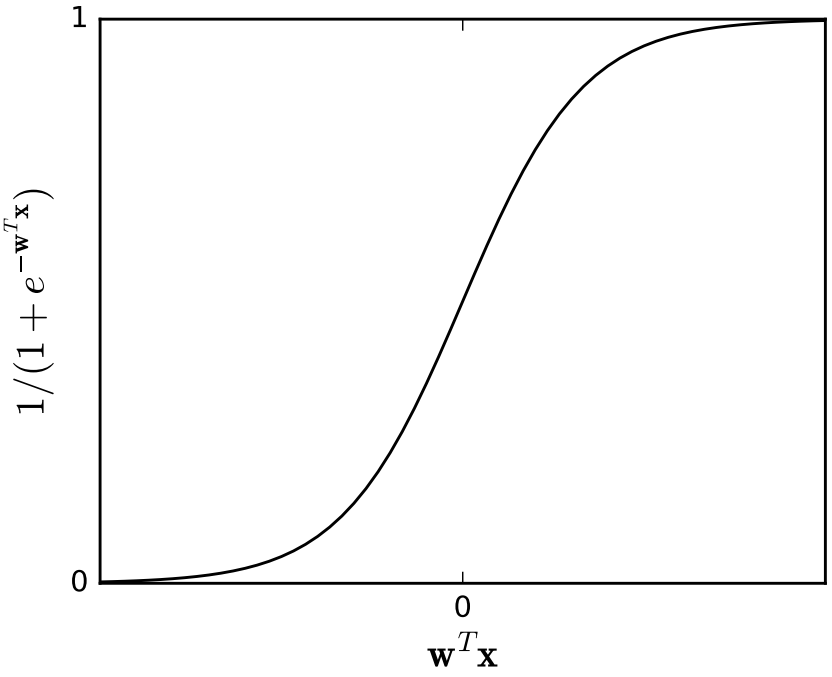

Predicting Patient Behaviour in Swedish Health Care Using Machine Learning

An Evaluation of Predictive Models and Their Utility in Risk Stratification

Master's Thesis in Software Engineering

PÄR LINDER

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY & UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2016

Predicting Patient Behaviour in Swedish Health Care Using Machine Learning

An Evaluation of Predictive Models
and Their Utility in Risk Stratification

PÄR LINDER

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Division of Software Engineering
Chalmers University of Technology & University of Gothenburg
Gothenburg, Sweden 2016

Predicting Patient Behaviour in Swedish Health Care Using Machine Learning
An Evaluation of Predictive Models and Their Utility in Risk Stratification
PÄR LINDER

© PÄR LINDER, 2016.

Supervisors:
FREDRIK JOHANSSON, Computer Science and Engineering
ERIK WIKLUND, Sahlgrenska University Hospital
Examiner:
REGINA HEBIG, Computer Science and Engineering

Department of Computer Science and Engineering
Division of Software Engineering
Chalmers University of Technology & University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Logistic function

Typeset in LaTeX
Gothenburg, Sweden 2016

Predicting Patient Behaviour in Swedish Health Care Using Machine Learning
An Evaluation of Predictive Models and Their Utility in Risk Stratification
PÄR LINDER
Department of Computer Science and Engineering
Chalmers University of Technology & University of Gothenburg

Abstract

There is a vast amount of patient data stored in health care record systems. Together with the rise of computing power this data could be used for advanced analysis of this data, and incorporate it in applications for use in daily operations. This is a case study in which unbalanced archival data from emergency room admissions is used for machine learning, in order to develop three models that predict the possibility of a patient returning to emergency room within 72 hours. The best of these model uses a logistic regression classifier and has a recall of 1% and a precision of 50%. The implementation of such a model in daily operation is discussed with a new approach to cost benefits. Despite the low predictability, the study is a proof of concept of predictive modeling in a health care context.

Keywords: Machine Learning, Predictive Models, Logistic Regression, Health Care, Patient Behaviour Prediction

Acknowledgements

I would like to thank my supervisors, Erik Wiklund at the Strategic Analysis Unit at Sahlgrenska University Hospital and Fredrik Johansson at the Department of Computer Science and Engineering at Chalmers University of Technology for their great enthusiasm and invaluable support.

Pär Linder, Gothenburg, October 2016

Contents

List of Figures

List of Tables

1

Introduction

Ever since the introduction of information and communications systems (ICT) in health care, the amount of health care data recorded has increased at a rapid rate. However, despite the large amount of observational data stored, the potential of using this data for strategic and operational decisions through the means of data analysis is, especially in Swedish health care, rarely acknowledged. In later years, the interest of the analysis of large amounts of data in general has become apparent under the common label big data and with it has emerged the recognition for many a technique related to prediction and data analysis, amongst them widely popular concepts including artificial intelligence and machine learning, which potentially could be used for assessing the aforementioned large amount of health care data. Also, in the absence of admitted machine learning related methods using health care data, the recognition of its potential and its development may be restricted due to lack of examples or limited resources.

One goal of this study is to advance knowledge within the area by applying machine learning to health care data, in order to develop a model that uses health care data to predict patient behavior, providing a basis for medical decisions, or in medical terms, risk stratification. In order to be able to achieve this goal, this study focuses on a prominent property of emergency patients, namely the fact that around nine percent of patients sent home from the emergency room (ER) return within 72 hours. Were these patients to be identified before they return, decisions regarding their care could be taken in order to reduce the risk of them returning, but more importantly improve patient safety and cut costs. Also, in order to gain new knowledge about what is causing the patient to return, it is important that the model provides a way of identifying important return factors.

In order for such a model to function, two criteria have to be fulfilled: It is possible, to some degree, to predict the possibility of a patient returning within 72 hours due to the patient's characteristics and previous medical history, and the data needed for doing such a prediction exists in current data records.

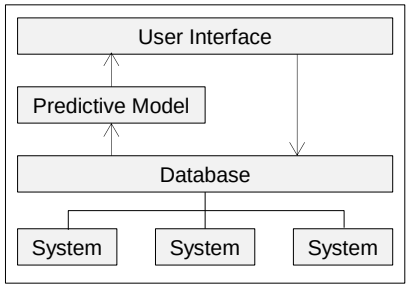

Figure 1.1: Application Architecture. A simplified concept. A user interface where the user inputs data which is communicated to the database, which in turn feeds the predictive model that returns its result to the UI. The database stores all data and interfaces all the different data record systems at the hospital.

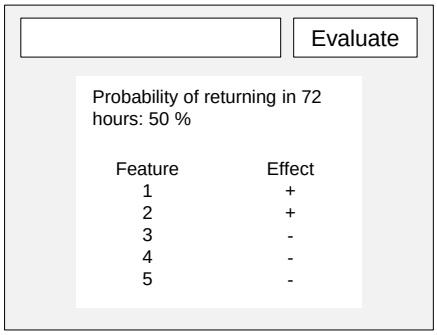

Figure 1.2: Mock-up UI. Upon entering a personal identification number, the probability of that person returning to the emergency within 72 hours appears, together with the features affecting this probability and their effect on the prediction.

Implemented in daily health care, the model could be part of an application in which health care personnel may assess the risk of the emergency patient returning within 72 hours. Such an application could, for example, have a simplified architecture as shown in 1.1 and a user interface as shown in 1.2 Four main components necessary for such an application's existence are recognized:

1. User Interface - The interface with which the user interact with the application
2. Model - The model used for predicting patient behaviour
3. Data Access - Data extraction from existing data record systems at the hospital
4. Database - Aggregation and storage of data

The second point is related to another goal of this study, namely to find what model to use for predicting patient behaviour. In order to elevate this study from a theoretical to a realistic concept, it is assumed that there already exists such an application where all components except the model already exist. The UI has the following components: A text field for entering the patients personal identification number, and a field in which the probability of a patient returning within 72 hours

together with the factors contributing to this probability appears upon entering a personal identification number. This is achieved through the following process: When a personal identification number is entered, model input data is extracted from the database and entered into the model, which communicates the predicted possibility to return to the UI.

Naturally, all the four aforementioned components need to be developed in order to get a fully functional, implementable product, not to mention the process of implementing the application in daily health care.

Summarized, the goal of this study is to simplify the software engineering problem regarding model selection, while at the same time proving the utility of predictive modeling in health care. It is important to remember that this study is performed with real world implementation in mind. The discussion about a real software implementation will be left out during the data analysis which focuses on model selection, but will return in the implementation chapter.

To cover all aspects, three research questions were formulated as follows:

- **RQ 1:** Is it possible to predict whether a patient admitted to any of the main emergency rooms at Sahlgrenska University Hospital (SU) will be come back within 72 hours?
- **RQ 2a:** Which predictive model is most suitable to implement in an actual application?
- **RQ 2b:** Why is that predictive model most suitable to implement in an actual application?

In the eyes of a software engineer, the most important takeaways from this study would be what type of predictive model to use in an implementation project, and why such a model would be suitable. It should be noted that the findings in this thesis are specific to the particular setting, but could be generalizable to similar environments, and at least helpful when taking decisions about predictive model implementation.

From a general health care point of view, this study serves as a proof of concept on what can be achieved when applying machine learning to large amount of data. Model selection is also of importance, but rather from a model understanding perspective, that is the assessment of the trade-off between predictability and how well the model may be interpreted. The overlap between the software engineering and health care perspectives is visualised in figure 1.3.

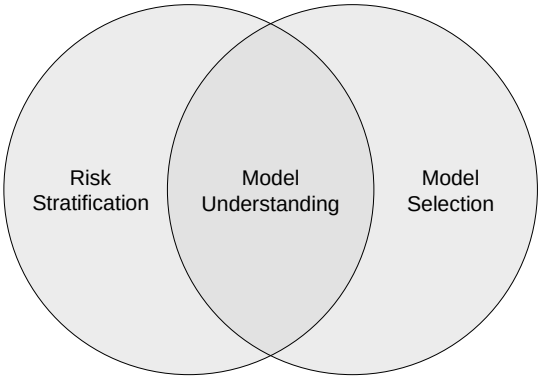

Figure 1.3: Health Care - Software Engineering overlap. From a health care perspective, the most valuable takeaway is how data and the model can be used to help medical decision-taking, or risk stratification. From the software engineering perspective, which model to choose and how to develop it is important. The overlap covers understanding: Health care personnel must be able to understand the model in order to interpret its output and how it relates to patient behaviour, and the software engineer must understand why the model is suitable for this specific setting and in what other settings it would work.

1.1 Limitations

Health care data access is sometimes difficult and time consuming, which is reflected in a relatively limited feature selection. Only three different models were evaluated in order to be able to fit within the time scope of this project. During work progress, many interesting ways of analyzing data were discussed. However, feasibility of analyses was carefully assessed in order to be relevant in the area of machine learning and within interest for the health care.

This study is a proof of concept, and the implementation of its result in real care is not assessed. The process of validating and securing a model for health care use is too extensive for the scope of this project. However, it is worth mentioning that this would be a natural follow-up to this project.

2

Background

In this chapter, key background knowledge related to this study is presented. This includes a general description of the current state regarding use of data analysis in health care, narrowing into the theoretical background of the methods used in this study, namely logistic regression and random forest. Lastly, related work within the field is described.

2.1 Data Analysis in Health Care

Murdoch and Detsky (2013) endorses the medicine field for recognizing the value of data earlier than many other industries, but that the use of newer techniques to analyze the vast data contained in electronic health records (EHR) is lagging, and that the data is largely regarded a by-product rather than a central asset to improve efficiency. This is important in relation to the study as to prove the usefulness of large sets of data by showing how it can be used as a basis for decision making.

Provost and Murray (2011) describe a lack of clarity on how to approach data. This leads to paralysis to take action, incorrect tool selection, sampling issues, incorrect analysis and incorrect conclusions regarding improvement work. Provost and Murray (2011) also mention that advanced methods for use with health care data are not available to policy makers. According to Faries, Obenchain, Haro and Leon (2010), the use and importance of data analysis in health care has grown but statistical analysis of observational health data is less well-developed compared to that of randomized clinical trials. They emphasize the lack of guidance on statistical analysis:

"Low-quality analyses and limited experience with such data by many decision makers has led to a lack of optimal use and even mistrust of such work. Several research groups, recognizing these analytical and reporting issues, are starting to provide general guidance on improving the quality of such analyses. However, there is still a lack of practical detailed guidance on implementing such methodology."

It is clear that the process of conducting advanced statistical analysis needs facilitating. Should this study prove the usefulness of such analysis, it will help bridge the

gap between the execution of analysis and data. However, there are recent examples of applying advanced statistical methods in health care contexts. Razavian et al. (2015) present a machine learning approach to predicting type-2 diabetes patients, and also mention a few other successful examples, such as Krumholz (2013) and Wang et al. (2013). This study attempts to further improve the use of statistical methods and knowledge in the area from a Swedish health care context.

2.2 Domain Background

Sahlgrenska University Hospital (SU) is the largest hospital in Sweden. It had about 128,000 emergency admissions in 2015, whereof 83,000 were at any of the three main emergency rooms. These emergency rooms are located at Sahlgrenska in the center of Gothenburg, at Östra Sjukhuset (Eastern Hospital) east of Gothenburg Center, and in Mölndal, south of Gothenburg.

For every admission at the hospital, health information about the patient is registered in any of the many registration systems. This information includes amongst other sex, age, diagnosis registration and timestamps. Above this, the emergency room visits also contain emergency department, emergency cause, mode of arrival and where the patient was sent after the admission - home, or admitted to another department.

A patient's diagnosis is described with the ICD 10 system, which is a standardized diagnosis dictionary, where each diagnose consists of a letter and some following numbers. Non ST-elevation myocardial infarction (I21) for example, is grouped under Diseases of the circulatory system (I00-I99), Ischemic heart diseases (I20-I25).

2.3 Machine Learning

In order to understand what the wide concept of machine learning means and why it is useful, it can be compared to statistics. An example, on how to use statistics from the field of physics, is how to calculate the gravitational constant. In order to do this, several measures are made of, for instance, the time it takes for an object to fall from a certain height (or different heights) to the ground. The measures are then used to calculate the gravitational constant. In other words, data is collected, and conclusions from the data are drawn.

Machine Learning shares a lot of similarities to statistical analysis from which it borrows many tools and concepts, but has a different focus on the outcome. A machine learning model uses a set of data, from which it learns how the data behaves. From that learning process, the model will be able to predict the outcomes of new measures.

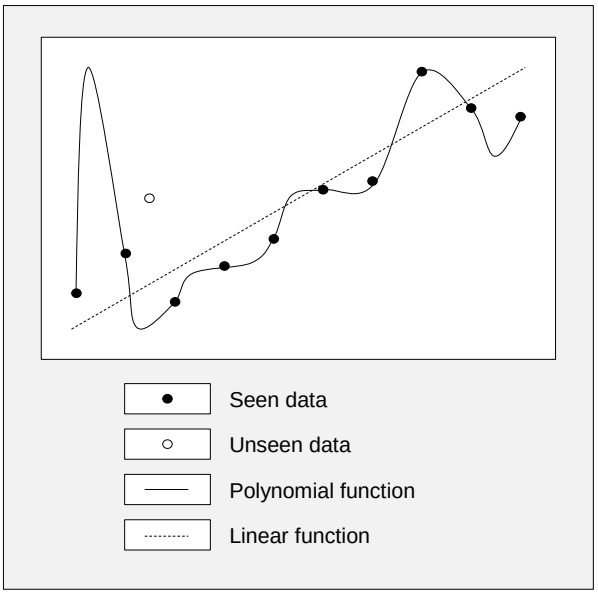

Figure 2.1: Linear Regression and Overfitting. A complex polynomial function can be fitted perfectly to seen data. However, it cannot predict the outcome of new data, for which the straight linear function of linear regression works better. The black line is thus overfitted.

The purpose of learning is to find a generalized pattern in the data, which then can be applied to new data to predict its outcome. In figure 2.1 some data points are plotted. If this is our training data, the pattern we found could be described with either the dotted curve or the black straight line.

The black line, which could be expressed as some high degree polynomial function, does indeed cross all the points, and is therefore very well fitted to the data - or as will be described, overfitted, but is also very complex. The dotted line however, follows some general pattern in the data - as the x value increases, the y value seems to increase too, linearly, and thus it has a high simplicity. In contrast to the black line, it only intersects a few of the points.

If a new data point is introduced, it seems unlikely that it falls on the dotted line, but it probably will lie somewhere close to the dotted line due to the pattern of the data. This means that the dotted line generalizes quite well and is applicable for predicting outcomes of new data, whilst the black line only describes our training data - it is overfitted. There is a chance, however, that some model generalizes even better to unseen data, that has a higher complexity than the linear model, but is not overfitted and as complex and as the polynomial. In other words, when selecting a model, there is a trade-off between model simplicity and data fit.

Logistic regression is a generalized linear model, sharing some similarities with linear regression, with the exception that the predicted value is binary (0 or 1). It uses the logistic function to map the outcome of a linear model to 0 or 1, hence the name *Logistic* regression (Murphy, 2012).

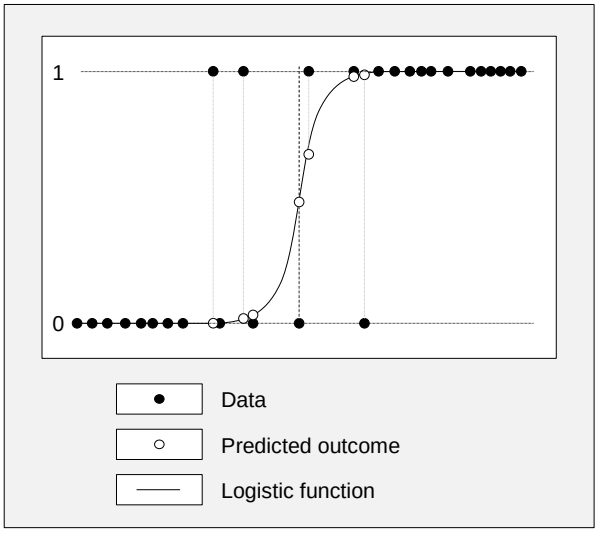

Figure 2.2: Logistic Regression. Binary outcome data is plotted on the lines $y = 0$ and $y = 1$. Using logistic regression, the points are mapped onto the logistic function and assigned a probability.

In figure 2.2 is an example of data where the outcome is binary. The higher the x-value, the larger the possibility of y being 1 or true. A linear function is not suitable to describe these data points. Instead, in logistic regression, the S-shaped sigmoid, or logistic, function is used to predict the outcome of y.

A simple linear regression model (only one explanatory variable) tries to minimize the error of the linear function

$$y(x) = \beta_0 + \beta_1 x$$

where y is the response variable, β_0 is the intercept, β_1 is the coefficient and x is the input variable. In a simple logistic regression model, the linear regression model is mapped to the logistic function, and gets the form

$$y(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

When more variables (features) are introduced, it becomes

$$y(x) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x})}}$$

where $\mathbf{w}$ is the vector of weights applied to each feature contained in the vector $\mathbf{x}$ (which also includes the intercept). Logistic regression can be trained on one or many independent categorical or continuous features. The logistic regression will calculate a weight for each feature. In order to filter out features that do not affect the outcome much, a penalty in the form of l1 regularization can be introduced:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \ell_{\mathbf{w}}(x, y) + \lambda||\mathbf{w}||_1$$

where ℓ is the original loss function to be minimized, and $\lambda||\mathbf{w}||_1$ is the introduced regularization term. λ can be adjusted in order to force weights for insignificant features to be zero, and thus reduce the number of features and increase the sparsity of the model.

Random forest is another model used for regression (or classification). This is achieved by constructing a number of randomly generated decision trees (with a predefined depth), from which the average prediction values are calculated. A large advantage with random forest over linear models is that its complexity (number of trees, tree depth, number of features) can be arbitrarily chosen. With great complexity however, comes often difficult interpretation. For a detailed description of Random Forest, see Zhang and Ma (2012) or Breiman (2001).

In machine learning, in order to train and evaluate the generalization capabilities of a model, data is normally split into two main data sets - a training set, on which the model is trained, and a test set, used to test the model's performance when it is optimized. Test data must not influence the parameters of the model, thus is it crucial that the test data is never used to train or to optimize the model.

During training, the model's performance is evaluated on different parameters (for example the size of the l1 penalty). These parameters are called hyperparameters. In order to evaluate the model with regards these parameter without using the test set, the training set is further split into a training set and a validation set, where the validation set is used to find the best hyperparameter. The model performance may be defined as one of many scores, but one common score is the Area Under Curve (AUC), which is the area under the Receiving Operating Characteristic curve (ROC). The ROC is a curve with the rate of true positives (correctly predicted positive outcomes) on the y-axis and the rate of false positives (incorrectly predicted positive outcomes) on the x-axis. The area under this curve is 50% when guessing randomly, and the larger it is the better the predictability.

One common technique to optimize the hyperparameters is cross validation. Here, the training data is split into different folds. Iteratively, each fold is left out from the training data and used to evaluate the model. Then, an average of the model performance is calculated ('cross-validation', 2016).

In order to compare different models statistically, one technique is bootstrapping. Bootstrapping means drawing samples from a test set repeatedly, and then the models' results using the sampled data are tested (with for example a paired t-test) in order to find whether the models' performances are different.

2.4 Related Work

Logistic regression is an often used technique when predicting patient outcomes, especially in combination with clinical trials. For example, Baan et al. (1999) use logistic regression to identify undiagnosed diabetes. Random Forest also occurs, Khalilia, Chakraborty and Popescu (2011) uses a random forest model together with data collected from publicly available health registers. These studies are good examples of that the logistic regression and random forest are functioning methods for performing data analysis in a health care setting.

When it comes to prediction of patient behaviour involving patients' own decisions, not much material can be found. An example, however, is Obermeyer, Powers, Makar, Keating and Cutler (2015) who predicts patient's decisions about end-of-life care, by looking at the characteristics of patient's physicians. This is important in relation to this study, as the decision to revisit an emergency room is indeed taken by the patient.

New opportunities have risen in the field of patient prediction due to available electronic health record systems (Jensen, Jensen & Brunak, 2012; Wu, Roy & Stewart, 2010; Roden, Xu, Denny & Wilke, 2012). These opportunities go in line with novelty claims of this study.

Using only readily available health record data in electronic health registers, Razavian et al. (2015) presents a retrospective study predicting type 2 diabetes, Louis et al. (2014) makes a prediction of risk of hospitalization or death; and Kontio et al. (2014) predicts patient acuity. As this study also uses that type of data, these studies and their methodology provided a solid background on how to process and use that type of data.

According to Hillestad et al. (2005), enhanced health data record systems could facilitate predictive modeling and thus decrease cost due to the insights of these predictions. However, these cost benefits are not assessed in the other studies mentioned in this section. This study, however, does indeed try to approach the cost benefits of implementing predictive models in daily decision making.

3

Method

This chapter describes how this study was carried out. Starting with a general description of the study design, it further details the different steps of the study in chronological order.

3.1 Study Design

This study is carried out in the form of an exploratory case study as described by Runeson and Höst (2009). However, the data used is quantitative, and archival. Despite this, this study is exploratory since it is done in order to generate new ideas and insights. Runeson and Höst (2009) mentions five major process steps for a case study:

1. Defining objectives and case study planning
2. Preparation for data collection
3. Collecting evidence: Execution with data collection on the studied case
4. Analysis of collected data
5. Reporting

The case outline follows these process steps (in parenthesis below) and was divided into three phases: In the preparatory phase a relevant health care issue with much accessible data suitable for machine learning was defined (1). After that, the research question was refined, and the preparation of data collection began (2).

The main phase can itself can be described as a retrospective cohort study as described by Duignan (2016). In this stage, Data was collected and prepared as described in the sections below. After that, the development of the models began. Three models were selected which were individually developed. Each model was trained and optimized with regards to adjustable hyperparameters. The models were then compared using statistical tests (3), and further evaluated with regards to practical use (4). The best model was then implemented in a prototype.

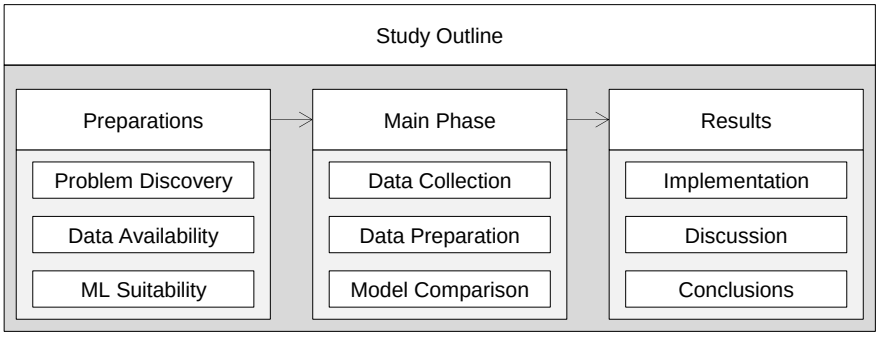

Figure 3.1: Study Outline. The study had three phases: Preparations, a Main Phase and Results. After the sub-tasks of a phase were finished, the study carried on to the next phase.

Lastly, as the last phase the results of the study were evaluated, including an implementation analysis, discussions, and conclusion (5). The study outline is visualized in figure 3.1.

3.2 Data Collection

The data was collected with help from the strategic analysis unit, which operates a database consisting of data from the years 2011 to present day from numerous registration system at the hospital. Most features were recorded in the data explicitly, while some had to be calculated by looking at previous data records, such as number of previous diagnoses the past 12 months.

Emergency admissions (registered emergency visits) during 2015 was selected as the cohort, with the number of previous diagnoses and visits counted one year back, as exemplified in figure 3.2. The data was stored in a table, as demonstrated in figure 3.3. Data that was regarded possible to have an effect on whether a patient would return or not was extracted, which resulted in a table with features, types and ranges as described in table 3.1.

The data was prepared by dichotomizing the categorical features, which means that each unique value of those feature were added as columns, and assigned a 1 or a 0, as exemplified in figure 3.4. This was done in order for the classifiers to be able to process the data. For use in the extended logistic regression model, a second data set was created, where the interaction between all features were included as well, as exemplified in figure 3.5

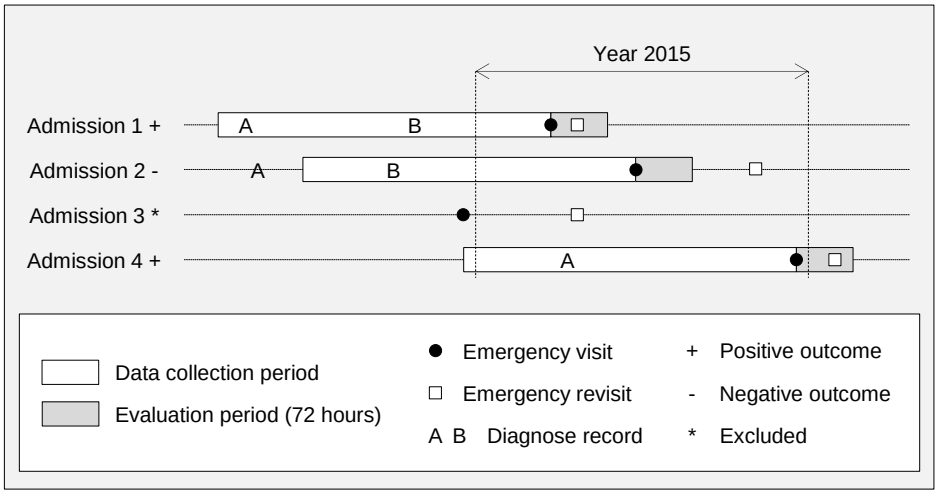

Figure 3.2: Data Collection Framework. All admissions to the emergency rooms during 2015 were considered. For each admission, twelve months' historical health data was considered in order to gather data for historical features. In the figure, an example of how diagnosis data collection was performed is shown, with diagnoses inside the data collection period considered. For each admission, a time of 72 hours forward was considered for whether the admission resulted in a patient revisit to the emergency room (denoted as +) or not (denoted as -). Admissions not occurring in 2015 were not considered (denoted as *). For the example in this figure, the resulting data would be as shown in table 3.3.

Admission	Revisit within 72 hours	Number of diagnoses	Diagnosis
1	1	2	B
2	0	1	B
4	1	1	A

Figure 3.3: Example Data. Resulting example data collected as described in figure 3.2. Only admissions during 2015 are included. The Number of diagnoses is the sum of distinct diagnoses during the data collection period, and the diagnosis column represents the last primary diagnosis.

3.3 Feature Details

Most of the selected features (shown in table 3.1) are trivial and easy to understand. Some, however, require a more detailed explanation. Emergency department includes the Mölndal, Östra and Sahlgrenska emergency rooms. These are not the only emergency rooms at the hospital, but they share similarities in registration processes and data formats. They also include all types of patients, while the other emergency rooms are more specialized (such as psychiatric emergency room and children emergency room). The children emergency room not being included will cause the inclusion of few admissions with a low age.

Emergency cause can be one of 132 different codes. Sometimes their meaning can be interpreted (e.g. *HÖGTB*: Högt Blodtryck; High Blood Pressure), but often their

Feature	Type	Range
Next 72 hours	Binary (Target)	$\{0, 1\}$
Age	Continuous	$\mathbb{N}$
Sex	Categorical	$\{F, M\}$
Emergency time	Continuous	$\mathbb{N}$
Emergency department	Categorical	See 3.3
Emergency cause	Categorical	See 3.3
Mode of arrival	Categorical	See 3.3
Previous primary diagnose	Categorical	See 3.3
#Previous diagnoses 12 months	Continuous	$\mathbb{N}$
#Admissions 12 months	Continuous	$\mathbb{N}$
Admission hour	Categorical	$\{0, 23\}$
Admission day	Categorical	$\{1, 7\}$
Admission month	Categorical	$\{1, 12\}$

Table 3.1: Feature Description. Description of the selected features for the model.

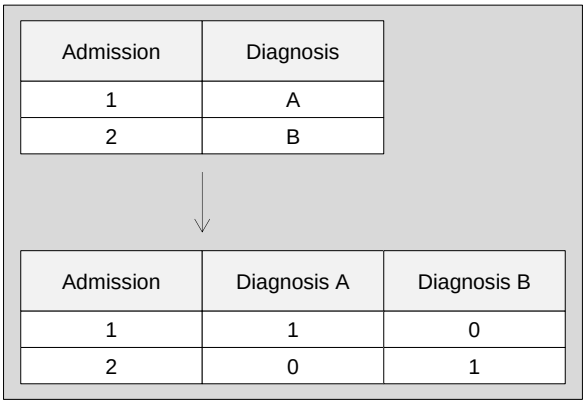

Figure 3.4: Dichotomizing Features. For each unique value in a categorical column, a new column is added. The admission entry is assigned a 1 in the column corresponding to its old value, and the old column is then dropped.

meaning is not easily decoded, for example *SVEXT*. However, the actual meaning is not regarded in this work, as it is not relevant for the prediction ability of the model. Mode of arrival includes five codes, depending on how the patient arrived to the emergency room, for example with ambulance or with helicopter.

Previous primary diagnose is, if there is any, the last previous diagnose the patient had in the past twelve months. In order to group diagnoses the different diagnoses were grouped under its ICD 10 chapter letter (22 letters, from A to Z with some excluded letters). A list of all features is provided in appendix A.

Admission	Diagnosis A	Emergency Cause A
1	1	0
2	0	1
3	1	1

Admission	Diagnosis A	Emergency Cause A	Diagnosis A ∩ Emergency Cause A
1	1	0	0
2	0	1	0
3	1	1	1

Figure 3.5: Extended Dataset. An extended data set, where each column is multiplied to the values of the others. Each multiplication creates a new column containing the new value, signifying the interaction between the multiplied columns.

3.4 Data Characteristics

In order to understand the data, a descriptive analysis of the features was made. This is important for constructive evaluation and validation of the results. Figure 3.6 through figure 3.8 visualize how the different values of the features are distributed. The y-axes of these figures represent the fraction of admissions resulting in a revisit. Each figure also contains an accumulative curve to provide an understanding of how the total number of admissions are distributed.

The emergency admissions are quite evenly spread out over the periodic features Admission Day and Admission Month. However, the Admission Hour has a more sinusoidal shape, which is also reflected in the accumulative curve. During the early hours of the day, between 00:00 and 07:00, the rate of admissions per hour is low, but the return rate is above average until around 06:00, reaching its lowest point at 10:00. From 07:00 the admission rate increases, and from 10:00 the return rate does too. After 19:00 both the admission rate and chance of return decreases.

For Number of admissions, number of previous diagnoses and emergency time the chance of returning greatly increases at higher values. However, the 90% of the population is reached fairly quickly. Age is more evenly distributed, and the accumulative curve tells us that very few are under 10 (due to the children's emergency room not being included) and over 90 years.

For the emergency cause and previous primary diagnosis, only the single features with the most difference with respect to the mean are plotted as these categorical features are numerous. Some of the Emergency causes contribute very little to the accumulation, especially those with probability below the mean. Exceptions to this are notably AKUTB and SVEXT.

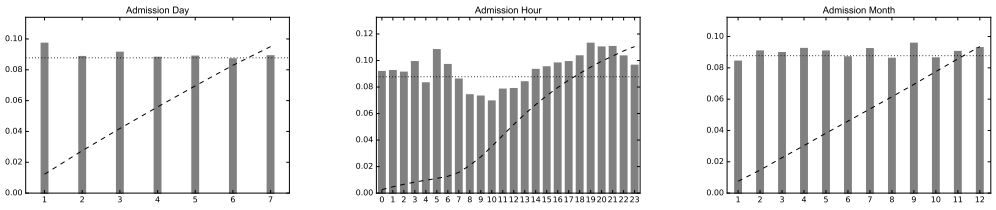

Figure 3.6: Periodic Features. Features are plotted in ascending order on the x-axis. The y-axis represents the fraction of admissions that resulted in a revisit, and the dotted line represents the percentage of admissions resulting in a revisit. The dashed line is the accumulated number of admissions (0% - 100%).

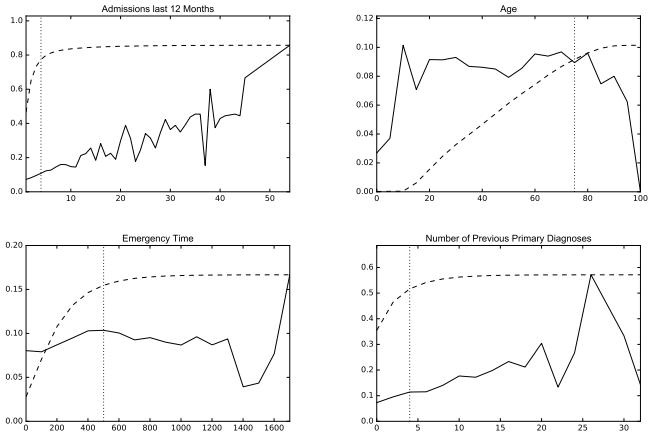

Figure 3.7: Continuous Features. Feature values are plotted in ascending order on the x-axis. The y-axis represents the fraction of admission that resulted in a revisit. The dotted line represent at which value 90% of the population is reached. The dashed line is the accumulated number of admissions (0% - 100%).

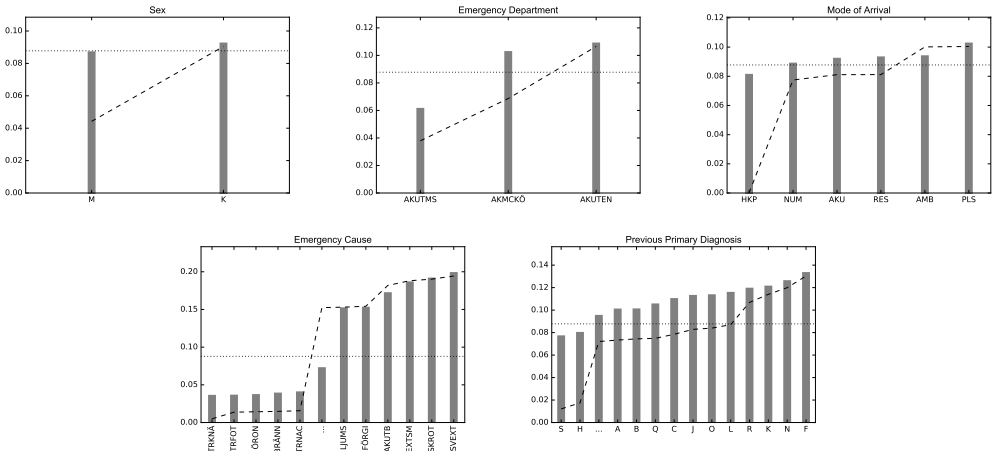

Figure 3.8: Categorical Features. Features are plotted on the x-axis, in ascending fraction of revisits order. The y-axis represents the fraction of admissions that result in a revisit, and the dotted line represents the mean fraction of admissions resulting in a revisit. The dashed line is the accumulated number of admissions (0% - 100%). In Emergency Cause and Previous Primary Diagnose, only the features deviating most from the mean are plotted.

3.5 Model Selection

Three models were selected initially with three different classifiers: Basic logistic regression (LR1), extended logistic regression (LR2) and random forest (RF). For a theoretical description of these classifiers, see section 2.3. The optimization of the three models is described in the sections below.

The rationale between selecting these three classifiers can be described from three perspectives: Prevalence, interpretability, suitability. First, the models chosen are widely used and proven models, meaning that they have a lot of easily accessible documentation and examples (Murphy, 2012), (Zhang & Ma, 2012). Interpretability, or how easily the results are interpreted, is important since the results from the models preferably should be easy to interpret. A classifier which does not reveal anything about how the different features affect the outcome cannot be used to gain knowledge about the features. Further, suitability means that the classifiers must be compatible with the data and the way the data is structured.

The logistic regression models and the random forest model have been optimized on AUC + sparsity and AUC respectively over a set of different hyperparameters using stratified 5-fold cross validation. Sparsity is included in the logistic regression optimizations since the L1 regularization allows for zeroing coefficients, and the less significant coefficients, the larger interpretability. For each model, a graph showing how the scores change depending on the hyperparameters in order to get an understanding of the relation between hyperparameters and scores. These initial analyses showed that the AUC score will not increase dramatically as the L1 penalty decreases. It should be noted that these graphs show scores of models that are trained on one part of the training data, and then tested on another part of the training data, in order for the test data to be strictly held out from any type of optimization. The Python based Scikit-learn package was used throughout the analysis, which contains classifiers for Logistic Regression and Random Forest.

Model	Classifier	Data set	Hyperparameter Set	Optimization Score
LR1	Logistic Regression	Basic	λ = (10 to the power of) -4, -3, -2, -1, 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 3, 4	AUC + Sparsity
LR2	Logistic Regression	Extended	λ = (10 to the power of) -4, -3, -2, -1, 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 3, 4	AUC + Sparsity
RF	Random Forest	Basic	Max Features = 1, 5, 20, 50, 100, 200 Tree Depth = 1, 3, 5, 7, 9, 11	AUC

Table 3.2: Model Settings Summary. Input used for optimization of the different models.

The basic logistic regression model was trained on the data set where combinations, or interactions between features, are not included. The initial score analysis suggested that it is expected to find good hyperparameters around $\lambda = 10^0$ and $\lambda = 10^2$. Thus, in order to capture this, this specific interval was more fine-grained in in the grid search. It also becomes clear that the AUC does not increase much or at λ values lower than 10^1. In this interval however, the sparsity greatly drops (as expected). Therefore, sparsity is included in the optimization score since sparsity is an important consideration for the model.

The extended logistic regression was trained on the data set where combinations, or interactions between features, are included hence calling it extended. The AUC score seems to reach its greatest value around $\lambda = 10^1$. Thus the evaluated hyperparameters for the extended model, too, was fine grained in the interval $\lambda = 10^0$ and $\lambda = 10^2$. The random forest model was trained on the data set where interactions between features are not included. For a summary of the model settings, see table 3.2.

3.6 Model Comparison

In order to compare the three models, data was sampled from the test set (bootstrapping), classified using the optimized model and tested in pairs using paired T-tests with the three hypotheses in table 3.3

Test	Hypotheses
LR1 v LR2	$\begin{cases} H_0 : AUC_{LR1} = AUC_{LR2} \\ H_1 : AUC_{LR1} \neq AUC_{LR2} \end{cases}$
LR1 v RF	$\begin{cases} H_0 : AUC_{LR1} = AUC_{RF} \\ H_1 : AUC_{LR1} \neq AUC_{RF} \end{cases}$
LR2 v RF	$\begin{cases} H_0 : AUC_{LR2} = AUC_{RF} \\ H_1 : AUC_{LR2} \neq AUC_{RF} \end{cases}$

Table 3.3: Paired T-test Hypotheses. Each of the models was tested against the other two.

4

Results

This chapter describes the results discovered during the study. Here, the findings regarding model comparison and selection are described, and how they lead to the decision of selecting the basic logistic model as the best model.

4.1 Model Performance

Accuracy is high in all three models, 91%. This is expected, since 91% of patients does not return within 72 hours and thus a model that predicted none of the patients returning would have an accuracy of 91%. Thus accuracy does not convey much information of model performance in this case.

Precision and recall is of greater interest here. The recall is low in all models, 1 %. This is interpreted as that 1 % of all patients returning are identified. Precision ranges from 48% to 53%. A precision of 50% means that of the patients predicted to return, 50% of them actually did. A precision of 50% is not trivial in this case, but considered high, as the data is very unbalanced. Random guessing would, for this data, result in a precision of 9%. Regardless, a 50% precision means that a predicted positive outcome will be wrong 50% of the time.

These low scores could be the product of three main factors. First, the model is bad for predicting patients returning. Second, the data might be insufficient for predicting this type of behaviour. Thirdly, there might be unmeasurable influences that determine whether the patients return, or that the returning patients do not share any distinguishable similarities.

The scores shown in figure 4.1 and 4.2 are calculated by extracting a small sub-test-dataset from the training data, and training the data on the remainder. Thus they are just an indication on how the scores are distributed across different hyperparameters. These score indications provided a base for choosing the hyperparameters selected during model training. The optimized hyperparameters for each model is shown in table 4.1. The optimized models' scores are shown in figure 4.3

The AUC score distributions for each test are shown in 4.4. The tests involving the

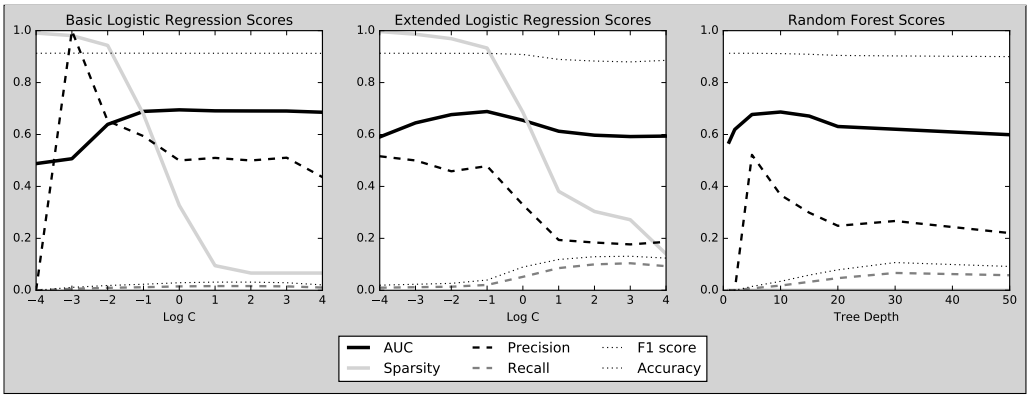

Figure 4.1: Scores for Different Hyperparameters. $\log C = -\log \lambda$

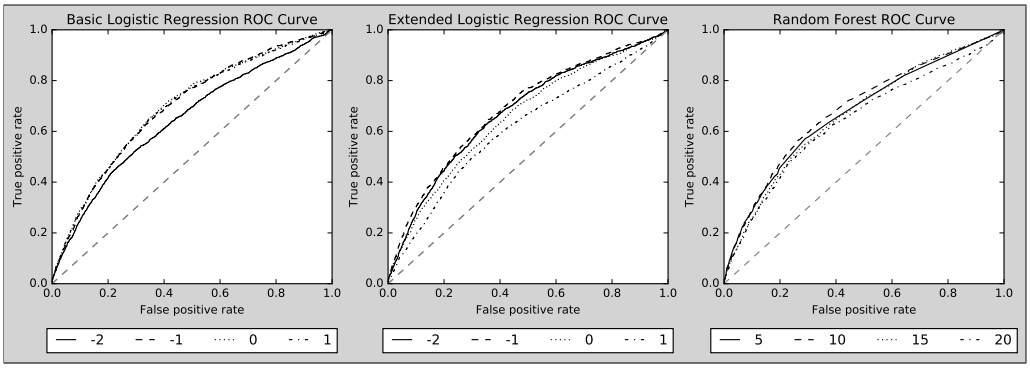

Figure 4.2: ROC Curves for Different Hyperparameters. Area Under the Curve score (AUC) is the area under the ROC curve, and should be over 0.5 to be better than random guessing.

LR1 model can be rejected due to a very low p-value. However, no conclusions can be drawn regarding the difference between the LR2 and RF model. See table 4.2 for a summary of the t-tests.

4.2 Statistical and Practical Significance

In order to understand the selection of the model in this particular study, a distinction between statistical and practical significance must be made. In this study, the T-test performed in order to prove that the different models are different, falls under the category of statistical significance. However, despite this proof of which model is the best, that model is not necessarily selected, since there may be other attributes of the model that makes it suitable.

In this study, the statistical significance is an important takeaway but not the sole determining factor. From a machine learning point of view, practical significance is much more dominant. That is, the selection of the model depends on what the situation actually requires. As an analogue to the three rationales of selecting models

Model	Best Hyperparameters
LR1	$\lambda = 10^{1.75}$
LR2	$\lambda = 10^{1.75}$
RF	$Maxfeatures = 100$ $Maxtreedepth = 9$

Table 4.1: Optimized Models Hyperparameters.

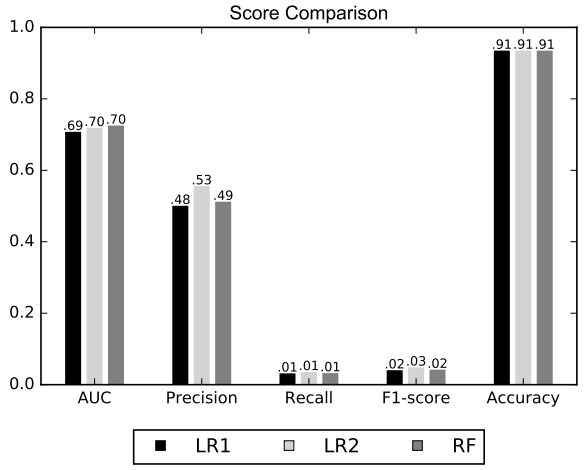

Figure 4.3: Models Scores Comparison.

mentioned in section 3.5, in this particular study the model should be preferably 1. interpretable, that is easy to draw conclusions from, 2. suitable, that is it needs to fit the prediction task and 3. efficient - which means that it should perform well without the need of excessive computer power. The LR1 model is easiest to interpret, since it is easy to understand how the features affect the outcome. In the LR2 and RF models, feature effect and meaning is harder to extract due to interaction between features and the tree split in a deep random forest. Suitability is similar in all models, since they are selected due to the prediction task. For efficiency, the LR1 model wins again over LR2 due to its lesser need of preprocessed data (no interactions). It is also preferable over the other two since its training time is shorter.

Table 4.3 shows a summarizing comparison of how well the models score from a statistical and practical point respectively.

4.3 Selected Model

Due to the importance of practical significance described in section 4.2, the LR1, or basic logistic regression model, was selected as the best model. In figure 4.5, the selected model's performance is visualized. This visualisation is powerful as it gives a graphical understanding of how good the model actually is at predicting. From

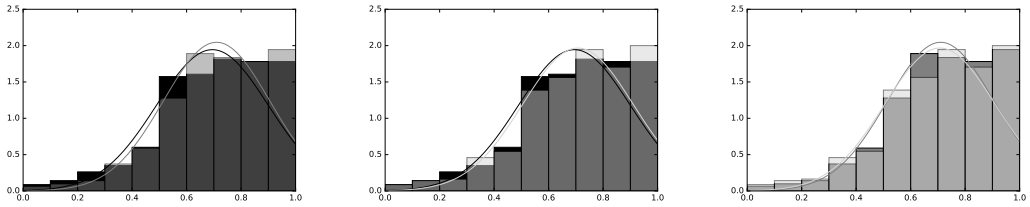

Figure 4.4: AUC Distributions: LR1 v LR2, LR1 v RF, LR2 v RF. The distributions show the normalized AUC distribution achieved during bootstrapping.

Test	P-value	Reject H_0?
LR1 v LR2	0.0030	Yes
LR1 v RF	0.0077	Yes
LR2 v RF	0.5932	No

Table 4.2: Paired T-test Result. Using a standard 5% cutoff, H_0 for LR1 v LR2 and LR1 v RF is rejected.

the lower left picture, it becomes clear that the model is very far from optimal. In an optimal model (100% precision and 100% recall), the area of box T would cover the area of box P, and the area of box F would cover the area of box N. Of the population, 9 % return. This is represented by the boxes N (Negative) and P (Positive) respectively in the upper left graph.

In the prediction, only 2% were predicted to return, that is the box T (True) in the upper right graph, whereas F (False) represents those predicted to returning. The bottom left box represents the union of the predicted and true outcome. In the bottom right box, the same graph is enlarged. The union between False and Negative results in TN (True Negatives), True and Positive results in TP (True Positives), False and Positive results in FN (False Negatives) and True and Negative results in FP (False Positives).

The length of the line A represents the recall ($Recall = \frac{TP}{TP+FN}$), and would increase were more True Positives to be predicted. The length of line B represents the inverse precision ($Precision = \frac{TP}{TP+FP}$), and would decrease were less False Positives to be predicted as precision increases.

Also, it is possible to see how Accuracy ($Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$), not included in the picture, is a very bad measure in this case, as it is will be very high as many True Negatives are predicted, which will always happen as long as the size of F is large, due to unbalanced data leading to N being large.

Features included in the selected model are shown in table 4.4. Interestingly, age affects the outcome negatively. That is, the older a patient is the less likely is it for him/her to return within 72 hours. An intuitive notion is that older patients would return more often, as elderly people usually are sicker than young people. However, it could also be that older patients are treated with more caution. Also, older people

Model	Statistical Significance	Practical Significance
Logistic Regression 1	-	+
Logistic Regression 2	+	-
Random Forest	+	-

Table 4.3: Statistical v. Practical Significance. The Logistic Regression 2 and Random Forest performs better according to a statistical test, but the Logistic Regression 1 has a better practical significance due to it being easier to interpret, which is the determining factor for choosing it as the best model.

Feature	Weight	Effect
Admissions last 12 months	0.0608	+
Number of Previous Primary Diagnoses	0.1434	+
Emergency Cause: AKUTB	0.6325	+
Emergency Cause: EXTSM	0.4995	+
Emergency Cause: SVEXT	0.3565	+
Age	-0.0029	-
Emergency Time	-0.0001	-
Emergency Cause: BRSM	-0.3029	-
Emergency Department: AKMCKÖ	-0.0026	-
Emergency Department: AKUTMS	-0.3206	-

Table 4.4: Significant Features. These are the features included in the model. Effect denotes whether the feature increases or decreases the probability of returning within 72 hours.

might have more severe illnesses which are easier to identify as compared to younger people.

As the Emergency Cause is a dichotomous feature, a patient may only have one of these features. The same goes for Emergency Department. Emergency time has the smallest impact of all the features.

Since the emergency departments only are three, coming to the Sahlgrenska emergency room seems to have a positive impact on the probability of returning within 72 hours. The number of previous admissions could be a sign of the patient being benign to return to the emergency room. The number of previous diagnoses is a sign of the patient being ill.

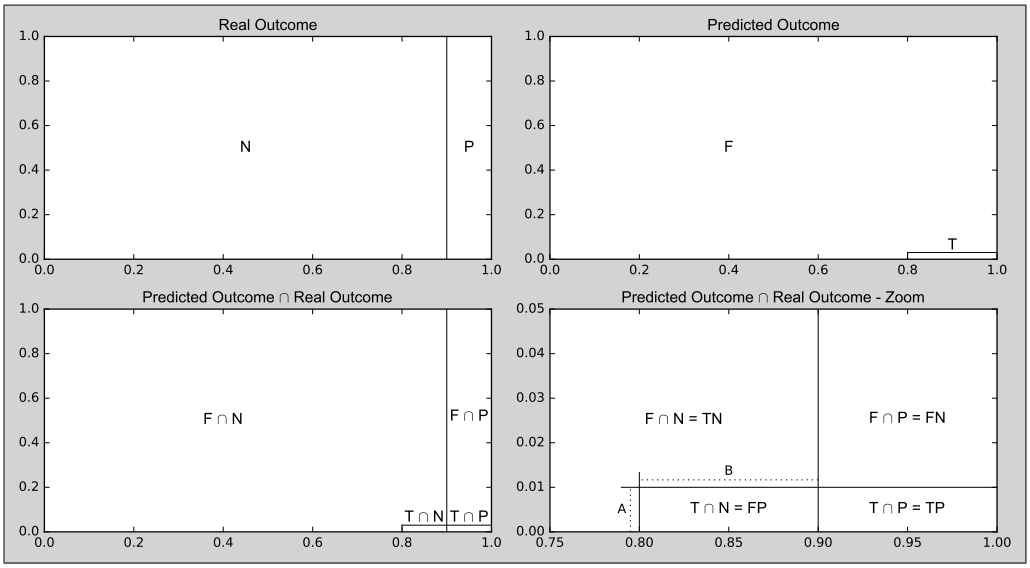

Figure 4.5: Visualisation of True/False Negatives/Positives. A box spanning from (0,0) to (1,1) represents 100% of the test population. Real Outcome represents the real distribution of negative (N) and positive (P) outcome. Predicted Outcome represents the predicted outcome, either true (T) or false (F). Predicted Outcome ∩ Real Outcome is the intersection between the two, visualising the overlay. The last box shows the overlay zoomed in, and how the different intersections represent True Negatives (TN), False Negatives (FN), False Positives and True Positives (TP). Line A and Line B represent Recall and inverse Precision respectively.

5

Implementation

In this chapter, implementation of the model and the application in health care is discussed, including ideas of what implications it might have and what technologies to use in order to develop a fully functional product. The most important function of this application would be to aid decision-makers in health care, not to be used to take decisions on its own.

Some might even argue that such an application could have an opposite effect - that the information would only confuse the personnel, as it is hard to know how to act on the information presented. Thus the information given would seem be contributing negatively to net information; or all the information the personnel has about the patient. However, information, supposing the information is correct, cannot contribute negatively. Either the decision-taker uses the information given, or rejects it. The confusion that would occur is because of inadequate description on how to act on the information. Therefore, it is important that all stakeholders, especially end users, are involved in the development of the application and how to interpret and make use of the information presented.

For prototyping, the best model was pickled i.e. preserved in a file using Sci-Kit Learn and incorporated in a Python application running a flask (a python web service library) web service. The application read from a CSV-file, acting as a database containing some example data. A front end graphical user interface (GUI) was developed using Java Swing (a Java GUI library), see figure 5.1, that communicates with the web service using HTTP requests.

5.1 Cost Score

After predicting that a patient will return to the hospital within 72 hours, an action would be taken to prevent that from happening. This will lead to an extra, unknown, cost above the normal visit cost. These actual costs will not be considered, but the size of such a cost would impact profitability of implementing a predictive model at the emergency rooms.

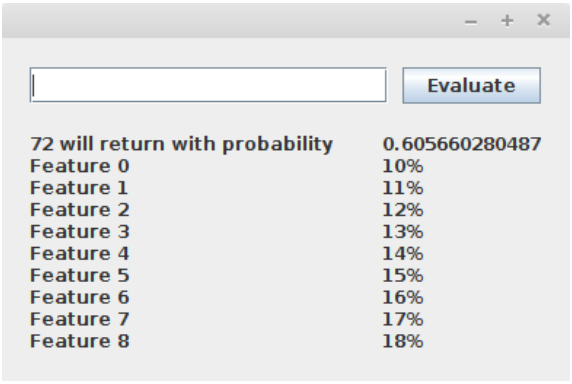

Figure 5.1: Prototype GUI screenshot. A Java Swing GUI following the concept visualised in figure 1.2.

This extra cost may be related to the "normal" cost of an emergency visit, to provide a measure which we may use to assess usability of the model from a cost perspective. In this analysis, one assumption is that all patients may be stopped from coming back by adding an extra cost, including for example hypochondriacs and those who utilize the emergency room as a substitute for primary care. How this cost would be added is not specified, but could be in the form of time or money, or extra treatment at the admission. Savings is also hard to define. For example, a patient predicted to return could be given treatment earlier before the condition worsens resulting in savings in closed care. Also, even though the induced cost for stopping patients from returning may not be profitable, it could for example increase patient value and overall community benefit, which is not included in the savings.

Let the average cost of a single emergency visit be

$$c_{visit}$$

and let the average extra cost, that would be applied above the single emergency cost to ensure that the identified patient does not return within 72 hours be

$$c_{extra}$$

The total cost without using a predictive model would therefore be the average cost times the total number of patients, or expressed with true/false positives/negatives in the model:

$$C_i = c_{visit}(T_P + F_P + T_N + F_N)$$

When using a predictive model, the extra cost would be applied to the patients that are predicted to return within 72 hours, that is the true positives and false positives. However, since the returning patients, the true positives, would now only visit once

(and not twice) the total cost of these is halved. The false positives however, would lead to an extra cost but no difference in the amount of visits. Costs for true negatives and false negatives will not change.

$$C_f = \frac{T_P(c_{visit} + c_{extra})}{2} + F_P(c_{visit} + c_{extra}) + c_{visit}(T_N + F_N)$$

Savings may be expressed as the total costs before implementation minus total costs after implementation. This has to be larger or equal to zero to be profitable;

$$Savings = C_i - C_f = c_{visit}(T_P + F_P) - [\frac{T_P(c_{visit} + c_{extra})}{2} + F_P(c_{visit} + c_{extra})] \geq 0$$

from which we may derive the relation between the average extra cost and the average visit cost;

$$c_{extra} \leq c_{visit} \times \frac{T_P}{T_P + 2F_P}$$

Where

$$\frac{T_P}{T_P + 2F_P}$$

can be interpreted as the Cost Factor, or the factor of the original visit cost that is profitable to add as an extra cost to prevent the patient from coming back. For coherence, it is called a score since a higher value is better, i.e. a higher extra cost is allowed.

The measure is related to the precision score, however it weights the false positives stronger. As mentioned earlier, this assumes that all patients, that would return in 72 hours and for whom an extra cost is added, will indeed not return. This is a non-conservative assumption, that could be managed by adding an extra error factor such that

$$CostScore = \frac{T_P}{T_P + 2F_P} \times \epsilon$$

However, as this factor is not evaluated in this work, the following analysis only considers the cost score without the error factor.

Figure 5.2 shows the cost score in relation to decision threshold. As the threshold increases, the cost score does too. This happens because a higher threshold means

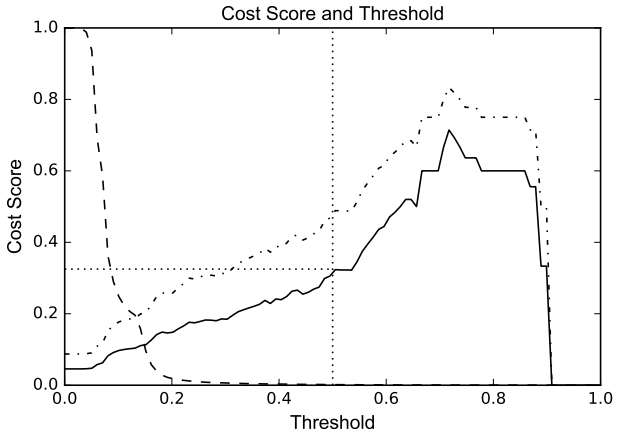

Figure 5.2: Cost Score - Threshold Relation. The black line represents the cost score, and the dot-dashed line is the percentage of identified returnees that would actually return. The dashed line is the fraction of the populating being predicted returnees. The dotted lines mark values of the default threshold and corresponding cost score.

fewer treated false and true positives, and the false positives are weighed stronger in the denominator of the cost score than the true positives. The steep drop at a very high threshold is due to no or very few true positives detected.

This graph gives important insights on how the model, incorporated in the application, could be used in daily operation. For logistic regression, the threshold for predicting a true outcome is 50%, and at that point the fraction of predicted returnees (dashed line) is very small - equal to the recall. At this point, about 50% of the patient given the extra treatment would actually have returned - visualised by the dot-dash line at this point, corresponding to the precision.

Giving the extra treatment to patients above a threshold of 0.5 gives a higher probability that the patient would have returned within 72 hours, but the number of patients at this point is very small.

Going below the threshold, especially below 20%, the fraction of predicted patients returning greatly increases, but the cost score also decreases. At a threshold of 0%, all patients would be predicted to return, and thus the values of the dashed and dot-dashed line corresponds to the total number of patients (100 %) and the fraction of those that return (9%).

This means that in daily operation, it is not suitable to use the application in order to take action purely based on its indications. Rather, it could provide just that: an indication. It is important to emphasize that this is in no way meant to replace the knowledge of medical personnel.

6

Discussion

In this chapter, the findings and their meaning are discussed, together with threats to validity of the results. Lastly, future work is discussed

The utility of the findings in this study, as is, is low. It is however, a proof of concept that advanced data analysis in the form of machine learning indeed has a large potential for use with observational data. Hopefully, it could lead to an increased understanding among policy makers that machine learning could be feasible for use within the health care sector. Also, the importance of electronic health systems becomes apparent and might influence the improvement of such. Data access and availability must be good in order to perform well-designed analyses.

6.1 Predicting Patient Behaviour

According to the findings in this study, there is a possibility to predict whether a patient will return to the emergency room within 72 hours - however with a very low predictability. The low predictability could be due to many factors. The data collected in this study is only what is readily available in the systems. There is a possibility that with more data regarding the patients' medical history, for example laboratory data and journal data, the predictability could be increased.

There is also a probability that the medical history data does not convey enough about the patients' behaviour, and that the behaviour is dependent on other factors, for example, personal characteristics, socio-economic background and so forth. Both cases are interesting. In the first case, improved electronic health record systems with easier access to more data records would improve the predictability. Should the second case be true, perhaps it would be worthwhile register this type of information regarding the in the electronic health system. Additionally, there could be other patient behaviours that are predictable using current data.

6.2 Patient Features

In the predictive model developed in this study, the following features are affecting whether the patient will return within 72 hours or not: Numbers of admissions the past 12 month, Number of unique diagnoses the past 12 months, Age, Emergency time, Emergency department and the emergency causes SVEXT, BRSM and EXTSM. It is important to remember that these features are just a subset of the selected features when defining the study, and that the predictive model using these features does not have a high performance. However, these findings are interesting, as these features do seem to affect patient behaviour.

6.3 The Predictive Model

The model was selected due to its simplicity and comparatively good performance as the best model to implement in an application. The discussion about how the use of it affects the cost score is important in order to understand the implications of such an application. The cost score analysis also provides a way of analysing the use of models that might not have satisfactory predictive capabilities, and could be interesting to apply to other models, for example in other contexts, as well. A similar cost analysis has not been found in any other papers.

6.4 Threats to Validity

In this section, the threats to validity of the study are assessed from the four perspectives: Internal validity, external validity, construct validity and reliability. These four perspectives are described by Runeson and Höst (2009) in order to get a complete picture of study validity threats.

There is a construct validity threat since the features found to affect the patient is in relation to the developed model - which does not have a very good prediction. For a perfect model, more, less, or other, features would probably have been relevant. The best model is thus actually just the best model with the prerequisites of this study.

As the data collected in this study was extracted from a data set containing only medical data, there is a very large chance that other factors influence the outcome. Also, the data collected consisted of a few features that were believed to affect the outcome, which undermines the internal validity. This is reflected in the low model performance. One specific example of an internal validity threat regarding the data, is a way of handling patients that cannot be diagnosed due to vague signs of specific diseases. These patients are often sent home, and told to come back soon, as their

condition worsens. This does not happen to a significantly large portion of the observed population, but is neither included in the data. This information was not known before the study was done, and perhaps other type work methods could be found to influential on the outcome.

On the topic of external validity, this study was conducted at one hospital with an access to data that is registered in a way unique to this specific hospital. In that case, doing the same study in another hospital might mean that another type and amount of data is accessible and thus the same study could not be conducted. It could also be a problem adopting this study within other areas of health care for the same reason. However, the method of selecting a good model is still a general approach, as is the cost score analysis.

The cohort in this study is well defined, and selected features are clearly stated. The cross validation performed is expected to minimize the errors during hyperparameter estimation. The optimization method for the classifiers may differ, which could yield somewhat different results during model optimization. However, three models give similar results and the model performance scores can be assumed to be accurate and consistent. It is also important that inclusion criteria during the data extraction are exactly the same as in this study in order for the study to be reproducible.

6.5 Future Work

This study opens up some different paths in relation to future work. Firstly, other behaviours could be studied, using available electronic health care records. Perhaps, the data can predict other behaviours better than the behaviour chosen in this study. Secondly, the behaviour selected in this study could be further investigated, but using more or other features. Thirdly, other machine learning models might prove to perform the prediction task better. Fourthly, the use of other types of data, perhaps from external sources may be a suitable approach. Of course all these could be combined in different ways.

When it comes to application development, further investigation of the different obstacles occurring when using health care data, such as data extraction from existing systems would be interesting. Here, many opportunities using machine learning emerge as well, such as image analysis of x-ray images, text analysis of journal data and so forth.

7

Conclusion

In health care, a vast amount of data is available in electronic health care records. However, the application of machine learning to this data for the purpose of predicting patient behaviour needs improvement. This is especially true when considering implementation of predictive models in decision making regarding patients' care.

In this study, a model that predicts patient behaviour is developed. It can predict whether a patient admitted to any of the main emergency rooms at Sahlgrenska University Hospital will revisit the emergency room in 72 hours with a recall of 1% and a precision of 50%. The model achieves an area under the curve score of 68%. In summary, its ability to predict patient behaviour correctly is very low. Moreover, the implication of implementing such a model in an application is assessed. It is clear that a model with such low predictability does not in any way replace the knowledge of medical personnel, and should only be used as a support for making decisions.

The selected model uses a logistic regression classifier and was trained on a data set without interactions between features. It was selected due to its simplicity in comparison with two other models; a model using random forest, and a model using logistic regression model trained with an extended data set including interaction between features. Simplicity was an important factor in order to be able to interpret the model. The data set consisted of features extracted from available electronic health records. In the selected model, the following features are affecting whether the patient will return within 72 hours or not: Numbers of admissions the past 12 month, Number of unique diagnoses the past 12 months, Age, Emergency time, Emergency department and the emergency causes SVEXT, BRSM and EXTSM.

The research questions are answered as follows:

- **RQ 1:** With the selected features and classifiers, it is not possible to predict whether a patient will return within 72 hours to a large extent. However, it is possible to make a prediction with a recall of 1%, and a precision of 50%, which is considerably better than random chance in this case.

- **RQ 2a:** The most suitable model incorporates a logistic regression classifier using data without interactions, in comparison to a logistic regression classifier

using data with interactions and random forest.

- **RQ 2b:** The selected predictive model is suitable to implement since it is easily interpreted, it fits the prediction task, and performs almost as well as the other two compared models.

Due to the low predictability of the model, it is not ready to be implemented in daily operations. However, it is a proof on concept that with some data from electronic health records at Sahlgrenska University Hospital, it is possible to predict patient behaviour. There might also be other behaviours that the data could predict more accurately. This study further emphasizes the importance of good electronic health record systems, and the utility of machine learning models in a health care context and the potential of such. Further work includes more advanced models, using different type of data, analyzing different patient behaviours.

In summary, this study is a proof of concept of predictive modeling using electronic health care records that despite low predictability highlights the utility of machine learning in a health care context.

Bibliography

Baan, C. A., Ruige, J. B., Stolk, R. P., Witteman, J., Dekker, J. M., Heine, R. J. & Feskens, E. (1999). Performance of a predictive model to identify undiagnosed diabetes in a health care setting. *Diabetes Care*, *22*(2), 213–219.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Duignan, J. (2016). Retrospective cohort study. Oxford University Press. Retrieved from http://www.oxfordreference.com/10.1093/acref/9780191792236.001.0001/acref-9780191792236-e-544

Faries, D. E., Obenchain, R., Haro, J. M. & Leon, A. C. (2010). *Analysis of observational health care data using sas*. SAS Institute.

Hillestad, R., Bigelow, J., Bower, A., Girosi, F., Meili, R., Scoville, R. & Taylor, R. (2005). Can electronic medical record systems transform health care? potential health benefits, savings, and costs. *Health affairs*, *24*(5), 1103–1117.

Jensen, P. B., Jensen, L. J. & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395–405.

Khalilia, M., Chakraborty, S. & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, *11*(1), 1.

Kontio, E., Airola, A., Pahikkala, T., Lundgren-Laine, H., Junttila, K., Korvenranta, H., ... Salanterä, S. (2014). Predicting patient acuity from electronic patient records. *Journal of Biomedical Informatics*, *51*, 35–40.

Krumholz, H. M. (2013). Post-hospital syndrome—an acquired, transient condition of generalized risk. *New England Journal of Medicine*, *368*(2), 100–102.

Louis, D. Z., Robeson, M., McAna, J., Maio, V., Keith, S. W., Liu, M., ... Grilli, R. (2014). Predicting risk of hospitalisation or death: a retrospective population-based analysis. *BMJ open*, *4*(9), e005223.

Murdoch, T. & Detsky, A. (2013). The inevitable application of big data to health care. *JAMA*, *309*(13), 1351–1352. doi:10.1001/jama.2013.393. eprint: /data/Journals/JAMA/926712/jvp130007_1351_1352.pdf

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. Cambridge, MA: MIT Press.

Obermeyer, Z., Powers, B. W., Makar, M., Keating, N. L. & Cutler, D. M. (2015). Physician characteristics strongly predict patient enrollment in hospice. *Health Affairs*, *34*(6), 993–1000.

cross-validation. (2016).

Provost, L. P. & Murray, S. K. (2011). *The health care data guide: learning from data for improvement* (1st ed.). San Francisco, CA: Jossey-Bass.

Razavian, N., Blecker, S., Schmidt, A. M., Smith-McLallen, A., Nigam, S. & Sontag, D. (2015). Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*, *3*(4), 277–287.

Roden, D., Xu, H., Denny, J. & Wilke, R. (2012). Electronic medical records as a tool in clinical pharmacology: opportunities and challenges. *Clinical Pharmacology & Therapeutics*, *91*(6), 1083–1086.

Runeson, P. & Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, *14*(2), 131–164.

Wang, L., Porter, B., Maynard, C., Evans, G., Bryson, C., Sun, H., ... Frisbee, K. et al. (2013). Predicting risk of hospitalization or death among patients receiving primary care in the veterans health administration. *Medical care*, *51*(4), 368–373.

Wu, J., Roy, J. & Stewart, W. F. (2010). Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, *48*(6), S106–S113.

Zhang, C. & Ma, Y. (2012). *Ensemble machine learning: methods and applications* (1. Aufl.;2012;). New York, NY: Springer-Verlag.

A

Appendix A - List of Categorical Features and Descriptive Numbers

See the following pages. Observe that all features are in dichotimized format.

emergency cause TRHUV	4,1%	2,6%	5,5%	emergency cause SYNKO	0,3%	0,2%	5,1%	emergency cause GIPS	0,6%	0,6%	7,9%	emergency cause VAGBL	0,0%	0,0%	60,0%
emergency cause SVEXT	2,1%	4,7%	19,4%	emergency cause TRTHO	0,8%	0,7%	7,3%	emergency cause HV	2,8%	2,5%	7,7%	emergency cause BRADY	0,0%	0,0%	0,0%
emergency cause BRSM	7,9%	4,0%	4,4%	emergency cause FÖRGI	0,6%	1,0%	14,8%	emergency cause SMRYG	0,2%	0,2%	9,0%	emergency cause SMNAC	0,2%	0,1%	5,1%
emergency cause TRKNÄ	2,6%	0,9%	3,1%	emergency cause TRFOT	4,5%	1,6%	3,2%	emergency cause TRUBE	0,5%	0,4%	6,4%	emergency cause DIARR	0,3%	0,3%	10,0%
emergency cause DYSPN	4,0%	3,4%	7,5%	emergency cause OSPEC	1,3%	1,2%	8,3%	emergency cause SENSB	0,1%	0,1%	9,9%	emergency cause FRKRO	0,4%	0,2%	4,2%
emergency cause UROSM	1,0%	1,0%	8,6%	emergency cause SÅRSK	1,8%	1,1%	5,4%	emergency cause HEMAT	0,5%	0,5%	9,5%	emergency cause SÅRBE	0,1%	0,2%	16,9%
emergency cause TRHÖF	0,9%	0,7%	7,2%	emergency cause HÖGTB	0,3%	0,2%	4,4%	emergency cause RECT	0,6%	0,8%	10,7%	emergency cause HYPOG	0,1%	0,1%	5,3%
emergency cause ANEMI	0,1%	0,1%	6,0%	emergency cause RYGGV	0,3%	0,4%	11,8%	emergency cause TRBÄC	0,0%	0,0%	15,0%	emergency cause TRLÅR	0,1%	0,1%	6,7%
emergency department AKMCKÖ	28,9%	33,0%	10,0%	emergency cause KOMA	0,0%	0,0%	11,1%	emergency cause ALLER	1,8%	1,3%	6,3%	emergency cause SKROT	1,1%	2,3%	18,7%
emergency department AKUTMS	35,6%	23,9%	5,9%	emergency cause UROST	0,8%	1,0%	11,8%	emergency cause GI-BL	0,2%	0,2%	11,4%	emergency cause LABB	0,1%	0,1%	8,9%
emergency department AKUTEN	35,5%	43,1%	10,6%	emergency cause INF	4,4%	5,9%	11,9%	emergency cause TRAUM	1,2%	0,7%	5,3%	emergency cause LJUMS	0,3%	0,5%	14,8%
mode of arrival HKP	0,0%	0,0%	7,9%	emergency cause TRARM	3,9%	1,9%	4,4%	emergency cause TRRYG	0,6%	0,4%	5,3%	emergency cause FEBER	0,3%	0,3%	10,5%
mode of arrival PLS	0,3%	0,3%	10,0%	emergency cause AKUTB	14,2%	27,2%	16,8%	emergency cause KRAMP	0,8%	0,8%	8,6%	emergency cause LEDVÄ	0,4%	0,2%	4,6%
mode of arrival RES	0,1%	0,1%	9,1%	emergency cause BORTF	2,2%	2,0%	7,8%	emergency cause KONFU	0,1%	0,1%	16,3%	emergency cause SMÄRT	0,1%	0,2%	10,4%
mode of arrival AKU	3,6%	3,7%	9,0%	emergency cause EXTSM	3,2%	6,7%	18,2%	emergency cause MELEN	0,4%	0,3%	8,5%	emergency cause TRIAG	0,6%	0,8%	12,0%
mode of arrival NUM	77,1%	76,1%	8,7%	emergency cause SJUKD	1,9%	2,3%	10,3%	emergency cause FALLB	0,0%	0,1%	18,2%	emergency cause SVIMN	0,7%	0,4%	4,7%
mode of arrival AMB	18,9%	19,7%	9,2%	emergency cause NÄSBL	0,4%	0,5%	10,8%	emergency cause HYPOT	0,0%	0,0%	0,0%	emergency cause LÅGBT	0,0%	0,0%	8,0%
sex M	48,9%	47,4%	8,5%	emergency cause TRHAN	6,1%	2,9%	4,2%	emergency cause YRSEL	2,6%	1,4%	4,6%	emergency cause NAUSE	0,3%	0,4%	11,7%
sex K	51,1%	52,6%	9,0%	emergency cause RYGGS	2,5%	1,9%	6,9%	emergency cause SMEXT	2,4%	1,7%	6,4%	emergency cause OPKOM	0,2%	0,3%	10,2%
next 72 hrs b(target)	8,8%	100,0%	100,0%	emergency cause URO	0,3%	0,4%	10,5%	emergency cause OHJRY	2,2%	1,4%	5,7%	emergency cause TRÖTT	0,1%	0,1%	8,8%
	% of admissions has this feature	% of admissions resulting in a revisit has this feature	% of admission with this feature result in a revisit		% of admissions has this feature	% of admissions resulting in a revisit has this feature	% of admission with this feature result in a revisit		% of admissions has this feature	% of admissions resulting in a revisit has this feature	% of admission with this feature result in a revisit		% of admissions has this feature	% of admissions resulting in a revisit has this feature	% of admission with this feature result in a revisit

emergency cause EXANT	0,0%	0,0%	0,0%	emergency cause ADDIS	0,0%	0,0%	0,0%	emergency cause KONSU	0,0%	0,0%	0,0%	previous pdiag Q	0,2%	0,2%	10,2%
emergency cause LÄKRE	0,0%	0,0%	2,6%	emergency cause FRÄTS	0,0%	0,0%	0,0%	emergency cause TRÖGO	0,0%	0,0%	0,0%	previous pdiag P	0,0%	0,0%	50,0%
emergency cause HALS	0,6%	0,4%	5,9%	emergency cause ELOLY	0,1%	0,0%	2,1%	emergency cause SÖMNS	0,0%	0,0%	0,0%	previous pdiag O	0,5%	0,6%	11,1%
emergency cause NÄSA	0,1%	0,1%	7,4%	emergency cause HÄLSU	0,0%	0,0%	20,0%	emergency cause SVAMP	0,0%	0,0%	0,0%	previous pdiag N	2,6%	3,7%	12,3%
emergency cause ALKOH	0,0%	0,0%	3,2%	emergency cause TRBUK	0,0%	0,0%	5,4%	emergency cause KEM	0,0%	0,0%	0,0%	previous pdiag M	5,2%	5,0%	8,4%
emergency cause BRÄNN	0,2%	0,1%	3,4%	emergency cause GYN	0,0%	0,0%	100,0%	emergency cause SPECU	0,0%	0,0%	50,0%	previous pdiag L	1,3%	1,7%	11,3%
emergency cause BRSMA	0,3%	0,2%	5,6%	emergency cause HUD	0,0%	0,0%	0,0%	emergency cause DRUNK	0,0%	0,0%	0,0%	previous pdiag K	3,1%	4,2%	11,8%
emergency cause INTOX	0,1%	0,2%	9,3%	emergency cause ASCIT	0,0%	0,0%	20,0%	emergency cause HALSB	0,0%	0,0%	0,0%	previous pdiag J	1,9%	2,4%	11,0%
emergency cause ÖNH	0,0%	0,0%	0,0%	emergency cause SMBUK	0,0%	0,0%	0,0%	emergency cause THORA	0,0%	0,0%	0,0%	previous pdiag I	4,0%	4,3%	9,4%
emergency cause EXTÖD	0,0%	0,0%	9,1%	emergency cause ÅNGES	0,0%	0,0%	25,0%	emergency cause POLIS	0,0%	0,0%	0,0%	previous pdiag H	2,4%	2,1%	7,7%
emergency cause TRAXE	0,0%	0,0%	18,8%	emergency cause PSYK	0,0%	0,1%	20,5%	emergency cause TONSB	0,0%	0,0%	0,0%	previous pdiag G	1,9%	1,9%	8,9%
emergency cause POSTO	0,3%	0,3%	8,9%	emergency cause ANAFY	0,0%	0,0%	100,0%	emergency cause TULL	0,0%	0,0%	0,0%	previous pdiag F	4,6%	6,8%	13,0%
emergency cause HEMOP	0,1%	0,0%	2,8%	emergency cause RÖKGA	0,0%	0,0%	0,0%	emergency cause X	0,0%	0,0%	0,0%	previous pdiag E	1,3%	1,5%	9,7%
emergency cause ÖRON	0,3%	0,1%	3,3%	emergency cause TREMO	0,0%	0,0%	8,3%	emergency cause DYKOL	0,0%	0,0%	14,3%	previous pdiag D	1,1%	1,2%	9,1%
emergency cause UNDER	0,0%	0,0%	0,0%	emergency cause TAKYK	0,0%	0,0%	10,5%	emergency cause HYPVE	0,0%	0,0%	0,0%	previous pdiag C	1,6%	2,0%	10,7%
emergency cause ÖDEM	0,0%	0,0%	16,7%	emergency cause AMBIT	0,0%	0,0%	0,0%	emergency cause NEURO	0,0%	0,0%	0,0%	previous pdiag B	0,5%	0,5%	9,8%
emergency cause BETT	0,3%	0,4%	9,6%	emergency cause SÅRIN	0,0%	0,1%	25,0%	emergency cause VÅLDT	0,0%	0,0%	0,0%	previous pdiag A	0,6%	0,6%	9,8%
emergency cause TRNAC	0,5%	0,2%	3,6%	emergency cause MISSH	0,6%	0,6%	8,2%	emergency cause SUICI	0,0%	0,0%	100,0%	emergency cause MISSB	0,0%	0,0%	0,0%
emergency cause HYPER	0,2%	0,2%	6,9%	emergency cause GULSO	0,0%	0,1%	62,5%	emergency cause INHAL	0,0%	0,0%	0,0%	emergency cause NJUR	0,0%	0,0%	0,0%
emergency cause ÖGON	0,1%	0,0%	3,4%	emergency cause DIAB	0,1%	0,1%	9,8%	emergency cause SJÄLV	0,0%	0,1%	27,8%	emergency cause PENET	0,0%	0,0%	0,0%
% of admissions has this feature	0,3%			% of admissions has this feature	0,1%			% of admissions has this feature	0,0%			% of admissions has this feature	0,0%		
% of admissions resulting in a revisit has this feature	0,4%			% of admissions resulting in a revisit has this feature	0,1%			% of admissions resulting in a revisit has this feature	0,0%			% of admissions resulting in a revisit has this feature	0,0%		
% of admission with this feature result in a revisit	9,6%			% of admission with this feature result in a revisit	25,0%			% of admission with this feature result in a revisit	0,0%			% of admission with this feature result in a revisit	0,0%		

	previous pdiag R	previous pdiag S	previous pdiag T	previous pdiag U	previous pdiag Z	admission hour 0	admission hour 1	admission hour 2	admission hour 3	admission hour 4	admission hour 5	admission hour 6	admission hour 7	admission hour 8	admission hour 9	admission hour 10	admission hour 11	admission hour 12	admission hour 13	admission hour 14
% of admissions has this feature	8,7%	5,3%	1,0%	0,0%	9,5%	2,3%	2,0%	1,7%	1,5%	1,3%	1,2%	1,5%	2,7%	4,8%	5,8%	7,1%	7,5%	7,4%	7,1%	6,5%
% of admissions resulting in a revisit has this feature	11,6%	4,5%	1,0%	0,0%	10,4%	2,4%	2,0%	1,7%	1,6%	1,2%	1,5%	1,6%	2,5%	3,9%	4,7%	5,4%	6,5%	6,4%	6,6%	6,7%
% of admission with this feature result in a revisit	11,6%	7,4%	9,0%	0,0%	9,6%	8,9%	9,0%	8,9%	9,7%	8,1%	10,6%	9,4%	8,3%	7,2%	7,1%	6,7%	7,6%	7,6%	8,1%	9,1%

	admission hour 15	admission hour 16	admission hour 17	admission hour 18	admission hour 19	admission hour 20	admission hour 21	admission hour 22	admission hour 23	admission day 1	admission day 2	admission day 3	admission day 4	admission day 5	admission day 6	admission day 7	admission month 1	admission month 2	admission month 3	admission month 4
% of admissions has this feature	6,1%	5,5%	5,1%	4,6%	4,4%	4,1%	3,4%	3,6%	2,9%	12,9%	16,1%	15,1%	14,7%	14,4%	14,0%	12,8%	8,1%	7,6%	8,4%	8,2%
% of admissions resulting in a revisit has this feature	6,5%	6,0%	5,6%	5,2%	5,5%	5,0%	4,2%	4,1%	3,1%	13,9%	15,8%	15,4%	14,4%	14,2%	13,6%	12,7%	7,6%	7,7%	8,4%	8,5%
% of admission with this feature result in a revisit	9,3%	9,6%	9,7%	10,1%	11,0%	10,8%	10,8%	10,1%	9,4%	9,5%	8,6%	8,9%	8,6%	8,7%	8,5%	8,7%	8,2%	8,9%	8,8%	9,0%

	admission month 5	admission month 6	admission month 7	admission month 8	admission month 9	admission month 10	admission month 11	admission month 12
% of admissions has this feature	8,5%	8,2%	8,3%	8,3%	8,5%	8,8%	8,6%	8,4%
% of admissions resulting in a revisit has this feature	8,6%	7,9%	8,5%	8,0%	9,0%	8,4%	8,7%	8,7%
% of admission with this feature result in a revisit	8,9%	8,5%	9,0%	8,4%	9,4%	8,4%	8,8%	9,1%