Machine Learning Engineer Capstone Project

Masinde Mtesigwa

2023-03-01

Contents

Healthcare

EHR/EMR

Electronic Health records or Electronic Medical Records data is the data being collected when we see a doctor, pick up a prescription at the pharmacy, or even from a visit to the dentist.

This data is used for a variety of use-cases. From personalizing healthcare to discovering novel drugs and treatments to helping providers diagnose patients better and reduce medical errors.

Definition

The objective of the project is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.

About Dataset

Context

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Content

The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on(Aamna 2023).

Project Statement

Can you build a machine learning model to accurately predict whether or not the patients in the dataset have diabetes or not?

Benchmark model

The model going to be is used in this project is Autogluon. The reason for using autogluon is because nany models are better than few and hyperparemeter tuning enhances learning. There is no train split manually the model does that internally. The model handles missing values.

Since it is Classification problem, the next model to use is classification models.

Metrics

A set of evaluation metrics

- The confusion matrix is a technique used for summarizing the performance of a classification algorithm i.e. it has binary outputs.
- Classification Report
- ROC - AUC ROC (Receiver Operating Characteristic) Curve tells us about how good the model can distinguish between two things (e.g If a patient has a disease or no). Better models can accurately distinguish between the two. Whereas, a poor model will have difficulties in distinguishing between the two

Platform

- The project will be done using AWS Sagemeker studio.
- Data Will be stored in Amazon S3 cloud storage
- The endpoint will be deployed in Amazon cloud

References

Aamna. 2023. "HW1 Machine Learning for EHR." Kaggle. https://kaggle.com/competitions/hw1-machine-learning-for-ehr.