Name: Sachinkumar Uttamrao Ranveer

Email address: ranveersachin143@gmail.com

Contact number: 7666968511

Anydesk address:

Years of Work Experience: **13 Months**

Date: **14th Oct 2020**

Self Case Study -1: ***Name of your Self Case Study***
Diabetes Dataset predicting whether a patient will readmit within 30 days or not.

"After you have completed the document, please submit it in the classroom in the pdf format."

Please check this video before you get started:
https://www.youtube.com/watch?time_continue=1&v=LBGU1_JO3kg

Overview

*** Write an overview of the case study that you are working on. ***(MINIMUM 200 words)*** ***

Introduction :

Diabetes is a chronic disease where a person suffers from an extended level of blood glucose in the body. Diabetes is affected by height, race, gender, age but a major reason is considered to be a sugar concentration. The present analysis of a large clinical database was undertaken 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. Reducing early hospital readmissions is a policy priority aimed at improving healthcare quality. In this case study we will see how machine learning can help us solve the problems caused due to readmission.

Business Problem :

It is estimated that 9.3% of the population in the United States have diabetes , 28% of which are undiagnosed. The 30-day readmission rate of diabetic patients is 14.4 to 22.7 % . Estimates of readmission rates beyond 30 days after hospital discharge are even higher, with over 26 % of diabetic patients being readmitted within 3 months and 30 % within 1 year.

Costs associated with the hospitalization of diabetic patients in the USA were $124 billion, of which an estimated $25 billion was attributable to 30-day readmissions assuming a 20 % readmission rate. Therefore, reducing 30-day readmissions of patients with diabetes has the potential to greatly reduce healthcare costs while simultaneously improving care.

Dataset overview and ML Formulation :

The Health Facts database (Cerner Corporation, Kansas City, MO), a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States. The actual database contain all the patients data but specifically I wanted to work with Diabetes related data so I have used the data that filtered using below mentioned five criteria and is available in UCI Machine Learning Repository link : https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008

1) It is an inpatient encounter (a hospital admission)
2) It is a “diabetic” encounter, that is, one during which any kind of diabetes was entered the system as a diagnosis.
3) The length of stay was at least 1 day and at most 14 days.
4) Laboratory tests were performed during the encounter.
5) Medications were administered during the encounter.

Almost 1 Lakh records were found to be satisfying the above five criterias and the dataset has 55 features like gender, weight, encounter_id, readmitted. Dataset can be used for Classification and Clustering tasks. I will be using it for predicting whether the patient will readmit within 30 days or not. That is a classification task.

Performance Metrics :

Our task is a classification problem so we can use performance metrics like precision, recall, Accuracy and F1-score.

1)Precision :

Precision is (TP/TP+FP) where TP True Positive and FP is False Positive We can think of precision as out of all points that are predicted as positive points by model how many of them are indeed positive points. Precision is a good measure when False positive cost is high which is the case here in our case study. For a patient which doesn't need readmission if our model predicts that the patient needs readmission that is False positive then the hospital will keep that patient in the hospital and that increases hospitalization cost.

2) Recall :

Recall is (TP/TP + FN) where TP is True Positive and FN is False Negative We can think of Recall as out of all points that are actually positive how many of them are predicted to be positive by model. Recall is used when False Negative cost is high that is indeed the case here. For the patient which needs readmission if the model predicts that it doesn't then the hospital will discharge him but the patient will eventually readmit again and that increases the cost.

3)F1-Score :

As from above we know that False Negative cost and False Positive cost both are important for us so it would be good if we have a measure which combines both. F1-score does the same for us; it combines Recall and Precision into single equation.

F1_score = (2 * Precision * Recall) / (Precision + Recall)

Research-Papers/Solutions/Architectures/Kernels

*** Mention the urls of existing research-papers/solutions/kernels on your problem statement and in your own words write a detailed summary for each one of them. If needed you can include images or explain with your own diagrams. it is mandatory to write a brief description about that paper. Without understanding of the resource please don't mention it***

1)https://towardsdatascience.com/predicting-hospital-readmission-for-patients-with-diabetes-using-scikit-learn-a2e359b15f0

This blog I found on towardsdatascience.com website. The author's name is Andrew Long. He did extraordinary data preprocessing. Few examples of data preprocessing steps I liked

1)The author managed to capture the problem Data Imbalance and dealt with it properly.

2)The author managed to deal with missing values in the dataset and replaced them with proper values

3)The author used feature binning techniques to deal with age like features.

I would use this already done data preprocessing to build my model. Though he hasn't done any data exploratory analysis I am pretty much satisfied with his work.

2) https://www.kaggle.com/iabhishekofficial/prediction-on-hospital-readmission

This is another blog that found to be helpful in solving this case study. The author's name is Abhishek Sharma. He did extraordinary data exploration and tuned the models properly to achieve very good accuracy. I liked the way he performed the univariate analysis. Univariate analysis done by the author actually tries to find the correlation between independent variable and dependent variable in other words how readmitted feature dependent and individual features like glucose, ethnicity, gender, time in hospital.

3) https://www.ijedr.org/papers/IJEDR1802080.pdf

From this paper I get to know more about how harmful diabetes disease is. According to the world health organization 382 Millions people have diabetes and by 2035 the numbers will double that is 587 Millions.

Different Types of Diabetes :

Type 1 Diabetes is called insulin-dependent diabetes mellitus (IDDM) or juvenile-onset diabetes.Type1 mostly occurs in young people who are below 30 years.

Type 2 Diabetes is called non-insulin-dependent diabetes mellitus (NIDDM) or adult-onset diabetes. In type 2 diabetes, the pancreas usually produces some insulin; the amount produced is not enough for the body's needs, or the body's cells are resistant to it.

Type 3 is called Gestational Diabetes is the third major form and occurs when pregnant women without a previous account of diabetes develop a high blood glucose level. It affects 4% of all pregnant women.

By studying the above types I get to know that “age and gender” play an important role in Diabetes so these features can be marked as important. The paper goes on implementing models like LR, Naive Bayes, SVM, ANN and concludes that SVM is not performing well. I am taking note of this point and will focus more on Algos other than SVM.

4)https://www.researchgate.net/publication/333015077_Feature_Engineering_FE_Tools_and_Techniques_for_Better_Classification_Performance

Feature engineering is the task of improving predictive modelling performance on a dataset by transforming its feature space. Feature engineering is usually conducted by a data scientist relying on her domain expertise and iterative trial and error and model evaluation.

Neural networks learn useful features automatically and have shown remarkable successes on video, image and speech data. However, in some domains feature engineering is still required. The features derived by neural networks are often not interpretable which is an important factor in domains like healthcare.

Following are the feature transformation that I will try to see if they can improve the performance of my models :

a) Mathematical : log function, square-root (both applied on the absolute of values), Square (Taking the square), tanh, sigmoid, round.

b) Statistical : isotonic regression, z score, normalization , frequency (count of how often a value occurs),

c) Mathematical Operations : sum, subtraction, multiplication and division

5) https://www.kaggle.com/aldrinl/interpretable-ml-for-diabetes-patient-readmission
https://www.analyticsvidhya.com/blog/2019/11/shapley-value-machine-learning-interpretability-game-theory/

Rather than using Machine Learning models as black box it would be great to have Machine Learning models which predicts results accurately and also explains how it arrived at such a conclusion. The models which have this ability are called “interpretable models”. In the case of the healthcare domain it is always good to have models which are interpretable.

I found this kernel to be helpful to get the interpretability of the models . The Author first discusses the models then uses libraries like LIME, SHAP etc to interpret each model . After done with model training I will definitely use three libraries to explain my models interpretability.

1) LIME (Local Interpretable Model-Agnostic Explanations) :

When a model predicts the output our main is to know why the model thinks the output should be whatever predicted.To achieve this LIME generates a new dataset consisting of permuted samples and the corresponding predictions of the black box model. On this new dataset LIME then trains an interpretable model. Interpretable models could be LR or DT.
Key idea here is ,We generate an explanation by approximating the underlying model by an interpretable one locally.

2)SHAP (SHapley Additive exPlanations) :

SHAP is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions.

The way it achieves the goal of explaining the black box is by calculating the shapley values. The higher the shapley value for a feature the more the importance it has and vice versa

Following are the steps to calculate importance of feature fx:

1)Get all subsets of features S that do not contain fx
2)Compute effect on our predictions of adding fx to all those subsets
3) Aggregate all contributions to compute the marginal contribution of the feature

3) ELI5 :

ELI5 is the algo which is responsible to make black box machine learning models interpretable. Using the following steps algo try to get weights to each feature according to its importance in decision making.

Take each feature at a time and do the following steps :

1)Shuffle values in the provided dataset
2)Generate predictions using the model on the modified dataset
3)Compute the decrease in accuracy vs before shuffling

If a model's accuracy happens to be not changing then it must be the case that feature is not important in decision making.

First Cut Approach

*** Explain in steps about how you want to approach this problem and the initial experiments that you want to do. ***(MINIMUM 200 words)*** ***

*** When you are doing the basic EDA and building the First Cut Approach you should not refer any blogs or papers ***

The given problem can be easily framed as a Machine Learning problem. We have given features and we want to predict whether patients will readmit within 30 days or not this is a classical Machine Learning classification problem.

I will first go for data exploratory analysis to get more insight about the dataset. Few data analysis that i would try are :

1)Outlier Analysis, Univariate Analysis, Bivariate Analysis

2)Balanced or Imbalanced dataset ?

3)Is there any data duplication ?

4)Missing data / check missing data can be a source of info or not.

5)Train, Test data has the same distribution or not ?

Next step in my approach will be to do feature engineering. I will try feature engineering techniques like binning, mathematical transform, feature orthogonality, domain specific features. There are a number of mathematical transforms like log, sqrt, square that we can apply on features. I will try to decide which mathematical transform is more effective by trial and error method.

Process of applying Machine learning in the healthcare industry must satisfy two conditions first one is model interpretability and second one is model accuracy. Model accuracy is important because health related decisions are sensitive. Any wrong decision might cause someone to lose his life. Model interpretability is also important because lets say for example when a model predicts that some personX has cancer but the model can't explain why it predicted yes for personX then it will be difficult for the doctor to believe such decisions and communicate the decision to personX (patient). I will use SHAPE, LIME, ELI5 to obtain the interpretability of the model.

Other authors have already explored methods like LR, Random Forest, Gradient Boosting. First I will try to create a few more features by analysing or studying more about diabetes related research papers. Once I have all the features I will try ML algos like SVM, Multilayer Perceptron, Naive Bayes, conv1D. I will also consider stacking these models or cascading them to see if that improves the performance of my model. I will also explore if I could improve the performance of already explored machine learning algos by other authors.

[Notes when you build your final notebook:

1. You should not train any model either it can be a ML model or DL model or Countvectorizer or even simple StandardScalar
2. You should not read train data files
3. The function1 takes only one argument “X” (a single data points i.e 1*d feature) and the inside the function you will preprocess data point similar to the process you did while you featurize your train data
    a. Ex: consider you are doing taxi demand prediction case study (problem definition: given a time and location predict the number of pickups that can happen)
    b. so in your final notebook, you need to pass only those two values
    c. def final(X):
    ```
            preprocess data i.e data cleaning, filling missing values etc
            compute features based on this X
            use pre trained model
            return predicted outputs
    final([time, location])
    ```
    d. in the instructions, we have mentioned two functions one with original values and one without it
    e. final([time, location])   # in this function you need to return the predictions, no need to compute the metric
    f. final(set of [time, location] values, corresponding Y values) # when you pass the Y values, we can compute the error metric(Y, y_predict)
4. After you have preprocessed the data point you will featurize it, with the help of trained vectorizers or methods you have followed for your train data
5. Assume this function is like you are productionizing the best model you have built, you need to measure the time for predicting and report the time. Make sure you keep the time as low as possible
6. Check this live session: https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/4148/hands-on-live-session-deploy-an-ml-model-using-apis-on-aws/5/module-5-feature-engineering-productionization-and-deployment-of-ml-models