Kaggleでの取り組み

I 類　メディア情報学プログラム
松浦 史明

データサイエンス演習 2024
July 20, 2024

アウトライン

- 導入
    - データ分析フロー
- 方法
    - データの理解
    - 特徴量の追加、削除
    - Kfold法
- 結果
    - スコア
    - 試した手法ごとの、スコアの変化
- 議論
- まとめ

アウトライン

- 導入
    - データ分析フロー
- 方法
    - データの理解
    - 特徴量の追加、削除
    - Kfold法
- 結果
    - スコア
    - 試した手法ごとの、スコアの変化
- 議論
- まとめ

導入

【データ分析フロー】
・主に取り組んだ点

データ理解	データ加工	モデル
・**データの外観** ・**各種統計量** ・**可視化**	・欠損値処理 ・外れ値処理 ・**特徴量作成**	・**モデルの作成** ・モデル評価 ・チューニング

第11回講義資料より

アウトライン

- 導入
    - データ分析フロー
- 方法
    - データの理解
    - 特徴量の追加、削除
    - Kfold法
- 結果
    - スコア
    - 試した手法ごとの、スコアの変化
- 議論
- まとめ

方法（1/3）：データの理解

データの理解:

・EDA（探索的データ解析）の実施

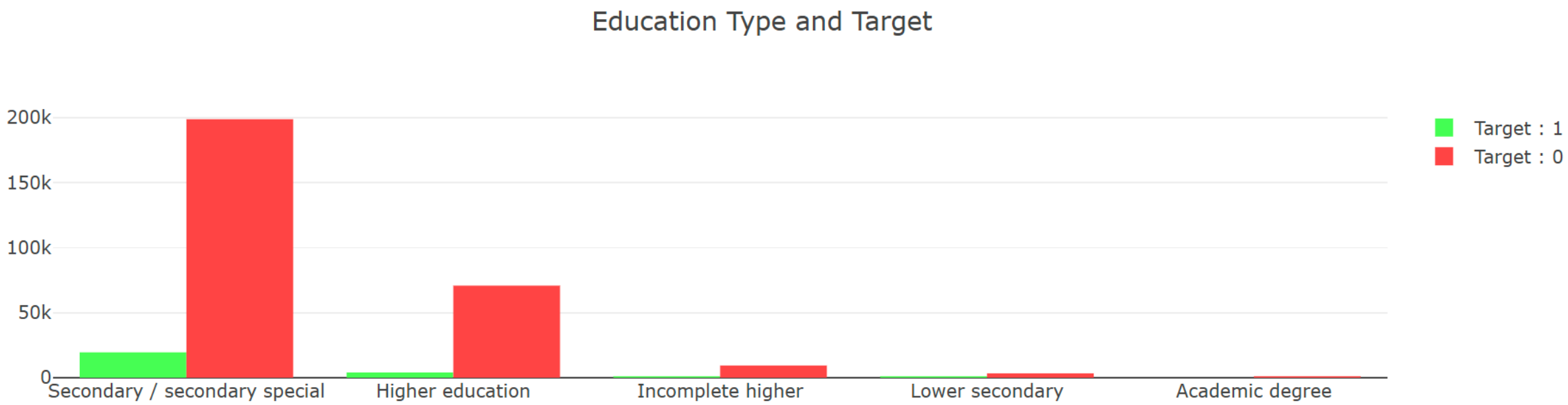
・Home Credit : Complete EDA + Feature Importance ?? (kaggle.com)を日本語に翻訳したHome Credit : Complete EDA(日本語訳) (kaggle.com)を参考

・最初の数行をcsvで出力し、どのようなカラムがあるのか眺めてみる

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	SK_ID_CU	TARGET	CODE_GE	FLAG_OW	FLAG_OW	CNT_CHIL	AMT_INC	AMT_CRE	AMT_ANN	AMT_GOO	REGION_F	DAYS_BIR	DAYS_EM	DAYS_RE	DAYS_ID_	OWN_CAR	FLAG_PH
2	100002	1	0	0	0	0	202500	406597.5	24700.5	351000	0.018801	-9461	-637	-3648	-2120		1
3	100003	0	1	0	1	0	270000	1293503	35698.5	1129500	0.003541	-16765	-1188	-1186	-291		1
4	100004	0	0	1	0	0	67500	135000	6750	135000	0.010032	-19046	-225	-4260	-2531	26	1
5	100006	0	1	0	0	0	135000	312682.5	29686.5	297000	0.008019	-19005	-3039	-9833	-2437		0
6	100007	0	0	0	0	0	121500	513000	21865.5	513000	0.028663	-19932	-3038	-4311	-3458		0

方法（1/3）：データの理解

・例：教育のタイプと、目的変数との関係



方法（2/3）：特徴量の追加・削除

【特徴量の追加と削除】

1. 第13回講義のipynbファイルを参考
2. Kaggler上位が追加、削除している特徴量を選択
3. importance値による特徴量の選択
4. 相関の高い変数の削除

方法（2/3）：特徴量の追加・削除

1, 2. 資料を参考にした、特徴量の追加・削除

・金利やローンの事については、正直初心者

・意味を持つ特徴量を作るために、資料を参照

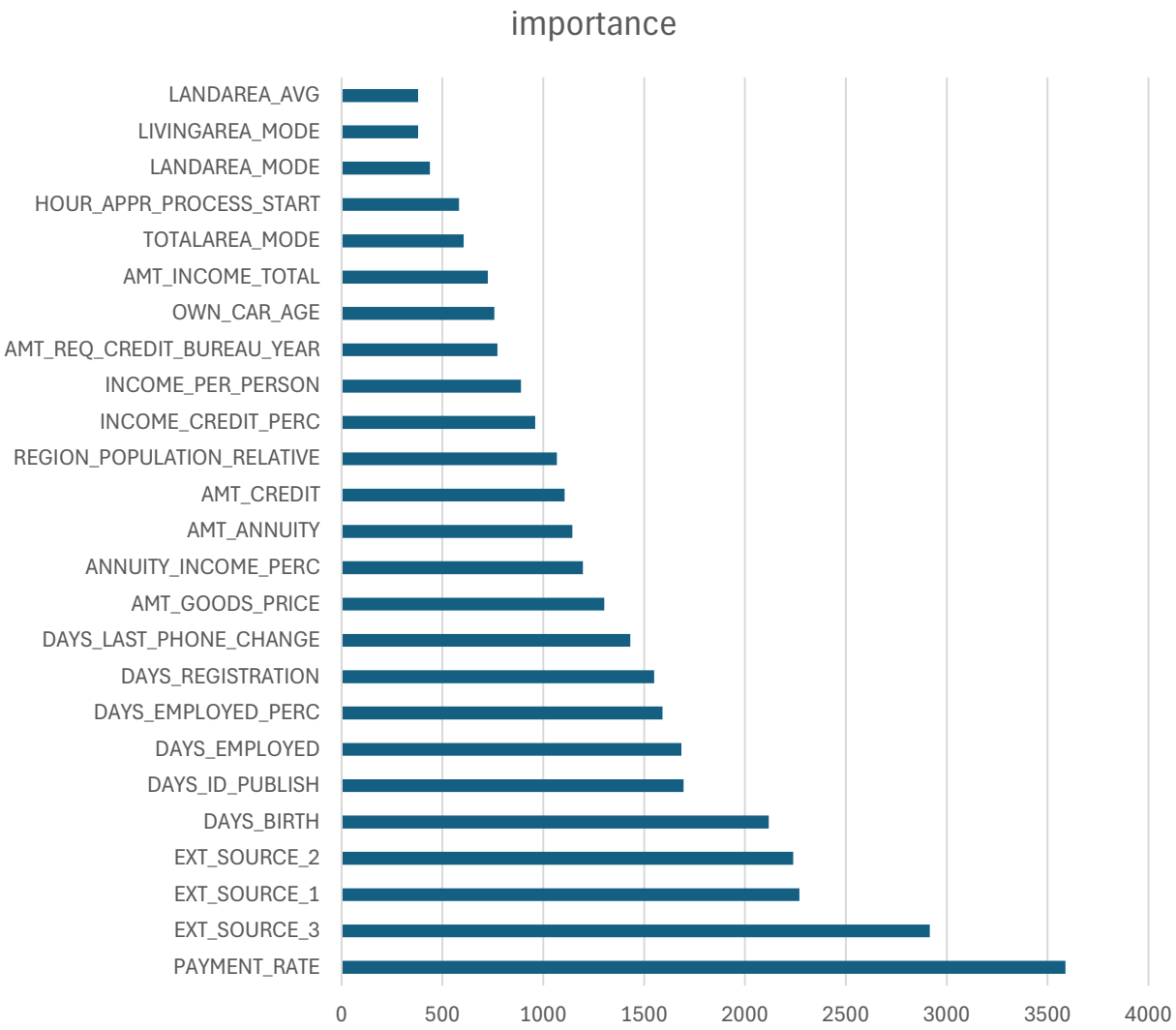
・基本的に、ある程度までは特徴量を削除した方がスコアが上昇した

ex)偏りが大きく（ほとんど 0 または 1）、情報量が少ないと考えられるような特徴量等

方法（2/3）：特徴量の追加・削除

3. Importance値による特徴量の選択：

・importance値と呼ばれる、特徴量の寄与度を示すグラフ



方法（2/3）：特徴量の追加・削除

4. 相関の高い変数の削除：

・一般的に、相関の高い変数はどちらか片方を消すと良い（とされている）

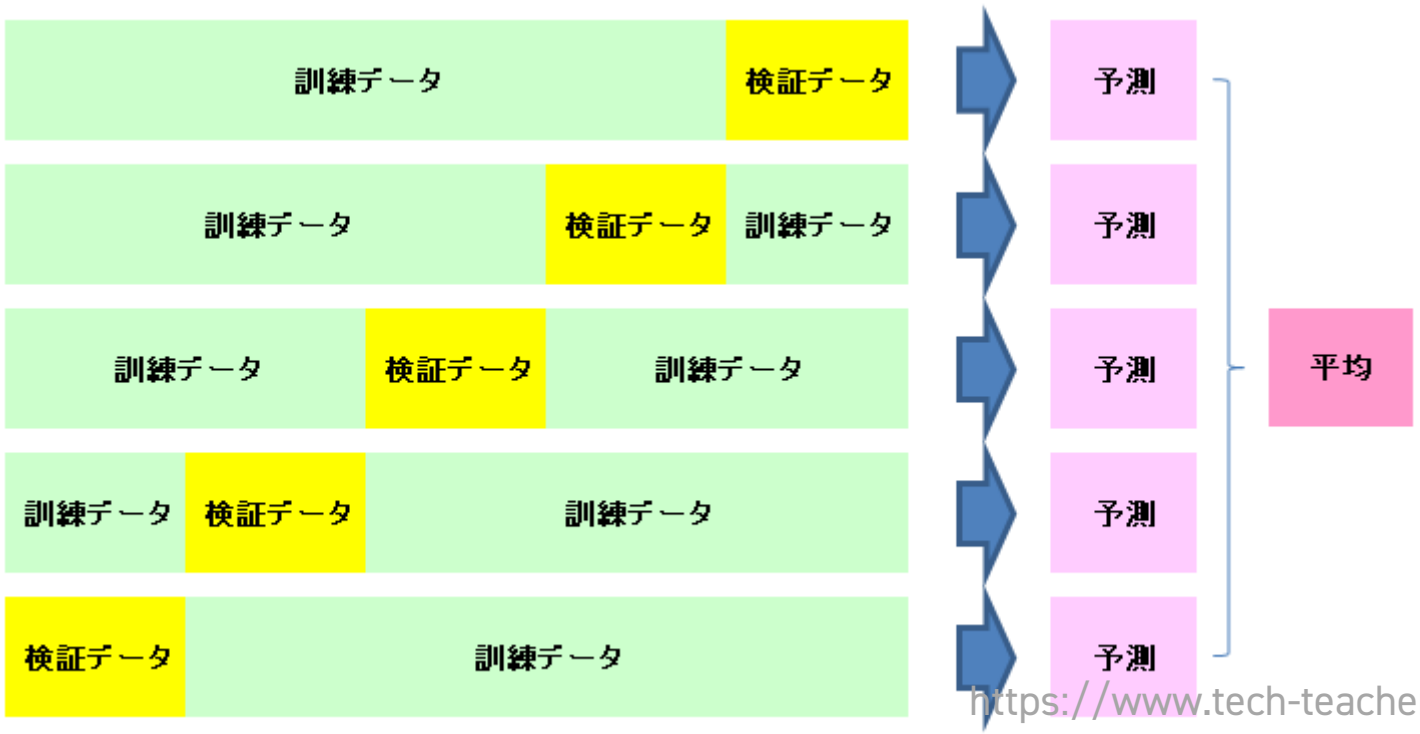
・相関行列を作成し、指定した閾値以上の相関を持つ特徴量を削除

・うまくいくはずだった...

方法（3/3）：Kfold法

【Kfold法を用いた、学習】

・訓練データと検証データを入れ替えつつそれぞれのモデルで予測を行い、その平均値を予測値として使用



https://www.tech-teacher.jp/blog/cross-validation/

アウトライン

- 導入
    - データ分析フロー
- 方法
    - データの理解
    - 特徴量の追加、削除
    - Kfold法
- 結果
    - スコア
    - 試した手法ごとの、スコアの変化
- 議論
- まとめ

結果 (1/2)：スコア

【現状のスコア】

・0.79136

20240715_v11_0.792182.csv

Complete (after deadline) · 4d ago

0.79136　0.79168

・更なる向上を目指したい

結果（2/2）：試した手法ごとの、スコアの変化

・初期段階（授業資料そのまま）

submit_tree_20210829_1.csv Complete (after deadline) · 12d ago · tree_model = DecisionTreeClassifier(criterion="gini", # Entropy基準の場合は"entropy...	0.66105	0.67113	☐

・データはapplicationのみを使用し、Kfoldを使用

20240709_test.csv Complete (after deadline) · 10d ago	0.76072	0.76536	☐

・applicationとbureau and balanceを使用し、特徴量選択を実施

20240710_v1.csv Complete (after deadline) · 10d ago · add bureau and balance to dataframe	0.77535	0.77500	☐

結果 (2/2)：試した手法ごとの、スコアの変化

- csvデータを全てデータフレームに導入

20240715_v5_0.792163.csv Complete (after deadline) · 4d ago	0.79086	0.78970	☐

- 特徴量選択を繰り返した後の結果

20240715_v11_0.792182.csv Complete (after deadline) · 4d ago	0.79136	0.79168	☐

アウトライン

- 導入
    - データ分析フロー
- 方法
    - データの理解
    - 特徴量の追加、削除
    - Kfold法
- 結果
    - スコア
    - 試した手法ごとの、スコアの変化
- **議論**
- まとめ

議論

【うまくいった点】

・特徴量の追加による、スコア向上

・特徴量の削除（変数名決め打ち）でのスコア向上

【改善点】

・今のところ、特徴量の統合や相関によるフィルタリングでは、スコアを改善させることができなかった

・importance値による特徴量の選別を深めたい

・Embedded Methodの変数選択を利用した特徴量選別も視野

アウトライン

- 導入
    - データ分析フロー
- 方法
    - データの理解
    - 特徴量の追加、削除
    - Kfold法
- 結果
    - スコア
    - 試した手法ごとの、スコアの変化
- 議論
- まとめ

まとめ

背景：
・データ分析フロー

方法：
・EDAによる、データの理解
・様々な資料を参考にした、特徴量の追加と削除

主な達成点：
・特徴量選択・削除による、スコアの向上

今後：
・特徴量の選択手法を検討し、より重要なものの選別

ご清聴、ありがとうございました