Predicting 30 Day Hospital Readmissions of Diabetic Patients

Homayoun Sadri

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

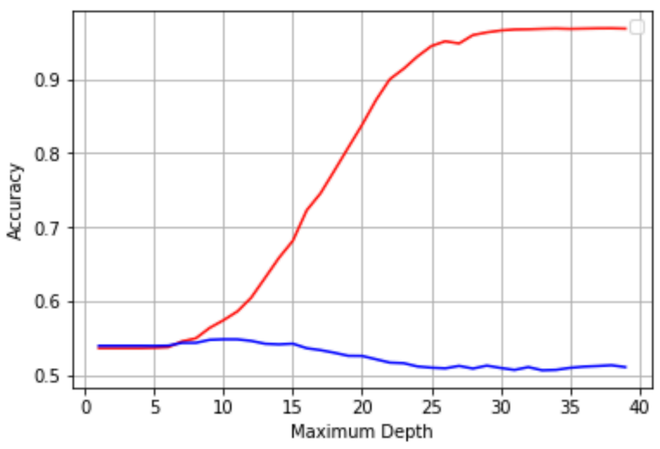
Introduction

- Caring for hyperglycemia patients has a significant effect both on morbidity and morality.
- Thus formal protocols has been put in place in intensive care units(ICUs)
- Here, an analysis of a large clinical database is performed to examine historical patterns of diabetes care in patients with diabetes admitted to a US hospital and to inform future directions which might lead to improvements in patient safety.

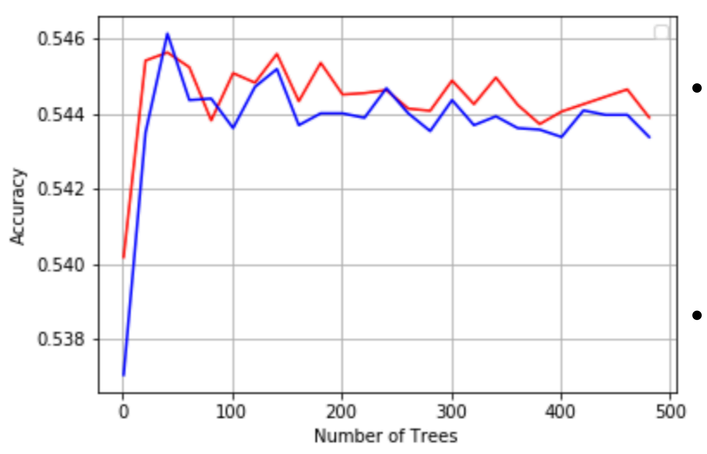
Data

- The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks.
- There are 101,766 instances of data.
- 14 features of the initial 50 features have been used for the classification task:
  - Age
  - Race
  - Admission Type
  - Discharge Disposition
  - Time in Hospital
  - Number of Lab Procedures
  - Number of Procedures
  - Number of Medications
  - Glucose Serum Test Result
  - A1c Test Result
  - Change in Medication
  - Diabetes Medications
- There are three classes:
  - No readmission, <30 day readmission, >30 day readmission

Random Forrest



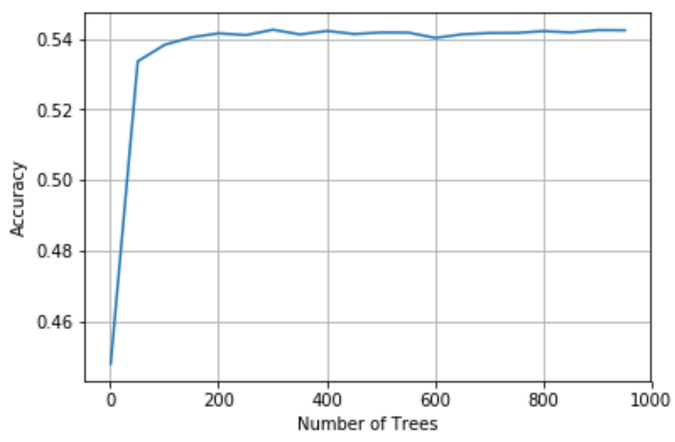
- Keeping the number of trees constant and looking for optimal maximum depth:
- After a maximum depth of ~12, the cross-validation accuracy drops.



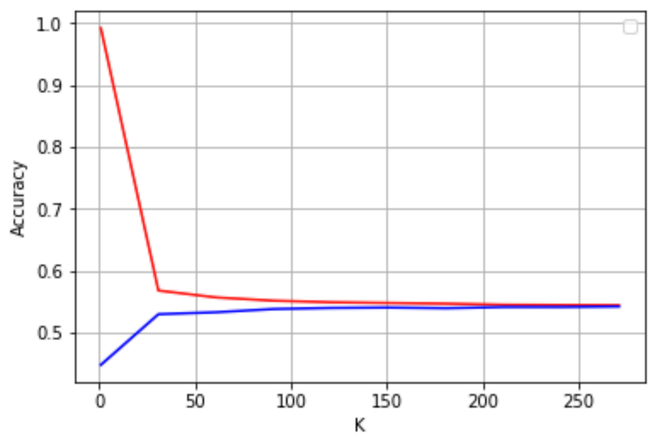
- Changing the number of trees and keeping maximum depth constant to find optimal number of trees:
- ~50 is a good number of tree.

- The baseline accuracy for number of threes = 10 and maximum depth = 5, is 0.5449
- The optimum accuracy accuracy for number of threes = 50 and maximum depth = 12, is 0.5553
- Random Forest does not show much improvement with respect to its baseline.

KNN



- We can see that increasing the number of trees after 300 doesn't increase the accuracy of classification.



- Comparison of training and cross-validation accuracy also shows that the accuracy is highest at around 300 for number of trees

- The baseline accuracy for k =5 is 0.5084
- The optimum accuracy accuracy for k = 300, is 0.5459
- KNN shows a better improvement, with respect to baseline, compared to Random Forest.

Future work

- The above accuracies are just slightly better than chance; Further analysis is needed for better classification.
- One example of this could be feature engineering:
- The graph on the right shows importance of each feature in correctly classifying the data instances.
- 5 features are more important than the others, with the top 2 -number of lab procedures, number of medication- being much more important