SIGNATECUP 2024
2nd Place Solution

tose

自己紹介

職場

三島信用金庫

主な業務

データの抽出

略歴

高専で数値解析を学ぶ

⇒大学で機械学習を学ぶ

⇒高校で数学を教える

⇒転職して三島信用金庫へ

⇒営業店からシステム部門へ

⇒データ分析を業者へ依頼するため、データの切り出しを担当

⇒もともと好きなので、自分でもやってみるか

コンペ概要

旅行会社の保有する顧客データ（属性や志向、営業担当との接触履歴等）を元に、旅行パッケージの成約率を予測するモデルを構築する

train.csv	test.csv
3,489件	3,489件

特徴量（17個）	
id(顧客ID)	ProductPitched(営業担当者のセールスした商品の種類)
Age(顧客の年齢)	PreferredPropertyStar(顧客の希望するホテルのランク)
TypeofContact(顧客への連絡・接触方法)	NumberOfTrips(顧客の年間旅行数)
CityTier(都市層(1>2>3))	Passport(パスポートの所持(0: 不所持、1: 所持))
DurationOfPitch(営業担当者による顧客への商品のセールス時間)	PitchSatisfactionScore(営業担当者のセールストークに対する顧客の満足度)
Occupation(顧客の職業)	Designation(顧客の役職)
Gender(顧客の性別)	MonthlyIncome(顧客の月収)
NumberOfPersonVisiting(予定している旅行の同行者の数)	customer_info(顧客の情報のメモ(婚姻状況や車の有無、旅行への子どもの同伴の有無))
NumberOfFollowups(セールス後に営業担当者が行ったフォローアップの回数)	

目標変数

ProdTaken(商品の契約状態(0:不成約、1:成約))

train.ｃｓｖのProdTakenの内訳

0:497件　1:2,992件

大体　1:6　の不均衡

精度評価

AUC

仕事でやれといわれている内容と近い！！けど・・・

コンペの内容を業務で考えてみる

おそらく運ゲーになるだろうと早々に精度(AUC)を一定以上あげることは辞める

いっそ仕事をこのコンペで置き換えて考えてみると・・・

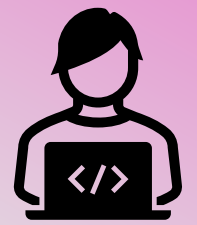
AIを使って購入しそうな顧客を予測し、
マーケティング戦略に活用しろ！

経営

数字がとりたいから、見込みが高い先を、高い順にリストにまとめて

マーケティング部門の営業

AUCが高いモデルを作ってリストを出せば
本当にいい？

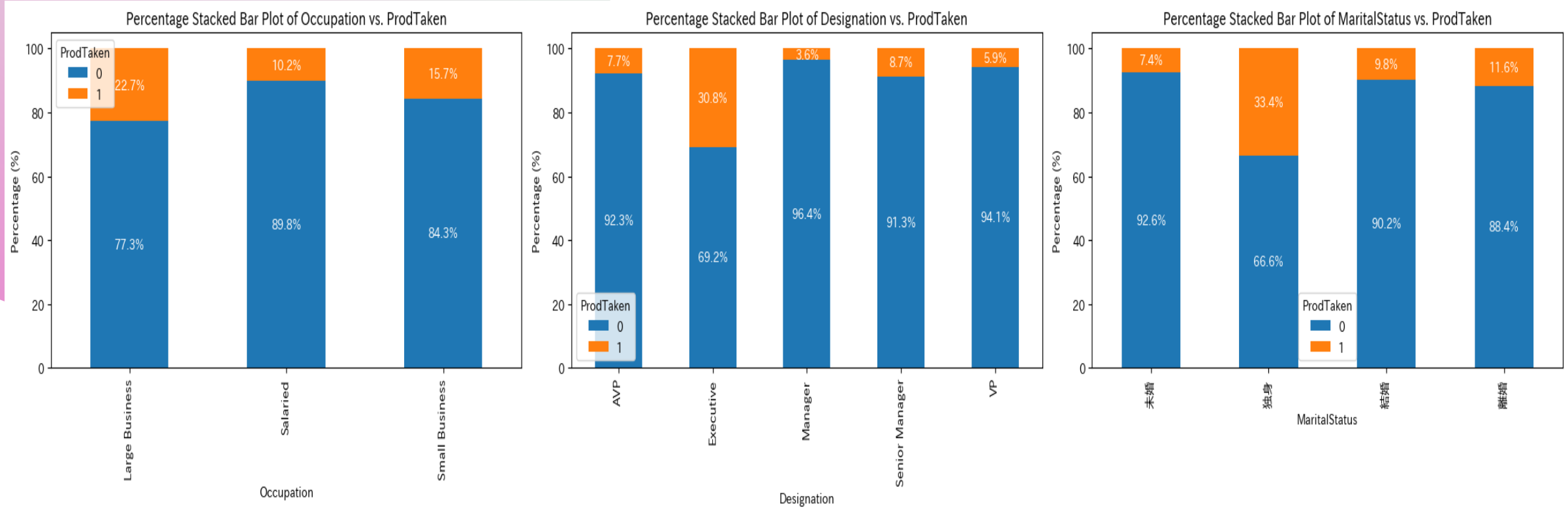


前処理

特徴量	処理内容
Age	漢数字の置き換え 〇代→各世代の最頻値
TypeofContact	Company Invitedは0, Self Enquiryは１
DurationOfPitch	秒から分へ置き換え
Gender	男性0,女性1
ProductPitched	表記の揺れを修正
Designation	表記の揺れを修正
NumberOfTrips	年単位の回数に変換（例：半年に1回→2)
NumberOfFollowups	１００以上は100で割る
MonthlyIncome	円単位で数字
customer_info	MaritalStatus, OwnCar, NumberOfChildrenを作成
HasChildren	子供が１人でもいれば１、いなければ０

欠損値があるAge, NumberOfChildren、DurationOfPitch、NumberOfFollowups、NumberOfTrips、MonthlyIncomeについては平均で埋める

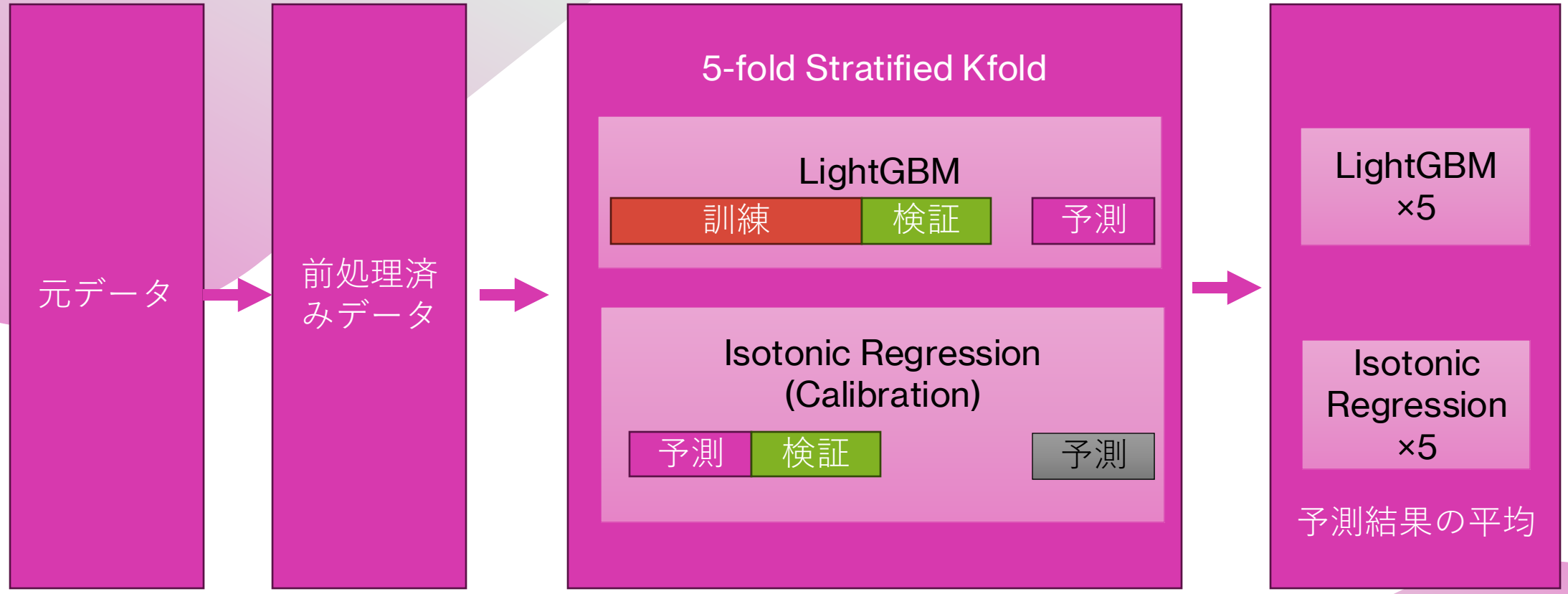
前処理



CityTier、Occupation、ProductPitched, PreferredPropertyStar,

PitchSatisfactionScore, Designation, MaritalStatusはOHEを実施

全体像



DiscordでCalibrationについての意見があったので、ChatGPTで作成しました

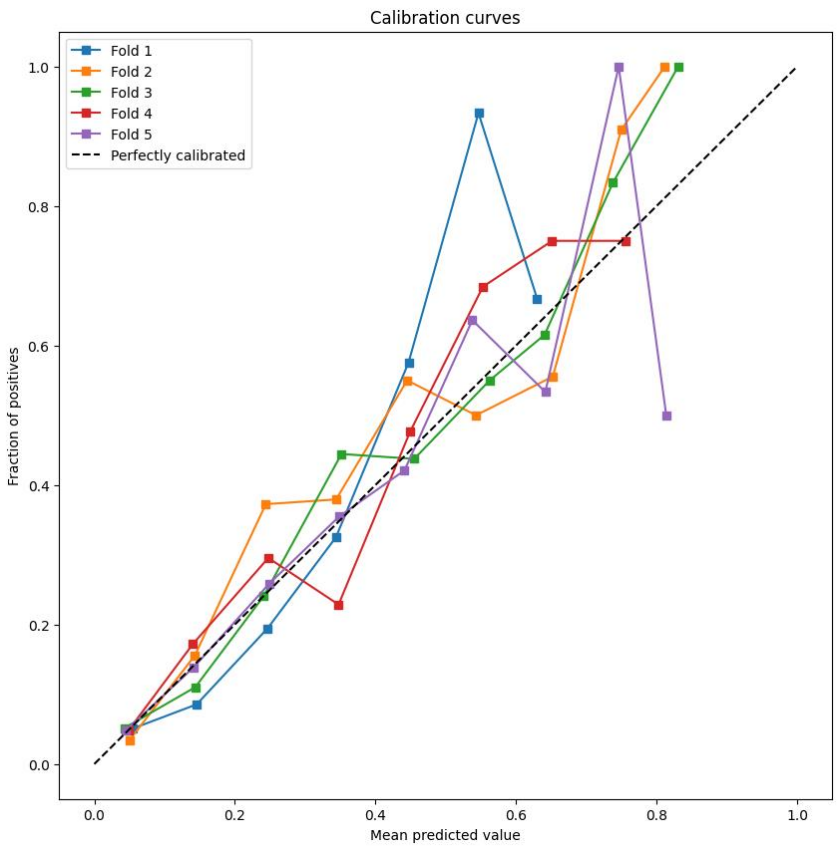
CV0.85487266（PublicLB 0.8431735 / PrivateLB 0.8466145）

lightGBMパラメータ

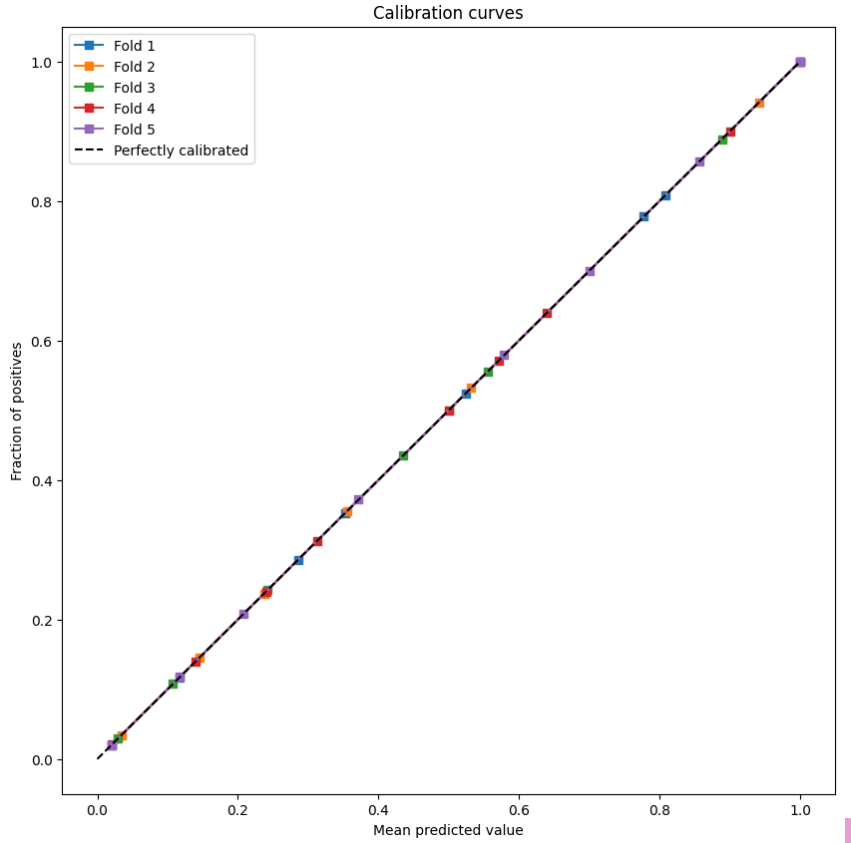
ChatGPTに問題を説明して、過学習しないでとお願いして作成

パラメータ名	設定	説明
objective	binary	問題内容より２クラスの分類
metric	auc	評価関数そのまま
num_leaves	20	複雑になりすぎない程度
feature_fraction	0.4	各ブースティングで使用される特徴量の割合
bagging_fraction	0.8	各ブースティングで使用するデータの割合
bagging_freq	3	バギングの頻度
lambda_l1	0.80	モデルの簡略化のため少し強めに
lambda_l2	90	モデルのパラメータを抑えるためかなり強めに
n_estimators	3000	最大学習サイクル数
learning_rate	0.01	学習率。ゆっくりと学習する設定
subsample_freq	1	サブサンプリングの頻度。1は各ブーストごとにサンプリングを行う
subsample	0.8	サブサンプルの割合。データ全体の80%を使って各ブーストを行う設定

Calibration



Isotonic Regression



出てきた値が１に近いほど数字が取れる見込みが高いとほとんどの場合いえるが、必ずしも１（数字が取れる）になる確率ではない

出てきた予測値を直に扱いたい→確率予測の信頼性をみるためCalibration Curveを確認

IsotonicRegressionで確率予測の補正を行うことで、”営業“が使いやすい数字になるのではないか

もっと速い馬が欲しい、訳ではない

		予測	
		1	0
実際	1	TP	FP
	0	FN	TN

“営業”が本当に求めているのは「精度の高いAI」ではない

まだ営業をかけてない、契約がない先で可能性が高い先はどこか？である

そして、予測値の高い先に実際に営業をかけたときにどれだけ取れたかが「営業」や「経営」にとっての使えるかどうかの”精度”の指標

まとめ

正直にいえば、精度を上げるために試行錯誤してshakedownしてしまった方には申し訳ない
おそらく早々にLBを上げることをやめたのがいい方向に働いたと思います
アンサンブルを試す前に適当に作ったモデルの１つでした
（アンサンブルしたものは手持ちで3位でした）

今回discord内でも運ゲーと言われていましたが、現実のデータ分析も運になることはあります
そういったときにはぜひ、精度だけではなくクライアントに“納得感”を与えてください
クライアントがどれだけAIを納得し、信用して”本気“で取り組んでくれたかで最終的な成果である
契約数は変わってくるでしょう

データコンペでいうことではありませんが、折角の機会ですので一クライアントとして、
実務では分析した結果をもとに働く人がいることも心に留めていただけると幸いです
データサイエンティストの皆様これからもご協力よろしくお願いします