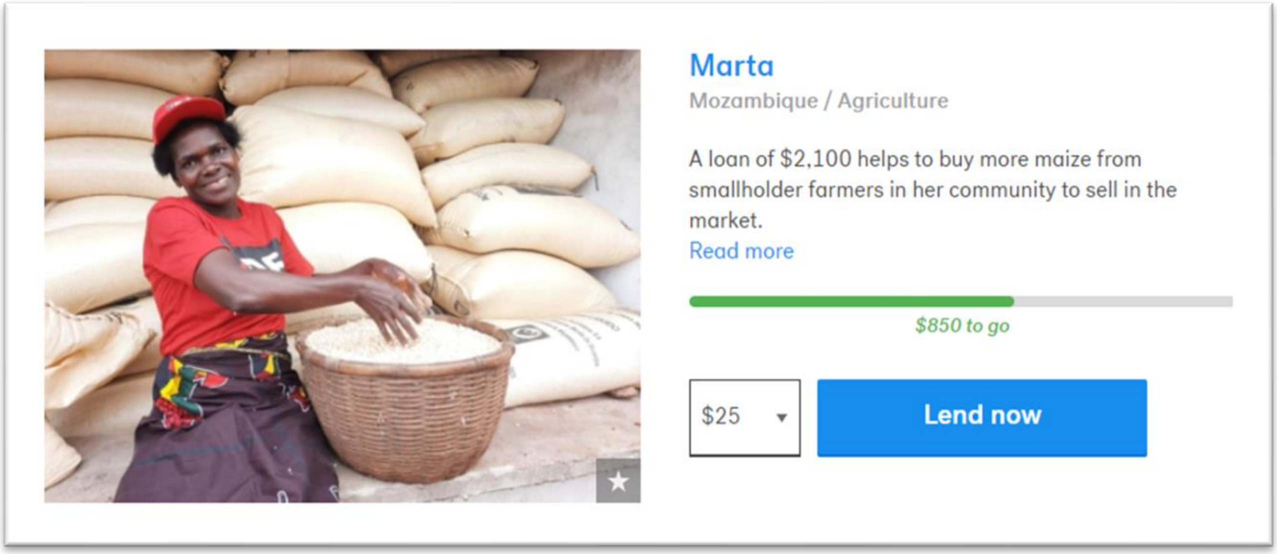
Microfinance Funding

A Datascience Capstone Project

Background:

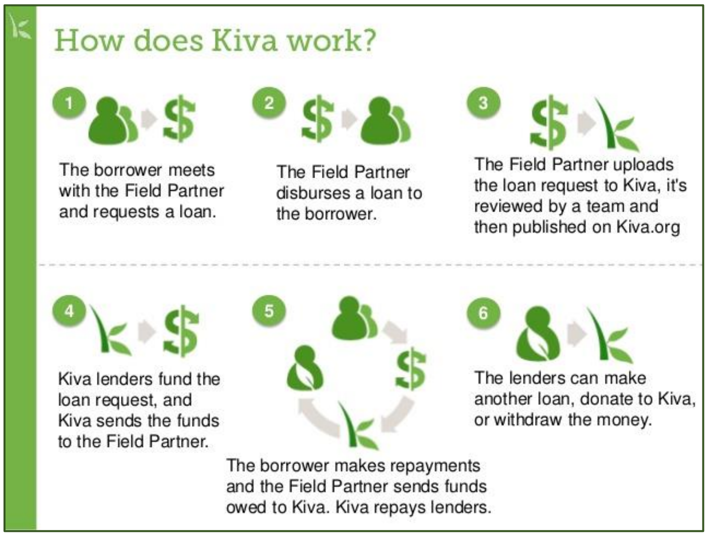
Kiva is a non-profit that provides an online platform that connects individual lenders to low-income entrepreneurs around the world to finance microloans. Each borrower has a profile page that shares their story with a picture, information about how the loan will be used, the amount requested and details about the loan (length, repayment schedule, funding model, etc.). The loans are posted for 30 days and during that time lenders can contribute in $25 increments until the loan is fully funded.



Marta

Mozambique / Agriculture

A loan of $2,100 helps to buy more maize from smallholder farmers in her community to sell in the market.



How does Kiva work?

1 The borrower meets with the Field Partner and requests a loan.

2 The Field Partner disburses a loan to the borrower.

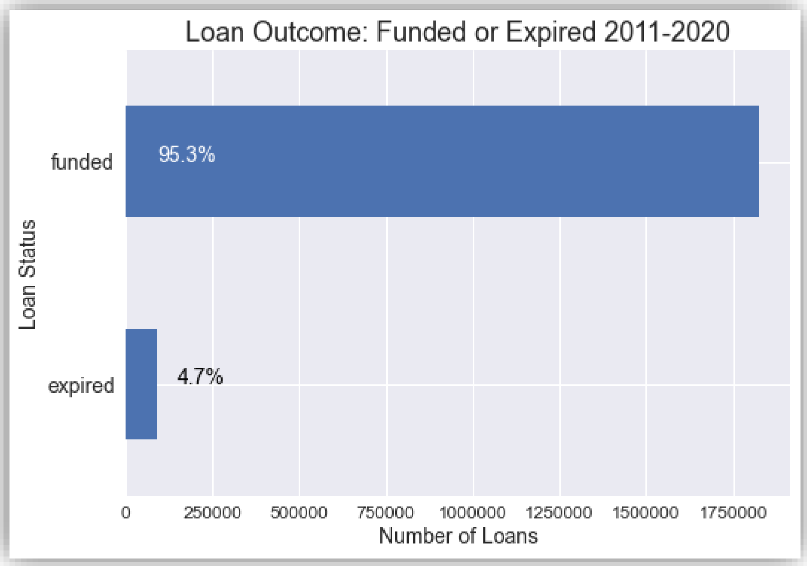
3 The Field Partner uploads the loan request to Kiva, it's reviewed by a team and then published on Kiva.org

4 Kiva lenders fund the loan request, and Kiva sends the funds to the Field Partner.

5 The borrower makes repayments and the Field Partner sends funds owed to Kiva. Kiva repays lenders.

6 The lenders can make another loan, donate to Kiva, or withdraw the money.

The median loan is for approximately $500, funded by 14 individual lenders and 95% of the loans are fully funded during the funding window. Each loan has characteristics that make it potentially appealing to different lenders. Distinguishing features include borrower demographics, how the loan will be used, and loan characteristics.

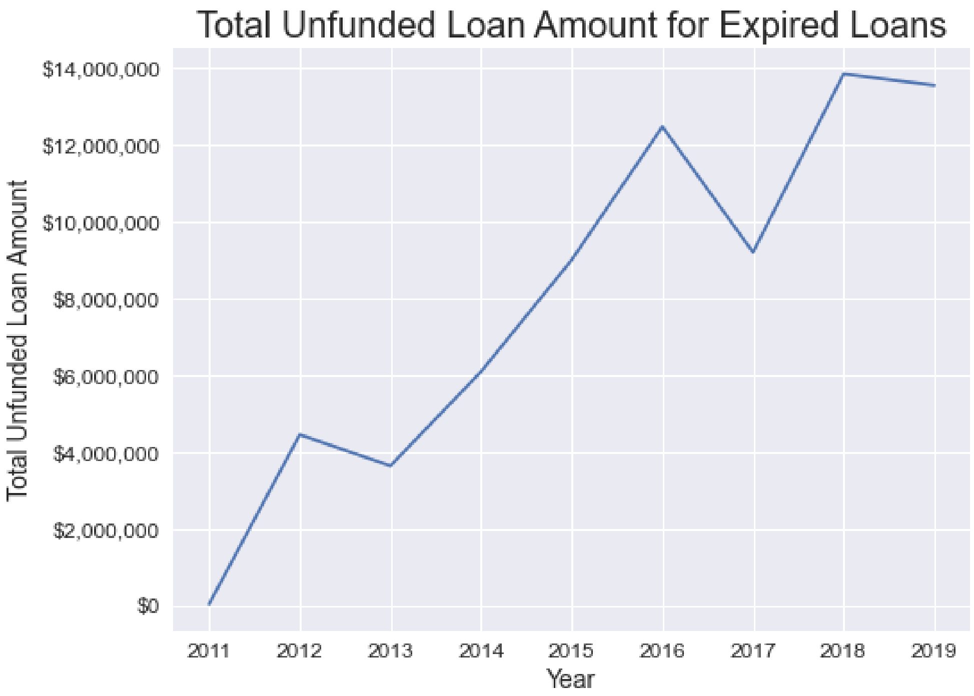


Problem Overview

In Kiva's current model all borrowers are provided with funds prior to the loan appearing on their platform for funding. This leaves them responsible to financially support and assume the risk for any individual loans that are not fully funded. As can be seen in the figure below, since 2016 Kiva has been responsible for at least $9M of expired loans that were under funded. The goal of this project is to minimize Kiva's overall financial risk by increasing the number of loans that are fully funded. To maximize the number of funded loans this project aims to:

- Predict the specific loans that will not be fully funded during the posting period.
- Understand the factors that are driving those loans to be less attractive to lenders.

Creating this model will allow Kiva to proactively help borrowers structure their loans as well as provide strategic placement (marketing) of loans on their platform to maximize the loans that are funded.



Data Overview

The data for this project comes from Kiva via their developer tools website which produces a nightly snapshot of detailed information about all loans since 2006. [https://www.kiva.org/build/data-snapshots] The entire dataset is over 2M records but was divided into smaller subsets for model development. Two separate approaches were used: utilizing the last 6 months of 2019 loan data (~100K loans) as well as taking 100,000 random loans from the entire dataset.

Data Features

The dataset provides a wealth of information that characterizes each one of the loans. The dataset can be subdivided into four sections: technical data, loan purpose, borrower profile, and funding information.

Technical Data

Each loan includes technical information: loan amount, lender terms (length of the loan), repayment interval, loan currency, currency exchange coverage rate (who is responsible for losses due to fluctuations in currency exchange rates), the distribution model (direct to borrower or through a partner) and the partner ID.

Loan Purpose

Each loan is described in detail in the dataset. The dataset includes a written free text description of how the loan will be used, and categorization of the loan, first into a sector (Agriculture, Retail, etc.) then further into an activity (farming, phone use sales, etc.) and finally into the most specific category of loan use (to buy farm supplies, to provide mobile services). Each loan was then tagged by Kiva to indicate popular types of loans like “woman owned biz” or “first time loan.”

Borrower Profile

Information was also provided about the borrower(s) (loans can be for a single borrower or multiple borrowers). The borrower's country and town name are included as well as the borrower's name(s) and gender(s). There was also information about the borrower's profile on the Kiva website including if the profile had a picture and/or a video, if the borrower was in the picture, and how many journal entries had been posted.

Funding Information

The funding process for each loan was also described in a series of four different time stamps: the date the loan was disbursed to the borrower, the date the loan was posted on the Kiva platform, the date the loan was planned to expire (funding period would be over), and the date when the loan was fully funded. The total amount that was funded and the number of lenders was included as well as an indicator if the loan was funded or if it expired prior to reaching the total loan amount.

Data Wrangling

Two subsets of data were extracted from the original dataset to be explored and modeled. The first dataset consisted of all the loans from the last 6 months of 2019. The second dataset was ~100,000 loans selected at random from the entire dataset. The two datasets were selected to ensure individual models could be applied to different subsections of the data and still provide valuable results.

Missing Data

The dataset overall had 2.7% missing cells and the majority of the missing-ness was explainable. For example, the loan tags field had missing cells for loans that were never tagged. The VIDEO_ID field which was close to 99% missing was dropped.

Feature Review

Each feature of the data was reviewed for the correct data type, outliers and to ensure that the range of values contained in the set were valid. All the date features were converted to date-time format. Less than 0.5% of the loans were categorized as "refunded" rather than expiring or being funded. These loans were dropped because the goal of the model was to distinguish loans that would expire before being fully funded and most of the associated data with these loans was missing.

Additional Features

To provide the model with more information to distinguish which loans will be funded additional features were created. The created fields were as follows:

- Total borrowers per loan
- Female only loan -- 1 if all borrowers were female and 0 if mixed or all male borrowers.
- Average amount loaned per lender
- Time to fund loan
- Funded percent

Several features were also expanded or modified to provide the model with more discrete, specific information.

- Loan tags – each individual tag was made into a discrete feature.
- Currency policy and currency exchange coverage rate features were combined to create a single feature that indicates the amount of currency fluctuation loan partners are responsible for
- Month – the month the loan was posted was extracted into a new feature to identify if there is seasonality for loan funding.

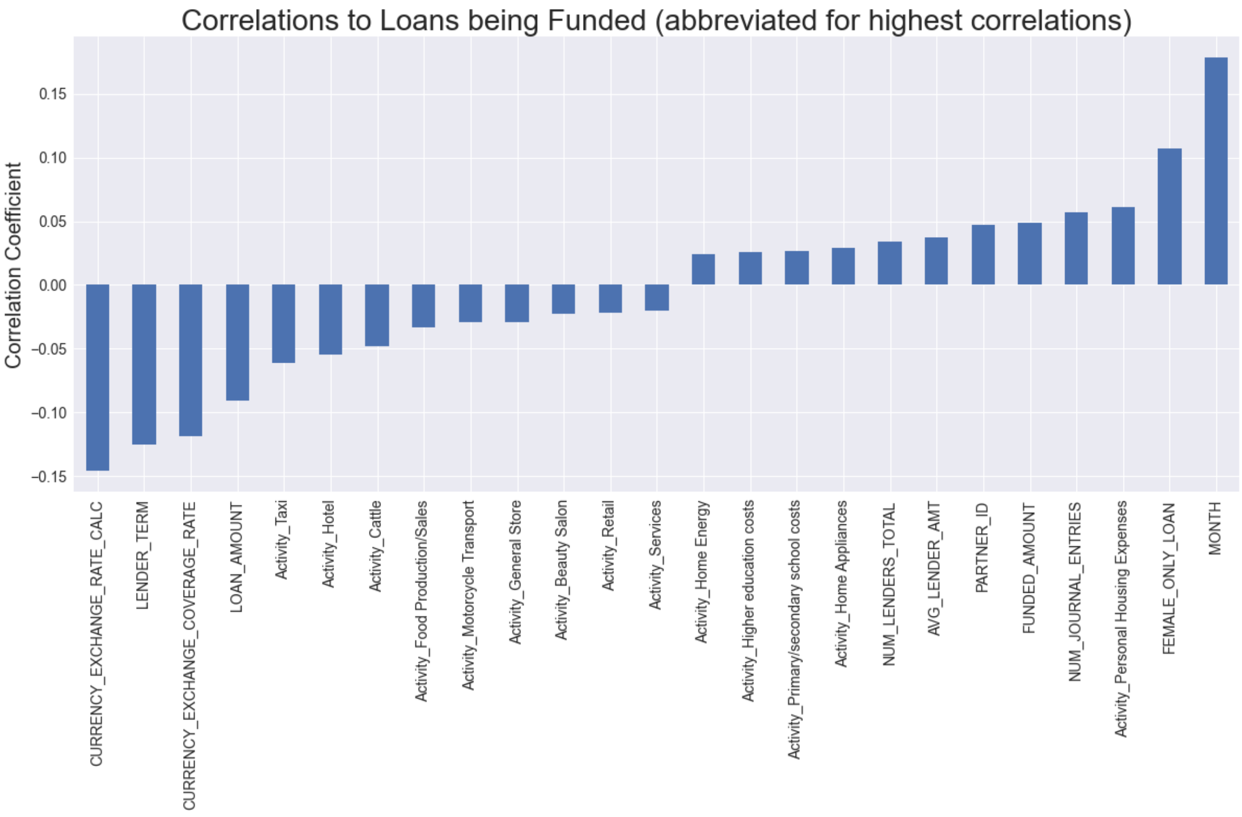
Categorical Data Encoding

Before the exploratory phase, the categorical features describing the loan purpose were expanded by one-hot encoding to identify potential relationships and the most meaningful aspects of the categorical variable.

- Loan Sector
- Loan Activity

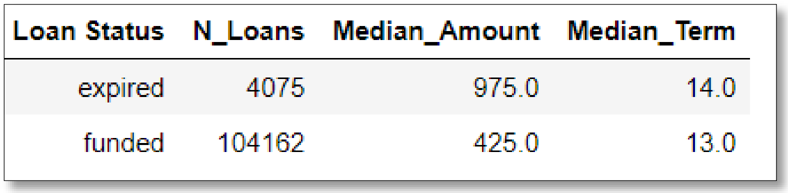
Exploratory Data Analysis

The purpose of exploratory data analysis is to gain a better understanding of the overall dataset and capture data patterns by visualizing the data. Due to the large number of features included in this dataset an initial visualization was completed to identify those features that most highly correlated with the outcome variable (loan status: funded or expired). The visualization shows the features that were most positively and negatively correlated. The exploration was completed with both datasets (2019, random) and any notable differences are included below.

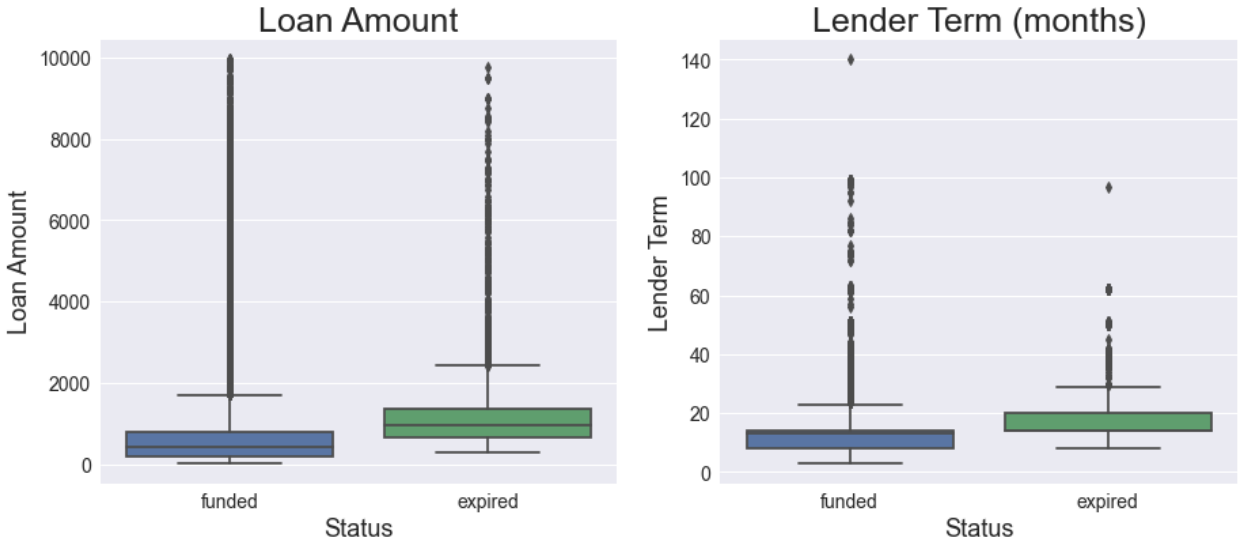


Loan Amount and Lender Term

Two features that were negatively correlated with a loan being funded were the loan amount and the lender term (length of the loan). Expired loans tended to be for a higher amount of money and had longer terms on average.

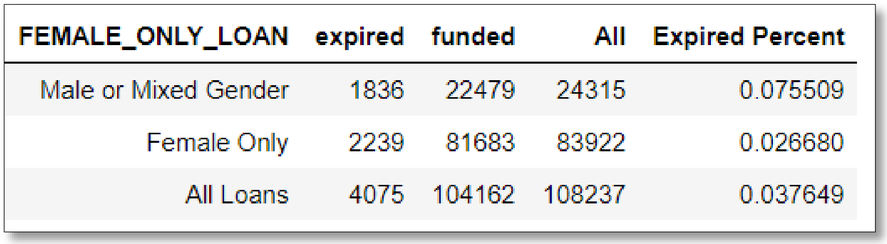


Loan Status	N_Loans	Median_Amount	Median_Term
expired	4075	975.0	14.0
funded	104162	425.0	13.0

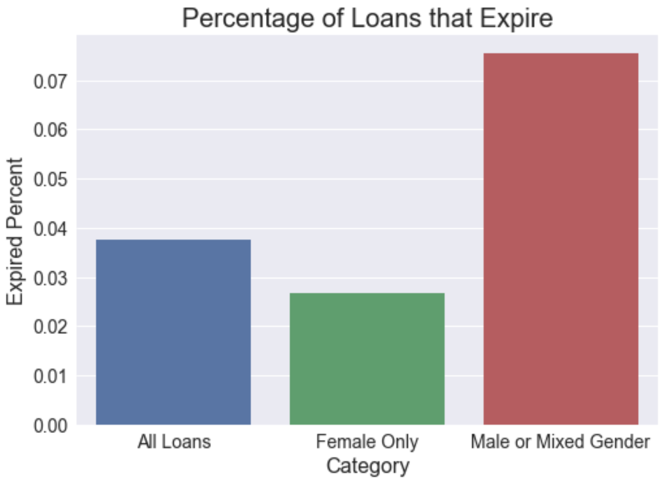


Female Only Loans

Part of Kiva's philosophy is to empower women and provide them with the financial means to improve their lives. On their website you can filter the loans to only female borrowers and a popular tag used is "woman owned biz". To further explore if a loan being a "female only loan" (only female borrowers, if more than one borrower), an additional feature was created that had one of the highest correlations with loans being funded or not. This correlation was strong across both data samples.

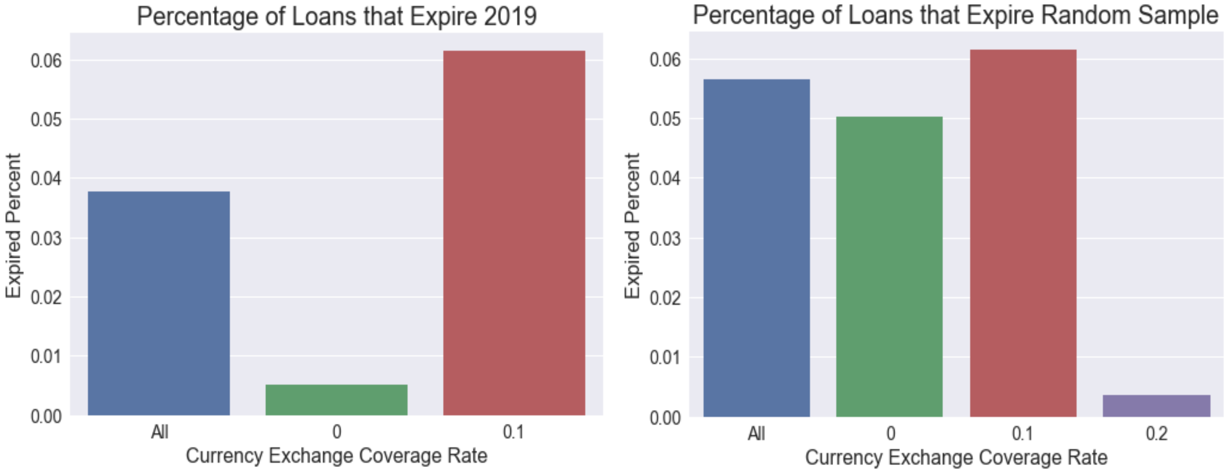


FEMALE_ONLY_LOAN	expired	funded	All	Expired Percent
Male or Mixed Gender	1836	22479	24315	0.075509
Female Only	2239	81683	83922	0.026680
All Loans	4075	104162	108237	0.037649



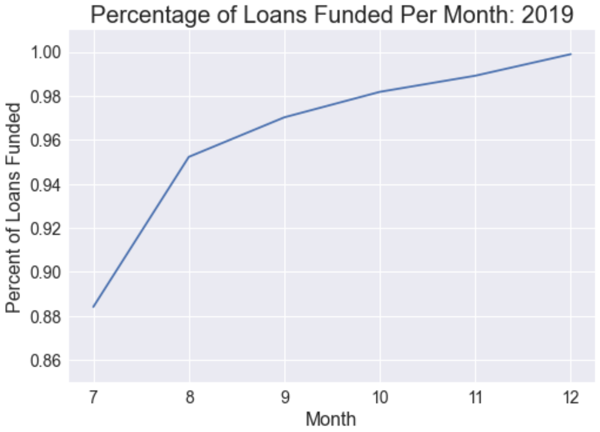
Currency Exchange Coverage Rate

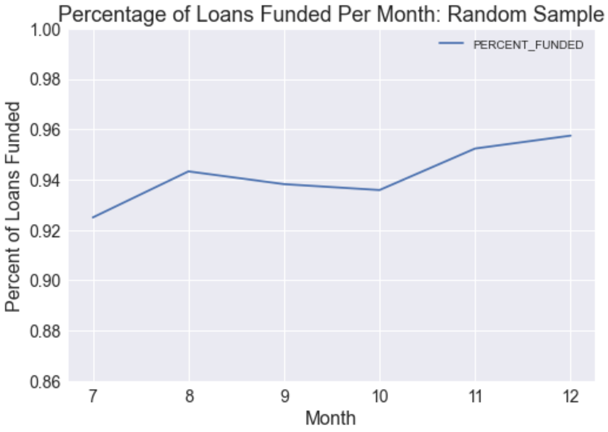
One of the negatively correlated variables was the Currency Exchange Coverage Rate for the 2019 dataset. Kiva has established two different options for their field partners to manage the fluctuations in exchange rates. The first policy is a shared model where the Field Partner takes on risk associated with currency fluctuations and is responsible for the first 10% or 20% of currency losses. The second model called the standard model is setup where all risk is shared, and any currency exchange losses are covered by all Kiva lenders rather than the field partners bearing the responsibility. This same trend did not hold out in the random sample dataset and this feature was shown to be positively correlated rather than negatively correlated. This policy has changed over time which might explain the difference between the two samples.



Time of Year

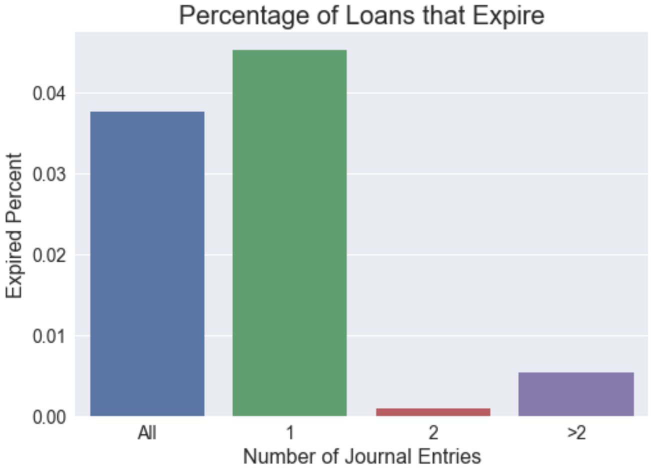
The month of the year in the 2019 dataset showed a high correlation with a loan being funded. However, this did not hold true when looking at the random sample or looking at the yearly trends for all years included in the dataset. It appears to be an anomaly and justifies why utilizing both datasets for model development is important.





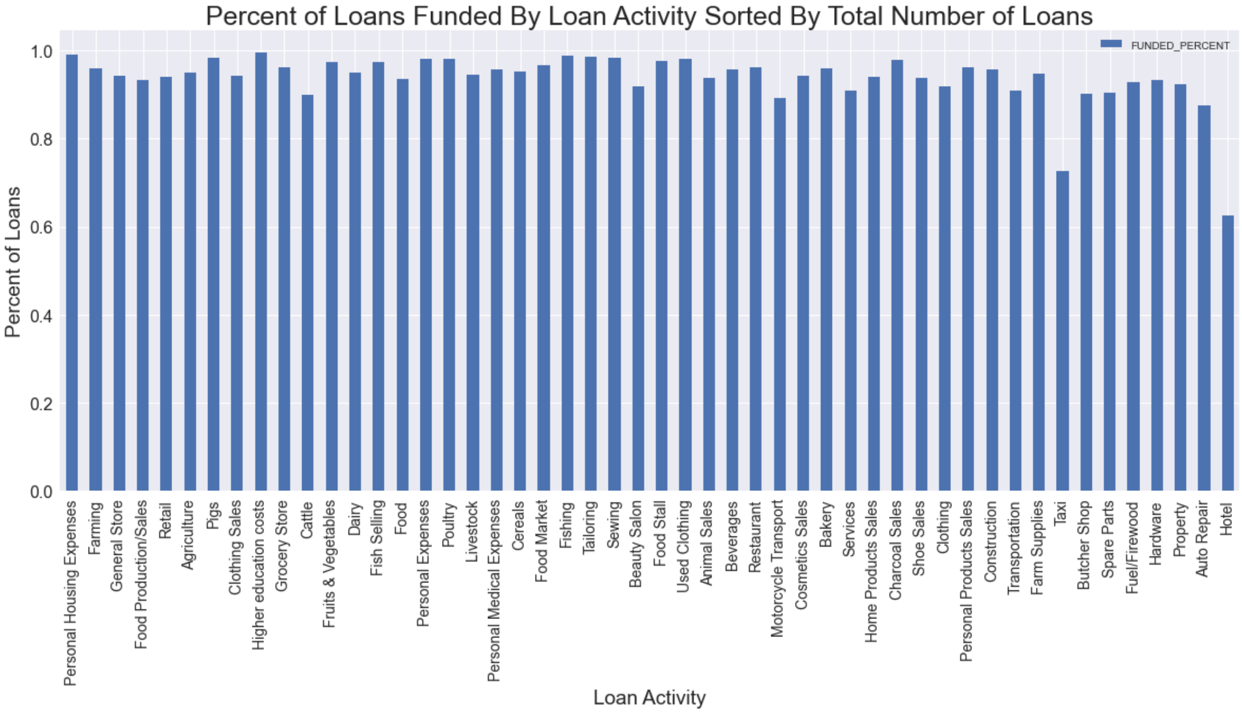
Journal Entries

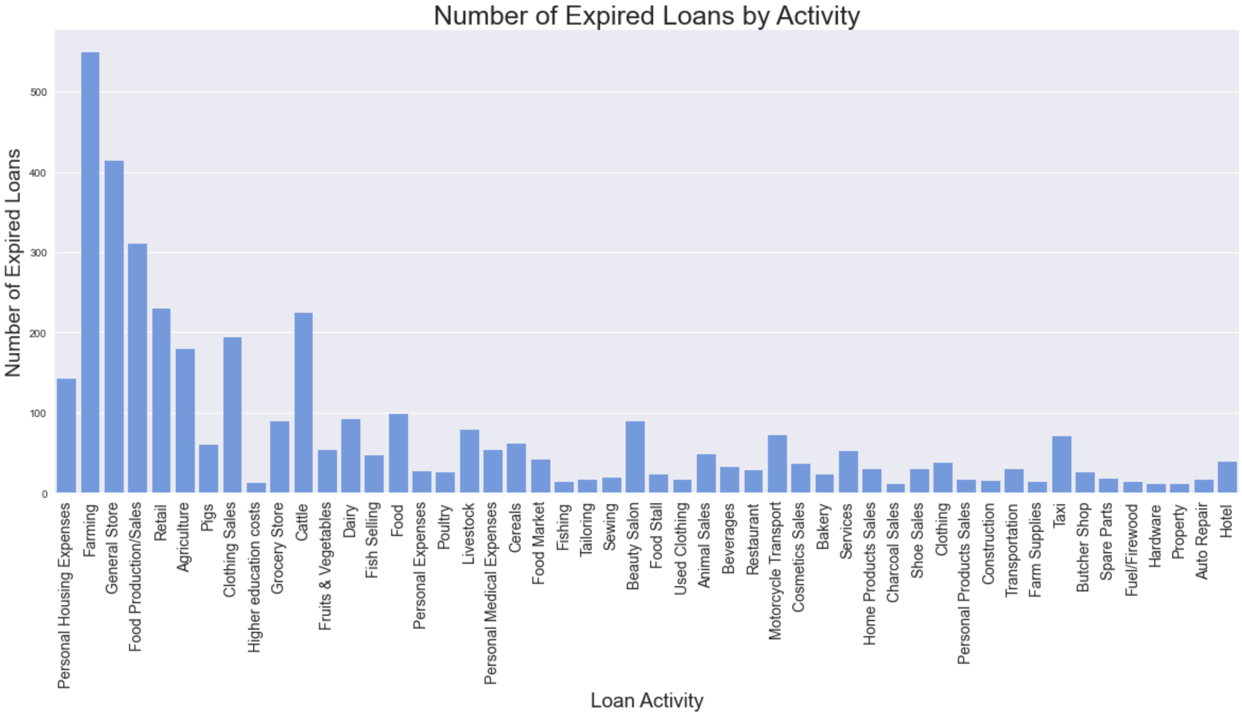
One feature on the Kiva website is the option for borrowers to provide updates via a "journal". This allows them to provide updates or provide a glimpse into their lives. Most loans only had a single (initial) journal entry that describes the loan. However, loans that had even just one additional journal entry rarely expired before being fully funded.



Activity Type

The remaining features that showed high correlation were all specific types of activities for how borrowers intended to use their loans. There are two visuals below. The first visual shows the percent funded for loans associated with a particular activity (sorted by the most frequent loan activity). The second graph shows the raw number of loans that were not funded by activity (sorted by the most frequent loan activity). The positive and negative correlation were more aligned with the percent funded rather than the number of loans that did not get funded.





Modeling

The goal of modeling in this project is to determine which loans will not be funded during the posting period. Only a small percentage of the overall loans are not funded and so the focus will be to ensure that the model can accurately predict the expired class rather than considering the overall accuracy of the model. When determining which models to test for this dataset the following characteristics were considered:

- Classification Prediction Model
- Class Imbalanced dataset – minority class is less than 4% of the records in the sample.
- Highly dimensioned dataset (522 features x 108237 records)

The following model types were selected as potential options that could address the specific needs of this problem.

- Random Forest
- Logistics Regression
- Support Vector Machine

Both the 2019 dataset and the random sample dataset were used to train and test models. The results shared here will be from the random sample dataset because these models proved to be more robust to new data.

Model Evaluation

Due to the imbalanced nature of the dataset, the evaluation of each model will be based on how well it is able to predict the minority class. Specifically, each model will be assessed on the F1 score for the minority class (expired loans). The F1 score is a balanced measure that considers both the recall (high true positive rate) and precision (low false positive rate) of the model.

$$\text{F1 Score} = \frac{2 \text{ x } (\text{Precision x Recall})}{\text{Precision + Recall}}$$

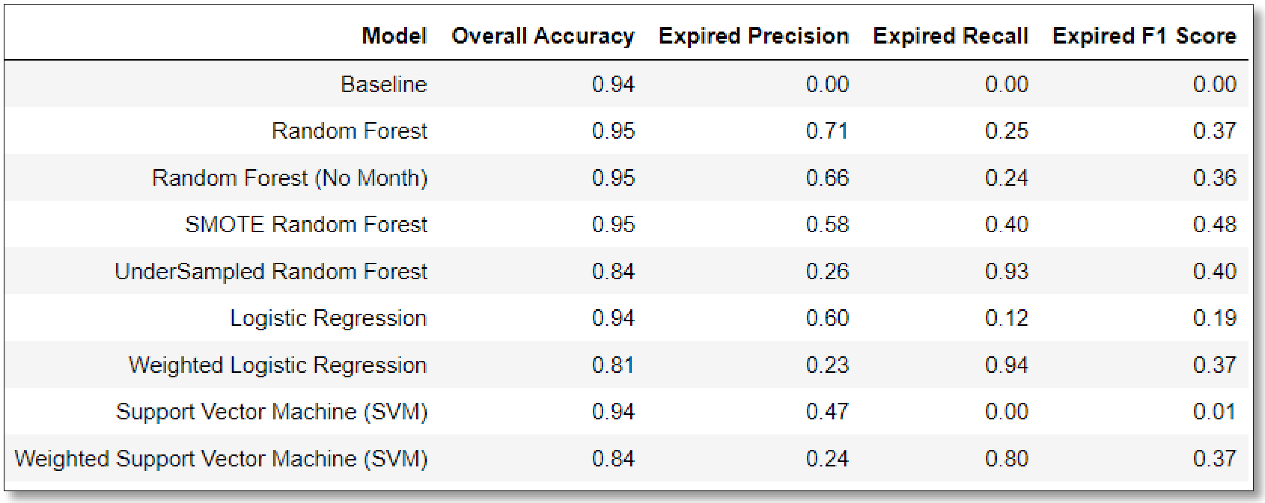
Model Comparison

All models were setup by splitting the dataset into a training set and a testing set (70/30 split). The test data was held back when preparing the model to ensure that there was no data leakage. The model evaluation was then assessed on its ability to handle the unseen test data.

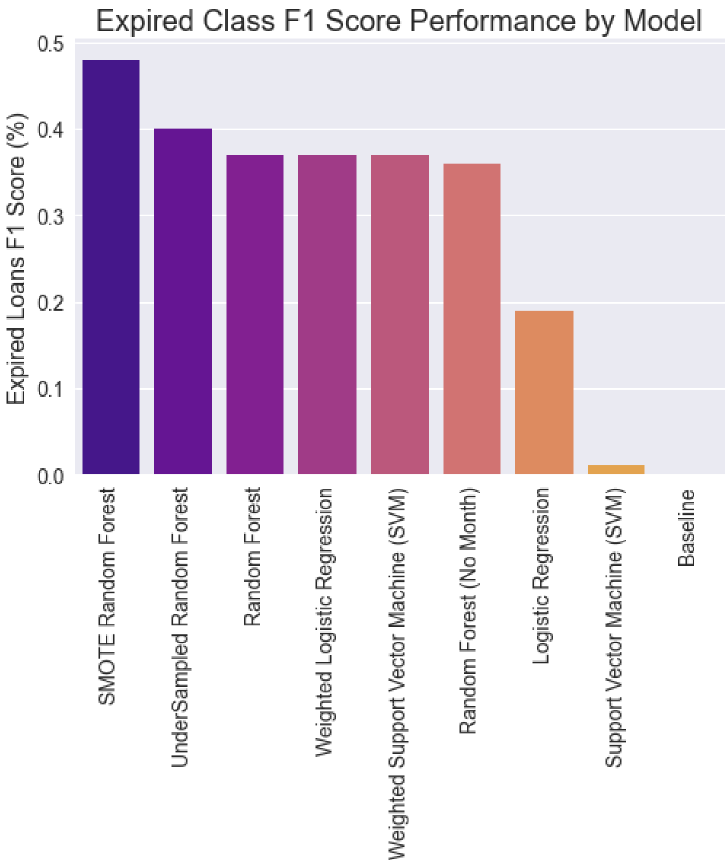
To provide a comparison an initial baseline model was created that predicted all loans for the majority class. The overall accuracy for this model was 94%, however the F1 score for the expired (minority) class was 0%. Eight additional models were tested to improve upon these results. The performance of each model is summarized below.

Additional techniques were tested to improve model performance. Two different sampling methodologies, SMOTE and Under Sampling, were applied to the training dataset for the Random Forest model to boost the model’s ability to learn the minority class. Weighting was also applied to the Logistic

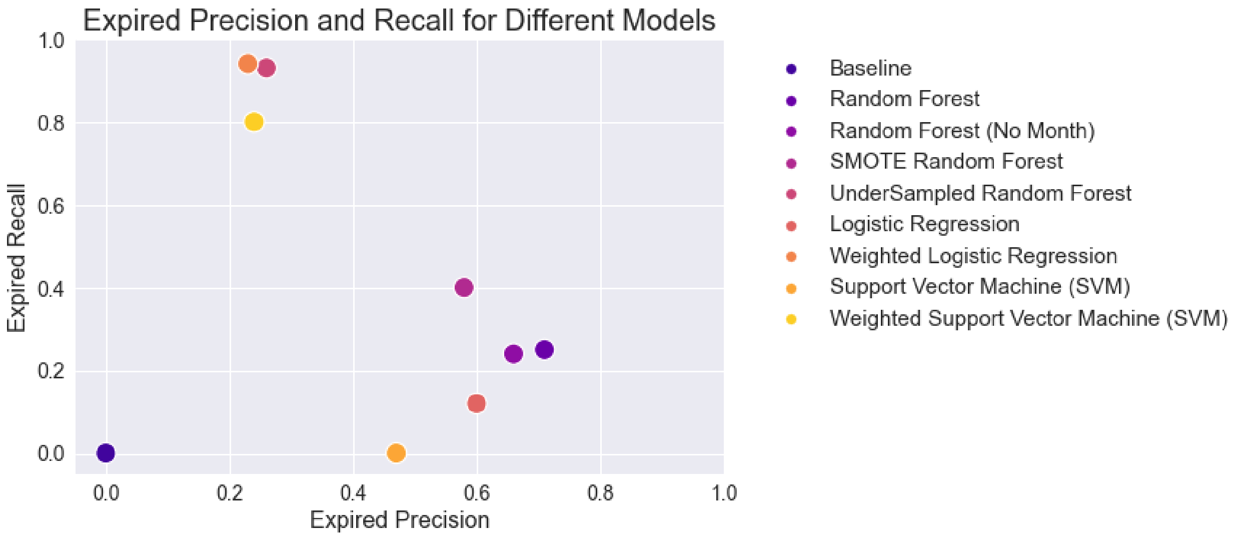
Regression and Support Vector Machine models to penalize the models for not predicting the minority class correctly.



Model	Overall Accuracy	Expired Precision	Expired Recall	Expired F1 Score
Baseline	0.94	0.00	0.00	0.00
Random Forest	0.95	0.71	0.25	0.37
Random Forest (No Month)	0.95	0.66	0.24	0.36
SMOTE Random Forest	0.95	0.58	0.40	0.48
UnderSampled Random Forest	0.84	0.26	0.93	0.40
Logistic Regression	0.94	0.60	0.12	0.19
Weighted Logistic Regression	0.81	0.23	0.94	0.37
Support Vector Machine (SVM)	0.94	0.47	0.00	0.01
Weighted Support Vector Machine (SVM)	0.84	0.24	0.80	0.37



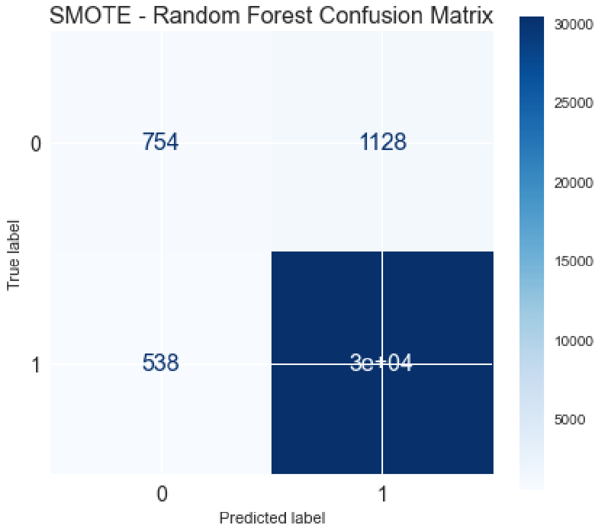
Expired Class F1 Score Performance by Model

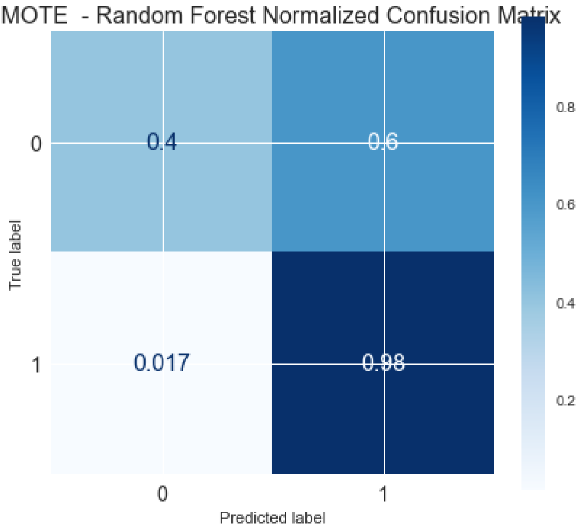


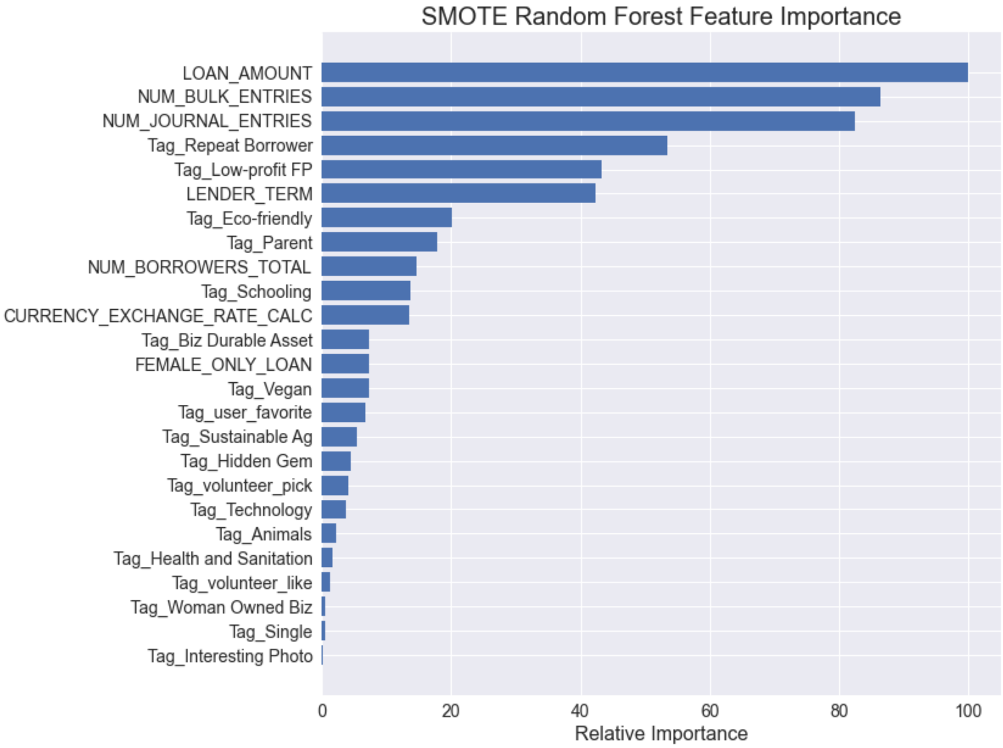
Model Performance

As with all models there was a tradeoff in performance. One set of models had high recall with lower precision. The other set tended to have better precision with lower recall. The most balanced model with the highest F1 score was the Random Forest model that used SMOTE.

SMOTE or Synthetic Minority Oversampling Technique can be used to balance the distribution of classes for an imbalanced dataset. It relies on increasing the size of minority class using existing data points to create "synthetic" neighbors drawing on information from the known feature space. This new balanced dataset is used to train the model which gives the model more data to learn about the minority class.







The features that the SMOTE Random Forest model selected to categorize the loans were technical data describing the loan (amount, terms, currency exchange coverage rate, etc.) and information that is being used to tag loans. These tags can describe how the loan will be used or highlight features about the borrower.

Recommendations

The intent of this project was to identify which loans would not be fully funded during the time that the loan was posted on the Kiva platform to allow for proactive interventions before posting the loan. The SMOTE model would allow for a little less than half of these loans to be identified while keeping the false positive rate relatively low. To understand the true opportunities with these loans Kiva should test the following intervention strategies:

1. Proactively run each loan through the SMOTE Random Forest model to identify the loans that are predicted to not be funded. These borrowers can then be counseled to modify the technical aspects of the loan (amount, terms, etc.) to align better with the profile of the loans that are funded.
2. Secondly, Kiva should adapt their website marketing techniques to feature the loans predicted by the model to not be fully funded.
3. Finally, Kiva should consider coming up with additional tags to allow for potential lenders to filter and more easily identify the loans that are close to expiring or that the model identified as at risk for not fully funding.

Future Work

Depending on the success of the interventions potential future work:

1. Explore the model's misclassifications in more detail to provide insight into how to strengthen the dataset or modify the modeling approach.
2. Compare how close the loan was to being funded (i.e. 90% funded before it expired) to the prediction probability score assigned by the model of it being funded or not, to understand if the model is sensitive enough to identify "almost funded" loans from "barley funded" loans.
3. Segment the data on the size of the loan and model each segment separately to determine if there are different predictors driving each segment.
4. Test traditional finance models to potentially improve predictive performance.