Predicting Relationship Quality from Relationship Attributes in 2022

Jasjot Parmar, Eugene Tse, Jade Chen, and Johnson Leung

Table of contents

1 Summary

In this project, we used a dataset on based on survey responses to common relationship questions to develop a logistic regression model to classify relationships into one of five relationship quality statuses: excellent, good, fair, poor, or very poor. The model developed below uses common relationship-based features such as whether the subject is married or not and how many children the subject has in the relationship.

2 Introduction

Relationship quality classification is an important topic for couples to be aware of, particularly when trying to maximize the amount of satisfaction each partner receives from the relationship. Accurate relationship quality classification allows couples in relationships to assess the quality of their relationship to develop better targeted strategies to improve or maintain that relationship quality. It is often difficult for couples to estimate the perceived quality of their relationship, so we investigate below if our machine learning model can correctly classify the quality of a relationship, based on common relationship attributes. The below analysis tries to answer: How well do relationship characteristics such as age, income category, marital status, relationship duration, and number of children predict relationship quality?

The dataset contains 1293 survey responses of common relationship characteristics such as the income category of the respondent and their estimated relationship quality (our target to predict). This dataset (Diverse Data Hub (n.d.)) is originally based on the How Couples Meet and Stay Together survey (Rosenfeld, Thomas, and Hausen (2023)). We accessed the CSV version of this dataset directly from CRAN (R Core Team (n.d.)). The README.md file contains seperate conda lock files that explain how to run the environment if you want to run the code. We used principles from the Reproducible and Trustworthy Workflows for Data Science textbook (Timbers et al. (n.d.)) on conda lock files to manage enviornments.

In our analysis below, we investigate whether a logistic regression model can correctly classify relationship attributes into one of five relationship quality statuses: excellent, good, fair, poor, or very poor.

3 EDA

We start by initially confirming our data was read in to a pandas DataFrame object, as shown in Table **??**.

	subject_age	subject_education	subject_sex	subject_ethnicity	subject_income_category	subject_e
0	53.0	high_school_grad	female	white	35k_40k	working_
1	72.0	some_college	female	white	75k_85k	working_
2	43.0	associate_degree	male	white	75k_85k	working_
3	64.0	some_college	male	white	75k_85k	working_
4	60.0	high_school_grad	female	black	75k_85k	working_

We can see from the distribution of the Relationship Quality categorical variable, that the dataset contains imbalanced classes, with a very large number of respondents reporting excellent or good relationship quality and a much lower number of respondents reporting fair, poor, and very poor relationship quality.

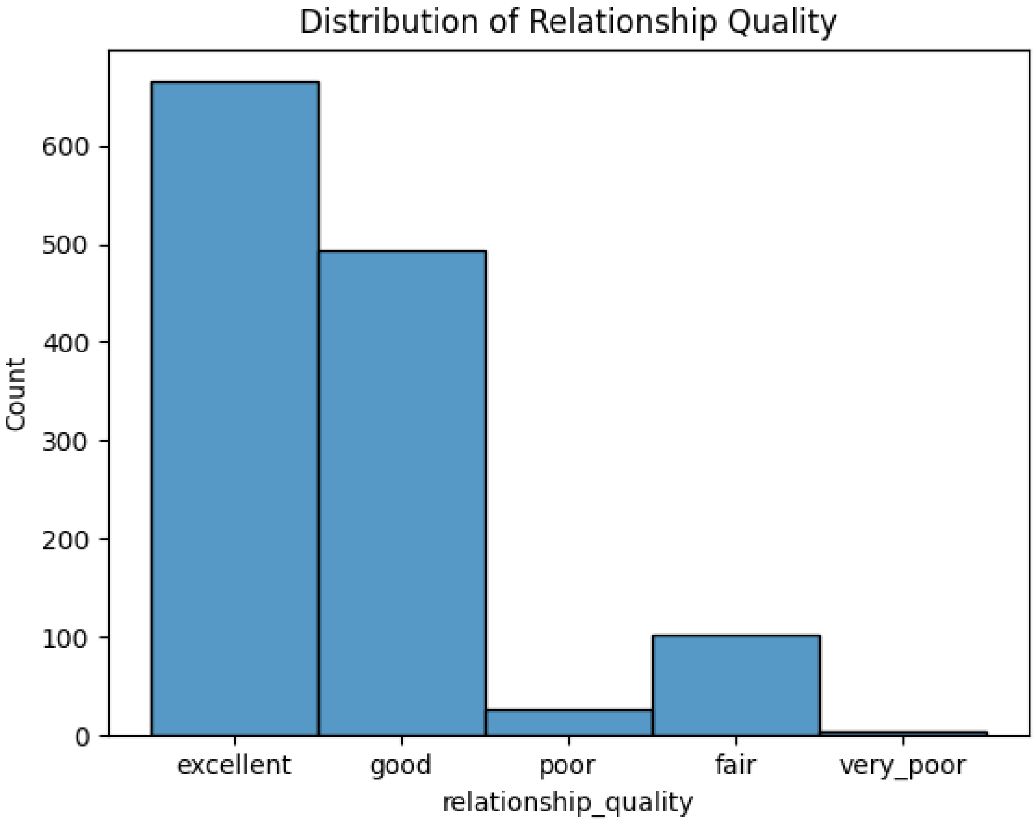

Figure 1: Relationship Quality Score Distributiom

Our numeric predictor / input features show a high correlation between subject age and relationship duration with a ρ value of 0.736, which means that as the subject's age increases, their relationship duration increases. Subject age and (number of) children show a weak negative correlation of -0.326, indicating that as subject age increases, the reported number of children slightly decreases.

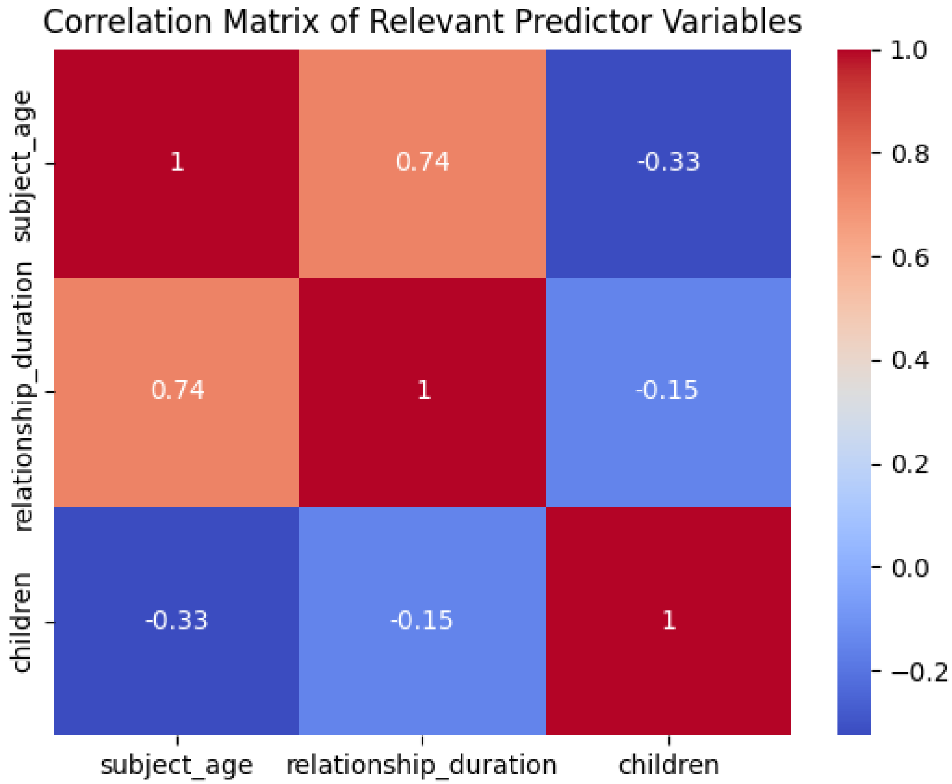

Figure 2: Correlation Heatmap

The distribution of income categories show that we have a left skewed distribution, with most respondents making over 50k per year. For respondents who earn >= 50k / year, income distribution between 50k and 250k+ seems to be roughly uniformly distributed, showing that there is a pretty even spread of incomes between respondents as income passes 50k.

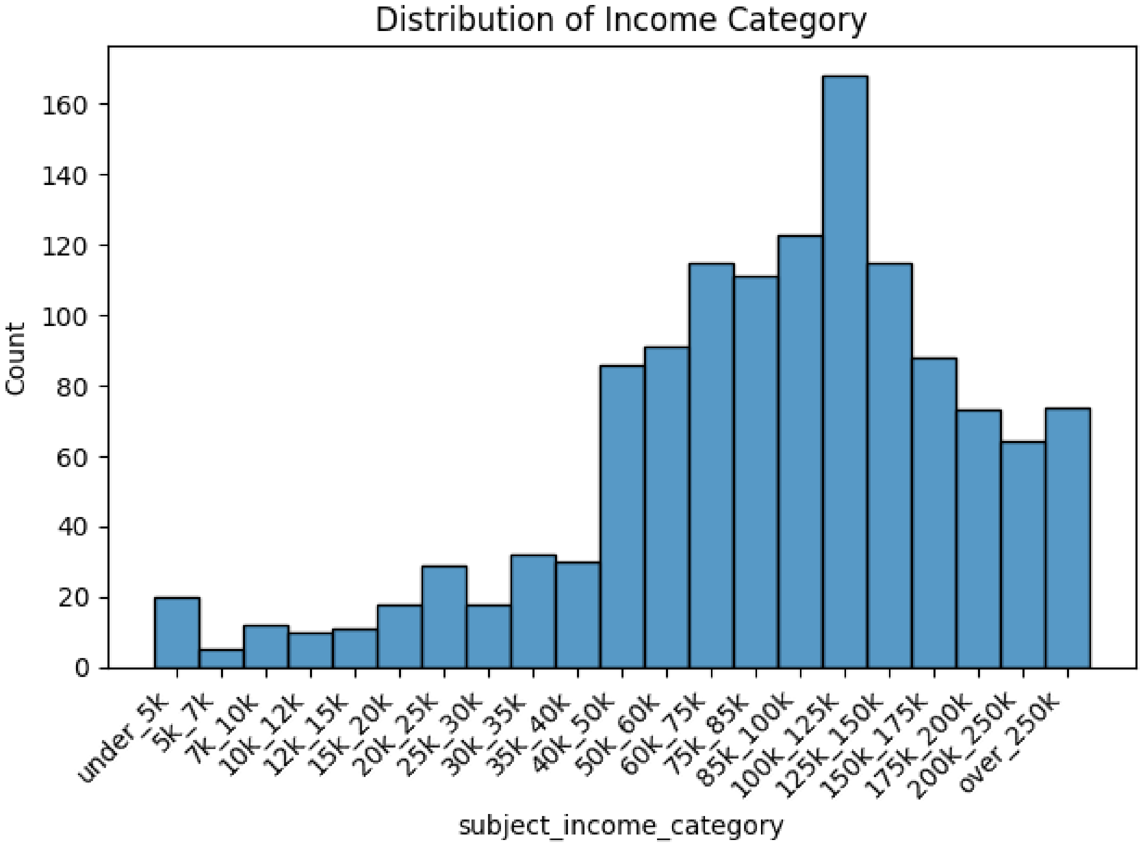

Figure 3: Income Category Distribution

We then split up the data into the relationship features that we want to predict relationship quality with. Input features include: Subject Age, Subject Income Category, Martial Status, Relationship Duration, and Number of Children, before splitting the data into train and test splits. We then conduct simple data cleaning through changing relevant numeric features such as age and number of children into integers, before reordering the income category feature to be ordered in ascending order by income.

Numeric features have different scales, with age having much larger values than relationship duration and number of children. Therefore, Standard Scaler is applied to numeric features so all numeric features contribute equally to the logistic regression model. Ordinal features such as subject income category are converted to ordinal categories, as their categories have an order based on the income of the subject. Categorical features such as marital status are one hot encoded, resulting in one column indicating martial status or not (0 / 1). Each transformation is wrapped in a column transformer.

A scikit-learn pipeline is used to preprocess and train the model on the training data in one step.