ISyE 601 Project Preliminary Report: Divorce Analytics

Akhilesh Soni (9079896453), Zach Zhou (9071567169)

Problem Statement

We aim to apply machine learning models towards predicting whether a couple will divorce or not based on responses to the Divorce Predictors Scale (DPS) of Yöntem and Îlhan (2018). We would also like to identify the most telling questions on the survey, in an attempt to explain what factors are key to understanding the likelihood of divorce.

Data

We will use the Divorce Predictors dataset. The dataset consists of 170 couples' responses to the DPS; 84 are divorced and 86 are married. The features correspond exactly to the 54 questions appearing on the survey. Each feature is a response in the form of a discrete rating on scale of 0 to 4, where 0 means "strongly disagree" and 4 means "strongly agree." As expected, we found in our EDA that many of the features are highly correlated. The target is a binary indicator of whether the couple is divorced.

Methods

Our first task is to identify effective classifiers for predicting whether a couple is divorced based on responses to the DPS. Classification models we will apply include logistic regression, KNN, decision trees (CART and optimal classification tree (OCT)), random forest, and support vector machine. As of this report, we have applied logistic regression, and decision trees (both CART and OCT).

Our second task is to identify the most important features in the dataset, i.e., questions which play the most crucial role in explaining whether a couple will divorce. We plan to use correlation statistics and selection methods to identify most significant features and train our model on this subset of features. We plan to study performance of the model with reduced features against the model inclusive of all the features.

Initial Findings

EDA

We show mean values of all the attributes for divorced and married couples in Figure 1. We observe that divorced couples seem to respond to all questions on the survey higher than married couples.

Next, we study correlation between different features (or attributes). Dark red color in Figure 2 shows that there is a high correlation while blue color means there is little correlation. Clearly,

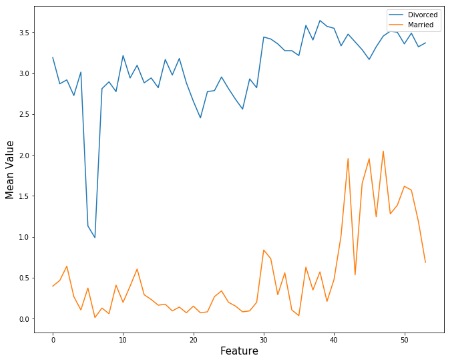


Figure 1: Mean attributes values

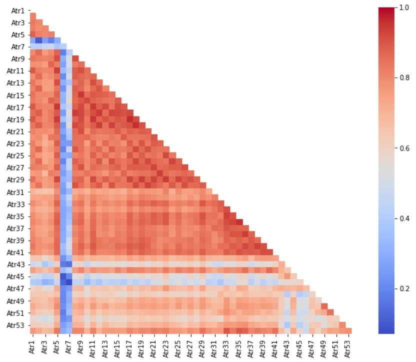
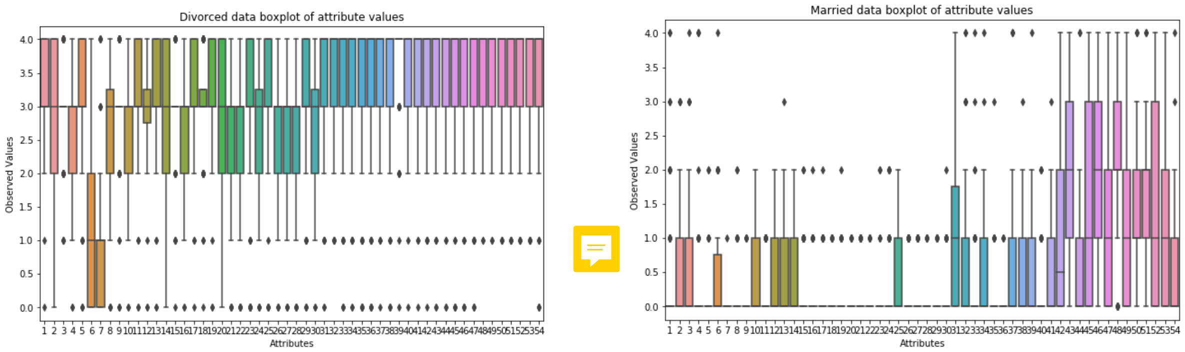


Figure 2: Correlation between attributes

we can see that there is a high correlation between most of the features. This behaviour indicates that there might be some key features in the model through which we might be able to predict the outcome i.e. whether the individual is still married or divorced.



(a) Divorced individuals

(b) Married individuals

Figure 3: Box plots to analyse attribute values

We next study the box plots of divorced and married individuals separately in order to get an idea of variance in the responses for each attribute. For most of the features, we see that interquartile range (25-75 percentiles) in both divorce and married classes lie within two points on DPS scale which suggest that both married and divorced couples are likely to respond in somewhat similar ways to the quiz.

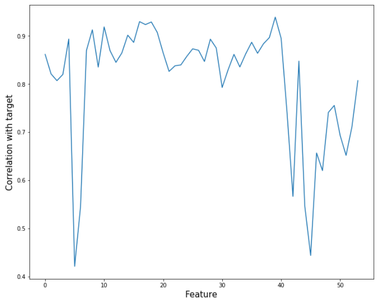


Figure 4: Correlation plot between attributes and target variable

Feature Filtering We plot a graph of each attribute's correlation with the target variable as shown in Figure 4. This figure suggests that there are few features which aren't highly correlated with the target variable. We can set a threshold, say *thr*, to filter features and then train our models with attributes having correlation with the target variables higher than *thr*.

Initial Results

Prediction

To perform prediction, we do not apply scaling as all features are on the same scale, from 0 to 4 (with the exception of OCT, which requires all features are in $[0, 1]$). We apply a 75/25 train/test split. We applied the following classification models in the described manner:

- Logistic regression: We train two models using ℓ_1- and ℓ_2-regularization. For each model, we search over 10 lambda values using 10-fold cross validation and AUC as the performance metric. We obtain the best lambda values and evaluate on the test set.
- CART: We perform a grid search cross validation over the max depth and alpha (cost-complexity pruning) parameters, searching over 5 values for each. We obtain the best hyperparameter values and evaluate on the test set.
- OCT: We only train with a max depth of 3 and lambda (cost-complexity pruning parameter) of 0.1. We then evaluate on the test set.

All models have been largely successful.

Prediction Accuracy		
Model	**In-sample accuracy (%)**	**Out-of-sample accuracy (%)**
Logistic regression (ℓ_1)	100	99.13
Logistic regression (ℓ_2)	100	100
CART	99.21	99.35
OCT	99.21	99.35

Completion Plan

The following tasks remain:

- **Develop an integer programming model for weighted KNN.** The problem is to find weights for each feature such that in-sample accuracy is maximized given those weights.
- **Identify which features do the best at explanation.** Unfortunately, the vast majority of the features are highly correlated with one another. While removing all highly correlated features in the dataset (by inspection, i.e., looking at the correlation plot and removing features that are mostly red, leaving around 10 features remaining) leads to no two features being correlated with one another, it seems the resulting dataset is not suitable for either explanation or prediction.