Divorce Prediction using SVM and Platt Scaling.

# **Jonathan Henin DSC 680 – Spring 2020** <https://github.com/JonHenin>

# Which Domain?

Marriage Therapy

Data

1. Dataset: <https://archive.ics.uci.edu/ml/datasets/Divorce+Predictors+data+set>
2. Original Paper: <https://dergipark.org.tr/tr/download/article-file/748448>

Research

1. APA Marriage & Divorce Statistics: <https://www.apa.org/topics/divorce/>
2. Marriage Counseling Statistics: <https://onlinedegrees.bradley.edu/blog/how-to-convince-your-partner-to-attend-marriage-counseling/>
3. Gottman basis for Divorce Predictor Scale: <https://www.gottman.com/blog/the-empirical-basis-for-gottman-method-couples-therapy/>
4. Further Gottman research which claims he can predict with up to 94% accuracy which couples will divorce: <https://reallifecounseling.us/predict-divorce-gottman/>

Modeling

1. Pratt Calibration Paper: <https://www.cs.cornell.edu/~alexn/papers/calibration.icml05.crc.rev3.pdf>
2. Calibrated Classification in Python using scikit-learn: <https://machinelearningmastery.com/calibrated-classification-model-in-scikit-learn/>
3. SVM in Python using scikit-learn: <https://scikit-learn.org/stable/modules/svm.html>

Survey Development

1. Docker SDK for Python: <https://docker-py.readthedocs.io/en/stable/>

# Which Data?

The dataset is provided by UCI Machine Learning, called Divorce Predictors. It is a dataset that has 54 attributes which are answers to a survey (Likert scale) and is split in classification between those who are recently divorced and those who are happily married.

# Research Questions? Benefits? Why analyze these data?

The divorce rate in the US is between 40-50%, with the divorce rate of subsequent marriages being higher. While counseling can work for some, the financial and emotional toll can prove too much for others. Being able to identify a problem through an anonymous algorithm might prove beneficial to people. The sooner a problem is identified and addressed, the more success couples usually have. Therefore, if we are able to identify potential issues early on before they become major issues, perhaps more marriages could be saved.

# What Method?

The classification of this problem is binary, however, the end result will be better given as a probability of divorce as opposed to a binary response of will the marriage fail or not. The way to do this is through Platt scaling by using a calibration scale on a support vector machine model. As a bonus add-on, I want to build an interface where someone could answer the survey and see the resulting percentage.

# Potential Issues?

**Technical:**

The modeling is rather straightforward so that shouldn’t be much of an issue. The challenge might be more in the interface. Because I want someone to be able to take a survey, that requires passing the results back through the model. I will build out a model and place it in a Docker container and run on Kubernetes, which will be hosted on Google. This should give me an API to call which I just need to hook up to the interface and then retrieve the results and present them to the user.

**Data:**

The data is rather clean, meaning no missing values. However, the data is incomplete in the sense we are only given the question answers and the classification. Missing is the metadata about the users such as did they have children, how long were they married, age, etc. Also, while I do have the questions to the original survey, I don’t have the setup from it. I can deduce that it’s a Likert scale based on the 0-4 values, and from my background in psychology, I am fairly confident that I can reproduce it, but it is not 100% guarantee.

# Concluding Remarks

Due to the data issues, presenting the resulting percentage to the user as a real-world indicator value would be unethical. I will explain this is for educational purposes only and not intended in any way to be used as an indicator in their marriage. In a real-world scenario, I would want to perform the survey myself. Also, I think presenting a user in general with their likelihood of divorce percentage is borderline unethical. Though a case could be made either way, I think the more ethical solution would be to develop an app that someone could fill out a survey and track their results over time, but instead of a percentage, translate that into some sort of Marriage Health score. For the benefit of the user, if it notices a slide the app could alert the user on things they could try in order to help, based on what areas they are sliding in. Of course, that’s in a real-world use. Since this is just an educational project, I’ll stick with just the percentages for now.