Cohort 12

Mortality Probability Team

Progress Report I

The Mortality Probability Project team broke down the our data science architecture into the following stages, each of which being worked on during and after taking a course providing necessary knowledge and expertise to complete it.

**Stage 1: Data Processing**

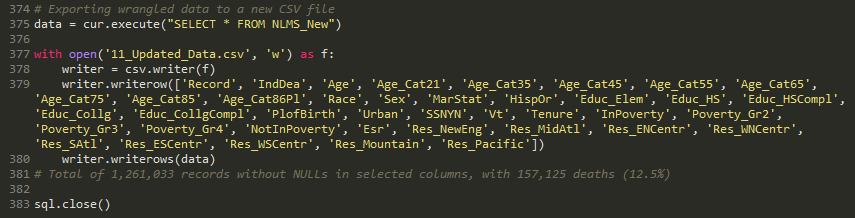
**Stage 2: Predictive Model Engineering**

**Stage 3: Execution**

**Stage 4: Model Deployment & Visualization**

Based on the courses we have taken thus far including Foundations of Data Science and Data Analytics, Software Engineering, Data Sources and Storage, and Ingestion and Wrangling (Session 1), we will be focusing on Stages 1 and 2 for the purposes of this report.

For the **Data Processing Stage**, our team achieved a consensus on using the SQLite Python package for data ingestion and data wrangling. While not ACID compliant, this tool proved to be ideal for our purposes because it provided the easiest possible set-up with no configuration, its free to use, open source and acts a single file database for our singular data set. The raw data set which we have obtained contained 1,835,072 observations with 41 features. There were no duplicate records; the data set has high integrity. Initial data analysis carried out by the team showed that the raw data set required some 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 deduced without compromising data integrity. Also, hierarchical variables were converted to categorical ones where applicable. 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 32. As a result of such data wrangling, we obtained a data set with smaller number of observations (1,261,033) and 100% of data presence for all of the selected features. One question that arose from this process related to the possibility of linking/joining new data to our set. During the Data Sources and Storage class on data ingestion our instructor talked about using multiple data sources (or at least more than one). Specifically, the Titanic Workshop example contained several variables associated with each passenger (gender, age, cabin location, etc.). Could it be possible to add additional information from another source to our data set that is so specific and unique to each record? However, given the level of difficulty and lack of additional insight in patient detail or state-level data, this could prove out of bounds from the scope of our capstone.



Upon completion of the data ingestion and wrangling stage, we have been focusing on selecting appropriate algorithm for **Predictive Model Engineering**. At this stage in the process is where we discovered a more complex question. With 12% mortality rate in the data set, we are potentially facing a class imbalance problem: number of observations where Indicator of Death dependent variable = 0 (alive) outweighs significantly another member of the class where Indicator of Death dependent variable = 1 (dead). Therefore, the model “knows” to predict alive in every time with a high level of accuracy of being correct. With this in mind, we needed to know the best approaches to fix class imbalance. Based on faculty member input, our team is currently researching different sampling techniques to fix our imbalance problem including applying oversampling, using a classifier, and adjusting performance measures. As a team, we are looking into data re-sampling techniques to solve the problem, if applying logistic regression without data re-sampling results in a high number of false positives and low number of true positives compared to the acceptable levels in the literature. If in the course of our trial-and-error approach we establish that data re-sampling is required, Synthetic Minority Over-sampling Technique and Edited Nearest Neighbours (SMOTEENN) algorithm from imbalanced-learn library will be applied to combine over- and under- sampling methods. Re-sampled data set will be further fed into the logistic regression model to generate predictions.

At this time our current hypothesis remains as follows:

“Probability of mortality by all and some specific causes among U.S. population can be predicted based on a number of socio-economic factors.”

Our team will continue working on model engineering stage, and will be further focusing on applying classification algorithms, as well as Cox PH model if appropriate. Our approach to model engineering will be reviewed and re-evaluated while taking the Statistics and Machine Learning 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 & test data;
* Examined data visually for variable relationships and their strengths using Tableau.

**Arnab:**

* Initial data analysis to understand the data set, finding correlation and suitability data for use in logistic regressions;
* Initial data analysis to suggest further analyses.

**Robert:**

* Coordinate schedules and times for team meetings, draft follow-up minutes and deliverable deadlines.
* Peer review and created drafts of project deliverables.

Arun:

* Independent hypothesis testing and ingestion research