**SQL-Based Data Preprocessing for Predicting Patient Appointment Attendance**

**Abstract**

This report presents a structured SQL-based data preprocessing workflow applied to patient appointment records, with the aim of supporting effective machine learning (ML) modelling for predicting non-attendance ("no-show") and enhancing healthcare service efficiency. The methodology involved rigorous data cleaning, validation, and feature engineering using SQLite to ensure analytical reliability. The preprocessed dataset was then used for exploratory visualisation in Tableau and served as the foundation for developing classification models in Python. These models aimed to identify patients at elevated risk of missing appointments, thereby contributing to improved healthcare resource planning and patient care delivery.

**Introduction**

Missed appointments are a persistent challenge in healthcare systems globally, contributing to wasted clinical resources, reduced operational efficiency, and extended patient waiting times. For example, the NHS reported over 7.8 million missed GP appointments in 2022 alone, at a cost of an estimated £216 million (Parsons et al., 2023). Predictive modelling provides a proactive approach to this issue by identifying patients at a higher risk of non-attendance and enabling pre-emptive interventions, such as reminders or rescheduling. However, the success of such predictive systems is highly dependent on the quality and structure of input data. This report details an SQL-driven preprocessing approach applied to a real-world dataset of medical appointments. The focus was on addressing data integrity issues, engineering informative features, and preparing the data for downstream machine learning (ML) applications, particularly in sensitive healthcare environments where reliability and fairness are paramount.

**Situation Analysis**

The original dataset comprised 110,527 appointment records, including variables such as patient and appointment identifiers, categorical indicators (e.g., gender, presence of chronic conditions), date-time values, and a binary target variable indicating appointment attendance. Initial analysis identified multiple data quality issues. These included inconsistent data types—particularly IDs and dates represented as text—implausible values such as negative waiting times and unrealistic patient ages, redundant records, and missing or duplicated entries. Furthermore, date fields were poorly formatted, limiting their usability in time-based analysis. Such inconsistencies posed significant risks to modelling accuracy and interpretability, necessitating a systematic and robust data cleansing strategy.

**Task Definition**

The primary objective was to transform the raw appointment data into a format suitable for machine learning through SQL-based preprocessing. Specific goals included the conversion and standardisation of data types (e.g., converting text to integer or date formats), identification and correction of anomalous records, and the creation of a derived feature called "LeadTime," which captured the interval between when an appointment was scheduled and when it occurred. Additional objectives included removing duplicates and records with null values to ensure dataset integrity, as well as exporting the final, cleaned dataset in a reproducible, version-controlled format. This foundation would then support visual analytics and predictive modelling using Python-based ML frameworks.

**Actions Taken**

A structured SQL pipeline was developed and implemented within a SQLite environment, consisting of four key stages. Version control and documentation were managed through GitHub to ensure transparency and reproducibility.

Data Ingestion

The raw .csv file containing the appointment data was imported into SQLite, forming a staging table (appointments\_raw). This preserved the original structure for auditing and enabled a seamless transformation process.

Data Type Conversion and Feature Engineering

Key preprocessing steps included casting identifier fields to text to prevent rounding errors, reformatting the date fields (ScheduledDay and AppointmentDay) to exclude unnecessary time components, and converting categorical flags (e.g., Hypertension, Diabetes) into integer format for compatibility with machine learning (ML) algorithms. The binary target variable (No-show) was transformed into a new integer column (NoShow\_Int) for model input. A new feature, LeadTime, was derived by calculating the number of days between the appointment scheduling and the actual appointment date, serving as a proxy for patient scheduling behaviour.

Data Cleaning and Validation

To ensure the dataset's validity, records with implausible values were removed. This included filtering out patients with ages less than zero or greater than 110. Cases with negative LeadTime values were examined: those representing same-day scheduling were retained, while others deemed erroneous were removed. Duplicate and null entries were identified using SQL queries and eliminated to maintain the dataset’s integrity.

Export and Reproducibility

The final cleaned dataset was saved as a .csv file (appointments\_cleaned). All scripts and associated outputs were versioned and uploaded to GitHub. This ensured reproducibility, a key consideration for projects involving healthcare data, and allowed for easy auditing of preprocessing steps.

**Results**

Data Quality Improvements

The data cleaning process resulted in a slight reduction in the dataset size, with the final record count totalling 110,521. This reduction was due to the removal of implausible or incomplete records. The LeadTime feature, created during the preprocessing phase, provided a valuable and interpretable metric for downstream modelling, reflecting scheduling patterns that could influence patient attendance.

Initial Exploratory Insights

Exploratory data analysis was conducted using Tableau. The gender distribution revealed that 65% of the patients were female, and 35% were male, consistent with broader healthcare usage trends. A significant proportion of appointments involved patients aged between 0 and 9 years, suggesting frequent paediatric consultations. Temporal analysis revealed a spike in no-shows in mid-May, suggesting potential seasonality or event-related factors. Additionally, the data revealed a class imbalance, with 79% of patients attending their appointments and 21% failing to attend. This imbalance highlighted the need for sampling techniques in the modelling phase to ensure fair and effective predictions.

**Predictive Modelling and Performance Evaluation**

Further preprocessing was conducted in Python to complement the initial SQL-based data preparation. This phase involved additional transformations to ensure compatibility with machine learning workflows, such as encoding categorical variables and scaling numerical features. Subsequently, four classification models—Logistic Regression, Decision Tree, Random Forest, and XGBoost—were developed to predict patient attendance at appointments. Given the class imbalance identified during exploratory analysis (with 79% attendance and 21% no-shows), three resampling strategies were implemented to mitigate bias: Synthetic Minority Oversampling Technique (SMOTE), Random Undersampling (RUS), and Neighbourhood Cleaning Rule (NCR). These techniques aimed to enhance model performance and fairness by improving sensitivity to the minority class.

Performance Metrics

Model performance was evaluated using three key metrics: Accuracy, which measures the overall correctness of predictions; Balanced Accuracy, which is more suitable for imbalanced datasets and averages the true positive and true negative rates; and AUC-ROC, which assesses the model’s ability to distinguish between classes across different threshold values.

Key Results

| **Model** | **Sampling** | **Accuracy** | **Balanced Accuracy** | **AUC-ROC** |
| --- | --- | --- | --- | --- |
| Decision Tree | NCR | -12.5% | +24.6% | 0.81 |
| Logistic Regression | RUS | -15.2% | +23.0% | 0.79 |
| Random Forest | SMOTE | -5.6% | +33.5% | 0.84 |
| XGBoost | NCR | -8.3% | +23.5% | 0.87 |

While overall accuracy occasionally declined following the application of sampling techniques, both balanced accuracy and AUC-ROC values improved significantly. Notably, the Random Forest model achieved the highest balanced accuracy, while the XGBoost model attained the highest AUC-ROC score of 0.87. These enhancements indicate that the models have become more effective at identifying no-show cases—an essential requirement in clinical settings, where missed appointments can lead to underutilised resources, disrupted care continuity, and increased operational costs.

Predictors

Feature importance plots and SHAP (SHapley Additive exPlanations) analysis consistently identified LeadTime, SMS\_received, and Age as the most influential variables across all models. These results align with existing literature, which suggests that patients are more likely to miss appointments when given short notice or when they do not receive reminders.

**Discussion**

The SQL-based preprocessing pipeline proved highly effective in preparing the appointment dataset for predictive modelling. By systematically addressing inconsistencies and engineering informative features such as LeadTime, the dataset’s quality and predictive value were substantially enhanced. The integration of GitHub for version control added a layer of transparency and reproducibility to the data preparation process. While some models experienced a decrease in overall accuracy following sampling adjustments, notable improvements were observed in fairness metrics, particularly in identifying patients at risk of non-attendance. This is crucial in real-world healthcare scenarios, where the consequences of false negatives can disrupt care delivery.

Limitations

There are two primary limitations to this approach. First, the aggressive removal of records with data anomalies may have resulted in the unintended exclusion of valid but atypical cases. Second, to maintain patient confidentiality, personally identifiable variables were excluded, limiting the ability to model patient-level patterns, such as historical attendance behaviour or specific demographic profiles.

Recommendations

To build upon the success of this project and enhance future predictive performance, several actions are recommended. Firstly, future data collection efforts should aim to capture additional contextual variables, such as transportation access, weather conditions, and a history of previous no-shows. Secondly, regular SQL-based data audits should be embedded into the clinical data pipeline to ensure ongoing data integrity. Finally, advanced feature engineering efforts should consider time-based trends (e.g., the day of the week of an appointment) and spatial data (e.g., distance to the clinic), which may improve model generalizability and interpretability.

**Conclusion**

This report demonstrates the critical role of SQL-based preprocessing in preparing healthcare appointment data for machine learning applications. Through systematic data cleaning, validation, and feature engineering, the project successfully generated a model-ready dataset that enabled accurate and fair prediction of patient attendance. The integration of sampling techniques and model interpretability tools further strengthened the clinical relevance of the findings. Future iterations should prioritise enriched data sources and continuous validation practices to improve further the robustness and scalability of predictive solutions in healthcare.

**Dataset**

Kaggle: No-show appointments <https://www.kaggle.com/datasets/joniarroba/noshowappointments/data>

**Reference**

Parsons, J., Abel, G., Mounce, L.T. and Atherton, H. (2023). The changing face of missed appointments. British Journal of General Practice, 73(728), pp.134–135. <https://doi.org/10.3399/bjgp23x732249>.