Datafest Report

Team UTSCA

Based on our analyze of the data, we have the following suggestions that could bring more profit:

**Recommendation 1** is to allocate more attorneys to Indiana, North Carolina, Massachusetts, Florida and Tennessee. That is because they have the least client attorney ratio among all states.

**Recommendation 2** is to recruit or train attorneys in the topics of "Family and Children" and "Housing and Homeless". That is because these are the two most frequently asked questions among all.

**Recommendation 3** is to construct a model to take in as input a client's info, then output the most likely topic this client will want to talk with the attorney about. We have done this successfully with a toy example. This is helpful to the attorney as before they even talk to the client, they can prepare better for what the client will want to talk about.

During the process of tackling the problem we first split tasks into the following two questions:

**1.** Allocate the correct number of lawyers to the correct state and category

**2.** Extract possible goals from questionposts.csv (PostText)

create a model that takes client information and predicts what their goals are likely to be

**First Question:** The techniques we used to resolve the first question is data visualization using Power BI. We use histogram, heat map and pie chart to help us visualize the data.

After analyzing the data, we realize that Indiana and North Carolina have the least client attorney ratio, while on the other hand, Texas and Florida have the most attorneys. Michigan and California are being assisted the most. Our clients are generally low-income, female, middle-aged, Caucasian and living by themselves. Hence we could have a better portrait of our clients.

**Second Question:** The techniques we used to resolve the second issue involve machine learning, specifically, natural language processing with R.

For this task we wanted to analyze what a client would be likely to talk about, given the client's info. To do this, we extracted keywords from conversations between clients and attorneys, then organized the keywords into topics. Then we mapped each conversation to its client, and each conversation to a topic (through the keywords). Now we have a mapping between clients, and topics. Now we feed it into a RandomForest model. From this we can predict the topic that an arbitrary client will talk about, given the client's info. (such as demographics, location, etc…) Now before an attorney even talks to a client, they can prepare for what the client will likely talk about, which would be of great use to attorneys.