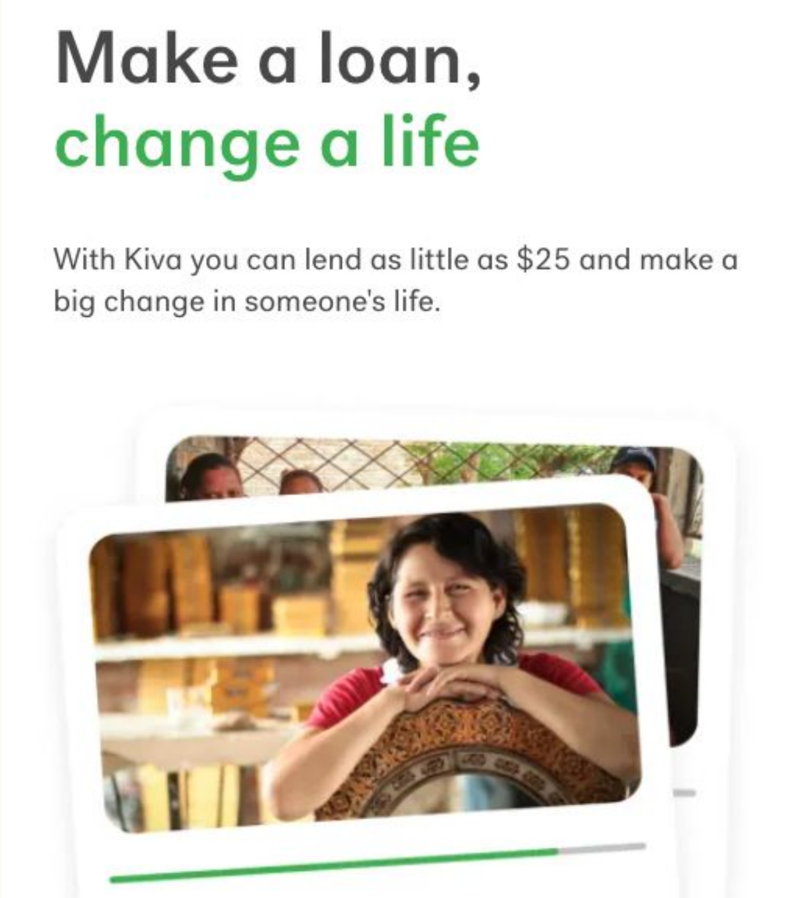
KivaMaxApprover App

Supporting Small Business Micro-Loan Field Partners in Kenya

Rachel Insler, Shashank Rajadhyaksha,
Carlos Rivera, Precious Smith

DSIR22221-E Group Project, Presented 4.26.21

Table of Contents



Background

"More than 1.7 billion people around the world are unbanked and can't access the financial services they need. Kiva is an international nonprofit, founded in 2005 in San Francisco, with a mission to expand financial access to help underserved communities thrive."

They do this by crowdfunding loans and unlocking capital for the underserved, improving the quality and cost of financial services, and addressing the underlying barriers to financial access around the world.

Source: Kiva.org

Background

Source: Kiva.org

Kiva Field Partners

are MFIs (microfinance institutions), schools, NGOs, & social enterprises that:

- screen borrowers
- post loan requests
- disburse loans on the ground
- collect repayments
- may offer additional services, like entrepreneurial training and literacy skills

Source: Kiva.org
Image Credit: juhudikilimo.com

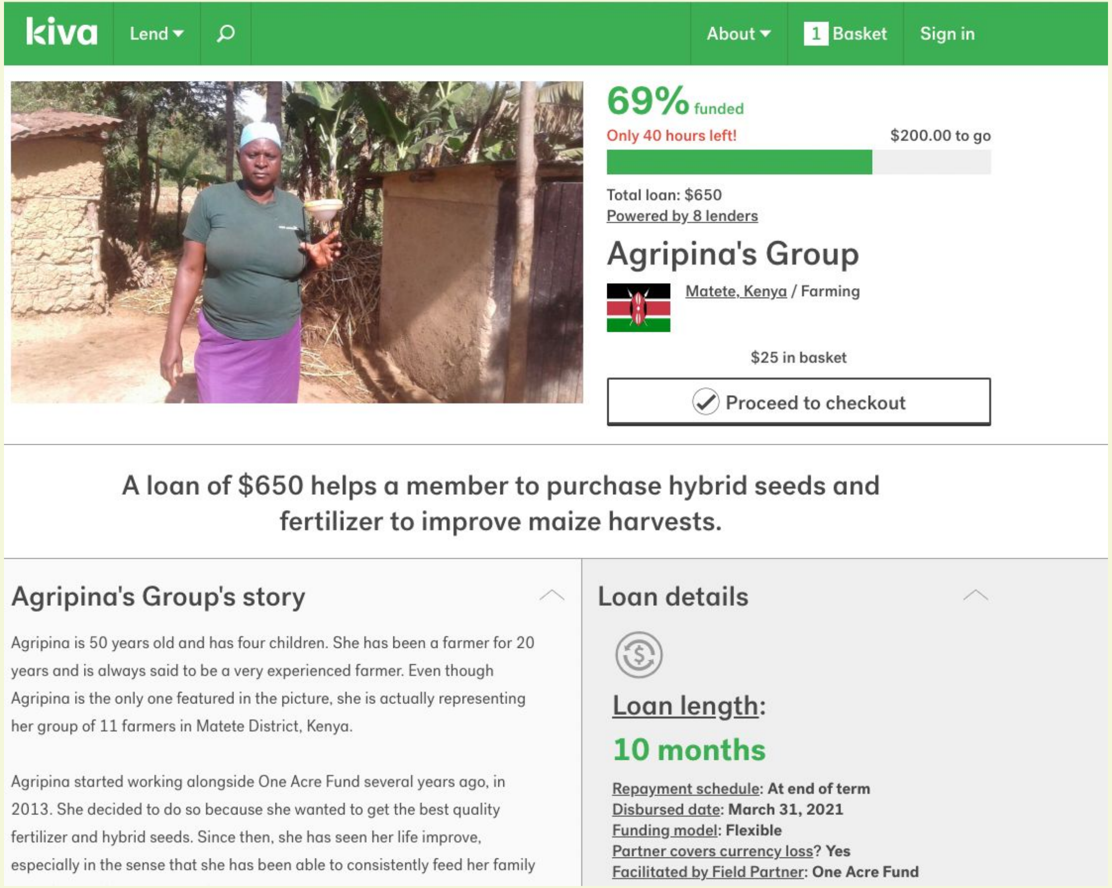
We aim to help Kiva Field Partners

with the **borrower screening process** by developing an online tool for prospective applications that will estimate the likelihood of a loan getting crowdfunded.

We will use machine learning models to **identify the key factors that predict loan funding success** and **develop a simple questionnaire** that allows field partners to quickly determine whether a loan application is likely to be funded or not.

Image Credit: Kiva.org

What's in a Loan Request?



- Borrower photo & video
- Brief loan use summary
- Borrower's personal story
- Loan request amount
- Country
- Repayment schedule
- Date of request
- Sector (e.g. Agriculture, Retail)
- Activity (e.g. Beauty Salon)
- And more...

Source: Kiva.org

02. Data

- Obtained 1,053,185 detailed records on loans from 2015 - 2019 at kiva.org

- Downsampled 419, 156 loan records, oversampling expired loans
  - 350,000 funded
  - 69,156 expired

Image Credit: Kiva.org

Kenya Focus

Why Focus?

- Eliminate effect of country
- Parse dataset to more manageable size (51,170 records) for NLP analyses

Why Kenya?

- Mean loan success rate (78.4%) slightly below overall downsampled population (83.5%) success rate
- Second-highest loan volume
- Kiva has an office in Nairobi for pilot
- 35 field partners in Kenya

Sources: Kiva.org, Kiva Blog,
Image Credit: Kiva.org

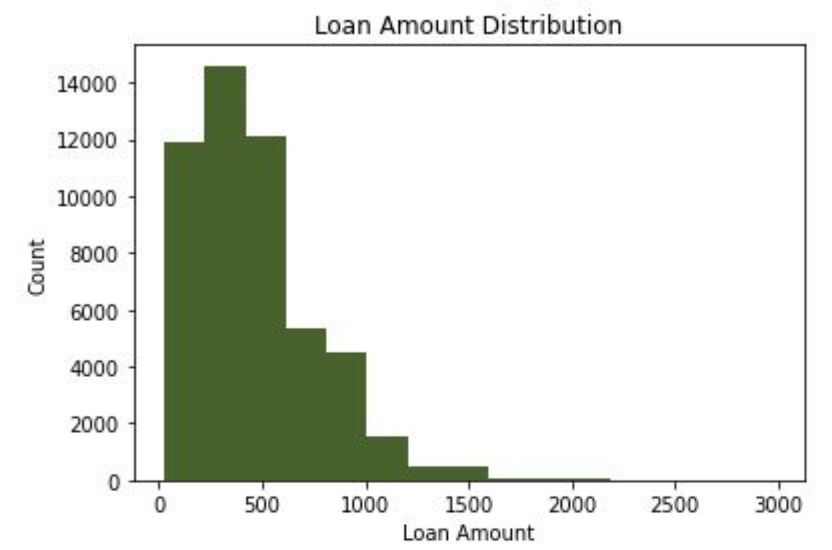
Data Cleaning & Preprocessing

	Numerical	Natural Language Processing (NLP)
Cleaning	• Filled null values • Transformed features • Dummified categorical values	• Removed rows with no text • Removed duplicates • Combined text columns
Preprocessing	• Converted text content to word length and char count • Created ‘misc’ activity names to reduce dimensionality • Created interaction terms for highly correlated features • Created ‘month’ variable	• Removed special characters, & stopwords • Lemmatized, tokenized (maintained tags), vectorized

- Final Numerical Dataset contained 28 features (pre-dummification)

03. Exploratory Data Analysis

EDA: Loan Amount (in USD)



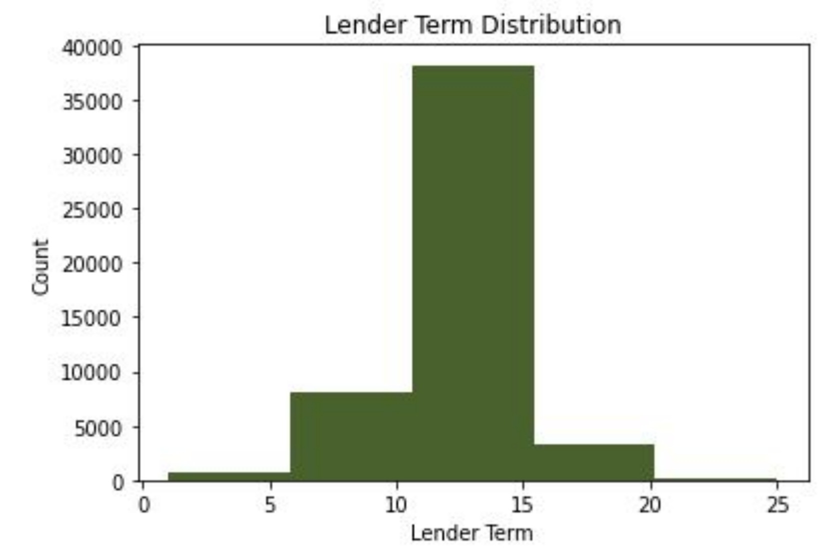
	LOAN_AMOUNT
STATUS	
0	727.790634
1	421.161366

0 = Expired (not-funded)
1 = Successful (funded)

Insight:

The mean loan amount for successful (funded) loans was $326 less than expired (not-funded) loans.

EDA: Lender Term



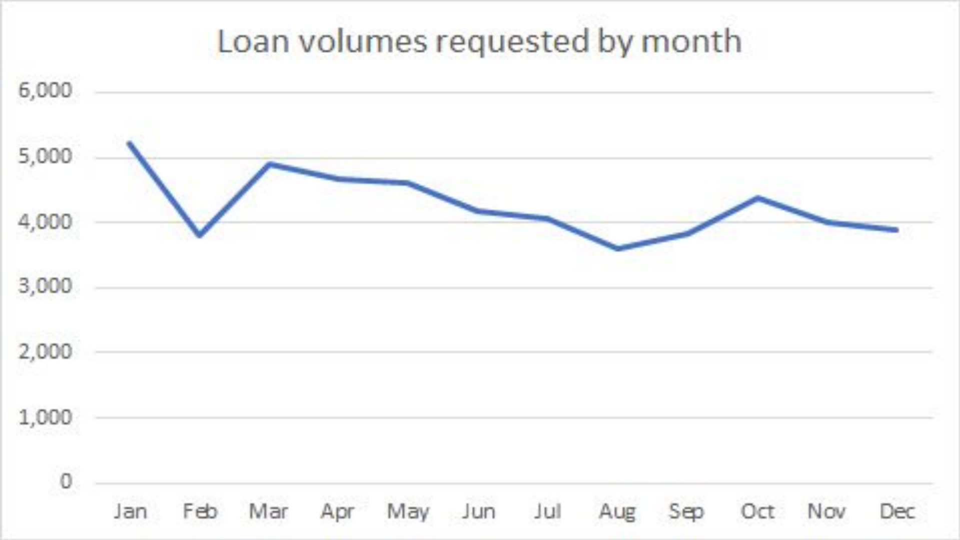
STATUS	LENDER_TERM
0	14.395499
1	12.815571

0 = Expired (not-funded)
1 = Successful (funded)

Insight:

Expired loan requests were on average two months longer than successfully-funded loans.

EDA: Month



Insight:

- Average Funded rate for the year: 78%
- Highest funding rate: **December** @89% (Holiday season!)
- Highest loan volume request: **January** (corresponds to the start of Kenya's farming season)

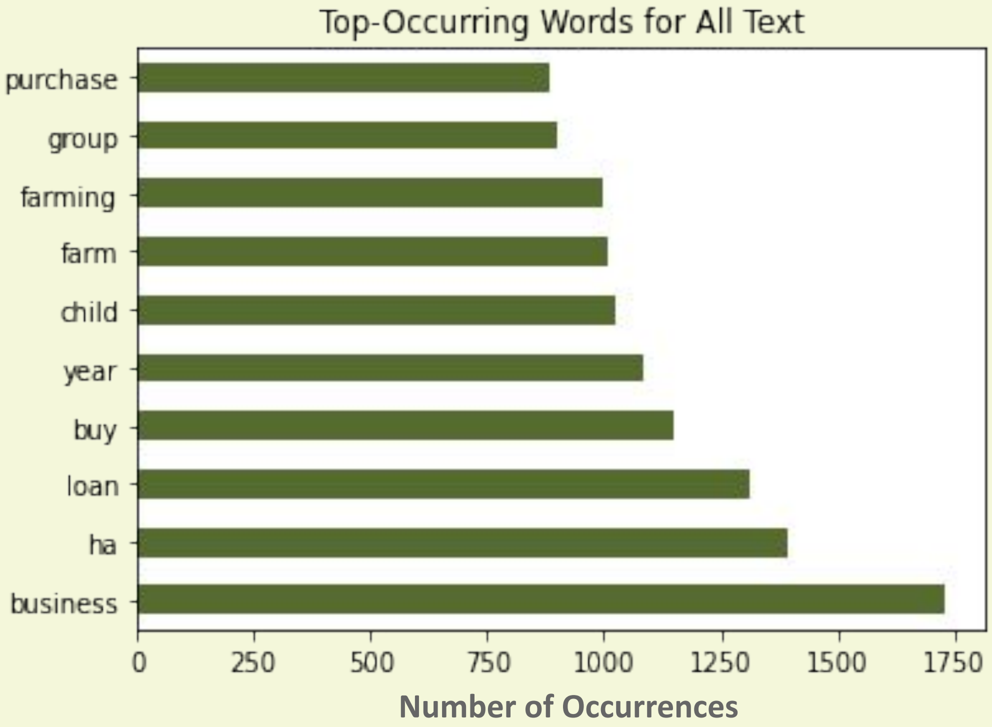
Sample Tags, Loan Use, & Description Text

TAGS: #Single Parent, #Woman-Owned Business

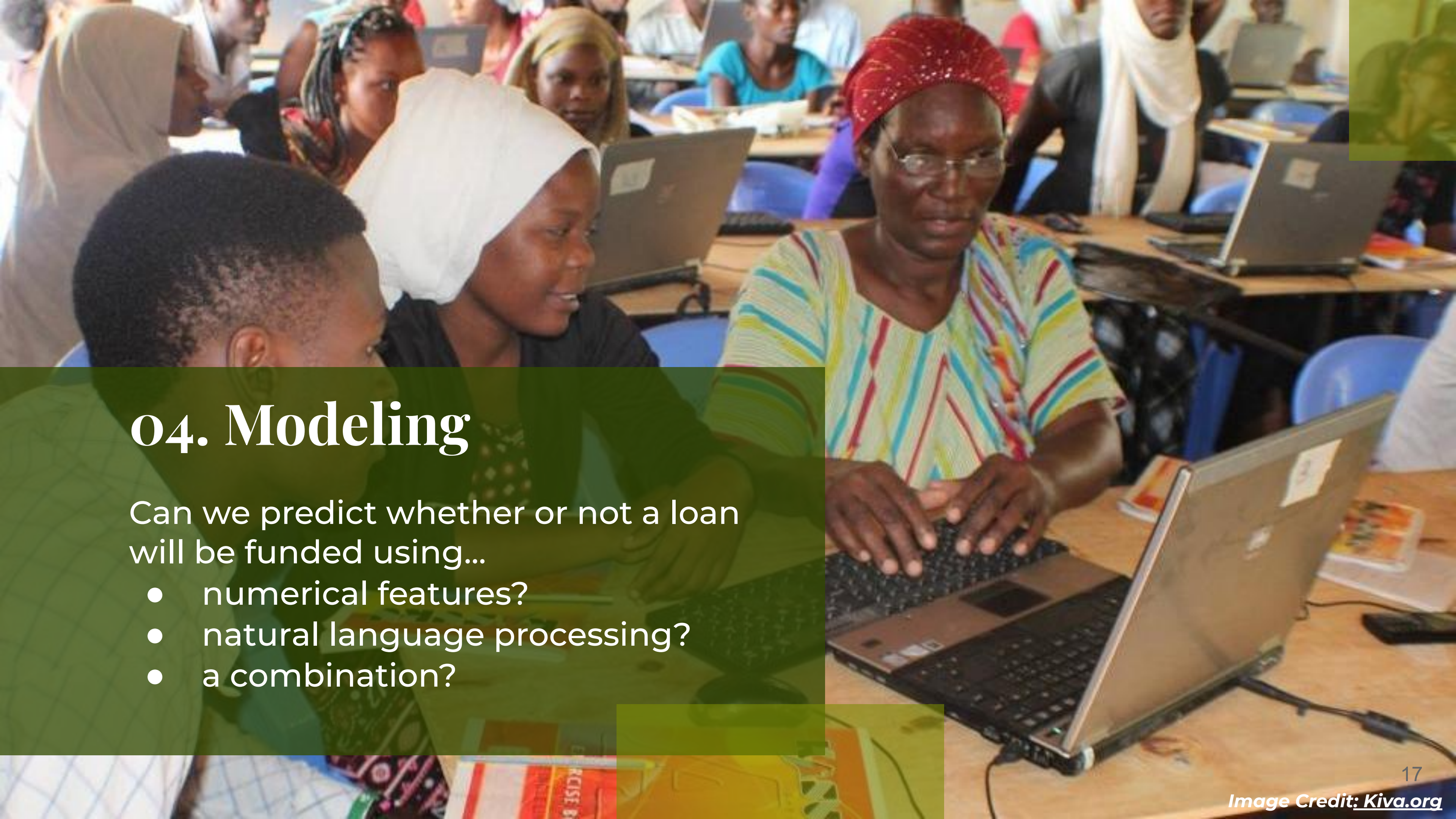
LOAN USE: to purchase new items for her retail shop to sell to her customers

DESCRIPTION: Carren is a single mother of one adorable child whom she adores. She is a hardworking and devoted woman when it comes to her business. Her positive attitude has earned her customers in her retail business. Her dream is to grow her business from a retailer to a wholesaler. She also aspires to educate her child to university level for a brighter future ahead. Carren is pleading for a fourth Kiva loan of KES 120,000 to make her dream come true. She promises to repay the loan with the earned profit from her business.

EDA: NLP



Common Words



04. Modeling

Can we predict whether or not a loan will be funded using…

- numerical features?
- natural language processing?
- a combination?

Image Credit: Kiva.org

Numerical Model

Development & Selection

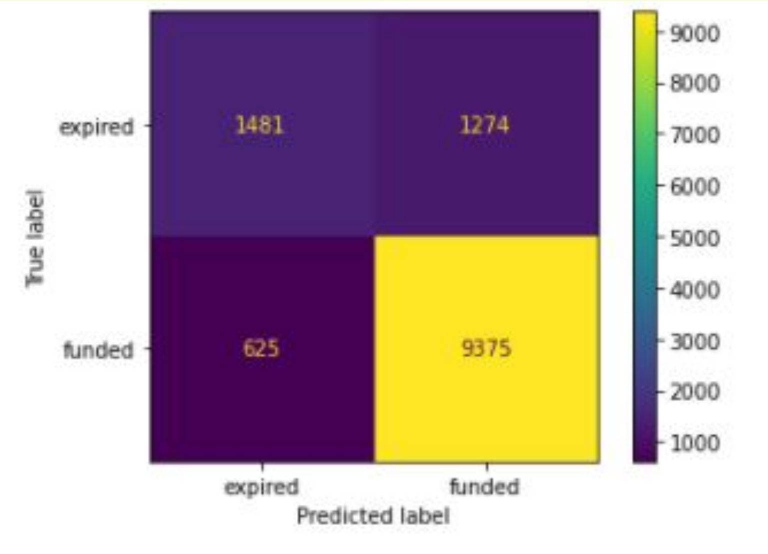
- Train-test assessment carried from NLP
- Scale features
- Modeled with multiple classifiers
- Extensively explored and hypertuned:
  - Logistic Regression
    - Best scores:
      - 0.834 test acc.
      - 0.834 train acc.
  - Random Forest
    - Best scores:
      - 0.848 test acc.
      - 0.98 train acc.
  - **Gradient Boost**

Best-Performing Model: Gradient Boost

- Parameters:
  - learning_rate: 0.1
  - max_depth = 5
  - n_estimators = 100
- Scores (*Baseline Model: 0.785)*
  - Train score: 0.866
  - Test score: 0.848
  - CV score: 0.852
- Best increase over baseline model without substantial overfitting

Numerical Model

Confusion Matrix



Classification Report

	precision	recall	f1-score	support
0	0.70	0.54	0.61	2755
1	0.88	0.94	0.91	10000
accuracy			0.85	12755
macro avg	0.79	0.74	0.76	12755
weighted avg	0.84	0.85	0.84	12755

- Overall accuracy at 0.85 better than baseline of 0.78
- Performs well when predicting funded loans (0.91 f1-score) but less well on expired loans (0.61).
  - Recall of expired loans an issue (i.e. higher False Negatives of 0s)
  - More expired loans to train could help

Numerical Model

- Gradient Boost & Logistic Regression suggest similar key variables
- #Tags are important
- Loan_Amount is one of the biggest predictors, negatively correlated with likelihood of funding
- Lender_Term: negatively correlated
- Having higher count of female applicants increases chances of funding
- Some activities like Home Energy/Education are more likely to be funded than Retail
- December has higher funded rate

Most Important Features

Gradient Boost

	importance	feature_names
0	0.475977	LOAN_AMOUNT
8	0.070460	FEM_COUNT
17	0.067644	MALE_FEM
1	0.056266	LENDER_TERM
6	0.042619	char_count_TAGS

Log Reg

	coefficients	features
15	2.896401	word_char_TAGS
20	2.159625	ORIGINAL_LANGUAGE_MISSING
3	-1.900395	word_count_TAGS
0	-1.858365	LOAN_AMOUNT
8	1.240612	FEM_COUNT
2	-0.793985	word_count_DT

NLP Model

Development & Selection

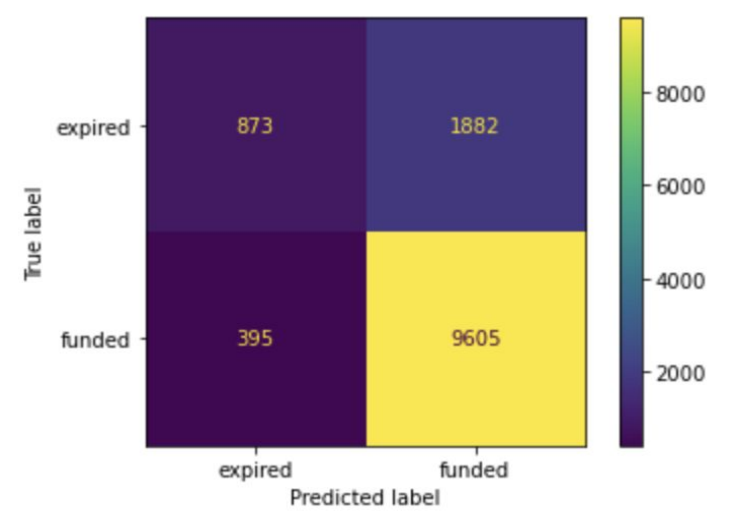
- Train-test-split
- Two separate modeling workstreams with text columns
    - as three individual features (Loan Use, Description, and Tags)
    - combined
- Modeled tags two ways:
    - joined (‘singleparent’)
    - individual words (‘single’, ‘parent’)
- Attempted CVEC & TVEC
- Explored all interpretable supervised learning models

Best-Performing Model: Logistic Regression with TfidfVectorizer, Combined Text Columns, Complete Tags

- Parameters:
    - ngram_range: (1,2)
    - max_features : 10,000
    - C = 1.0
- Scores (*Null Model: 0.785)*
    - Train score: 0.837
    - Test score: 0.821
    - CV score: 0.816

NLP Model

Confusion Matrix



Classification Report

	precision	recall	f1-score	support
0	0.69	0.32	0.43	2755
1	0.84	0.96	0.89	10000
accuracy			0.82	12755
macro avg	0.76	0.64	0.66	12755
weighted avg	0.80	0.82	0.79	12755

- With 0.82 accuracy, performed slightly less well than numerical model overall
- Performs well when predicting funded loans (0.89 f1-score) but poorly on expired loans (0.43).
  - More expired loans to train on might help.

NLP Model

Words & Phrases Most Correlated With Funded Loans

	coefs	word_combinations
	5.614132	20 000
	3.470010	30 000
	3.381368	kes 20
	3.338394	singleparent
	3.326476	20
	3.211537	user_favorite
	3.091468	single mother
	2.488434	widowed
	2.448559	15 000
	2.440288	kes 000
	2.439046	widow
	2.396601	single mom
	2.304716	10 000
	2.266018	30
	2.167488	grew

Words & Phrases Most Correlated With Expired Loans

	coefs	word_combinations
	-4.320953	100 000
	-4.180876	100
	-3.664855	man
	-3.484908	repeatborrower
	-3.371616	repairrenewreplace
	-3.265535	bizdurableasset
	-3.145058	parent
	-2.864306	150 000
	-2.810579	150
	-2.748698	120 000
	-2.706491	80 000
	-2.569605	nhe
	-2.551535	80
	-2.547320	120
	-2.531583	shop

Combination Model

Q: Can we combine the predictive power in the two models to further improve the performance?

Process & Summary

- Combination model entailed bringing together the Numeric model and the NLP model
- Train/test were made identical in the 2 models and each application scored on both models
- Resulting probabilities from each model were then weight-averaged to translate to an overall score and predictions based on that.
- Eg: Overall score: w1*(p_num)+w2*p(nlp)
- Combined model slightly underperformed the Numeric model and outperformed the NLP model
- Next step is to figure out better way of calculating integrated predictions

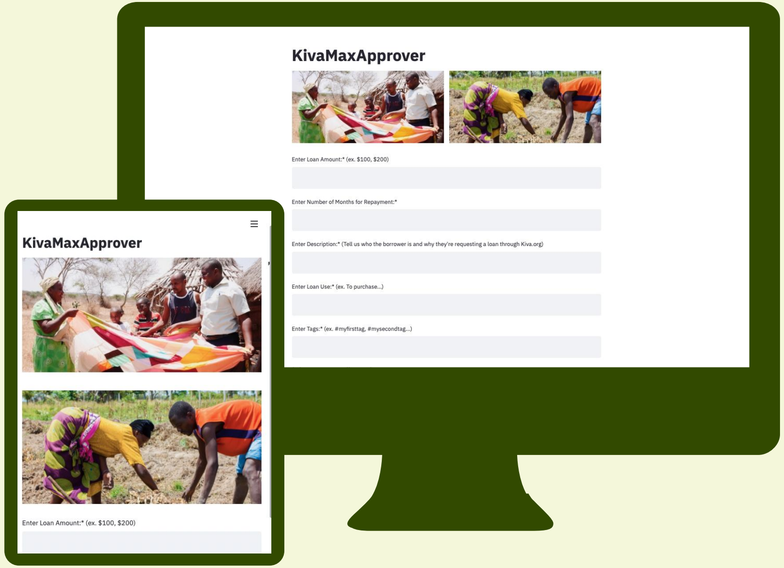
Numerical Results:

	precision	recall	f1-score	support
0	0.70	0.54	0.61	2755
1	0.88	0.94	0.91	10000
accuracy			0.85	12755
macro avg	0.79	0.74	0.76	12755
weighted avg	0.84	0.85	0.84	12755

Combined Results (70% Num/30% NLP):

	precision	recall	f1-score	support
0	0.75	0.48	0.58	2755
1	0.87	0.96	0.91	10000
accuracy			0.85	12755
macro avg	0.81	0.72	0.75	12755
weighted avg	0.84	0.85	0.84	12755

05. KivaMaxApprover App



KivaMaxApprover App

- Built on a simplified prediction algorithm that minimizes false positives
- Requires only 5 user entries
- Requires minimal training
- Returns predictions rapidly
- Predicts on both models: Numeric & NLP
- DEMO TIME!

06. Conclusions & Future Directions

Conclusions

Yes!

We can predict with **85% accuracy** whether a loan will be crowdfunded or not using our numerical model. Our app should help the Field Partners with screening efficiency.

What Matters Most?

- Amount of loan,
- Loan term
- Gender/number of applicants
- Amount of text
- Month of request

So Words Don't Matter?

Not so. They just matter in their own model. The NLP model was able to predict with 82% accuracy on its own.

NLP Has Its Place

In fact, there are several hundred cases (228 expired and 280 successful) in which NLP accurately predicted loan status, but Numerical fails.

Future Directions

App Refinement

Enhance functionality so that the app can automatically generate suggestions for application improvement.

Model Combination

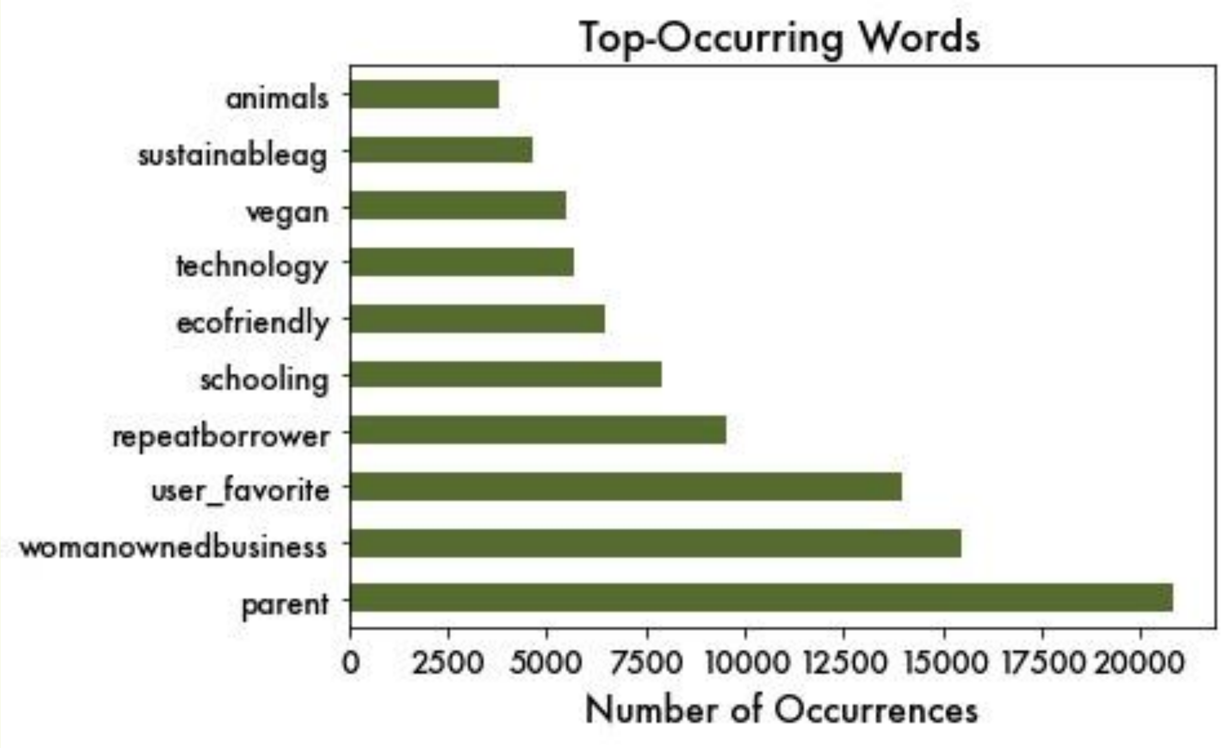
Explore additional modeling combination techniques to harness the additive predictive power from both models.

Thanks!

Do you have any questions?

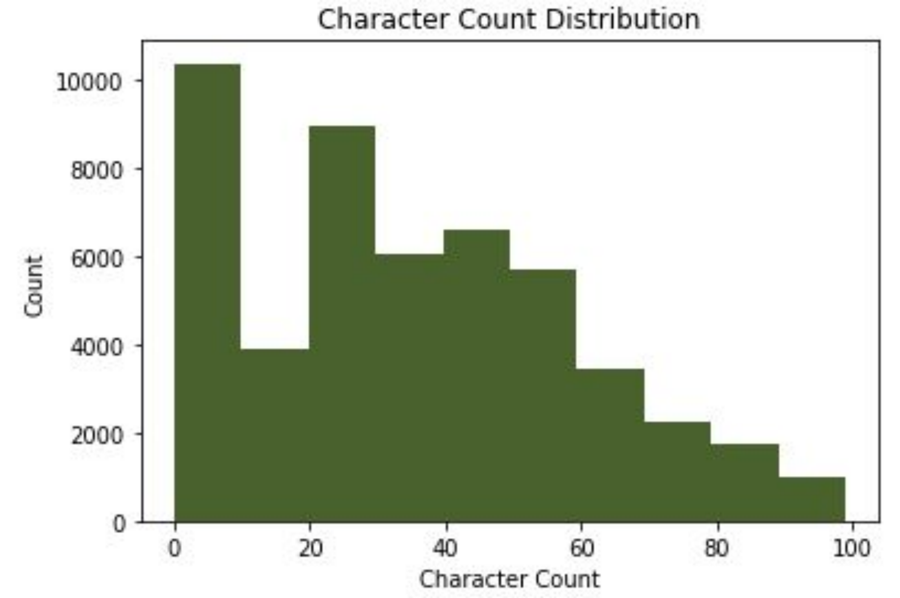
CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik.**

EDA: NLP



Common Tags

EDA: Tags



STATUS	char_count_TAGS
0	42.618386
1	35.236258

0 = Expired
1 = Successful

Insight:

Applications that were funded had an average TAG of 35 characters.