**Predicting the Probability of a Loan being Crowdfunded on Kiva**

**What is Kiva?**

Kiva Microfunds is a non-profit organization that allows people to lend money via the internet to low-income entrepreneurs and students throughout the world. Kiva's mission is "to connect people through lending to alleviate poverty." Since 2005, Kiva has crowd-funded more than a million loans, totaling more than $1 billion. The Kiva platform has attracted a community of well over a million lenders from around the world.

**How does the business model work?**

Kiva is a crowdfunding website where people all over the world can make interest free loans to entrepreneurs worldwide. These Funds are posted and disbursed by Kiva’s field partner organizations worldwide.

**What is the problem we are trying to solve?**

At any point in time there are 4000 + entrepreneurs who need funding. However, if loans are not fully funded within 30 days of posting, **they expire, and no one gets the money.** In this scenario, the field partner has to make the loan themselves, increasing a risk of loss exposure on their books

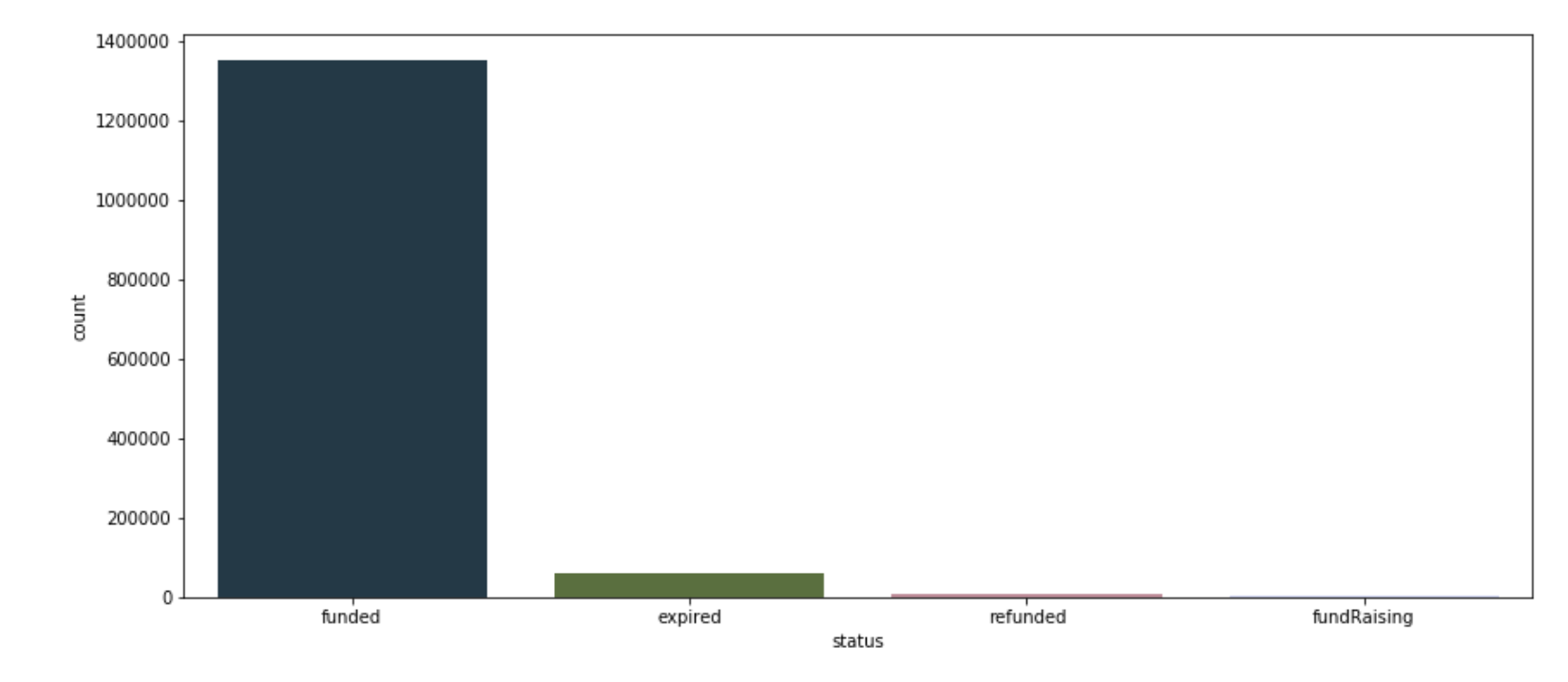
**How will this analysis help?**

This analysis will help them post the loans most likely to get funded on Kiva and reduce exposures of losses on their books. By learning about which loans are mostly likely to get funded, they can focus their energy on promoting those aspects of the loan on the webpage.

**Exploratory Analysis:**

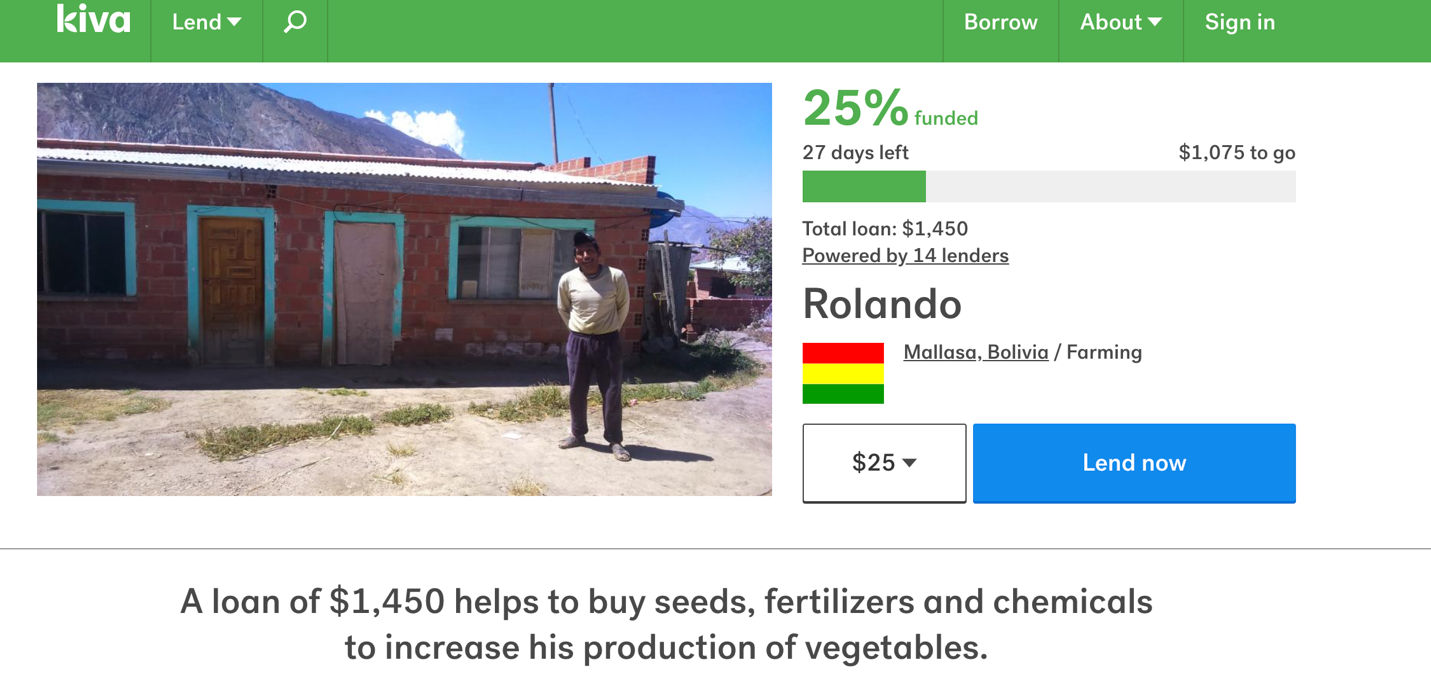
I began by exploring various facets of the dataset itself. I wanted to see , whats the proportion of loans on Kiva that are funded?

Composition of funded vs unfunded loans

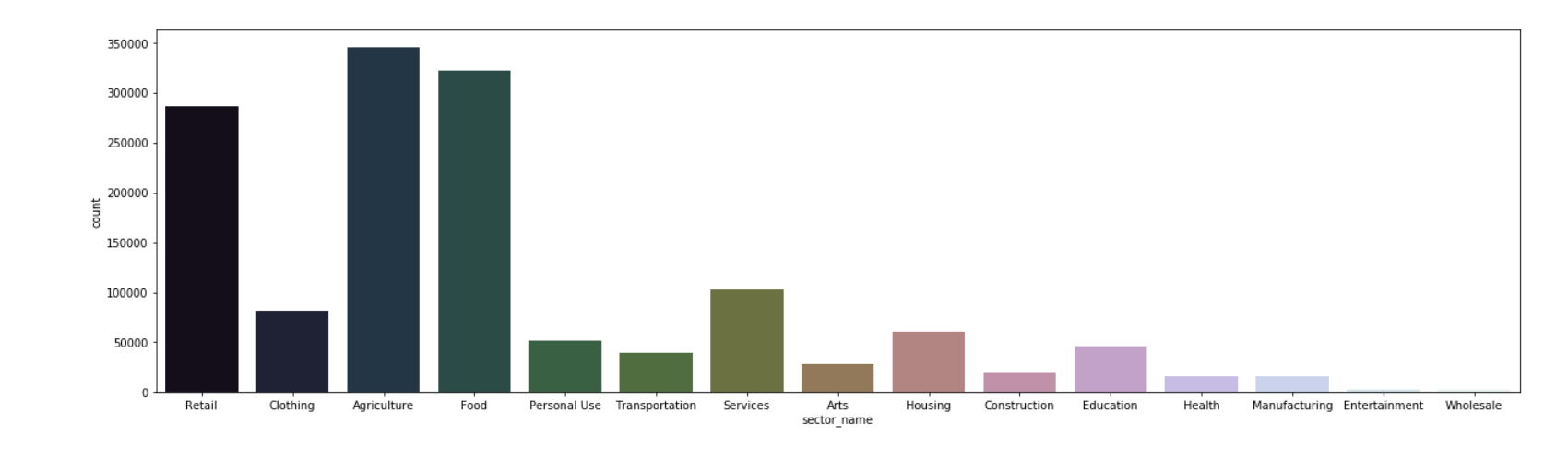
****

Most loans in this sample are funded, skewing our data considerably. This makes me think that this is a quirk of the dataset provided by Kiva, as it unlikely they have such a low expiration rate.

Just by browsing the Kiva.org website, I can intuitively tell that many of loans requested are for purposes of agriculture and growing food. An example is this:

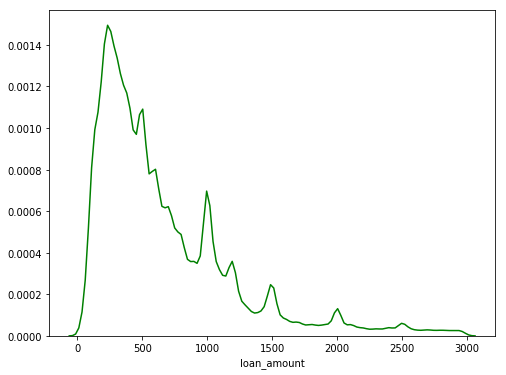


So what does the data tell us?

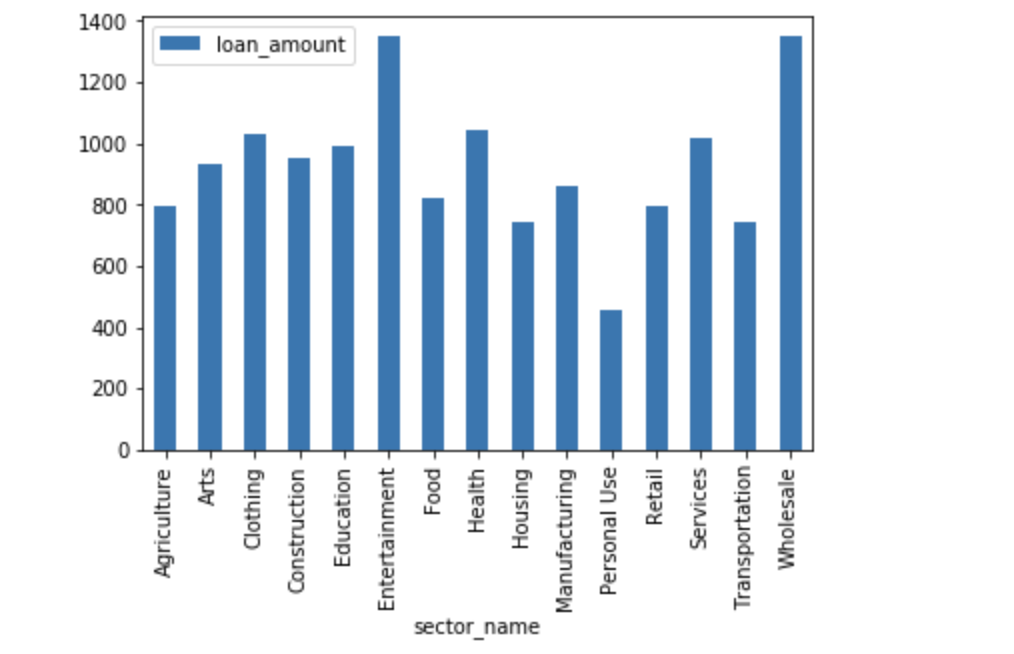


Agriculture and Food are indeed the most requested loans, followed by retail. Intuitively this makes sense as many of Kiva’s field partners are focused on agricultural loans, leading to a skewness in this data.

What do average loan amounts look like?



Overall, the mean loan amount is $800. But what does this distribution by sector?

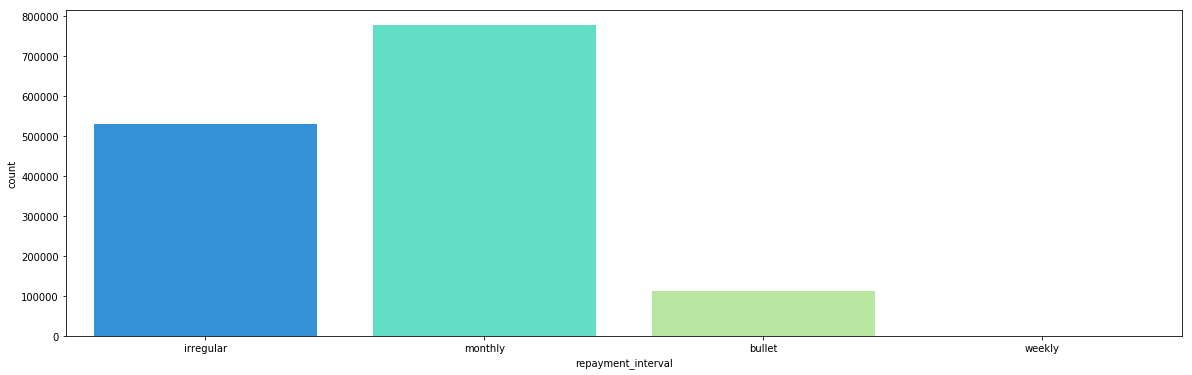


Agriculture and Food have requested amounts on the lower end of the spectrum, their sheer volume skews the entire average to be around $800.

Do Repayment intervals matter?

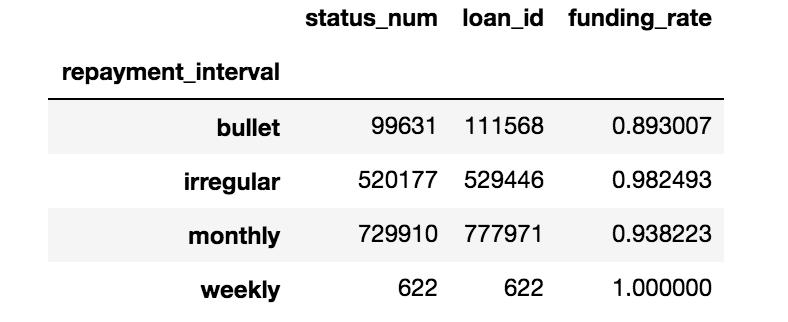
Kiva allows borrowers to select various repayment intervals, these include irregular (based on ability to pay back), monthly, weekly or bullet (lump sum paid towards the end of the loan’s life).

What does this distribution look like?



Repayment intervals skew monthly, but sometimes they can be irregular due to the seasonality and variation in farming products.

Does this make a difference in funding rate? My suspicion is that this may be a valid input to the model as people prefer regular payment intervals.

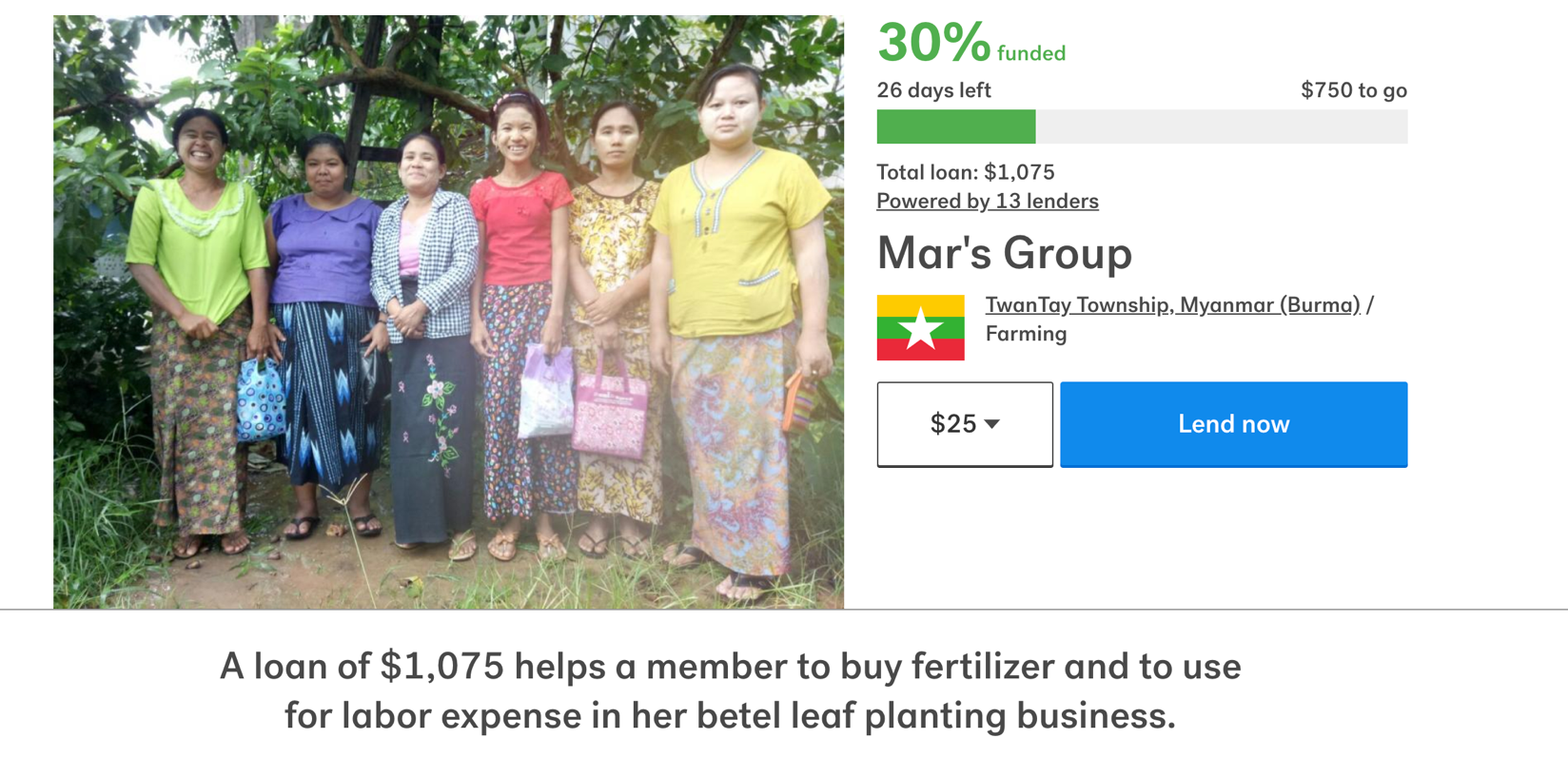


Voila! It looks like while people are actually OK with irregular intervals, they do not prefer bullet loans as this means their money is locked up for a long time.

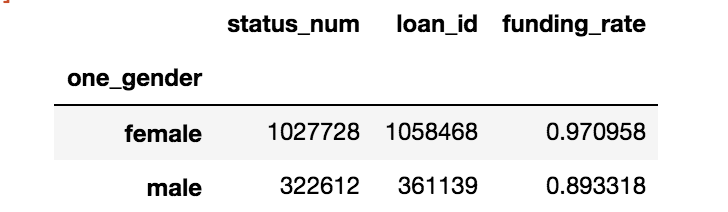
Does Gender matter?

Anecdotally, women tend to fund at a higher rate than men. Can we confirm this with our dataset?

This is an interesting conundrum as Kiva has many “group loans” where many people apply for a larger loan amount as a group. An example of this is below:



In this case, the “gender” of the borrower is a series of strings. I got around this problem by assigning the group loan the “dominant” gender. This means that if a group has 3 women and 2 men, the gender assigned to this group will be “female” as they make up more than 50% of the group

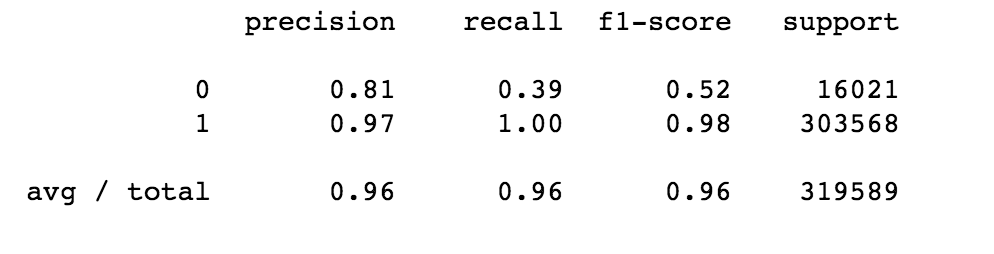
****

**Wow, this is a massive difference! Females fund 8 percentage points higher than men!** The ‘one\_gender’ variable will definitely be an important input to our model.

**The Models:**

1. Logistic Regression

**Logistic regression** is the regression analysis to conduct when the dependent variable is binary. In this case, loan\_funded =1 and expired=0.

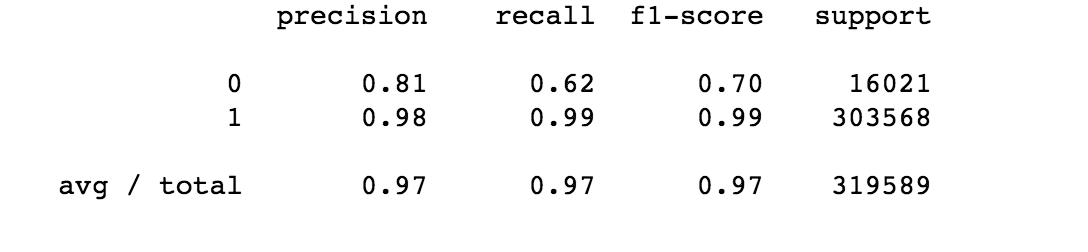


This model is good at determining if a loan will be funded, but not good at determining which loans will not fund. this is very likely to the class imbalance issue mentioned earlier.

The model correctly identifies 40% of all expired loans as "not funded"

2. Random Forest

**Random Forest** is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it’s simplicity and the fact that it can be used for both classification and regression tasks.



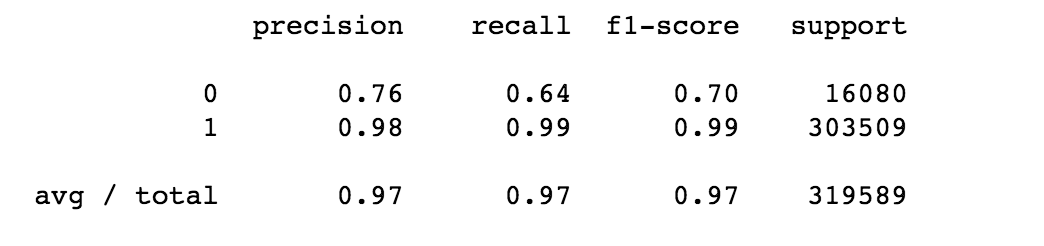
This model is better than the previous one at determining which loans won’t be funded. It correctly identifies 62% of expired loans as “not funded”

**Class Imbalance:**

However, because we identified class imabalance as a big issue, I use the SMOTE algorithm to recreate a dataset that is more finely balanced. SMOTE is defined as follows:

Synthetic Minority Oversampling (SMOTE) works by creating synthetic observations based upon the existing minority observations (Chawla et al., 2002). For each minority class observation, SMOTE calculates the k nearest neighbors. Let’s assume we consider the crossed square and pick the 5 nearest neighbors represented by the black squares. Depending upon the amount of oversampling needed, one or more of the k-nearest neighbors are selected to create the synthetic examples.

After SMOTE, our results change slightly:



As you can see above, our recall rate for unfunded loans increase by 2% and precision dropped by 4%. Unfortunately, this is as good as it gets. This model is good at predicting which loans will be funded, but not good at predicting which loans will not be funded.

**Conclusions & Recommendations:**

* **Field partners should prioritize posting smaller loans amounts** to increase probability of funding. It may also help to break up one loan into smaller pieces if it is a large amount.
* **Field partners should post loans that are shorter in length to increase chances of funding.** Longer loan durations are less likely to get funded as this locks up lenders’ money for a long time.
* **Journal entries are significant!** The partner should make an effort to write these regular updates as lenders are more likely to feel connected to the borrowers this way. **This is also a great way to get lenders engaged in funding other loans from the same field partner.**
* **Females are more likely to be funded than males!** Partners should prioritize posting those loans.