Project Part 3 Report

You Wang, yw6127@nyu.edu, Luoyao Chen, lc4866@nyu.edu

1 EDA Physical Design

1.1 Documentation of Designing Techniques

Indexing for Tables in DataBase

After the creation of MS Azure MySQL server, the database is managed and optimized by the local MySQL workbench, which is connected to Azure DataBase. In order to facilitate the searching process, each table has been indexed. For instance, (name, address) together form a primary key for table *account_admin*, and we use the combination of the pair as the index for the table. On the other hand, *contractbenefit* has the contract_number being the identifying value, which we set as the primary key. Screenshot shows the columns as well as indexes related to table *account_admin* and *contractbenefit*.

List Partitioning for Tables in DataBase

After creating tables, we partitioned the tables. Since most groups distinguish from another based on discrete features, we partitioned the tables by list partitioning so that each partition does not overlap with each other. For instance, in order to partition the *account* table, we used command like follows:

PARTITION BY LIST(Location) (
PARTITION pEast VALUES IN (NY, MA),
PARTITION pWest VALUES IN (CA, WA),

PARTITION pNorth VALUES IN (AK, ND, MN),
PARTITION pSouth VALUES IN (FL, TX));

Materialized Views

Compared with Views of tables, which only stores the queries command, materialized views store also the data returned by the views in order to facilitate complicated query operations. The materialized views can be created with the table, and updated during usage when needed.

1.2 Explanation of a Business Case using the DataBase

The tables shown in the screenshot were designed to be used by insurance companies. For instance, a customer can create an account by establishing an insurance plan contract with the company, (creates records in *Customer, ContractBenefit*, and *Account* tables), furthermore, the associate information who is responsible for the customer is also documented in the *Associate* table. In addition, customer's chronic diseases/health conditions are also investigated, including whether they have diabetes/heart disease/asthma. Stored in *cdcasthma, cdcdiabetes*, and *heart* tables. Such information can subsequently be analyzed by machine learning models to obtain the risk of a customer, i.e. whether customer might require more coverage in the future, potentially caused by being readmitted to hospital.

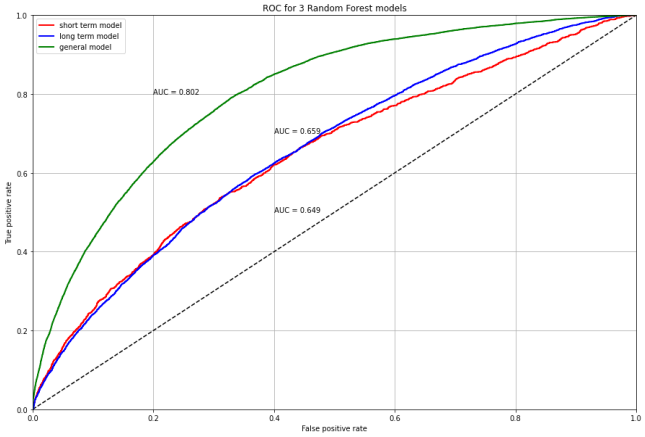
2 Machine Learning

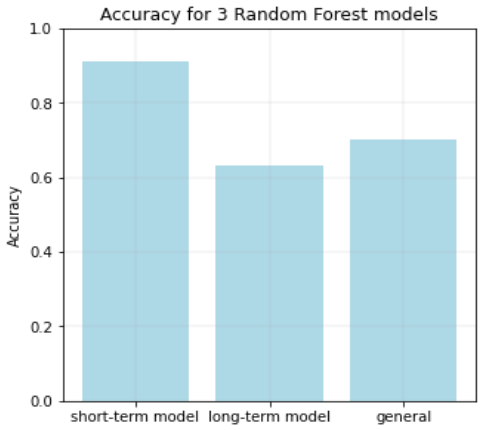
In the machine learning part of our project, we developed predictive models to identify diabetic patients who are likely to be readmitted to hospital in the future. We built and trained our model using the *Diabetes 130-US hospitals for years 1999-2008* dataset UCI Machine Learning Repository. We used Random Forest and Logistic Regression to predict which hospitalized diabetic patients will be readmitted in the future, and we used accuracy and AUC score to evaluate our model.

2.1 Predictive Models

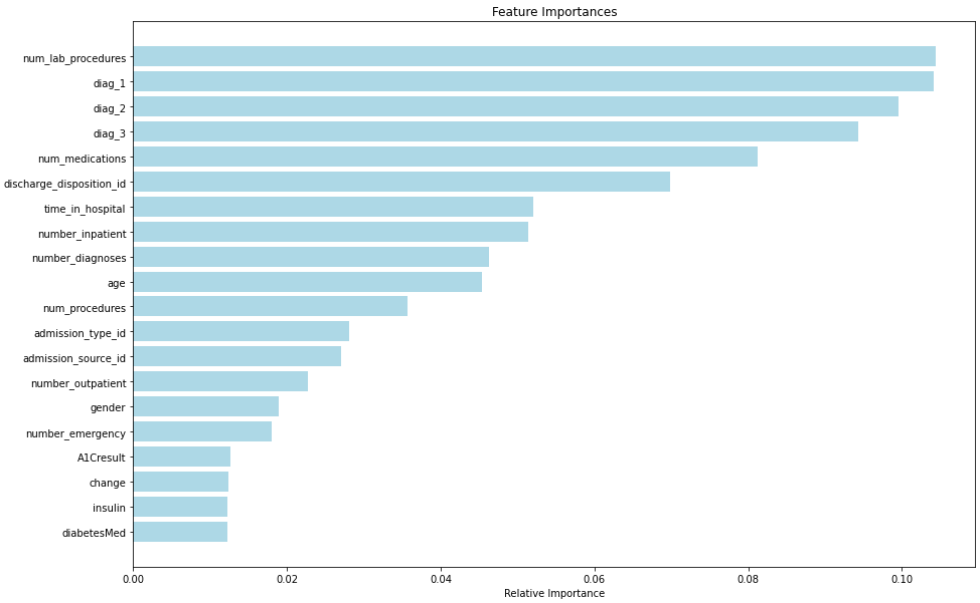
The dataset provided labels that indicate whether a patient is readmitted in the next 30 days, after 30 days, or not readmitted at all. We built 3 types of model using Random Forest: short term readmission prediction model, long term readmission prediction model, and general prediction model. For short term prediction, we achieved an accuracy of 0.91, and 0.65 AUC score. For long term prediction, we achieved an accuracy of 0.64, and 0.66 AUC score. Finally, for the general readmission prediction, we achieved accuracy of 0.62 and 0.8 AUC score. We can see that the most accurate model is the short-term model, however, the general model has the highest AUC score. In this case, we believe AUC is a better metric for model evaluation because true/false positive rate

matters a lot in medical services. Therefore, we believe the general model we created outperforms the other two models.

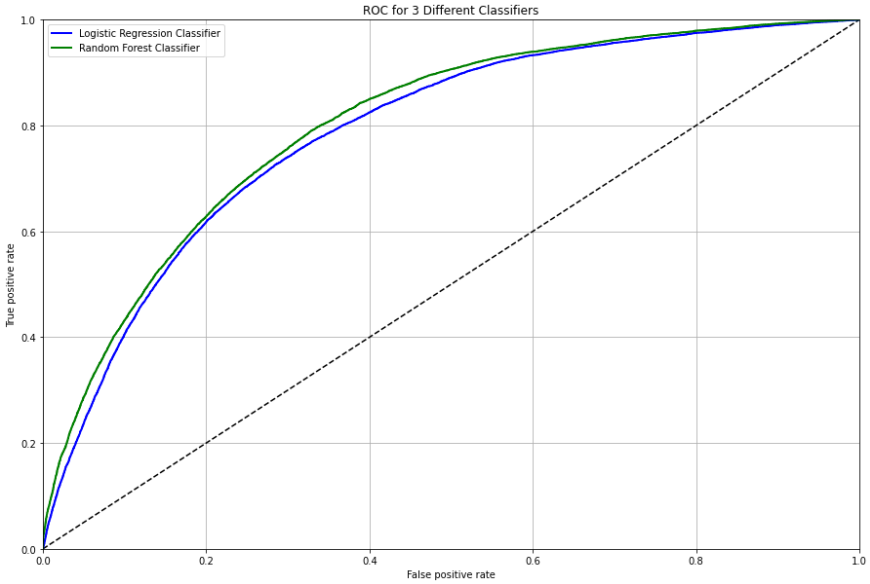




According to the general prediction model, we also find out the 20 most important features for predicting the readmission, using the feature_importances_ attribute of the random forest classifier. The most important features are the number of lab tests performed during the encounter. Also, the type of the diagnosis and the number of medications are also important predictors for readmission prediction.



In this part, we have built 3 random forest classifiers. We try to see whether a different model would yield better results. For simplicity and more general results, we only consider the multi-class models that classify the patients into 3 classes. Moreover, as we have discussed above, we are going to use AUC scores for model evaluation in the following models. By using the logistic regression model in sklearn, we run the logistic regression. The AUC value for logistic regression classifiers model equals to 0.79. We compare the logistic regression model with the random forest model we built and plot the ROC of each of these models. We can see that the random forest model yields the best results.



2.2 Combining the machine learning result with the database

In this part, we use the diabetes dataset to develop models that identify patients(or potential customers for insurance companies) who have higher risks. With this model, the insurance company can use the medical record of the customers and make predictions about their likelihood of being readmitted to hospital because of diabetes. If a customer has a higher probability of being readmitted, the insurance can adjust their policy provided for this customer. On the other hand, if a customer has a lower risk, the insurance company can provide lower prices on diabetes related insurance policies for this customer.

The random forest model not only provides predictions, but also yields the importance of the features for predicting the diabetes readmission. Finding the most useful predictor is useful because in reality, the insurance company can pay more attention to these features in the customers' information, use these features as flags for potential high risks, and also, anyone can make a rough prediction about whether a customer is going to be hospitalized again.