A Problem in Brazil's National Healthcare System–No Show to Doctor's Appointment: Why and How to Predict It

1 Abstract

This paper aims to design and implement prediction models using different machine learning algorithms for patients' no-shows, thereby helping Brazil's National Healthcare System reduce the probability of patients' no-show behavior and improve clinical resource allocation. There are 110,527 medical appointments and 14 associated variables in the dataset. We trained six different machine learning models. Among the six models, KNN has the best performance in terms of modeling accuracy, which tells us that using the 'Medical Appointment No Shows' is feasible to predict the patients' no-show behavior in Brail's Healthcare System.

2 Introduction

Medical resource scarcity presents a significant challenge for countries operating on a universal healthcare system, especially for medically underserved areas. Brazil adopted national healthcare in 1988, and a substantial waste of medical resources is the high rate of absences from reserved medical appointments. To illustrate the gravity of the problem, we observed roughly 30% of appointments as no-shows in the dataset we will use. To shed light on the issue and ultimately help conserve medical resources, we employed six machine learning algorithms — k-nearest neighbors, logistic regression, decision tree, naive bayes, perceptron, and random forest— to predict potential medical appointment no-shows and identify factors that hinder patients from attending their appointments faithfully. Consequently, when patients book an appointment, our model can gather data from their medical history and indicate whether a no-show is likely to happen. If so, we can send them more reminders of the appointment and encourage them to save medical resources by canceling appointments that they do not plan to attend.

Most research on medical no-show predictions currently focuses on developed countries' national healthcare systems. However, few researchers or data scientists delve into the medical inefficiency problems in developing countries like Brazil. Our project sets a precedent by looking deeply into Brazil's medical misallocation and making predictions to help the developing countries' governments reduce the risk of patients' no-show behavior. Aside from the topic itself, our techniques for processing data are also novel. We extracted the days of the week when patients made the appointments to observe how they might affect the possibility of patients' no show. Finally, we introduced many scoring metrics to evaluate the effectiveness of different models, including accuracy, f1 score, AUROC, Etc., which helped us get a complete picture of the modeling performances.

3 Background

Patients' no-shows cause substantial medical resource wastage and have profound impacts on patients' health. Therefore, it is becoming a growing problem that the universal healthcare systems must deal with. There have been previous attempts to predict the patients‘ no-shows with supervised missing data imputation techniques, logistic regression, k-nearest neighbors, boosting, decision tree, random forest, Etc. For instance, data scientist Adeline Ong used the same dataset as we did to predict the medical appointment no-shows using random forest. She adopted the binary coding method in the data cleaning part and dropped the observations with logical inconsistency. Furthermore, she used 5-fold cross-validation, and the data were fitted on four different models: logistic regression, naive bayes, k-nearest-neighbors, and random forest. The models were assessed based on mean accuracy and ROC AUC scores. Eventually, her works showed that random forest outperformed the others.

Similarly, the research team led by Guorui Fan, Zhaohua Deng, Qing Ye, and Bin Wang from China used 382,004 original online outpatient appointment records as a dataset to design a prediction model for patient no show. In addition to the machine learning algorithms adopted by Adeline Ong mentioned above, they also used boosting, decision trees, and bagging for prediction. The result was no surprise that bagging, random forest, and boosting had the highest area under the ROC curve and the AUC score, which were 0.990, 0.987, and 0.976.

The researchers who used a deep neural network-based predictive model combined with ten-fold cross-validation also achieved a surprising performance in terms of AUROC score. They published the results in Nature and introduced machine learning strategies that estimate the likelihood of no-shows even in missing data.

In general, many past works, including but not limited to the ones above, have indicated that deep neural networks, random forest, and bagging are the models that outperform the others in predicting medical appointment no-shows.

4 Methods

K-nearest neighbors, perceptron, Logistic Regression, decision tree, random forest, and naive bayes are this project's six main learning algorithms. K-nearest neighbor is a supervised classifier that assumes every data point near each other belongs to the same class. For example, to determine whether a data point belongs to group X or Y, the known algorithm looks at the state of data points around it. If most of the data points around it belong to group X, then it is very likely that the data point(query) is also in group X and vice versa. As patients who were absent for appointments multiple times are more likely to be absent at their next appointment, we hypothesized that K-nearest neighbors, as a model that classifies unforeseen points based on the values of the closest existing points, would have outstanding performance in predicting patients' no show.

Perceptron is a supervised learning of binary classifiers that learns the weights for each input to draw a linear decision boundary. Hence, it is an algorithm suitable for binary decision prediction(no-show). Logistic Regression works by estimating the probability of the occurrence of an event, such as whether patients show up or not, based on a given dataset of independent variables. The algorithm uses a logistic function(log odds) to estimate probability between 0 and 1, which can be interpreted as the probability of each example belonging to a particular class. Hence, logistic Regression might be helpful for binary classification regarding patients' no show.

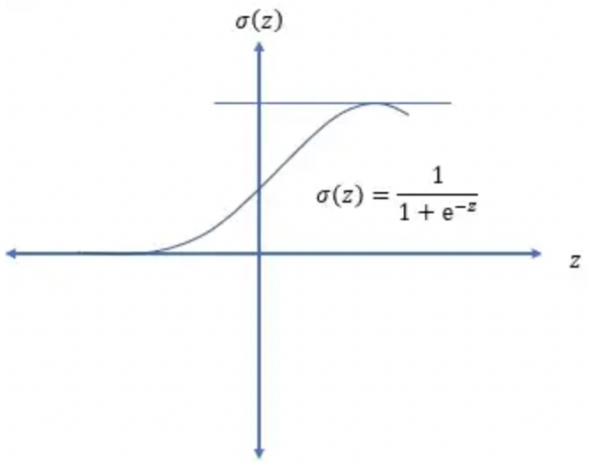

Figure 1

Decision tree works by classifying instances by sorting them down from the root node to the leaf nodes. Each internal node represents a test on a feature, and each branch denotes a value. Every leaf node represents a class label. Every path along the branch stands for a decision rule following successive choices. As for decision tree, it comes without saying that its natural structure, which traverses sequentially through the branch by evaluating the truth of each logical statement until the final prediction(leaf node), makes itself well to predict the binary result of no show.

Building upon decision tree, random forest model combines several decision trees to prevent overfitting and, therefore, gives accurate and precise prediction results. Due to its similar algorithmic feature with decision tree, random forest is suitable for binary prediction and meanwhile could produce better performance by eliminating overfitting problems.

Naive bayes algorithm puts the assumption, which is the independence among the features, to the Bayes' theorem. In gaussian naive bayes algorithm, continuous values associated with features are assumed to be distributed in gaussian distribution. Naive bayes algorithm works exceptionally well with categorical variables and could be used for binary class prediction. Since the dataset we chose contains multiple categorical variables and binary class labels, we decided to take this algorithm into our consideration.

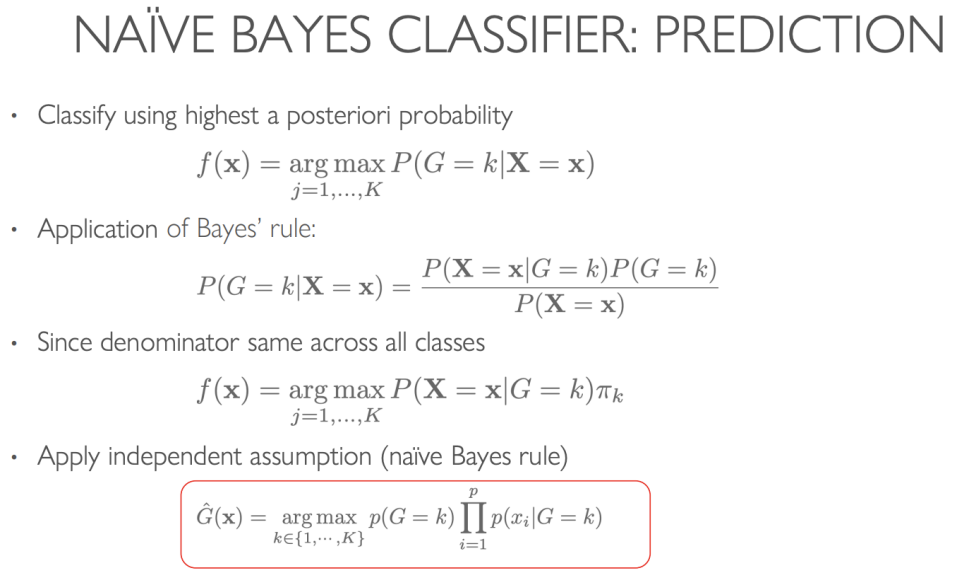

Figure 2

5 Experiments/Results

5.1 Data Preprocessing & Modeling Choice

The 'Medical Appointment No Show' dataset is provided by kaggle. The full dataset has 110, 527 medical records from public healcare system in a Brazilian city and contains 14 features related to patients' no show such as gender, age, patientID, alcoholism, diabetes, handcap, etc. We extracted 'schedule hours,' 'appointment weekday,' and the days in between the time patients scheduled the appointment and the appointment's actual date. We believed that these three new features extracted from the time stamp could be part of the reasons why some of the patients missed their appointments. Additionally, we applied one-hot encoding to the 'gender' and 'no show' variable because it makes our training data more expressive and useful, and it can be rescaled easily. By using numerical values to replace 'True' and 'False,' we were able to more easily determine the probability for our values, increasing the training accuracy.

We also dropped columns that we no longer needed or columns with logical inconsistency such as ages below zero and when scheduled dates were after the appointment dates. To bring all features to the same magnitude, we standard scaler to ensure that features like 'age,' 'Dats_in_between,' 'ScheduleHour' have zero as mean and one as variance.

The dataset was split into 40% test and 60% training sets. The models being applies are k-nearest neighbors, logistic regression, decision tree, gaussian naive bayes, perceptron, and random forest. All models can fit exceptionally well with binary prediction problems, and we would evaluate the effectiveness of each model through different metrics and eventually select the one with the highest accuracy by comparisons.

5.2 Models Performances & Evaluation

Using accuracy, the models' performances are pretty similar and consistent. With knn ranked the highest(0.797) and the perceptron ranked the lowest(0.689), all other models have pretty decent performance scores ranging from 0.771 to 0.797.

For f1 score, KNN has the lowest performance(0.010), and, surprisingly, perceptron and naive bayes have relatively better performances, which are 0.222 and 0.180. Other than these three algorithms, other models have poor performances: with random forest (0.017) having a similar low score as KNN, while logistic regression(0.030) and decision tree(0.058) perform slightly better but still towards the lower end.

For AUROC, random forest(0.7188) and decision tree(0.7095) outperform the other models. Meanwhile, models like perceptron and naive bayes have relatively low performances, which are 0.546 and 0.630.

5.3 Parameter Tuning

Out of all six models, only decision tree, k nearest neighbors, logistic regression, and random forest have parameters. And we left logistic regression and random forest untuned. The former simply takes too much computational time during the tuning process and improvements are limited even with the optimal iteration size. Also for random forest, while the larger the tree sizes, the lower the variances, the tree size doesn't really affect the modeling performance once it gets to a certain number. As the depth and minimum sample leafs both belong to parameters of decision tree and random forest, we could simply tuned once to find the optimal parameters for both models. Hence, we only perform hyperparameter tuning for k nearest neighbors and decision tree in this project.

With 5 fold cross validation, we used GridSearchCV(5 folds, [{'n_neighbors': range(90, 150, 5), 'metric': ['euclidean', 'manhattan']}] to find the optimal parameters for KNN algorithms. After the parameter tuning, we found that the optimal metric for knn is 'euclidean distance, and the optimal number of neighbors is 100. By applying the optimal parameters to the algorithm, we got new knn performance with AUROC scoring metric is 0.6881, the same as the original result. The reason behind getting the same performance result after the parameter tuning is that we happened to use the optimal metric and number of neighbors before the tuning.

For decision tree parameter tuning, we also used GridSearchCV(5 folds, [{'max_depth': range(1, 11,1), 'criterion': ['gini', 'entropy'], 'mini_samples_leaf': range(1, 11, 1)]. Afer the parameter tuning, we found that the optimal metric for decision tree is gini index, the optimal maximum depth is 6, and the optimal min_sample_leaf is 2. By using the optimal parameters, the new decision tree performance using AUROC scoring metric is 0.7177. The accuracy is approximately one percent higher than the original decision tree accuracy.

Model	Accuracy	F-1 Score	AUROC
K-Nearest Nerighbors	0.797	0.010	0.688
Decision Tree	0.794	0.058	0.709
Logistic Regression	0.793	0.030	0.66
Naïve Bayes	0.771	0.180	0.63
Perceptron	0.689	0.222	0.546
Random Forest	0.797	0.017	0.719

Figure 3. Modeling Performances Before Parameter Tuning

Model	AUROC
K-Nearest Nerighbors	0.688
Decision Tree	0.718

Figure 4. Modeling Performances After Parameter Tuning

6 Discussion

Based on our preliminary results, in terms of raw accuracy, the k-Nearest Neighbors model tends to outperform every other model, just short of the major class percentage. This is quite surprising because kNN is a relatively primitive model, and we didn't expect it to dominate other more advanced models such as perceptron and assemble predictors such as random forrest. However, we theorized that since one individual could make multiple visits and get logged, the person may have exhibited habitual behavoir, be it show or no-show, that caused the kNN model to locate the same person's previous entry. Therefore, a person who is likely to make no-shows demonstrated similar patterns of attendance which was picked up by the kNN model. On the other hand, the alternative explanation could be that kNN looked at closeast k neighbors, and, due to the dataset being imbalanced, tend to always predict negatively. Therefore, the kNN model wins by more blindly predicting negative result.

From the F1 score, moreover, we could discern that kNN fell short compared to other models such as random forest, naive bayes, and logistic regression. As this score reflects more of a balance between precision and recall, a model that blindly predicts one label over the other will be scored very low, and that was the case for kNN. Therefore, it can be discerned that our second explanation stands that kNN outperformed all other models in accuracy because accuracy is not a fit evaluation metric for such imbalance dataset. AUROC, as a metric, reflected similar trends by placing kNN on the lower end and putting random forest at the higher end. Something that is interesting and worth noting is taht logistic regression and naive Bayes tend to outperform other models in F1 score, meaning that these two are likely to pick up the minor positive class while sacrificing the accuracy by increasing false positives. Moreover, in the aspect of predicting no-show in Brazil with the intention to curb such behavior, a higher weight or sensitivity towards positive results are preferred. Therefore, AUROC was used as the evaluation metric for kNN and decision tree models after parameter tuning.

We performed parameter tuning only on kNN and decision tree models because of the size of the dataset, which gets really slow on other predictors such as the random forest. Comparing the AUROC after tuning with that before tuning, however, we observe no increase in kNN model while significant boost from the decision tree model. There was no remarkable change in AUROC of the kNN model because we happened to have picked a value of the number of neighbors that was close to the final tuned parameter. On the other hand, parameters for the decision tree, such as the maximal depth and the minimal leaf sample size, were changed drastically after parameter tuning, which resulted in the noticeable boost in performance evaluated by AUROC.

7 Contribution

In terms of contribution, both of us worked on the project altogether and so there there was not any specific part that only a certain group member worked on separately. From the initial brainstorming process to the final deliverable part, we all worked together and contributed equally. For every task concerning the project, such as data preprocessing, machine learning models selections, slides preparation, final report, etc, we always met in person to collaborate and and work together to constantly provide feedback to one another and help each other out whenever we face difficulties in our project. For the initial process of brainstorming, we went through a lot of datasets on kaggle and eventually agreed on using 'No show to Doctor's appointment' because of the significance of improving patients' cate and clinical resource allocation. We later on worked together to develop and implement the machine learning models and the evaluation metrics that we thought would be suitable for this specific dataset. Whenever we faced a bug, we helped each other out by looking up resources and debugging the problems.

Code:
https://drive.google.com/file/d/1sBIWYBEyl0_R8JEDEnxG7EK3_ONWoQ-m/view?usp=sharing

Dataset:
https://drive.google.com/file/d/1VCdTiviaTVUQ-rM5qdOzt1ia6j8m7VsD/view?usp=sharing

8 Work Cited

Adeline Ong. 'Using RandomForest to predict Medical Appointment No-Shows.,' Towards Data Science, Nov 19, 2019.

Liu, D., Shin, WY., Sprecher, E. et al. 'Machine learning approaches to predicting no-shows in pediatric medical appointment.' npj Digit. Med. 5, 50 (2022).

Guorui Fan, Zhaohua Deng, Qing Ye, Bin Wang, Machine learning-based prediction models for patients no-show in online outpatient appointments, Data Science and Management, Volume 2, 2021, Pages 45-52, ISSN 2666-7649, https://doi.org/10.1016/j.dsm.2021.06.002.

Liu, D., Shin, WY., Sprecher, E. *et al.* Machine learning approaches to predicting no-shows in pediatric medical appointment. *npj Digit. Med.* 5, 50 (2022). https://doi.org/10.1038/s41746-022-00594-w