DTSC 691 Project Submission
Lawrence Ejime Iweriebor

Goals of the Project

This project aims to build a machine learning model capable of predicting the likelihood of a doctor prescribing opioids. The following objectives would be achieved:

1. Understand the opioid dataset using Exploratory Data Analysis (EDA)
2. Build different classification models
3. Compare the model performance using different evaluation metrics
4. Select the best model based on the evaluation metrics performance

The response variable for this research is the opioid prescriber, and the predictor variables are the various medications prescribed by the healthcare practitioners. The research question I have developed for this project is: what is the probability of a medical practitioner prescribing opioids?

Data description

Data Description

The dataset for this project was retrieved from the archives of the UC Irvine machine learning repository (https://archive.ics.uci.edu/ml/datasets.php), which is a publicly available dataset and repository. The dataset contains medical practitioners' information: gender, specialization, list of prescriptions containing the different types of medication prescribed for other patients, and different cases. Specifically, the information about the different areas of medical practice/specializations is Urology, Surgeon, Dentist, Psychiatry, Nurse Practitioner, etc. The dataset contains information about the different types of prescription medication administered to patients, and the number of times it was prescribed.

In this project, the response variable is the opioid prescriber and has two outcomes (Yes, No). The predictor variables are the 239 medications prescribed by medical practitioners in their area of specialization. Also, the response variable is the opioid prescriber with two outcomes (Yes, No). The predictor variables are the 239 medications prescribed by a medical practitioner in the different healthcare

specializations. The relationship between the response variable and the predictor variables is assumed to be a cause-and-effect relationship. This is so because prescribing a high dosage of opioid medication would likely result in addiction to the opioid medication.

Software

To complete my project, I will be utilizing python in Jupyter notebook, and several other packages to complete my capstone project.

Analyses

Analysis Description: Please see below for the analysis description.

Week 1 goals: My focus in week 1 is to complete my project proposal fully and then submit it to my mentor for review and feedback. Also, I will try to collect my dataset ready for week 2.
Week 2 goals: During week 2, I hope to have received some feedback from my mentor. I will incorporate any necessary changes and revise my proposal as required to ensure my proposal is ready for submission.

Week 3 goals: My goal for week 3 is to start working on my project. I will begin by importing the various libraries required for conducting Exploratory Data Analysis (EDA). Next, I will import my dataset, clean the data by filling missing values in features and complete the Exploratory Data Analysis. I will also conduct some visualization and graphical representations of my dataset to gain insights and learn more about the variables in my dataset.

Week 4 goals: My focus in week 4 is encoding and dimensionality reduction. Since my dataset contains some categorical variables, it is necessary to transform categorical features into numeric features. Next, I will do feature importance and selection by removing less important features to address the issue of correlation issues.

Week 5 goals: During week 5, I will perform the model selection by building several models and then decide on a model that can generalize the training dataset and make accurate predictions.

Week 6 goals: After selecting my desired model, I will do hyperparameter tuning and interpret my model.

Week 7 goals: Complete a final review of my project, prepare a video presentation and submit my project for grading.
Delivery Plan: The delivery and presentation of my project will be completed using Jupyter Notebook. Utilizing Jupyter Notebook allows for a more engaging and interactive presentation.

Deliverables

Understand the opioid dataset using Exploratory Data Analysis (EDA)

1. From the graphical exploratory data analysis (EDA) using histogram, the graph revealed that all the values in the numercial columns are distributed between 0 and 1000.
2. Also, the density plot revealed that all the feature variables are slightly skewed toward the right which implies that their mean values are greater than their corresponding median values.
3. Looking through the various correlation matrices and heatmaps, a correlation coefficient values greater than 0.8 indicates the presence of multicollinearity. However, most of our values were less than 0.8, hence no presence of multicollinearity.
4. Exploratory data analysis also revealed that the opioid data set contains more men than women. Men were 62.3% and women were 37.7% respectively. Further EDA analysis revealed that opioid was prescribed more for men than women.
5. Exploratory data analysis for the state showed that Carlifonia had the highest number of entries in the dataset with a 10.2 percentage, followed by New York with a 7.8 percetage point. The following states Virginia, (VI), Outside

USA, (ZZ), Guam, (GU), Armed Forces Europe (AE), and Armed Forces America (AA) account for an insignificant amount of entries in this data.

6. A bivariate EDA between state and opioid prescriber showed that Virginia had the highest number of opioids prescribed followed by New York, whereas Washington DC had the least count followed by Georgia.
7. Exploratory data analysis revealed that Internal Medicine specialists had the highest number of entries in this data, with a percentage of 12.8, followed by Family Practice Specialists, with a percentage of 11.9.
8. A bivariate Exploratory data analysis between Specialty and opioid Prescriber was conducted and the EDA revealed that Family Practice specialists prescribed more opioids than any other specialist group. This was followed by Internal Medicine. Also, some specialist groups such as Optometry, Certified Nurse Midwife, did not prescribe opioid as they had a Yes count of 0.
9. Finally, exploratory data analysis revealed the presence of class imbalance in the target variable. Therefore, class balancing techniques must be applied to balance the data.

Build different classification models

The following five models were built and tested:

1. Decision Tree Model
2. Logistic Regression Model
3. Light Gradient Boost Machine Model
4. XGBoost Classifier
5. Random Forest Model

Compare the model performance using different evaluation metrics

	Models	Accuracy_ score	Precision_ score	Recall_score	f1_Score	False_Positive_Rate	False_Negative_ Rate	ROC_ AUC_ Score
0	**Decison_Tree_Model**	93.504902	96.296296	90.508567	93.312772	3.490592	9.491433	96.870633
1	**Logistic_Regression_ Model**	93.750000	98.347356	89.012782	93.447537	1.499864	10.990207	97.058811

2	**LGBM_Model**	94.825708	98.272328	91.270057	94.641850	1.608945	8.729943	97.949021
3	**XGBM_Model**	94.594227	97.591410	91.460430	94.426506	2.263431	8.539570	97.718955
4	**Random_Forest_Model**	94.934641	97.581342	92.167528	94.797203	2.263431	7.914060	98.507608

Select the best model based on the evaluation metrics performance

Judging from the Accuracy score, f1 score, ROC_AUC score and the False Negative Rate of the different models, the Random Forest Model is the best model for the Opioid Prescriber.