Predicting Medical Appointment No-Shows Using Machine Learning

A Healthcare Operations Analytics Case Study

Pablo Wills

EXECUTIVE SUMMARY

Healthcare providers lose operational efficiency, revenue, and patient access when individuals fail to attend scheduled appointments. This project analyzes **110,527 real medical appointments** and builds a machine learning model to predict the likelihood that a patient will miss their visit.

Key outcomes:

- Built an end-to-end ML pipeline (cleaning → feature engineering → modeling → evaluation)
- Compared Logistic Regression, Random Forest, and XGBoost
- Identified actionable drivers of no-shows
- Translated insights into **operational recommendations**

Best performing model: Random Forest
AUC: ~0.73
Accuracy: ~0.79

This model provides meaningful predictive power that can directly support scheduling optimization and patient outreach.

DATASET OVERVIEW

Dataset size: 110,527 appointments
Target variable: no_show_flag (1 = missed, 0 = attended)

Primary features:

- **Demographics:** age, gender
- **Clinical indicators:** hypertension, diabetes, alcoholism, handicap
- **Scheduling fields:** scheduled date, appointment date
- **Behavioral signals:** SMS received
- **Geographical:** neighborhood

FEATURE ENGINEERING

To strengthen predictive power, several new variables were engineered:

- **days_wait**: days between scheduling and appointment
- **appt_weekday**: weekday of appointment
- **appt_hour**: appointment hour (morning, mid-day, afternoon patterns)
- **age_bucket**: child → youth → young adult → adult → senior
- Removal of patient ID and raw timestamps

These transformations improved signal clarity and modeling quality.

nder	age	neighbourhood	scholarship	hipertension	diabetes	alcoholism	handcap	sms_received	no_show_flag	days_wait	appt_weekday	appt_hour	age_bucket
F	62	JARDIM DA PENHA	0	1	0	0	0	0	0	-1	4	0	senior
M	56	JARDIM DA PENHA	0	0	0	0	0	0	0	-1	4	0	adult
F	62	MATA DA PRAIA	0	0	0	0	0	0	0	-1	4	0	senior
F	8	PONTAL DE CAMBURI	0	0	0	0	0	0	0	-1	4	0	child
F	56	JARDIM DA PENHA	0	1	1	0	0	0	0	-1	4	0	adult
...	...	...	...	...	...	...	...	...	...	...	...	...	...
F	56	MARIA ORTIZ	0	0	0	0	0	1	0	34	1	0	adult
F	51	MARIA ORTIZ	0	0	0	0	0	1	0	34	1	0	adult
F	21	MARIA ORTIZ	0	0	0	0	0	1	0	40	1	0	youth
F	38	MARIA ORTIZ	0	0	0	0	0	1	0	40	1	0	young_adult
F	54	MARIA ORTIZ	0	0	0	0	0	1	0	40	1	0	adult

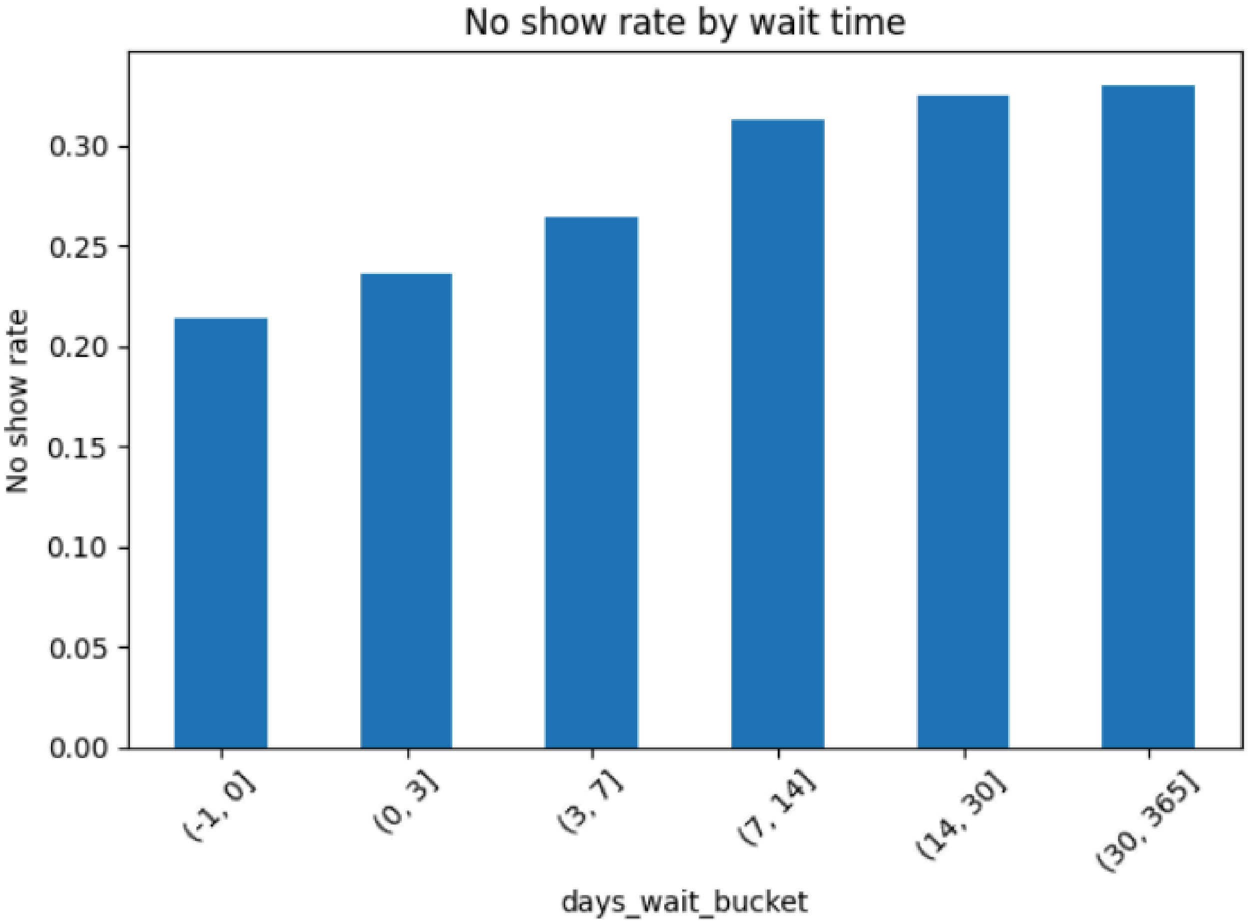
EXPLORATORY ANALYSIS

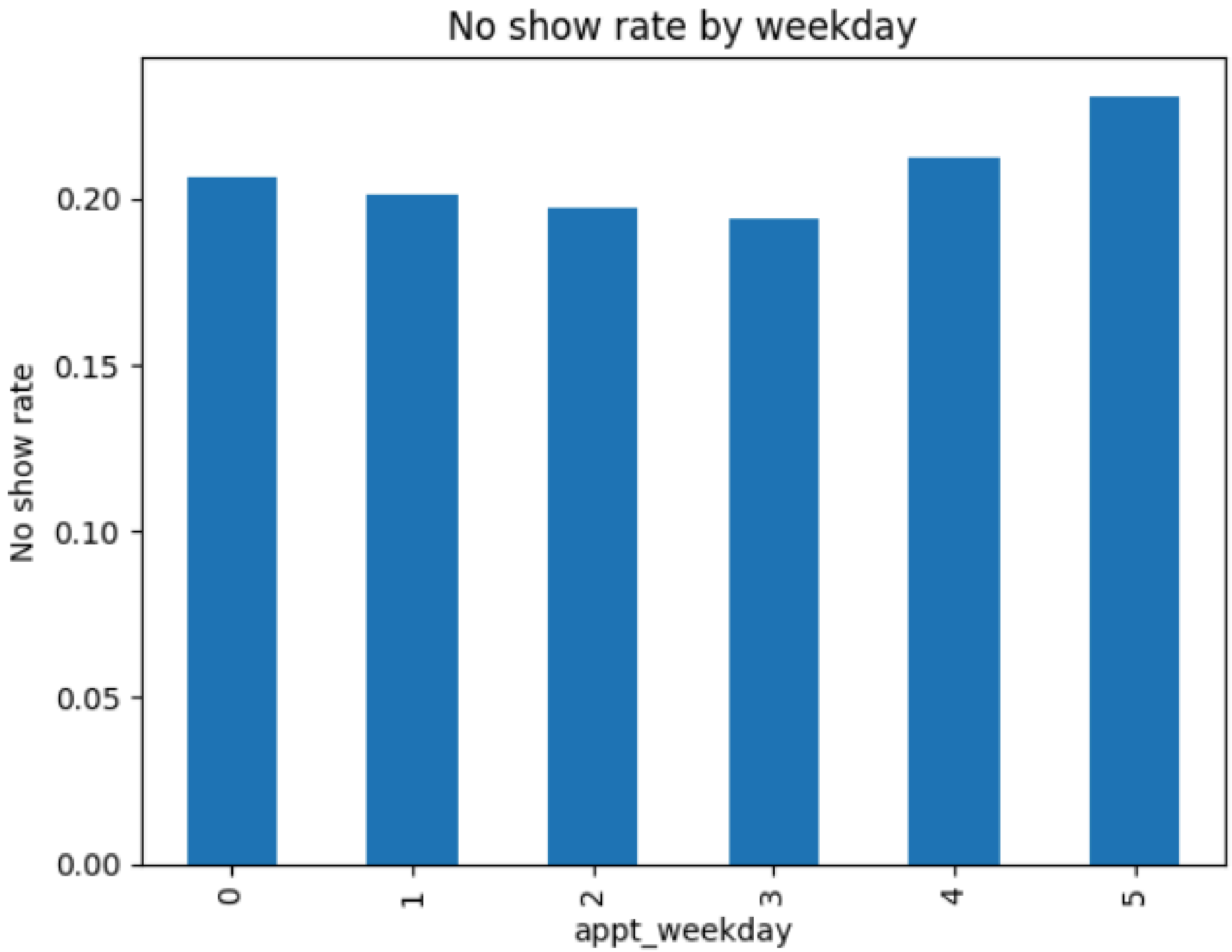
1. No-show rate increases with wait time

- Same-day appointments: ~21% no-shows
- 30+ day wait: ~33% no-shows

2. Weekday effects

- Fridays and certain mid-week days show higher no-show rates
- Operational scheduling windows influence attendance





MODELING APPROACH

Models evaluated:

- Logistic Regression
- Random Forest
- XGBoost

Shared preprocessing pipeline:

- Numeric imputation (median)
- Categorical imputation (most frequent)
- Scaling (numeric)
- One-hot encoding (categorical)
- 80/20 stratified train-test split

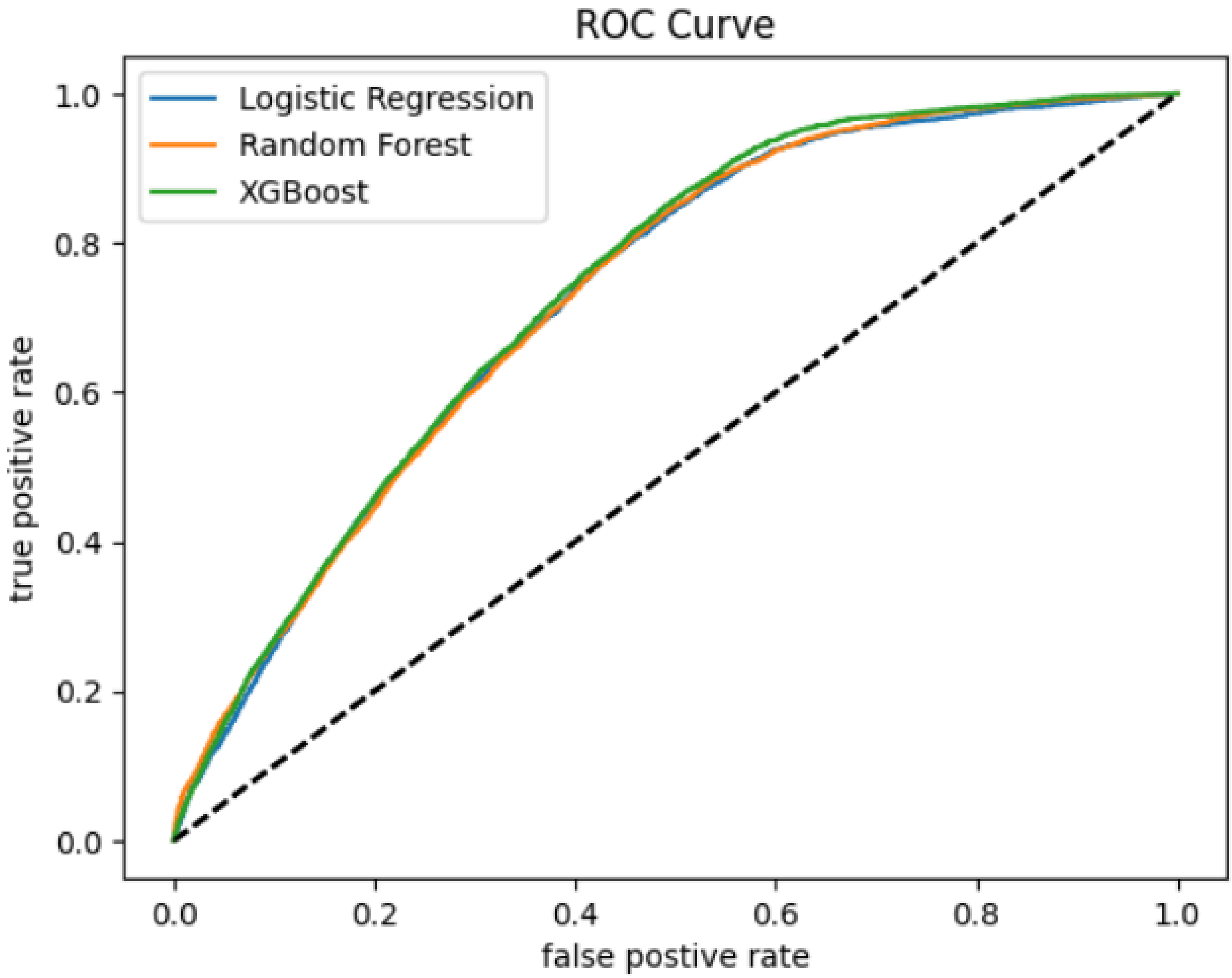
Evaluation metrics: Accuracy, Precision, Recall, F1, ROC AUC

```
=== Logistic Regression===
Accuracy:0.799
Precision:0.500
Recall:0.019
F1 Score: 0.036
ROC AUC: 0.727

Confusion Matrix:
[[17586    83]
 [ 4354    83]]
----------------------------------------
=== Random Forest===
Accuracy:0.794
Precision:0.461
Recall:0.160
F1 Score: 0.237
ROC AUC: 0.730

Confusion Matrix:
[[16839   830]
 [ 3728   709]]
----------------------------------------
=== XGBoost===
Accuracy:0.796
Precision:0.460
Recall:0.094
F1 Score: 0.155
ROC AUC: 0.736

Confusion Matrix:
[[17182   487]
 [ 4022   415]]
----------------------------------------
```

MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.799	0.500	0.019	0.036	0.727
Random Forest	0.794	0.461	0.160	0.237	0.730
XGBoost	0.796	0.460	0.094	0.155	0.736

Best practical model: Random Forest
It provides the strongest balance between performance and interpretability.



Feature Importance Interpretation

1. Days Wait is the Dominant Predictor

The Random Forest model highlights days_wait as the strongest signal across the dataset.

Interpretation:
This is a structural scheduling issue — reducing wait time is the single most impactful operational improvement.

2. Appointment Weekday Reflects Behavioral Patterns

Weekday features (encoded via one-hot vectors) show high importance.

Interpretation:
This suggests that no-shows are not random — they follow weekly behavioral cycles.
Healthcare ops can exploit these patterns for smarter scheduling.

3. Neighborhood Encodes Environmental Constraints

High importance of neighborhood features indicates geographic and socioeconomic drivers.

Interpretation:
Area-specific interventions could dramatically reduce no-shows:

- Telehealth for certain zip codes
- Transportation vouchers
- Localized reminder strategies

4. Gender & SMS Behavior Reflect Engagement Trends

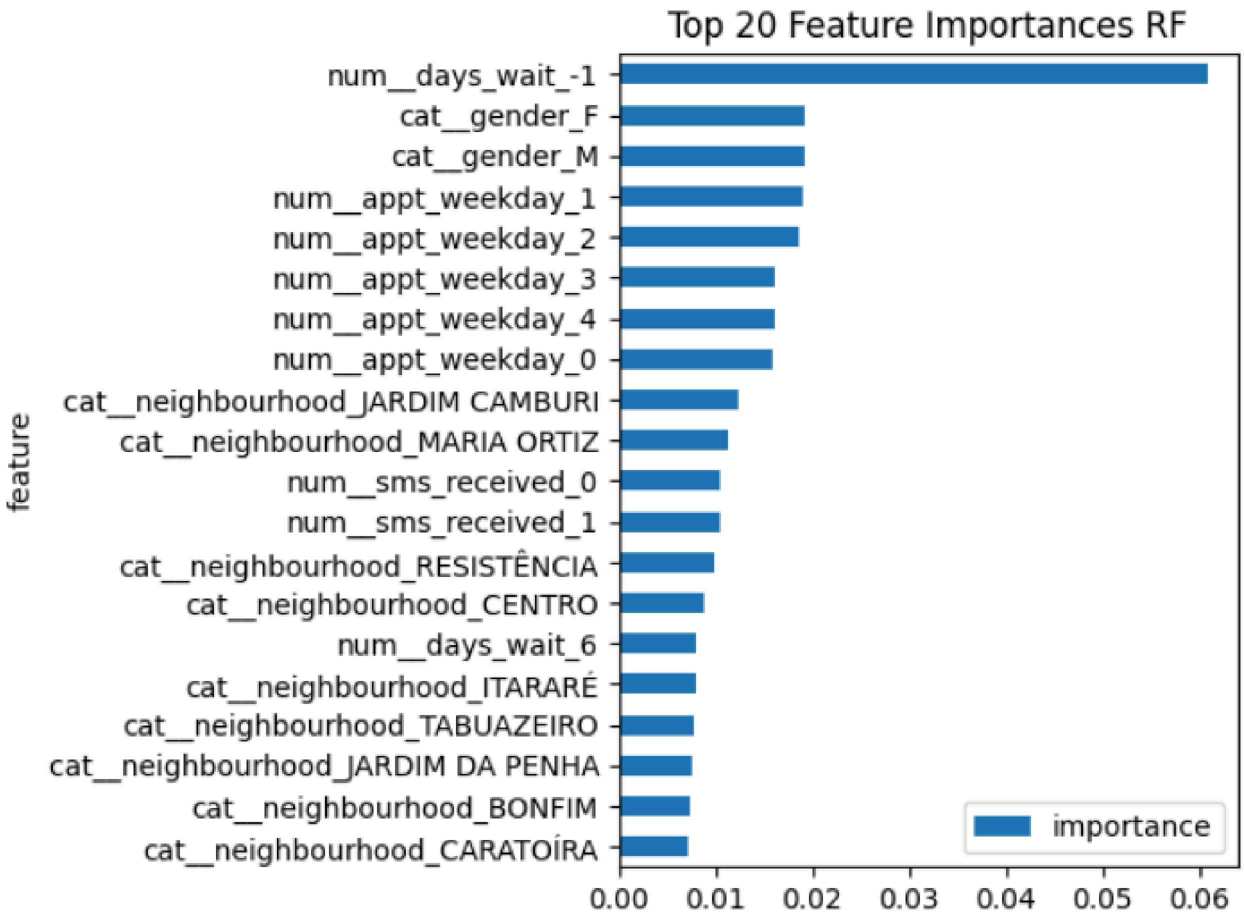
Gender and SMS-related features show moderate predictive power.

Interpretation:
Certain demographic groups may require tailored communication strategies.

Top drivers of no-shows (Random Forest):

1. **Wait time (days_wait)**
2. **Appointment weekday**
3. **Neighborhood**
4. **Gender**
5. **SMS received**

Interpretation:
These are **operationally controllable**, making the model valuable for scheduling decisions



BUSINESS RECOMMENDATIONS

Behavioral Drivers of No-Shows

1. Wait Time Strongly Predicts Attendance Behavior

Across the dataset, no-show probability increases steadily as the time between scheduling and appointment grows.

- Same-day / short wait: lowest no-show rate (~21%)
- 15+ day wait: noticeable increase
- 30+ days: highest risk (~33%)

Interpretation:
Long delays reduce patient commitment. Life circumstances change, patients forget, or they find alternatives.
This is not just statistical, it suggests a behavioral decay curve.

2. Appointment Day of Week Influences Commitment

Fridays show consistently higher no-show rates than mid-week days.

- Monday–Wednesday: more stable patterns
- Friday: peak missed appointments

Interpretation:
Patients may deprioritize healthcare commitments heading into the weekend. Work-week fatigue and schedule conflicts also play a role.

3. Neighborhood-Level Differences Reflect Accessibility Factors

Some neighborhoods exhibit significantly higher no-show rates.

Interpretation:
This likely reflects structural barriers:

- Transportation reliability
- Work schedules
- Socioeconomic constraints
- Distance to clinic

This is actionable, not just observational.

4. SMS Reminders Help, But Their Effect is Limited

While receiving an SMS reduces no-show likelihood, the effect is modest.

Interpretation:
SMS alone is not a strong behavioral lever.
Enhanced or personalized communication strategies may be necessary.

Operational Risks & Opportunities

1. High Wait Times = High Operational Waste

Long wait periods correspond directly to no-show volume.
Operational takeaway:

- Clinics should prioritize scheduling efficiency.
- Opening more short-term availability or reallocating staff could reduce no-show risk.

2. Targeted Overbooking Can Increase Provider Utilization

Based on model predictions and historical trends:

- Clinics can selectively overbook high-risk time slots to counteract expected no-shows.
- Avoid blanket overbooking → leads to overcrowding and poor patient experience.

Key idea: *Data-driven overbooking, not guessing.*

3. Reminder Systems Should Be Tiered, Not Uniform

Instead of sending SMS to everyone equally, clinics can:

- Prioritize **high-risk patients** (based on predicted score)
- Use multi-channel reminders for the top risk percentiles (call + SMS)
- Increase frequency as appointment day gets closer

This reduces cost while increasing impact.

4. Staffing & Scheduling Can Be Adjusted by Day-of-Week Risk

Because certain days consistently show higher no-show rates:

- Fridays may need more reminders, more flexible scheduling, or adjusted staffing levels
- Good day-of-week scheduling may reduce bottlenecks and unused provider slots

1. Reduce wait times for high-risk patients
Even a 10-day reduction can meaningfully reduce no-show likelihood.

2. Prioritize SMS reminders and follow-up calls
Especially for high-risk neighborhoods + Fridays.

3. Implement strategic overbooking
Use predicted risk scores to avoid empty provider slots.

4. Integrate risk scoring into scheduling tools
Support front-desk staff with real-time insights.

Limitations & Future Enhancements

1. Appointment Type is Unknown

Different appointment types have naturally different no-show rates
(check-ups vs chronic care vs acute visits).
Adding this feature would likely improve accuracy.

2. SMS Timestamp Missing

The dataset only indicates whether an SMS was sent — not when.
Reminders closer to the appointment may be more effective.

3. No Social Determinants of Health (SDOH)

Factors like income, transportation availability, or work schedules are not included.
These are powerful predictors in real-world healthcare models.

4. Class Imbalance Limits Recall

No-shows are a minority class, which leads to:

- Models predicting “show” too often
- Lower recall for the no-show class

Techniques like oversampling, cost-sensitive learning, or SMOTE could help.

5. Future Work

- Segment patients and build **persona-specific models**
- Build a **real-time risk scoring tool** for scheduling staff
- Deploy a **reminder optimization model** based on predicted risk
- Experiment with **gradient boosting or neural nets**

CONCLUSION

Predicting no-shows is not simply a modeling exercise — it is a **solution to operational inefficiency**. This project demonstrates:

- How data science can directly support scheduling
- How healthcare operations can reduce waste
- How predictive analytics leads to better patient access

This case study combines technical rigor with practical, real-world recommendations.

Insights Summary

Top Insights:

- Longer wait times strongly increase no-show risk
- Certain weekdays consistently produce more missed appointments
- Neighborhood-level patterns indicate accessibility barriers
- SMS reminders help, but not enough for high-risk groups

Top Recommendations:

- Reduce wait times through scheduling optimization
- Prioritize reminders for high-risk patients
- Implement targeted overbooking
- Adjust staffing and scheduling based on weekday risk

Bottom line:
Predictive analytics gives clinics a practical, data-driven strategy to reduce missed appointments, increase capacity, and improve patient access.