Project Deliverable 2

Wei Luo

Goal

As the Project description mentions, the initial goal of this project was to examine the relationship between YES rate, the demographic features of youth and the caseworkers’ note.

Before Project Deliverable 1, we accomplished the Topic Modeling of caseworkers’ note and did some preliminary analysis on the result of Topic Modeling.

By far, all three questions in project description has been answered. The attributes of the young people’s lives are summarized in the Codebook\_Case Management Notes.xlsx, which was done in the previous group. The role of relationships with DYS staff play in a young person’s decisions to engage with YES has been examined, while there is by far no obvious relationship between these two taken into consideration of the caseworker’s notes(we will proceed trying until end of project). The outcomes of young people who engage with YES has not been examined yet since the lack of dataset.

Data Aggregation

This step was done by previous group (see Codebook\_Case Management Notes.xlsx)

Analysis

The topic modeling was first done in order to put caseworkers’ notes into different topics, and below is a visualization of the result of topic modeling:

Chart, bubble chart

Description automatically generated

Followed by the topic modeling, I tried different combinations of features and compared with result of Topic Modeling until I found no apparent associations.

The next thing I tried was a decision tree algorithm of all the features on the labels produced by topic modeling, but the depth of decision tree was huge (around 69, see DTresult.dot.pdf), meaning that there is no clear relationship using the approach of decision tree.

Then I, with the project team realized that there are two main issues:

1. The result of Topic modeling was in fact not pure, meaning that TFIDF of different groups are mainly differed by words that are not important (eg. also, well, etc)
2. Decision tree algorithm was not useful if the goal is only to find out the effects of caseworker’s notes on Yes engagement rate.

Therefore, following actions will be taken:

1. Use Jieba package to again preprocess the texts so that only keywords are taken into consideration, and the topic modeling result will only be compared against Yes engagement rate.
2. Groupby caseworkers notes by Yes engagement rate and summarize the term frequencies and punctuations and manually find out the differences between the caseworker’s notes that lead to a Yes engagement and caseworker’s notes that did not lead to a Yes engagement.
3. A sentiment analysis prediction of caseworker’s notes will be taken force.