

家の説明文からの家賃予測

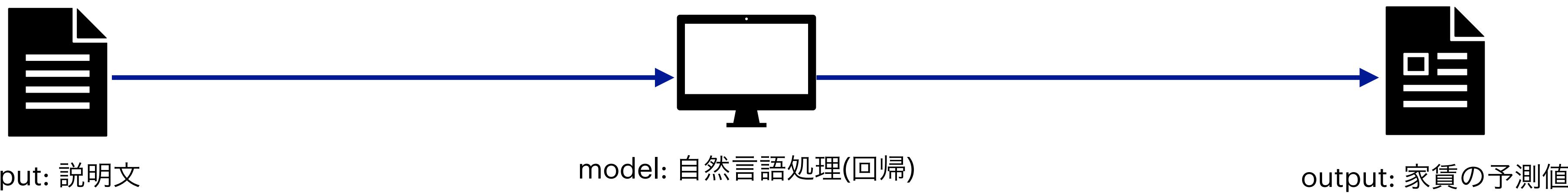
2021/11 Sekikawa

課題設定

- 取り組む課題
 - 一人暮らしをする際に、適切な家賃が分からぬいため、お得な物件が分からぬ
 - 賃貸の説明文から適切家賃を出力できるようなモデルを作成する
- 取り組む意義
 - 賃貸を探す側は、物件を選ぶ際の指標が見える化できるようになる
 - 賃貸を提供する側は、その賃貸の目安の家賃を考慮できるようになる
 - 賃貸サイトに導入すれば、他の賃貸サイトとの差別化が図れる
 - コロナが落ち着いてきたことによる出社率の増加により、一人暮らしニーズも増加しているはず
 - テーブルデータでモデル作成する場合と比べて、精度は劣りそうだが自然言語をインプットにできる点で差別化出来そう

作成するモデル

- 賃貸の説明文から家賃を予測するモデル
 - ユーザが家の説明文を入力すると、そのテキストの適切値(モデルの予測値)を返してくれるイメージ
 - テーブルデータではなく、テキストを入力としてモデルを作成することで、ユーザの入力テキストをそのまま受け取って予測を出力できる



バストイレ別、バルコニー、エアコン、ガスコンロ対応、シャワー付洗面台、TVインターホン、浴室乾燥機、室内洗濯置、シューズボックス、システムキッチン、南向き、追焚機能浴室、角住戸、温水洗浄便座、洗面所独立、洗面化粧台、駐輪場、宅配ボックス、3口以上コンロ、ペット相談、グリル付、保証人不要、cs、ネット使用料不要、浄水器、ダブルロックキー、24時間換気システム、複層ガラス、人感照明センサー、都市ガス、室内物干機、BS、IT重説 対応物件、初期費用カード決済可、家賃カード決済可

損失関数 : MSE
→ 結果を評価する際は MAE で見る

家賃の予測値 : 78000円など

作成するモデル

- モデル(自然言語処理)部分の詳細

- 単語の羅列を入力とするため、RNN等の順序性を考慮するモデルは使用しない
- 単語列 → エンベディング → BoW → Linear(3層) → 家賃の予測値

入力(単語列)	エンベディング	BoW(平均)	Linear	出力(家賃の予測値)
バストイレ別	[0.3, 0.2, 0.1, 0.8, 0.1, 0.5, 0.3, 0.1]		Linear(8, 32)	
ペット相談	[0.2, 0.1, 0.1, 0.2, 0.4, 0.1, 0.3, 0.4]	[0.2, 0.2, 0.1, 0.4, 0.2, 0.3, 0.3, 0.2]	Linear(32, 16)	78000
浴室乾燥機	[0.1, 0.3, 0.1, 0.2, 0.1, 0.3, 0.3, 0.1]	(BoWに関しては加算・平均両方試しました)	Linear(16, 1)	

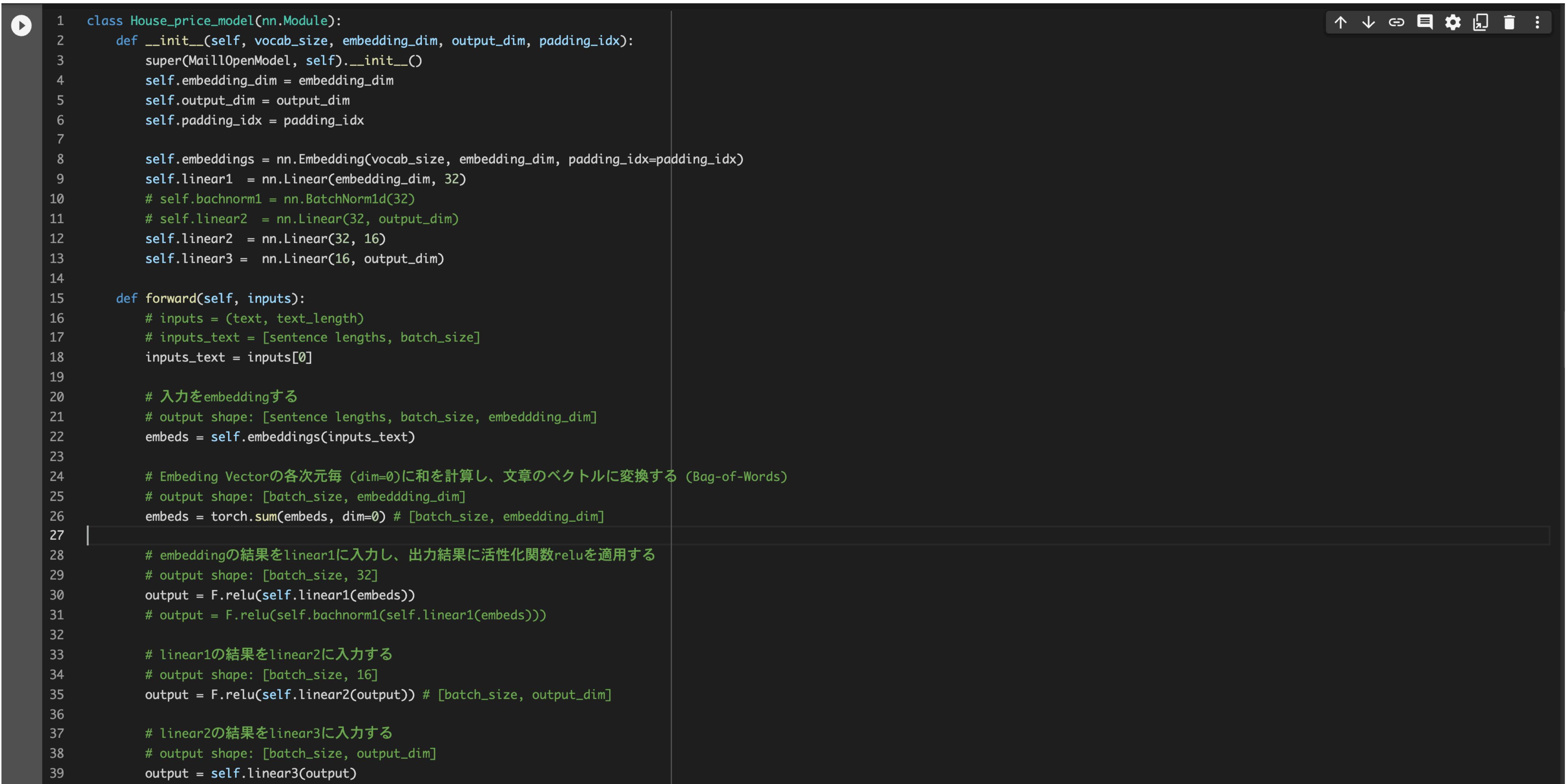
今回作成するモデル

入力(単語列)	エンベディング	BoW(平均)	Linear	出力(家賃の予測値)
バストイレ別	[0.3, 0.2, 0.1, 0.8, 0.1, 0.5, 0.3, 0.1]		Linear(8, 32)	
ペット相談	[0.2, 0.1, 0.1, 0.2, 0.4, 0.1, 0.3, 0.4]	[0.2, 0.2, 0.1, 0.4, 0.2, 0.3, 0.3, 0.2]	Linear(32, 16)	78000
浴室乾燥機	[0.1, 0.3, 0.1, 0.2, 0.1, 0.3, 0.3, 0.1]		Linear(16, 1)	

• 考えていたこと

- ・ エンベディングにおいて、家賃に対して「どの程度プラスに働くか」「どの程度マイナスに働くか」の項が単語ごとに学習される
- ・ 「プラスに働く項」「マイナスに働く項」を元にして、家賃の予測値を出力するネットワークが学習される
- ・ そこまで複雑な関係性がなさそうなので、複雑なネットワークは必要なさそう
- ・ 精度の面では特徴量加工して作成した表データを入力としたMLモデルの方が良さそう(NLPはテキストそのまま入れられるので差別化はできるはず)

モデルのソース



```
1 class House_price_model(nn.Module):
2     def __init__(self, vocab_size, embedding_dim, output_dim, padding_idx):
3         super(House_price_model, self).__init__()
4         self.embedding_dim = embedding_dim
5         self.output_dim = output_dim
6         self.padding_idx = padding_idx
7
8         self.embeddings = nn.Embedding(vocab_size, embedding_dim, padding_idx=padding_idx)
9         self.linear1 = nn.Linear(embedding_dim, 32)
10        # self.bachnorm1 = nn.BatchNorm1d(32)
11        # self.linear2 = nn.Linear(32, output_dim)
12        self.linear2 = nn.Linear(32, 16)
13        self.linear3 = nn.Linear(16, output_dim)
14
15    def forward(self, inputs):
16        # inputs = (text, text_length)
17        # inputs_text = [sentence lengths, batch_size]
18        inputs_text = inputs[0]
19
20        # 入力をembeddingする
21        # output shape: [sentence lengths, batch_size, embeddding_dim]
22        embeds = self.embeddings(inputs_text)
23
24        # Embeding Vectorの各次元毎 (dim=0)に和を計算し、文章のベクトルに変換する (Bag-of-Words)
25        # output shape: [batch_size, embeddding_dim]
26        embeds = torch.sum(embeds, dim=0) # [batch_size, embedding_dim]
27
28        # embeddingの結果をlinear1に入力し、出力結果に活性化関数reluを適用する
29        # output shape: [batch_size, 32]
30        output = F.relu(self.linear1(embeds))
31        # output = F.relu(self.bachnorm1(self.linear1(embeds)))
32
33        # linear1の結果をlinear2に入力する
34        # output shape: [batch_size, 16]
35        output = F.relu(self.linear2(output)) # [batch_size, output_dim]
36
37        # linear2の結果をlinear3に入力する
38        # output shape: [batch_size, output_dim]
39        output = self.linear3(output)
```

データ収集

- データの収集先

- 賃貸サイトからスクレイピング(scrapy)

- 武蔵小杉駅沿線に絞ってデータ取得
 - 説明文は「、」区切りの単語列

- データの量

- **5190レコード**

- train: 75% (3892レコード)
 - valid: 25% (1298レコード)

部屋の特徴・設備

バストイレ別、バルコニー、エアコン、ガスコンロ対応、クロゼット、フローリング、TVインターホン、浴室乾燥機、オートロック、室内洗濯置、シューズボックス、システムキッチン、角住戸、温水洗浄便座、脱衣所、エレベーター、洗面所独立、2口コンロ、駐輪場、宅配ボックス、外壁タイル張り、2面採光、防犯カメラ、ペット相談、照明付、分譲賃貸、グリル付、保証人不要、敷金1ヶ月、24時間緊急通報システム、2沿線利用可、ディンプルキー、CS、ネット専用回線、ネット使用料不要、ダブルロックキー、24時間換気システム、人感照明センサー、耐火構造、2駅利用可、3駅以上利用可、3沿線以上利用可、駅徒歩10分以内、24時間ゴミ出し可、敷地内ごみ置き場、セキュリティ会社加入済、都市ガス、洗面所にドア、BS、礼金1ヶ月、保証会社利用可、初期費用カード決済可

物件概要

 情報の見方

間取り詳細	洋6.6	構造	鉄筋コン
階建	3階/7階建	築年月	2015年11月
損保	2万円2年	駐車場	-
入居	相談	取引態様	仲介
条件	ペット相談	取り扱い店舗 物件コード	doki 215494
SUUMO 物件コード	100258898297	総戸数	-
情報更新日	2021/11/12	次回更新日	次回更新日は情報更新日より8日以内
契約期間	普通借家 2年		
保証会社	保証会社利用必 指定保証会社 初回保証委託料：月額総賃料の50%、1年後以降：年間保証料10,000円		
ほか初期費用	合計4.4万円（内訳：24時間サポート1.65万円 鍵交換代2.75万円）		
ほか諸費用	更新料 新賃料1.50ヶ月分 口座振替手数料 330円		
備考	巡回管理		

データ例

```
In [330]: def print_data_info(df):
    print("データ数: ", len(df))
    display(df.head())
    print("others: ")
    print(data.loc[100, "others"])

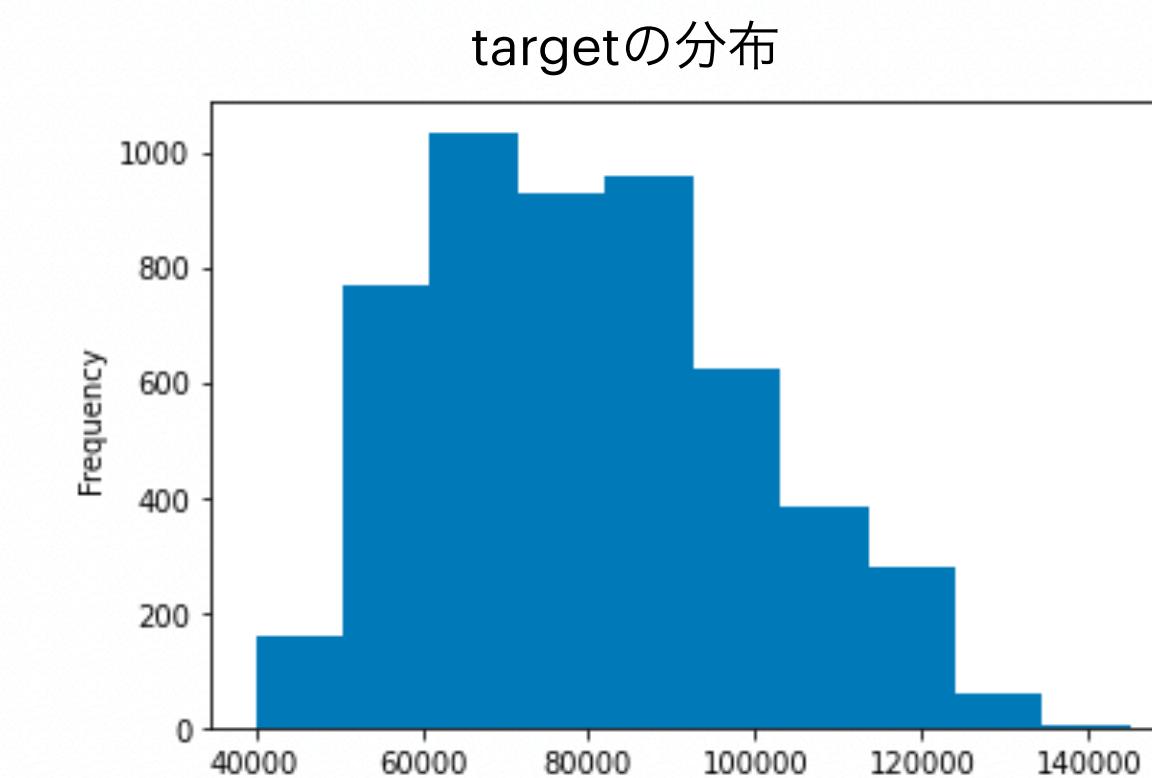
print_data_info(data[["others", "target"]])
```

データ数: 5190

	others	target
0	バストイレ別、バルコニー、エアコン、ガスコンロ対応、クロゼット、浴室乾燥機、オートロック、室...	114000.0
1	バストイレ別、バルコニー、エアコン、ガスコンロ対応、クロゼット、フローリング、シャワー付洗面...	92000.0
2	バストイレ別、バルコニー、エアコン、ガスコンロ対応、クロゼット、TVインターホン、浴室乾燥機...	92000.0
3	バストイレ別、バルコニー、エアコン、クロゼット、フローリング、TVインターホン、浴室乾燥機、...	104000.0
4	バストイレ別、バルコニー、エアコン、ガスコンロ対応、クロゼット、フローリング、TVинтер...	107000.0

others:

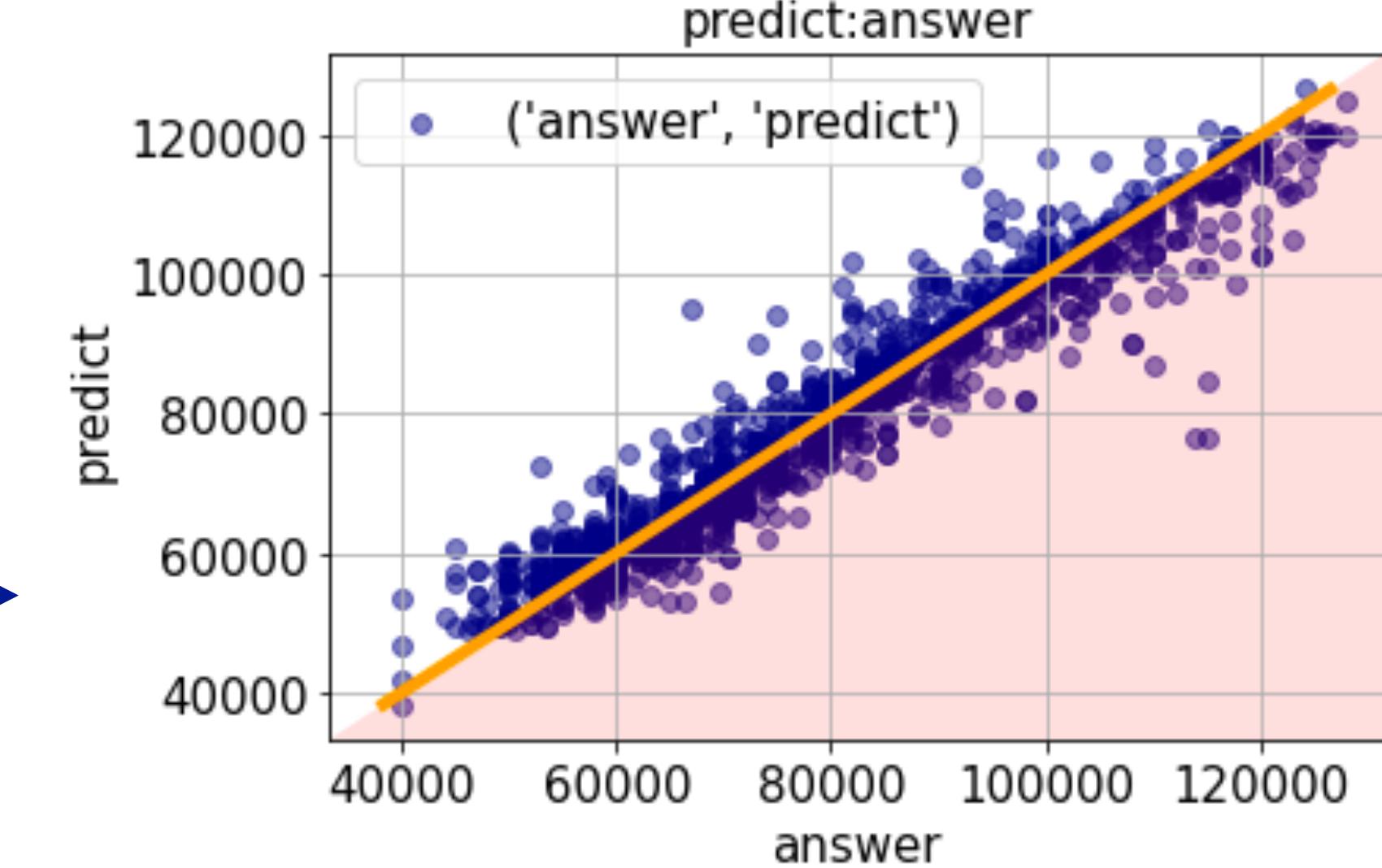
バストイレ別、バルコニー、エアコン、クロゼット、フローリング、TVインターホン、浴室乾燥機、オートロック、室内洗濯置、シューズボックス、システムキッチン、温水洗浄便座、脱衣所、エレベーター、洗面所独立、2口コンロ、駐輪場、宅配ボックス、押入、即入居可、礼金不要、敷金不要、防犯カメラ、ペット相談、分譲賃貸、保証人不要、バイク置場、2沿線利用可、CS、ネット使用料不要、築2年以内、築3年以内、一部フローリング、ペット専用設備、トイレ未使用、2駅利用可、駅徒歩5分以内、駅徒歩10分以内、24時間ゴミ出し可、敷地内ごみ置き場、築5年以内、都市ガス、BS、敷金・礼金不要、保証会社利用可、IT重説 対応物件、初期費用カード決済可



理想の結果

- 右図は同じデータをテーブルデータとし、LightGBMでモデル作成した結果
- オレンジ線より上側に存在する点 = 実際の家賃より高いと見積もった = お得度の高い物件
 - 「Garbage in Garbage out」にはなっていなそう
 - 自然言語インプットでもこうなって欲しい、、

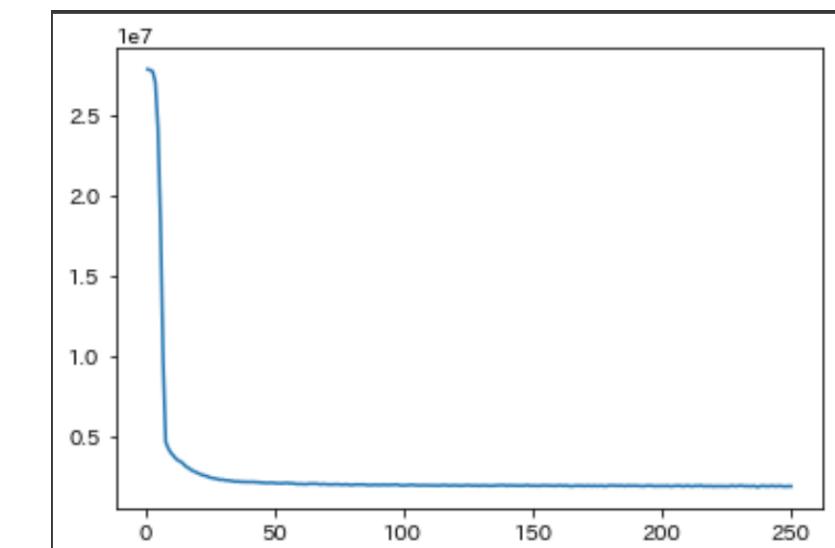
- predict = answer : 妥当な値段の物件(オレンジ線)
- predict < answer : 割高物件(オレンジ線の下)
- predict > answer : 割安物件(オレンジ線の上)**



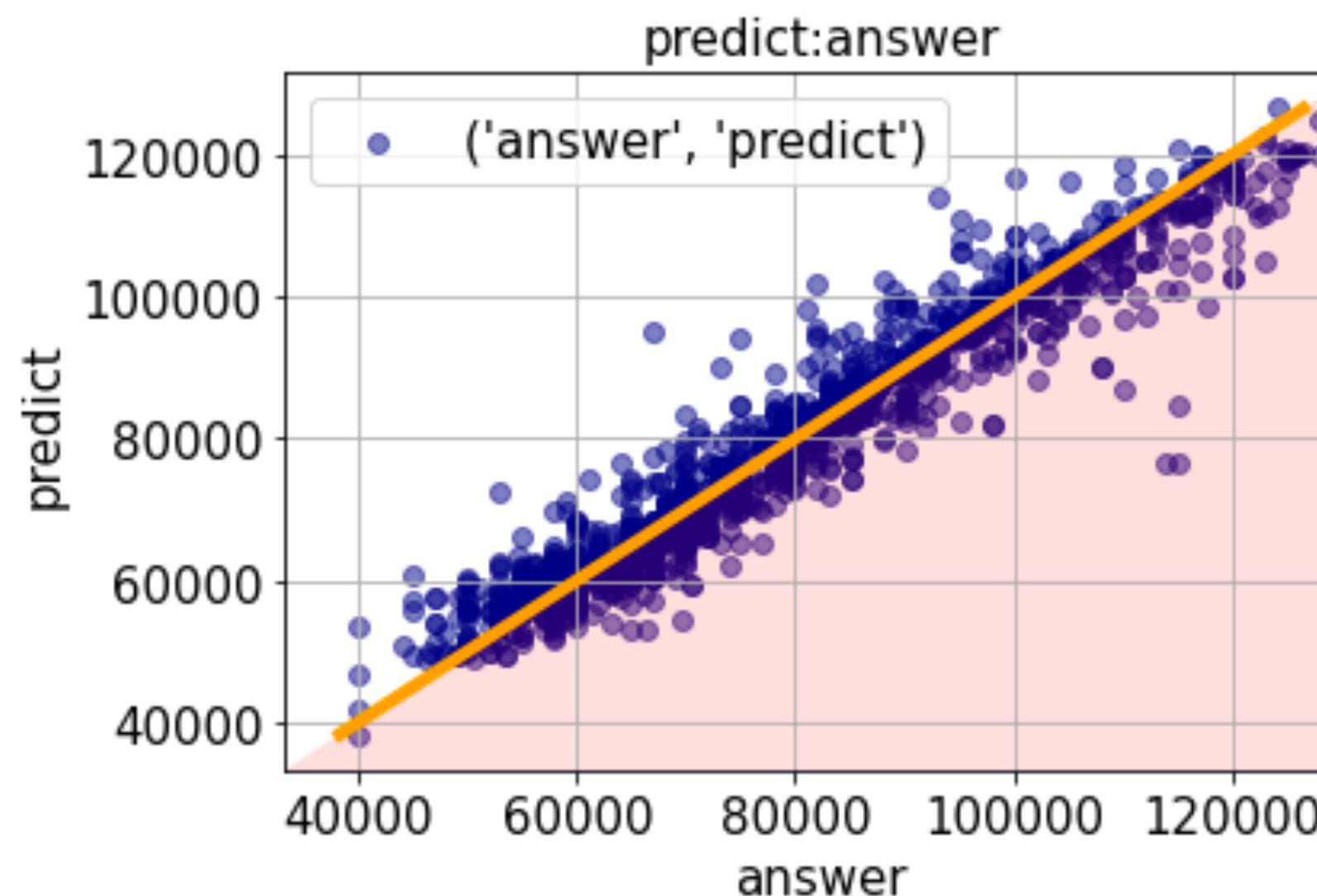
LightGBM(Valid MAE: 3614)
オレンジ線 : predict = answer

モデルの作成結果

- うまくいきませんでした、
 - ・ 全データに対して、同じような予測値を出力してしまう（学習ができていない）

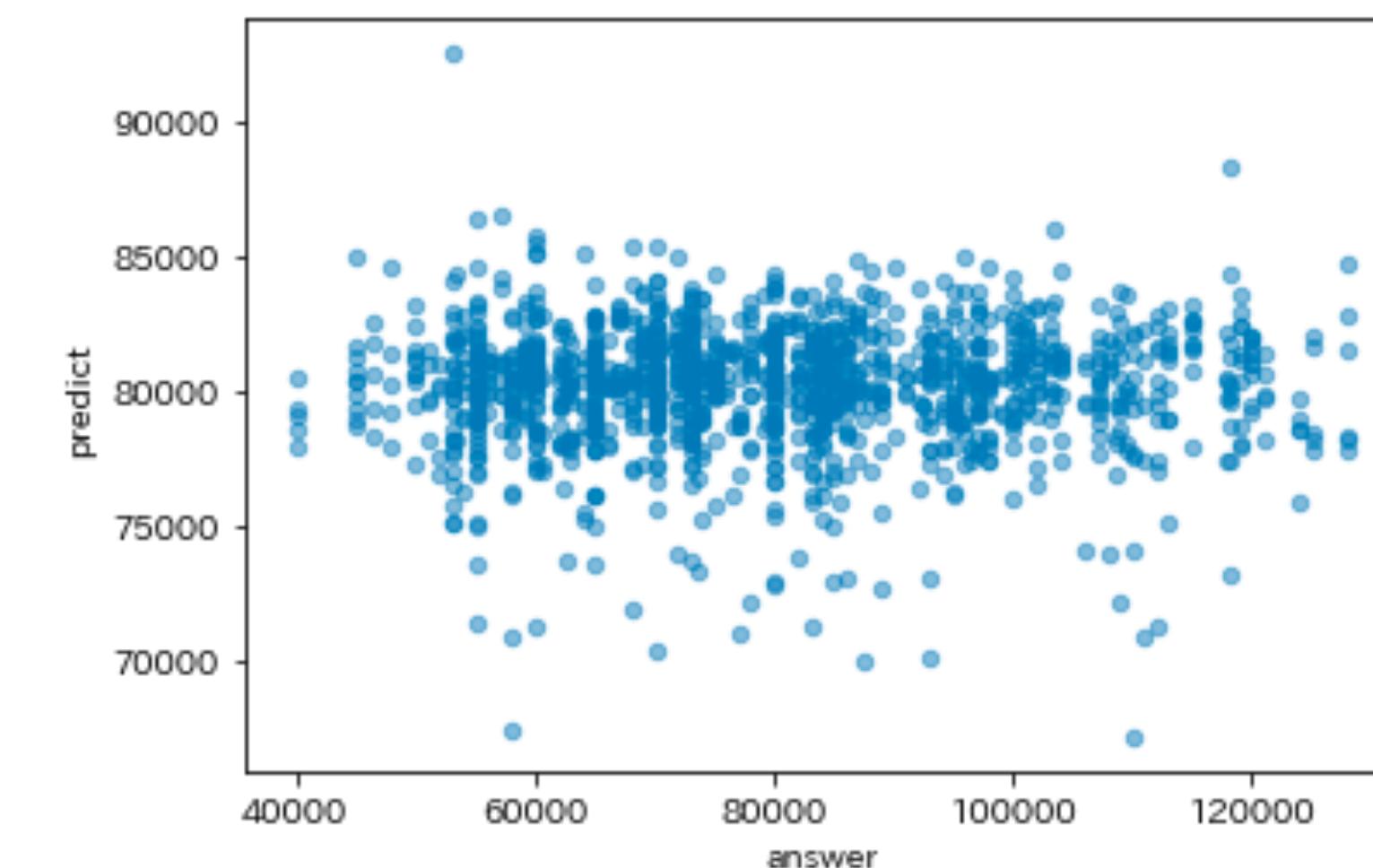


理想 オレンジの線より上にあるデータが割安物件



LightGBM(Valid MAE: 3614)
オレンジ線 : $\text{predict} = \text{answer}$

現実 全データに同じような予測をしてしまっている



自然言語処理(Valid MAE: 17417)

モデルの出力

```
input text: バストイレ別バルコニーエアコンフローリングTVインターホンオートロック室内洗濯置シューズボックス追焚機能浴室角住戸温水洗净便座エレベーター洗面所独立駐輪場宅配ボックス押入外壁タイル張り即入居可防犯ナ  
answer label: 100000.0  
predicted label: tensor([78972.0312], device='cuda:0', grad_fn=<SelectBackward>)

input text: バストイレ別バルコニーエアコンガスコンロ対応クロゼットフローリングTVインターホン浴室乾燥機オートロック室内洗濯置シューズボックスシステムキッチン温水洗净便座脱衣所エレベーター洗面所独立洗面化粧台2  
answer label: 108500.0  
predicted label: tensor([82894.6953], device='cuda:0', grad_fn=<SelectBackward>)

input text: バストイレ別エアコンガスコンロ対応フローリング室内洗濯置陽当り良好角住戸即入居可閑静な住宅地2面採光最上階出窓全居室収納全居室フローリング2沿線利用可ディンプルキー駅まで平坦ダブルロックキー1フロア  
answer label: 70000.0  
predicted label: tensor([81419.7734], device='cuda:0', grad_fn=<SelectBackward>)

input text: バストイレ別バルコニーエアコンクロゼットフローリングシャワー付洗面台TVインターホンオートロック室内洗濯置システムキッチン追焚機能浴室温水洗净便座脱衣所洗面所独立洗面化粧台駐輪場CATV礼金不要敷金不要  
answer label: 79500.0  
predicted label: tensor([79218.7188], device='cuda:0', grad_fn=<SelectBackward>)

input text: バストイレ別バルコニーエアコンクロゼットフローリングシャワー付洗面台TVインターホン浴室乾燥機オートロック室内洗濯置陽当り良好シューズボックスシステムキッチン追焚機能浴室温水洗净便座脱衣所エレベータ  
answer label: 112000.0  
predicted label: tensor([77325.5391], device='cuda:0', grad_fn=<SelectBackward>)

input text: バストイレ別バルコニーエアコンガスコンロ対応シャワー付洗面台室内洗濯置陽当り良好シューズボックス角住戸温水洗净便座脱衣所洗面所独立洗面化粧台2口コンロ駐輪場押入CATV礼金不要閑静な住宅地最上階照明付  
answer label: 59000.0  
predicted label: tensor([83236.8906], device='cuda:0', grad_fn=<SelectBackward>)

input text: バストイレ別バルコニーエアコンクロゼットフローリング室内洗濯置陽当り良好シューズボックス駐輪場光ファイバー即入居可礼金不要閑静な住宅地IHクッキングヒーター敷金1ヶ月2沿線利用可駅まで平坦平坦地3駅以  
answer label: 60000.0  
predicted label: tensor([82847.4219], device='cuda:0', grad_fn=<SelectBackward>)

input text: バストイレ別バルコニーエアコンクロゼットフローリングTVインターホン浴室乾燥機オートロック室内洗濯置陽当り良好シューズボックスシステムキッチン角住戸脱衣所エレベーター洗面所独立洗面化粧台2口コンロ宅  
answer label: 83000.0  
predicted label: tensor([80801.4922], device='cuda:0', grad_fn=<SelectBackward>)

input text: バストイレ別バルコニーエアコンガスコンロ対応クロゼットフローリングシャワー付洗面台TVインターホン浴室乾燥機オートロック室内洗濯置陽当り良好シューズボックスシステムキッチン追焚機能浴室角住戸温水洗净  
answer label: 93000.0  
predicted label: tensor([80810.5625], device='cuda:0', grad_fn=<SelectBackward>)

input text: バストイレ別バルコニーエアコンガスコンロ対応クロゼットフローリング陽当り良好シューズボックス追焚機能浴室2口コンロ駐輪場閑静な住宅地全居室フローリング2沿線利用可駅まで平坦キッチンに窓雨戸平坦地都  
answer label: 55000.0  
predicted label: tensor([79019.2578], device='cuda:0', grad_fn=<SelectBackward>)
```

モデルの改善 1

• 考えられる原因

- ▶ ネットワークのパラメータが調整できていない
 - ネットワークやバッチサイズなどは試行錯誤済み
- ▶ 部屋の説明文だけでは情報が足りない
 - **説明文に単語を追加する**
 - ▶ 「間取り」「面積」「築年月」など
 - ▶ (先程のLightGBMは追加してモデル作ってしまってました)

部屋の特徴・設備

バストイレ別、バルコニー、エアコン、ガスコンロ対応、クロゼット、フローリング、TVインターホン、浴室乾燥機、オートロック、室内洗濯置、シューズボックス、システムキッチン、角住戸、温水洗浄便座、脱衣所、エレベーター、洗面所独立、2口コンロ、駐輪場、宅配ボックス、外壁タイル張り、2面採光、防犯カメラ、ペット相談、照明付、分譲賃貸、グリル付、保証人不要、敷金1ヶ月、24時間緊急通報システム、2沿線利用可、ディンプルキー、CS、ネット専用回線、ネット使用料不要、ダブルロックキー、24時間換気システム、人感照明センサー、耐火構造、2駅利用可、3駅以上利用可、3沿線以上利用可、駅徒歩10分以内、24時間ゴミ出し可、敷地内ごみ置き場、セキュリティ会社加入済、都市ガス、洗面所にドア、BS、礼金1ヶ月、保証会社利用可、初期費用カード決済可

物件概要

 情報の見方

間取り詳細	洋6.6	構造	鉄筋コン
階建	3階/7階建	築年月	2015年11月
損保	2万円2年	駐車場	-
入居	相談	取引態様	仲介
条件	ペット相談	取り扱い店舗 物件コード	doki 215494
SUUMO 物件コード	100258898297	総戸数	-
情報更新日	2021/11/12	次回更新日	次回更新日は情報更新日より8日以内
契約期間	普通借家 2年		
保証会社	保証会社利用必 指定保証会社 初回保証委託料：月額総賃料の50%、1年後以降：年間保証料10,000円		
ほか初期費用	合計4.4万円（内訳：24時間サポート1.65万円 鍵交換代2.75万円）		
ほか諸費用	更新料 新賃料1.50ヶ月分 口座振替手数料 330円		
備考	巡回管理		

テキストに単語を追加する

others:

バストイレ別、バルコニー、エアコン、クロゼット、フローリング、TVインターホン、浴室乾燥機、オートロック、室内洗濯置、シューズボックス、システムキッチン、温水洗净便座、脱衣所、エレベーター、洗面所独立、2口コンロ、駐輪場、宅配ボックス、押入、即入居可、礼金不要、敷金不要、防犯カメラ、ペット相談、分譲賃貸、保証人不要、バイク置場、2沿線利用可、CS、ネット使用料不要、築2年以内、築3年以内、一部フローリング、ペット専用設備、トイレ未使用、2駅利用可、駅徒歩5分以内、駅徒歩10分以内、24時間ゴミ出し可、敷地内ごみ置き場、築5年以内、都市ガス、BS、敷金・礼金不要、保証会社利用可、IT重説 対応物件、初期費用カード決済可

others add features:

バストイレ別、バルコニー、エアコン、クロゼット、フローリング、TVインターホン、浴室乾燥機、オートロック、室内洗濯置、シューズボックス、システムキッチン、温水洗净便座、脱衣所、エレベーター、洗面所独立、2口コンロ、駐輪場、宅配ボックス、押入、即入居可、礼金不要、敷金不要、防犯カメラ、ペット相談、分譲賃貸、保証人不要、バイク置場、2沿線利用可、CS、ネット使用料不要、築2年以内、築3年以内、一部フローリング、ペット専用設備、トイレ未使用、2駅利用可、駅徒歩5分以内、駅徒歩10分以内、24時間ゴミ出し可、敷地内ごみ置き場、築5年以内、都市ガス、BS、敷金・礼金不要、保証会社利用可、IT重説 対応物件、初期費用カード決済可、1K、K6.9、鉄筋コン、7階、8階建、25m、東急目黒線、新丸子駅、歩4分、新丸子東1、2021年

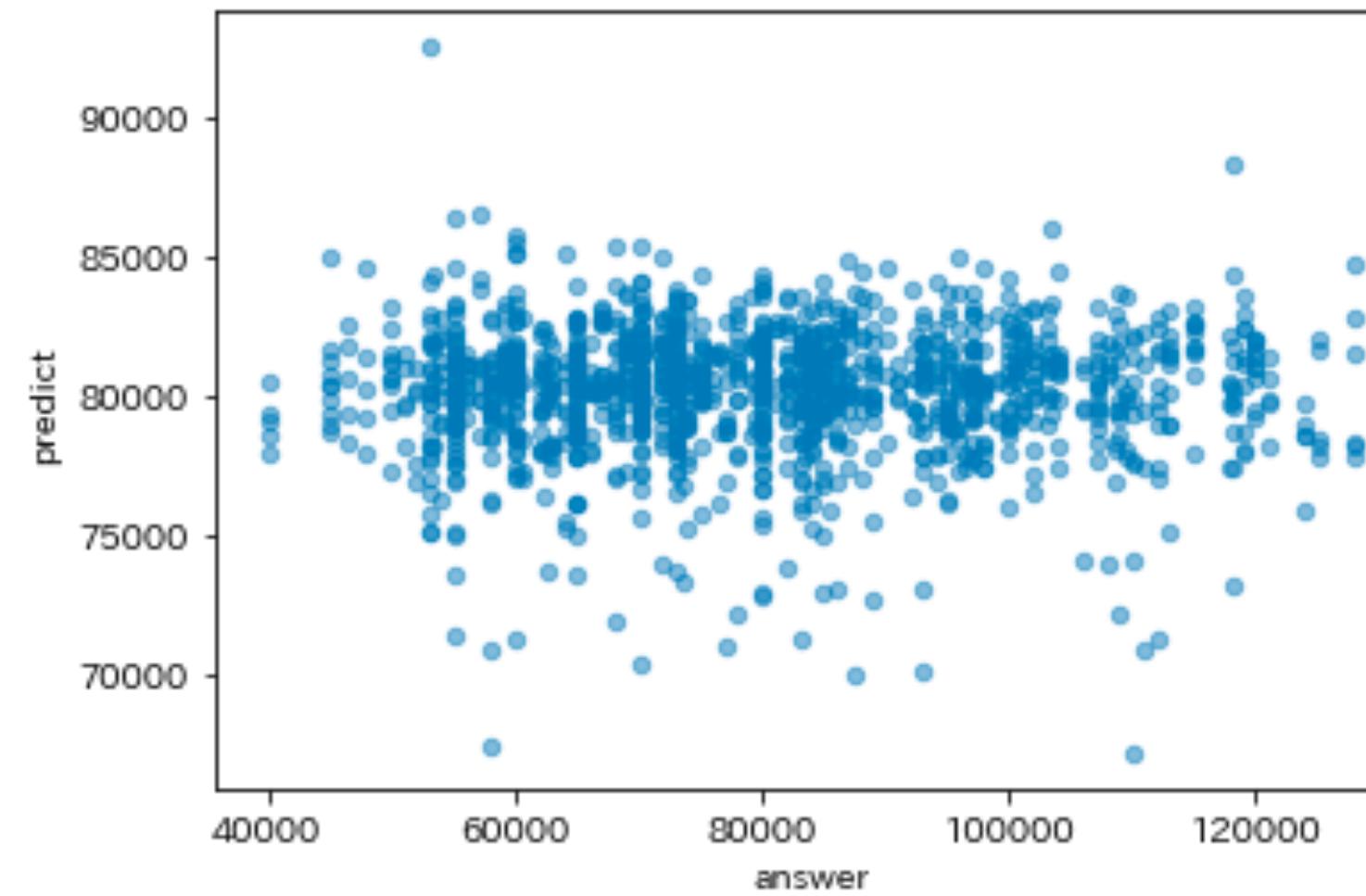
間取り、構造、階数、階建、面積、最寄り路線、最寄駅、駅徒歩、町名、築年数

重要そうなデータが抜けていたので、影響がありそう

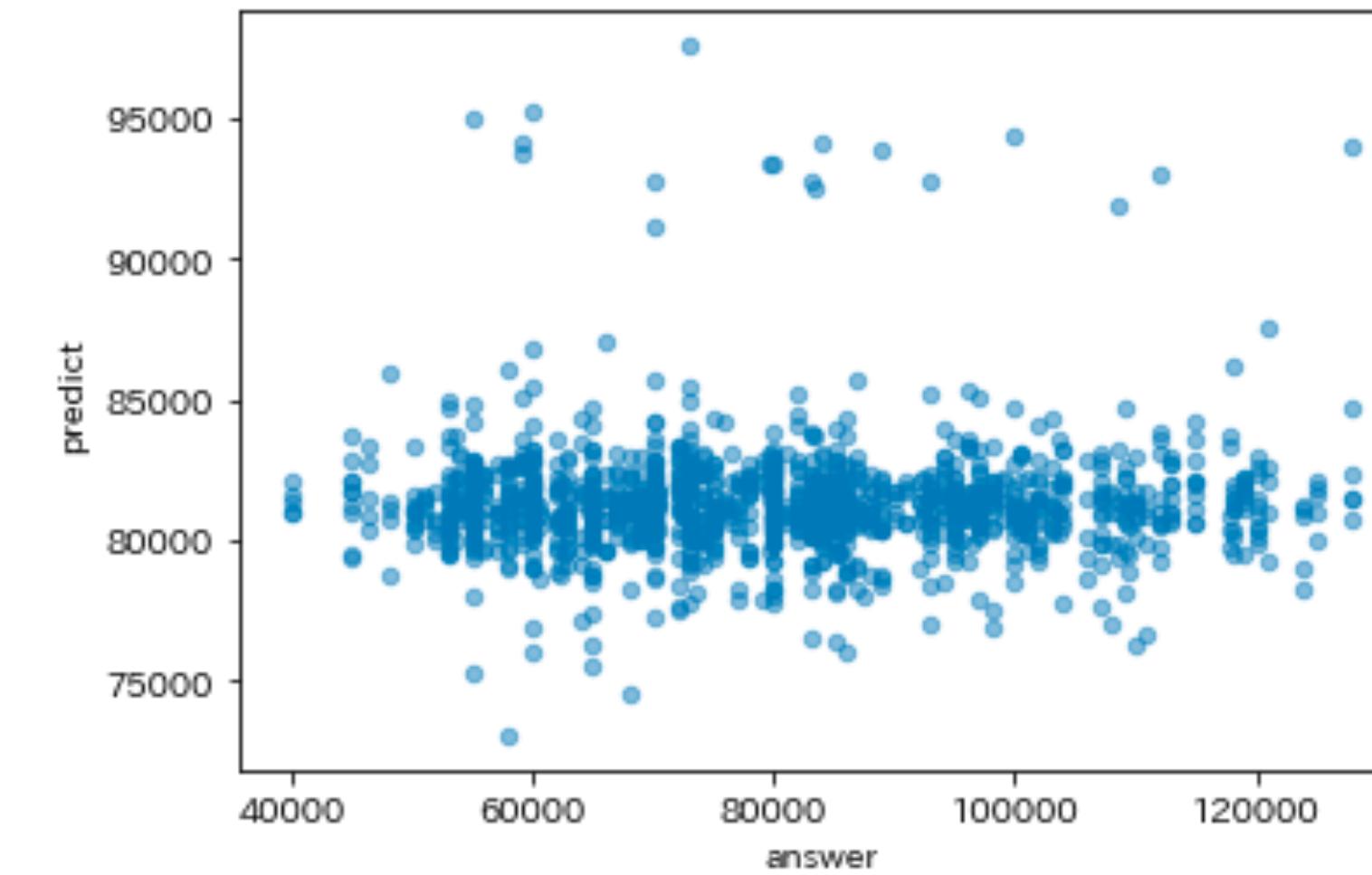
(別の場所にこのデータが入っていたので、結合忘れてNLPモデル作ってしまってました、、)

モデルの再作成

- データを変更して再度モデル作成
 - 結果は殆ど変わらず、全データに対して同じような予測値を出力してしまう
 - 数値情報(面積など)は、自然言語処理だと難しい可能性がある → 数値情報はMLでモデル作成する



Before (Valid MAE: 17417)



After (Valid MAE: 16342)

モデルの改善 2

1. MLモデルでベースモデルを作成して、その予測値と実際の家賃の誤差を求める

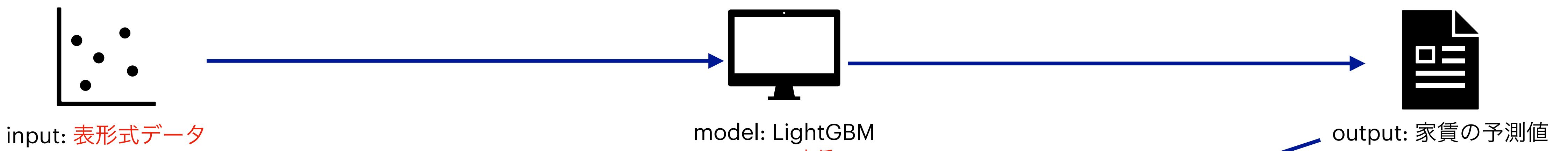
- ・ ホールドアウト：train, test (targetの値はスコア計算時のみ参照し、学習では一切使用しない)
- ・ trainのみ使って交差検証でモデル作成 → スタッキングの要領で予測値を新しい特徴量として埋める
- ・ 予測値と実際の値の誤差を求める
- ・ testでスコアを確認しておく

2. 「予測値と実際値の誤差」をターゲット、MLで使用していない「家の説明文データ」から自然言語処理モデルを作成する

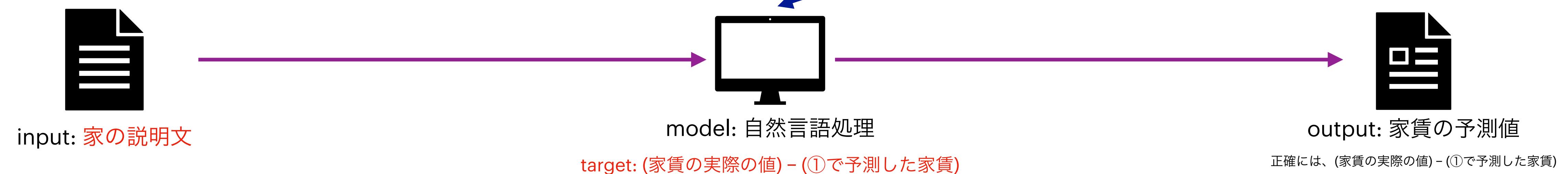
- ・ MLでは見ていない情報を見るので、誤差を補正してくれるのではないか
- ・ testでスコア検証する → MLモデルの結果よりスコアが良ければ意味があると考えられる

モデルの改善 2

1



2



モデルの改善 2

- こちらもうまいきませんでした → 時間切れ

- MAE: 3614 → 3616

```
input text: パストイレ別バルコニーエアコンガスコンロ対応クロゼットフローリングTVインターホン浴室乾燥機オートロック室内洗濯置シューズボックスシステムキッチン温水洗净便座脱衣所エレベーター洗面所独立洗面化粧台2
answer label: -1851.5977783203125
predicted label: tensor([-70.2231], device='cuda:0', grad_fn=<SelectBackward>)

input text: パストイレ別エアコンガスコンロ対応フローリング室内洗濯置陽当り良好角住戸即入居可閑静な住宅地2面採光最上階出窓全居室収納全居室フローリング2沿線利用可ディンプルキー駅まで平坦ダブルロックキー1フロア
answer label: 4997.5751953125
predicted label: tensor([-79.1154], device='cuda:0', grad_fn=<SelectBackward>)

input text: パストイレ別バルコニーエアコンクロゼットフローリングシャワー付洗面台TVインターホンオートロック室内洗濯置システムキッチン追焚機能浴室温水洗净便座脱衣所洗面所独立洗面化粧台駐輪場CATV礼金不要敷金不要
answer label: 2071.78857421875
predicted label: tensor([-47.0575], device='cuda:0', grad_fn=<SelectBackward>)

input text: パストイレ別バルコニーエアコンクロゼットフローリングシャワー付洗面台TVインターホン浴室乾燥機オートロック室内洗濯置陽当り良好シューズボックスシステムキッチン追焚機能浴室温水洗净便座脱衣所エレベーター
answer label: -1178.7869873046875
predicted label: tensor([-63.2259], device='cuda:0', grad_fn=<SelectBackward>)

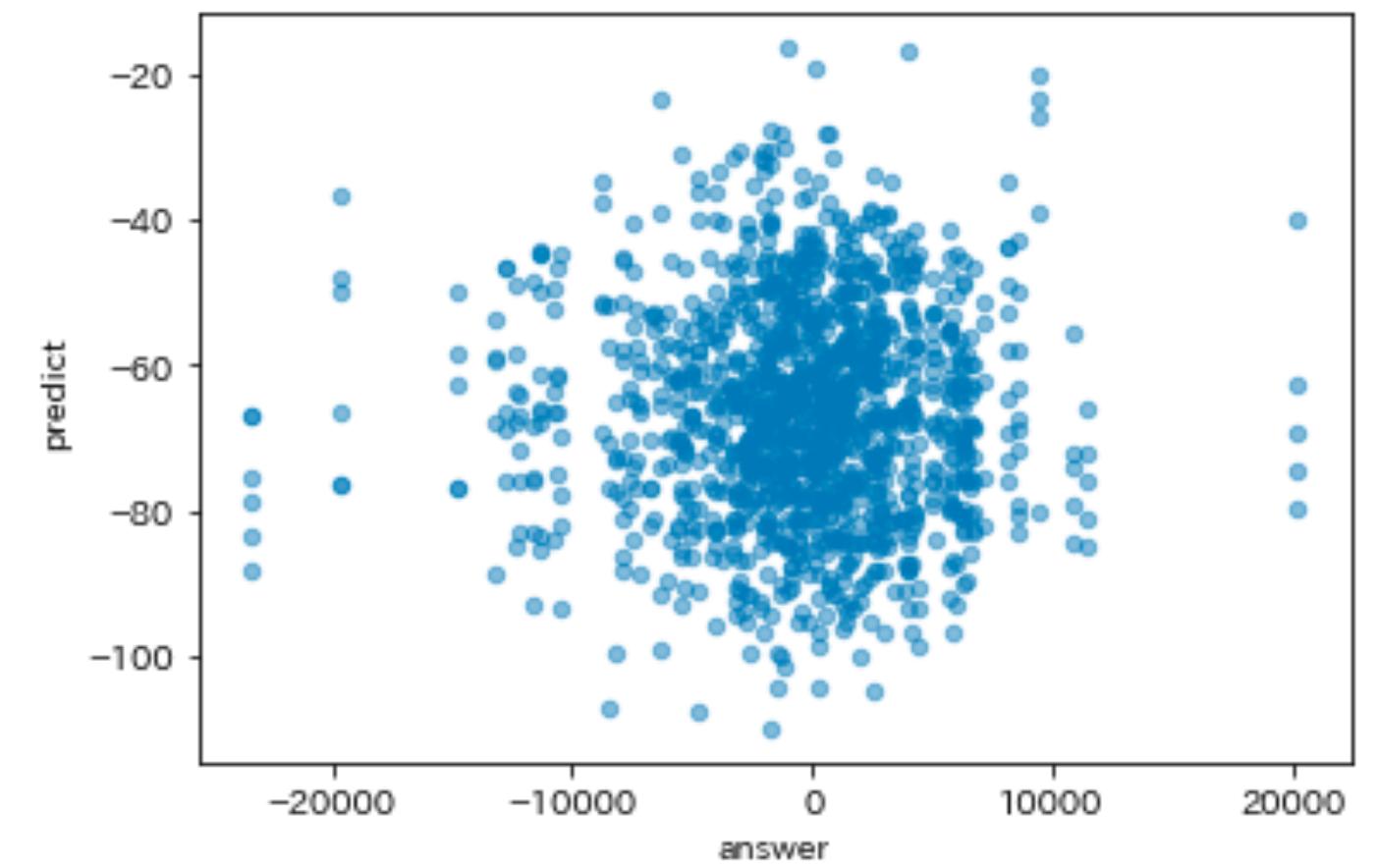
input text: パストイレ別バルコニーエアコンガスコンロ対応シャワー付洗面台室内洗濯置陽当り良好シューズボックス角住戸温水洗净便座脱衣所洗面所独立洗面化粧台2口コンロ駐輪場押入CATV礼金不要閑静な住宅地最上階照明付
answer label: -11705.1103515625
predicted label: tensor([-82.7508], device='cuda:0', grad_fn=<SelectBackward>)

input text: パストイレ別バルコニーエアコンクロゼットフローリング室内洗濯置陽当り良好シューズボックス駐輪場光ファイバー即入居可礼金不要閑静な住宅地IHクッキングヒーター敷金1ヶ月2沿線利用可駅まで平坦平坦地3駅以
answer label: -2101.796630859375
predicted label: tensor([-81.8765], device='cuda:0', grad_fn=<SelectBackward>)

input text: パストイレ別バルコニーエアコンクロゼットフローリングTVインターホン浴室乾燥機オートロック室内洗濯置陽当り良好シューズボックスシステムキッチン角住戸脱衣所エレベーター洗面所独立洗面化粧台2口コンロ宅
answer label: -3095.697998046875
predicted label: tensor([-89.7619], device='cuda:0', grad_fn=<SelectBackward>)

input text: パストイレ別バルコニーエアコンガスコンロ対応クロゼットフローリングシャワー付洗面台TVインターホン浴室乾燥機オートロック室内洗濯置陽当り良好シューズボックスシステムキッチン追焚機能浴室角住戸温水洗净便
answer label: -23375.3125
predicted label: tensor([-83.3283], device='cuda:0', grad_fn=<SelectBackward>)

input text: パストイレ別バルコニーエアコンガスコンロ対応クロゼットフローリング陽当り良好シューズボックス追焚機能浴室2口コンロ駐輪場閑静な住宅地全居室フローリング2沿線利用可駅まで平坦キッチンに窓雨戸平坦地都
answer label: -5362.99609375
predicted label: tensor([-56.9642], device='cuda:0', grad_fn=<SelectBackward>)
```



考察

- 自然言語処理でうまくモデルが作成できない原因の考察
 - ▶ 数値情報を連続値として扱えない
 - 面積を例に挙げると、「25m」「26m」はそれぞれ別の単語としてエンベディングされてしまう
 - ▶ BoWとして加算する部分で必要な情報が消えてしまう
 - 加算では表せないような情報が存在した可能性がある
 - ▶ 実装部分でのミス
 - 私自身の経験値が低いため、どこかしらで実装上のミスがあった可能性がある
 - ▶ ここまでうまくいかないとは思っていなかった...
 - ▶ ソースコード一式：https://github.com/Sekikawa318/house_price