Machine Learning Progress Report

Paula Lozano Gonzalo

April 9, 2025

1 Model Overview

This report documents my progress in developing a classifier for diabetes prediction. I began with an SGDClassifier as a baseline model and have now progressed to evaluating a Random Forest classifier with polynomial features.

The task is binary classification - predicting diabetes outcomes based on patient characteristics. The Random Forest model was chosen based on prior experience suggesting it performs better than linear models for this dataset.

2 Dataset & Preprocessing

The diabetes dataset contains both numerical and categorical features. Key preprocessing steps included:

- Handling missing values:
  - Categorical features: Most frequent imputation
  - Numerical features: Median imputation
- Feature engineering:
  - Added polynomial features (degree=2) to capture non-linear relationships
- Feature encoding:
  - Categorical features: Binary encoding

3 Training Progress

The Random Forest model achieved a little better performance than the initial SGDClassifier:

Here are the confusion matrices:

```
==== CM Training Data ====   ==== CM Validation Data ====
 t/p      F     T              t/p      F     T
    F 65887.0  28.0              F 7288.0   0.0
    T 1990.0 4095.0              T 240.0 472.0
```

Metric	Training	Validation
Accuracy	0.999	0.969
Precision	0.993	1.000
Recall	0.673	0.663
F1 Score	0.802	0.797

Table 1: Random Forest performance metrics

```
Precision: 0.993          Precision: 1.000
Recall:    0.673          Recall:    0.663
F1:        0.802          F1:        0.797
```

Again, we can't be happy with 99% precision because if we look at the confusion matrices we can see that the model predicted 1990 non-diabetic when they were for the training data and 240 for the validation data. These results show me that I have long ways to go.

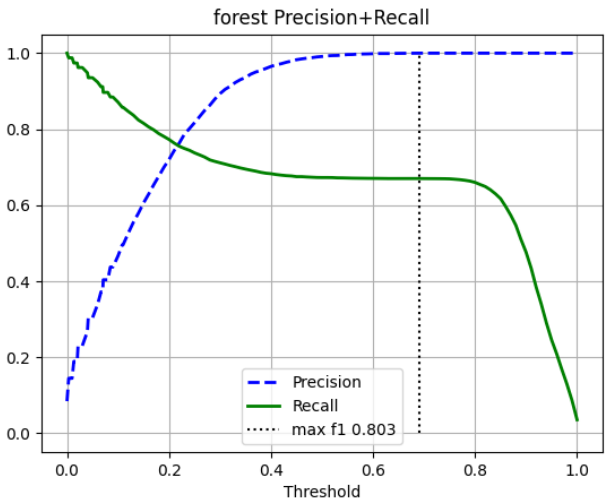

Figure 1: Precision-Recall Plot for Random Forest

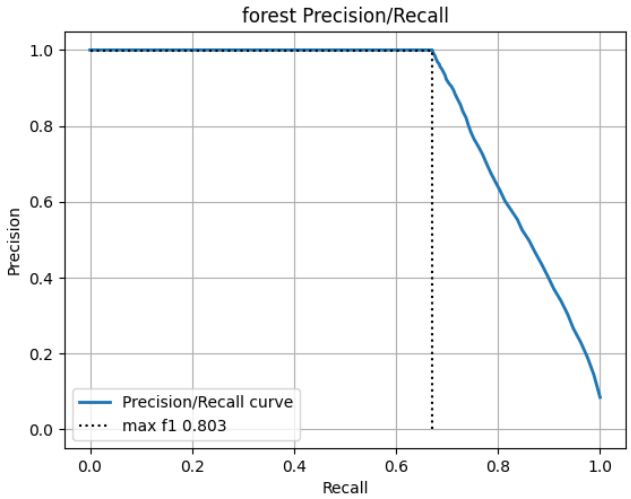

Figure 2: Precision-Recall Curve for Random Forest

Key observations:

- The model achieves near-perfect training accuracy (0.999)
- Validation performance remains strong (0.969 accuracy)
- The model is excellent at avoiding false positives (precision=1.0 on validation)
- Recall remains moderate (0.663), indicating some difficulty identifying all positive cases, which is not good if we told someone they are not diabetic when they really are

4 Challenges & Solutions

One significant **challenge** identified during model evaluation is the class imbalance shown in the confusion matrices. The dataset contains significantly more negative cases than positive ones, which introduces a prediction bias toward negative classifications. For instance, the model incorrectly classified 1990 diabetic cases as non-diabetic. To address this issue, I am considering two potential solutions:

1. Preprocessing the dataset to balance the class distribution by having approximately equal numbers of positive and negative cases
2. Being more critical when interpreting accuracy metrics—recognizing that a 95% accuracy score can be misleading when the model systematically misses positive cases. This awareness is particularly important given the medical context where false negatives carry serious implications.

5 Next Steps

Based on the identified class imbalance challenge, my immediate next steps will focus on data cleaning and comparative evaluation:

- **Data Cleaning**: Restructure the dataset to balance positive and negative cases, then reevaluate both SGDClassifier and Random Forest models to measure the impact of balanced classes
- **Model Comparison**: After data cleaning, conduct side-by-side comparisons of:
  - Original imbalanced dataset performance
  - Balanced dataset performance
- **Feature Analysis**: Investigate feature importance in both balanced and imbalanced scenarios to identify any shifts in predictive patterns
- **Evaluation Metrics**: Develop a more comprehensive evaluation framework that prioritizes recall and F1-score given the medical context