ECON 1660 PS4: Prediction

Josh Baum, Amy Li, Jake Sokol
October 12, 2021

Models

Model 1: Date and Time

MSE: 86.29
Our first model uses date and time as the main considerations to predict days_until_funded. We chose this because we thought some days and times of loan posting would engender more generous responses. In particular, we created additional variables to represent the day of the week of the posting, a binary variable representing whether the posting was made during waking hours (defined as 9am-5pm), and a binary variable representing whether the posting was made during the holiday season (defined as the 15 days before and after the new year). We thought the day of the week may produce interesting patterns in days_until_funded because some days of the week may produce more traffic (i.e. weekdays or otherwise). We thought the holiday season would produce more traffic because people may be more generous then. Lastly, we thought there would be more traffic if the posting was made during waking hours because that is when people are more often online. To do this, we parsed the loan posting variable to find the dateTime characteristics.

Model 2: Loan and Repayment

MSE: 91.45
Our next model uses loan and repayment information to determine days_until_founded. We thought this was relevant because certain loan and repayment amounts/schemas would may be more likely to be contributed to. For example, risk-averse people may be more likely to make smaller loans that are repaid quickly. With that in mind, we created a binary variable to represent whether the loan amount was greater than or less than the mean loan amount and another variable to represent whether the repayment term was greater or less than the mean repayment term. We decided to use a binary variable because we thought adding more levels would not provide significantly more additional information than simply splitting on the mean.

Model 3: Country and GDP

MSE: 89.41

Our next model delves into how the country of the loan affects days_until_funded. We thought this would be a good variable to delve into because certain countries may be more or less likely to be seeking loans and certain countries may be more or less likely to engender loans from other people. In particular, we thought that countries that have lower GDPs would be correlated with lower days_until_founded because people may feel more generous towards them. We created categorical variables for the country region and the country subregion, as well as a binary variable representing whether the GDP of the country is “small” (defined as below the world mean GDP in 2016). We decided to use a binary variable for GDP rather than levels because we thought that would be sufficient to encapsulate our perception of countries as “lower or higher” GDP. That is, we doubted that a lender would be able to differentiate between more discrete differences in GDP (ex. Lowest 25th percent) and would instead associate with a country as wealthy or not. Lastly, we used the mean GDP from 2016 because that was the most recent data that we could find.

Model 4: Age, Gender, Name, and Pictured

MSE: 85.07

Our next model combines standard demographics like age and gender along with the name variable and the pictured variable to determine days_until_funded. This seemed like a reasonable approach because there are naturally inherent biases along demographic lines that we would like to look into further. First, we used the binary variables for gender and pictured. Next, we created a binary variable representing whether the name of the loan requester was popular or not. The intuition for this was that people may be more likely to lend to people whose names are familiar. We determined the popularity of the name of the loan requester by finding online a list of the most popular names in the US since 1980. We chose the year 1980 because we thought that people born since then would be most likely to go on Kiva. We chose names from the US because we made the assumption that people most likely to make loans on Kiva are from the US. Lastly, we created a categorical variable to represent the age range of the loan requester. We decided to use four levels because it is reasonable to assume that different age ranges (unknown, young, adult, and unknown) behave differently regarding loans for various financial and social reasons.

Model 5: Description Sentiment

MSE: 90.62

Our last model performs sentiment analysis on the description section of the data to determine days_until_funded. Naturally, lenders will be more or less inclined to provide a loan based on the sentiment or topic of the description. In our sentiment analysis we created three categories of words that we associate with family values (ex. “Children”, “family”), entrepreneurship (ex. “Entrepreneur”, “business”), and sympathy (ex. “Necessary”, “single”). We then parsed the descriptions of each datapoint and counted how many words they scored in each category. From there, we created three binary variables for each sentiment category representing whether they had scored in the corresponding category or not. We thought that a binary variable would suffice because the existence of some sentiment-inducing words should be enough to make a lender feel a certain way.

Model Chosen

We chose to use the fourth model (age, gender, name, and pictured) because this had the lowest MSE. See our output csv file for our predictions!