ESTIMATING KIVA LOAN AMOUNTS

SHAFIQ JADALLAH

CAPSTONE 2: SUPERVISED LEARNING MODELS

OCTOBER, 2019

QUESTIONS ABOUT KIVA'S LOANS

- Kiva started in 2005
- An crowdsourced microlender established as a 501(c)(3) that allows people to lend money to students and entreprenuers in low-income
- Have lent over 1.6M loans totaling over $1.3Bn

- ***Question 1*:**
  - What would be a typical Kiva loan amount?
- ***Question 2:***
  - What are the strongest features so that a borrower should know that impacts his/her expected loan amount?

THE DATA SETS

- Pulled data from Kaggle.com
  - https://www.kaggle.com/mhajabri/kiv-me-a-loan/data
- Found additional dataset to complement country profile information
  - https://www.kaggle.com/codename007/a-very-extensive-kiva-exploratory-analysis/data

KIVA Dataset
Features: 20
Rows: 671,205

id 671205
funded_amount 610
loan_amount 479
activity 163
sector 15
use 424912
country_code 86
country 87 region 12695 currency 67
partner_id 366
posted_time 667399
disbursed_time 5719
funded_time 498007
term_in_months 148
lender_count 503
tags 86719
borrower_genders 11298
repayment_interval 4
date 1298

Country Stats Dataset
Features: 13
Rows: 174

country_name 174
country_code 173 country_code3 173
continent 174
region 174
population 174
population_below_poverty_line 152
hdi 171
life_expectancy 168
expected_years_of_schooling 168
mean_years_of_schooling 168 gni 168
kiva_country_name 174

EDA: EXPLORE, CLEAN, EXPLORE SOME MORE…

- Eliminated <NaN> values through combined technique of filling in through means and dropping rows once <5% of data fields were complete
- Treated outliers using winsorize function for one-tailed distribution
- Removed any blank spaces
- Created categories to convert categorical features from 'str' to 'int' in order to perform analysis and plotting functions

Kiva data set: pre adjustments

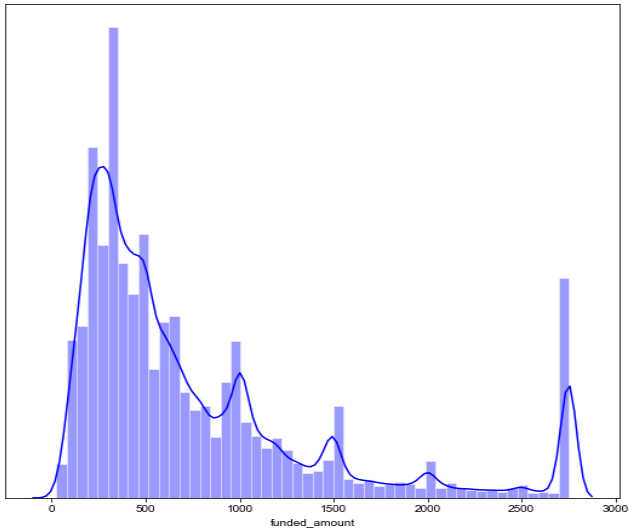
	funded_amount	loan_amount	term_in_months	lender_count
count	423081.00	423081.00	423081.00	423081.00
mean	849.45	849.45	14.18	23.47
std	1134.65	1134.65	8.63	30.70
min	25.00	25.00	2.00	1.00
25%	300.00	300.00	8.00	9.00
50%	500.00	500.00	14.00	15.00
75%	1000.00	1000.00	14.00	29.00
max	100000.00	100000.00	158.00	2986.00

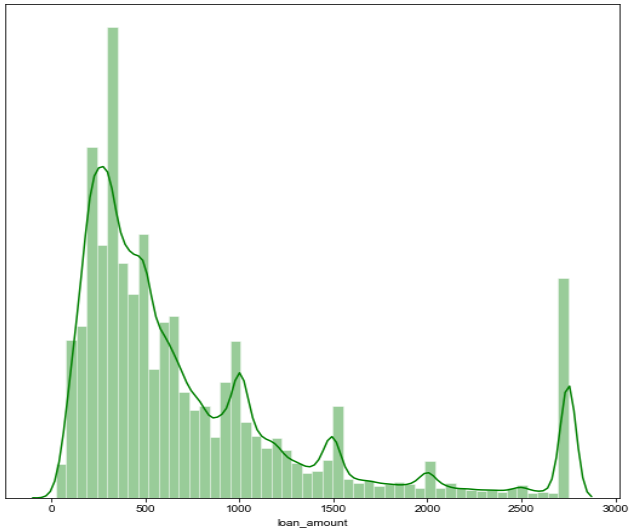
Final data set: post adjustments. Final shape (427764, 16)

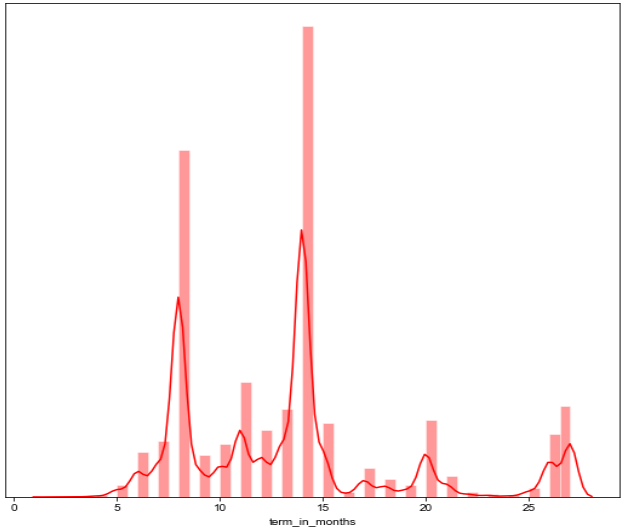
	count	mean	std	min	25%	50%	75%	max
funded_amount	426764.0	773.48	678.80	25.00	300.00	525.00	1.000000e+03	2.750000e+03
term_in_months	426764.0	13.35	5.38	2.00	8.00	14.00	1.400000e+01	2.700000e+01
lender_count	426764.0	21.57	18.01	1.00	9.00	15.00	2.900000e+01	7.000000e+01
population	426764.0	54250459.02	46310359.43	611343.00	11051600.00	42862958.00	1.049181e+08	1.439898e+08
population_below_poverty_line	426764.0	29.61	12.04	6.60	21.60	25.60	3.490000e+01	6.300000e+01
hdi	426764.0	0.63	0.09	0.42	0.55	0.65	6.800000e-01	9.200000e-01
life_expectancy	426764.0	67.87	5.67	55.48	62.16	68.34	7.206000e+01	8.254000e+01
expected_years_of_schooling	426764.0	11.56	1.57	8.11	10.90	11.73	1.287000e+01	1.735000e+01
mean_years_of_schooling	426764.0	7.35	2.01	2.97	5.78	6.94	9.330000e+00	1.270000e+01
gni	426764.0	6759.38	5105.73	1262.17	2880.74	6154.89	8.395090e+03	4.360882e+04
loan_count	426764.0	31415.25	31572.07	1.00	5596.00	19597.00	4.500800e+04	8.696000e+04
gender_group	426764.0	0.53	0.73	0.00	0.00	0.00	1.000000e+00	2.000000e+00
country_class	426764.0	2.27	0.64	1.00	2.00	2.00	3.000000e+00	3.000000e+00
sector_category	426764.0	4.74	3.24	1.00	2.00	4.00	6.000000e+00	1.300000e+01

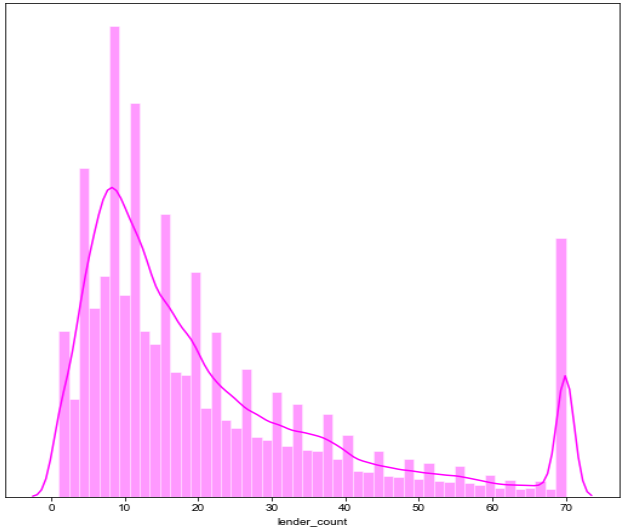
EDA: EXPLORE, CLEAN, EXPLORE SOME MORE…

- Plotted distributions of the four continuous variables within the Kiva dataset.



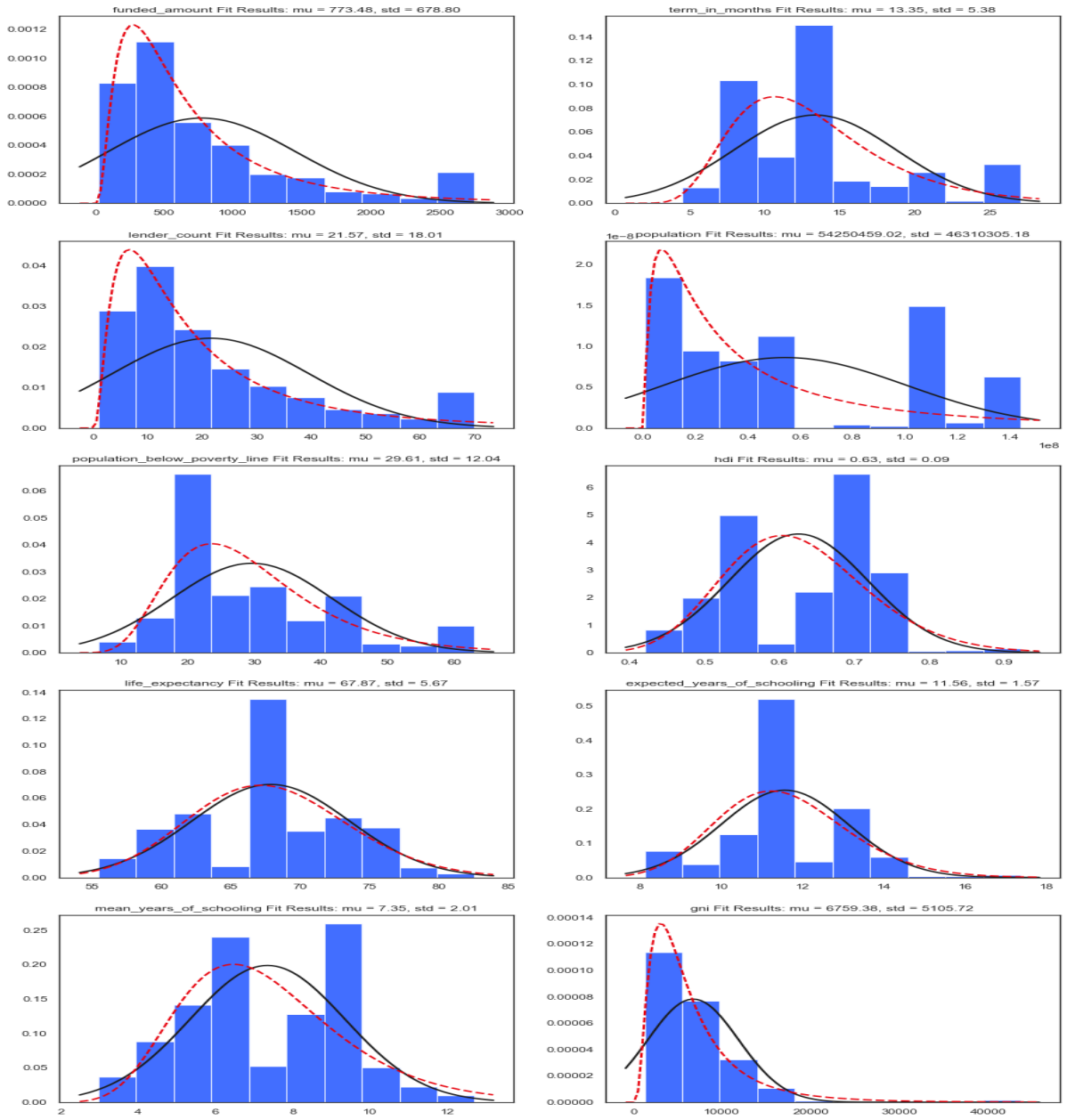






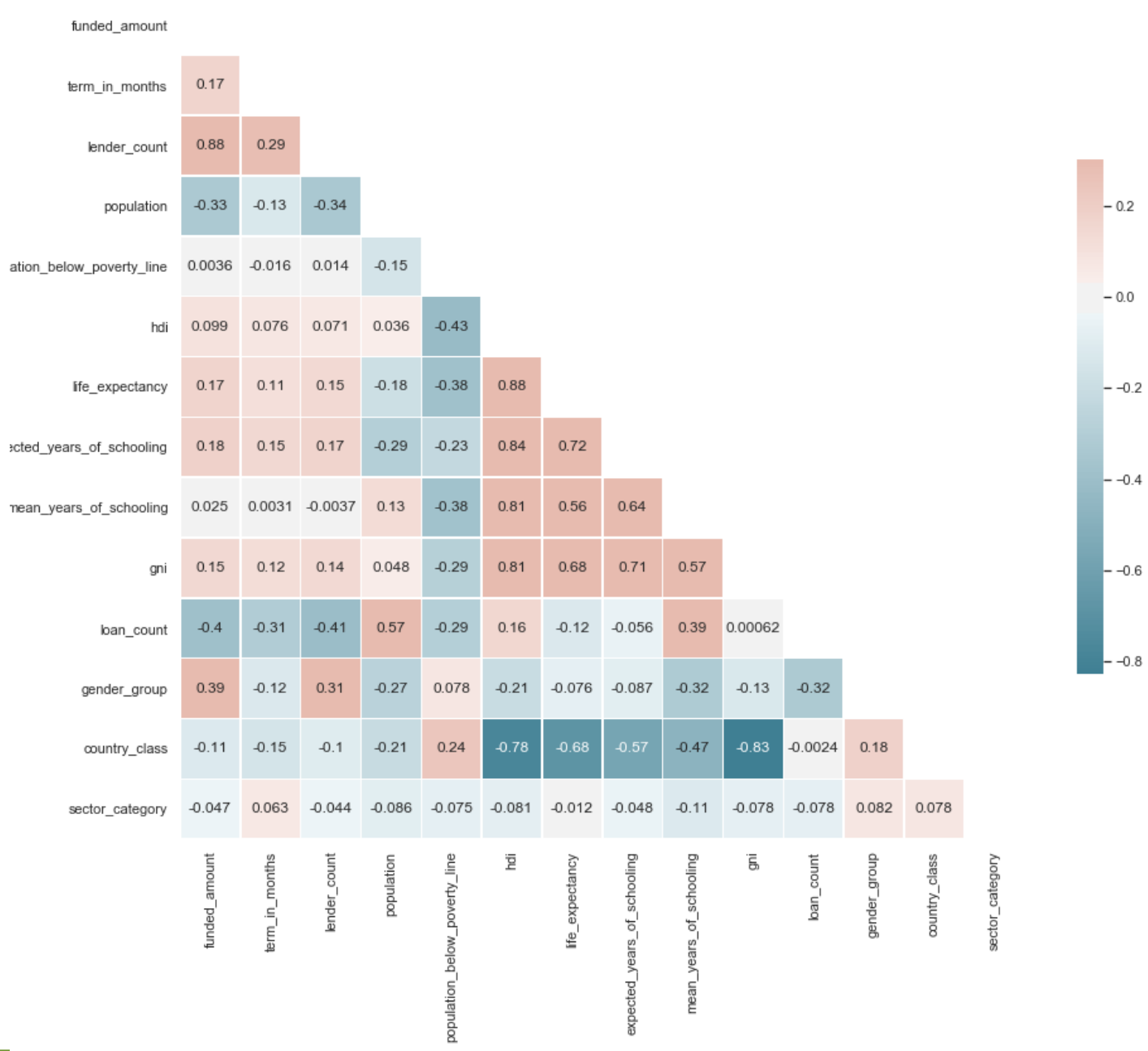
EDA: EXPLORE, CLEAN, EXPLORE SOME MORE…

- Combined country_stats dataset with Kiva dataset
- Converted categorical object features into categorical numerical features
- Plotted distributions of combined continuous variables
- Looked at regular, normal and lognormal distribution of datasets
  - Black – PDF normal
  - Red – PDF lognormal



CORRELATIONS

- Prior to building models ran correlations on existing data set
- Identified the highest (and lowest) correlated features
- Did not consider these for the purposes of the first pass at building the models
- Used data for PCA analysis



BUILDING THE MODEL(S):

CREATED TRAINING & TEST SETS USING 80/20 SPLIT
THEN, FOR PROCESSING SPEED PURPOSES, RANDOMLY SAMPLED 50K DATA POINT

DECISION REGRESSION TREE

Sample size : 50K

Y = Funded Amount

X features= 16

Results
Mean Absolute Error: 158.88
Mean Absolute Percentage Error: 0.30
Mean Squared Error: 67798.09
RMSE: 260.38

Cross Validation Score: .839 > <.865

RANDOM FOREST

Sample size: 50K

Y = Funded Amount

X features =16

Results
Mean Absolute Error: 138.35
Mean Absolute Percentage Error: 0.22
Mean Squared Error: 55113.60
RMSE: 234.76

XG BOOSTING

Stats to include:

Sample size

Y = Funded Amount

X features = 16

Results
Mean Absolute Error: 141.312
Mean Absolute Percentage Error: 0.247
Mean Squared Error: 50369.01
RMSE: 224.43

BUILDING THE MODEL(S):

CREATED TRAINING & TEST SETS USING 80/20 SPLIT
THEN, FOR PROCESSING SPEED PURPOSES, RANDOMLY SAMPLED 50K DATA POINT

KNN REGRESSION

Sample size = 50K

Y = Funded Amount

X features = 4
(Lender Count, HDI, Term in Months, Gender Group)

Results

Unweighted Accuracy: 0.86 (+/- 0.012)
Weighted Accuracy: 0.85 (+/- 0.010)

OLS REGRESSION

Sample size = 50K

Y = Funded Amount

X features = 16

Results

R^2: .809
Mean Absolute Error: 190.841
Mean Absolute Percentage Error: 0.32
Mean Squared Error: 91027.29
RMSE: 301.70

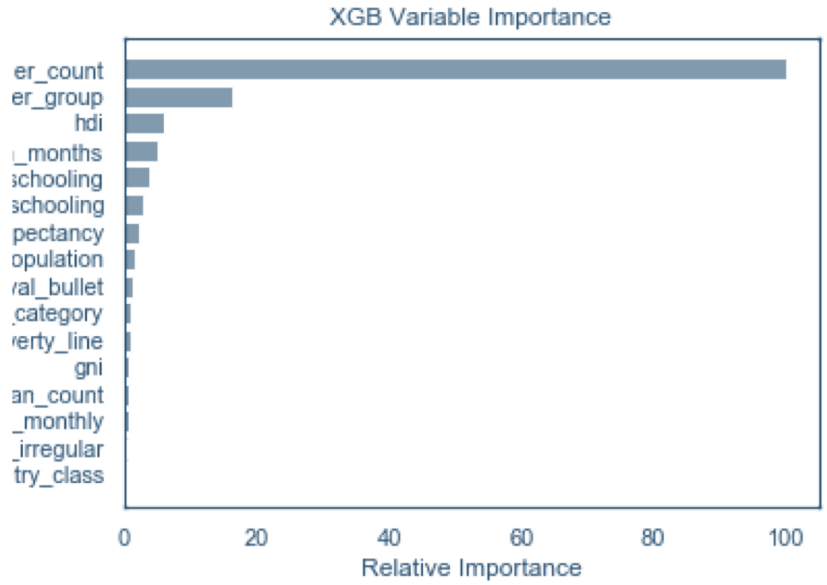
CONCLUSIONS:

- On average, an entrepreneur could expect to receive $773 in funds from a Kiva loan
  - This amount has an error of +/- 22%

- To achieve this estimated loan amount an entrepreneur would be better served to have the following characteristics:
  - Live in a country with a high number of lenders on the Kiva platform
  - Gender of borrower is important
  - The country should have a high Human Development Index number as per the UN
  - The population should have schooling past middle school

- ***Choose the KNN Regression model over the others basis the higher accuracy score***

Model	Actual Funded_Amount	Predicted Funded_Amount	MAPE
KNN Regressor	***766.67***	***772.86***	***14%***
Random Forest	765.08	770.46	22%
OLS	778.31	775.81	32%
Average	770.02	773.04	22%



IMPROVEMENTS

- Improve accuracy of model(s). Shortcomings include:
  - The data is not normally distributed
  - Only model the amount lent not the lending decision
  - Dissection of the top 5 features
    - Which countries?
    - What level of HDI is the breaking point?
    - Which gender is most important for the lending decision?
- Introduction of new features to improve accuracy
  - What other decisions do people make prior to lending?
  - Model the default rate or impact of lending term to funds lent
- Create Ensemble model
  - Classification for predicting if people will lend and if so towards what?
  - Input from classification to better predict how much is lent and for how long

QUESTIONS & ANSWERS

THANK YOU