

# Kiva/クラウドファンディングの資金調達額予測 2nd Place Solution

2022年3月2日

team-maruyama (maruyama\*, yo-zef2, ko)

# 本日本話しする内容 ～モデル改善の過程～

1. ドメイン知識の獲得
  - 1-1. 利子と仲介料
  - 1-2. 為替リスク
2. データ観察
  - 2-1. リークの有無
  - 2-2. 訓練データとテストデータの乖離
  - 2-3. 融資額を決める要因
3. ベースラインモデルの作成
4. モデル改善
  - 4-1. 特徴量の抜き差し
  - 4-2. データセットシフトの解消
5. モデルのアンサンブル
6. 後処理の追加

# 本日本話しする内容 ～モデル改善の過程～

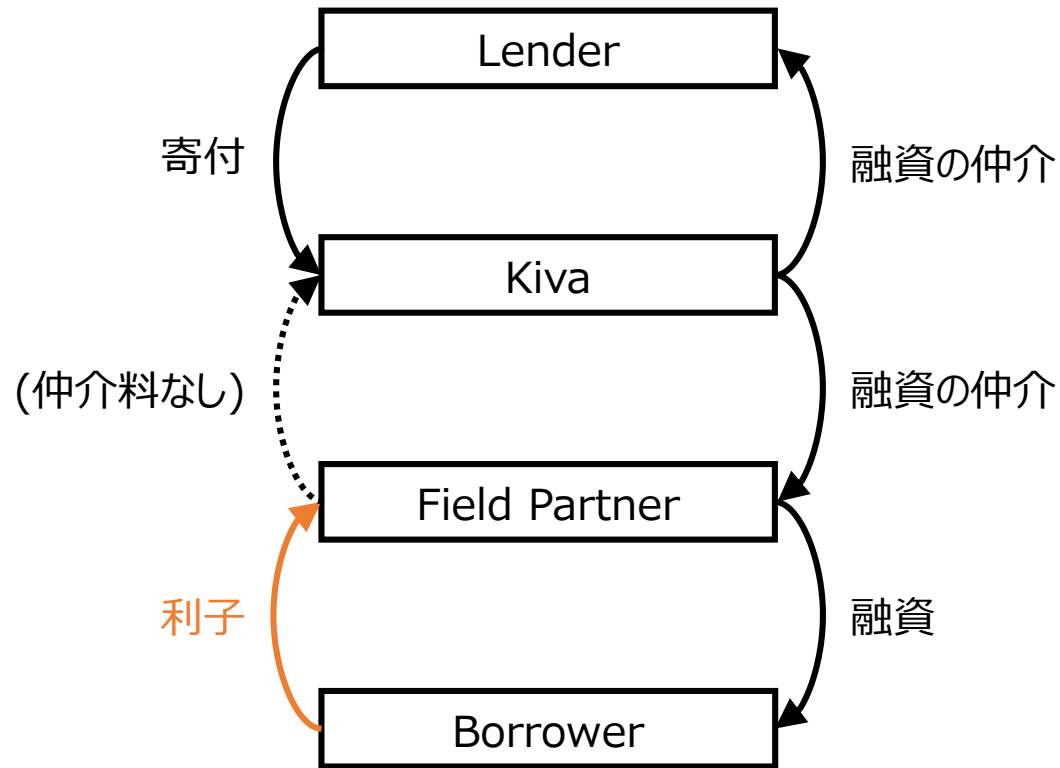
1. ドメイン知識の獲得
  - 1-1. 利子と仲介料
  - 1-2. 為替リスク
2. データ観察
  - 2-1. リークの有無
  - 2-2. 訓練データとテストデータの乖離
  - 2-3. 融資額を決める要因
3. ベースラインモデルの作成
4. モデル改善
  - 4-1. 特徴量の抜き差し
  - 4-2. データセットシフトの解消
5. モデルのアンサンブル
6. 後処理の追加

# 事業活動以外の費用が融資額に含まれるか？ | 利子と仲介料

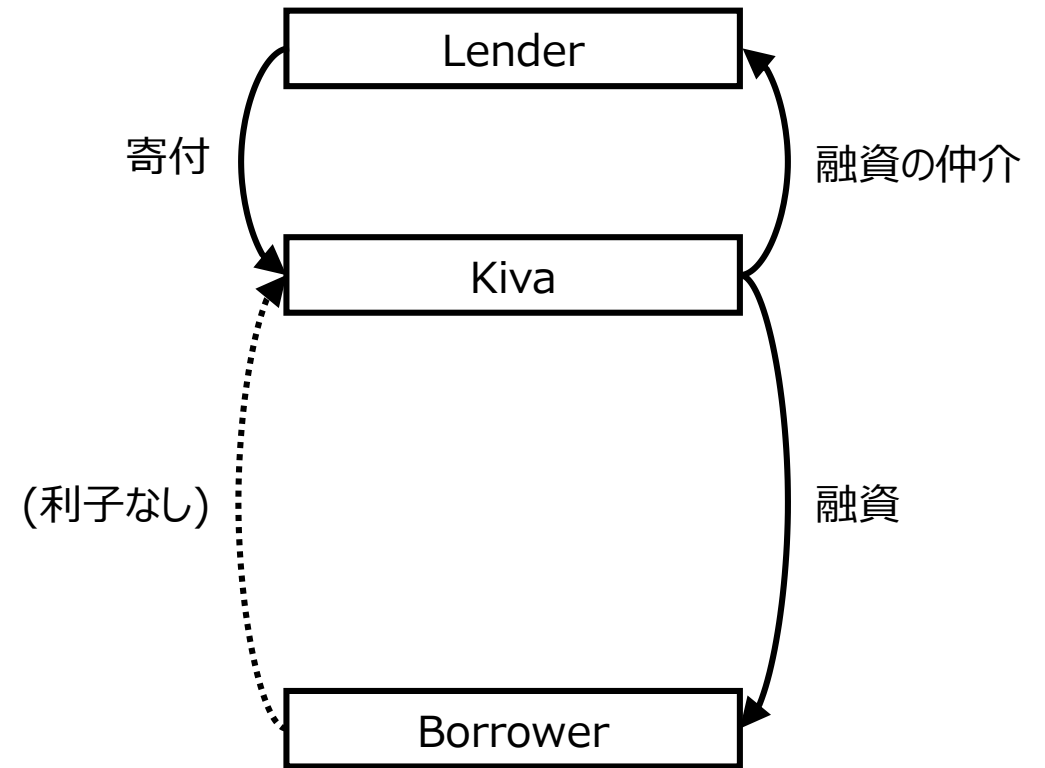
利子が発生するのは、借り手とフィールドパートナーの間だけ。

→ 利子や仲介料が上乗せされることはないため、純粋に事業に必要な額を推定すればよさそう。

Partner Loan



Direct Loan



# 事業活動以外の費用が融資額に含まれるか？ | 為替リスク

為替リスクを借り手が負担するオプションと、貸し手と共有するオプションが用意されている。  
→ 借り手負担ありのオプションを選んだ場合、為替リスクを加味した融資額が設定されていそう。

#	CURRENCY_POLICY	CURRENCY_EXCHANGE_COVERAGE_RATE	借り手の負担割合	貸し手の負担割合
1	standard	NA	100%	0%
2	shared	0.1	10%	90%
3	shared	0.0	0%	100%

# 本日本話しする内容 ～モデル改善の過程～

1. ドメイン知識の獲得
  - 1-1. 利子と仲介料
  - 1-2. 為替リスク
2. データ観察
  - 2-1. リークの有無
  - 2-2. 訓練データとテストデータの乖離
  - 2-3. 融資額を決める要因
3. ベースラインモデルの作成
4. モデル改善
  - 4-1. 特徴量の抜き差し
  - 4-2. データセットシフトの解消
5. モデルのアンサンブル
6. 後処理の追加

# リークの有無

DESCRIPTIONに希望額が書かれている。LOAN\_IDが時系列を表している。

→ テスト期間に近い訓練データから為替レートを求め、希望額をドル換算すれば、正確な融資額を特定できそう。

## Kiva APIから、テストデータの融資額がわかる。

LOAN\_IDが仮名化されていないため、以下のURLにアクセスすると融資額がわかる。

[https://www.kiva.org/lend/<LOAN\\_ID>](https://www.kiva.org/lend/<LOAN_ID>)

→ 外部データの利用は禁止されているため、使えない。

## DESCRIPTIONから、一部のテストデータの融資額がわかる。

一部の案件ではDESCRIPTIONに希望額（ただし、現地通貨）が書かれており、かつ基本的に希望通りに融資が行われているため、DESCRIPTIONから融資額がわかる。

→ 希望額を予測値とすることで、正確に予測できる。

## LOAN\_IDから大まかな時期がわかる。

新しい案件ほど大きな値になるようLOAN\_IDが振られているため、

LOAN\_IDで昇順にソートすると時系列順に並び変えられる。

→ テスト期間に近い訓練データだけ使うことで、データセットシフトを緩和できる。

# 訓練データとテストデータの乖離 | コロナ禍

訓練期間はコロナ禍前 (2018～2019)、テスト期間はコロナ禍 (2020～2021)。

→ 訓練データとテストデータで分布が大きく異なっていそう。ただし異なっていることに気付くのは難しそう。

コロナ関連案件の  
特設ページが設けられており  
他の案件より目立つ

グループ案件だが  
コロナ対策で  
集合写真を取れないため  
代表者のみの  
画像が記載されている

The screenshot shows the Kiva website's 'Covid-19' category page. At the top, there's a navigation bar with 'kiva', 'Lend', a search bar, and links for 'Borrow', 'About', and 'Log in'. Below this, the page is titled 'Covid-19' with a sub-header 'All Loans > Covid-19'. A paragraph explains that the COVID-19 pandemic has economically impacted people worldwide and that Kiva is providing a financial safety net. Below this, three loan cards are displayed. The first card is for Samuel in Uganda, with a photo of him in front of a shelf of water bottles. The second card is for Josiah's Group in Kenya, with a photo of a man in a blue shirt. The third card is for Salima in Tajikistan, with a photo of her standing next to a horse. Salima's card is highlighted with an orange border. Each card includes a loan amount, a description of the loan's purpose, a progress bar, and a 'Lend now' button. To the right of the cards, there is a 'Salima's story' section with text about her background and her request for a loan.

## Salima's story

Salima is a hardworking, patient Tajik woman with big potential and interest in agriculture. She was born in 1966 year in Shahristan, Tajikistan. Just look at the photo how beautiful Shahristan region is, mountains are covered with snow, fresh air and an excellent tourist place in winter and summer.

She is married and has six adult children. She has been actively engaged in livestock breeding and farming activities. She has favorable conditions for her business and with great zeal and desire does her favorite work. She wants to expand her business, acquire farming supplies.

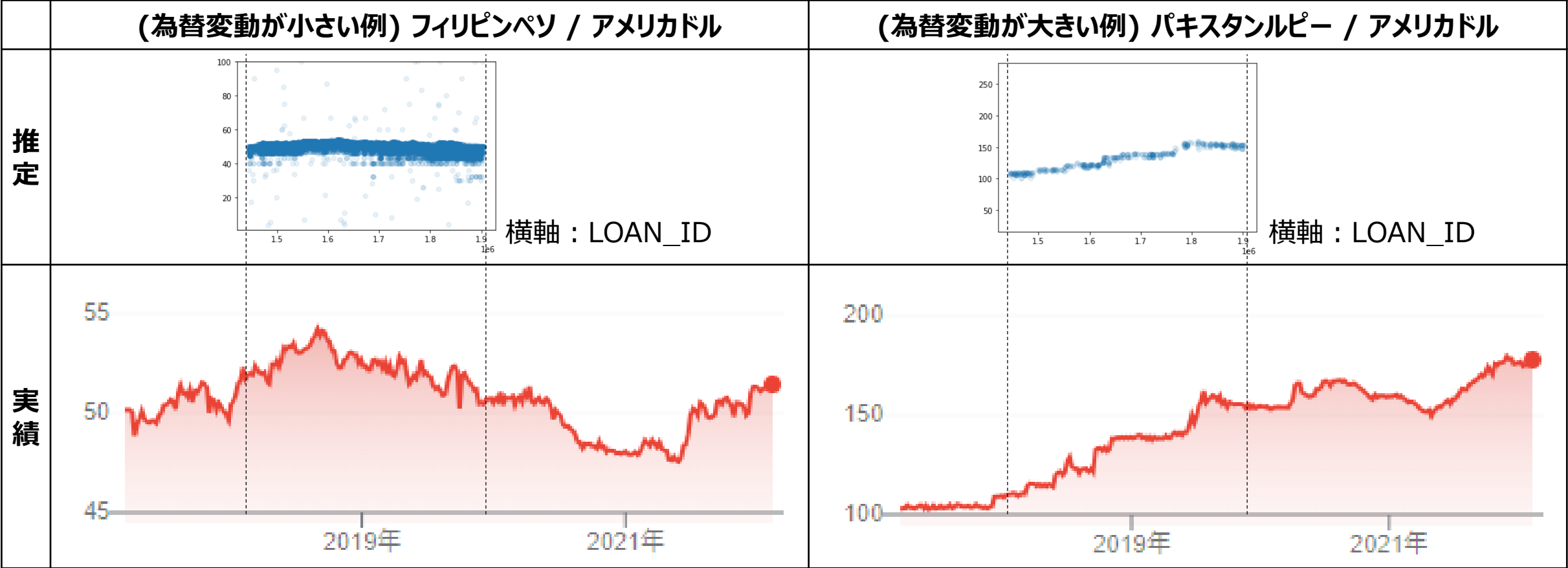
Salima is asking for a loan with Kiva's partner IMON for the first time. This loan will help her purchase seeds, minerals and necessary farming supplies. She hopes for your financial support.

コロナ関連案件として  
ピックアップされていても  
本文中に「Covid-19」という  
キーワードが含まれているわけではない  
(テストデータ91822件中、  
「Covid-19」を含むのは6569件だけ)



# 訓練データとテストデータの乖離 | 為替変動

通貨によっては、為替レートが大きく変動している。  
→ 同じ融資内容でも為替の影響で時期によって融資額が変わるはず。補正したほうがよさそう。



# 訓練データとテストデータの乖離 | 為替リスクプログラムの変化

「為替リスクの借り手負担なし」オプションが始まったのは、2019年8月から（＝訓練データ終盤）。  
→「為替リスクの借り手負担なし」オプション導入後の訓練データを重視したほうがよさそう。

CURRENCY_POLICY				shared		standard	
CURRENCY_EXCHANGE_COVERAGE_RATE		0.0		0.1		NaN	
usage		train	test	train	test	train	test
CURRENCY							
訓練期間は まだ事例が少なく テスト期間に 本格導入されている	KES	1336.0	9155.0	11174.0	2684.0	NaN	605.0
	TJS	619.0	4386.0	3029.0	NaN	NaN	NaN
	UGX	656.0	2790.0	4895.0	297.0	NaN	1.0
	PHP	511.0	2704.0	16857.0	316.0	3897.0	25165.0
	XOF	NaN	2387.0	2548.0	1042.0	NaN	NaN
	PYG	367.0	1771.0	1301.0	27.0	NaN	NaN
	KHR	131.0	1289.0	2586.0	848.0	51.0	374.0
	KGS	193.0	1207.0	787.0	NaN	NaN	NaN
	HNL	146.0	660.0	1093.0	51.0	NaN	NaN
	RWF	39.0	603.0	816.0	407.0	NaN	NaN

CURRENCY_POLICY				shared		standard	
CURRENCY_EXCHANGE_COVERAGE_RATE		0.0		0.1		NaN	
usage		train	test	train	test	train	test
CURRENCY							
訓練期間には 事例がなく テスト期間に初めて 事例が発生している	MZN	41.0	423.0	469.0	15.0	NaN	NaN
	CRC	26.0	344.0	233.0	NaN	NaN	NaN
	ALL	23.0	336.0	307.0	NaN	NaN	NaN
	COP	NaN	316.0	4758.0	1649.0	NaN	NaN
	INR	169.0	308.0	1954.0	NaN	NaN	NaN
	GEL	NaN	248.0	375.0	NaN	NaN	NaN
	BRL	NaN	233.0	221.0	29.0	NaN	1.0
	MDL	NaN	224.0	201.0	NaN	NaN	NaN
	EUR	NaN	116.0	132.0	NaN	NaN	NaN
	GHS	NaN	87.0	778.0	446.0	NaN	NaN

# 訓練データとテストデータの乖離 | Kiva導入地域の変化

地域によっては、訓練期間とテスト期間で案件数が大きく異なる。

→ テスト期間に多く現れる国を重視したほうが良さそう。

テスト期間のほうが多い国 上位10件

	usage	train	test	diff
COUNTRY_NAME				
Philippines	21265	28185	6920	
Nicaragua	1406	2675	1269	
Madagascar	1583	2664	1081	
Tajikistan	3664	4388	724	
The Democratic Republic of the Congo	473	923	450	
Burkina Faso	606	1011	405	
Ecuador	3424	3818	394	
Sierra Leone	122	495	373	
Indonesia	709	1082	373	
Zambia	98	436	338	

テスト期間のほうが少ない国 上位10件

	usage	train	test	diff
COUNTRY_NAME				
Colombia	4758	1965	-2793	
Uganda	5551	3088	-2463	
India	2124	308	-1816	
Pakistan	1547	316	-1231	
El Salvador	4100	3041	-1059	
Peru	1725	879	-846	
Honduras	1239	711	-528	
Nigeria	1183	701	-482	
Samoa	1352	880	-472	
Ghana	786	548	-238	

# 融資額を決める要因 | 用途

事業拡大の費用を募る場合と、生活に充てる費用を募る場合がある。  
→ 事業用途のほうが融資額が大きく、生活用途のほうが融資額が小さそう。

## LOAN\_USE 出現頻度上位20件

---

to build a sanitary toilet for her family	3074
to build a sanitary toilet for her family.	1583
to buy a water filter to provide safe drinking water for their family.	931
to buy a water filter to provide safe drinking water for her family.	573
to buy ingredients for her food production business	449
to build a sanitary toilet	415
to buy feed and other supplies to raise her pigs.	305
to purchase hybrid seeds and fertilizer to improve harvests of maize	250
to buy items to sell like beverages, canned goods, junk food, and other groceries	229
to purchase hybrid seeds and fertilizer for the cultivation of maize, as well as a solar light.	223
to buy feeds, vitamins, and other supplies to raise her livestock	186
to buy construction materials.	179
to buy fertilizers and other farm supplies	170
to buy feed, vitamins, and other supplies to raise her livestock.	164
to buy items like canned goods, personal care products, etc. to sell in her general store.	157
to purchase hybrid seeds and fertilizer for the cultivation of maize	155
to access premium seeds and high quality fertilizer for 0.5 acres of maize, in addition to advice and insurance, optimizing for increased productivity and profits	145
to buy feed and other supplies to raise her livestock.	141
to buy feeds and other supplies to raise her livestock	139
to buy a water filter to provide safe drinking water for his family.	123

# 融資額を決める要因 | 用途

用途によって融資額が異なる。

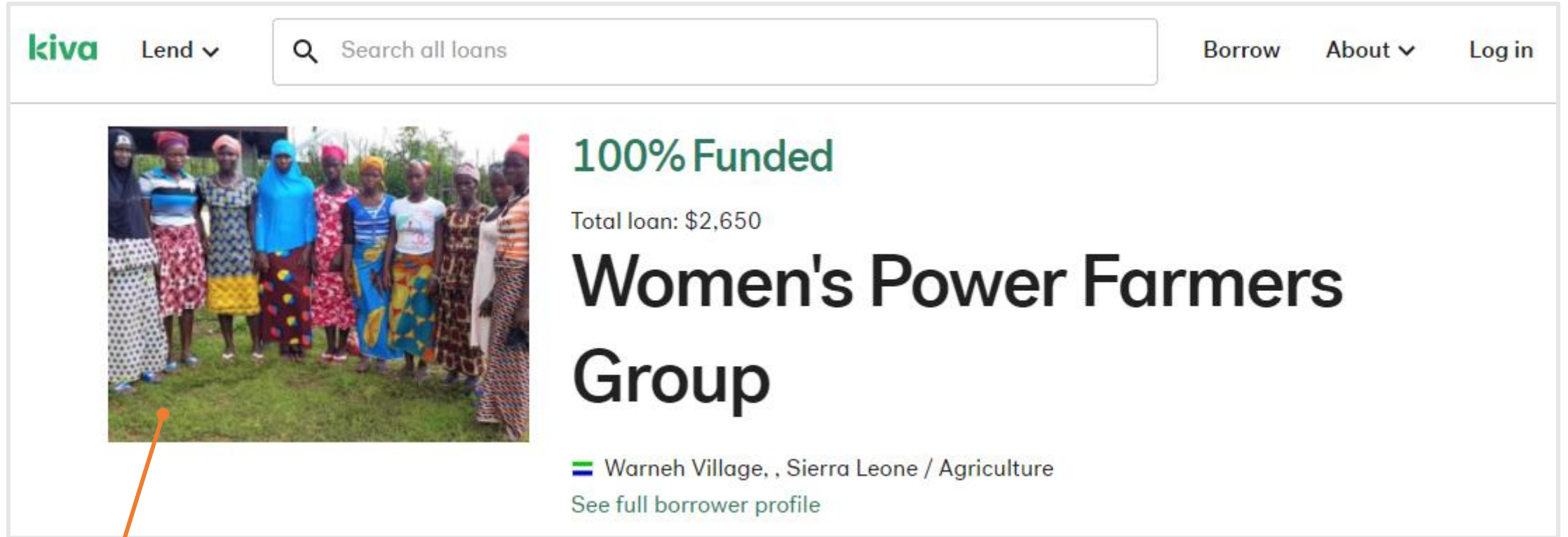
→ 用途やその背景が含まれているDESCRIPTIONとLOAN\_USEを重視するのがよさそう。

LOAN\_USEの先頭から2番目に出現する単語 出現頻度上位20件

	count	mean	std	min	25%	50%	75%	max		count	mean	std	min	25%	50%	75%	max
verb									verb								
buy	52384.0	768.772431	1018.473454	25.0	250.0	475.0	875.00	10000.0	to	393.0	1099.491094	1667.178202	25.0	350.0	650.0	1025.00	10000.0
purchase	16224.0	721.355708	827.109932	25.0	300.0	500.0	875.00	10000.0	help	329.0	684.194529	1075.113879	75.0	300.0	400.0	700.00	10000.0
build	6323.0	237.707576	238.121868	25.0	125.0	200.0	200.00	6350.0	repair	282.0	804.964539	891.761022	75.0	300.0	575.0	1000.00	10000.0
pay	5008.0	738.138978	796.581121	25.0	275.0	500.0	925.00	10000.0	renovate	244.0	939.036885	994.687636	50.0	325.0	675.0	1150.00	6600.0
expand	1009.0	502.874133	758.929426	50.0	300.0	375.0	450.00	10000.0	restock	191.0	649.345550	581.227346	125.0	325.0	550.0	825.00	5750.0
invest	828.0	682.216184	885.056655	50.0	250.0	475.0	825.00	10000.0	a	190.0	1178.289474	2035.652553	50.0	200.0	375.0	1000.00	10000.0
stock	740.0	865.202703	1425.847667	100.0	225.0	425.0	825.00	9850.0	start	180.0	653.055556	752.587044	100.0	300.0	500.0	731.25	7825.0
add	736.0	583.118207	594.531049	25.0	300.0	500.0	781.25	8025.0	borrower	160.0	611.093750	470.495414	75.0	225.0	462.5	956.25	2000.0
access	682.0	149.046921	73.041039	75.0	100.0	150.0	150.00	1000.0	more	136.0	1001.470588	1595.275910	100.0	300.0	512.5	1000.00	10000.0
increase	523.0	648.231358	948.689850	25.0	275.0	400.0	700.00	10000.0	make	127.0	1119.094488	1306.483590	225.0	425.0	825.0	1237.50	10000.0

# 融資額を決める要因 | 人数

事業を立ち上げるためにグループを作って融資を募っている場合がある。  
→ 人数が多いほど大規模な事業になるから、融資額も増えそう。



The screenshot shows the Kiva website interface. At the top, there is a navigation bar with the Kiva logo, a 'Lend' dropdown menu, a search bar containing 'Search all loans', and links for 'Borrow', 'About', and 'Log in'. Below the navigation bar, the main content area features a loan listing for the 'Women's Power Farmers Group'. On the left of the listing is a photograph of a group of approximately 10 women standing outdoors in a grassy area, wearing colorful traditional patterned dresses and headwraps. An orange line points from the bottom of this photo towards the explanatory text below. To the right of the photo, the text reads '100% Funded' in green, followed by 'Total loan: \$2,650'. The group's name, 'Women's Power Farmers Group', is displayed in large, bold black font. Below the name, it says 'Warneh Village, , Sierra Leone / Agriculture' with a small flag icon, and a link 'See full borrower profile'.

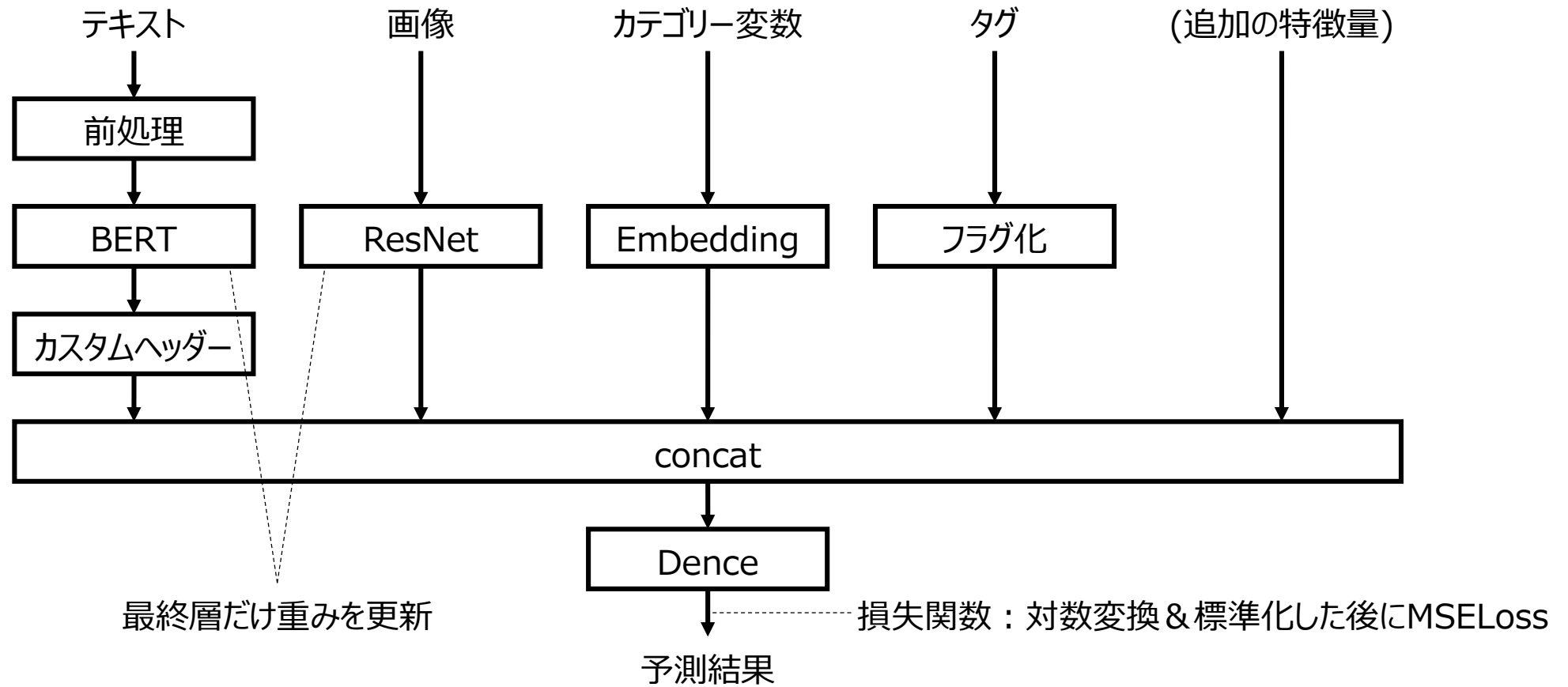
DESCRIPTIONに人数が書かれていないことが多いため、画像から数えるしかなさそう。

# 本日本話しする内容 ～モデル改善の過程～

1. ドメイン知識の獲得
  - 1-1. 利子と仲介料
  - 1-2. 為替リスク
2. データ観察
  - 2-1. リークの有無
  - 2-2. 訓練データとテストデータの乖離
  - 2-3. 融資額を決める要因
3. ベースラインモデルの作成
4. モデル改善
  - 4-1. 特徴量の抜き差し
  - 4-2. データセットシフトの解消
5. モデルのアンサンブル
6. 後処理の追加

# ベースラインモデル

データの種別に合わせて特徴量を抽出し全結合層を通して予測値を得る、NNによるベーシックなマルチモーダルモデルを作成。この時点でのMAEは約270。





# 本日本話しする内容 ～モデル改善の過程～

1. ドメイン知識の獲得
  - 1-1. 利子と仲介料
  - 1-2. 為替リスク
2. データ観察
  - 2-1. リークの有無
  - 2-2. 訓練データとテストデータの乖離
  - 2-3. 融資額を決める要因
3. ベースラインモデルの作成
4. モデル改善
  - 4-1. 特徴量の抜き差し
  - 4-2. データセットシフトの解消
5. モデルのアンサンブル
6. 後処理の追加

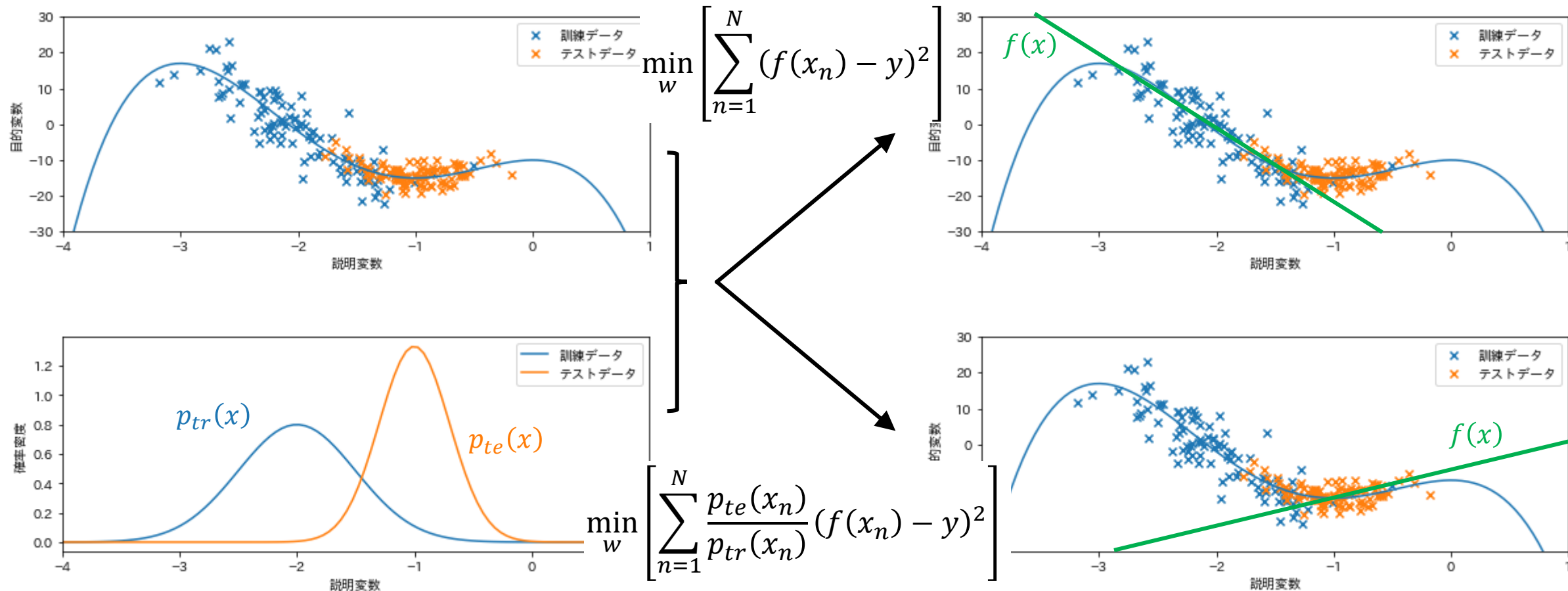
# 特徴量の抜き差し

画像の削除、手作業で設計した特徴量の追加、BERTのカスタムヘッダーなど、予測器に入れる特徴量をいろいろ変えてみたものの、どれもMAEは270前後で変化なし。

- 画像  
→ 抜いても精度に変化なし。
- 画像から抽出した人数  
→ 入れても精度に変化なし。
- DESCRIPTIONから抽出した希望額  
→ 入れても精度に変化なし。
- DESCRIPTIONに対する前処理 (HTMLタグの削除など)  
→ やってもやらなくても精度に変化なし。
- BERTのカスタムヘッダー  
→ CLSのみとConv1dを試したが、精度に変化なし。

# データセットシフトの解消

訓練データとテストデータの乖離を解消するため、訓練データとテストデータの密度比を推定し損失関数を重み付け※したもの、MAEはあまり変わらなかった。



# 本日本話しする内容 ～モデル改善の過程～

1. ドメイン知識の獲得
  - 1-1. 利子と仲介料
  - 1-2. 為替リスク
2. データ観察
  - 2-1. リークの有無
  - 2-2. 訓練データとテストデータの乖離
  - 2-3. 融資額を決める要因
3. ベースラインモデルの作成
4. モデル改善
  - 4-1. 特徴量の抜き差し
  - 4-2. データセットシフトの解消
5. モデルのアンサンブル
6. 後処理の追加

# モデルのアンサンブル

これまでに得られたモデルの予測結果を、加算平均を取ってアンサンブルした。  
MAEは約15改善した。

- 異なる特徴量で学習させたモデル
  - 画像のあり/なし、追加の特徴量のあり/なし
  - BERTの各種カスタムヘッダー
- 異なる学習済みモデルで学習させたモデル
  - RoBERTa、DeBERTa
- 異なる損失関数で学習させたモデル
  - 密度比による重み付き損失関数
- 学習過程で得られたモデル
  - 10エポックまで回したが、どのモデルも8エポックあたりで学習が収束していたため、8エポック目と9エポック目のモデルもアンサンブルに加えた。

# 本日本話しする内容 ～モデル改善の過程～

1. ドメイン知識の獲得
  - 1-1. 利子と仲介料
  - 1-2. 為替リスク
2. データ観察
  - 2-1. リークの有無
  - 2-2. 訓練データとテストデータの乖離
  - 2-3. 融資額を決める要因
3. ベースラインモデルの作成
4. モデル改善
  - 4-1. 特徴量の抜き差し
  - 4-2. データセットシフトの解消
5. モデルのアンサンブル
6. 後処理の追加

# 後処理

テスト期間に近い訓練データから推定した為替レートで希望額をドル換算し、予測値を置換。  
MAEは5程度改善した。

## 1. 為替レートの推定

1-1. テスト期間に近い訓練データに絞る

(訓練データをLOAN\_IDでソートしたときの末尾25%程度を使った)。

1-2. 訓練データのDESCRIPTIONから希望額を抽出する。

1-3. 希望額を融資額で割り、平均を取る。平均値を為替レートとみなす。

## 2. 予測値の補正

2-1. テストデータのDESCRIPTIONから希望額を抽出する。

2-2. 希望額に為替レートをかけ、ドルに換算する。

2-3. 希望額を抽出できた事例の予測値を、希望額に置き換える。

# まとめ

## 1. ドメイン知識の獲得

1-1. 利子と仲介料

→ 気にする必要なし

1-2. 為替リスク

→ 気にする必要あり

## 2. データ観察

2-1. リークの有無

→ 希望額で予測値を置換することを決定

2-2. 訓練データとテストデータの乖離

→ 密度比で損失関数を重み付けすることを決定

2-3. 融資額を決める要因

→ テキストに重点を置くことを決定

## 3. ベースラインモデルの作成

→ MAE 約270

## 4. モデル改善

→ MAE 約270 (特に改善せず)

4-1. 特徴量の抜き差し

4-2. データセットシフトの解消

## 5. モデルのアンサンブル

→ MAE 約255

## 6. 後処理の追加

→ MAE 約250