Instructor - Prof. Anirban Dasgupta

Predicting patient-health

Ayush Anand 19110074
Bhoomika Mandloi 19110076
Ramireddy Lakshmi Nageswari 19110097

Overview

- Problem Description
- Performance Evaluation Metric
- Dataset Analysis
- Dataset Preprocessing
- Different Techniques
- Results
- Challenges
- Conclusion
- References

Problem Description

- WiDS Datathon 2021
- Maintaining and inferring important data about the patients' health at hospitals is a matter of utmost importance.
- Objective is to create a model to determine whether a patient admitted to an ICU has been diagnosed with a particular type of diabetes.
- Diabetes mellitus.
- Knowledge about chronic conditions such as diabetes can inform clinical decisions about patient care and ultimately improve patient outcomes.

Performance Evaluation Metric

- For visualizing performance of multi-class classification problem, we use the AUC(Area Under The Curve) ROC (Receiver Operating Characteristics).
- Area Under the Receiver Operating Characteristics (AUROC)
- AUROC curve is a evaluation metric for the classification problems at various threshold settings.
- AUC represents the degree of separability & ROC is a probability curve.
- Higher the AUC, the better the model is at predicting between patients having a disease and no disease.

Dataset Analysis

- Data is obtained from MIT's GOSSIS (Global Open Source Severity of Illness Score)
- Dataset can be accessed from kaggle.
- There are 3 files : TrainingWiDS2021.csv (the training data), UnlabeledWiDS2021.csv (the unlabeled data - data without diabetes_mellitus provided), DataDictionaryWiDS2021.csv (supplemental information about the data)
- Dataset consists of 180 features.
- 7 categories of features in the data.

- Lot of missing values in the feature set.
- Some features are missing in pairs.
- Example - number of missing values for d1_glucose_max and d1_glucose_min is the same.
- Lot of common values in some apache and vital columns.
- Dataset is imbalanced, where around 23% of the patients have Diabetes Mellitus.
- Sample Submission file in the correct format and Solution Template with list of all the rows (and encounters) are provided.

Data Preprocessing

- Start with an examination of how imbalanced the data is.
- Imbalanced data with 102006 data points belonging to the class zero(not having diabetes) and the rest belonging to the class one.
- Drop features which do not contribute much to the prediction we aim to make.
- Training on many features wouldn’t be efficient.
- Certain features consist of many missing values.
- Analyze what are the categorical and numerical features.
- Pick top 10 features obtained after fitting random forest classifier on the data.

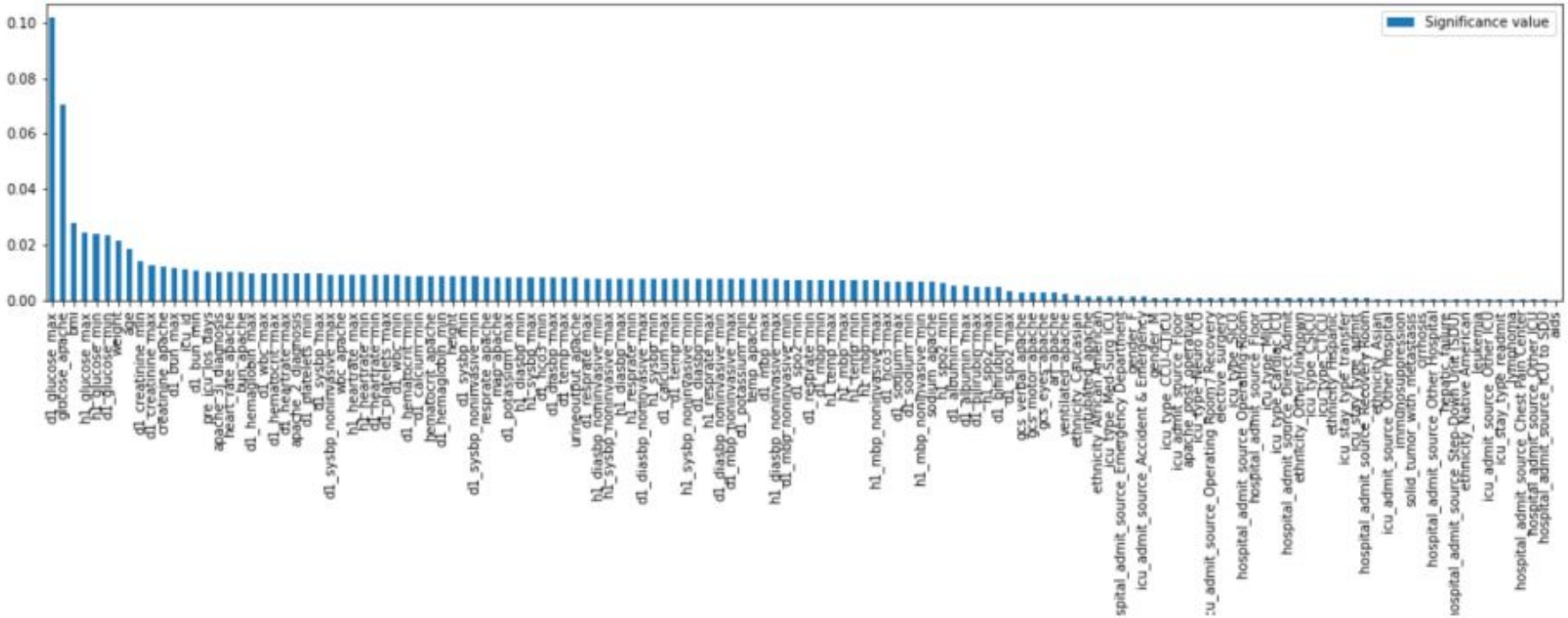


Figure 1 Relative Feature importances obtained after fitting random Forest Classifier on the data

Different Techniques

We give a brief description of the classification models used by us in the prediction task.

Random Forest

- An ensemble of many decision trees
- Can act as either classifier or regressor
- In case of regressor it takes the output as the mean of all the decision trees
- Classifier takes the majority output from the decision trees
- For measuring the loss either Gini index or Cross entropy loss is used
- Helps to deal with the high variance in decision trees.
- Better accuracy due to the combined result of multiple decision trees.

Adaptive Boosting (AdaBoost)

- AdaBoost classifier combines weak classifier algorithm to form strong classifier.
- Combines the predictions from short one-level decision trees(decision stumps).
- Choose the training set for each new classifier that you train based on the results of the previous classifier.
- Determines how much weight should be given to each classifier's proposed answer when combining the results.
- Combining the predictions made by all decision trees in the ensemble.

Gradient boosting

- Powerful ensemble machine learning algorithm.
- Greedy algorithm and can overfit a training dataset quickly.
- Each predictor tries to improve on its predecessor by reducing the errors.
- Fits a new predictor to the residual errors made by the previous predictor.
- Used when we want to decrease the Bias error.
- In classification problems, the cost function is Log-Loss.
- Ensemble algorithm that fits boosted decision trees by minimizing an error gradient.

eXtreme Gradient Boosting (XGBoost)

- Decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework
- Execution Speed
- Model Performance
- Dominates structured or tabular datasets on classification and regression predictive modeling problems.
- Automatic handle missing value
- Interactive feature analysis
- Winning model for several kaggle competitions
- Efficiency
    - Can be run on a cluster
    - Automatic parallel computation on a single machine

Category Gradient Boosting (CatBoost)

- Implements binary decision trees where same features are used for making splits on both sides of the tree.
- Decreasing the features in a level decreases the prediction time
- Automatically deals with categorical input variables
- Default parameters are a better starting point than in other Gradient boosting decision tree algorithms
- Missing values need to be processed separately

Results

Table 1 AUROC scores of various models

Model	AUROC Score
Random Forest with default parameters	0.8169
Random Forest with better parameters	0.8489
Gradient boosting classifier	0.8273
Ensemble of Random Forest and Gradient boost	0.8399
Adaboost classifier	0.8156
Ensemble of Adaboost+RandomForest+Gradient boost	0.8398
XGBoost average over 5 Stratified folds	0.7143
Catboost classifier	0.8474

XGBoost K-Stratified results

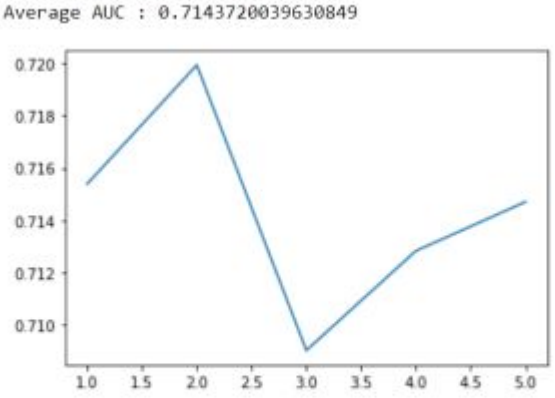
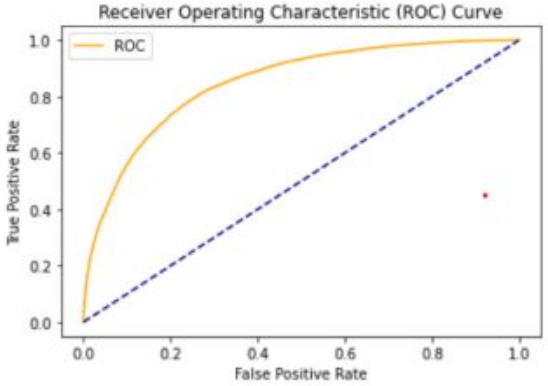
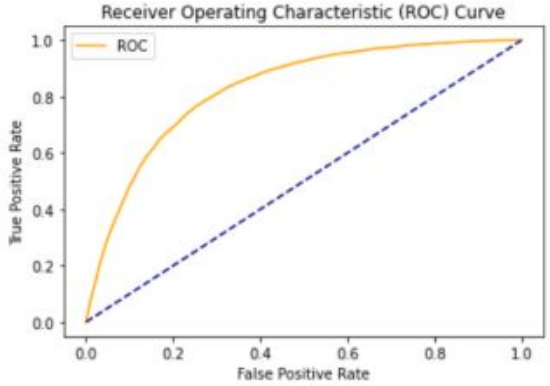


Figure 2 XGBoost AUROC scores over the 5 folds



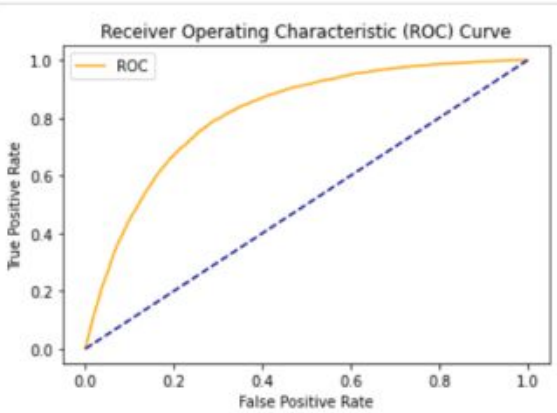
AUC : 0.8489250203531864

Figure 3 ROC of Random forest



AUC : 0.8273413661117092

Figure 4 ROC of Gradient boost



AUC : 0.8156722910741385

Figure 5 ROC of AdaBoost

Challenges

- The dataset contains a lot of features and missing values which need to be handled
- Would be better if could train on the cloud; the local machine had limited processing power
- Could not do hyperparameter tuning for all the models
- Could expect better performance if could train on more features

Conclusion

- Implemented and tested the performance of several machine learning models on patient health data to predict if a patient has diabetes or not.
- Used the evaluation metric AUROC score for comparison.
- Random Forest, Adaboost and Gradient Boost all give almost the same auroc score of around 0.82.
- XGBoost slightly underperforms as compared to the other models with an AUROC score of around 0.71.
- CatBoost performs slightly better than the other models with an AUROC score of around 0.84.
- Combine the predictions of some of these models and return their average as the final prediction.

References

1. https://www.kaggle.com/c/widsdatathon2021/overview
2. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
3. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/
4. https://medium.com/almabetter/catboost-the-fastest-algorithm-c21d44f8b990
5. https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/
6. https://towardsdatascience.com/understanding-adaboost-2f94f22d5bfe
7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6942113/
8. https://github.com/Precillieo/WIDS-Datathon-2021
9. https://github.com/ahanagangopadhyay/wids-datathon-2021-diabetes-prediction
10. https://github.com/KanchanSatpute/WiDS-Datathon-2021/blob/main/WiDS-Datathon.ipynb

Thank You !