MedClick: Last Minute Medical Appointments No-Show Management

Inês Tormenta Pinheiro Duarte Ferreira
ines.d.ferreira@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

May 2019

Abstract

A no-show is one of the phenomena that leads to an efficiency decrease in various sectors, including in the health care sector. When a scheduled patient misses an appointment without cancelling, it will not only waste the clinic's resources, but it will also deny medical service to another patient who could have benefited from the respective time slot. This paper describes the research that has been developed in the context of MedClick, an online platform that aims to help medical service providers increase the efficiency of their practices. The solution supports the reduction of no-shows by using supervised learning techniques to predict their occurrence and also by finding replacements to fulfill last-minute vacancy slots. The prediction is performed using a classification algorithm that computes the probability of no-show for each patient based on features that have shown to influence his decision, such as the waiting time, the day of the appointment and the number of previous no-shows. These and other features were extracted from two distinct healthcare datasets that were considered in this research. To reduce the occurrence of no-shows, the system sends reminders and then, the prediction of no-show is performed enough days before each appointment so that there is still enough time to find a replacement, if necessary. In order to select the most suitable classification algorithm to be applied in this research, a 10-fold cross validation was used to perform a comparative analysis between some of the most commonly used algorithms in this type of classification problems, in which the Gradient Boosting proved to have the best performance.

Keywords: No-show, Health Care, Supervised Learning, Classification Algorithms, Cross Validation.

1. Introduction

The world is going through a phase of rapidly escalating costs which implies an efficient use of resources. However, the efficiency of various sectors is increasingly affected by no-shows. This research focuses specifically in the health care sector, in which there are at least two negative effects whenever a scheduled patient misses an appointment without cancelling: firstly, the clinic's resources are wasted and secondly, medical service is denied to patients who could have benefited from the respective time slot. MedClick is an online platform that aims to help medical service providers increase the efficiency of their practices. All the work of this thesis was developed in the context of that platform and it is focused on helping MedClick to achieve their goals by providing tools that, among other features, allow for predicting no-shows and also for fulfilling "last-minute" vacancy slots, by notifying patients whose needs and restrictions are best suited to the time slot. The implementation of these features was rewarding because it will not only help the Portuguese health care providers but also the patients. Furthermore, there are not many systems using techniques based on machine learning to reduce no-shows so if this project proves to be a reliable solution, it may be useful for other businesses.

1.1. Objectives

The goal of this thesis is focused in one of MedClick differentiation functionalities: the reduction of no-shows from patients in medical appointments in order to increase the productivity and the resource usage in health care services. To achieve the desired goal, this research is focused on providing a solution that must be able to:

- Minimize the occurrence of no-shows by using strategies to reduce their probability, such as reminder notifications;
- Build a supervised learning model capable of predicting no-shows based on a given set of features. For this step, several classifications algorithms must be explored in order to find the most suitable for no-shows problem;
- In the case of detecting a future no-show, the system must try to find a suitable replacement;

- Extract data from health care datasets and preprocess it until it is ready to be sent through the learning model and provide reliable predictions;

The above aims are expected to complement and improve the no-show algorithm structure that was previously implemented in MedClick [1].

1.2. Document Outline

This paper is structured as follows: Section 2 provides a review of no-show literature. Section 3 introduces the concept of supervised learning. Section 4 describes the no-show approach that was previously implemented in Medclick. Section 5 describes the proposed solution. Section 6 presents the evaluation tests that were performed along with the respective results and, finally, section 7 concludes the paper by summarizing the developed work.

2. Literature Review on No-Shows

The efficient use of resources is increasingly important so several studies have arisen, focused on detecting the origin of no-shows and finding possible solutions to this problem. Regarding the causes of no-shows, the most commonly reported reason is when the patient forgets his appointment [2]. Therefore, appointment reminders are commonly used to prevent that from happening [3]. Several other reasons are reported for no-shows, such as: financial problems, lack of transportation, competing priorities, bad quality of the service and patient health status [4].

Besides some scheduling systems aimed at reducing no-shows, such as overbooking and open-access [5], there are some strategies such as patient education or patient sanctions. The first consists of providing all the important information in order to ensure that patients feel secure about their appointment. The second is used as an attempt to change the patient's behaviour [6]. In addition, the field of supervised learning has been increasingly explored to reduce and predict no-shows. The previous solution implemented in MedClick uses a hybrid approach that combines logistic regression, as a population-based method, and Bayesian Inference, as an individual-based method. This approach has already been used for reducing no-shows in the healthcare sector [7]. There are many other studies focused on predicting no-shows in different sectors, such as in airline companies [8] and in the hospitality sector [9].

Ample literature is available discussing predictors of no-shows, which can be divided into two categories: patient's characteristics and appointment's characteristics. Regarding the first, several studies have demonstrated that no-show patients tend to be younger [10], unmarried [11], uninsured [12], with psychosocial problems [13] and finally, with prior no-show history. The second category includes the day of the scheduled appointment, the clinic's proximity and, finally, the waiting time, which corresponds to one of the major problems in healthcare services [14]. Although several studies proved the impact of these features, it is important to bear in mind that the results may vary depending on where the study is done.

3. Supervised Learning Algorithms

The idea of supervised learning is to analyze a set of training data and to learn a function capable of predicting the output given new input data. Supervised learning problems can be further divided into regression and classification problems. The prediction of no-shows corresponds to a classification problem, in which a function must predict the class of a given observation. The effectiveness of this techniques depends on the performance of the chosen algorithm and, therefore, it is important to test and consider different approaches. For this research, four classification algorithms were considered, namely Logistic Regression, k-Nearest Neighbors (k-NN), Random Forests and Gradient Boosting. The Logistic Regression applies a logistic function that receives a set of features along with their respective coefficients and outputs the probability of no-show. The coefficients are estimated during the training phase and their values are log odd ratios which may give information on the impact of each feature. Positive coefficients correspond to higher odds of occurring no-show and negative coefficients corresponds to lower odds. A feature with a coefficient near 0 has a low impact on the prediction. The k-NN is one of the simplest algorithms and it predicts the class of a given instance based on the classes of the k nearest neighbors [15]. Random Forest and Gradient Boosting are both ensembles of decision trees, in which the idea is to produce a strong learner by combining a group of weak learners. The first is an extension over bagging which consists of building multiple trees in parallel and combining them together to obtain a more stable and accurate prediction [16]. The second is an extension over boosting, in which the learners are built in a sequential way and each tree corrects the classification error of the previous tree [17].

4. MedClick Previous Solution

A no-show algorithm had already been implemented in MedClick, in which the goal was to find patients interested in filling ”last-minute” vacancy slots [1]. After detecting a slot to be filled, the system would start by getting the filtered list of candidate patients, who would later be notified, from the least likely to miss the appointment to the one with the greatest probability of missing it. This requires a prior computation of the no-show probability associated with each candidate patient, which was being performed using a hybrid approach that combines Logistic Regression, as a population-based method, and Bayesian Inference, as an individual-based method [7]. After the patients were ordered accordingly to their no-show probabilities, the algorithm would go into a loop until it finds a replacement or until there are no more candidates left. At the end, if no replacement was found, the system would notify the healthcare center that the algorithm was unable to fulfill the time-slot.

Despite the satisfactory results, there are some aspects that should have been considered in order to improve the quality of the system. One of the major limitations of this solution is that the algorithm that estimates the no-show probabilities was not being used to its full potential since it was only used to sort the candidate patients list according to their probabilities of missing the appointment. Instead of that, the algorithm could also have been leveraged to predict no-shows.

5. No-Show Management Approach

In this research, a no-show module was developed primarily using NodeJS, with the exception of some machine learning tasks that were implemented using Python due to its highly optimized libraries. The BPMN diagram in Figure 1 presents the workflow of the implemented module, in which two distinct tasks stand out.

The one that is represented at the upper part of the Figure corresponds to the starting point of the module and it consists of preprocessing and using the available data to train the learning model, which is subsequently persisted into the data source in order to be used whenever the system needs to compute a no-show prediction. In the lower part of the figure it is represented the task that will be performed daily (e.g., everyday at 8am), which consists of two sub-tasks. One is responsible for sending reminders to the patients with appointments 5 days from current date and the other one computes the probability of no-shows on the appointments 3 days from the current date, using the model that was previously trained. In the case of detecting a no-show, the system is responsible for notifying possible replacements, from the least likely to miss the appointment to the one with the greatest probability of missing it.

Both learning and prediction may use online or offline approaches. In this research, predictions are computed in real-time using new input data, which corresponds to an online approach. Regarding the learning, the offline approach consists of training the model once on historical data, remaining constant after being deployed to production. However, it is important to ensure that the model does not become unstable, which may happen very often. Considering this, it should be used a batch learning technique which consists of combining both online and offline approaches, since using exclusively an online learning requires to constantly update the model as new data arrives, which is not viable. With a batch approach, the model is re-trained only after a certain number of observations have been inserted into the dataset.

5.1. Dataset

During the development process of this research, two distinct datasets were considered. The first was provided by a portuguese clinic, *MD Clínica*, and it was already considered in MedClick to test the previous algorithm [1]. The second was downloaded from *Kaggle* and it contains data related to 110k medical appointments from Brazil.

Real-world data should not be sent through a model without first being preprocessed since it is often incomplete and it is likely to contain noisy and unreliable data.Considering that, the following sequence of preprocessing techniques was applied

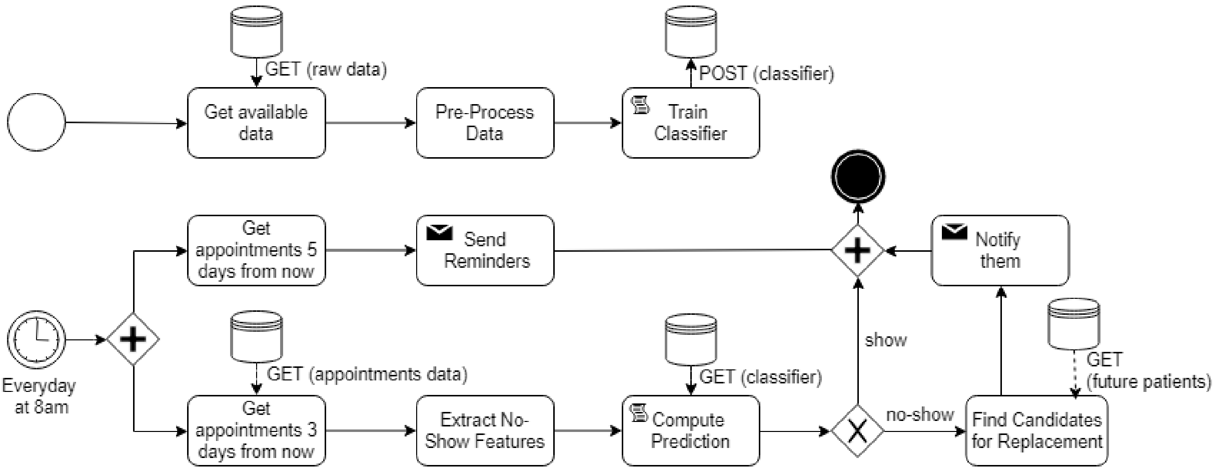

Figure 1: Implementation of No-Show Module

to both datasets:

- Removal of instances containing missing values;
- Removal of instances with inconsistent values, such as postal codes with less than 7 digits,negative ages or scheduling days after the respective day of the appointment;
- Using the available data to extract features that already proved their negative impact on patient's no-show probability;
- Balancing the data by applying SMOTE [18] after splitting the data into training and test sets.

Following, there is a list of the final considered features for each dataset:

MD Clínica Features: postal code, age, marital status, gender, insurance ID, number of previous appointments, number of prior no-shows, physician ID, appointment's day;

Brazil Features: age, gender, scheduling day, waiting time, handicap level, number of diseases, number of previous appointments, number of prior no-shows, physician ID, appointment's day;

6. Evaluation and Results

There are several methods that can be used to evaluate the performance of learning models. In this thesis, a 10-Fold Cross Validation was used as it is one of the most efficient methods that allows model hyper parameters optimization and it also evaluates the model performance with different subsets of data [19]. The idea is to split the data into 10 folds, one of which is used for testing the model and the remaining 9 are used to train the model. Then, the process is repeated 10 times so that each fold will be used once as a test set.

The choice of the metrics depends on the type of problem. Regarding the problem of no-shows, it is known that there are typically more shows than no-shows and, therefore, the dataset used in this project is highly imbalanced since there is a negative majority class highly dominating over a positive minority class. This means that measuring only the accuracy would not be sufficient because, considering a dataset in which 90% of the data belongs to one class, it is easy to create a classification model that gets an accuracy of 90% by simply assigning all data to the majority class. In order to get more reliable results, three additional metrics were also measured: precision, recall and f1-score.

6.1. Impact of Pre-Processing Techniques

As mentioned on section 5.1, raw data should not be sent through a model without first being preprocessed since it is often incomplete and likely to contain noisy and unreliable information. Considering that, several preprocessing steps were performed and this section will cover the impact of those steps on models performance, which will be supported by the results presented in Table 1 and in Table 2, in which the several measures from *10-Fold Cross Validation* are organized based on different pre-processing levels.

Initially, each model was evaluated with raw data and, as expected, the accuracy was the only measure getting good results due to the imbalanced data. This proves that accuracy is not a reliable measure in this type of problems since any model can get good results by simply assigning all data to the majority class (show), which happened in the logistic regression model, whose precision, recall and f1-score had values of 0.

As mentioned above, SMOTE technique was applied in order to balance the data. With this amendment, the models improved in general and their performance measures became more reliable, including the accuracy measure whose scores have decreased.

Finally, some features were extracted from the available data and others were discarded for being irrelevant to the problem. With this final step the models increased their performance, as shown on the lower part of both tables.

In addition, the importance of each feature was analyzed. In contrast to *Brazil* values, the features from *MD Clínica* dataset, have shown a balanced distribution of importance (figure 2). The features from *Brazil* that prove to have the biggest impact on predicting no-shows is the time that the patient had to wait to see their physician and the patient's age (figure 3).

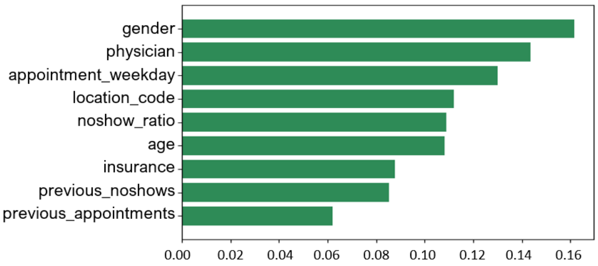

Figure 2: Feature Importance (MD Clínica Data)

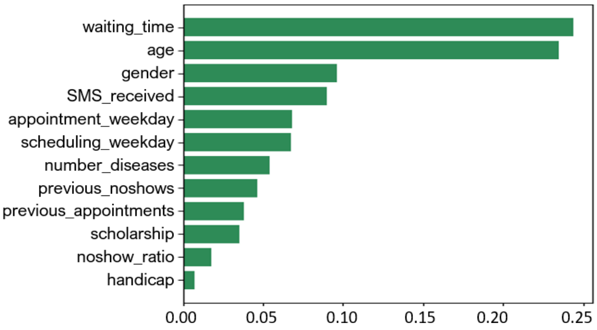

Figure 3: Feature Importance (Brazil Data)

	Evaluation Metric	Classification Algorithm			
		Logistic Regression	k-Nearest Neighbors	Random Forest	Gradient Boosting
Raw Data	Accuracy	0.8278	0.8236	0.8160	0.8274
	Precision	0.0000	0.3211	0.2517	0.5178
	Recall	0.0000	0.0180	0.0315	0.0154
	F1-Score	0.0000	0.0336	0.0540	0.0295
Raw Data (Resampled)	Accuracy	0.5846	0.5660	0.6824	0.7155
	Precision	0.2399	0.2021	0.2205	0.2720
	Recall	0.5293	0.5959	0.3330	0.3307
	F1-Score	0.2943	0.3005	0.2639	0.2798
Processed Data	Accuracy	0.7508	0.6527	0.7601	0.7681
	Precision	0.3552	0.2510	0.3618	0.3861
	Recall	0.3925	0.5049	0.4981	0.5098
	F1-Score	0.3591	0.3334	0.4163	0.4328

Table 1: Performance obtained at different levels of pre-processing (MD Clínica Data)

	Evaluation Metric	Classification Algorithm			
		Logistic Regression	k-Nearest Neighbors	Random Forest	Gradient Boosting
Raw Data	Accuracy	0.7981	0.7605	0.7951	0.7981
	Precision	0.0000	0.2583	0.3356	0.5296
	Recall	0.0000	0.1024	0.0163	0.0010
	F1-Score	0.0000	0.1429	0.0309	0.0022
Raw Data (Resampled)	Accuracy	0.6246	0.5761	0.6009	0.6369
	Precision	0.2621	0.2259	0.2531	0.2709
	Recall	0.4756	0.4579	0.5099	0.4803
	F1-Score	0.2933	0.2991	0.3268	0.3222
Processed Data	Accuracy	0.7304	0.7416	0.8071	0.8091
	Precision	0.4035	0.4015	0.5210	0.5285
	Recall	0.6304	0.5793	0.4373	0.5342
	F1-Score	0.4820	0.4734	0.4732	0.5176

Table 2: Performance obtained at different levels of pre-processing (Brazil Data)

6.2. Choosing an Optimal Threshold

With the exception of k-NN algorithm that directly creates a class output (0 or 1), the remaining models return probability outputs that are subsequently converted into classes by using a threshold probability. The default value for this threshold is 0.5, which means that a probability above that value indicates positive class and a probability below indicates negative class. However, each problem must find their optimal threshold. When the data contains a negative majority class highly dominating over a positive minority class (such as in the no-show problem), the threshold must be chosen considering the precision-recall trade-off, since the precision does not depend on the number of true negatives. Precision and recall are inversely related, which means that decreasing the threshold leads to a decreased precision but to an increased recall. When choosing the optimal threshold for this research, it is important to consider the several clinics in which the solution will be applied since each approach may lead to different consequences. Hence, the threshold must be chosen considering three possible approaches:

- **High Precision & Low Recall:** the model is not able to detect many no-shows but it is highly trustable when it does. This means that the clinic's resources will continue to be wasted but, at least, there will be no overbooking since the system will not schedule replacement patients in slots whose original patient will not fail the appointment. This may be an advantage since it decreases the waiting lists and, consequently, does not decrease patient satisfaction.

- **Low Precision & High Recall:** most of the no-shows are detected but the model also classifies some shows as no-shows. This means that the clinic's resources will not be wasted with last-minute vacancy slots but, the system will accidentally overbook appointments since it will try to find replacements to slots in which the no-show was incorrectly detected. This may lead to long waiting lists and, consequently, to decreased patient satisfaction.

- **Precision $\simeq$ Recall:** In this case, since both have a similar formula, saying that precision is equal to recall is the same as saying that the number of false positive (FP) is equal to the number of false negatives (FN). In other words, the number of no-shows that were incorrectly classified as show (FP) is equal to the number of shows that were incorrectly classified no-shows (FN).

When choosing the threshold value, it is recommended to use precision-recall curves, which uses different probability thresholds to summarize the trade-off between the precision and recall [20]. The Figure 4 presents the precision-recall curve of the three algorithms over the MDClínica data.

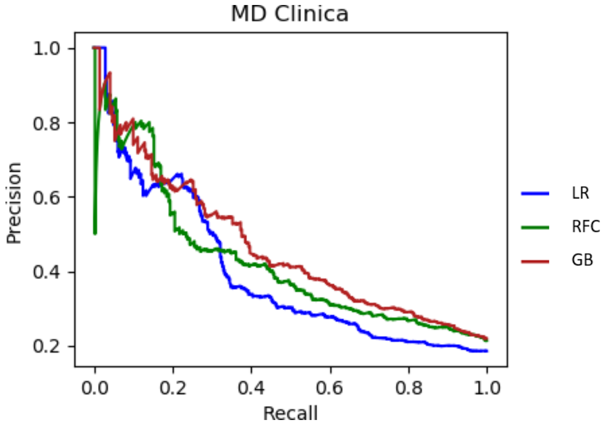

Figure 4: Precision-Recall Curve

In this research, the thresholds were chosen as a way of getting similar precision and recall, which has resulted in a threshold of 0.5 for Brazil data and a threshold of 0.3 for *MD Clínica data*. As mentioned above, the value of this threshold must be adapted, in the future, according to each clinic approach.

6.3. Comparative Analysis of Classification Algorithms

There is a wide range of classification algorithms and each particular problem must find their most suitable algorithm since the effectiveness of the solution will depend on their performance. For this reason, it is important to test and consider different options and as such, four different algorithms were tested in this thesis, namely Logistic Regression, k-Nearest Neighbors, Random Forests and Gradient Boosting. As mentioned above, the evaluation method was a 10-fold Cross Validation. The mean of the results of each algorithm over the 10 iterations is presented in Table 1 and in Table 2, according to the considered dataset.

As shown in the lower part of both tables, the results obtained with *MD Clínica* data proved to be consistent with the results from Brazil data. From these results, despite the slight difference, it is clear that Gradient Boosting outperforms the remaining algorithms in each of the considered metrics. In general, the models achieved good accuracy results but the remaining metrics showed lower values, which might seem an indicator of a bad performance but it is important to consider that the human behavior is extremly complex, which makes it hard to predict. Also, this learning model will be running as a part of a no-show algorithm that supports others strategies aimed at reducing no-shows, such as the reminders approach. Nevertheless, the algorithm that is preferable to implement in MedClick system is the Gradient Boosting whose recall results showed that around 50% of no-shows will be predicted, which leads to an increase in the efficiency of the clinic's resources.

Standard deviations were also computed and their low values shown that there is low variance among the different folds of cross validation, which means that the algorithms would perform similarly with different data sets of the same clinic. In addition, several box plots were provided to illustrate the variability and dispersion of the results of each evaluation metric through the 10 iterations of cross-validation. As shown on Figure 6, a box plot is a standardized way of displaying the distribution of results based on five values [21]: minimum, first quartile (Q1), median, third quartile (Q3), and maximum: The Q1 value, also known as 25h Percentile, is the middle result value between the smallest value and the median of the results, which, in turn, corresponds to the middle value of the results. Q3, also known as 75h Percentile, is the middle value between the median and the highest result.

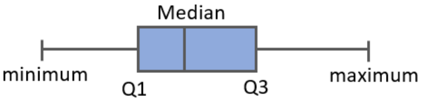

Figure 6: Box Plot Interpretation

Despite the results from this research, it is important to repeat these tests with the new data that will be provided by the Portuguese clinics since the choice of the algorithm depends on the given data.

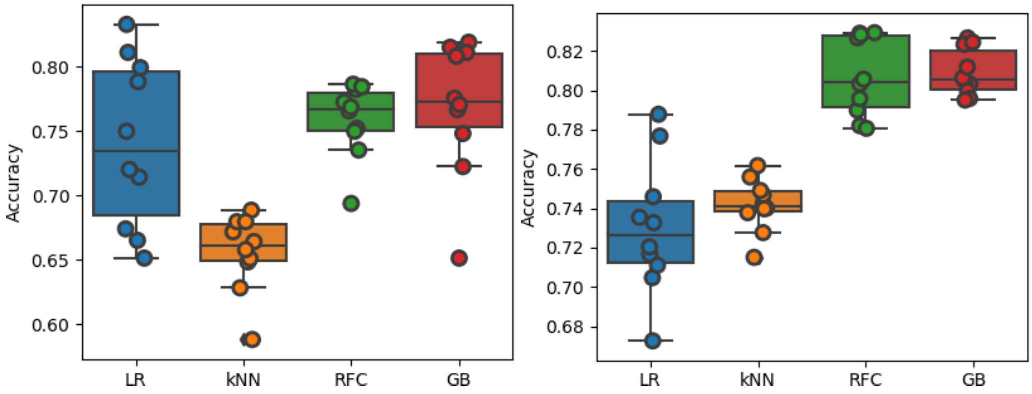

Figure 5: Accuracy results over 10 iterations of cross-validation (MD Clínica data on the left and Brazil data on the right)

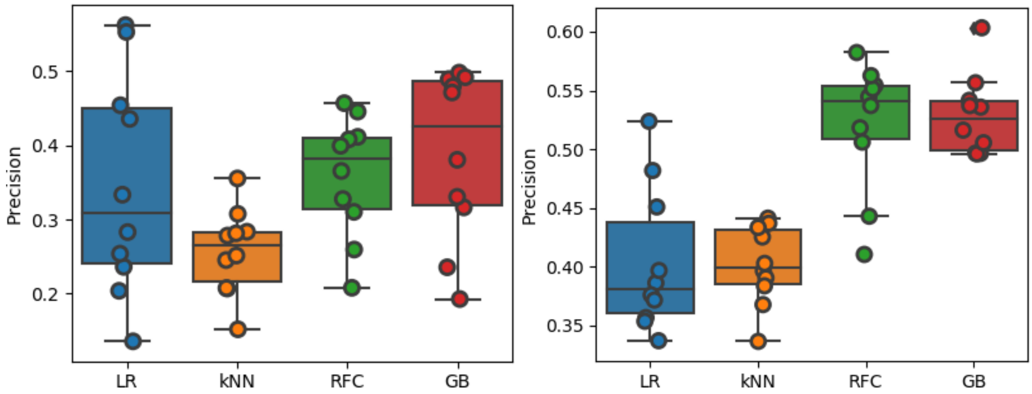

Figure 7: Precision results over 10 iterations of cross-validation (MD Clínica data on the left and Brazil data on the right)

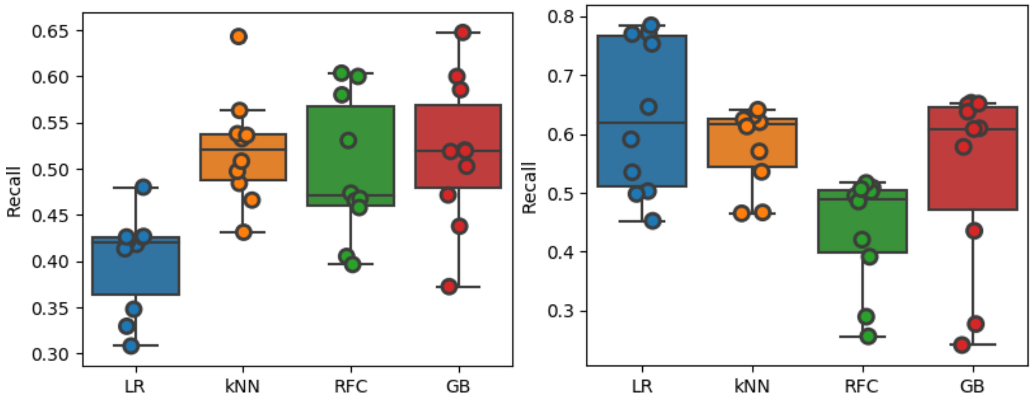

Figure 8: Recall results over 10 iterations of cross-validation (MD Clínica data on the left and Brazil data on the right)

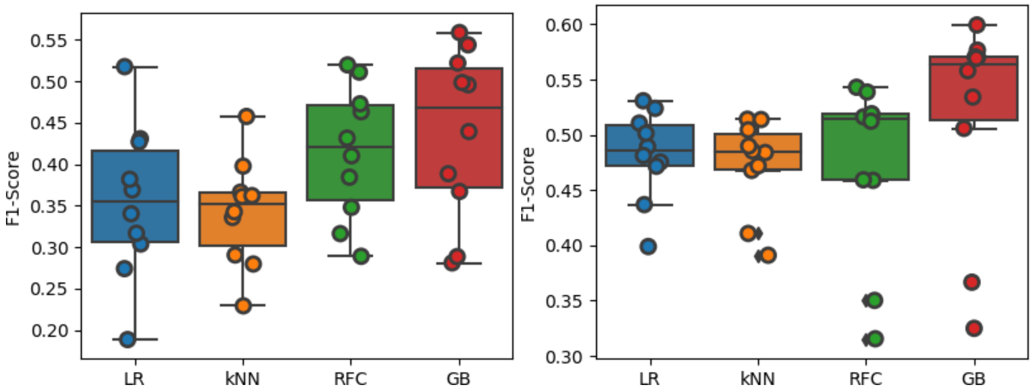

Figure 9: F1-Score results over 10 iterations of cross-validation (MD Clínica data on the left and Brazil data on the right)

6.4. Proposed Solution vs. Previous Solution

This research was applied in the context of MedClick, in which a no-show approach had already been implemented [1], which consists of using Logistic Regression, as a population based method, and Bayesian inference, as an individual method. In this research, four classification algorithms were compared and the one that showed the best performance for the considered datasets was Gradient Boosting, an ensemble of decision trees.

Although both solutions had considered data from the same clinic (*MD Clínica*) during the evaluation processes, the provided data was actually different, not only in terms of quantity but also in terms of features. Also, the tests that were performed for the previous solution did not considered some important aspects, such as the fact that the data is unbalanced, which, despite the seemingly satisfactory results, lead to unreliable conclusions. All these factors make the comparison of both solutions a complex process, making it impossible to perform an objective analysis of the impact of new implemented strategies. However, performing the same tests for both solutions provides a better understanding of which solution is capable of achieving the best performance according to the respective conditions. As such, some tests from the previous solution were repeated using the current available data over the Gradient Boosting algorithm, results of which will be revealed in the following sections.

6.4.1 Model Accuracy

In this test, the data was divided into different portions of test and training sets, ranging from 10% of test data to 100%. In the latter case, all the data was simultaneously used to training and testing the model. The table 3 presents the model accuracy for each split of the previous solution, which, despite the seemingly satisfactory results, lead to unreliable conclusions.

Test data	Accuracy
10%	0.73
20%	0.72
30%	0.7
40%	0.7
50%	0.68
60%	0.67
70%	0.67
80%	0.68
90%	0.67
100%	0.78

Table 3: Results from Previous Solution

From 10% to 90% of test data, the previous model reaches around 70% of accuracy which can be easily achieved by assigning all the data to the majority class (show). As previously discussed, to avoid these unreliable results, other metrics should be measured, namely precision, recall and f1-score. In order to compare both solutions, those tests were repeated with the Gradient Boosting model, results of which are presented in table 4. This time, to ensure the reliability of the results, the accuracy was measured along with other metrics.

Test data	Accuracy	Precision	Recall	F1-Score
10%	0.79	0.43	0.50	0.46
20%	0.79	0.40	0.55	0.47
30%	0.77	0.40	0.46	0.43
40%	0.77	0.42	0.46	0.44
50%	0.75	0.38	0.47	0.42
60%	0.79	0.41	0.43	0.42
70%	0.74	0.34	0.55	0.42
80%	0.76	0.37	0.54	0.43
90%	0.73	0.32	0.50	0.39
100%	0.87	0.85	0.89	0.87

Table 4: Results from Proposed Solution

At first glance, comparing the accuracies from table 3 and table 4, one may assume that the model from this research is better than the previous one. However, in addition to the lack of evaluation metrics from the previous solution, the features that were considered in each solution were not the same, so the performance of both models should not be compared as a way of concluding which one is the best. In the previous solution, only 4 features were considered while the model from this research have considered 10 features, which may justify the difference between both performances.

In both tables, it is possible to notice that the performance lowers slightly as the training data diminishes in size. Also, as expected, with 100% of data being simultaneously used to training and testing the model, the global performance increases.

6.4.2 Using No-Show History for Profile Personalization

The approach from the previous solution consisted on first computing the probability of no-show based on a given set of features (e.g. age, gender, day of the appointment, etc.) and then apply Bayesian Inference to adapt that initial probability to each patient, using their record of no-shows. In table 5 is presented the impact of the patient attendance behavior on the prediction of no-show. The first row corresponds to the initial probability, that was previously computed by Logistic Regression, while the following rows correspond to the first ten appointments of that patient, in which is possible to compare the actual outcome of patient's decision with the predicted probability of no-show. Before comparing both solutions, it is important to notice that, in the previous solution, a probability threshold of 50% was considered while this solution has chosen a threshold of 30%.

Actual Outcome	No-Show Probability
	37%
Attended	18%
Attended	12%
Attended	9%
Attended	7%
Attended	5%
Missed	17%
Missed	26%
Attended	23%
Attended	21%

Table 5: Impact of Patient's Attendance Behavior (Previous Solution)

From the results in table 5, it is possible to notice that as the patient keeps attending to his appointments, his probability of no-show gradually decreases, reaching 5%. However, human behavior is extremely complex so there is no guarantee that in the next appointment will not occur a no-show, as shown on the table 5. For this reason, such importance should not be attached to the patient's attendance behavior, which means that is preferable to implement the solution of this thesis in which the history of no-shows is only considered as a feature of the learning model, along with the remaining features. To support the previous statement, a patient with the same attendance behavior was created in order to repeat this test with the new proposed solution. The respective results are presented in the following table:

With the solution from this research, the patient

Actual Outcome	No-Show Probability
	33%
Attended	17%
Attended	29%
Attended	32%
Attended	29%
Attended	33%
Missed	24%
Missed	29%
Attended	14%
Attended	13%

Table 6: Impact of Patient's Attendance Behavior (Proposed Solution)

probability of no-show does not continuously decrease when he attend to 5 appointments in a row, as presented in the table 6. This time, the prior history of no-shows is considered without being given to much importance, which makes the model more prepared for a sudden shift in patient behavior.

7. Conclusions

The world is going through a phase of rapidly escalating costs which implies an efficient use of resources that can be achieved by reducing the occurrence of no-shows. This section summarizes the work of this thesis which is focused on implementing a solution in the context of MedClick platform in order to prevent and predict no-shows in medical appointments. The section concludes by revealing the major limitations and providing some suggestions for future work.

7.1. Contributions

This research is focused on no-shows of the health care sector and seeks to gather all the necessary information to implement a solution capable of reducing no-shows and, consequently, increase the efficient use of clinic resources. The proposed solution was applied in the context of the MedClick application and aims to improve their system by using the following strategies:

- **Improve the classification algorithm:** The previous solution was based on a hybrid approach which uses both logistic regression for population-based features and bayesian inference for individual features. The only individual feature that was being used to personalize the initial patient probability of no-show was the respective prior history which could be used as a feature of the first classification model. As a way of comparing both options, this thesis performed a comparative analysis between four classification algorithms in order to choose the most suitable for the no-shows problem. Gradient Boosting was the one with the best performance, in which the patient prior history was being sent to the model as one of the many features. After comparing both solutions, the new approach have shown to be more suitable to the no-shows problem, since it proved to be more prepared to sudden shifts on patient behavior.

- **Add relevant features:** in the previous solution, only two features were considered relevant (patient's age and the day of the appointment). In order to provide more information to the model, this solution extracted the following features: patient's age, patient's gender, waiting time, day of the appointment, scheduling day, number of previous appointments, number of previous no-shows, number of patient diseases, scholarship status and finally, patient's handicaps. Most of these features already proved a negative impact on no-show probability in previous studies so, as expected, they improved the model performance.

- **Use the algorithm to detect no-shows:** the previous solution was only using the classification algorithm to sort the candidates list, from the least likely to miss the appointment to the one with the greatest probability of missing it. This solution, in addition, leverages the algorithm to predict no-shows.

- **Use strategies to reduce no-shows:** This solution supports sending notifications before each appointment, in which the patient must confirm their presence. This implementation aims at reducing the probability of no-show and it can be seen as a reminder mechanism since it prevents the patient from forgetting their appointment. In addition, it is also useful for avoiding last minute vacancies because if the patient is already planning to miss their appointment, this mechanism encourages him to notify the clinic in advance.

- **Improve the method of selecting candidates for replacements:** the previous method that was being used to get the list of candidates was not the most appropriate since it was sending numerous notifications to patients who may not be interested. To improve this, this solution uses a list that includes all patients who have already scheduled an appointment at a later date in the same health care center and with the same health professional.

7.2. Conclusions

Unfortunately, it was not possible to perform a more objective analysis of the impact of new implementations since there are many factors that make the comparison between this solution and the previous solution implemented in MedClick a complex process. However, some tests from the previous solution were repeated, which, combined with the tests that were exclusively performed on this research, had resulted in a thorough evaluation, from which the following conclusions have arisen:

- Each classification problem must find their most suitable model and, for that, there are several approaches focused on validating each model. Before choosing the evaluation method and the performance metrics that will be used, it is important to analyze the particularities of the dataset. Regarding the no-shows problem, it is known that there are typically more shows than no-shows, and as such, the data is highly imbalanced. The performance metrics should be chosen taking this into consideration since there are some that may return unreliable results, namely, the accuracy.

- During the pre-processing process, information from the existing studies can be used to understand which features have the biggest impact on no-shows. After extracting those features from the available data, their influence may be confirmed by computing the respective feature importance.

- Some classifiers return probability outputs that are subsequently converted into classes by using a threshold probability. In order to improve the performance of these classifiers, it is important to choose an optimal threshold, which must provide a solution that best fits their needs. When dealing with imbalanced datasets, the threshold is typically chosen considering the precision-recall tradeoff, in which different approaches may be considered. For this research, it is important to consider each clinic in which this solution will be applied since each approach may lead to different consequences, as mentioned in section 6.2.

- The effectiveness of this solution depends on the performance of the chosen algorithm and therefore, it is important to test and consider different options. In this thesis, four different algorithms were tested, namely, Logistic Regression, k-Nearest Neighbors, Random Forests and, finally, Gradient Boosting, which was the one showing the best performance.

7.3. Limitations and Future Work

One of the major limitations of this thesis is that there are features that were implemented in the system whose impact can only be measured once the MedClick application is finished and deployed in a real clinic environment. For example, the sending of reminder notifications and the mechanism of finding replacements. As mentioned on section 5, the classification model should be re-trained to prevent it from becoming unstable. The solution that should be implemented is based on a batch approach which consists of only re-training the model after a certain number of observations have been inserted into the dataset. Another limitation is related to a feature that was not concluded due to lack of time but that must be considered in future work: the waiting lists. After detecting a no-show, the previous solution was trying to find a replacement by sending numerous notifications to patients who may not be interested. This solution improved that method by only notifying patients who have already scheduled an appointment at a later date in the same health care center and with the same health professional. However, the goal was to also allow patients to add themselves in waiting lists and once the system detected a no-show, these patients would be notified. Finally, throughout this research, several parameters were considered, to which default values were assigned, such as, the day of sending reminders, the day of predicting no-shows and the threshold for no-show probabilities. All these variables must be further explored in the future in order to finding their most suitable values.

References

[1] Daniel Sousa. Medclick: Last minute medical appointments no-show. Master's thesis, Instituto Superior Técnico, Lisbon, 2017.

[2] Richard D. Neal, Mahvash Hussain-Gambles, Victoria L. Allgar, Debbie A. Lawlor, and Owen Dempsey. Reasons for and consequences of missed appointments in general practice in the uk: questionnaire survey and prospective review of medical records. *BMC Family Practice*, 6(1):47, Nov 2005.

[3] Kwok Chi Leong, Wei Seng Chen, Kok Weng Leong, Ismail Mastura, Omar Mimi, Mohd Amin Sheikh, Abu Hassan Zailinawati, Chirk Jenn Ng, Kai Lit Phua, and Cheong Lieng Teng. The use of text messaging to improve attendance in primary care: a randomized controlled trial. *Family Practice*, 23(6):699–705, 2006.

[4] Francesca Gany, Julia Ramirez, Serena Chen, and Jennifer C. F. Leng. Targeting social and economic correlates of cancer treatment appointment keeping among immigrant chinese patients. *Journal of Urban Health*, 88(1):98–103, Feb 2011.

[5] Stewart Cameron, Laura Sadler, and Beverley Lawson. Adoption of open-access scheduling in an academic family practice. *Canadian Family Physician*, 56(9):906–911, 2010.

[6] Clare E. Guse, Leanne Richardson, Mariann Carle, and Karin Schmidt. The effect of exit-interview patient education on no-show rates at a family practice residency clinic. *The Journal of the American Board of Family Practice*, 16(5):399–404, 2003.

[7] Adel Alaeddini, Kai Yang, Pamela Reeves, and Chandan K. Reddy. A hybrid prediction model for no-shows and cancellations of outpatient appointments. *IIE Transactions on Healthcare Systems Engineering*, 5(1):14–32, 2015.

[8] Richard D. Lawrence, Se June Hong, and Jacques Cherrier. Passenger-based predictive modeling of airline no-show rates. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 397–406, New York, NY, USA, 2003. ACM.

[9] Nuno Antonio, Ana De Almeida, and Luís Nunes. Predicting hotel booking cancellation to decrease

uncertainty and increase revenue. *Tourism and Management Studies*, 13:25–39, 04 2017.

[10] Suzanne Cashman, Judith Savageau, Celeste Lemay, and Warren Ferguson. Patient health status and appointment keeping in an urban community health center. *Journal of health care for the poor and underserved*, 15:474–88, 08 2004.

[11] Joanne Daggy, Mark Lawley, Deanna Willis, Debra Thayer, Christopher Suelzer, Po-Ching DeLaurentis, Ayten Turkcan, Santanu Chakraborty, and Laura Sands. Using no-show modeling to improve clinic performance. *Health Informatics Journal*, 16(4):246–259, 2010.

[12] Kevin Bennett and Elizabeth Baxley. The effect of a carve-out advanced access scheduling system on no-show rates. In *Family medicine*, volume 41, pages 51–6, 02 2009.

[13] Michael T Compton, Bruce Rudisch, Jason Craw, Tina Thompson, and Dwight Antonio Owens. Predictors of missed first appointments at community mental health centers after psychiatric hospitalization. *Psychiatric services (Washington, D.C.)*, 57:531–7, 05 2006.

[14] Ajay George and Greg Rubin. Non-attendance in general practice: A systematic review and its implications for access to primary health care. 20:178–84, 05 2003.

[15] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[16] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[17] Hastie Trevor, Tibshirani Robert, and Friedman JH. The elements of statistical learning: data mining, inference, and prediction, 2009.

[18] Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813, 2011.

[19] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. pages 1137–1143. Morgan Kaufmann, 1995.

[20] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):1–21, 03 2015.

[21] D L. Massart, Johanna Smeyers-Verbeke, X Capron, and K Schlesier. Visual presentation of data by means of box plots. *LC-GC Europe*, 18:215–218, 04 2005.