Predicting Readmission of Diabetes Patients: An ML Analysis

Stuti Pande (1082167)

1 Introduction

Hospital readmission is a critical healthcare challenge, especially for chronic conditions like diabetes. Preventing unnecessary readmissions, not only improves patient outcomes, but also reduces the financial burden on hospitals significantly. In 2017, diabetes patients accounted for 1/5th of unplanned re-hospitalisations, costing US hospitals approximately 123 billion [2]. This report aims to analyse the ability of machine learning classification models in predicting readmission risk of diabetes patients within 30 days of their discharge. The data for this study comes from the Diabetes 130-US hospitals for years 1999–2008 Data Set, available on the UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008 [3]. My personal interests in the intersection of data science and healthcare motivated me to explore how predictive analytics can be leveraged to assist clinicians in identifying at-risk patients before they return to the hospital.

2 Literature Review

This project uses the publicly available **Diabetes 130-US hospitals dataset**, originally by Strack et al. (2014). The dataset includes 101766 admissions across 130 U.S. hospitals from 1999 to 2008, and 47 features containing demographic, clinical, and hospital encounter data for diabetic patients. Recent research has demonstrated that ensemble methods (e.g., random forests, gradient boosting) often achieve superior performance compared to linear models in this domain. These findings highlight the critical role of robust feature engineering and strategic model selection [1]. This report builds upon these insights by evaluating three distinct classifiers across the bias-variance spectrum: Logistic Regression, Random Forest, and Gradient Boosting.

3 Methods

3.1 Data Preprocessing and Feature Construction

Data Cleaning:

- Removed duplicate patient records (ensuring statistical independence)
- Standardised missing value representations (?, None, N/A → NaN)
- Dropped redundant/non-predictive features (`patient_nbr`, `encounter_id`, `payer_code`), and columns with >50% missing values (`weight`)
  - An exception was made for two features: A1C and Glucose Serum test result as background research showed that they were crucial in understanding the risk of readmission for diabetes patients [3]. The missing values may suggest that no tests were undertaken. Thus, these two features were kept and their missing values converted to ”Not Available’
- Removed 3 records with unknown gender
- Dropped expired/hospice cases (disposition IDs 11,12,16-20) as they cannot readmit

Feature Engineering:

- Medical Specialties: Grouped 71 specialties into 7 clinical categories (Diabetes, Internal, Surgical, etc). Applied target encoding with 5-fold cross-validation to prevent data leakage.
- Diagnosis Codes: Mapped ICD-9 codes to 8 clinical categories (circulatory, respiratory, diabetes, etc) to reduce dimensionality [3]. One-hot encoded with the first category dropped to reduce multicollinearity. Applied a similar mapping and encoding to discharge disposition
- Medication Features: Selected 6 medications with > 0.1 correlation to diabetes treatment. Created aggregate features including total dosage adjustments across medications, an insulin change flag, and a medication change ratio.
- Novel Derived Features: Health index (inverse of total visits), severity score (weighted sum of procedures and medications), visit-to-diagnosis ratio, and medical complexity metric.

- Log Transformation: Applied log transformation to length of stay and procedure counts to reduce skew and stabilise variance.
- Encoding:
  - One-hot encoding: gender, race (Caucasian, African American, Other), admission type (Emergency, Scheduled, Other), discharge disposition.
  - Ordinal encoding: medication changes (No, Steady, Up, Down).
  - Numeric conversion: age brackets converted to midpoint values.
- **Target Variable:** The target variable mapped readmitted patients within 30 days to `1`, and readmitted patients after 30 days and non-readmitted patients to `0`.

The final dataset comprises 34127 instances and 89 features, where each instance represents the hospital admission of a unique patient diagnosed with diabetes, which does not result in patient death or discharge to a hospice. The class is highly imbalanced with 8.77% in class `1` and 91.23% in class `0`.

3.2 Algorithms

The following three algorithms were evaluated after taking into consideration their varying complexity and overall coverage of the bias-variance spectrum.This selection also reflects a balance between predictive performance and interpretability, which is especially important in healthcare applications.

Logistic Regression (High Bias / Low Variance): A linear model that offers interpretable coefficients, making it well-suited for healthcare domains where interpretability is crucial. Logistic regression assumes linear relationships between predictors and the log-odds of the outcome, resulting in high bias but low variance.

Random Forest (Moderate Bias / Moderate Variance): An ensemble method that aggregates predictions from multiple decision trees. It captures non-linear feature interactions, handles both categorical and numerical variables effectively, and provides feature importance rankings. Its averaging mechanism helps reduce overfitting, offering a balance between model complexity and robustness.

Gradient Boosting (Low Bias / High Variance): A sequential ensemble method that builds trees iteratively, each one correcting the errors of its predecessor.

Models Considered but Not Selected:

- **Support Vector Machine (SVM):** While SVMs are effective in high-dimensional spaces and can model complex non-linear decision boundaries, they are computationally intensive for large datasets. Additionally, PCA analysis indicated that the data is not linearly separable, further limiting SVM performance.
- **Naïve Bayes:** Excluded due to its strong independence assumptions, which are unrealistic for this dataset containing correlated features. It also tends to perform poorly on mixed-type medical data involving both categorical and continuous variables.
- **Neural Networks:** Not chosen due to the relatively small feature set and limited dataset size, which increases the risk of overfitting. Additionally, neural networks are less interpretable, which is a drawback in medical applications requiring transparency.

3.3 Cross Validation

5-fold cross-validation (CV) was uniformly applied across all models. Each fold preserved class balance using stratified splitting. A nested CV approach was taken to ensure the tuning decisions did not leak into final evaluation metrics, yielding reliable and unbiased performance estimates. This included 2 loops: Outer loop (5 fold) for model evaluation and Inner loop (3 fold) for hyperparameter tuning using grid search Nested CV

was chosen because tuning hyperparameters on the same data used for model evaluation can cause data leakage and overly optimistic results. By isolating hyperparameter tuning (inner loop) from model testing (outer loop), data leakage is prevented, ensuring unbiased performance metrics. The 5-fold stratified splitting maintains class distribution across folds, essential for the class imbalance in this dataset, while balancing computational efficiency and model robustness. Feature normalisation was applied within the CV pipeline to ensure consistent scaling.

3.4 Hyperparameter Tuning

The following hyperparameters were tuned for each model:

- Logistic Regression: type of regularisation (`penalty` e.g., `l1`, `l2`), inverse of regularisation strength(`C`)
- Random Forest: number of trees (`n_estimators`) and depth of trees (`max_depth`)

- Gradient Boosting: number of boosting stages (`n_estimators`), step size shrinkage used to prevent overfitting (`learning_rate`)

Grid search was used within the inner cross-validation loop, optimising auc score to account for class imbalance. Hyperparameter ranges were chosen after several iterations to ensure adequate parameter coverage. After tuning, the hyperparameters underlined in Table 1 were chosen.

Model	**Hyperparameters**	**Analysis**

Logistic Regression	C = [0.1, **1**, 10]
	penalty = [**l1**, l2]
		C = 1 offers balanced regularisation, while the l1 penalty promotes sparsity by selecting the most predictive features from the high-dimensional dataset.

Random Forest	n_estimators = [100, **200**, 300]
	max_depth = [2, **5**, 10]
		Relatively shallow trees capture key patterns without overfitting, suggesting complex interactions were not beneficial.

Gradient Boosting	learning_rate = [0.01, **0.05**, 0.1]
	n_estimators = [50, **100**, 150]
		A low learning rate (0.05) ensures gradual learning, and a moderate number of estimators (100) helps avoid overfitting while maintaining performance.

Table 1: Optimal hyperparameters selected for each model and their interpretative analysis

3.5 Results and Discussion

3.5.1 Experimental Results

The models were evaluated based on 4 performance metrics: accuracy, precision, recall and ROC-AUC. A confusion matrix was created for each model, to highlight the classification performance across the three models (Figure 1)

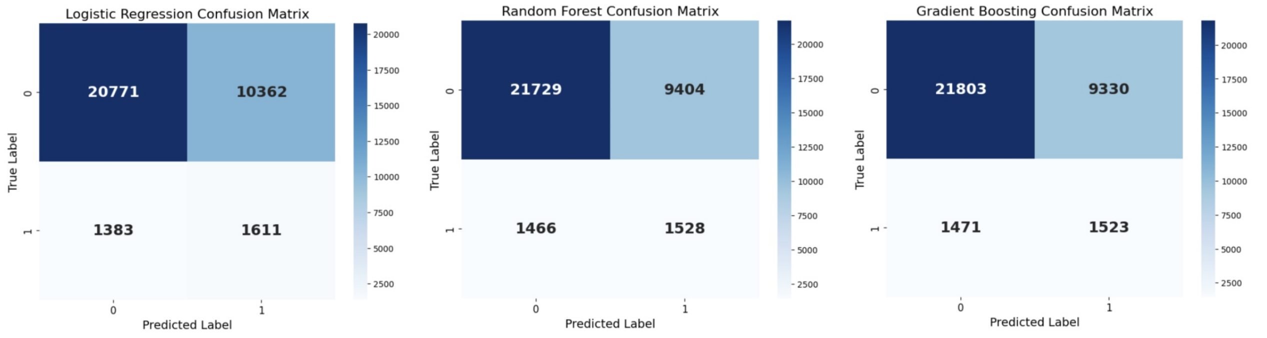

Figure 1: Confusion Matrix

Figure 1 reveals that all models had a high number of false positive rates, incorrectly classifying a high proportion of patients as readmitted. This can be attributed to the extreme class imbalance in the dataset, where very few patients are actually readmitted, resulting in the model struggling to learn what makes a true positive. Gradient Boosting achieved the highest number of true positives (21,803), indicating superior sensitivity in identifying actual readmissions, closely followed by Random Forest (21,729). Gradient Boosting also maintained the lowest false positive rate among the three models (9330) . For true negatives, Logistic Regression outperformed the others (1611), however all three models had marginal differences. These results suggest that while Gradient Boosting offers the best balance between sensitivity and precision, Logistic Regression may be preferable in scenarios where minimising false positives is critical. In terms of false positives, Logistic Regression yielded the lowest count (10,362), suggesting worst precision. Gradient Boosting and Random Forest showed lower values , however, were still overall reasonably high.

Table 2 summarises the key metrics for each model, averaged across outer folds. Gradient Boosting demonstrated the highest overall accuracy, marginally outperforming Random Forest. Logistic Regression, while less accurate than the other models, had the highest recall (0.5381). This high-bias, low-variance characteristic of Logistic Regression models is well-suited for clinical settings, where failing to detect true re-admissions is more costly

Metric	**Logistic Regression**	**Random Forest**	**Gradient Boosting**
Accuracy	0.6558	0.6815	0.6835
Recall	0.5381	0.5104	0.5087
F1 Score	0.2153	0.2194	0.2199
ROC AUC	0.6433	0.6434	0.6436

Table 2: Key classification metrics across 3 models

than false alarms. All three models had low F1 scores (0.2), which indicates a poor balance between precision and recall. This is attributed to the significant class imbalance and limited predictive features

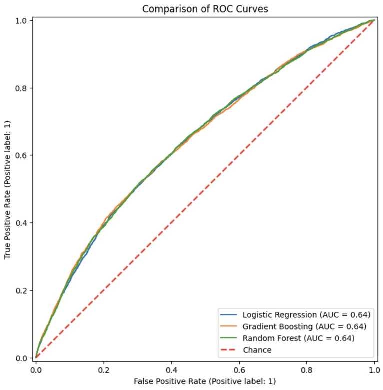

Figure 2: Comparison of ROC Curves across 3 models

Figure 2 shows that all three methods have a relatively modest performance with ROC AUC scores clustering around 0.64, indicating that overall, the models predict approximately 1.5 times better than randomly selecting patients. To confirm this, I ran mutual information scores across features to identify how informative they were. The results showed that the 'best' features were still close to zero, indicating a weak predictive relationship.

4 Conclusion

This study demonstrates that machine learning models have a moderate ability to predict hospital readmission for diabetes patients. Overall, the results suggest a trade-off between sensitivity and precision where Gradient Boosting offers a balanced performance, making it the most suitable model for early intervention strategies, while Logistic Regression could be favoured in contexts where false alarms must be minimised and interoperability is preferred. Moving forward, incorporating stronger, more targeted features such as temporal data, socio-economic factors and patient behavioural patterns to further enhance predictive power.

References

[1] Yu-Wei Lin, Yuqian Zhou, Faraz Faghri, Michael J Shaw, and Roy H Campbell. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PloS one*, 14(7):e0218942, 2019.

[2] Jade Gek Sang Soh, Wai Pong Wong, Amartya Mukhopadhyay, Swee Chye Quek, and Bee Choo Tai. Predictors of 30-day unplanned hospital readmission among adult patients with diabetes mellitus: a systematic review with meta-analysis. *BMJ open diabetes research & care*, 8(1), 2020.

[3] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014(1):781670, 2014.