Mid-Semester Progress Report

DSA5900 – Fall 2021

Haixiao Lu

10/09/2021

# Introduction

A central challenge facing the Federal Aviation Administration (FAA) is how to determine if an airman will be flight ready. Advancements in machine learning technology and improvements in text mining capabilities allow for researchers and federal authorities to examine a variety of data in new ways. This project proposal will focus on data mining medical records to recommend who should get their pilot’s license renewed and for how long based on features in the data.

# Objectives

Project Objectives:

* Analyze the existing data to process medical into useful features by using text analysis, such as NLP (Natural Language Processing)
* Work with Subject Matter Experts to determine how they make this assessment
* Train machine learning models based on processed medical record data
* Identify which model produce the best predictions
* Visualizing the outputs.

Personal Learning Objectives:

* Become more familiar working with text data using different natural language processing techniques, data mining and other algorithms to learn more about the relationship between extracted features and output labels.
* Acquire experience by working with real-world data to know how it affect people’s day to day life, and how to extract sample data from a huge dataset.
* Carry out a complete modeling process, starting with the problem understanding, through data analysis, model training and testing, and ending with the model’s validation.
* Finally, to get experience with a professional team that is working towards making peoples life better using technology, in my case data science.

# Data Ingestion

There were some uncontrollable external factors to get the dataset on time since this dataset contains sensitive medical data and it comes from Federal government. However, while I was waiting for the dataset to arrive. I used a medical text dataset from Kaggle to get familiar with the text data and get some insights about what I should expect from the real dataset.

The dataset from Kaggle is very small(17MB). I downloaded to my local computer.

# Exploration

The dataset has 4999 rows and 6 columns. All the rows with records of text information.

Some processes I have taken so far:

* Data extraction, extracting features from text notes.
* Data Cleansing, cleaning up the dataset by extracting feature and adding labels
* Exploratory Data Analysis
  + Bar Chart
  + Countplot
  + Histogram
  + Heatmap (visualize the missing values)

# Preparation

* Defined functions to clean up the dataset, such as remove punctuation, length count of certain words in a text file, etc.
* Used TFIDF (term frequency-inverse document frequency) to perform feature extraction
* Visualized the tf-idf features using t-sne plot
* Used PCA to reduce feature’s dimension

# Methodology

Steps I took for Kaggle dataset:

* Supervised Learning
* Split train and test dataset (80/20 rule)
* Build Initial Logistic Regression model to check the importance of the feature
* Adjust dataset, dropped some features
* Train more models:
  + Naïve Bayes
  + Random Forest
* Use Confusion matrix to evaluate the model performance

Steps prepare to take for FAA dataset:

* Supervised leaning
* Extract 10% (1 GB) of the sample data from the dataset (100 GB)
* Split train – validation-test dataset (80 – 10 – 10)
* Build models
  + Logistic Regress
  + Random Forest
  + KNN
* Evaluate method
  + Accuracy
  + Confusion matrix
  + ROC curve and AUC

# Process Validation

This project is a classification problem, the outcome is going to be binary, either the pilot’s license should Renew or not (YES OR NO)

I am planning to use Random Forest model to check the feature importance. Then based on the performance, some features might need to be adjusted. Then, I will use Logistic Regression and KNN to compare with each other’s performance.

All these models’ performance will be evaluated by confusion matrix, classification accuracy, ROC curve and AUC.

These evaluation methods are propitiate for classification problem.