1. Problem Statement

High no-show appointment rates have a negative impact on healthcare by disruption in daily operations and decreasing provider revenue. A no-show occurs when a patient fails to attend his/her appointments without prior notice or on short notice. As early as 1983, Bigby et al found that "about two thirds of the savings realized from reminders was generated in 23% of the patients whose prior predicted probability of a no-show appointment was above 20%" [1]. Therefore, it is crucial to develop a model to accurately predict the patient no-shows so that the clinic can overbook an appointment where a patient is predicted to be a no-show. The aim of this project is to use machine-learning to build a model that predicts no-shows for individual appointments, based on the data provided by Kaggle.

I will test the data with different models including Logistic Regression (LR), Linear Discrimination Analysis (LDA), Gaussian Naïve Bayes classifier (GNB), and Support Vector Machine (SVM). By comparing the cross-validation scores, we identify the most appropriate model to predict the no-show appointments. Also, by implementing the feature-selection models using lasso regression and Recursive Feature Elimination, I can identify which features play more significant roles in predicting the no-shows.

2. Methodology

In this study, various methods were used to identify the factors affecting the no-show hospital appointments. This section begins with an overview of the dataset, followed by data preparation, and a discussion on variable selection of the machine learning models. I use Lasso and Recursive Feature Elimination for feature selections, and the parameters in each learning model are selected by K-fold cross-validation. The most successful algorithm is Logistic Regression, which is defined as the one with highest cross validation score.

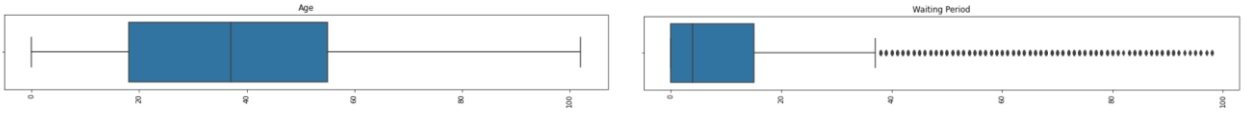
2.1 Data Source

The data source of this project is from Kaggle, a website offers datasets for machine learning, data visualization, exploratory analysis, and neural network projects. The dataset contains 110.527 medical appointments from April 2016 to June 2016 with its 14 associated variables (characteristics). The dataset is a combination of data about an appointment and patient demographic such as gender, age of the patients, social welfare benefit, or the chronic diseases they have, etc. The most important variable is whether the appointment occurs (no-show variable).

2.2 Data Cleaning

2.2.1 Missing Values and Outliers: The first step is locating and correcting missing values as well as outliers. I removed the data with negative age, and negative waiting period, which is the duration between the schedule day and the appointment day. In this dataset, most of the patients are between 18 and 55 years old. There is only one patient who is 115 years old, which is the outlier that should be removed from the dataset. I also remove the data with waiting

periods longer than 100 days. After removing the outliers, we have a total of 110,378 data. The boxplots of "Age" and "Waiting Periods" are shown as below:



2.2.2 Duplicate Values: The next step is to identify the duplicates in the dataset. After running the check, I figure out that there are no duplicates appearing in the dataset.

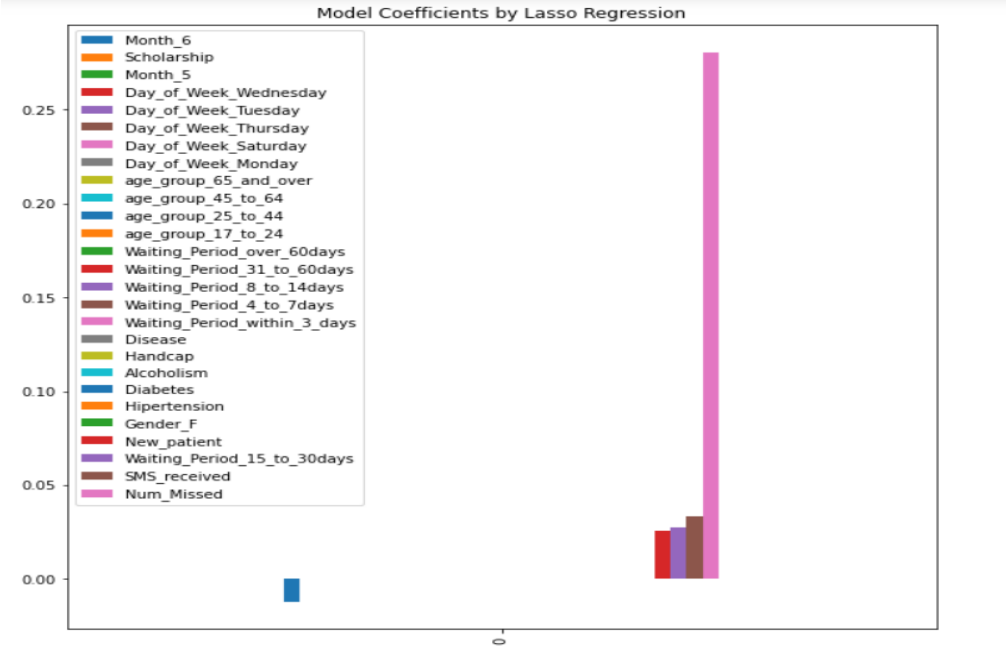
2.2.3 New Variables: I add a few more variables to the dataset such as time elapsed between booking date and appointment date (Waiting Periods), cumulative sum of past missed appointments (num_missed), new patient (whether the patient is new or existing patient), age group (age range of the patients).

2.2.4 Features without predictive power: Remove features with no predictive power, such as Patient ID, Appointment ID, Scheduled Day, Appointment Day, etc.

2.3 Feature Selection

2.3.1 Lasso Regression

Chart 1: Feature Selection by Lasso Regression



The top features from Lasso Regression: Time elapsed between booking date and appointment date (Waiting Periods), Cumulative sum of past missed appointments (num_missed), SMS received, new patient, chronic disease, age, scholarship, handicapped, Day of the Week, Month of the appointment date, hypertension.

2.3.2 Recursive Feature Elimination

Usually, RFE is popular because it is easy to configure and effective at selecting the relevant features to predict the target variable.

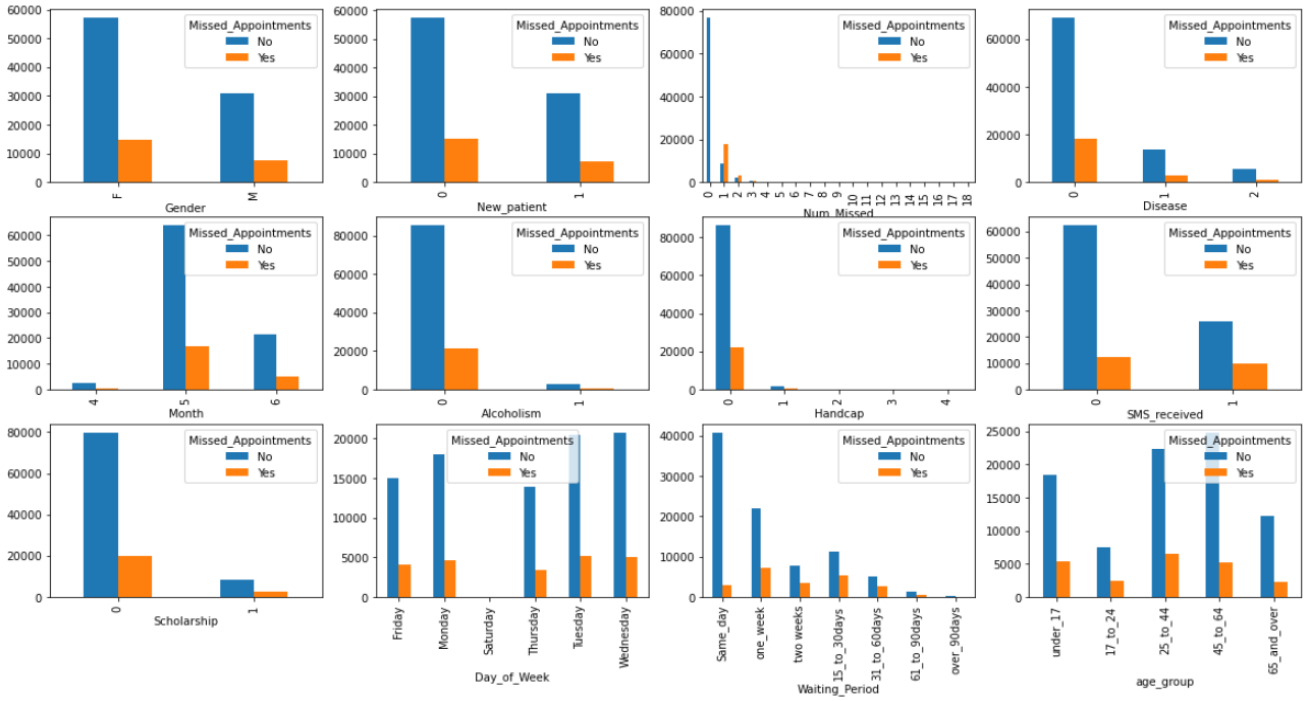
Table 1: Feature Selection by RFE

	Ranking	Variables
0	1	Month_5
1	1	Month_6
2	1	New_patient
3	1	Num_Missed
4	1	Waiting_Period_15_to_30days
5	1	Waiting_Period_31_to_60days
6	1	Waiting_Period_4_to_7days
7	1	Waiting_Period_8_to_14days
8	1	Waiting_Period_over_60days
9	1	Waiting_Period_within_3_days
10	1	age_group_25_to_44
11	1	age_group_45_to_64
12	1	age_group_65_and_over
13	2	Alcoholism
14	3	Handcap
15	4	Diabetes
16	5	SMS_received
17	6	Day_of_Week_Tuesday
18	7	Gender_F
19	8	Scholarship
20	9	Hipertension
21	10	Disease
22	11	age_group_17_to_24
23	12	Day_of_Week_Thursday
24	13	Day_of_Week_Wednesday
25	14	Day_of_Week_Saturday
26	15	Day_of_Week_Monday

From Table 1, the first fifteen variables with ranking from 1 to 4 are chosen to fit the model.

2.3.3 Data Visualization

The top features found in the below charts are: Time elapsed between booking date and appointment date (Waiting Periods), Cumulative sum of past missed appointments (num_missed), SMS received, new patient, chronic disease, age, scholarship, handicapped.



Among those features, SMS reminder is negatively correlated to the no-show appointments, which is a surprising result. However, we do not know how the SMS worked in Brazil; therefore, I decided not to include this variable in the models.

3. Evaluation and Final Results

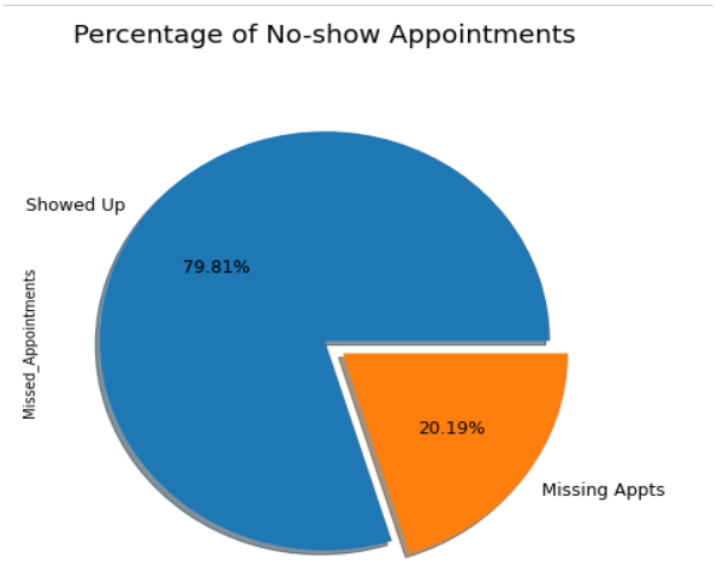
3.1 **Model Selections:** With 10-fold cross validation, we have the following results:

	Cross Validation Accuracy Score	
Model	Feature Selection from Lasso	Feature Selection from RFE
Logistic Regression	0.892 (+/- 0.006)	0.902 (+/- 0.004)
Linear Discrimination Analysis	0.839 (+/- 0.010)	0.876 (+/- 0.007)
Gaussian Naïve Bayes	0.807 (+/- 0.011)	0.824 (+/- 0.009)
Linear SVM	0.842 (+/- 0.014)	0.887 (+/- 0.006)

From the above results, Logistic Regression was selected with features selected from Recursive Feature Elimination. Logistic Regression is used when the dependent variable is categorical. The type of Logistic Regression in this project is Binary Logistic Regression.

3.2. **Prediction of No-show Hospital Appointments**

3.2.1 **Data Visualization**



The above chart shows that 20.19% of the appointments are no-show appointments.

3.2.2. **Machine Learning Model – Logistic Regression**

One of the factors affecting the no-show appointment is the duration between the booking date and appointment date. The patients have many reasons not to attend the appointments such as significant work or family commitments; therefore, it is hypothesized that the waiting period is positively associated with the missing appointments.

The second influential factor is the patient's missed appointment history, which is the cumulative sum of prior no-show appointments. If the patients used to miss the appointment previously, it is hypothesized that they will continue missing the appointments.

The third impactful factor is the age of the patient. When the patient is older, he/she is less likely to miss the hospital appointment, as their health are associated with more risks. Parsons et all also concluded in their research that patients less than 21 years old are more likely to miss the appointment[2]. Therefore, it is hypothesized that the age of the patient is negatively associated with the missing appointments.

The fourth factor might impact the no-show appointment is whether they are new patient or existing patient. Marbouh et al[3] mentioned in their research that the no-show appointment rate is higher among new patient visits.

According to "3.1 - Model Selections" part, we choose Logistic Regression to predict the no-show appointments, and use this algorithm for our dataset. We randomly split the data set into training and testing set by assigning 70% of data points to the former and the remaining 30% to the latter. We will train the Logistic Regression model on the training set and compute the test set accuracy result, as well as predict the no-show in real-time.

The logistic regression model is as the following, in which $\hat{r}$ is the show probability:

$$\text{Log}\left[\frac{\hat{r}}{1-\hat{r}}\right] = \text{Intercept} + \textit{(History of prior no-show)} + \textit{(New Patient)} + \textit{(Waiting Period Range)} + \textit{(Age Range)} + \textit{(Appointment month)} + \textit{(Handicap)} + \textit{(Diabetes)} + \textit{(Alcoholism)}$$

Logistic regression model to predict the patient show rate using the feature selection from RFE is shown in Table 2. The most significant predictor from Table 2 is the history of no-show appointments (i.e number of previous missing appointments).

Table 2: Logistic regression model to predict the patient show rate using the feature selection from RFE. Each level for the categorical variables is assigned the values of either 0 or 1. The coefficients represent the relative ratio to the reference level.

```
Optimization terminated successfully.
         Current function value: 0.267265
         Iterations 8
                            Results: Logit
================================================================================
Model:                 Logit                    Pseudo R-squared:  0.468
Dependent Variable:    Missed_Appointments_Yes  AIC:               59034.3031
Date:                  2022-08-01 11:35         BIC:               59197.7014
No. Observations:      110378                   Log-Likelihood:    -29500.
Df Model:              16                       LL-Null:           -55502.
Df Residuals:          110361                   LLR p-value:       0.0000
Converged:             1.0000                   Scale:             1.0000
No. Iterations:        8.0000
--------------------------------------------------------------------------------
                             Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
--------------------------------------------------------------------------------
const                       -4.2522   0.0682 -62.3516 0.0000 -4.3859 -4.1186
Num_Missed                   2.9623   0.0206 143.8362 0.0000  2.9219  3.0027
New_patient                  0.9993   0.0239  41.7419 0.0000  0.9524  1.0462
Waiting_Period_within_3_days 1.7491   0.0383  45.6513 0.0000  1.6740  1.8242
Waiting_Period_4_to_7days    1.8536   0.0376  49.2556 0.0000  1.7798  1.9274
Waiting_Period_8_to_14days   2.0098   0.0391  51.3660 0.0000  1.9331  2.0865
Waiting_Period_15_to_30days  2.2088   0.0358  61.6621 0.0000  2.1386  2.2790
Waiting_Period_31_to_60days  2.2850   0.0429  53.2462 0.0000  2.2009  2.3691
Waiting_Period_over_60days   2.2401   0.0746  30.0255 0.0000  2.0939  2.3863
age_group_25_to_44          -0.1564   0.0271  -5.7607 0.0000 -0.2096 -0.1032
age_group_45_to_64          -0.3982   0.0285 -13.9946 0.0000 -0.4540 -0.3424
age_group_65_and_over       -0.4462   0.0378 -11.8029 0.0000 -0.5202 -0.3721
Month_5                     -0.4514   0.0608  -7.4283 0.0000 -0.5705 -0.3323
Month_6                     -1.1847   0.0642 -18.4451 0.0000 -1.3106 -1.0588
Handcap                     -0.1363   0.0722  -1.8888 0.0589 -0.2778  0.0051
Diabetes                     0.1232   0.0442   2.7850 0.0054  0.0365  0.2099
Alcoholism                   0.1479   0.0654   2.2629 0.0236  0.0198  0.2761
================================================================================
```

The p-values for all variables are smaller than 0.05, which means that the alternative hypothesis is true. The results from Table 2 show that the no-show appointment is higher among patients with new patient visits, prior patient's missed appointments, time elapsed between the booking date and appointment date, and patients with diabetes and alcoholism. Meanwhile, the no-show rate is negatively associated with the age of the patients and handicapped patients.

3.2.3. **Use the Model to Make Predictions**

Once we've fit the logistic regression model by training dataset, we can calculate the accuracy on test set. Let $\in$ is the threshold of the show probability, p_i is the predicted show probability from the model for patient i, and Y_i be the predicted label of no-show (0) or show-up (1):

$$Y_i = \begin{cases} 0 & IF\ p_i < \in \\ 1 & IF\ p_i > \in \end{cases}$$

In this study, we choose threshold of 0.5 (default one) and 0.25 to compare:

Table 3: Test Set Accuracy with Threshold = 0.5

	precision	recall	f1-score	support
0	0.91	0.96	0.93	29060
1	0.80	0.61	0.69	7365
accuracy			0.89	36425
macro avg	0.85	0.78	0.81	36425
weighted avg	0.88	0.89	0.88	36425

Table 4: Test Set Accuracy with Threshold = 0.25

	precision	recall	f1-score	support
0	0.99	0.90	0.94	29060
1	0.72	0.96	0.82	7365
accuracy			0.91	36425
macro avg	0.85	0.93	0.88	36425
weighted avg	0.93	0.91	0.92	36425

Based on above classification report, we can see the precision and recall of each label:

- Precision is a measure of the accuracy provided that a class label has been predicted.
  Precision = TP / (TP + FP)

- Recall is the true positive rate.

  Recall = TP / (TP + FN)
- F1 score: the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

Overall, the accuracy scores of the test set with threshold of 0.5 and 0.25 are 89% and 91% respectively, which mean that our model is a good one. From the result, we decided to choose the threshold of 0.25 to make our model better.

We can then use this model to make predictions about whether a patient will show up at the appointment. With the function "predict_proba" to return estimates for all classes, the first column of the result is the probability of class 0 (i.e predicted no-show), and second column is probability of class 1 (predicted show-up) for each patient.

Assume patient A is scheduled as a return visit patient, and he has three prior no-show appointments. Other factors to consider in the model are the patient is 47 years old, and he called to schedule for a 4-day-later appointment. The patient does not have any chronic diseases such as diabetes or hypertension. He is not alcoholic or handicapped. From the model, it is predicted that he has 99.22% chance of not attending the appointment, and 0.77% chance of showing up at the appointment.

Assume patient B is scheduled as a return visit patient, and he has no prior record of no-show appointments. The patient is 23 years old, and he has no chronic diseases. He is not alcoholic or handicapped. His appointment is 18 days from the day he called to make the appointment. From the model, it is predicted that he has 91.20% chance of attending the appointment, and 8.8% chance of not showing up at the appointment.

It is very highly likely that patient A will not show up for this appointment; therefore, we may overbook this appointment as the scheduled appointment is predicted to be a no-show appointment. This overbooking could help to reduce the other patient's waiting period as well as reduce the cost that hospital incurred for no-show appointments.

4. Conclusions

In this study, we attempt to identify the key factors to predict the probability of no-show appointments using regular available hospital data. The most significant factors in this study are the history of previous missing appointment, the waiting periods, whether they are new or existing patients, or whether they are old or have chronic illnesses.

There are some limitations to this study as other important factors for no-show such as the distance from the patient's house to the hospital or whether they have insurance or not, are not included in this dataset. It will be more accurate if those factors are studied; therefore, further work should explore these factors. Another limitation of this study is that the dataset is only for three months from April to June 2016.