Predicting A&E Attendance from Primary Care Data

NON-TECHNICAL EXPLANATION OF YOUR PROJECT

This project investigates factors predicting A&E (Accident & Emergency) attendances in the UK using a dataset of 4,495 patients from a single practice. It includes demographic, health, and deprivation data, alongside A&E attendance records over three years. The aim is to identify links between primary care factors (e.g., long-term conditions, deprivation index) and A&E visits to improve resource allocation and preventive care. Mental health and access to primary care are highlighted for future exploration. The dataset is anonymised to comply with data protection laws, with a focus on avoiding stigmatisation or overgeneralisation in its application.

DATA

The dataset was created

* by South West London BI service who securely emailed attendances to A&E over the last 3 years of patients from my surgery only
* A data analysis tool I use in my surgery to risk stratify and link population groups to deprivation and population health metrics.

MODEL

I settled on Random Forest with Oversampling. This gave me the best outcome on the most important aspect which is predicting A+E Attendance

HYPERPARAMETER OPTIMSATION

Various hyperparameters in Random Forest were used as show below.

'n\_estimators': [150, 200, 250],  
'max\_depth': [8, 12, 16], *# Lower depth reduces noise*'min\_samples\_split': [10, 20, 30], *# Increase split size to prevent overfitting*'min\_samples\_leaf': [5, 10, 15], *# Larger leaves reduce false alarms*'class\_weight': [{0:1, 1:1.5}, {0:1, 1:2}], *#*

The main focus after this was to see if we could optimise the threshold around Precision and Recall (True Positives)

A graph of different colored lines

AI-generated content may be incorrect.

It was decided to settle on Threshold value of 0.3 which gave the best results

RESULTS

Enclosed are the results of the final Confusion Matrix using Random Forests. I also attempted to look at Gradient Boosting but got better results from Random Forest with some hyperparameter Tuning

A green squares with numbers and labels

AI-generated content may be incorrect.

|  |  |  |
| --- | --- | --- |
| **Category** | **Raw Count** | **Percentage** |
| **True Negatives (TN)** ✅ | 377 | **59.67%** |
| **False Positives (FP)** ❌ | 255 | **40.33%** |
| **False Negatives (FN)** ❌ | 181 | **31.00%** |
| **True Positives (TP)** ✅ | 403 | **69.00%** |

**Interpretation**

✅ **True Positives improved (69.00%)**, meaning **more A&E attendees were correctly detected**.  
✅ **False Negatives are still low (31.00%)**, meaning **fewer missed A&E attendees**.  
❌ **False Positives are slightly higher (40.33%)**, meaning **some non-A&E cases are still misclassified**.  
📌 **The model is favouring Recall (detecting A&E attendees) over Precision**.

This is the best fit for the model and comprise between True Positives and False Positives

CONTACT DETAILS

Pending IG approval I plan to publish my results on GitHub

<https://github.com/RazaToosy/Imperial_Portfolio_Project>

Also

<https://www.linkedin.com/in/razatoosy/>

<https://www.razatoosy.com>