Mid-Term Progress Report

DSA5900 – Fall 2021

Haixiao Lu

10/10/2021

# 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.

# Literature Review

The FAA project (Aeromedical Certification Prediction) is being leveraged to improve future forecasting for airman whether in her/his current state, has sufficient capacity to adequately perform safety-critical tasks in the aviation environment with a binary outcome. Large datasets (100 GB) can be used to train and test machine learning models. However, this project planning to focus on using smaller portion (10%) of the data to train and test machine learning algorithms.

Since a significant portion of the data in a textual format. NLP (Natural Language Process) will be used to leverage the dataset. The data will be split into train (80%) – validation (10%) – test (10%). Some potential machine learning algorithms will be used, such as Neural networks to classify individuals, Random Forest to define and select feature importance, Logistic Regress and KNN to compare model performance.

# Issues and Challenges Encountered

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 textual data and get some insights about what I should expect from the real dataset.

Some issues while I was exploring Kaggle dataset:

* Domain knowledge is required, (so many medical words I never heard of)
* Data mining technique is required, (how to define and extract which feature is useful)
* Text data is really mess

Issues will be encountered in FAA dataset:

* How to extract a portion of the data from large dataset (100 GB)
* How to make the data clean in a useful way
* What kind of method can be used to extract 10% of the data from the large dataset
* How to get familiar with medica domain knowledge and define which features are important to the model
* Which evaluation methods to use to evaluate the model performance

# Work to date

## Data Ingestion

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
  + Neural networks
  + Logistic Regress
  + Random Forest
  + KNN
* Evaluate method
  + Accuracy
  + Confusion matrix
  + ROC curve and AUC