Predicting Divorce

Robert George

I work part time as a data collector with the Institute for Research on Poverty here at the University. A large portion of my time at work is spent reading through divorce proceedings from courts across different Wisconsin counties. The diverse demographics, familial traits, and ages from the cases that I see in Wisconsin alone pose questions about the underlying causes for divorce. In this project we will explore a few major familial and individual influences upon divorce in an attempt to provide insight regarding the underlying causes of this phenomenon.

The case data that the IRP uses for its research analysis are not publicly available, and I do not have clearance to use them for my own research projects. As a result, I was unable to source my divorce data from that which I see at work. However, the issue of divorce is well documented across the country. The National Longitudinal Survey of Youth 1979 enacted by The US Bureau of Labor (Access Data / Investigator) records the lives of a sample of American youth born between 1957-64 from 1979 until 2016. I selected 16 variables from their large database, several of which represented the same variable represented bi-yearly over a ten year period from 2000 to 2010. These variables included Family size, Age of Youngest Child, and Age of Subject. I averaged these variables across the five separate columns to create one aggregate variable for each. Marital Status was aggregated across the same five years to create a "divorce" variable, which gives a boolean value indicating the presence of a divorce in the time frame. Finally, I mapped string representations to the numerical values under the Region and Race variables. Cases where Regional dwelling changed across the years in question were dropped, as were rows with na values for any remaining variables (these were few). After these modifications, our calculations are dependent upon six variables: Family Size, Race, Region, Age of Youngest Child, Age of Subject, and Divorce. The objective of this experiment is to determine, from our data, the variables with the greatest influence on divorce.

We will approach the subject by examining the effect of family size and race on divorce. **Figure 1** depicts differences in family size frequencies between the three race categories present in the data, as well as between subjects with and without a divorce in the time frame. We see that the mean family size of divorced subjects tends to be slightly smaller, indicated by the black dashed line, than that of undivorced subjects. The mean family size of the latter is always over 4 (for other, black, hispanic: 4.1, 4.1, 4.4), whereas the mean family size for the former is always under 4 (for other, black, hispanic: 3.5, 3.8, 3.9).

To begin trying to explain the variance in the data, we will perform a Principal Component Analysis (PCA). Because there are two columns in our data with categorical values, we will assign them symbolic numerical values. To proceed we create an sklearn Pipeline with One-Hot Encoding followed by a PCA. In doing this, we increase the number of variables in the data to 10, with a different variable for each race and for each region. The result of the PCA is indicated by the orange line in **Figure 2**. As we can see, over 90% of the variance can be explained by only two components. We can attribute this to a lack of scaling, however, because our age column contains much higher values than our One-Hot Encoded columns. To mitigate this phenomenon we apply a StandardScaler to the Pipeline. The scaled PCA is represented by the blue line; we can see that, when scaled, the data is much more difficult to reduce in dimension. It now takes six components to explain 90% of the variance.

We continue by performing a logistic regression on the data, in consideration of the boolean value type of our target prediction variable, divorce. We begin by splitting the dataset into two ¾:¼ train/test dataframes, stratifying on the divorce column. Next we create a new sklearn Pipeline using One-Hot Encoding, StandardScaler, and LogisticRegression with class weights balanced to 0.75:0.25, which is the inverse ratio of divorce to no-divorce present in the training data. Without this action, our model achieved 75% accuracy, however this value is a reflection of the ratio found in the actual training data. The model rarely predicted divorce and our recall was below 1%. After rebalancing, our model produces a 62% accuracy score and recalls of 63% and 59% for false and true predictions respectively. **Figure 3** portrays the regression coefficients for our model. We can see that family size was the most influential predictor of divorce with a weight of -0.4, an exhibition of a strong inverse relationship. The age of the youngest child in the subject's family exhibited an opposite effect on our prediction, with a higher age being more likely to predict divorce. We can conclude that divorce can be found much more frequently in our data among subjects with small family size, indicating larger family sizes as a buffer for a successful marriage.

Figures

Figure 1: Frequency of Family Size by Race and Divorce

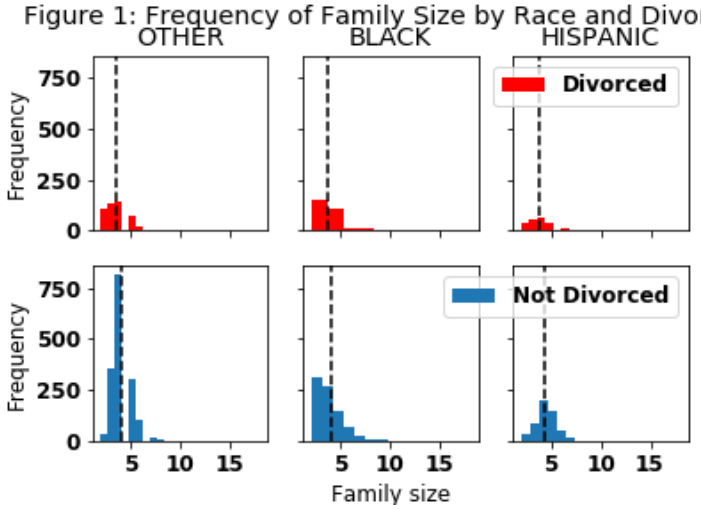

Figure 2: Principal Components for Divorce

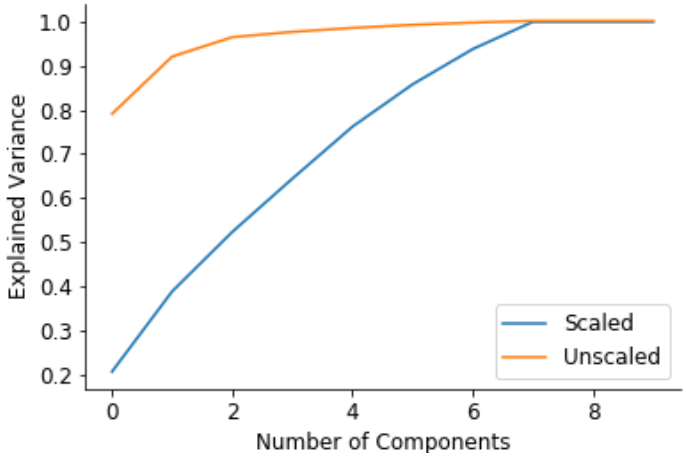

Figure 3: Logistic Regression Coefficients