|  |
| --- |
| Springboard Capstone One |

Predicting Divorce

A Social Problem Yields to Machine Learning Binary Classification

11/12/2017



# Contents

1. The Problem
2. The Data
3. Exploratory Data Analysis
4. Machine Learning: The Models
5. The Results
6. Results Analysis
7. Recommendations
8. References

### The Problem

Divorce is a major problem in our society. Strong marriages produce healthier children and greater economic security. Identifying factors which may contribute to marital dissolution is useful for couples thinking about marriage, family therapists, as well as married couples. Divorce prediction may be also be of interest to social scientists, credit card/loan approval companies as well as retailers looking to tailor advertising to families in crisis.

The Question:

Is it possible to use machine learning algorithms to predict the success or failure of a first marriage from a dataset which includes the results of a comprehensive survey of socioeconomic and family background?

### The Data

The National Survey of Family Growth is a survey conducted by the National Center for Health Statistics, a division of the Centers for Disease Control. The survey gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraceptives, and men’s and women’s health. Its stated purpose is “to understand trends related to fertility, family structure, and demographics in the United States.” [1]

The data used for this investigation was imported from the NSFG Cycle 6 survey which was conducted in 2002-2003. The Cycle 6 survey collected responses from 7643 female respondents varying in age from fifteen to forty-five. Male respondents were not included in this project. A total of 3087 responses were recorded or derived in the dataset. The participants answered questions about their socioeconomic and family background, as well as queries about marriages and divorces. For the machine learning test the data was limited to women who had been married at least once at the time of the survey. This brought the sample from 7643 women to 4,126 (54% of the original sample).

The distribution of respondents who had ever been married to those who had never been married was 4126:3517

evrmarry variable:

0 3517

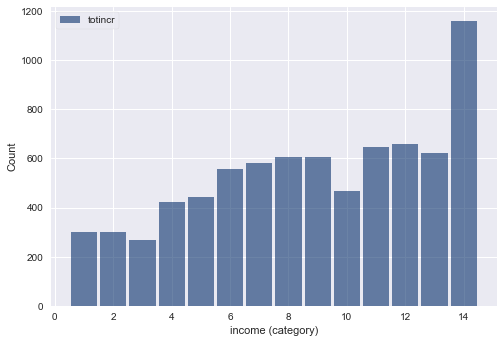
1 4126

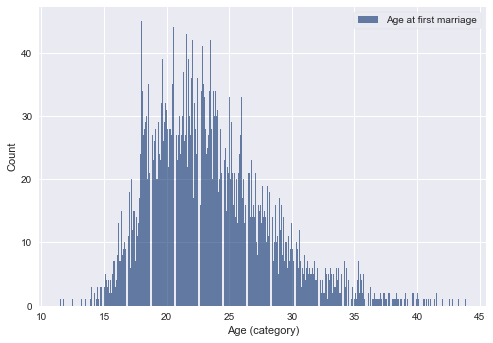
A detailed description of each variable can be found here:

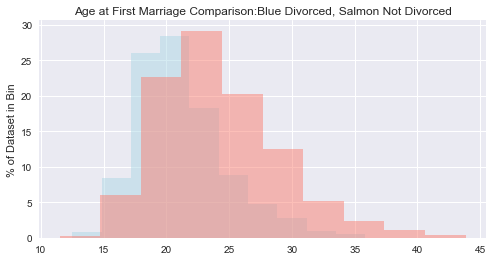
Link to online codebook: <http://www.icpsr.umich.edu/nsfg6/Controller?displayPage=femaleResp>

### Exploratory Data Analysis

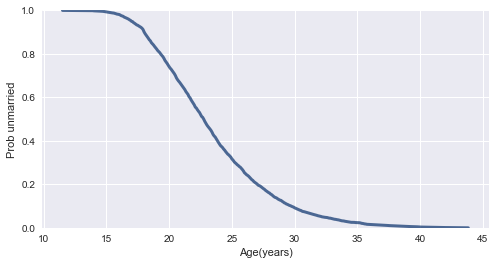
With so many variables to choose from (3087), the first order of business in exploring the data is to find the variables which would likely be most predictive of divorce status. Two variables which seem likely to have such value are: age at first marriage and income. The histograms below show the income level of the respondents and a comparison of the age at first marriage. The results show that the ever divorced cohort has more instances in the younger age brackets and less in the more mature age brackets.

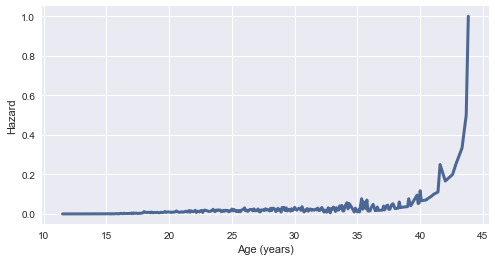






I also explored the concept of survival and hazard curves with regard to age at first marriage. The first graph is a survival curve which shows that the probability of having never been married decreases from 100% at around age 10 to close to 0% as the age approaches 45. The hazard curve shows the inverse of the survival curve, starting with a near 0% hazard of having ever been married to near 100% hazard as the age approaches 45.





### Dimensional Reduction

Many of the 3087 variables are redundant, and many them would predict with 100% accuracy the end of a first marriage since they pertain to second, third, fourth, and so on marriages.

Many of the variables contain a significant number of null values since they only pertain to a subset of the dataset.

After dropping redundant, non-predictive, and variables with significant number of null values, we are left with 78 significant variables for the analysis. Age at first marriage had a number of null values, but since I thought it was important to keep it in the dataset, I filled in missing values with the median age.

The top ten most predictive variables are:

evrdivorced 1.000000

ager 0.212419

rstrstat 0.141280

brnout 0.102303

better 0.101545

attndnow 0.100393

hispanic 0.078107

intctfam 0.077006

intact18 0.077006

rwant 0.074193

Since everdivorced is the target variable, it has 100% predictive value.

### The Models

Since the target variable is binary, the machine learning algorithms used will be classifiers. We will use the following models from the scikit learn machine learning library for our predictive accuracy results comparison:

Logistic Regression

Gradient Boost Machine

Decision Tree

Random Forest Classifier

1. Logistic Regression

Logistic regression fits a logistic model to data and makes predictions about the probability of an event (between 0 and 1).

1. Gradient Boosting Machine Classifier is an ensemble method that combines multiple decision trees. Gradient boosting works by building trees serially. Each tree tries to correct the previous one and more and more shallow trees are added to iteratively improve performance.
2. Decision Tree Classifier ask and learn a hierarchy of if/then else questions which lead to a decision.
3. Random Forest Classifier is an ensemble model which aggregates the results of a specified number of decision trees to predict the most likely classification. Decision trees are randomly generated, and each prediction is summed up the winning prediction is chosen.

### The Results

The prediction results for Logistic Regression Classification are:

accuracy score: 0.823244552058

roc\_auc\_score: 0.857295815667

**Logistic Regression Confusion Matrix**

[[602 43]

[103 78]]

**Logistic Regression Classification**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | precision | | recall | f1-score | support |
| 0 | | 0.85 | 0.93 | 0.89 | 645 |
| 1 | 0.64 | | 0.43 | 0.52 | 181 |
| avg / total | 0.81 | | 0.82 | 0.81 | 826 |

Gradient Boosting Classifier

accuracy score: 0.808716707022

roc\_auc\_score: 0.818553611921

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.84 | 0.92 | 0.88 | 630 |
| 1 | 0.64 | 0.45 | 0.53 | 196 |
| avg/total | 0.79 | 0.81 | 0.80 | 826 |

**Decision Tree Classifier**

DecisionTree: Area under the ROC curve = 0.7865808228053126

DecisionTree: Accuracy score = 0.674334140436

**Random Forest Classifier**

accuracy score: 0.817191283293

roc\_auc\_score: 0.842140413077

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.85 | 0.93 | 0.89 | 635 |
| 1 | 0.65 | 0.45 | 0.53 | 191 |
| avg/total | 0.80 | 0.82 | 0.80 | 826 |

**Model Accuracy Comparison**

|  |  |  |
| --- | --- | --- |
| Model | Accuracy\_Score | ROC\_Score |
| Logistic Regression: | 0.823244552058 | 0.857295815667 |
| Gradient Boosting Classifier: | 0.808716707022 | 0.818553611921 |
| Decision Tree Classifier: | 0.674334140436 | 0.7865808228053126 |
| Random Forest: | 0.817191283293 | 0.842140413077 |

### Analysis

## There are various ways to evaluate the success of a predictive model. Some of them are listed below:

* **Predictive Accuracy**: How many does it get right? This is generally the most important metric and is shown in the above chart. However, some other important considerations are:
* **ROC curve**: A receiver operating characteristic **curve**, or **ROC curve**, is a plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It measures the ability of the model to separate binary variables.
* **Speed**: How long does it take for the model to deploy? Since the dataset we are evaluating is not overly large, 5 x 1600, 5 columns, or variables, and 1600 rows, or instances, the time for the model to deploy is not much of a factor. For larger datasets this would be more of an issue.
* **Scalability**: Can the model handle large datasets? Given the size of the dataset, this is not a significant evaluative issue.
* **Robustness**: How well does the model handle outliers and missing values? The dataset comes already curated within certain parameters with no outliers or missing values. Therefore robustness is not an evaluative factor.
* **Understandability**: Is the model easy to understand? Linear classification models work by dividing the data into two classes and drawing a line separating the two classes, making them some of the easiest algorithms to understand and interpret.

**For this binary classification use-case, the accuracy score is the most important evaluative tool.**

### Recommendations

Apply tuning to the model parameters to improve accuracy.

Apply a balancing algorithm such as SMOTE to balance the classes between ever divorced and never divorced.

Investigate male responses as well as female to determine if there are significant differences in predictive variables between genders.

Add more recent cohorts to the dataset.

Investigate differences in predictive variables based on age cohorts.

### References

[1] NSFG Website: <https://www.cdc.gov/nchs/nsfg/index.htm>