for time windows of 4 hours, 28 hours, or both depending on the type of variable [3].

- **Categorical variables**
  For variables that are categorical in nature the mean value over the full time period is calculated. This is performed after transforming the variables to binary or ordinal types. This kind of extraction aims at capturing the history of the variables. Further details can be found in [3].

- **Medications**
  For medication variables the mean dose of each medicine administered during the full time period is calculated.

- **Input/Output variables**
  For Input/Output variables the same approach is followed. The mean value over the full time period is calculated for each variable.

- **Demographics**
  Variables that describe demographics are constant and therefore do not need to be aggregated over time.

- **Derived variables**
  Additional variables were calculated to capture information from the data. Meta-information that describes the presence or absence of variables, ratio's, and temporal behavior were calculated as described in [3].

The Stationary Daily Acuity Score (SDAS) approach from [3] was implemented. 11 PM is chosen as the time at which a new day starts to be able to systematically assess results. The data is aggregated using different periods of time. We start with the aggregation of the data generated during the first 24 hours at the ICU starting from 11 PM. This is extended with data from the next 24 hours and so on. This type of aggregation allows us to compare models trained on data from one or multiple 24 hour periods.

To perform variable selection Pearson correlation is used. Variables with the highest correlation with the target variable are selected when they are not highly correlated (≥ 0.2) with a variable that already has been included.

4.3 Feature extraction for k-nearest neighbors

The second approach relies on feature extraction on the measurement level. To achieve this, statistical measurements are calculated. For the k-nearest neighbors algorithm a different feature extraction and selection approach is used. For each day that a patient stays at the ICU entries are added for each measured time point with granularity of 15 minutes. For each entry a predefined set of variables are filled in from the dataset. Furthermore, missing data is imputed using the hold technique as explained in Section 3 and features are normalized to a range of 0 to 1.

The dataset contains two sorts of features: time series data and measurement features. The features that belong to the first category are: the heart rate, respiration, the nocturnal, systolic, and diastolic blood pressure, and the oxygen saturation. All remaining features are not time varying and thus belong to the second category. For this category the average over the entries of patients is taken and the Euclidean distance is used to calculate the similarity distance. For the time-varying features two different approaches were implemented. Both approaches are based on the Dynamic Time Warping algorithm for determining patient similarities. The first variant uses the Keogh lower bound [4] to determine the similarity between waveform data from two different patients. This technique makes it feasible to calculate the distance between these features for large datasets. Since this approach approximates a lower bound a second variant with the name FastDTW was implemented. FastDTW gives an approximation that is near-optimal in linear time [9]. A detailed description of the methods is given in Section 4.5.

4.4 Logistic regression model

The first patient model we developed is based on the logistic regression algorithm. This model is known for its transparency, understandability within the field of medicine, robustness and tractability [3],[6]. Logistic regression belongs to the class of models called generalized linear models (GLMs). Given n observed data x_1, x_2,..., x_n and n independent target variables Y_1, Y_2,..., Y_n our goal is to model the relationship between the target variables and p non-random covariates. With logistic regression one can model the log odds of the binary target variable (e.g. mortality). These odds are modeled using a linear combination of covariates X. The logistic regression model for the independent target variables Y_1, Y_2,...,Y_n is defined as follows:

$$n_i Y_i \sim Bin(n_i, \mu_i), \tag{1}$$

$$\eta_i = X_i^T \beta, \tag{2}$$

$$\eta_i = g(\mu_i) = log\left[\frac{\mu_i}{(1-\mu_i)}\right], \tag{3}$$

$$\mu_i = P(Y_i = 1|X_i), \tag{4}$$

for i = 1,...,n.

The first formula describes the random component of the model. This component specifies the distribution of the target variables Y_i. The second formula defines the systematic component of the model which describes a vector of predictors η_i for each observation x_i and the way the covariates are contained into the model. Finally, the link-function is defined in the third formula. This function specifies the link between the first and the second components of the model. After rewriting the probability $P(Y = 1|X)$ becomes:

$$log\left(\frac{P\left(Y=1 \mid X\right)}{1-P\left(Y=1 \mid X\right)}\right) = X\beta \tag{5}$$

$$P\left(Y=1 \mid X\right) = \frac{1}{1+e^{-X\beta}} \tag{6}$$

With this probability we can obtain the risk of mortality for each patient.

4.5 K-nearest neighbors model

The second model we developed is based on a custom implementation of the k-nearest neighbours algorithm. This algorithm uses patient similarity to perform instance based learning. The customisation is needed to calculate similarity scores for features that have a time component (e.g. time varying features). These features can be highly time varying, shifted or different in size. For the calculation of the similarities the Dynamic Time Warping algorithm is used as mentioned in Section 4.2. For the remaining features the Euclidean distance is used as a similarity measure.

Define a set DTW of time series feature, a set EUC of the remaining features and a set $FEAT$ that contains all features. Let $t_{i,j}$ be a vector of values of patient i for feature j during a predefined time period of n time points. The similarity between two patients A and B is described by equation 7.

$$distance(A,B) = \frac{\sqrt{DTW(A,B)^2 + Euc(A,B)^2 + Penalty(A,B)}}{features\_matched(A,B)} \tag{7}$$

$$DTW(A,B) = \sum_{i \in DTW} \begin{cases} keogh\_LB(t_{A,i}, t_{B,i}) & \text{if } (|t_{A,i}| > 0) \wedge (|t_{B,i}| > 0) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

$$Euc(A,B) = \sum_{j \in EUC} \begin{cases} (\langle t_{A,j} \rangle - \langle t_{B,j} \rangle) & \text{if } (|t_{A,j}| > 0) \wedge (|t_{B,j}| > 0) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$Penalty(A,B) = C \cdot (1 + |DTW| + |EUC| - features\_matched(A,B)) \tag{10}$$

$$features\_matched(A,B) = \sum_{k \in FEAT} \begin{cases} 1 & \text{if } (|t_{A,k}| > 0) \wedge (|t_{B,k}| > 0) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Due to the sparsity of the data it is common that measurements are missing. Therefore, features are only compared when a least one value is present for both patients. Furthermore, the total distances between patients are averaged by dividing over the total number of matching features. Finally, a weighting scheme is applied by adding a penalty factor where C determines the weighting. Given the number of nearest neighbors (k) we look up the k patients with the lowest similarity scores and calculate the risk of mortality using the number of patients from class 1 (patients that unfortunately died at ICU) and k.

5. EXPERIMENTS

In this section we describe the machine learning approaches for setting up the models, validation methods and parameter settings.

5.1 Experimental Setup

To assess the results obtained with the experiments in a systematic way, choices about the validation methods and the accuracy metrics need to be made. With stratified cross-validation one can assess how well the obtained models will generalize on data from new patients. 5-fold cross-validation is used and the two classes in each folds are kept in proportion to each other (i.e. stratified cross-validation). To investigate the influence of the size of the dataset on the accuracy as well as the scalability the number of patients is varied between 150 and 2500. Furthermore, we experiment with different ways of calculating the similarity between patients for KNN. The accuracy metric used is the AUC. For the run times we use seconds to quantify the scalability of the algorithms.

5.2 Parameter Settings

For the Logistic Regression model a grid search over the set of parameters is run. L1 and L2 regularization are tested and the number of features to be used by the model is varied. The cost parameter C was varied between 0 and 1000 with step size 50 and the the following tolerance values were tested: 1e-6 and 1e-4. For the KNN model we create different subsets of features based on the percentage of missing values (85%) and the type of data. Furthermore, we optimize for the number of nearest neighbors.

6. RESULTS

In this section we will present the results we obtained with our models. Sections 6.1 and 6.2 describe the results obtained with respectively logistic regression and KNN. In Section 6.3 we compare the two approaches.

6.1 Logistic regression

For the model trained with logistic regression we applied the Stationary Daily Acuity Score (SDAS) approach as described in Section 3. Here, we use data of the first 24 hours at the ICU starting from 11 PM. Furthermore, with feature engineering we derived additional features from the original

Feature	Coefficient
neosynephrinek	37.83487
min_tidvolset	32.72291
natrecor	9.366317
slope_windows_8_sbp	8.004235
max_spo2	4.49859
min_cl	3.511009
insulin	1.542605
slope_windows_16_spo2	1.530865
max_wbc	1.137946
iabp_bin	-1.52952
std_tidvolobs	-1.60236
pale_skin_bin	-2.0729
white	-2.19223
slope_windows_16_nbpmean	-3.40944
slope_windows_24_nbpsys	-23.9617

Table 1: Top 15 features with highest coefficients

dataset. The final dataset used to train this model contained 748 features. Next, feature selection was applied. Here we take the full dataset after preprocessing and select the subset of features with the highest correlation with the target variable. We also take correlations between features into account. Hereby, we leave out features that are highly correlated with features that already have been selected for their high correlation with the target variable. The size of the subset of features was varied between 30 and 70. Furthermore, different correlation thresholds were tested.

To further increase the accuracy of the model, L2-regularization was used with a cost of 150 and a tolerance of 1e-6. These parameters were optimized using the grid search method. A subset of 50 features and a correlation threshold of 0.2 yielded the highest AUC. Stratified cross-validation with 5-fold was used to validate the models. The results shown in figure 1 were obtained with a model trained with data from 2000 patient (3.5 million rows) and 50 features. An AUC of 0.84 was obtained with this set of patients. Table 1 shows the top 15 features with the highest coefficients for this model.

The size of the dataset was varied to test the influence of the number of patients on the accuracy of the model. The same 50 features that were selected before are used to train these models. Figure 2 shows that the accuracy of the model increases as we increase the number of variables. After reaching 2000 patients the accuracy stagnates and even decreases from 0.84 to 0.83 when we increase the number of patients. Regarding the run times of the models trained with logistic regression we can state that they all finish training in under then 40 seconds. 50% of the time is spent on reading the files, while the other 20 seconds is spent on feature selection and training.

6.2 K-nearest neighbor

For the model trained with the k-nearest neighbor algorithm we applied a different approach for feature extraction. For this model the data is not aggregated but just stored in a time uniform manner. Section 4 goes more in depth about the precise steps of feature extraction. Furthermore, with

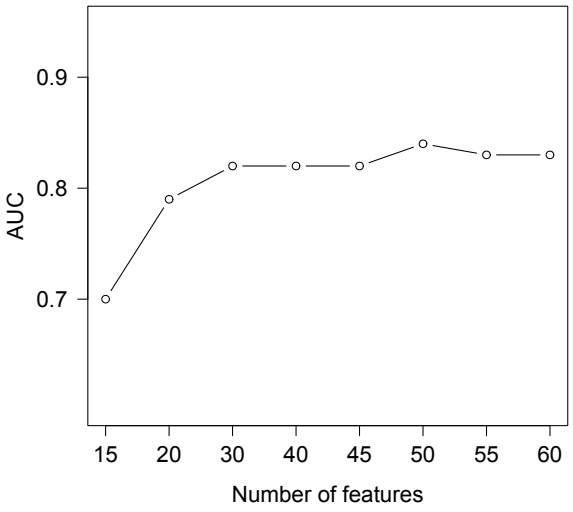

Figure 1: Sensitivity of LR to the number of features.

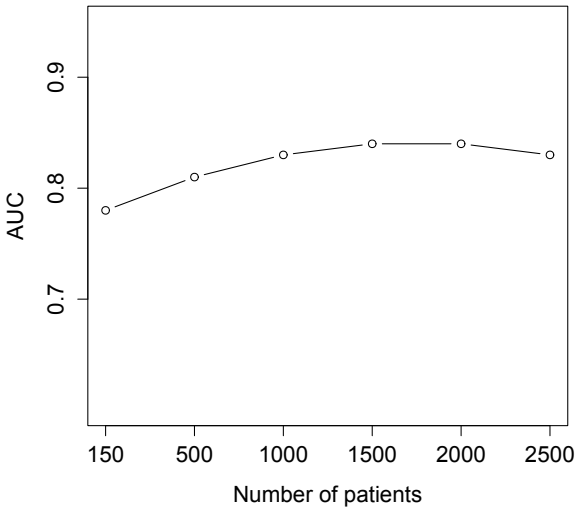

Figure 2: Sensitivity of the LR model to the number of patients used for training.

feature engineering we derived additional features from the original dataset. The final dataset after preprocessing contained 350 features.

The feature selection approach used here is also different in comparison with the first model. Based on this dataset we can distinguish between two types of features. These are time-varying and not time-varying features. The following time series features were selected manually: the heart rate, respiration, the nocturnal, systolic, and diastolic blood pressures, and oxygen saturation. For these features the lower bound of Keogh was used to calculate the similarities as ex-

plained in Section 4.

Further, another version of the dynamic time warping algorithm was implemented. This algorithm is called FastDTW and is implemented in such a way that it performs faster than the original DTW algorithm and usually yields more accurate similarity scores than the lower bound of Keogh [4]. FastDTW was tested and did not improve the accuracy significantly. Additionally, it made the algorithm much slower compared to the lower bound of Keogh. Therefore, the lower bound of Keogh seemed the better choice.

For the remaining features we use the Euclidean distance as a similarity measure. Different ways of calculating similarities were experimented with. The Cosine distance, the Mahalanobis distance and a different version of the Euclidean distance where the number of observations in a certain time period is taken into account, were implemented and tested. None of these alternative distance calculations yielded an improved accuracy compared to the standard Euclidean distance. Therefore, we opted for the latter a distance measure for its simplicity.

Finally, the parameter k (e.g. the number of nearest neighbors used to calculate the risk of mortality) was optimized using randomly drawn datasets of sizes up to 1000 patients. Figure 3 shows the accuracies yielded with the KNN models after varying this parameter. Stratified cross-validation with 5-folds was used to validate the models. From this figure we can see that $k = 1$ yields the highest accuracy while for higher values of k (up to 15) the accuracy seems to decrease.

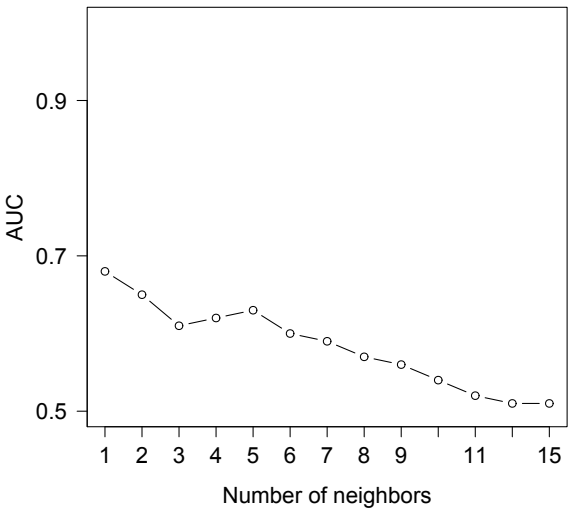

Figure 3: Sensitivity of the KNN model to the number of neighbors.

Based on these results we can now test how the accuracy of the model changes after increasing or decreasing the size of the dataset. Different models were trained with varying sizes of datasets (from 150 to 2500 patients). The number of neighbors is set to 1 and the number of features used to train the models is kept at 132. Figure 4 shows the result obtained with varying datasets. From this figure we can clearly see that the accuracy of the model decreases as we increase the number of patients. Reasons for this unexpected behavior could be the fact that many features will not have a match due to missing data, and that the distance function is not robust in approximating the true distance given the number of features we use and the missing data.

k	With DTW	Without DTW
1	0.68	0.66
2	0.65	0.63

Table 2: Accuracy with and without Dynamic Time Warping obtained with 150 patients

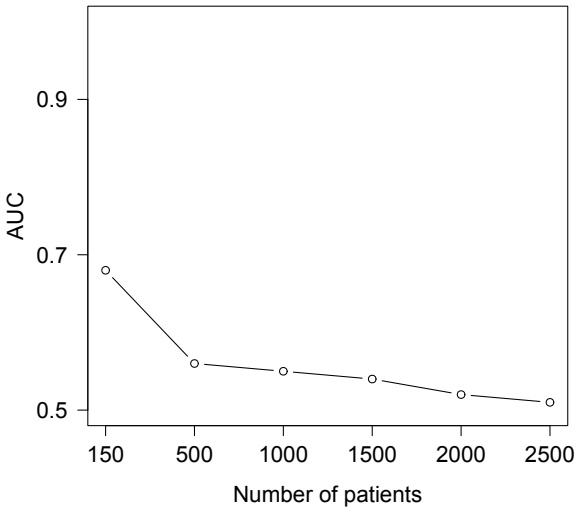

Figure 4: Sensitivity of the KNN model to the number of patients.

A final experiment was carried out to test whether the DTW approximation algorithm (e.g. lower bound of Keogh) has any added value. Hereto, similarities for time varying features were calculated using the Euclidean distance. Table 2 shows the accuracies obtained with the original model and the new model that used the Euclidean distance on all 132 features. From this table we can see that using an approximation of the DTW algorithm gives a significant improvement compared to using only the Euclidean distance. Using the data obtained with the 5-fold cross validation a t-test was applied and showed that the difference is significant ($p<0.05$). Regarding the run times of the models trained with the KNN algorithm, we can state that these seem to increase exponentially with the size of the dataset. Figure 6 shows the run times of both models.

6.3 Logistic regression versus KNN

Figures 6 & 7 show the best ROC curves obtained with the logistic regression and k-nearest neighbors algorithms. To compare the two algorithms a paired t-test was applied. The data used for this test was taken from the different runs with different numbers of patients. The alternative hypothesis is that the different in means of the accuracy (AUC) is greater than zero. The test yielded a p-value of 0.00275. Based on conventional criteria one can conclude that the logistic regression model performs significantly better than the k-nearest neighbor algorithm.

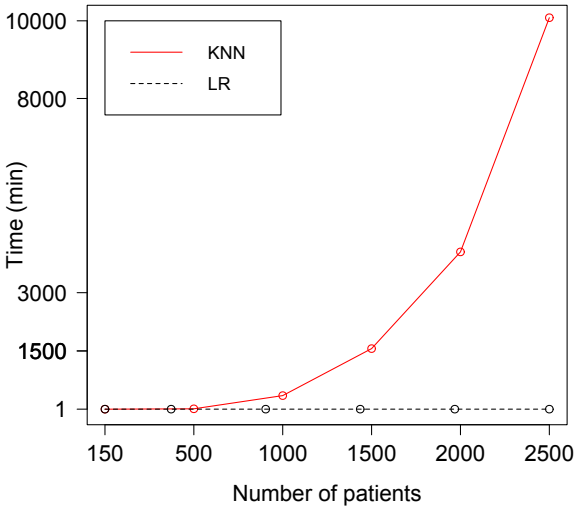

Figure 5: Run times of KNN and LR

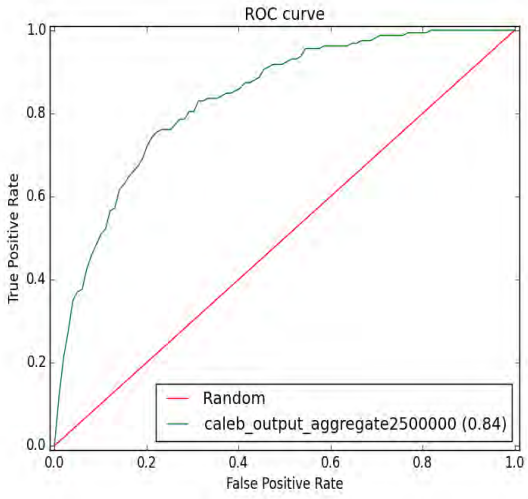

Figure 6: ROC curve for logistic regression

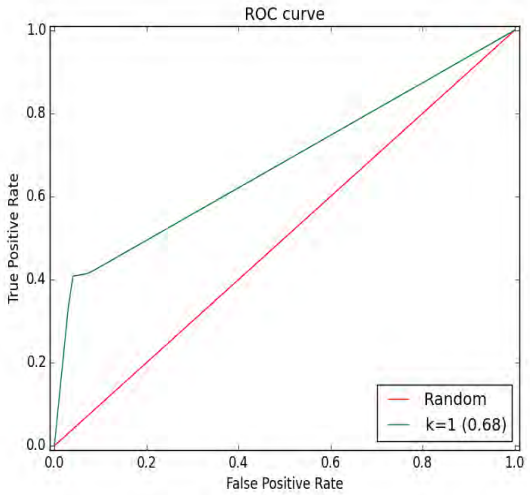

Figure 7: ROC curve for KNN

7. DISCUSSION

In this research paper two different modeling approaches were implemented and compared. The first approach follows a pipeline that prepares the data for predictive modelling with logistic regression. The second pipeline allows the use of the instance-based learning model KNN. For the first approach data of patients was aggregated over a time dimension. In the second approach data was stored in a time uniform manner. For each model well though out experiments were performed to increase the quality of the results. In this paper we aimed at comparing the two approaches based on their accuracy and scalability.

Predictive modelling resulted in a model that could predict the outcome of an ICU stay based on data of at most 1 day with an AUC of 0.84. This model was trained on 50 features that have been preselected. The top 15 features with the highest coefficients obtained with this model are shown in table 1. Neosynephrinek has been selected by the model as the most important feature. This feature indicates whether the drug Neosynephirine is injected into the blood of a patient or not. Research shows that this drug is intended to maintain the blood pressure of patients during inhalation anesthesia or to treat vascular failure in cases of shock, hypotension or hypersensitivity. During all these scenarios patients are in a life threatening state which explains why this feature is very important for the model. Furthermore, features that indicate minimum, maximum, mean, standard deviation, or slope values were the most selected features. This indicates that the aggregation of data over the time dimension had added value. When making a comparison with [3], one can see that several features have been selected by both models while the remaining features are different. The model obtained by [3] achieves an AUC of 0.89 when predictions are limited to day 1. Furthermore, we see that [3] uses 35 features that are selected using a slightly different feature selection method. Also, the model obtained by [3] contains features that were not used in this paper. Regarding the run time of logistic regression, it was shown that this model runs relatively fast. Training and testing takes under one

minute to finish. This means that this model can be useful in practical situations.

Instance based learning resulted in a model that was able to predict the outcome of an ICU stay based on data of at most one day with an AUC of 0.68. This model was trained with 132 features that were selected based on the percentage of missing values. This algorithm searches for the k-nearest patients from the set of training patients. Based on the proportions of the classes of the these patients, the risk of mortality is calculated. As a consequence this algorithm is not able to generalize well. Also, it lacks transparency of knowledge which is represented by the sum of information gains in favor or against a given class [6]. This also makes it difficult to understand the behavior of the algorithm that seems to become less accurate as the number of patients in the training set is increased as can be seen in figure 4. Furthermore, a value of 1 for the parameter k yielded the highest AUC. A hypotheses to explain such behavior would be the problem of sparsity and the effect this has on the calculation of the similarity scores. Using a similarity score function that is not robust would result in an increased chance of calculating scores with high deviance compared to the true ones. This would mean that the calculation of the risk of mortality gets less accurate. Furthermore, the fact that highest AUC is obtained with k=1 indicates that the algorithm relies on the smallest similarity scores which have a lower chance of being less accurate.

Besides the accuracy of the model, its scalability was also measured. In Section 6 we showed that the run time of the k-nearest neighbor increases exponentially as we increase the number of patients in the training set. This could be explained by looking at the way the KNN algorithm compares patients and performs similarity calculations. Each patient in the training set should be compared to every other patient to calculate the similarity scores. An attempt was made to use clustering to decrease the set of patients that each instance will be compared to. The clusters were created by looking at the age of the patient and the Glasgow Coma Scale at the moment of arrival at the ICU. The resulted clusters were highly imbalanced in their sizes which made the process of cross-validation unreliable. This would also make the process of analyzing and trying to understand the behavior of the algorithm more complicated.

8. CONCLUSION

A physician should be able to understand the predictions made by the algorithm. This can lead to new insights that were not recognized by the physician before for the given problem. Another important point is the ability of the model to explain its diagnosis or prediction. Physicians will usually not accept the predictions of a black box model unless it outperforms their assessments largely. On these two points predictive modelling proves to be better than instance based learning. The algorithm should also be able to make a prediction in a practical amount of time. On this point logistic regression clearly outperforms KNN. Finally, one can conclude that even though KNN is known for its simplicity and its ability to perform well on medical datasets, it does not do well on the MIMIC-II dataset when we compare it to logistic regression. More sophisticated ways of selecting features and imputing missing values for KNN would be worthwhile to explore in feature research. Additionally, approaches such as clustering could be tested where a bigger dataset is used. All these suggestions could help improve the scalability and accuracy of the algorithm and provide an instance based model that is more accurate and much faster.

9. ACKNOWLEDGEMENTS

This research paper has been written for the course Research Paper Business Analytics as part of the Business Analytics Master program at the Vrije Universiteit Amsterdam. The goal of this course is to carry out a research on a subject that is related to the program Business Analytics. The individual skills of researching, writing a report and giving an oral presentation are tested during this course.
I would like to thank my supervisor Dr. Mark Hoogendoorn for introducing me to this subject and for providing me with insights and support. Furthermore, I would like to thank Marzyeh Ghassemi and Dr. Peter Szolovits for providing me with the data to conduct this research.

10. REFERENCES

[1] B. Eftekhar, K. Mohammad, H. Ardebili, M. Ghodsi, and E. Ketabchi. Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC Medical Informatics and Decision Making, 5(1):1–8, 2005.

[2] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. Circulation, 101(23):e215–e220, 2000.

[3] C. Hug. Detecting hazardous intensive care patient episodes using real-time mortality models. 2009.

[4] E. Keogh and C. A. Ratanamahatana. Exact indexing of dynamic time warping. Knowledge and information systems, 7(3):358–386, 2005.

[5] I. Kononenko. Estimating attributes: analysis and extensions of relief. In Machine Learning: ECML-94, pages 171–182. Springer, 1994.

[6] I. Kononenko. Machine learning for medical diagnosis: History, state of the art and perspective. Artif. Intell. Med., 23(1):89–109, Aug. 2001.

[7] M. Makar, M. Ghassemi, D. M. Cutler, and Z. Obermeyer. Short-term mortality prediction for elderly patients using medicare claims data. International Journal of Machine Learning and Computing, 5(3):192, 2015.

[8] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. Critical care medicine, 39(5):952, 2011.

[9] S. Salvador and P. Chan. Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis, 11(5):561–580, 2007.

[10] Y. Xiao, M. P. Griffin, D. E. Lake, and J. R. Moorman. Nearest-neighbor and logistic regression analyses of clinical and heart rate characteristics in the early diagnosis of neonatal sepsis. Medical Decision Making, 30(2):258–266, 2010.