SIB - Hospitalisation Length Modeling

Lucas de La Brosse, Anurag Miglani, Thomas Zamojski

30 March 2016

1 Introduction

We have seen since the end of last century rapid development in the information technologies. Many institutions and businesses are taking advantage of the possibilities offered by these technologies. In particular, health care institutions have computerised their information system in the interest of providing ever better experience to their patients. This has only significantly improved the efficiency of administration, but has also brought digitalisation in the health care system, bringing new opportunities.

Santé Informatique de Bretagne (SIB), created in 1973, is a publicly funded group dedicated to the information systems of health care institutions, from retirement homes to regional hospital centers. Its wide range of activities include conception and deployment of solutions, storage of personal confidential data, hosting services, and consulting. Around 350 institutions are serviced by SIB.

Moving forward with its time, SIB is looking to implement Big Data solutions for its partners. Amongst many things, it is in the possession of many databases of patients' data that is being mostly used for description requests. The current project stems from the will to use these accumulated data towards predicting bed usage and stay duration of patients in order to improve the service to patients through more efficient bed management.

In this report, we investigate patients data kindly provided to us by SIB and its partner with the goal of predicting the duration of hospitalisation in order to do more efficient bed management. The project is at the stage of infancy. It is desirable at the moment to provide a proof of concept before any implementation of solutions. At this stage in time, the project is at the phase of data preparation. With that in mind, we make a series of recommendations as to which direction the project might go in the near future.

2 Description of the Data

In this section, we propose to describe the data provided for the project in relation with the strategies for the models. The data provided comes from the information on patients of a single hospital.

2.1 Structure of the Data

It is divided into two main categories: data from the programme de médicalisation des systèmes d'informations (PMSI) and data from the dossier patient informatisé (DPI). PMSI data is generated 5-10 days after the stay and is therefore useful to generate outcome variables. DPI data on the other hand is permanent and actualised along the stay. It is used to get explanatory variables.

PMSI is a provision part of a reform of the french health care system aiming at reducing the ressources inequalities between different health care institutions in France. PMSI consists of quantified and standardised information on institutions' activity in Médecine Chirurgie-Obstétrie (MCO). It is gathered by the institution itself based on the information contained in the résumé de sortie standardisé (RSS), which is entered in the system at the end of the hospitalisation. It contains a summary of the patient's stay, including medicals acts, entry dates, exit dates and diagnostics. In particular, the duration is deduced from this information.

The DPI is permanently linked to the patient, available at entry date and updated along the stay. It consists of all the information recorded about the patient in the database. In this category of data, we look for establishing a profil for a patient that would correlate with a hospitalisation duration. The logic is that prior information on the patient's medical visits should influence the reason why he is coming to the hospital and thus the duration of a stay.

Patients have a unique ID in the information system called identifiant permanent du patient (IPP). The IPP is common to both PMSI and DPI dataset. Moreover, each time a patient visits the hospital, a identifiant externe du patient (IEP) is assigned to his stay. This ID thus corresponds to a unique *stay*. At the patient's exit, the RSS is produced and also assigned an ID, called RSSid. This also corresponds to a unique *stay*.

Thus, we have a key identifying a patient, the IPP, and two keys identifying a stay, the IEP and RSSid. In the PMSI, the IPP is given, therefore linking the outcome variables with the explanatory variables.

The information found in the categories can be summarised in the following list:

- PMSI: entry/exit dates and modes, principal/associated diagnostics, medical acts performed during the stay. Also given are the date of birth, gender and residence of the patient.
- DPI: date of birth, gender, country, family situation, number of children, profession (rarely given), allergies (extremely rare), body mass index (bmi), weight/height, previous diagnostics code/text, reason for emergency in short free text format, reason for hospitalisation in 9 labels.

Number of PMSI patients	92'568
Number of Stays in PMSI	273'999
Number of DPI patients	223'208

Figure 1: Sizes of patients in PMSI and DPI and number of stays.

2.2 The Target: Duration of Hospitalisation

2.2.1 Global Output

The goal of the project is to make predictions about the duration of hospitalisation. The first question is: how is this variable represented in the dataset? In the PMSI data, there are two ways to infer the duration: from the entry and exit date and from the duration column. Ideally, there should not be any discrepencies between the two. However, we have found cases where there is a difference. Most notably, if a patient comes in the hospital and leaves the same day, should the stay length be considered as 0 or 1? Both occurs[1]:

[1]In all our examples, the ID's are faked to preserve confidentiality of patients

IEP	RSSid	Entry Date	Exit Date	Duration
AG03	R74	31/01/12	31/01/12	1
AT45	R1023	31/08/11	31/08/11	1
BF34	R2459	06/01/16	06/01/16	0
HT78	R2788	11/08/11	11/08/11	1
UI89	R2898	26/02/15	26/02/15	0

This shows that more care has to be taken in the future by the model as to what duration to assign to stays. We made two possible hypotheses to resolve this. First, if we are interested in bed usage overnight only, it makes more sense to take the difference in days between entry and exit date, making the duration all 0 in the above table. On the other hand, we could also understand from this that a patient with same entry and exit dates having a duration of 1 means that he has used a bed, as opposed to 0 indicating that no bed was needed. In the following, we are opting for the second interpretation.

Another issue we encountered is examplified by the following: patients having the same IEP, but different RSS id:

IEP	RSSid	Entry Date	Exit Date	Duration	Feature Merged
DR79	R3490	17/03/06	17/03/06	1	Alcoolised
DR79	R3491	21/03/06	21/03/06	1	Alcoolised
DR79	R3492	24/03/06	24/03/06	1	Alcoolised
DR79	R3493	28/03/06	28/03/06	1	Alcoolised
DR79	R3494	31/03/06	31/03/06	1	Alcoolised

The problem here is that stay's features are only available through the IEP. When merging other features with the output table according to the IEP key, the new features will be copied over all the lines as demonstrated in the table. This might not be appropriate in the context. From discussions with SIB, it also appears that this could be a patient following therapy, and might have his duration set to 0.

To summarise, we see that bed management raises some questions as to what part of the patients we need to consider and what duration should be taken in different scenarios. That means that the duration should be refined in the future to reflect the actual bed occupation rather than facturation or other.

2.2.2 Output visualisation and Statistics

The distribution of the duration of hospitalisation across all patients is plotted in Figure 3 and Figure 4. Two properties of this distribution are important to mention. First, there is high concentration around the median of 1 day. The second is that the distribution is highly dispersed towards the right, going as far as a maximum value of 283 days. Standard statistics are provided in Figure 2

Min	1st Qt	Median	3rd Qt	Max	Mean	Std	MAD
0	1	1	4	283	3.226	6.00	2.72

Figure 2: Statistics for Duration of Hospitalisation

Because of the large dispersion of values, it makes more sense to use mean absolute error as a mesure of prediction performance rather than mean square error, especially given that there is no reason to penalise heavily long stays. In that case, predicting the median of 1 day for everyone would result in an error of the median absolute deviance (MAD). That is, the model would give an error of 2.72 days on average.

This seems quite low considering that stay length ranges from 0 to 283 days. However, 68% of the patients have stay length between 0 and 2 days, 72% between 0 and 3, and 88.5% stay no more than a week. Although longer stays are very dispersed, and therefore hard to predict, they are nonetheless rare. This is the reason why the median performs well *on average*.

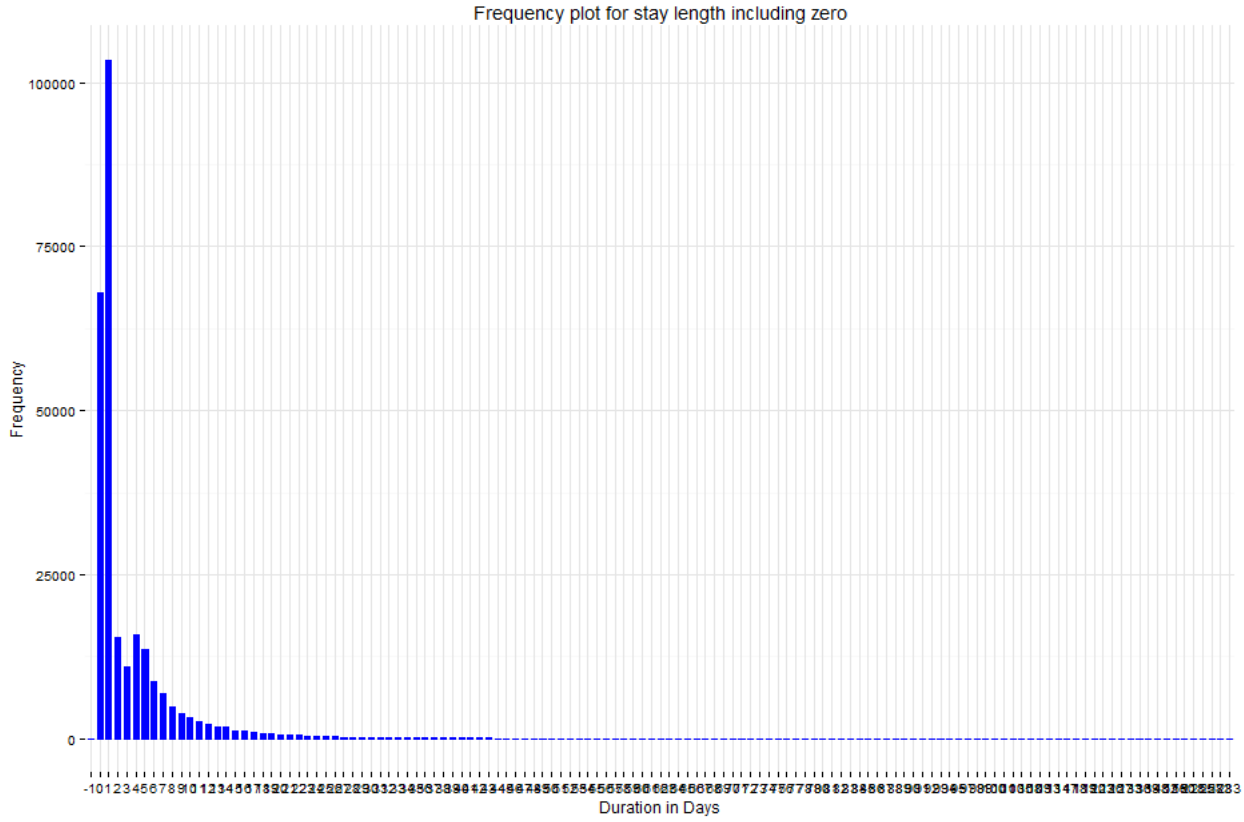


Figure 3: Distribution of Duration of Hospitalisation

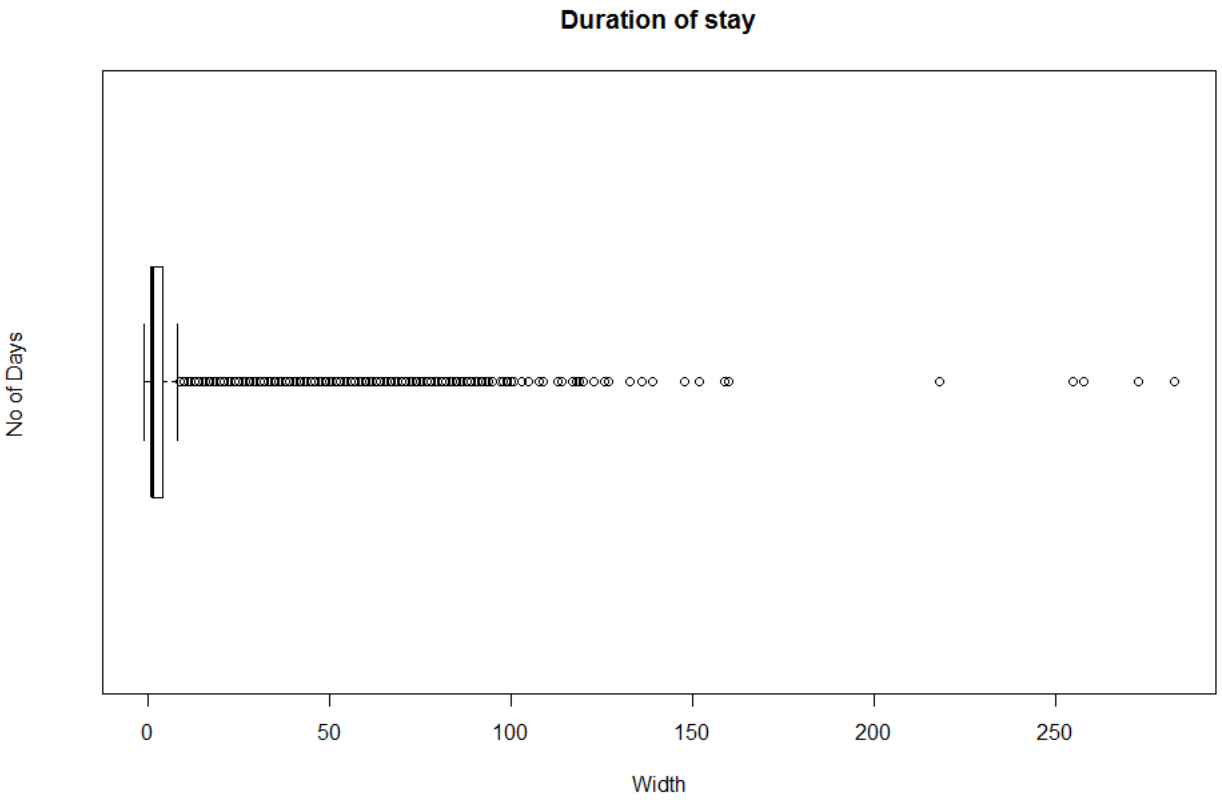
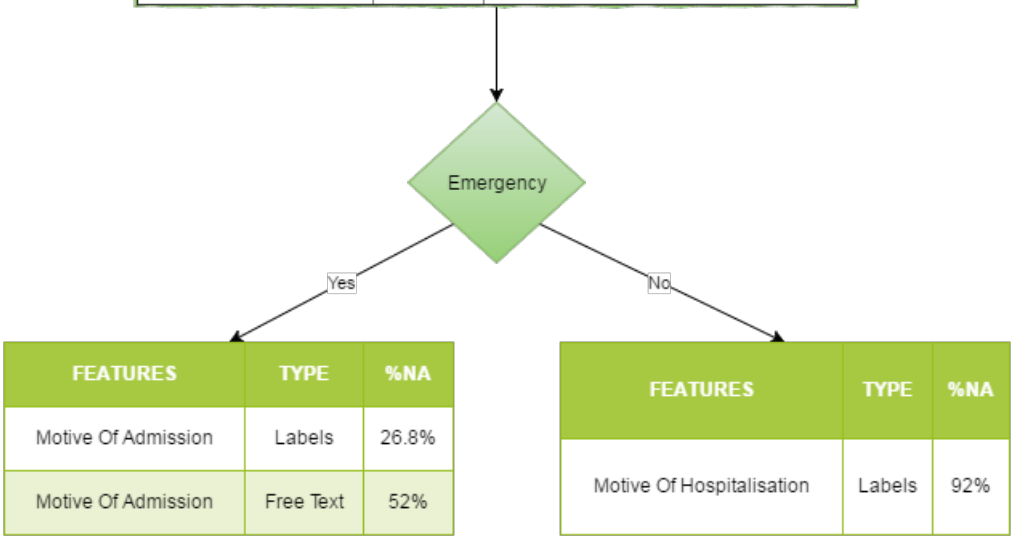


Figure 4: Boxplot of Duration of Hospitalisation

Common Features	Type	Percentage Of Missing Values
date of birth	date	0%**
gender	labels	0%
allergies	labels	0%**
emergency?	boolean	0%
family situation	labels	9.3%
weight	numeric	70.5%
height	numeric	76.6%
bmi	numeric	78.9%
profession	labels	92.6%
antecedents	text	98%**
children	numeric	99.1%
antecedents	labels	99.8%**
country	labels	100%



FEATURES	TYPE	%NA
Motive Of Admission	Labels	26.8%
Motive Of Admission	Free Text	52%

FEATURES	TYPE	%NA
Motive Of Hospitalisation	Labels	92%

Figure 5: Global Features Actually Available in the DPI.

2.3 Input Features

As described earlier, the PMSI is generated 5-10 days after the patient has left the hospital, therefore it is not available at the patient entry in the hospital. Past PMSI from say at least 14 days could be considered. Such historical data is available for a small portion of the patients though, so we postponed this discussion for later.

As for the DPI, some features are common to all patients and others apply only to certain groups. They are summarised in Figure 5 and Figure 6.

Patients can be admitted to the hospital in two ways: through the emergencies or classical way. This distinction gives raise to two different ways of filling information on the patient's stay. If the patient is coming from emergencies, then we find information about his stay in the emergency files. Two pieces of information then are available: the fact that he came through the emergencies and the motive of admission to the emergencies (if available). Otherwise, a motive of hospitalisation is available for other patients. If you want to make use of this information, the patients will need to be split into those groups.

It appears that a lot of data is missing. However, one has to be careful about this claim. For example, missing value for allergy can be filled in with the following logic: a patient that is not part of the allergy data might simply have no allergy. It would be better to distinguish no allergies from missing entry, but currently, we can only assume no entry means no allergy. The same logic applies to antecedents. Missing values for emergencies on the other hand should mean that the patient is not

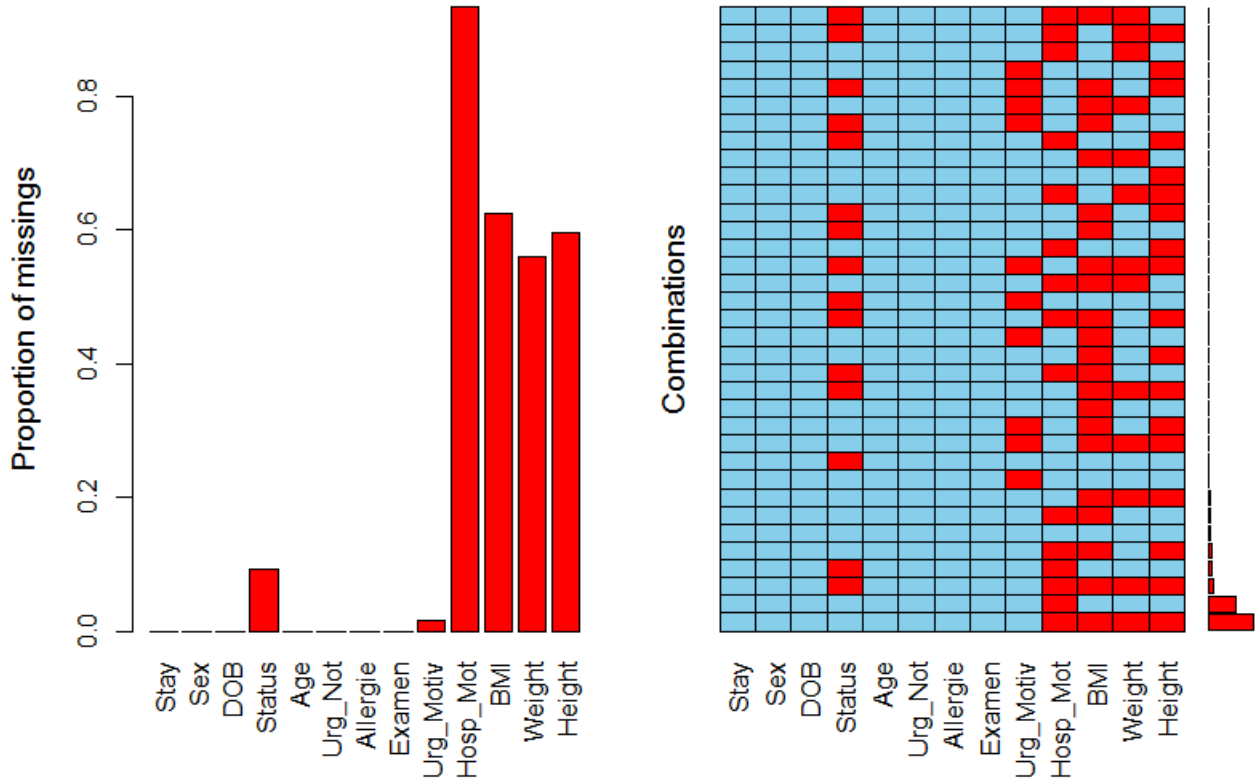


Figure 6: Visualisation of Missing Data. In red are the missing values, in blue the available values or not concerned.

concerned by this feature as discussed above. However, amongst the patients that we know are coming from the emergencies, a small portion of them have a motive of admission given. We do not know of a good way to make up for this missing data.

Another issue is that even though some values are given, they might be erronous. We have spotted an issue in estimating the age. Although all dates of birth are given, they are poorly formatted. They are given in a dd/mm/yy format, having only two digits for the year. This causes noticeable problems as children and centenarians are both present in the population, but difficult to distinguish.

3 Analysis of the Features

The question we try to address in this section is: where can we find relevant information related to the duration of hospitalisation? The DPI information provided comes from a standardized part of a much larger DPI that might contain much data available for patient profiling. However, amongst the feature we were given from the DPI, age, gender and marital status are the ones that were most available. More promising variables are sometimes also present like motive of emergency. In that case, we have to restrict our attention to a subgroup of the population in order to assess the variable relevance.

3.1 Models Based on DPI

We first look at the common variables that are present in DPI as these are readily available. In Figure 7, we see the relation between age, gender and stay length. We notice first that newborns are heavily represented in the dataset, probably due to the presence of maternity in the hospital. The second is that between 21-30 years old, patients are dominantly women, perhaps due to pregnancies. Above 80 years old, women dominate again, but that was expected as they have a longer life expectancy. Also, we note that linear regression exhibits only a very small increasing trend in the stay length with age. However, we note that the outliers, those patients with very long stays, will drag the regression line a

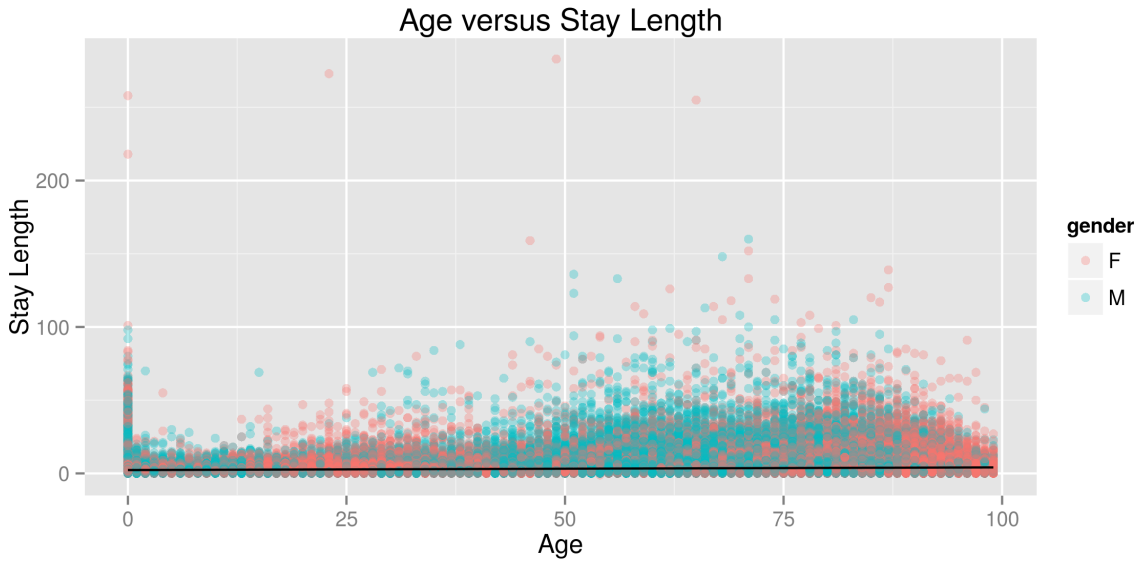


Figure 7: Linear model of stay length against age and gender. The performance is average. We can also distinguish newborns as a special category.

little higher than should be using mean absolute error. The performance of this model is at any rate average.

We then ran several models on age, gender, status, and emergency or not, which are the common features to all patients. Our best model was with random forest. Note that it was optimized for mean square error, but even with that measure, we see that the performance is average.

Model	MAE	MSE
Constant	2.72	36.08
Random Forest	3.06	35.39

As for emergency motives, we first note that only a small proportion of the total population has a motive of emergency, even amongst patients coming from emergency. Therefore, in order to get an idea of the value of that information, we restrict our attention to the group of patients having a motive of emergency. There are 11'164 such patients, and the models' performance is a little more significant:

Model	MAE	MSE
Constant	4.48	50.49
Ridge Regr	4.10	47.15
Random Forest	4.18	49.22

The motive of emergency are divided into 129 labels. As the results show, it is not enough discriminative to significantly improve performance. The free text part of the motive of emergency was not used (too many missing values). One should perhaps consider structuring this information to get better models.

To conclude, in the actual state of the data, the models based on the actual DPI do not show any clear added value.

3.2 Models Including Other Variables from the PMSI

Medical acts data are found in the PMSI. As mentioned earlier, this is produced a week after the patient's exit. However, we can imagine that such data could also be added in realtime to the DPI, although not in our actual data. Therefore, it is interesting to investigate whether medical acts on the patients performed at the beginning of their stay affects the duration of hospitalisation. We do so in relation to classifying whether the patient will stay more than a day or not, that is we bucket the stay

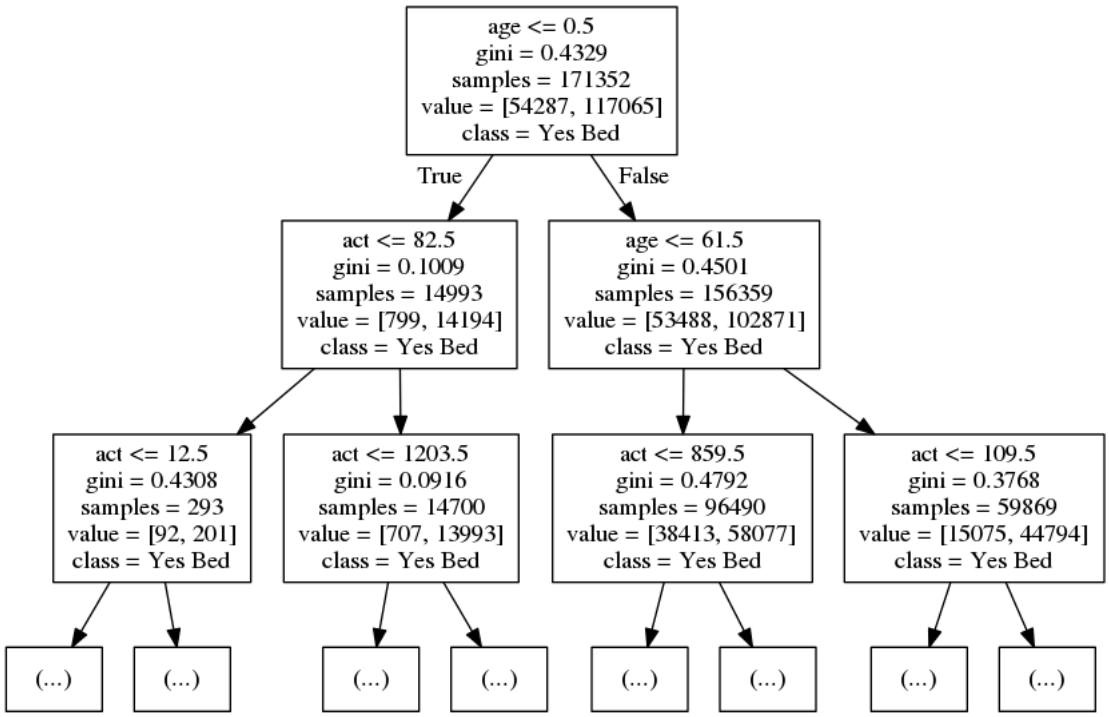


Figure 8: Decision Tree Model for bed usage against age, gender and medical acts performed at the beginning of the stay

length in two categories. We then fit a decision tree model, as well as random forest, to observe this effect. We input three features, age, gender and act. The model is visualised in Figure 8

From this visualisation, we learn something interesting. Although the age seems not important at first, in fact the very first decision is whether the patient is a newborn or not. As mentioned above, newborns are well represented in the dataset and the model tells us that they really should be considered as a separate group.

Another thing we learn from this model is that medical acts have a certain degree of relevance that merits more investigation. In fact, the misclassification rate of the model is 25.3% versus 31.8% for the best constant model. We note that there were 2428 different acts in the datasets. In order to be able to use so many labels, we had to encode them as ordered numbers, and the splitting are based on this ordering. The labels have a significance, the first letter is related to the body part, second letter with the function, and so on. Therefore, the order we chose was alphabetical, so as to reflect that labels starting with the same letters are closer to each other. This relation of labels could be investigated further, and even maybe split into categories themselves before hand.

Finally, we make predictions based on all the variables that we have in the dataset: age, gender, marital status, from emergency or not, acts, exams. The goal is to estimate the value of an aggregation of all information that could be gathered sequentially during the stay. The results are presented in the following table:

Model	MAE	MSE
Constant	2.72	36.08
Random Forest	1.51	14.98

It appears that exams and acts do provide a significant added information and decision trees perform well.

4 Recommendations and Future Directions

We would like our experience with the data to be as beneficial as possible for future use cases. This is why we summarise here some recommendations emanating from our study, as well as possible directions

this project might take in the near future. As we understood, a lot of data is stored and distributed from the databases, but very often in non-standardised formats like free texts and varies from one site to the other. Therefore, there is much work left to do in preparing the data for the project and we are hopeful that relevant information can be found and used to build a performant model. The few quick models that we have generated for this report seem to indicate that medical exams and medical acts performed at an early stage of the stay do influence its length.

Our first recommendations pertain to the data quality so that it becomes easier to work with for future use cases.

1. All dates format should be switched from 2-digits years to 4-digits years to avoid confusion in date of births.

2. The motives of emergencies were given as short free text. We noticed that the bag of words used for the descriptions were not that big. We therefore believe that it would be possible to categorize this data in labels beforehand, like for motive of hospitalisation. Although this would make the data less specific, it increases its usability. Labels could be an addition to the actual data.

3. The motive of hospitalisation were categorised and represented with labels. However, about 90% of these labels were the NA label, or missing label (NP for non-précisé in the database). This begs for investigation as to why this is the case, because as it stands, it is mostly useless.

4. Historical exams of patients were not dated, and therefore impossible to relate to hospitalisation periods. It would be helpful to have these data with timestamps.

5. The absence of a condition could be recorded as much as possible and entered in the database rather than entering only positive patients. For example with allergies, knowing that a patient has no allergy is more informative than having no entry for allergies for him/her. Although helpful, we are aware of memory constraints and that this might not always be feasible.

In terms of modelling, as mentioned before, medical exams and acts seem to be a good place to start to build a model. The acts were taken out of the PMSI, which is unavailable at entry. However, they should be given in the DPI of the patient and we hope that it is indeed possible to include the information in the model early on. Also, surgeries tend to become more and more ambulatory. Since these operations are programmed, information should be available.

On the other hand, it suggests also that the model will need to refine the target. We have seen profiles suggesting that the patient is frequently visiting the hospital, probably for therapy sessions. These patients should not be included in the model, as they surely do not need a bed. This suggests that a better patient profiling would be beneficial.

We summarise in the following list the directions that we have envisioned for building a model, that due to lack of time we could not investigate further:

1. Our decision trees performed better than average. Splitting rules in such trees are based on values of variables only. Perhaps patient profiling could be further used to explicite splitting criteria that are more complex. For example, patients having a profile indicating that they are following a therapy should probably be put into a separate group.

2. The meaning of the codes for medical exams and acts should be investigated further so that the model knows or learns their proximities. For example, two acts performed on the head should perhaps be considered closer to each other than an act on the head and another on an ankle.

5 Conclusion

The real insight from this project was that gathering and preparing the data from the databases is of highest priority at this stage of the project. The delivered data had too many missing entries and the relation among different datasets was incomplete. Nonetheless, we gained insight towards possible

relevant features on which a model could be based. We are therefore hopeful for the success of the project, leading to better hospitalisation management.

With a humble heart, we would like to thank Nicolas Bulteau & Bertrand Cujard from SIB for all their efforts put into this project and their availability. We appreciated working with them and we learned from this wonderful experience.