**os**

Strategic Capstone Project

No-show Medical Appointment Research

By Huu Lam Nguyen

**DSC 789-001 Fall 2020**

Strategic Capstone Project

**Mentor** Pro. Zahra Sedighi Maman

**INTRODUCTION**

“Missed appointments cost the U.S. healthcare system $150B each year”, according to Jamie Gier, Chief Marketing Officer at SCI solution. Nowadays, we manage a professional responsibility for ourselves and with friends and family, in addition to our professional obligations, it’s no surprise that some appointments, meetings, and reservations just get missed. We understand the impact these missed opportunities have on our personal lives, but in the professional world (especially healthcare), the impact can be costly in both time and money. The total cost of missed healthcare appointment in the United States every year is $150 billion and individual physicians an average of $200 per unused time slot. whether or not patients show up, healthcare organizations and medical practices still have to pay their staffs and cover expenses like rent and the cost of equipment. But above and beyond the economic implication, no-shows have a direct impact on individual’s health. A missed medical appointment could pose serious health risks for patients as it could mean the difference between catching a disease early on or too late. In addition, an inefficient scheduling process can wreak havoc and raise stress levels for both a health systems’ staff and patients.

**Problem Description**

Previous research indicates patients who miss appointments tend to be, and of lower socioeconomic status. Historically, they also often have a history of failed appointments, government-provided health benefits, and psychosocial problems. They are also less likely to understand the purpose of the appointment. No-show rates increase with increasing time between scheduling and the actual appointment. Longer waiting times have been shown to be related to lower satisfaction, which, in turn, leads to less reliable appointment keeping. In addition to forgetting appointments, patients have provided several reasons for no-shows. Logistical issues include trouble getting off work, childcare, transportation, and cost. In addition, both patients who felt better and patients who felt too unwell to come failed to show.

**Background**

This report also considers some methods which were utilized in the past. For instances, I will list some previous researching reports in the same scope of this research in order to highlight the methods in the past. First, I have chosen research article named “Why We Don’t Come: Patient Perceptions on No-Shows” by Naomi L. Lacy, Audrey Paulman, Matthew D. Reuter and Bruce Lovejoy (November 2004). They chose a qualitative design to allow broader exploration of issues and to limit the impact of the researchers’ preconceptions of the causal basis for failed appointments. The participants were patients at an urban, university-affiliated family practice clinic. The clinic is located in an ethnically diverse neighborhood and serves a predominantly low-income population. Interviews, conducted in the clinic’s examination rooms, were tape recorded and later transcribed verbatim. The second research article is “Time and Money: Effects of No-Shows at a Family Practice Residency Clinic” by Charity G. Moore, PhD; Patricia Wilson-Witherspoon, MD; Janice C. Probst, PhD (2001). The method of this research article is Schedule information was retrieved for 4,055 visits over 20 business days. Data were collected on appointment status (show, no-show, cancel, walk-in), time allocated for the appointment, charges for visit, date and time of the visit, and other appointment information. The third article research is ”No-shows in appointment scheduling – a systematic literature review” by Leila F.Dantas, Julia L.Fleck, Fernando L.Cyrino Oliveira, Silvio Hamacher. This research utilized data collection from Scopus database, which is the largest online database of peer- reviewed literature. Additionally, the analysis of surveyed studies followed a stepwise approach that included pre-analysis, material exploration, and treatment, inference and interpretation of results. The fourth research article is “The Effectiveness of Outpatient Appointment Reminder Systems in Reducing No-Show Rates”, its method they surveyed patients who arrived in at the clinic. The next article is “Predicting No-Shows in Radiology Using Regression Modeling of Data Available in the Electronic Medical Record”. The purpose of this research to test whether data elements available in the electronic medical record (EMR) can be effectively leveraged to predict failure to attend a scheduled radiology examination. They used descriptive statistics and logistic regression models to test whether these data elements could predict failure to attend a scheduled radiology examination.

**Dataset Description:**

The dataset for the capstone project is obtained from Kaggle.com, which is a healthcare provider dataset from the state in Brazil. Kaggle.com is known to hold data science project competitions.

The dataset is on Moodle. In this dataset, there are 72,607 cases (patients) and 16 variables in total (See Table 1). Two of these 15 variables are the ID (patient and appointment ID) variable while one of them is the dependent (outcome) variable (Show up).

Table 1. Dataset Description

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| Age | Patient age |
| Gender | Patient sex |
| Month | Month of the appointment |
| Appointment day | Day of the appointment |
| Lead time | Waiting time for the appointment in minute |
| Calling time | The hour of calling to schedule an appointment |
| Appointment reminder | Status of receiving reminder message or call |
| Alcoholism | Status of being an alcoholic |
| Financial aid | Status of having a financial support |
| Handicap | Status of being handicap |
| Hypertension | Status of having high blood pressure |
| Diabetes | Status of being diabetic |
| Time between appointments | Number of days between two consecutive scheduled appointments |

**Research Questions:**

1. Can we predict the no-show statuses using data analytical techniques?
2. What are the important features in order to predict the no-show status?
3. What is the importance of the features?
4. How do the different sampling techniques affect the prediction of no-show statuses?
5. How do the different feature selection methods affect the prediction of no-show statuses?
6. How does the prediction performance for the no-show statuses change for more complex analytical models?
7. How does the presence of outliers change the prediction performance for the no-show statuses?

**METHOD**

This report applies CRISP-DM, which stands for CRoss-Industry Standard Process for Data Mining.

Diagram

Description automatically generated

Figure 1 CRISP-DM methodology framework

**Step 1: Business Understanding**

No-show in medical appointment has significant impact on our personal lives as well as costly in both time and money. For instance, the total cost of missed healthcare appointment in the United States every year is $150 billion and individual physicians an average of $200 per unused time slot. whether or not patients show up, healthcare organizations and medical practices still have to pay their staffs and cover expenses like rent and the cost of equipment.

**Step 2: Data Understanding**

* Language and libraries: Python, NumPy, pandas, seaborn, and matplotlib
* In this dataset, there are 72,607 cases (patients) and 16 variables in total. Two of these 15 variables are the ID (patient and appointment ID) variable while one of them is the dependent (outcome) variable (Show up).
* Female is 68% and Male is 32%

Chart, pie chart

Description automatically generated

* Percentage of persons that do not miss appointments: 79.06% and percentage of persons that miss appointments: 20.94%

Chart, pie chart

Description automatically generated

* Age:
  + Child (age < 12) : 19.5%
  + Teenager (12 <= age < 18): 5%
  + Young Adult (18 <= age < 25): 6%
  + Adult (25 <= age< 60): 47%
  + Senior( age >= 60) : 22.5%

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

* Correlation between all variables in the dataset:

Chart, histogram

Description automatically generated

**Step 3: Data Preparation**

* Binary coding (1,0) no-shows, gender and handicap. I assumed that handicap was suppose to be binary.
* Dropping observations that had logical inconsistencies, such as negative ages
* Set gender, age, month as integer
* Binning age, waiting\_time\_minute and day\_before\_app (days before next appointment)
* Feature Engineering:
  + Total conditions was the sum of hypertension, diabetes, handicap and alcoholism. In other words, it indicates the number of conditions a patient suffers from.
  + Age has been binning in five group which are child, teenager, young adult, adult, and elder.

**Step 4: Model**

**Language and libraries:** Python, matplotlib, seaborn, pandas, numpy, and scikit learn.

**Feature Selection Method:** weight of evidence (WOE) and Information value (IV), powerful techniques to perform variable transformation and selection.

**Train and test dataset:** test 20% and train 80%

**Balancing:** K-fold Cross-Validation

**Models:** logistic regression, K-Nearest Neighbor, Naive Baye, Random Forest

**Retrain and retest:**  binning Age to five categories child, teenager, young adult, adult, and elder to improve models’ performance.

**RESULT**

**Selected feature:**

|  |  |
| --- | --- |
| **Feature** | **IV Score** |
| age | 0.06 |
| calling\_time\_hour | 0.02 |
| waiting\_time\_minute | 0.66 |
| financial\_aid | 0.01 |
| sms\_received | 0.09 |
| day\_before\_app | 0.01 |
| total\_conditions | 0.01 |

**Model Performance Before Retrain and Retest:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Sensitivity** | **Specificity** | **ROC AUC** | **F1-Score** | **Accuracy** | **Precision** | **Recall** |
| **Logistic Regression** | 0.012558 | 0.991547 | 0.504116 | 0.880514 | 0.787294 | 0.791997 | 0.991313 |
| **K-Nearest Neighbor** | 0.049393 | 0.978544 | 0.513856 | 0.878863 | 0.785969 | 0.795314 | 0.982040 |
| **Naive Baye** | 0.176224 | 0.905505 | 0.545396 | 0.854954 | 0.756340 | 0.807526 | 0.908305 |
| **Random Forest** | 0.007953 | 0.997724 | 0.502906 | 0.882790 | 0.790531 | 0.791583 | 0.997757 |

**ROC Curve Analysis:**

**Chart, diagram

Description automatically generated**

**Answer Research Questions:**

1. Can we predict the no-show statuses using data analytical techniques?

Yes. We can predict no-show statuses using data analytical techniques. There are different techniques for Data Analysis depending upon the question at hand, the type of data, and the amount of data gathered. Each focuses on strategies of taking onto the new data, mining insights, and drilling down into the information to transform facts and figures into decision making parameters. In this research paper, I have applied Regression Analysis to predict outcome of the dataset.

1. What are the important features in order to predict the no-show status?

There are 7 important features in order to predict the no-show status:

* waiting\_time\_minute: patient have to wait in the clinic (minutes)
* sms\_received: whether patient received message reminder to come to the clinic
* age: patient’s age
* calling\_time\_hour: the clinic contacted patient during the day (hours)
* financial\_aid: patient received any financial support
* day\_before\_app: days before next appointment at the clinic.
* total\_conditions: patient have at least handicap or alcoholism or diabetes or hypertension

1. What is the importance of the features?

In my analysis, it shows waiting\_time\_minute feature has the highest IV score (0.66) which means waiting\_time\_minute has the most significant effect on no-show status. The second rank is sms\_received (IV score 0.09), third is age (IV score 0.06).

1. How do the different feature selection methods affect the prediction of no-show statuses?

In my research, I have used 2 methods: IV Score and Random Forest.

IV Score, selected features are financial\_aid, sms\_received, day\_before\_app, total\_conditions, age, calling\_time\_hour, waiting\_time\_minute, show\_up

Radom Forest, selected features are gender, app\_day, month, calling\_time\_hour, sms\_received.

The two methods above show there are two groups of important features which mean each method affects the prediction of no-show statuses.

1. How does the presence of outliers change the prediction performance for the no-show statuses?

There are three variables have outlier which are age, waiting\_time\_minute, and day\_before\_app.

**DISCUSSION AND CONCLUSION**

**DISCUSSION**

In the Logistic Regression, the sensitivity is 0.012558, which means 1.12558% of total number of no-show people was predicted correctly by the model. The specificity is 0.991547, which also means 99.1547% of total number of show-up people was predicted correctly by the model.

In the KNN model, the sensitivity is 0.049393, which means 4.9393% no-show people was predicted correctly by the model. On the other hand, the specificity is 0.978544, which also means 97.8544% show-up people was predicted correctly by the model

In the Random Forest, the sensitivity is 0.007953, which means 0.7953% no-show people was predicted correctly by the model. However, the specificity is 0.997724, which also means 99.7724% show-up people was predicted correctly by the model.

Finally, Naïve Baye has different performance than the others. The sensitivity is 0.176224, which means 17.6224% no-show people was predicted correctly by the model. The specificity is 0.905505, which means 91.5505% show-up people was predicted correctly by the model.

In conclusion, specificities of all models are from 90% to 99%, which means the difference between all predicted-correctly show-up people models is not large, it quietly same. Therefore, I would like to focus on sensitivity. Base on models’ performance, Naïve Baye is better than the others since its sensitivity is higher (17.66224%) then the others (logistic regression 1.12558%, KNN 4.9393%, 0.7953%)

**CONCLUSION**

For this dataset, which is no-show medical appointment, we can apply analysis techniques to predict patient statuses. There are 7 factors that the clinic has to consider decreasing the number of no-show people. Firstly, the clinic should have a better plan to reduce waiting time at the clinic. The percentage of no-show people who are adult is 47%, which means these people mostly have jobs or working, they probably do not have enough time to spend in the clinic for waiting. My suggestion is that the clinic should improve the schedule between each patient and reduce waiting time. Secondly, as we all known that smartphone is now an essential item, the clinic can improve reminder system by SMSs. I suggest that the clinic should check patient’s phone numbers to make sure that their record is correct. Then, keeping remind them more often especially who is adult, senior, child, young adult and teenager. Third, age is also the matter of no-show status at the clinic. There 47% of no-show people who is adult, then senior, child, teenager, and young adult. Additionally, the clinic should consider time in a day for calling, and whether the patient have financial aid or not, how many days between their two separate meeting and their conditions (hypertension, alcoholism, diabetes, and handicap).

Finally, I would say to reduce no-show rate. The clinic should more consider patients who are adult and senior. Improving some services such as SMSs reminder system, reducing waiting time for those people who are in the age of adult and senior categories.