Predicting Hospital Readmissions for Diabetic Patients using Machine Learning

Abhishrut Khanna
IIIT Delhi

Bhupesh Singh Kainth
IIIT Delhi

Pankaj Yadav
IIIT Delhi

Abstract

The project is based on a binary classification problems wherein our aim is to predict whether a diabetic patient will be readmitted to the hospital or not based on the treatment received by the patient.The dataset contains 50 features and 1,01,766 instances representing the patient and hospital outcomes. At first,data preprocessing was performed to remove redundant/invalid values.The dataset was trained using Logistic Regression,SVM(both rbf and polynomial kernel),Random forest classifier,MLP(3 hidden layers and 4 hidden layers) and K-Nearest Neighbour(KNN) method.The predicted accuracies are 0.6283(Logistic Regression),0.6315(SVM with 'poly' kernel),0.6153(Random forest),0.8317(MLP(3 hidden layers) with sigmoid) and 0.9169(KNN with 5 neighbours).The MLP and KNN models performed significantly better than the other three model.KNN proved to be the best model with accuracy close to the state of the art accuracy of 95 %.

1. Introduction

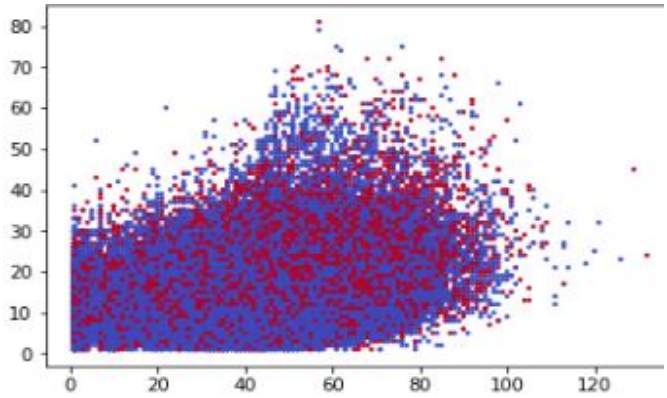
In our project, our training model is predicting whether a diabetic patient will have to be readmitted to the hospital in the future or not based on treatment received by the patient. It is important and useful to know this as it will help us know which treatments provide better results for the patients.
The dataset that we are going to use is obtained from https://www.kaggle.com/brandao/diabetes.The output condition of readmission would be categorized into 3 parts: No readmission, readmission in greater than 30 days and readmission in less than 30 days. This criterion will help us in identifying the kind of treatment that a patient has received. The reason for us to try and solve this problem is we want to reduce the number of diabetic patients getting readmitted to the hospitals, based on the previous medications and results we can determine which medication is best based on the extent of diabetes.

2. Related Work

The concept of predicting early readmission is of wide importance in USA. There have been few researches that used machine learning algorithms to predict readmission of a patient. Beata Strack(2014) used logistic regression models to assess the impact of features on readmission. In this,there was not enough emphasis on utilizing the logistic regression model or other machine learning algorithms.C. Chopra, S. Sinha, S. Jaroli, A. Shukla, and S. Maheshwari(2017) used Recurrent Neural Network and produced an accuracy of 81.12%. Though the model described used 33 of the 55 available features, including a few of which were redundant.A work by C.-Y.Lin, H.S.Singh, R.Kar and U. Raza(2018) conducted at Berkeley produced a better performance in comparison to other model.
The results produced used AUC(Area-under-Curve) with an accuracy of 94%.A.Hammoudeh, G. Al-Naymat, I. Ghannam and N. Obied(2018) used Convolutional Neural Network to produce the state of the art performance. The state of the art model obtained the c-statistics performance of 95 %.

3. Methodology

Since the dataset for large, we first performed preprocessing steps on the dataset. We observed data in every column and row by considering their value distribution and by plotting them as bar graphs and then used the feature extraction techniques to remove ambiguous and uneven values. This reduced the dataset to 99492 rows and 22 features. After that, we observed that the resulting dataset was non-linear and noisy and thus can't be separated linearly. As there were a lot of string values, we designated "Yes" as 1 and "No" as 0.



We then trained our different models on the processed dataset. We first trained our dataset using Logistic Regression wherein we first divided our dataset into training and testing dataset using train_test_split. We used sklearn's linear_model to get the accuracy score, confusion matrix, recall score and precision value on the testing data.

Since Logistic Regression is a naive method, we then used SVM to train our dataset. The performance of our model was tested over 5 folds. Accuracy score, confusion matrix, recall, precision and f1 score was recorded for each kernel(rbf and polynomial(degree 2)) for each of the 5 folds using sklearn's in-built methods. We used both One-vs-rest(ovr) and One-vs-all(ovo) decision functions.After obtaining the best fold and decision function, the same was applied to the testing data to obtain accuracy score.

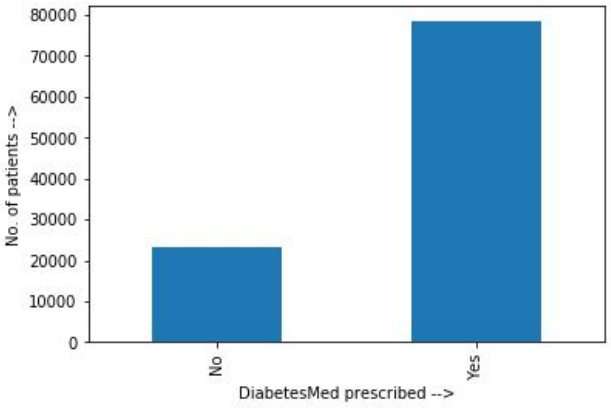
We trained our dataset using Multilayer Perceptron(MLP). This was done using MLPClassifier and two different hidden layer sizes and numbers. Firstly, the number of hidden layers was 3 and the number of nodes in each layer was (256,128,64). The accuracy score was predicted for tanh, relu, logistic(sigmoid) and linear activation function. The same model was then trained using the number of hidden layers as 4 and the number of nodes in each layer as (512,256,128,64). Both of the above models used stochastic gradient descent(sgd) as the solver.

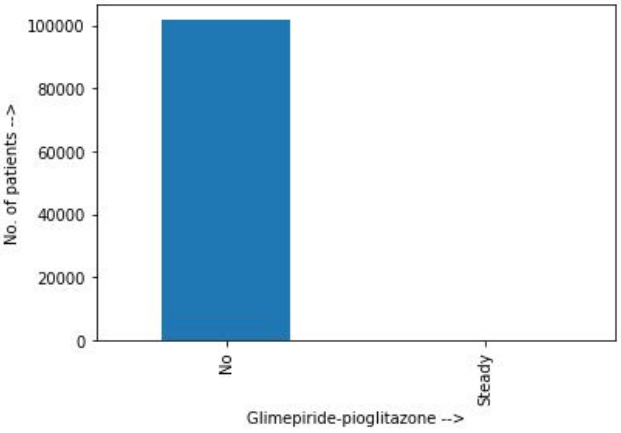
We also trained our dataset using the RandomForestClassifier. We used 80 % of the data set for training and 20% for testing. First, we kept the n_estimators (number of trees) as 100 and a random state of 50 while keeping the rest of the parameters as default. Then kept on increasing the value of n_estimator till 1000.

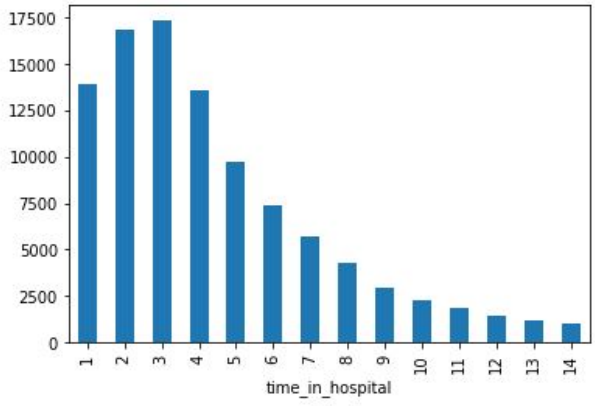
Since our dataset was large, we tried K nearest neighbours (KNN) model too with our dataset. It considers the closest K neighbours for a given test data and predicts to which class it belongs by calculating the probability of positive class divided by K. Since KNN is very sensitive to the value of K, we tried different values of K for an efficient result. We used K = 5,50,100,200 and obtained accuracies corresponding to each value of K.

4. Result

The first step that we performed on our dataset was pre-processing.For this,we kept a threshold value of 0.95(95%). This value signifies that if we have a redundant/invalid value occuring more than 95% in a particular column(feature), then we drop that feature.







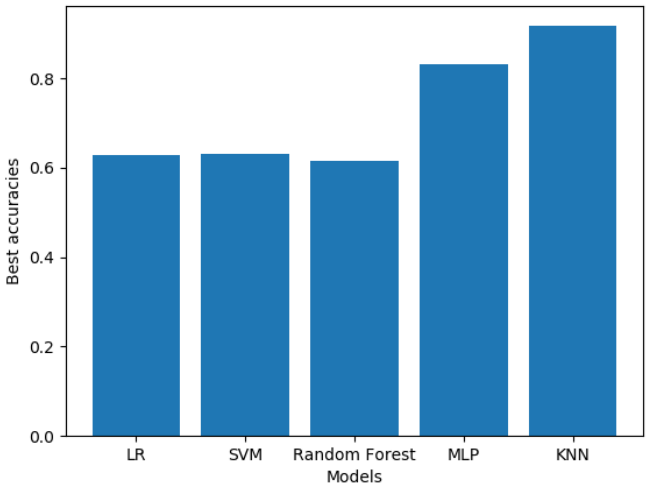
The Logistic Regression model gave an accuracy of 0.6283,recall score of 0.5665 and a precision score of 0.628.The confusion matrix was $\begin{Bmatrix} 7928 & 2825 \\ 4711 & 4445 \end{Bmatrix}$ .Then we ran SVM models on the dataset to obtain the best accuracy as 0.6315, best kernel as 'poly'(degree 2) and best deci-

sion function shape as 'ovr'.The performance observed on the same dataset with rbf kernel was significantly lower.We observed that though the accuracy of the LR model was slightly less than that of SVM but LR was a lot faster than SVM.

The Random forest classifier gave an accuracy of 0.6153 with the number of estimators as 1000. It was also noted that increasing the number of estimators did not change the accuracy. The model resulted in the same accuracy even after increasing the number of estimators by 10 folds.

The MLP Classifier with 3 hidden layers and number of hidden layer nodes as (256,128,64).The best accuracy obtained was 0.8317(83.17%) with sigmoid activation function.The accuracies obtained for other activation functions were 0.7590(tanh) and 0.7070(relu).The above values were obtained with learning rate as 0.1 and number of iterations as 100.Then,the MLP Classifier with 4 hidden layers and the number of hidden layer nodes as (512,256,128,64) was used.The best accuracy obtained was 0.708(70.8%) with sigmoid activation function.The other accuracies obtained were 0.7071(relu) and 0.7070(tanh). We can see from the above that the accuracy for the model with 3 hidden layers was better than that for 4 hidden layers.

The K-Nearest Neighbour(KNN) model with the number of neighbours in the range [5,50,100,200] gave accuracy between 0.8205 to 0.9169. The highest accuracy was obtained for the number of neighbours as 5.The confusion matrix was $\begin{Bmatrix} 13557 & 494 \\ 1159 & 4689 \end{Bmatrix}$ .Increasing the number of neighbours decreased the accuracy.



5. Conclusion

The best results on our dataset were obtained using KNN(0.9169) with the number of neighbours as 5 and using MLP Classifier(0.8317) with sigmoid activation function. These accuracies were significant improvements over the previously trained Random Forest classifier, SVM and Logistic regression models.The classification reports for each trained model helped us in improving our accuracy and model selection.Our trained model accuracies were close to that specified as the state of the art model(0.95).