Foundation of Data Science - Project

Completed by: Anurag Mishra(aa0279), Laksh Kataria(lvk8525), Harsh Harwani(hh2752)

Problem Statement

It's a widely accepted reality that there is a rampant substance abuse problem in our world today. It has become an essential task to devote resources and technology to help tackle this crisis. Data Science can be leveraged to study and find patterns that would help hospitals to understand patients, their needs, and if they can be discharged or not. Studying various features can help doctors predict the discharge status of a patient.

A previous **study** compared various machine learning models with Super learning and reported results based on comparing the models. Our project aims to apply predictive machine learning models to help hospitals identify the likely course that the treatment provided to a patient coming for substance abuse disorders will take, i.e, whether the treatment regime would be successful or unsuccessful. Our final dataset will have a target variable '$DischargeStatus'$ with class '1' indicating successful treatment completion and discharge and class '0' for an unsuccessful treatment program. Further, we will also be providing descriptive analysis on - state-wise death rates, most lethal addictions, and most susceptible age groups. Moreover, we will be using explainable AI tools such as LIME, and Shapely Values, to interpret our model results.

We aim to create a model which would provide hospitals with an intuitive idea, as to whether a certain treatment program chalked out for an incoming patient, would be successful or not, based on prior observations.

Background

A study published in the journal Plos One, in April 2017, looked at the TEDS-D 2006–2011 data set. The TEDS-D dataset contains a huge amount of information, for different patients that are admitted to hospitals across the United States, specifically for substance abuse-related health issues. This data is collected by the Substance Abuse and Mental Health Services Administration, under the United States Department of Health and Human Services. The organization collects and organizes data into two data sets - TEDS-D and TEDS-A. TEDS-D is the data set that is prepared for patients that have been discharged from the medical institution, and TEDS-A is the dataset that is created by collecting their information, at the time of admission.

The article referred to above aims to introduce Super Learning, which is a Machine Learning ensemble method, and compare its performance against other predictive algorithms, commonly used in ML, such as deep neural networks, random forests, and logistic regression, to match the patients, based on their characteristics encapsulated in the TEDS-D dataset, to the appropriate SUD (substance use disorder) treatment that they must receive. This paper forms the inspiration for our work, where we aim to deploy predictive ML algorithms, to predict the discharge status of incoming patients and identify at-risk groups and the most lethal addictive substances, amongst other exploratory analyses.

Value	Label	Frequency	%
1	Treatment completed	725,929	42.1%
2	Dropped out of treatment	432,610	25.1%
3	Terminated by facility	95,254	5.5%
4	Transferred to another treatment program or facility	369,503	21.5%
5	Incarcerated	26,005	1.5%
6	Death	3,576	0.2%
7	Other	69,626	4.0%
	Total	*1,722,503*	*100%*

Figure 1: Dataset description - Initial

Data Description

As alluded to earlier, we will be using the TEDS-D dataset, collected by the Substance Abuse and Mental Health Services Administration, for the year 2019. The data set in its entirety contains 1.7 million rows and 76 columns, where each row corresponds to the data collected for a single patient, and each column holds information for that patient. Each of the features in the data set is categorical and has been encoded to display numerical values corresponding to their class. The feature columns are also nominal.

Data Reduction Steps

Feature Selection

The data set contains a total of 76 features. We apply pre-analysis feature selection techniques such as correlation analysis, and domain knowledge to drop 27 features, thereby

ending up with a data set with 1.7million rows and 49 features

Downsampling

Figure 1 shows the description of our target variable, as included in the original dataset. The reasons for handing a discharge to a patient are specified and encoded, as shown in the table. In our project, we narrow down the objective to a binary classification, wherein we predict if the treatment provided to the patient was completed, i.e, a 'Successful Discharge' or incomplete, i.e, an 'Unsuccessful Discharge'.

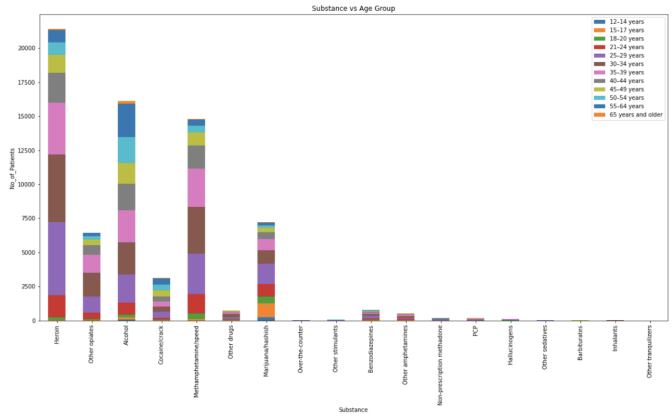

Figure 2: Chart showing most common age group per substance

- To reduce the number of data points within our dataset, we start by dropping all rows having target values - 2,6,7 (see assumptions for reasoning)
- To convert the problem to a binary classification, we encode classes - 3,4,5 as class '0', and class '1' remains the same We now end up with a dataset having 1.2 million rows and 49 features
- Now, we downsample our dataset further, making it easier to process and store:

  We drop all rows containing any 'Null' values. This approach leaves us with a dataset having 71k rows and 49 features, with a class split of 41k and 30k for classes '0' and '1' respectively
- The entire dataset has been documented in this pdf, and a description of each of the features included in the dataset can be obtained from the same

Exploratory Data Analysis

Before we set up a model pipeline and perform predictions on the dataset, we carry out some basic Exploratory Data Analysis. Since the dataset being used for this project is very descriptive and holds a lot of valuable information, we believe that performing EDA would help us identify certain key features which could be of great value to hospitals and medical facilities.

We perform our analysis in two ways -

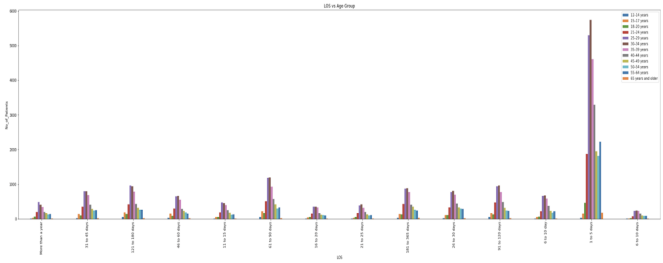

Figure 3: Chart showing most common age group per substance

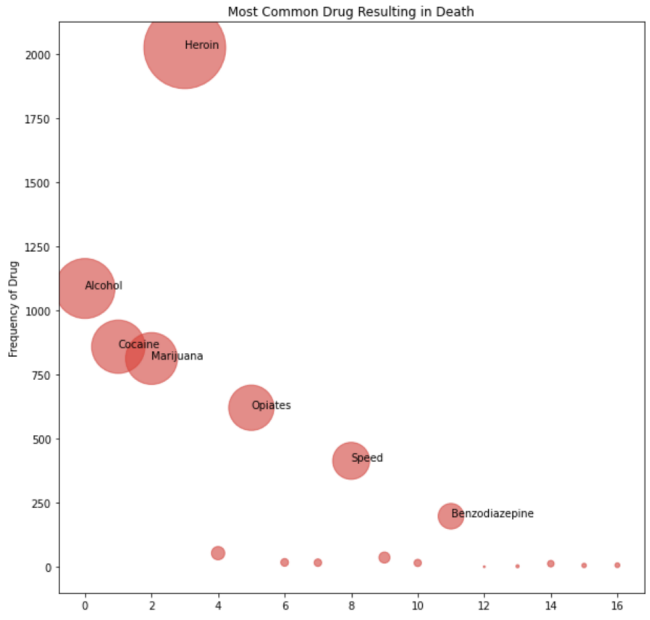

Figure 4: Bubble Plot showing most fatal drug addictions

- Exploratory Analysis on the entire dataset: For this analysis, we use the downsampled dataset which is being used for prediction, with 71k rows and 49 features. Through this analysis we identify:
  - The most common substance addictions per Race, for example - In our dataset, it was identified that category 'White' were more prone to Heroin addiction, and 'American-Indian' 'Asians' were more susceptible to alcoholism.
  - The most common substance addictions per age group. It was identified that people in the age range '25-29' were the primary Heroin abusers, followed by the age group '30-34'. Similarly, most Methamphetamine abusers also belonged to the age group '30-34'.
  - Further, we also performed analysis to identify the 'Length of Stay' within the medical facility, and visualized the treatment outcomes per age group.
- Analysis on specific treatment outcome - 'Death': In our dataset, class '6' within the column 'Reason' corresponds to a fatal treatment outcome. Upon data analysis,

we realized that aproximately 3500 patients died within the treatment facility. We felt the need to perform a special analysis on this specific dataset to identify any common patterns or addictions which could lead to this fatal outcome.

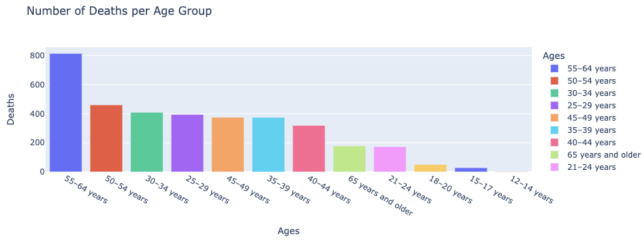

Figure 5: Chart showing death rates per age group

- First, we try and find the most common substances that these patients were addicted to, to find out the most lethal drg addictions. Heroin addiction was found to be the most fatal and by quite some margin.
- Second, we identify the age groups having most fatalities. Unsurprisingly, most deaths occured in the age group '55-64', however, perhaps the most surprising result from the analysis was the fact that age groups '21-24' and '65+' had almost the same number of deaths.
- Finally, we visualize the death rates across the various states of the USA and present our findings in a chloropleth map. California and New York have the highest death rates, with New York having almost double the count of California.

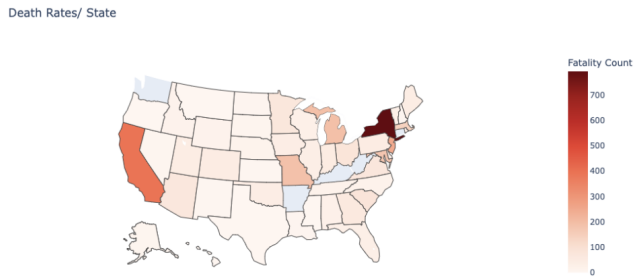

Figure 6: Chloropleth map depicting death rates across the USA

Model Type and Description

In creating our project, we will adopt different methodologies to ensure that we can extract the maximum relevant information from a data set, as descriptive and well-documented, as ours is. As mentioned earlier, our primary goal is to predict the successful treatment of patients admitted to the healthcare facility and identify the key features which could help provide hospitals with prior knowledge at the time of patient intake, regarding the possible outcomes for that case. For this purpose, we will be building a predictive model.

- **Scaling and Splitting:**After handling the dataset by reducing and getting it to proper form we wanted to ensure that its scaled so we used MinMax Scaler as it works well on categorical dataset. After completing the scaling we splitt the dataset and designed a pipeling for multiple classification models.
- **Model Pipleline and Hyperparameter Tuning:**Further we wanted to create the Grid Search parameters for each model. We are doing this so that we can pass these in the GridSearchCV function that will take a pipeline and test out each parameter value we pass through in the following parameter list. Note: Before we created the grids for each, we created parameter values within lists so that I could pass in the lists, rather than hardcoded values.
- **GridSearch and Cross Validation** Now we used the GridSearchCV function and pass in both the pipelines we created and the grid parameters we created for each model. In this function, we are also passing in cv = 3 for the gridsearch to perform cross-validation on our training set and scoring = 'precision' in order to get the accuracy score when we score on our test data.

Evaluation Metric

The evaluation metric that we will be using for our model, would be the Precision of the model. For our model, reducing the number of 'False Positives' is more important as compared to 'False Negatives' (check assumptions). To measure the ability of our classifier, we will be using the ROC-AUC curve.

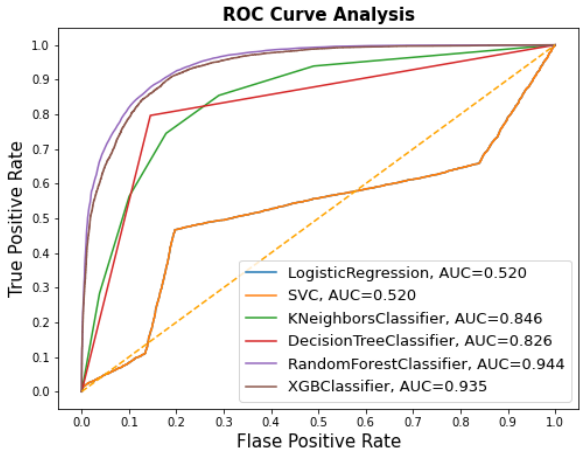

Figure 7: ROC-AUC curve for all classification Models)

Results

Based on the pipeline structure as explained above, Random Forest classifier was found to be the best model, in

terms of the evaluation metrics that we aimed to maximize (Precision). Apart from achieving a precision value of 0.88, the model also achieved an accuracy of 87% which gave it the edge over XGB, which was the second best performing model. To visualize our results, we created an ROC-AUC curve for the Random Forest model

	precision	recall	f1-score	support
0	0.88	0.88	0.88	10297
1	0.84	0.84	0.84	7667
accuracy			0.87	17964
macro avg	0.86	0.86	0.86	17964
weighted avg	0.87	0.87	0.87	17964

Figure 8: Precision and Recall for each class

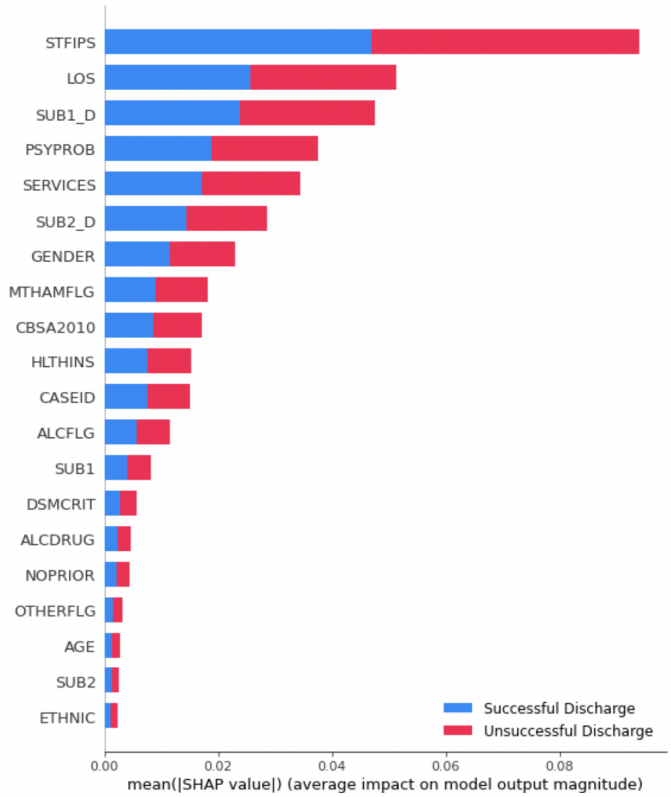

Figure 9: Summary Plot

Model Interpretation using Lime and Shap

Machine Learning models are often black boxes that makes their interpretation difficult. In order to understand what are the main features that affect the output of the model, we need Explainable Machine Learning techniques that unravel some of these aspects.

Two of these techniques is the SHAP and LIME method, used to explain how each feature affects the model, and allows local and global analysis for the dataset and problem at hand.

POSTIVE CLASS	PRECISION	RECALL	F1-SCORE
LOGREG	57	1	73
RANDOM FOREST	88	88	88
KNN	81	81	82
SVM	57	1	73
XGB	88	87	88

Figure 10: Performance of Classifiers

Shapely Values

SHAP provides global and local interpretation methods based on aggregations of Shapley values. We wont go into detail about these libraries in this paper, but just using few plots try to understand what features contributed to the preidiction at hand and also for a specific case.

- SHAP values of a model's output explain how features impact the output of the model.
- In Figure 9 Summary plot explains the impact of a feature on the classes.In other words, the summary plot for binary classification can show you what the machine managed to learn from the features.
- In figure 10 we can see a waterfall plot which is a local analysis plot of a single instance prediction (in this case instance 8).
  - f(x) is the model $predict_proba$ value: 0.65 and E[f(x)] is the base value = 0.575.
  - On the left are the feature value and on the arrows the feature contribution to the prediction.
  - Each row shows how the positive (red) or negative (blue) contribution of each feature moves the value from the expected model output over the background dataset to the model output for this prediction

LIME

LIME stands for Local Interpretable Model-agnostic Explanations. It is a Python library help you understand the behavior of your black-box classifier model.

Steps that LIME takes:

- **Input Data Perturbation** The first step that LIME would do is to create several artificial data points that are close with the data
- **Predict the class of each artificial data point** Next, LIME will predict the class of each of the artificial data point that has been generated using our trained model
- **Calculate the weight of each artificial data point** The third step is to calculate the weight of each artificial data to measure its importance.

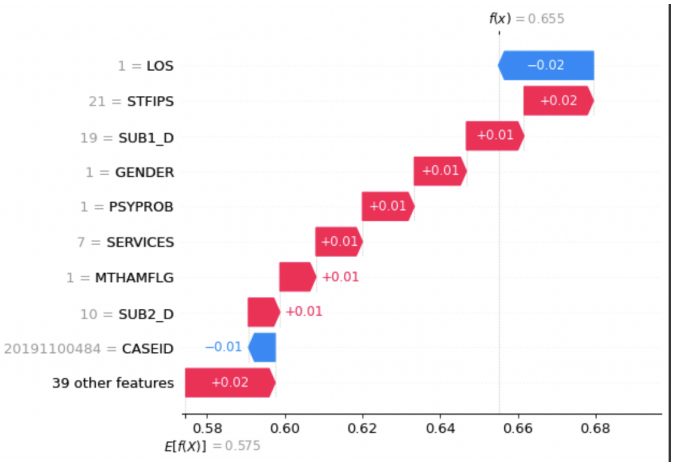

Figure 11: Waterfall Plot

- **Fit a linear classifier to explain the most important features** The last step is fitting a linear regression model using the weighted artificial data points. After this step, we should get the fitted coefficient of each feature, just like the usual linear regression analysis.

  Here for the 10th index it predicts it to lie in class '0' with 75% confidence and the reasons of attributes are shown on the side.

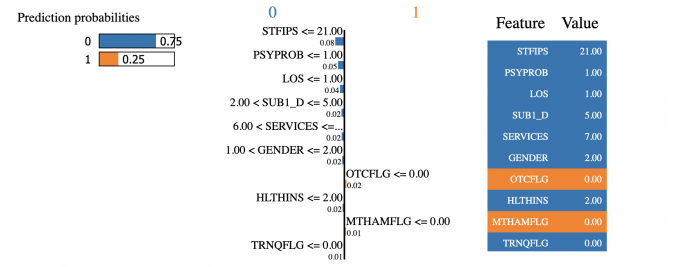

Figure 12: Lime summary

Assumptions

- Since the dataset at hand comes from a government source, we assume the validity of the dataset is implied. We also assume that the risk factors and other important metrics which we identify through analyzing the information, albeit based on 2019 data, can be extrapolated to the current year and will provide a useful knowledge base for current world scenarios.
- We dropped classes 2 - 'Dropped out of treatment', 6 - 'Death', and 7 - 'Other' from our dataset, as for classes 2 and 7, the treatment facility does not have information regarding the final status and whereabouts of the patient. For each of the included classes, the final status and condition of the patient are known. We drop class 6 as we perform a separate descriptive analysis on these patients, and rows corresponding to class 6 are only 3.5k which does not disturb the dataset.
- Our reasoning for using Precision as our evaluation metric is that, since the model is being developed with a primary goal of helping hospitals identify if a specific treatment course would be successful or not, and not optimize the logistic costs of the hospital, falsely classifying a course as successful could have greater repercussions than a negative classification, in terms of the patient's health and well-being. Further, for an imbalanced dataset, precision would be the most appropriate choice for evaluation.

Future Steps

- Performing similar analysis on more recent and complete data
- Using deeper models such as neural networks to handle high volume data
- Optimizing algorithm to reduce cost on patient and hospital end
- Learning more about overfitting and how we can deal with the overfitting in our analysis

  Another major consideration as mentioned in class, would be the high Recall values for Logistic Regression and SVM, despite poor accuracy and F1 scores. Fig 10 shows how these models achieve artificially high, perfect recall scores. We attribute this to a class imbalance within the dataset and possibly because of the data preprocessing steps. However, an in-depth analysis into identifying the reasons for the bad performance of these models, in comparison to XGB and RF would be beneficial.

References

[1] Acion, Laura, Kelmansky, Diana, der Laan, Mark V., Sahker, Ethan, Jones, DeShauna, and Stephan Arndt. "Use of a machine learning framework to predict substance use disorder treatment success." PLOS ONE 12, no. 4 (2017): e0175383. Accessed December 12, 2022. https://doi.org/10.1371/journal.pone.0175383.

[2] $DataFiles$: https://www.datafiles.samhsa.gov

[3] $SHAP$: https://github.com/slundberg/shap

[4] $LIME$:https://github.com/marcotcr/lime

Team Evaluation

Points to distribute - 12
Distribution:
1. Anurag Mishra (aa9279) - 4
2. Harsh Harwani (hh2752) - 4
3. Laksh Kataria (lvk8525) - 4