Cohort 12

NLMS Team

Progress Report I

Project which our team has been working on was broken down into the following stages with each of them being worked on during and after taking a course providing necessary knowledge and expertise to complete it.

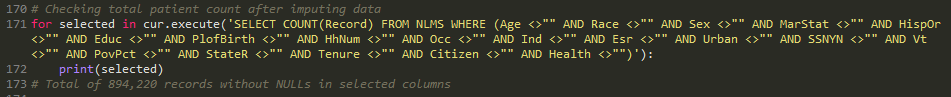
**Stage 1: Data Processing**

**Stage 2: Predictive Model Engineering**

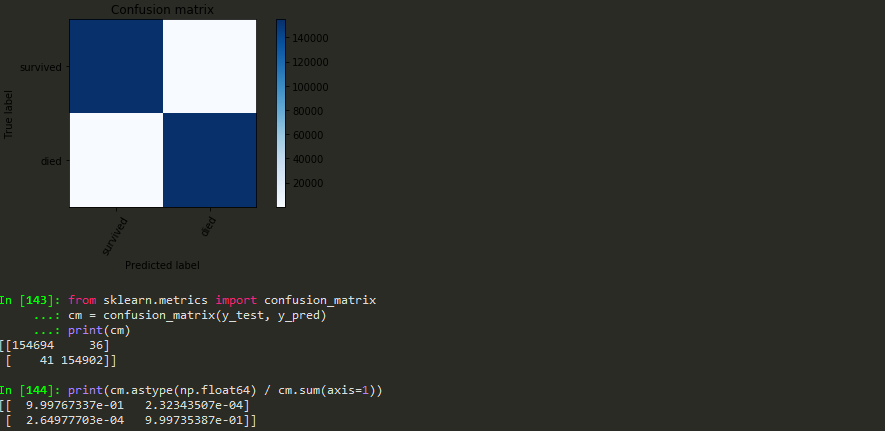
**Stage 3: Execution**

**Stage 4: Model Deployment & Visualization**

For the **Data Processing Stage**, our team have achieved consensus on utilizing SQLite Python package for data ingestion and data wrangling. The raw data set which we have obtained contained 1,835,072 observations with 41 features. High level data analysis carried out by the team showed that raw data set required substantial data manipulation, as missing data rates for some of the features were very high, thus reducing the number of observations suitable for being fed into the machine learning models. The decision was taken to remove features with high rates of missing data, the values for which cannot be imputed or deducted without compromising data integrity. In the same manner, features which demonstrated high collinearity were removed as well, keeping only one of them, and, thus, reducing our set of features from 41 to 19. As a result of such data wrangling, we obtained a data set with smaller number of observations (894,220) and 100% of data presence for all of the selected features.



Upon completion of the data ingestion and wrangling stage, we have focused on choosing appropriate algorithm for **Predictive Model Engineering**. This task had been complicated for us due to the class imbalance problem which we encountered – number of observations where Indicator of Death dependent variable = 0 (alive) outweighed significantly another member of the class where Indicator of Death dependent variable = 1 (dead). Based on the input provided by the faculty members and some independent research, it was decided to attempt applying data re-sampling technique to solve the problem. For that we have combine over- and under-sampling methods using Synthetic Minority Over-sampling Technique and Edited Nearest Neighbours (SMOTEENN) from imbalanced-learn library. Re-sampled data set was further fed into the logistic regression model to generate predictions.



At this time we continue working on model engineering stage, and will be further focusing on applying classification algorithms, as well as Cox PH model to the re-sampled data set. Our approach to model engineering will be reviewed and reevaluated when taking Machine Learning and Statistics classes.

Individual project contributions of each team member are listed below:

**Yara**: - Created drafts of project deliverables: Project Proposal, Design & Architecture, Progress Report I;

* Created Python code for data ingestion and wrangling using SQLite library;
* Researched existing approaches to dealing with class imbalance issue;
* Implemented Python code for re-sampling data set utilizing SMOTEENN method;
* Developed Python code for building logistic regression model with re-sampled data.

**Don:** - Ingested/wrangled initial data set (700K+ records);

* Analyzed acknowledged methods for addressing class imbalance within data;
* Implemented Python logistic regression model with training &amp; test data;
* Examined data visually for variable relationships and their strengths using Tableau