Medical Appointment No Shows Issue

Data Visualization Application

Hsuan-ju Lin

Context

- Introduction
- Research question
- Dataset
- Preprocess
- Exploratory Data Analysis
- Findings
- Future works
- Appendix

Introduction

- Managing patient scheduling is important to clinics. Overbooking could lead to employee burnout and patient dissatisfaction. In additions, patients no-show to the appointment causes opportunity loss for clinics. Thus, in this research I will address medical appointment no-show issue

Research Question

- Build a model to discover who will not show up for medical appointment
    - 1. Which model has a better performance to classify a no-show appointment?
        - logistic regression model vs. Random Forest model
    - 2. What are the most important features for the appointment no-show?

Dataset

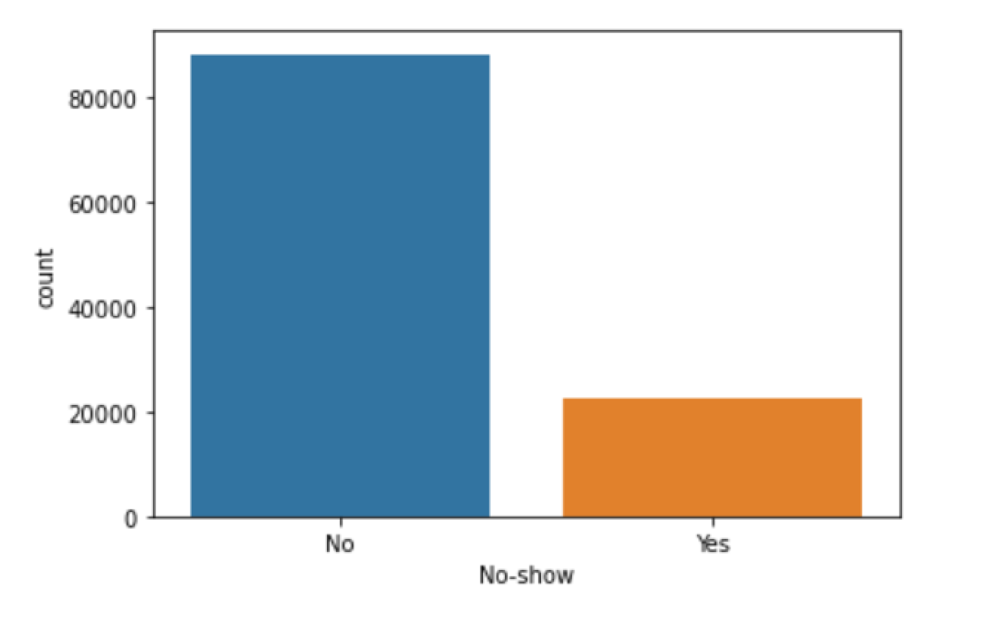
- The dataset contains 110,527 medical appointments and its 14 associated variables. The most important one is no-show variable, the target variable for models, to indicate if the patient show-up or no-show to the appointment.
- Features includes demographic variables(age, gender, neighborhood…), health status(diabetes, handicap…), schedule day, appointment day, and text message reminder

Preprocess

- The dataset is clean without any null value
- Create a “lead time for appointment” variable by subtracting schedule day from appointment day
- Create dummy variables for categorical variables in order to build models using Python

* Dataset file name is “out.csv”

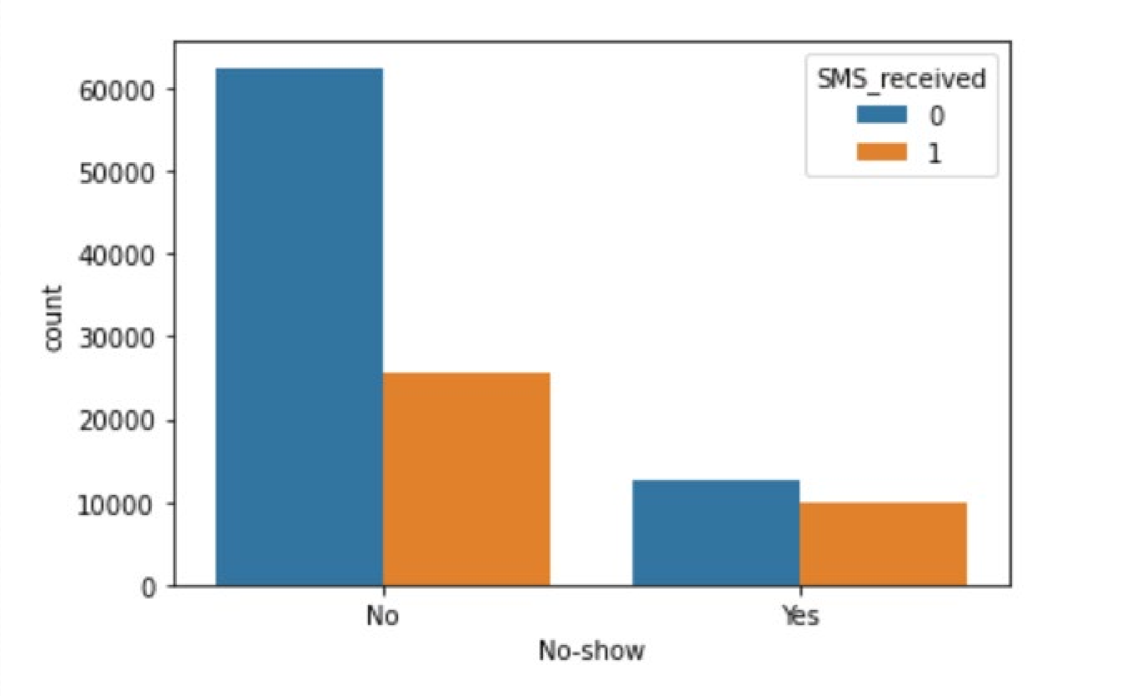
Exploratory Data Analysis
Imbalanced Dataset



* No-show label “Yes” means that the patient no-show to the appointment

- The distribution of 2 classes within the target variable, No-show, is unequal
- Thus, I use oversampling to deal with imbalanced issue and choose recall to evaluate the performance instead of accuracy

Exploratory Data Analysis
No-show across SMS_received



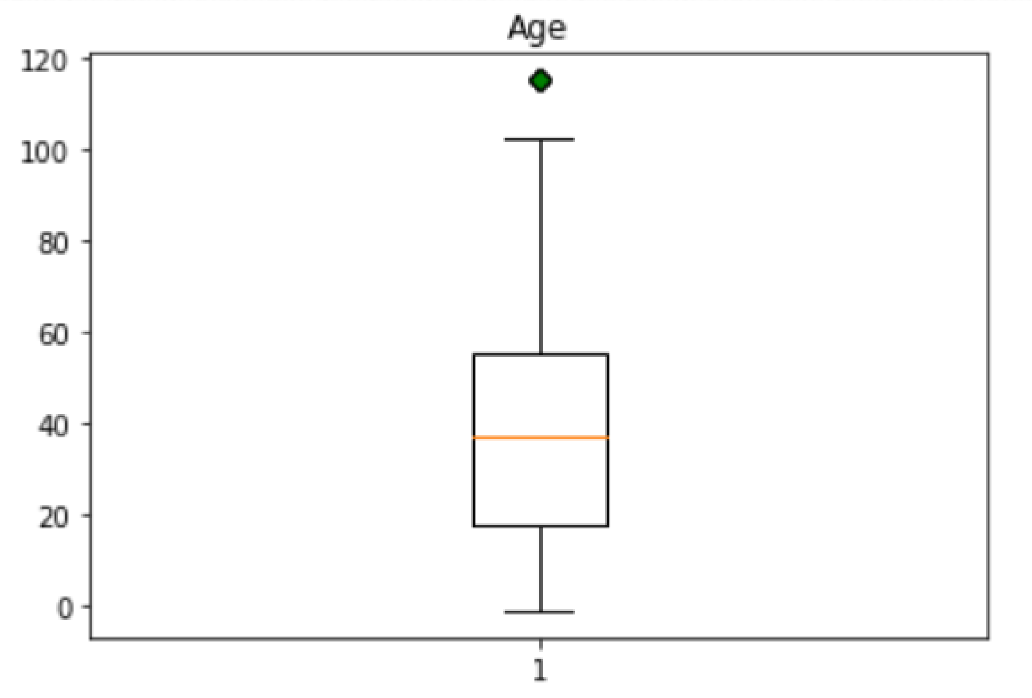
* SMS_received =1 means the patient get text message reminder

- It seems that there is a relationship between No-show and SMS_received because the bar charts differ across classes

p.s. bar plots for 2 categories have a strong visual impact than cross table to show the relationship

Exploratory Data Analysis

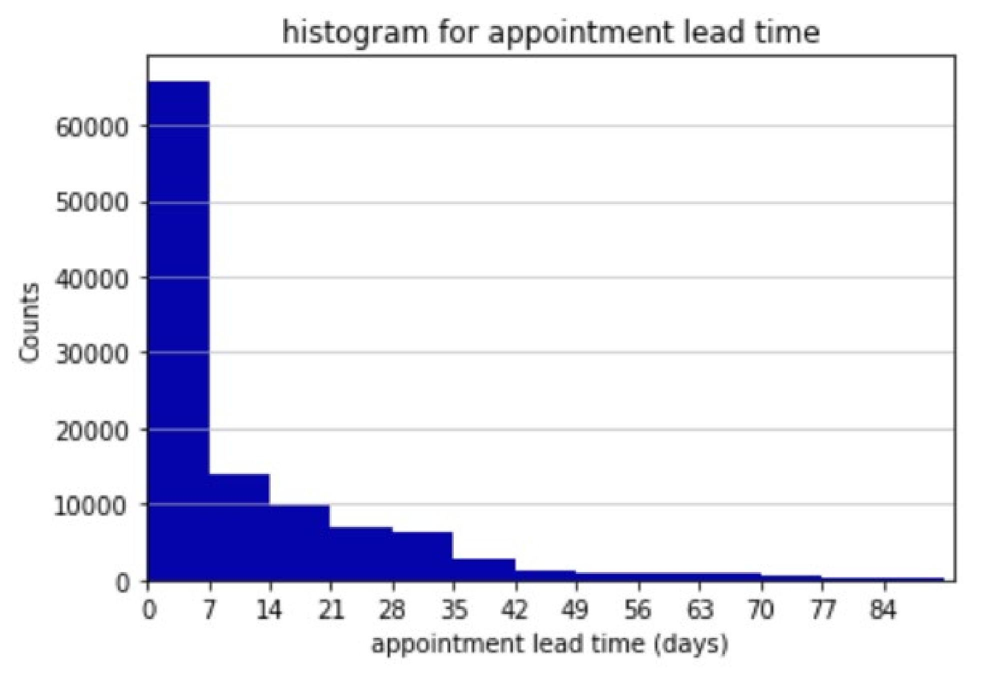
50% of the patients' age are between 20 and 50 years old



p.s. I have tried box plot and histogram to visualize the distribution of age, and find that box plot could present more information such as quantile

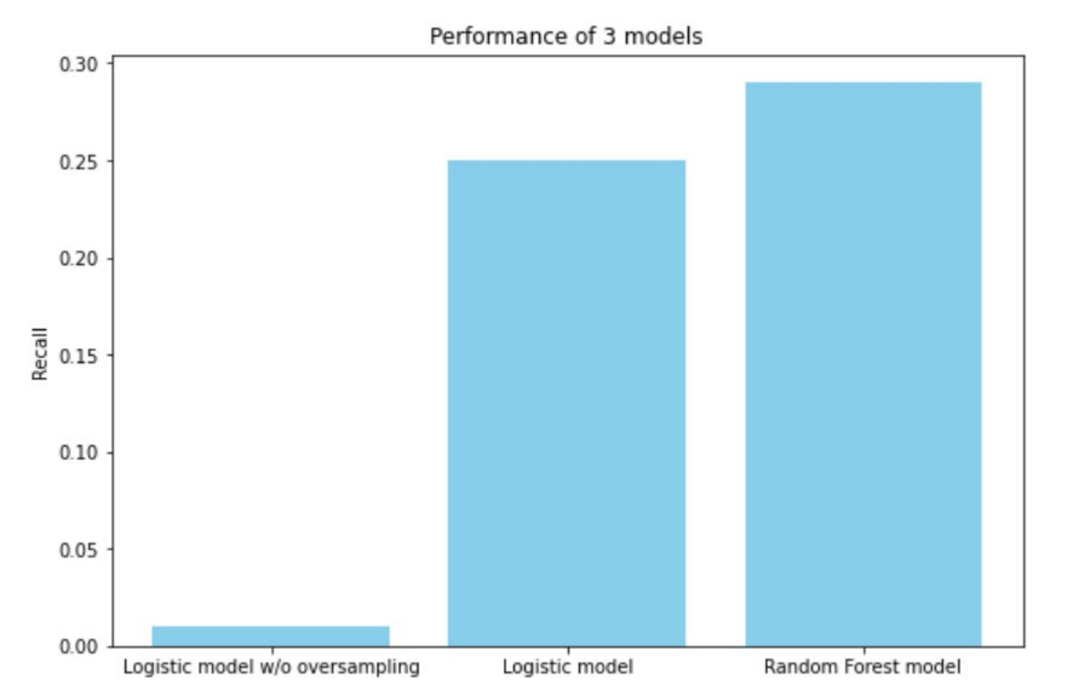
Exploratory Data Analysis

Most of the appointments are made within 1 week



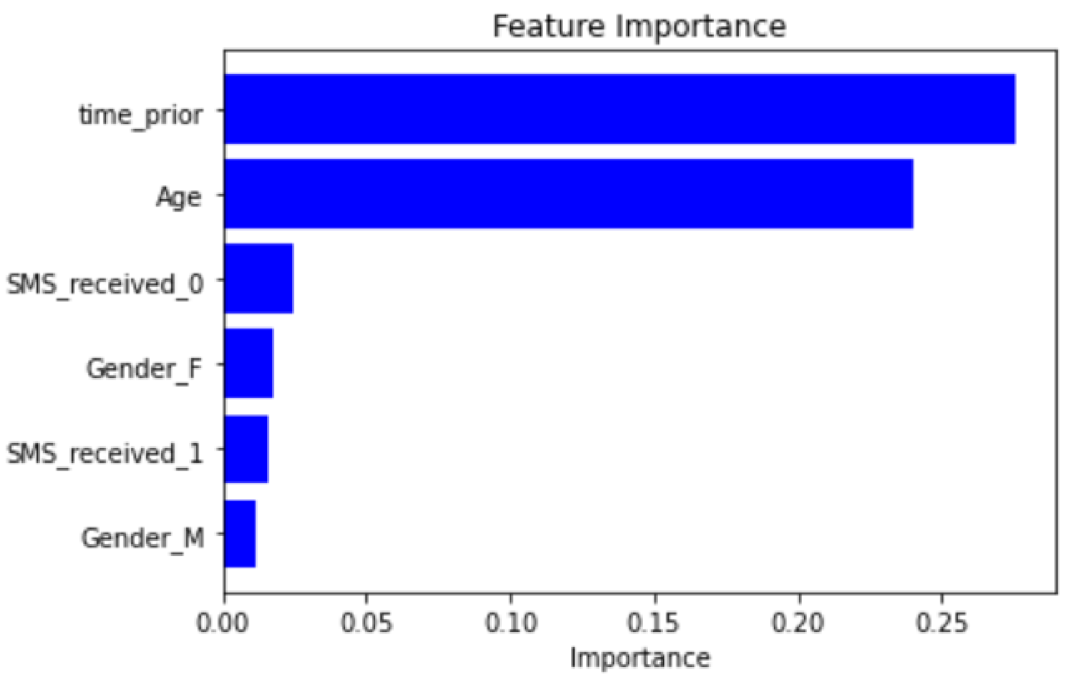
Findings

Compare Performance of 3 models



- The Random Forest model perform better than Logistic regression model
- Oversampling helps to reach a better recall

Findings
the most important features



* Feature importance is derived from Random Forest

- Time_prior (appointment for lead time) and age are important features to predict an appointment no-show

p.s. bar chart could show the difference clearly among features

Future works

- Use grid search to find the optimal hyperparameters of models
- Use different algorithms to get better performance

Appendix

- Full Python code and graphs:

https://github.com/shanrulin/Individual_project-Medical-Appointment-No-Shows/blob/master/FE%20550-Individual%20project.ipynb