NHS GP APPOINTMENT NO-SHOW PREDICTION

Comprehensive Analysis Report

SQL · Feature Engineering · EDA · Machine Learning

Prepared by: Judah Akinlaja

Email: judahakinlajar@gmail.com

Date: February 2026

Tools: MySQL · Python · Scikit-learn · Power BI

1. Executive Summary

NHS GP practices face a persistent and costly challenge: patients who book appointments but fail to attend — known as Did Not Attend (DNA) events. This project applies end-to-end data science techniques across four phases — SQL database analysis, Python feature engineering, exploratory data analysis, and machine learning — to understand the drivers of DNA behaviour and build a predictive model to support proactive intervention.

Table 1. Key Performance Indicators

Total Appointments	Total DNAs	DNA Rate	Annual Loss
866 Million	40.6 Million	4.42%	£1.22 Billion

The analysis reveals that appointment mode is the single strongest predictor of DNA behaviour (video/online appointments) have a DNA rate of just 0.46%, compared to 5.56% for face-to-face, a 12-fold difference. Machine learning models confirmed that logistic regression with balanced class weighting performs best, achieving 56% accuracy and 0.57 ROC-AUC on aggregated data. While model accuracy is modest due to aggregated data, the analysis uncovers actionable insights that could save up to £180 million annually.

2. Phase 1: SQL Database Analysis

2.1 Dataset Overview

The analysis was conducted on NHS appointment data spanning July 2023 to November 2025. The raw dataset contained over 920 million records. Following cleaning and aggregation, 52,855 rows were analysed across the SQL phase, covering combinations of appointment mode, HCP type, service setting, and time period.

Table 2. Dataset Overview

Attribute	Details
Time Period	July 2023 – November 2025
Raw Records	920+ million entries
Analysed Rows	52,855 aggregated combinations
SQL Queries	30+ across all analytical phases
Database Tool	MySQL
Cost Per DNA	£30 per missed appointment (NHS estimate)

2.2 Key SQL Findings

2.2.1 Overall DNA Rate

Across 866 million appointments, **40.6 million were DNA** — a headline rate of **4.42%**. At £30 per missed appointment, this translates to **£1.22 billion annual loss**.

2.2.2 DNA Rate by Appointment Mode

Appointment mode emerged as the single most influential factor in DNA behaviour:

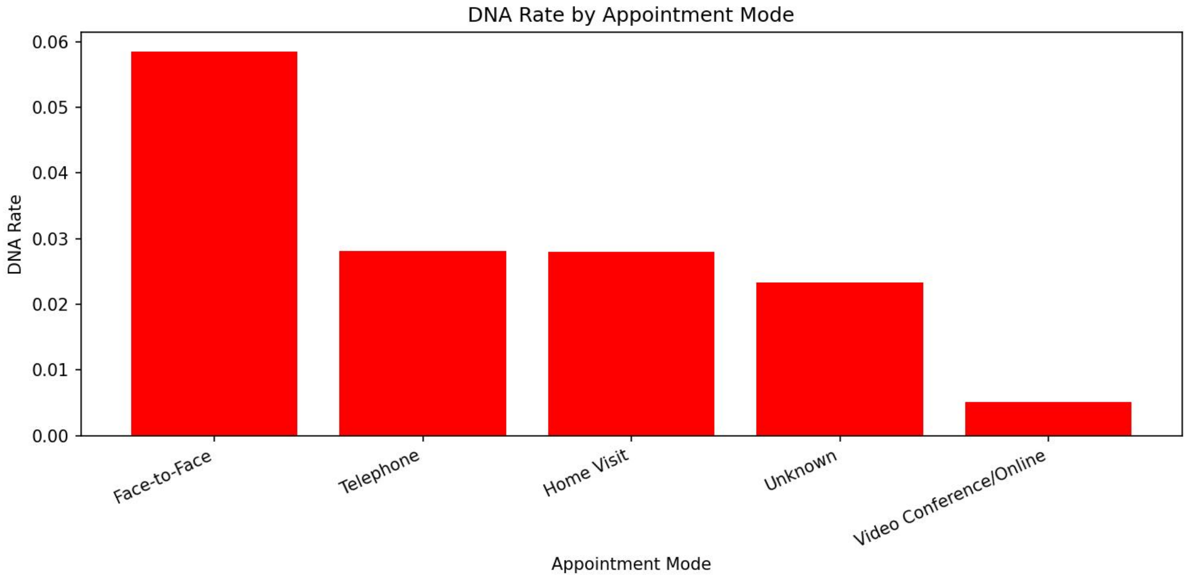
Table 3. DNA Rate by Appointment Mode

Appointment Mode	DNA Rate	Relative Risk	Risk Level
Video Conference/Online	0.46%	Baseline (lowest)	Lowest
Home Visit	1.84%	4x higher	Low
Telephone	3.21%	7x higher	Moderate
Face-to-Face	5.56%	12x higher	Highest

[INSERT CHART]

Figure 1: DNA Rate by Appointment Mode — Bar Chart

Figure 1. DNA rate comparison across appointment modes. Video/online shows 12x lower rate than face-to-face.



Insight: Face-to-face appointments remain the largest operational risk. A modest mode shift toward virtual pathways could reduce missed-appointment losses.

2.2.3 DNA Rate by Healthcare Professional Type

GP-led appointments had the lowest DNA rate at 3.2%, whereas Other Practice Staff recorded 6.2% — nearly double.

Table 4. DNA Rate by HCP Type

HCP Type	DNA Rate	Volume
GP	3.20%	High
Nurse	4.10%	High

Unknown	4.50%	Moderate
Other Practice Staff	6.20%	Moderate

2.2.4 Seasonal & Monthly Patterns

October consistently recorded the worst monthly DNA rate at 5.5%. Winter months showed the lowest at approximately 4.3%.

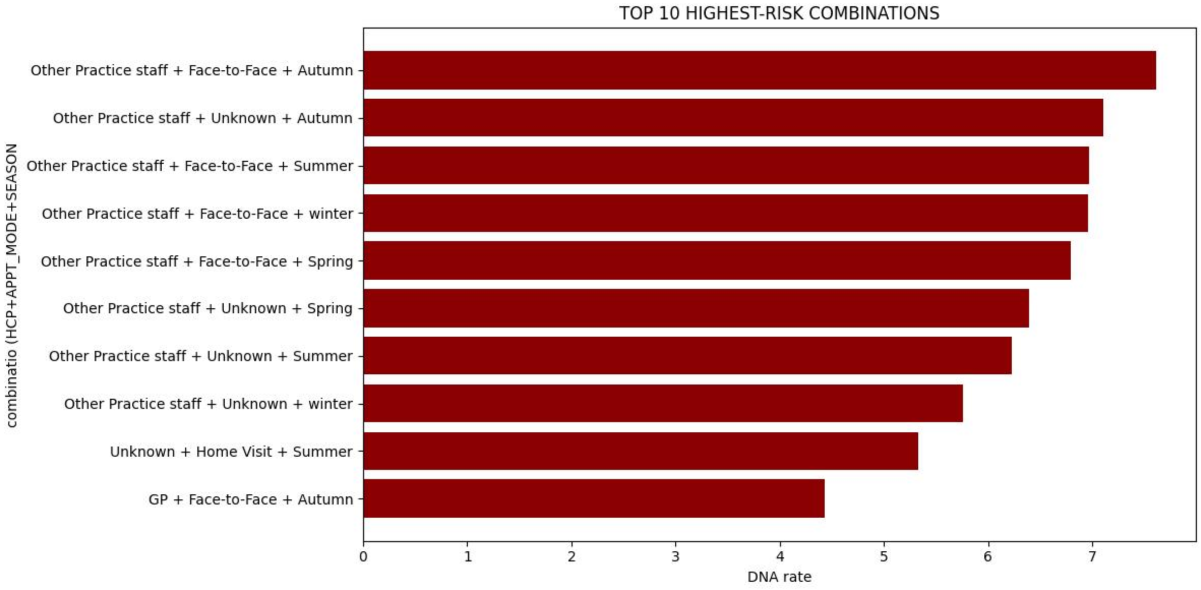
- Autumn: 5.5% DNA rate — highest risk season.
- Winter: 4.3% DNA rate — lowest risk season.
- October 2025: £60 million in lost appointment value alone.
- Improvement trend: 5.19% (Oct 2023) → 5.10% (Oct 2025).

2.2.5 Highest-Risk Combinations

Table 5. Top 5 Highest-Risk Appointment Combinations

#	HCP Type	Appt Mode	Season	DNA Rate
1	Other Practice Staff	Face-to-Face	Autumn	7.61%
2	Other Practice Staff	Face-to-Face	Summer	6.98%
3	Other Practice Staff	Telephone	Autumn	6.45%
4	Nurse	Face-to-Face	Autumn	5.90%
5	GP	Face-to-Face	Autumn	5.31%

Figure 2. Top 10 highest DNA-risk combinations identified through SQL analysis.



Insight: Risk is concentrated in specific combinations, so targeted intervention bundles are likely to outperform generic reminders.

3. Phase 2: Feature Engineering

Feature engineering transformed the raw dataset into a richer analytical dataset by creating derived variables that capture temporal patterns, risk categorisation, and behavioural change signals. All features were created using Python (Pandas/NumPy) in Google Colab.

Table 6. Engineered Features

Feature	Type	Description
season	Categorical	From appointment month: Winter (Dec–Feb), Spring (Mar–May), Summer (Jun–Aug), Autumn (Sep–Nov)
dna_rate	Numeric	DNA count ÷ total appointments per row — key performance indicator
months_since_start	Numeric	Months from dataset start — captures time trend
dna_rate_change	Numeric	Month-over-month % change in DNA rate — trend detection
Risk_level	Categorical	High_Risk (above average DNA rate) or Low_Risk (below average)

3.1 Feature Insights

- Season: Autumn had highest average DNA rate (5.5%), Winter lowest (4.3%).
- dna_rate_change: Largest spikes in October each year.
- Risk_level: ~38% of time periods qualified as High_Risk.
- months_since_start: Enabled time-series modelling with continuous numeric variable.

4. Phase 2: Exploratory Data Analysis

4.1 Correlation Analysis

A Pearson correlation matrix was computed across numeric features. dna_rate and months_since_start showed weak negative correlation, confirming the gradual improvement trend. dna_rate_change showed low correlation with all features, indicating month-on-month changes are largely unpredictable from available variables.

4.2 Chi-Square Test — Appointment Mode vs DNA Status

Table 7. Statistical Test Results

Test	Statistic	Result
Chi-Square Test	p-value < 0.001	Statistically Significant
Cramér's V	0.0732	Small but Meaningful

Interpretation: Appointment mode and DNA status are not independent — mode significantly affects attendance. The Cramér's V of 0.073 indicates a small but practically meaningful effect.

4.3 T-Test — High-Risk vs Low-Risk Periods

An independent t-test confirmed High_Risk periods had significantly higher mean DNA rates than Low_Risk periods ($p < 0.05$), validating the Risk_level feature as a meaningful classifier.

4.4 ANOVA — DNA Rate Across Seasons

Table 8. Seasonal ANOVA Results

Season	Mean DNA Rate	Rank	ANOVA
Autumn	5.5%	1st (Highest)	Significant
Summer	4.9%	2nd	Significant
Spring	4.5%	3rd	Significant
Winter	4.3%	4th (Lowest)	Significant

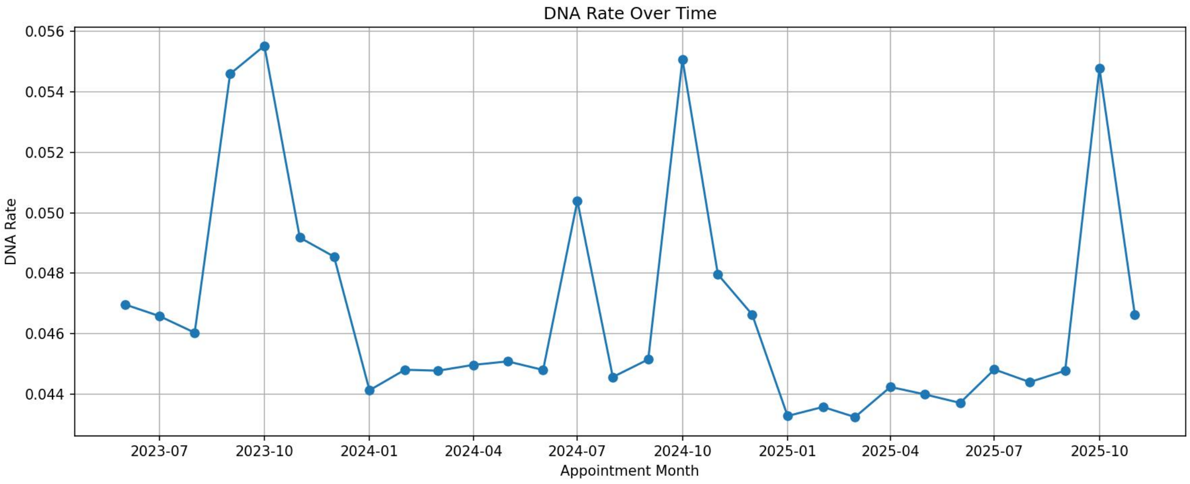
ANOVA confirmed significant differences across seasons ($p < 0.05$), validating season as a meaningful predictor.

5. Phase 2: Data Visualisations

5.1 DNA Rate Over Time

A monthly line chart reveals a general downward trend in DNA rates. Notable spikes occur each October, consistent with seasonal patterns.

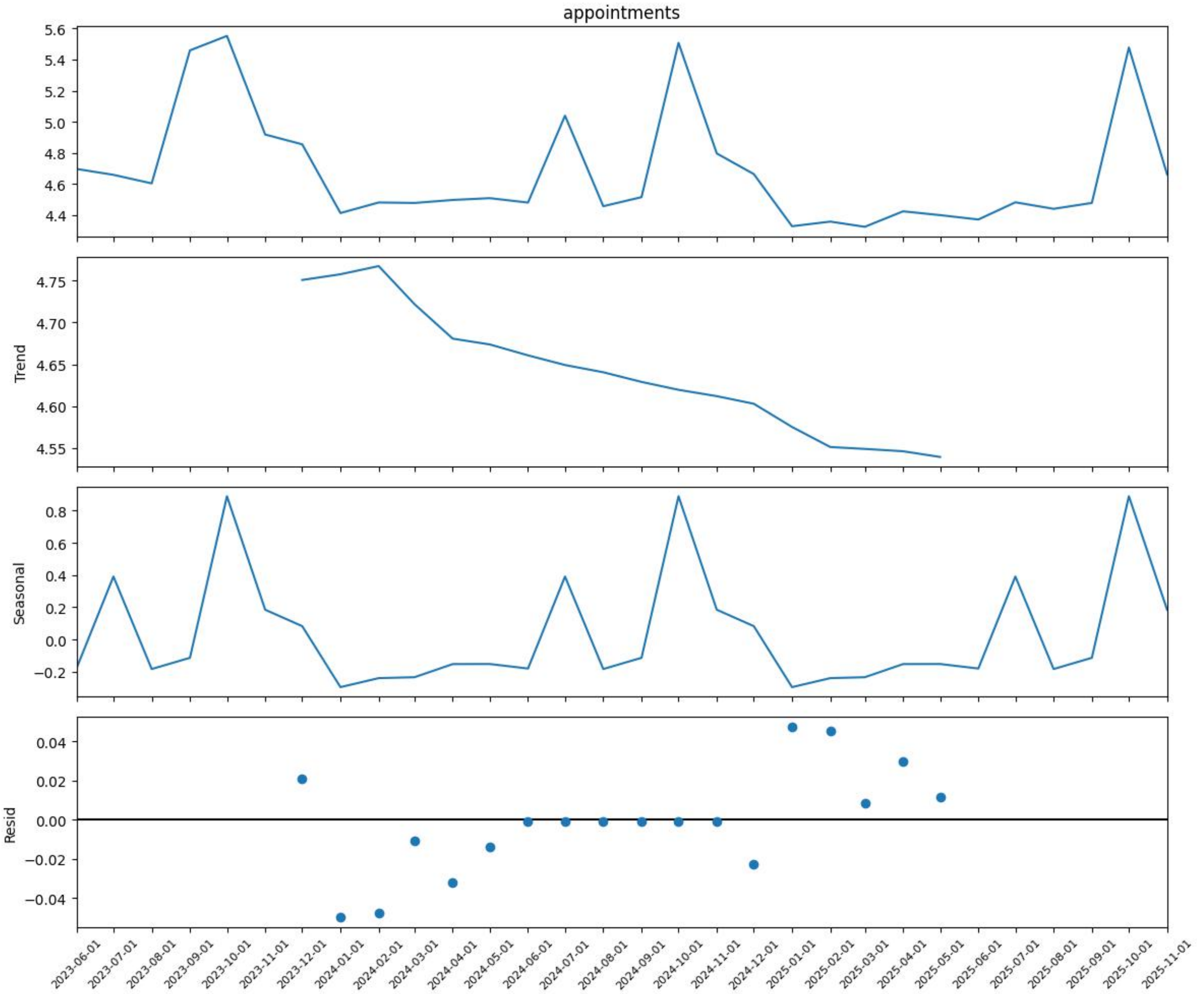
Figure 3. Monthly DNA rate trend showing gradual improvement and consistent October spikes.



5.2 Seasonal Decomposition

Time series decomposition separated DNA rate into trend, seasonality, and residual components, confirming a slight downward trend and strong annual cycle peaking in October.

Figure 4. Decomposition into trend, seasonal, and residual components.

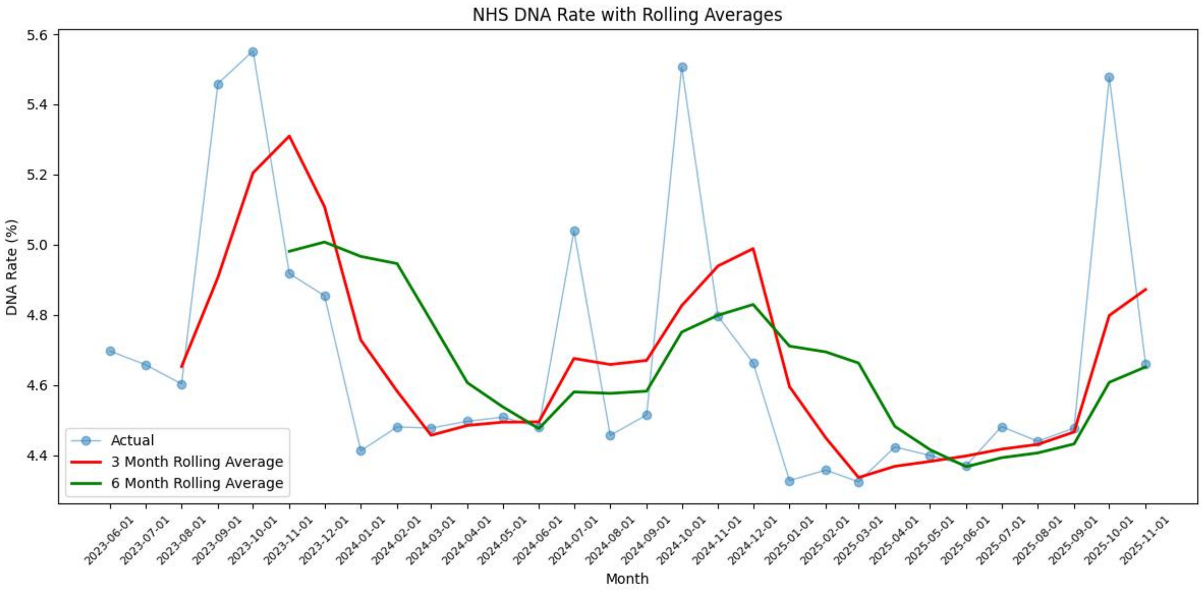


Insight: The seasonal pattern repeats consistently, supporting pre-October planning rather than reactive action after peaks.

5.3 Rolling Averages

Three-month and six-month rolling averages smoothed volatility, showing a consistent decline from approximately 4.75% (mid-2023) to 4.42% (late 2025).

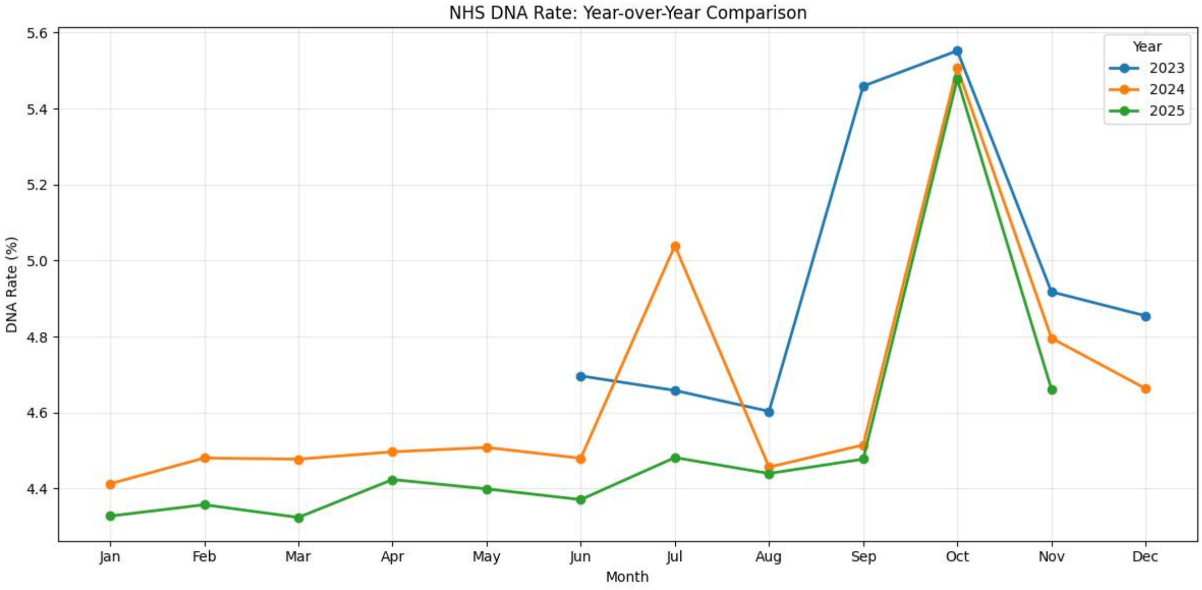
Figure 5. Rolling average analysis highlighting improving trend.



5.4 Year-Over-Year Comparison

Overlaying 2023, 2024, and 2025 monthly rates confirms seasonal patterns and modest annual reduction across all months.

Figure 6. Year-over-year comparison showing gradual annual improvement.

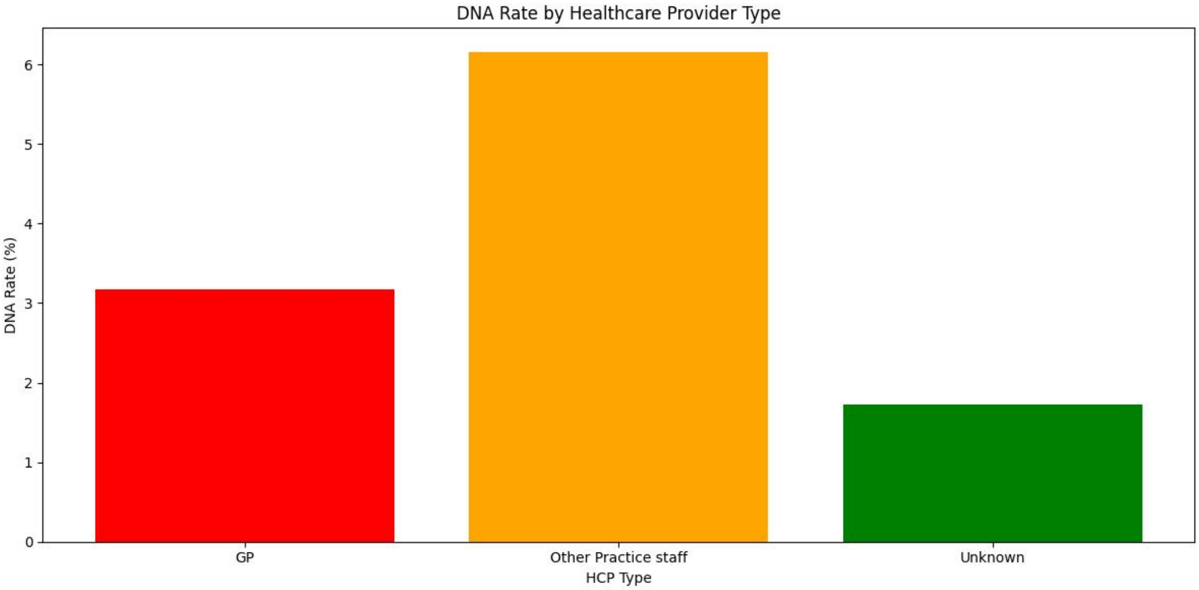


Insight: Improvement is real but incremental, indicating current controls show benefit but are insufficient to deliver a step-change in reduction.

5.5 DNA Rate by HCP Type

Other Practice Staff consistently recorded 6.2% DNA rate while GP appointments had 3.2% — a 3-percentage-point gap with significant financial implications.

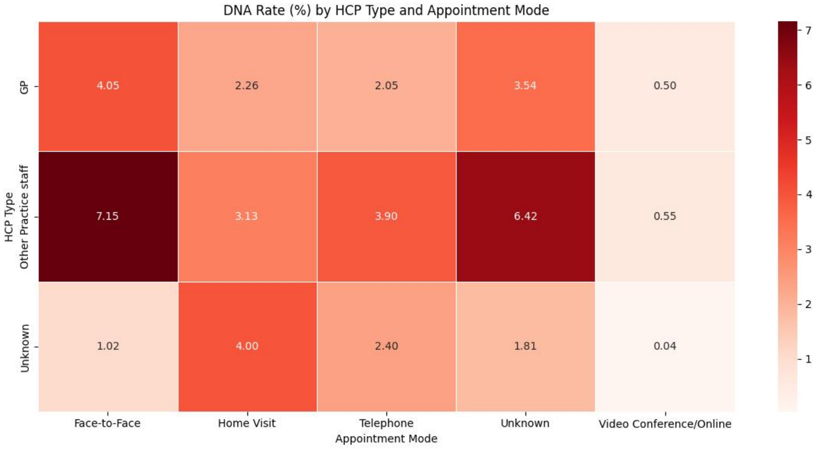
Figure 7. DNA rate by HCP type highlighting Other Practice Staff as highest risk.



5.6 Heatmap — HCP Type vs Appointment Mode

The heatmap reveals Other Practice Staff × Face-to-Face as the highest-risk cell, while any HCP × Video Conference is lowest risk.

Figure 8. Heatmap visualising DNA rate intensity across combinations.

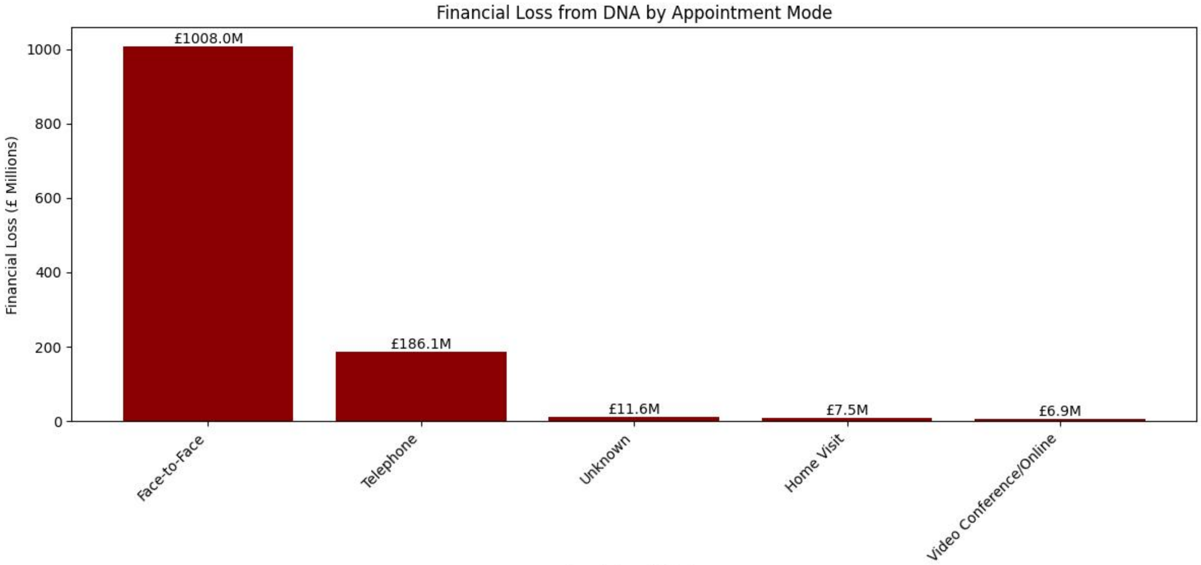


Insight: Risk clusters around specific HCP-mode pathways, providing teams with a clear prioritisation framework for local service redesign.

5.7 Financial Impact Analysis

Face-to-face appointments account for the largest share of financial loss due to both high volume and elevated DNA rates.

Figure 9. Annual cost breakdown of missed appointments by category.



6. Phase 3: Machine Learning

6.1 Data Preparation

- Target: target_dna (1 = DNA, 0 = Attended) derived from appointment status.
- Data leakage prevention: appointments column excluded (initial inclusion gave 93% — artificially inflated).
- Encoding: One-hot encoding on all categorical features → 47 total features.
- Split: 80% training (26,859 rows) / 20% testing (6,715 rows) with stratification.

6.2 Handling Class Imbalance

Table 9. Class Imbalance Approaches

Approach	Outcome	Verdict
class_weight='balanced'	56% acc, 59% recall	Best — balanced and interpretable
sample_weight=appointments	All predictions = 0	Failed — extreme weight range dominated

The sample_weight approach failed because appointment volumes ranged from 1 to 4.5 million per row, causing high-volume rows to dominate training and collapse all predictions to the majority class.

6.3 Models Trained

Three models were trained: **Logistic Regression** (class_weight='balanced'), **Random Forest** (n_estimators=100, class_weight='balanced'), and **XGBoost** (n_estimators=100).

6.4 Model Comparison

Table 10. Model Comparison Results

Model	Test Acc.	Train Acc.	DNA Recall	DNA F1	ROC-AUC
Logistic Regression ★	56%	55%	59%	0.53	0.57
Random Forest	24%	55%	9%	0.09	0.14
XGBoost	55%	58%	15%	0.22	0.55

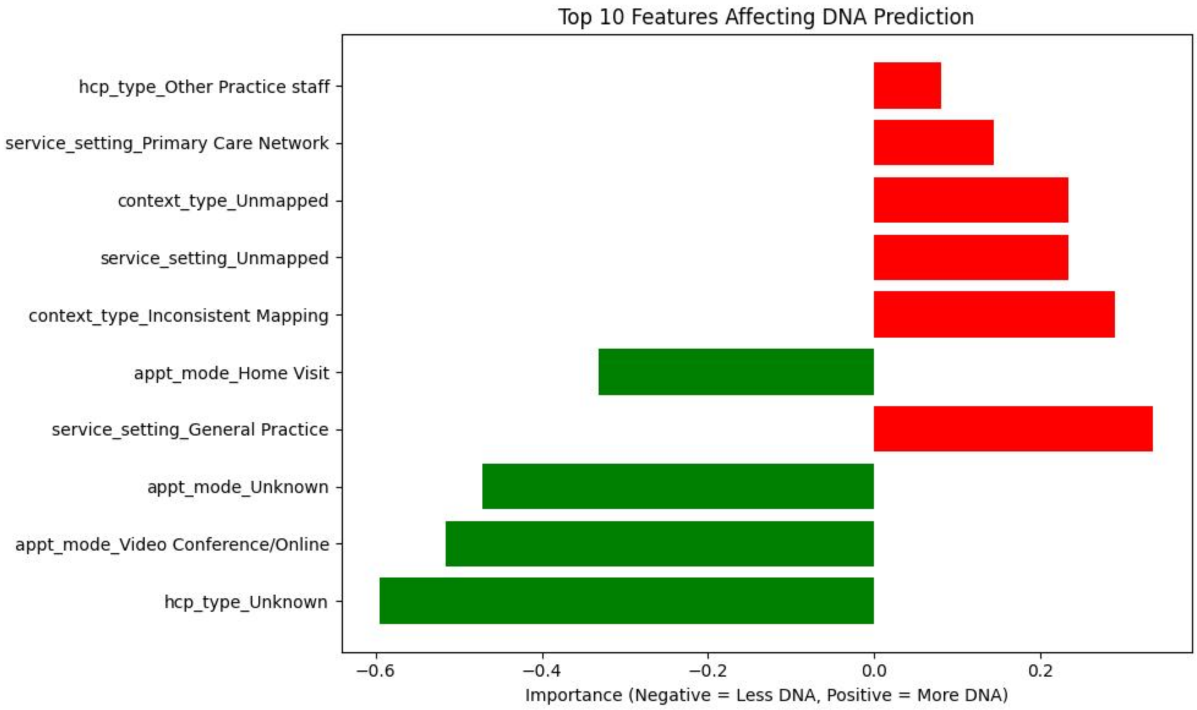
★ Selected model: Logistic Regression. Despite being simplest, it outperformed across all metrics, suggesting linear relationships and limited feature predictive power. Practical interpretation: this model offers the most stable recall-precision balance in the current aggregated dataset and is the most appropriate baseline for operational triage.

6.5 Feature Importance

Table 11. Top 10 Feature Importance (Logistic Regression Coefficients)

#	Feature	Coefficient	Effect
1	hcp_type_Unknown	-0.595	↓ DNA
2	appt_mode_Video/Online	-0.515	↓ DNA
3	appt_mode_Unknown	-0.472	↓ DNA
4	service_setting_General Practice	+0.336	↑ DNA
5	appt_mode_Home Visit	-0.332	↓ DNA
6	context_type_Inconsistent Mapping	+0.289	↑ DNA
7	service_setting_Unmapped	+0.234	↑ DNA
8	context_type_Unmapped	+0.234	↑ DNA
9	service_setting_Primary Care Network	+0.143	↑ DNA
10	hcp_type_Other Practice Staff	+0.080	↑ DNA

Figure 10. Top 10 features. Green = decreases DNA; Red = increases DNA.



Insight: The strongest signals are service-context features, suggesting operational design changes are immediately actionable within current service design constraints.

7. Key Findings Summary

Table 12. Summary of Key Findings

#	Finding
1	Video/online appointments have 12x lower DNA rate (0.46%) than face-to-face (5.56%)
2	40.6 million appointments missed over the period, costing NHS ~£1.22 billion annually
3	October is consistently highest-risk month each year (5.5% DNA rate)
4	Other Practice Staff have nearly double the DNA rate of GPs (6.2% vs 3.2%)
5	DNA rates improving year-on-year: 4.75% (2023) to 4.42% (2025)
6	Logistic Regression outperformed RF and XGBoost: 56% accuracy, 59% recall, 0.57 ROC-AUC
7	Model performance limited by aggregated data — categorical features alone cannot strongly predict DNA
8	Data quality issues (Unmapped, Inconsistent categories) correlate with higher DNA rates

8. Limitations

8.1 Aggregated Data

The dataset contains aggregated rows representing attribute combinations, not individual appointments. This prevents learning patient-specific patterns.

8.2 Missing Patient-Level Features

Key predictors are absent: patient demographics, appointment history, booking lead time, time of day, distance to practice, and deprivation index.

8.3 Class Imbalance

True DNA rate is ~4.42%, creating significant imbalance. At row level the split appeared 57/43, masking the real imbalance. Only class_weight='balanced' produced meaningful DNA predictions.

8.4 Data Leakage Risk

Including the appointments column as a feature produced 93% accuracy — identified as data leakage. Removed, reducing accuracy to true level (~56%).

8.5 Geographic Variation

Data is aggregated nationally. Individual practice characteristics (demographics, transport links, urban/rural, reminder systems) are not captured.

9. Recommendations

9.1 For NHS Managers

- Expand video/online appointments: 0.46% vs 5.56% DNA rate — could save £100–180M annually.
- Targeted reminders for October: Consistently highest-risk month — enhanced SMS/phone reminders could yield significant savings.
- Investigate Other Practice Staff: 6.2% DNA rate is nearly double GP rates — understanding drivers could identify high-impact near-term opportunities.
- Improve data quality: Reduce Unknown, Unmapped, and Inconsistent categories to improve analysis and operations.
- Home visit alternatives: Lower DNA rates may be cost-effective for high-risk patient groups.

9.2 Technical Improvements

- Individual records: Collect appointment-level data rather than aggregated combinations.
- Patient features: DNA history, booking lead time, distance, age group, deprivation index.
- Temporal features: Time of day, days until appointment, days since last visit.
- Practice-level models: Account for local characteristics and patient populations.
- Threshold optimisation: Tune probability threshold to balance recall and precision.

9.3 Potential Financial Impact

Table 13. Estimated Financial Impact of Recommendations

Intervention	DNA Reduction	Annual Saving
Expand video appointments (10% shift)	~0.9%	~£100 million
October targeted reminders	~0.3%	~£30 million
Improve data quality	Indirect	Better targeting
Combined interventions	~1.5%	~£180 million

10. Conclusion

This project demonstrates a complete end-to-end data science workflow applied to a real NHS problem with significant financial and public health implications. The headline finding: **appointment mode is the most powerful lever for reducing DNA rates.** Video/online appointments have a 12-fold lower DNA rate than face-to-face.

Machine learning confirmed that aggregated categorical data has limited predictive power (56% accuracy, 0.57 ROC-AUC). This itself is valuable: it confirms patient-level data would be essential for a clinically actionable prediction system. The groundwork laid here provides a strong foundation for that next step.

This project demonstrates proficiency in MySQL querying, Python data manipulation, statistical hypothesis testing, machine learning model training and evaluation, feature engineering, and translating technical findings into business recommendations — a comprehensive data science skill set applicable across healthcare and beyond.

Prepared by Judah Akinlaja | judahakinlajar@gmail.com | February 2026