No show to **Appointment**

Agenda

BUSINESS UNDERSTANDING

METHOD

RESULTS

DISCUSSION

CONCLUSION

Problem Description

- The no-show appointment problem refers to bail on appointments and forgetting scheduled appointments.
- Economic and efficiency problems in the healthcare sector.

Background - Business Understanding

- Previous analytical works have shown a positive correlation between consultation length and patient no-shows.
- Identified age, gender, appointment type, and the number of previous appointments as significant contributors to appointment no-shows.
- Simulation models analyze overbooking and double-booking can reduce negative impacts on the healthcare industry.

Dataset Analysis:

- Our dataset consists of 16 variables:

  Age, gender, scheduled day, appointment day, month, calling time, waiting time, financial aid, hypertension, diabetes, alcoholism, handicap, appointment reminder, the time between appointments, prior no-show, and show-up.
- Each variable was analyzed and transformed using methods such as factoring, labeling, and other techniques.

Research Question

1. Can we predict the Show-up variable using data analytical techniques?
2. What are the important features in order to predict the Show-up variable?
3. What is the importance of the features?
4. How does the presence of outliers change the prediction performance?
5. How do the different feature selection methods affect the prediction of the Show-up variable?
6. What strategy outline will we propose to increase the performance of the No-Show to appointments?

METHOD

Phase 2:
Data Understanding

- Used Tableau to visualize all variables in the dataset.
- Tableau provided valuable insights into the characteristics and distributions of each variable.
- Statistical reports generated for every variable helped us gain a thorough understanding of their basic properties.
- Developed a clear idea of the necessary actions for subsequent stages of the data processing pipeline.

•Analyze the data collected.
- Visualize all variables of the dataset
- Bar Chart
- Pie Chart
- Box Plot
- Apply general statistical methods.
- Detect Null Values

METHOD

**Phase 3:
Data Preparation**

- Imported the dataset
- Removed unimportant columns Patient ID and "Appointment ID.“
- Identified zero-variance variables Alcoholism and Handicap.
- Encoded categorical variables (Gender, Scheduled Day, and Appointment Day) using label encoding.
- Detected outliers in variables such as "Age", "Calling_time..hour.in.a.day.", "Waiting_time minute.", "Time_b_appointment..day.", and "Prior_noshow" through the binning method.
- Transformed the value into numerical

- Delete no important data:
  - Patient ID
  - Appointment ID
- Check Zero variance
  - Alcoholism
  - Handicap
- Encoding (Label Encoding)
  - Gender
  - Scheduled Day
  - Appointment Day
- Outliers Detection
  - Binning
    - Age
    - Waiting Time Minute
    - Time b Appointment
    - Prior No Show
- Transforms values into numerical

METHOD

Phase 4: Modeling

- Used balancing techniques (SMOTE and ADASYN) to address dataset imbalance.
- Used Lasso regression and Pearson correlation for feature selection.
- Identified the most relevant and significant variables for show-up prediction.
- Tested various classification algorithms (Logistic Regression, Naïve Bayes, Random Forest) to build predictive models.
- Techniques and models
- Creation of an effective predictive model for show-up output.

- Training and Testing Set (70% - 30%)
- Balancing Method:
  - SMOTE
  - ADASYN
- Feature Selection:
  - LASSO
  - PEARSON's CORRELATION
- Models:
  - LOGISTIC REGRESSION
  - NAIVE BAYES
  - RANDOM FOREST

METHODOLOGY FRAMEWORK

Phase 1: Business Understanding

- Previous Analytical Works

Analyze the relationship between previous reports

- Finds out the best approach for the project
- Knowing more about the industry

Phase 2: Data Understanding

•Analyze the data collected.

- Visualize all variables of the dataset
- Bar Chart
- Pie Chart
- Box Plot
- Apply general statistical methods.
- Detect Null Values

Phase 3: Data Preparation

- Delete no important data:
  - Patient ID
  - Appointment ID
- Check Zero variance
  - Alcoholism
  - Handicap
- Encoding (Label Encoding)
  - Gender
  - Scheduled Day
  - Appointment Day
- Outliers Detection
  - Binning
    - Age
    - Waiting Time Minute
    - Time b Appointment
    - Prior No Show
- Transforms values into numerical

Phase 4: Modeling

- Training and Testing Set (70% - 30%)
- Balancing Method:
  - SMOTE
  - ADASYN
- Feature Selection:
  - LASSO
  - PEARSON's CORRELATION
- Models:
  - LOGISTIC REGRESSION
  - NAIVE BAYES
  - RANDOM FOREST

Phase5: Evaluation

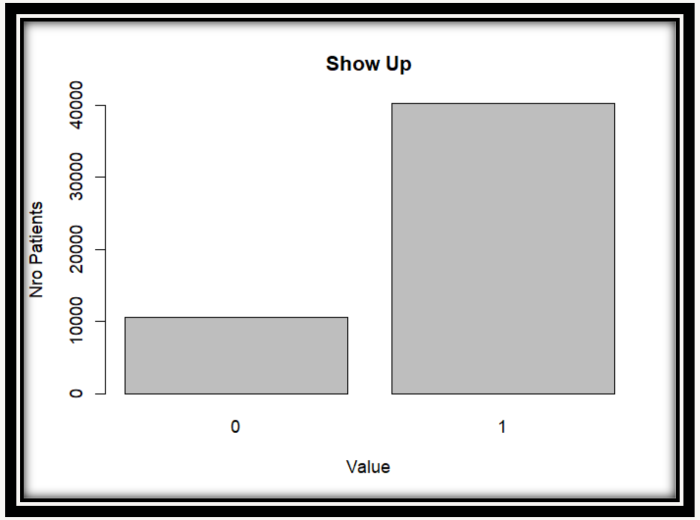
- Metrics:
  - Sensitivity
  - Specificity
  - Precision
  - G-Mean
  - Accuracy
  - AUC
- Select the best model

Phase 6: Deployment

- Final Report
- Discuss Best Models
- Review Model Evaluation

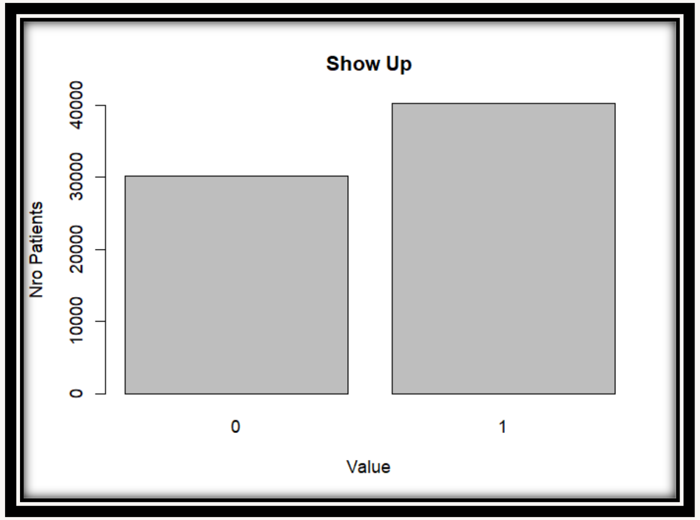
RESULTS

No Balancing



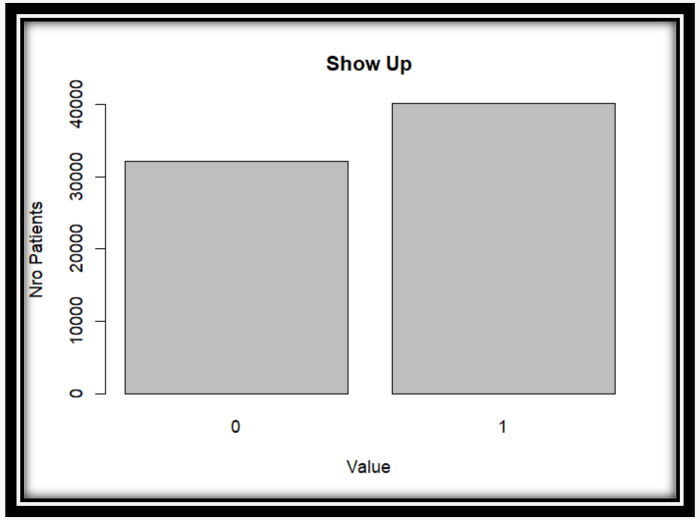
0 (NO)	10611
1 (SI)	40214

SMOTE



0 (NO)	30161
1 (SI)	40214

ADASYN

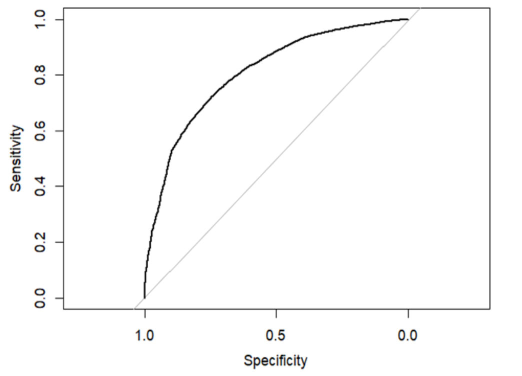


0 (NO)	32127
1 (SI)	40158

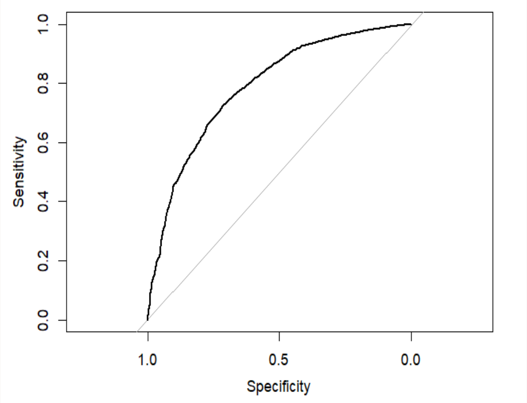
RESULTS

	Scenario	Model	Encoding approach	Imputation	Outlier handling method	Balancing approach	Feature selection	Sensitivity	Specificity	Precision	Gmean	Accuracy	AUC
1	1	LR	Label Encoding	NA	Binning	SMOTE	Lasso	0.647	0.798	0.457	0.719	0.767	0.807
	2	NB	Label Encoding	NA	Binning	SMOTE	Lasso	0.726	0.613	0.329	0.667	0.636	0.669
3	3	RF	Label Encoding	NA	Binning	SMOTE	Lasso	0.528	0.894	0.566	0.687	0.818	0.711
	4	LR	Label Encoding	NA	Binning	SMOTE	Pearson	0.615	0.795	0.440	0.670	0.758	0.788
	5	NB	Label Encoding	NA	Binning	SMOTE	Pearson	0.715	0.628	0.334	0.670	0.646	0.671
	6	RF	Label Encoding	NA	Binning	SMOTE	Pearson	0.507	0.900	0.572	0.676	0.819	0.704
2	7	LR	Label Encoding	NA	Binning	ADASYN	Pearson	0.654	0.773	0.431	0.711	0.748	0.787
	8	NB	Label Encoding	NA	Binning	ADASYN	Pearson	0.738	0.592	0.322	0.661	0.622	0.665
	9	RF	Label Encoding	NA	Binning	ADASYN	Pearson	0.531	0.883	0.544	0.685	0.810	0.707

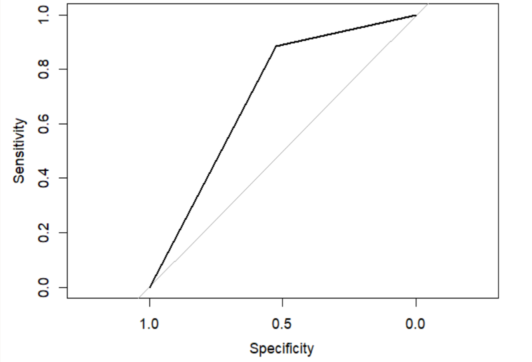
SCENARIO 1



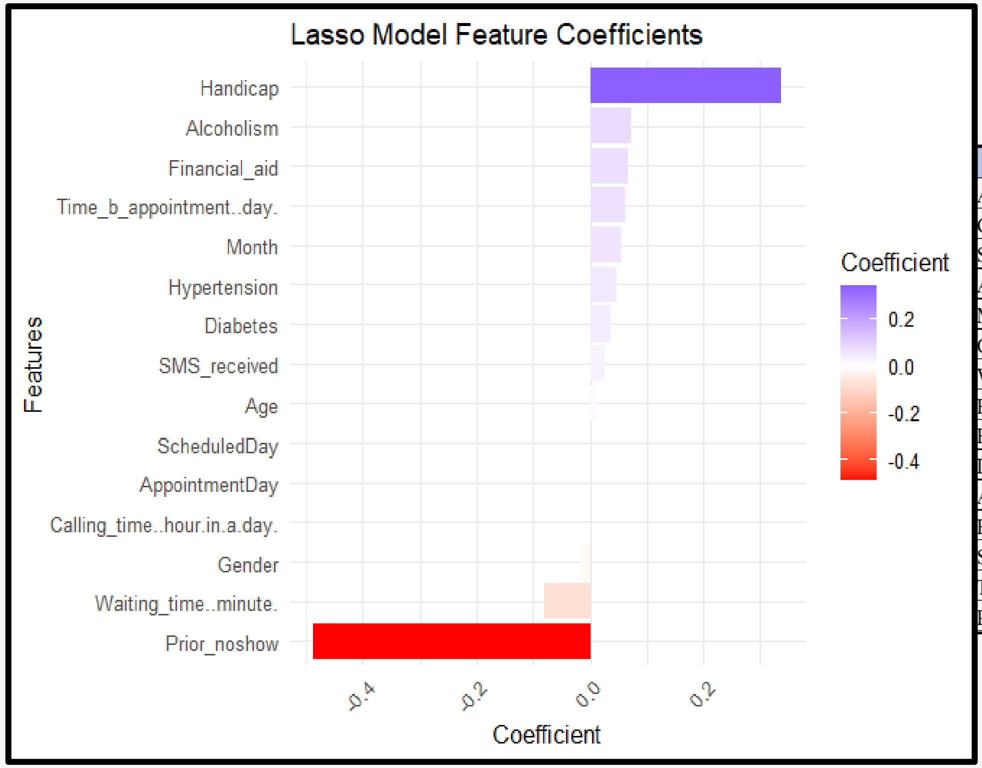
SCENARIO 7



SCENARIO 3



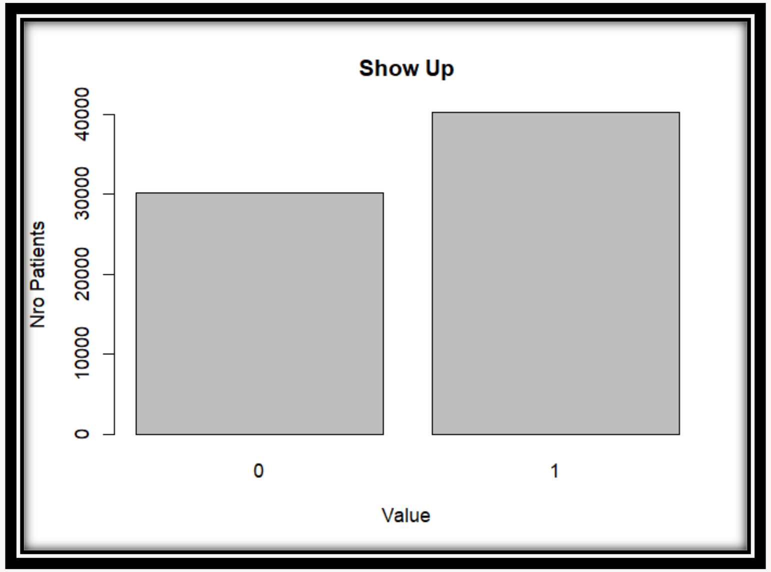
RESULTS



Feature	Coefficients
Age	0.011367586
Gender	-0.015367918
ScheduledDay	0.001028447
AppointmentDay	0.000657967
Month	0.055739646
Calling_time..hour.in.a.day.	0.000322619
Waiting_time..minute.	-0.079499915
Financial_aid	0.066590356
Hypertension	0.046239667
Diabetes	0.03796823
Alcoholism	0.072562827
Handicap	0.338156804
SMS_received	0.026935681
Time_b_appointment..day.	0.062889999
Prior_noshow	-0.487907185

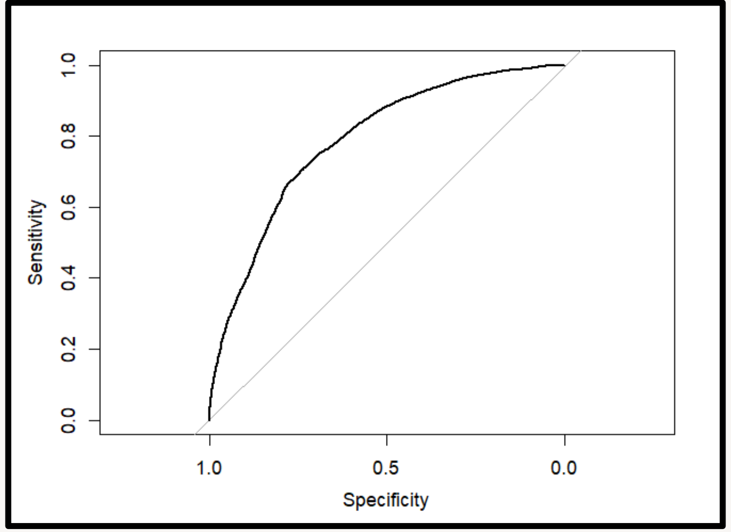
- **Prior_noshow** (-0.4879071846) – Negative Correlation: prior history of not showing up for appointments is associated with a higher likelihood of not attending future appointments
- **Handicap** (0.3381568039) - Positive Correlation: the presence of a handicap is associated with an increased probability of attending an appointment
- **Waiting_time..minute.**(-0.0794999147), **Alcoholism** (0.0725628274), and **Financial_aid** (0.0665903564), also have a notable relationship with Show Up.

RESULTS



0 (NO)	30147
1 (SI)	40196

Metrics	Outliers	No Outliers
Sensitivity	0.543	0.647
Specificity	0.854	0.798
Precision	0.501	0.457
G-Mean	0.681	0.719
Accuracy	0.788	0.767
AUC	0.787	0.807



- The presence of outliers significantly impacts the model's performance.
- The model without outliers is better at identifying true positives.
- The model with outliers performs slightly better in identifying true negatives
- The model without outliers provides a better balance between sensitivity and specificity, making it more suitable for handling imbalanced datasets
- Removing outliers leads to a more balanced and better-performing model

DISCUSSION AND CONCLUSION

Final Model Selection – Scenario 1

- Scenario 1

- Highest G-mean value- .719
- Parameters
- Smote data balancing
- Lasso Feature Selection
- Logistic Regression Model

DISCUSSION AND CONCLUSION

Most influential features

- Prior no show
- Handicap
- Waiting Time in Minutes
- Alcoholism
- Financial Aid

DISCUSSION AND CONCLUSION

Outlier evaluation

- Excluding outliers led to a more efficient model
- Better at dealing with an imbalanced dataset

DISCUSSION AND CONCLUSION

Future Improvements

- Further Feature Engineering
- Finding a Versatile Outlier Treatment Strategy
- Exploring more Data Balancing Techniques
- Tuning HyperParameters
- Consider Exploring More complex models (GBM's, Neural Networks, etc)

Thank you

Yulimar Rivero