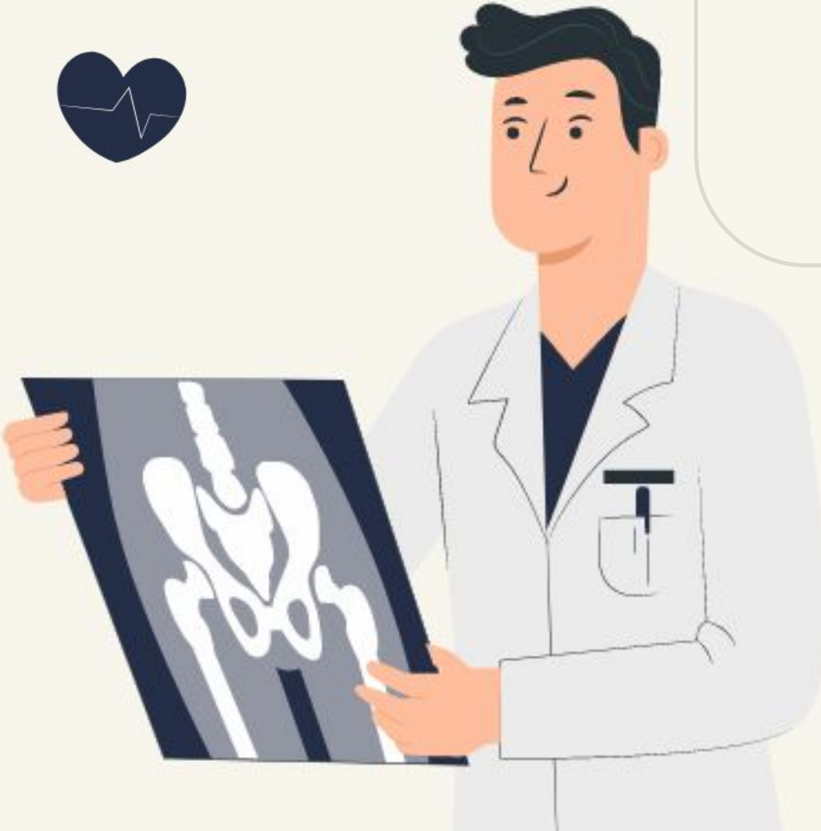
STRATEGIC CAPSTONE PROJECT

ADELPHI UNIVERSITY

“No Show to Appointments”

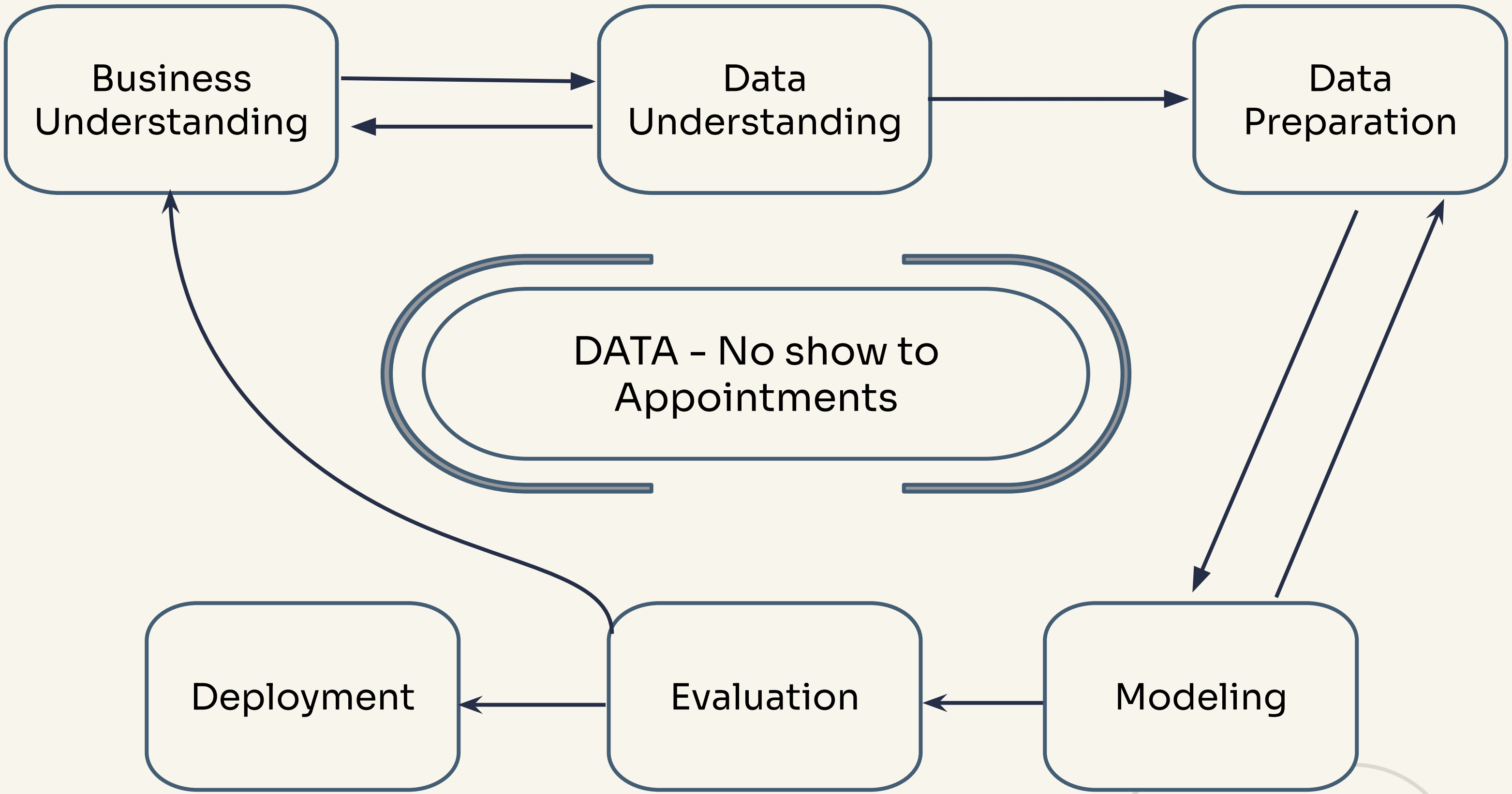


Group 1:-

1. Sandra Cordoba
2. Thanh Huyen Dau
3. Lauren Toland
4. Sai Pranay Tummala
5. Jeongseok Yu

Contents of the project

- Followed CRISP-DM Framework -



1. Business Understanding

<u>Objective:</u> *Clearly define the business problem, set project goals, and identify key success criteria*

PROBLEM STATEMENT

What is the issue?

High no-show rates in medical appointments are a critical challenge in the healthcare industry.
They disrupt operational efficiency, waste valuable resources, increase patient wait times, and lead to financial losses.

Why does it matter?

By analyzing and predicting no-show appointments, healthcare providers can optimize scheduling, reduce waste, and minimize financial losses.
Ultimately, this will improve overall system efficiency and enhance patient care.

BUSINESS UNDERSTANDING FRAMEWORK

ADELPHI UNIVERSITY

Type of problem ?

Classification Problem

Technical goal

Build an efficient and interpretable model to predict patient no-shows

Computing and data storage needs

1. Dataset size (72,607 records) can be handled by a standard PC (8 GB RAM, multi-core CPU) for simpler models.
2. For more complex models like XGBoost, a cloud-based or server environment is preferred to ensure faster processing and scalability.

Success criteria

Evaluation: Sensitivity, Specificity, precision, G-mean, Accuracy, AUC
Subjective criteria: The model should be interpretable, actionable. It should provide actionable insights

2. Data Understanding

Objective: *Explore the dataset to understand its structure, quality, and key patterns for effective data preparation.*

Data Overview

- **Dataset size**: 72,607 records, 18 variables
- **Source:** Medical appointment data from Brazilian healthcare system
- **Target variable:** Show-up (Yes/No)
- **Features include:**
  - Patient demographics (Age, Gender)
  - Health conditions (Diabetes, Hypertension, Alcoholism, Handicap)
  - Appointment-related details (Scheduled Day, Appointment Day, SMS received, Waiting time, Time between Appointments, Prior no-shows)

Key Variables & Patterns

Age

- 0 to 115 years
- Most common: 0 years
- 75% of patients ≤ 55 years

Gender

- 66.3% Female- 33.7% Male

Month

- Data only from Apr to Jun.
- May has highest volume

Scheduled Day & Appointment Day

- Most scheduled and Appointment days on Tuesday, Wednesday
- Very few on Saturday

SMS Received

- 31.4% received SMS
- No-show rate higher with SMS

Waiting Time

- Avg: 9.25 min
- Same-day appointments (0 minutes) are very common
-Data error: Some records have -6 minutes

Time Between Appointments

- Avg: 5.6 days
- Longer gap → Higher no-show risk

Prior No-show

- Avg: 13.5%
- Strongest predictor of future no-shows

Data Quality & Issues

(i). No missing values

(ii). Outliers & Data Errors

- **Waiting time:** -6 minutes detected (invalid value)
- **Age:** Values from 0 to 115 years (suspicious outliers)

(iii). Imbalanced Variables

- **Handicap:** 97.9% have value 0 (severe imbalance)
- **SMS Received:** 68.6% did not receive SMS reminders

(iv). Imbalanced Target Variable

- **Show-up:** 80.2%
- **No-show:** 19.8%

3. Data Preparation

Objective: *Prepare a clean, structured dataset suitable for modeling*

Data Preparation

Initial Checks

- No missing (null) values
- Converted data types and formatted columns

Encoding Categorical Variables

- Day Variables:
    - Manual ordinal encoding (Mon–Sat → 0–6)
    - One-hot encoding chosen (better for models, avoids implied order)
- Gender & Show-up: Encoded manually (F/M → 0/1, No/Yes → 0/1)
- Dropped redundant columns (e.g., ScheduledDay, Show.up)

Identify and address near-zero variance variables

- Flagged using:
    - Frequency Ratio > 20
    - % Unique < 10%
- Variables:
    - Alcoholism: Kept (some predictive value)
    - Handicap: Binarized (0 = none, 1 = any level)

Outliers & Noise Handling

ADELPHI UNIVERSITY

Outlier Detection:

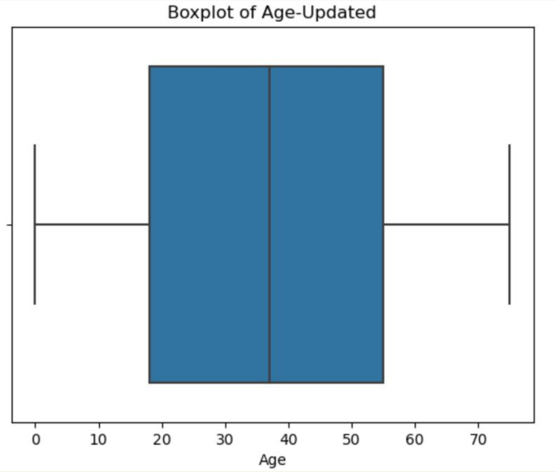
- Used IQR Method: Outliers = Values < Q1 – 1.5×IQR or > Q3 + 1.5×IQR
- Key variables: Age, Waiting_time, Time_b_appointment, Calling time, Prior_noshow

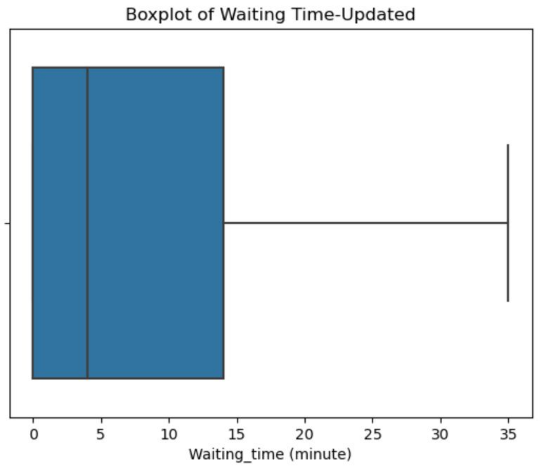
Outlier Handling Method: Capping:

- Values above the 95th percentile → set to the 95th percentile value
- Values below the 5th percentile → set to the 5th percentile value

Skewness Reduction

Applied log transformation post-capping to improve distributional characteristics.





4. Modeling

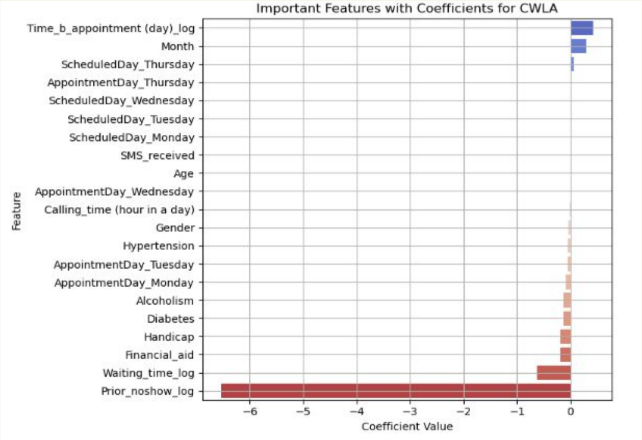
<u>Objective:</u> *Balance the data, extract important features and apply different models*

Modeling Part 1.

The division used to split our data was 70/30

Class imbalance techniques

- SMOTE
- Class-Weight

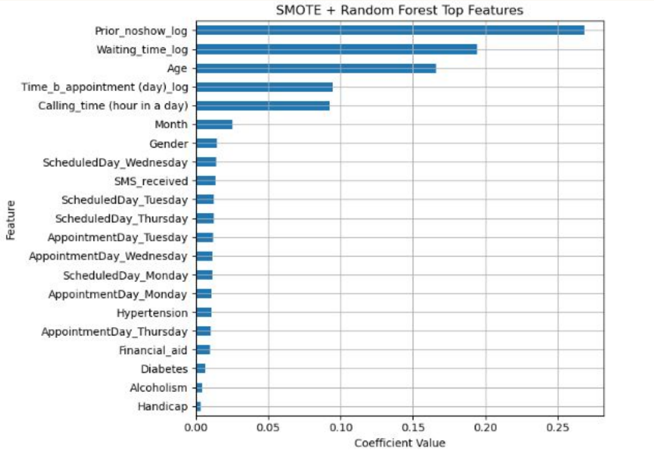


Feature Selection techniques

- LASSO
- Random Forest

Combinations:

- SMOTE + LASSO
- SMOTE + Random Forest
- Class-weight + LASSO
- Class-weight + Random Forest



R.F selects features based on importance scores from tree splits, while LASSO selects by shrinking less useful coefficients to zero.

Modeling Part 2.

3 Analytical Models Selected:

- XGBoost
- Ridge Regression
- Naive-Bayes

4 Class Imbalance + feature selection combinations:

- SMOTE+LASSO
- SMOTE+Random Forest
- Class-weight+LASSO
- Class-weight +Random Forest

Top 3 original metrics:

XGBoost
Class weight+LASSO

```
Evaluation Metrics:
Sensitivity (Recall): 0.7553
Specificity: 0.7467
Precision: 0.9185
G-Mean: 0.7510
Accuracy: 0.7535
AUC: 0.8401
```

Ridge Regression
SMOTE+LASSO

```
Evaluation Metrics:
Sensitivity (Recall): 0.7930
Specificity: 0.6443
Precision: 0.8938
G-Mean: 0.7148
Accuracy: 0.7619
AUC: 0.8059
```

Naive-Bayes
SMOTE+LASSO

```
Sensitivity: 0.8031121175172734
Specificity: 0.6160087719298246
Precision: 0.8876339600847077
G-Mean: 0.703366269616177
Accuracy: 0.763944360280953
AUC: 0.7847567856018727
```

5. Evaluation

Objective: *To Evaluate and choose the best model for Predicting patient* ***no-shows*** *in order to reduce missed appointments and optimize healthcare operations.*

-> **Model of Choice :-** XGBoost + Class Weight + LASSO

-> **Feature Selection :-** Using LASSO's selected non - zero variables.

-> **Tuning the Threshold :-**

- Adjusted classification threshold from 0 to 1
- Optimal threshold found at 0.6
- Improved specificity from 0.75 to 0.82, enhancing no-show prediction (Business Understanding Goal)

-> **Outcomes :-**

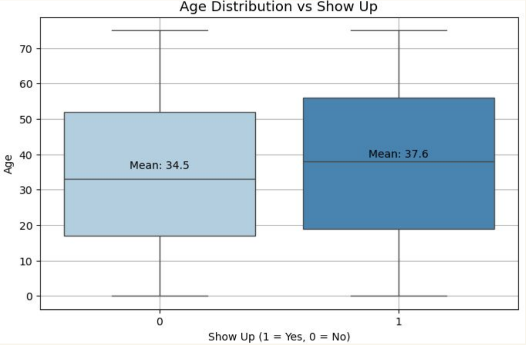
- More accurate no-show predictions (**Specificity** of 82%), leading to fewer unnecessary reminders.
- Supports targeted intervention (based on **G-mean:** 75%), which will help staff focus on high-risk patients (basically the model has a good balance between classes and is not favoring none of them).
- Based on the **Accuracy** (72%) and **AUC** (84%) → which means that the model has a high capacity of differentiate the two classes.
- Model ready for deployment and monitoring.

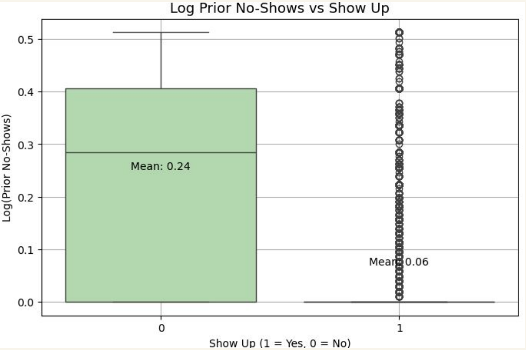
Original model

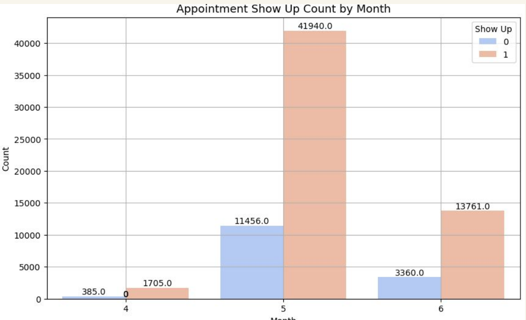
```
🔍 Confusion Matrix:
[[ 3405  1155]
 [ 4214 13009]]

📊 Evaluation Metrics:
Sensitivity (Recall): 0.7553
Specificity: 0.7467
Precision: 0.9185
G-Mean: 0.7510
Accuracy: 0.7535
AUC: 0.8401
```

Updated model

```
Confusion Matrix:
[[ 3749   811]
 [ 5341 11882]]

Evaluation Metrics:
Threshold used: 0.6
Sensitivity (Recall): 0.6899
Specificity: 0.8221
Precision: 0.9361
G-Mean: 0.7531
Accuracy: 0.7176
AUC: 0.8420
```



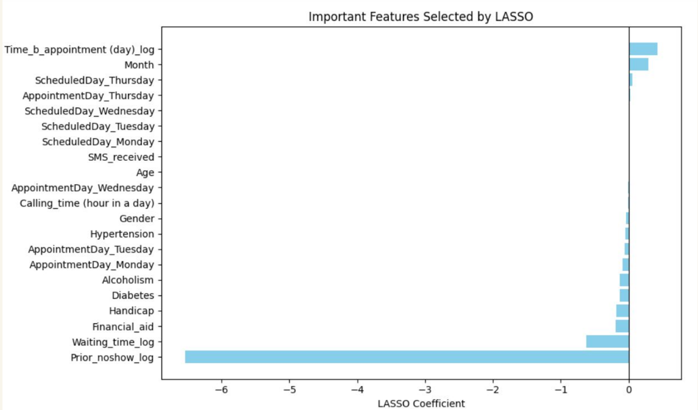


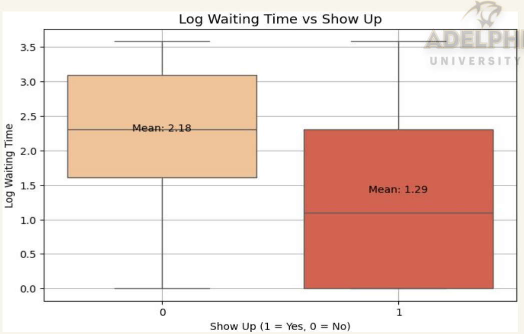


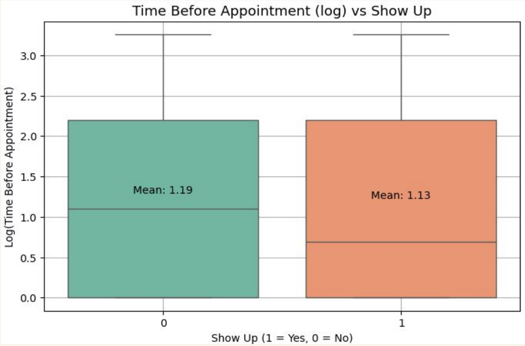
Visualizations of/for the main features vs Target Variable, based on Exploratory Analysis.

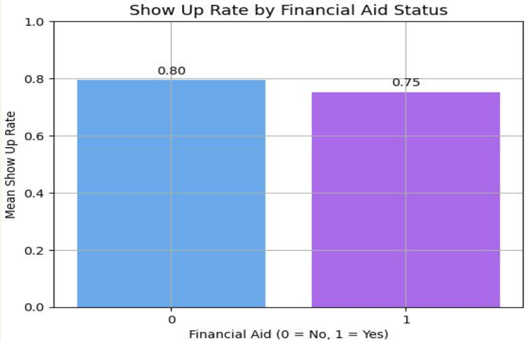
Based on the historical, demographic, and behavioral data, the most influenced features extracted from LASSO technique are :-

- Age
- Month
- Financial_aid
- Waiting_time_log
- Time_b_appointment (day)_log
- Prior_noshow_log
- Calling_time (hour in a day)









6. Deployment

<u>Objective:</u> *Propose actionable recommendations for data-driven decision making*

Recommendations

- **Recommendation 1: Target Outreach for High-Risk Patients**
  - Insight: ‘Prior_no_show_log’ has a high negative coefficient
  - Action: Implement targeted outreach program for patients with a history of no-shows, including personalized reminders and follow-up calls
- **Recommendation 2: Reduce Waiting Times**
  - Insight: ‘Waiting_time_log’ has a significant negative coefficient
  - Action: Optimize scheduling practices by adjusting appointment slots based on peak hours and ensuring adequate staffing
- **Recommendation 3: Send Reminders Closer to Appointment Date**
  - Insight: ‘Time_b_appointment(day)_log’ has a positive coefficient
  - Action: sending reminders (via SMS, calls, or email) one day before the appointment can offset forgetfulness or lack of commitment

Recommendations

ADELPHI UNIVERSITY

- **Recommendation 4: Seasonal Engagement Campaigns**
    - Insight: ‘Month’ has a positive coefficient
    - Action: Develop campaigns that promote importance of attending appointments throughout the year, possibly aligning with health awareness events or seasonal health tips
- **Recommendation 5: Implement a Flexible Rescheduling Policy**
    - Insight: Patients might be deterred from attending if they feel that they cannot easily reschedule
    - Action: Create a policy that allows for changes without penalties, encouraging patients to communicate rather than simply not show up
- **Recommendation 6: Provide Transportation Assistance**
    - Insight: Patients may face barriers related to transportation
    - Action: Offer partnerships with local transport services to help patients reach their appointments, especially for those with prior no-show histories

Recommendations

- **Recommendation 7: Conduct Patient Surveys**
  - Insight: Understanding the reasons behind the no shows can provide valuable insights
  - Action: After a no show, reach out to patients to gather feedback for their reason, use this data to refine scheduling
- **Recommendation 8: Monitor and Adjust Based on Data Analytics**
  - Insight: Continuous analysis of no show data can reveal trends and patterns we haven’t detected yet
  - Action: Regularly review appointment data and adjust strategies based on factors most strongly associated with no shows. Be proactive not reactive!

THANK YOU FOR LISTENING!

Presented by "No show to Appointments" Team