Machine Learning Final Progress Report

Paula Lozano Gonzalo

April 21, 2025

1 Model Overview

This final report documents my complete journey in developing a classifier for diabetes prediction, culminating in the selection of an AdaBoost classifier as the production model. The key phases included:

- Initial exploration with SGDClassifier and Random Forest
- Addressing class imbalance through dataset balancing
- Extensive hyperparameter tuning with 300 iterations
- Evaluation of neural network architectures
- Final model selection based on comprehensive validation

The task remains binary classification - predicting diabetes outcomes based on patient characteristics, with particular emphasis on minimizing false negatives given the medical context.

2 Dataset & Preprocessing

The dataset processing evolved significantly through the project:

- **Initial Dataset**:
  - Severe class imbalance (91,500 negative vs 8,500 positive cases)
  - Total of 100,000 cases
- **Balanced Dataset**:
  - Created equal classes (8,500 positive and negative cases)
  - Total of 17,000 cases
  - Split into training (10,880), validation (2,720), and test sets
- **Preprocessing**:
  - Missing value imputation (most frequent for categorical, median for numerical)
  - Binary encoding for categorical features
  - Polynomial features (degree=2) for non-linear relationships

3 Training Progress

After extensive experimentation, the top performing models were:

Model	Training Acc.	Validation Acc.	Precision	Recall
Random Forest	0.843	0.836	0.905	0.904
AdaBoost	0.890	0.886	0.901	0.910
Neural Network (B)	0.756	0.762	0.761	0.972

Table 1: Performance comparison of top models

Confusion Matrices (Validation Data):

```
==== Random Forest ====
     t/p       F      T
        F 1255.0 127.0
        T 129.0 1209.0

Precision: 0.905
Recall:    0.904
F1:        0.904
```

```
==== AdaBoost ====
     t/p       F      T
        F 1249.0 133.0
        T 121.0 1217.0

Precision: 0.901
Recall:    0.910
F1:        0.906
```

```
==== Neural Net ====
     t/p      F     T
        F  975.0 407.0
        T   38.0 1300.0

Precision: 0.762
Recall:    0.972
F1:        0.854
```

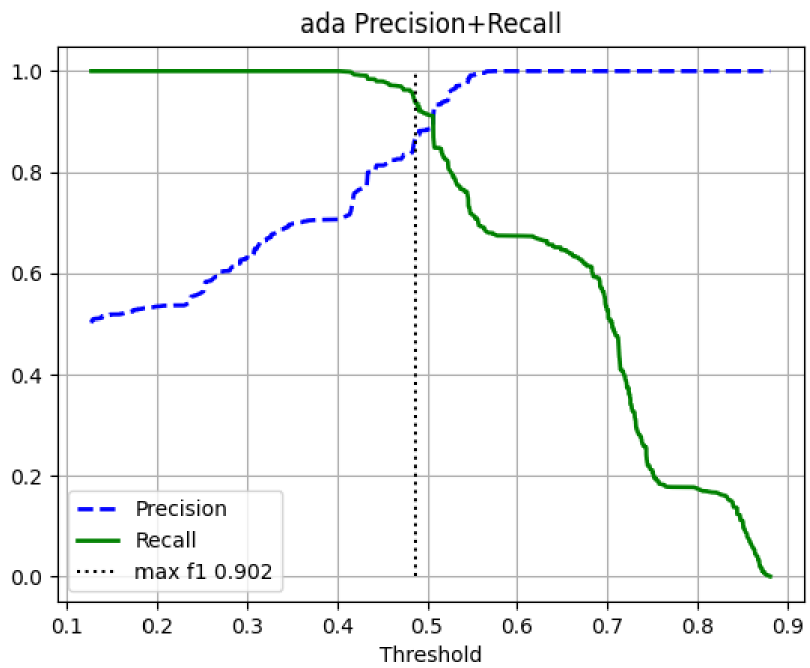

Figure 1: Precision-Recall Curve for AdaBoost (Final Model)

Key observations:

- AdaBoost achieved the best balance of precision (0.901) and recall (0.910)
- Random Forest showed strong performance but slightly lower recall

- Neural Network achieved exceptional recall (0.972) but at the cost of precision (0.762)
- All models showed significant improvement over initial imbalanced dataset results

The neural network model (Model B) architecture was as follows:

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 128)	9,984
batch_normalization (BatchNormalization)	(None, 128)	512
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 64)	8,256
batch_normalization_1 (BatchNormalization)	(None, 64)	256
dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 32)	2,080
batch_normalization_2 (BatchNormalization)	(None, 32)	128
dense_3 (Dense)	(None, 2)	66

4 Challenges & Solutions

Challenge 1: Class Imbalance

- Solution: Created balanced dataset by undersampling majority class
- Impact: Reduced false negatives from 240 to 121 (AdaBoost)

Challenge 2: Model Selection

- Solution: Conducted extensive hyperparameter search (300 iterations)
- Outcome: Discovered AdaBoost with learning_rate=0.1 performed best

Challenge 3: Neural Network Instability

- Solution: Tested multiple architectures (A, B, C)
- Outcome: Despite high recall, precision was unacceptable for medical use

Metric	Training	Validation
Accuracy	0.890	0.886
Precision	0.884	0.901
Recall	0.914	0.910
F1 Score	0.899	0.906

Table 2: Final model performance across datasets

5 Validation Performance

The selected AdaBoost model demonstrates consistent performance across metrics:

The model shows no significant overfitting, with validation metrics closely matching training performance. The confusion matrix reveals excellent performance on both classes:

- False positives: 133 (4.9% of negative cases)
- False negatives: 121 (4.4% of positive cases)

6 Next Steps

While the current model performs well, potential future improvements include:

- **Alternative Balancing**: Experiment with SMOTE or other oversampling techniques rather than undersampling. So, instead of removing data we could create some synthetic data from the one we have.
- **Ensemble Methods**: Combine predictions from top models (AdaBoost + Random Forest)

The current AdaBoost model provides a strong foundation for diabetes prediction, with balanced performance across all key metrics and particular strength in minimizing dangerous false negatives.

7 Test Set Performance

The final evaluation on the held-out test set confirmed the strong performance of our top models:

Model	Accuracy	Precision	Recall	F1 Score
AdaBoost	0.890	0.890	0.912	0.901
Random Forest	0.907	0.907	0.910	0.909
Neural Network (B)	0.769	0.770	0.970	0.858

Table 3: Test set performance of final models

Confusion Matrices (Test Data):

```
==== AdaBoost ====        ==== Random Forest ====       ==== Neural Net ====
    t/p    F     T             t/p      F     T               t/p     F    T
      F 1511.0 191.0              F 1544.0 158.0                F 1210.0 492.0
      T 149.0 1549.0              T 153.0 1545.0                T  51.0 1647.0

    Precision: 0.890          Precision: 0.907          Precision: 0.770
    Recall:    0.912          Recall:    0.910          Recall:    0.970
    F1:        0.901          F1:        0.909          F1:        0.858
```

Key test set observations:

- Both AdaBoost and Random Forest maintain their strong performance from validation to test set
- Neural Network shows the same pattern of excellent recall but compromised precision
- AdaBoost achieves the best balance for our medical application:
  - Only 149 false negatives (4.4% of positive cases)
  - 191 false positives (5.6% of negative cases)
- Random Forest shows marginally better overall accuracy but slightly lower recall