Predicting Divorce Likelihood using SVM and Platt Scaling

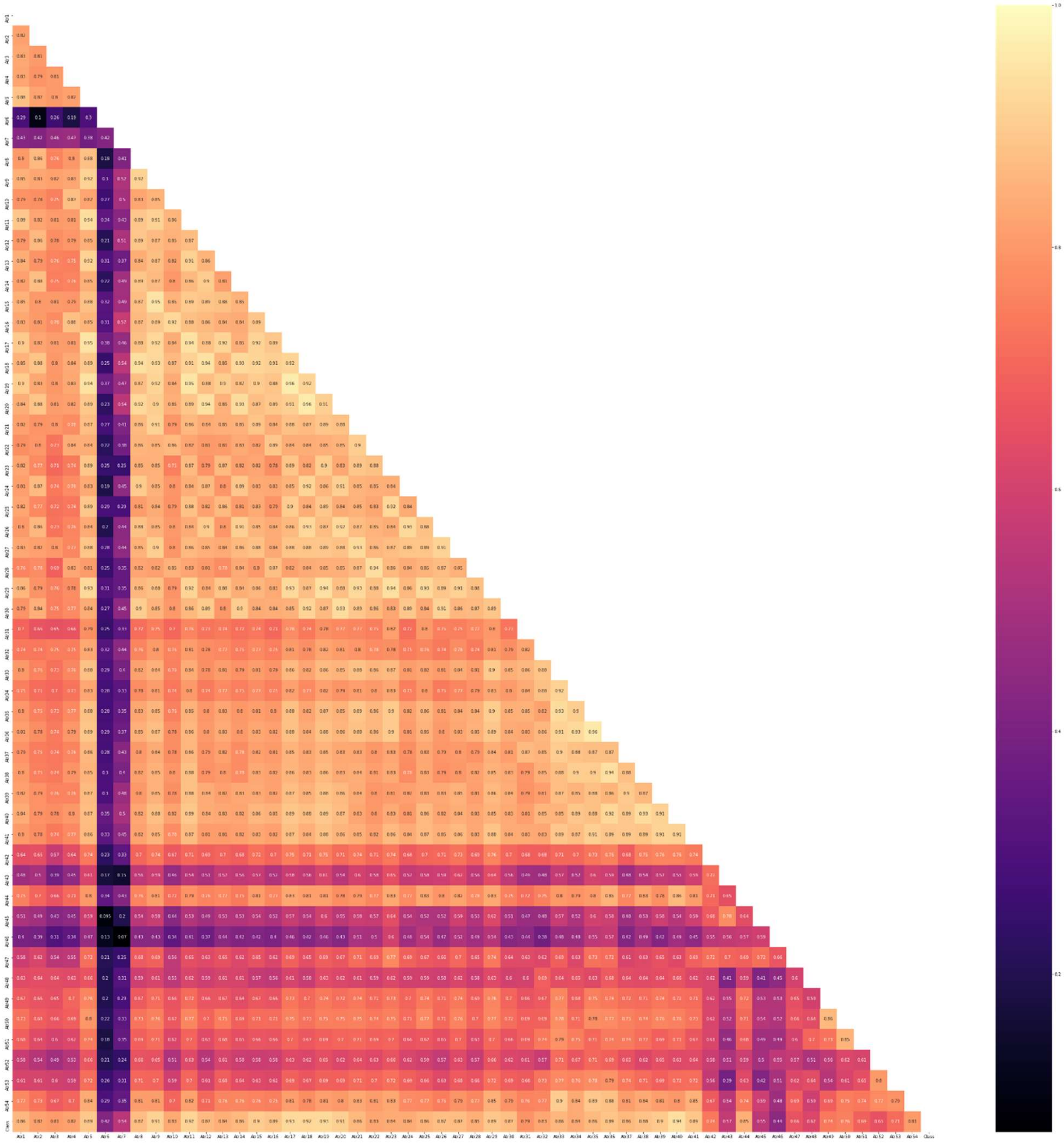
- **Any surprises from your domain from these data?**

  Outside the data which I am using for this project, it's quite interesting what has been done with divorce prediction in the field. The most well known is from John Gottman, who is an American psychological researcher that has published many papers on the matter. He also founded the Gottman Institute which has turned this data into a relationship toolset for couples. Similar to what I'm trying to do, they've created their own survey to gauge couple health. There are a few other companies out there that also claim to be doing something similar.

  One of the more surprising studies was done by the University of Southern California where rather than gauge health off a survey, they base it off analyzing the tone of voice the couple uses when speaking to one another. Madhumita Murgia says, "The algorithm broke the conversations down, using speech processing technology, into its acoustic characteristics - volume, pitch, intensity, jitter and shimmer, which can measure whether someone's voice is shaking or warbling, perhaps due to emotion". The algorithm turned out to be accurate 79% of the time, which beat out the accuracy of professional therapists.

- **The dataset is what you thought it was?**

  The dataset itself is pretty straightforward. It is well cleansed, with no missing values or data prep necessary. Part of what I'm doing is recreating the survey that went along with the study. While the questions were provided, the exact scoring wasn't. I have a background in psychology so I'm familiar with these types of surveys. It's very obvious that it is built around a 5-step Likert scale, most likely along the lines of Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree. I went through each question and assigned the score values to each one. Now the interesting thing is when the question inverses (and thus the points switch, e.g., 0 goes from Strongly Disagree to Strongly Agree), which is very common for psychology surveys as it's a way to weed out the people who fill it out haphazardly. In these situations, these questions plummet in correlation with the predictor as seen below in the black line. Most likely it probably indicates that these questions are poorly worded and confuse people.



Just to make it clearer, a positive question (which are the norm) would be something like question 5:

The time I spent with my wife is special for us.

Which is right before that first dark line, which is the first negative question (question 6):

We don't have time at home as partners.

Again, they don’t really provide the specific methodology in the paper which is a shame. In a real survey they are also supposed to balance the order of the questions (random), and each ‘random’ order is supposed to be balanced among the individuals. There are 54 questions and the sample size is only n = 171 so it seems to me that it is not exactly balanced. If this were something I wanted to do a real study on I would probably redo the test and balance it better.

- **Have you had to adjust your approach or research questions?**

  Not yet, I still think I will be able to do what I set out to do / answer the question I wanted to ask. The approach still remains the same, build a model and then pass through a set of answers to get a prediction. If I have time I want to host this in a docker container using Kubernetes on Google servers. This will allow me to build out a web interface for people to take the survey and provide a result.

- **Is your method working?**

  I have just done the exploratory data phase and started to build out the pipeline for building the model and providing test answers. I will know soon if the method is working but I don’t anticipate any issues.

- **What challenges are you having?**

  Anaconda, I swear it breaks every time I take a pause from using it and go back to using it. I really enjoy using Jupyter notebook for quick prototyping and I use Visual Studio Code to build out my code. VS Code has an interface Jupyter Notebook to run them within the program and for some reason I just can’t get it to work. I think the latest build is breaking but I just don’t have the time to keep troubleshooting it. I’ll have to go back and forth between Jupyter and VS Code.