Signate Career Up Challenge 2nd

team_challenge

自己紹介(team_challenge)

都内の受託分析企業（同じ会社）に勤務する2人でチームを組み参加しました。なかやまはマネージャー的な立場でたけむらは新卒1年目です。

チームメンバー	参加の動機など
なかやま	たけむらのマネージャー的な立場。データサイエンティストとしての実力を身に着けてもらうのに良さそうなコンペでしたのでたけむらと参加
たけむら	新卒1年目、データサイエンティストとしての実力を身に着けたい

コンペへの取り組み方

前半はたけむらが分析（モデリング）実施、なかやまがアドバイス
後半はそれぞれ別々に分析実施

サマリ

今回のコンペにおいて実施した工夫の内容です。主に4つの工夫を実施しました。モデリングの工夫が3個とモデリングの工夫に注力しました。

工夫の内容	詳細
前処理の工夫	木構造モデルが解釈しやすい形へのカテゴリ変数の数値変換、 本質的に同じデータの同質化
モデリングの工夫①	ハイパーパラメーターの異なる3モデルを構築
モデリングの工夫②	損失関数MAPE
モデリングの工夫③	目的変数(価格)の分布が特徴的であり、これへの対応として価格帯 を判定した後に、それぞれに対応する価格予測モデルを適用

基本情報ーコンペ概要

本コンペでは、中古車の販売実績のテーブルデータを用いて中古車の価格を予測し、
MAPEで予測精度の評価を行われました。最終的な提出物は価格予測のデータです。

参加コンペ	SIGNATE Career Up Challenge　モデリング部門
課題	中古車の価格予測
データ	中古車の販売実績のテーブルデータ
評価指標	予測精度（MAPE） $$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left\lvert\frac{\hat{y}_i - y_i}{y_i}\right\rvert$$ （$\hat{y}_i$：予測値、y_i：正解の値）
提出物	価格予測の結果（テーブルデータ）

基本情報ー学習データ

学習データの概要は以下の通りで、それらを用いて中古車の予測を行いました。
説明変数は14個あり、うち2個が数値変数で残り12個がカテゴリ変数です。

レコード数	27,532件
説明変数の数	14個
説明変数	1. region、地域 2. year、製造年 3. manufacturer、メーカー 4. condition、状態 5. cylinders、シリンダーの種類（数） 6. fuel、燃料の種類 7. odometer、走行距離 8. title_status、タイトル 9. transmission、トランスミッション 10. drive、駆動方式 11. size、サイズ 12. type、車の種類 13. paint_color、色 14. state、州 （製造年と走行距離のみ数値変数）

EDA、データの前処理

データの前処理はカテゴリ変数の変換、データの調整を実施しています。モデルが学習しやすくなることや本質的な解釈の誤りを減らす目的で実施しています。

内容	詳細	目的
カテゴリ変数の変換	カテゴリ変数毎に目的変数の平均値が低い順に数値をラベル付	木構造モデルが学習しやすくする
データの調整	• メーカー及びサイズ：表現、文字コードの統一 • 年：2,100年以上について1,000年マイナス • 州：地域と州のペアから地域があり州が欠損しているデータについて、州の欠損値を埋める	本質的に同じデータを別のデータと誤解させないようにする（本質的な解釈の誤りを減らす）

カテゴリ変数の変換

データの前処理はカテゴリ変数の数値への変換を実施しました。目的変数の平均値が低い順に数値をラベル付しています(カテゴリ数に対してデータ量が多かったのでこのような簡単な処理をしています)。

カテゴリ変数の変換	カテゴリ変数毎に目的変数の平均値が低い順に数値をラベル付

カテゴリ変数の変換の例

州	平均価格	ラベル
ニューヨーク	300	1
カリフォルニア	500	3
フロリダ	400	2

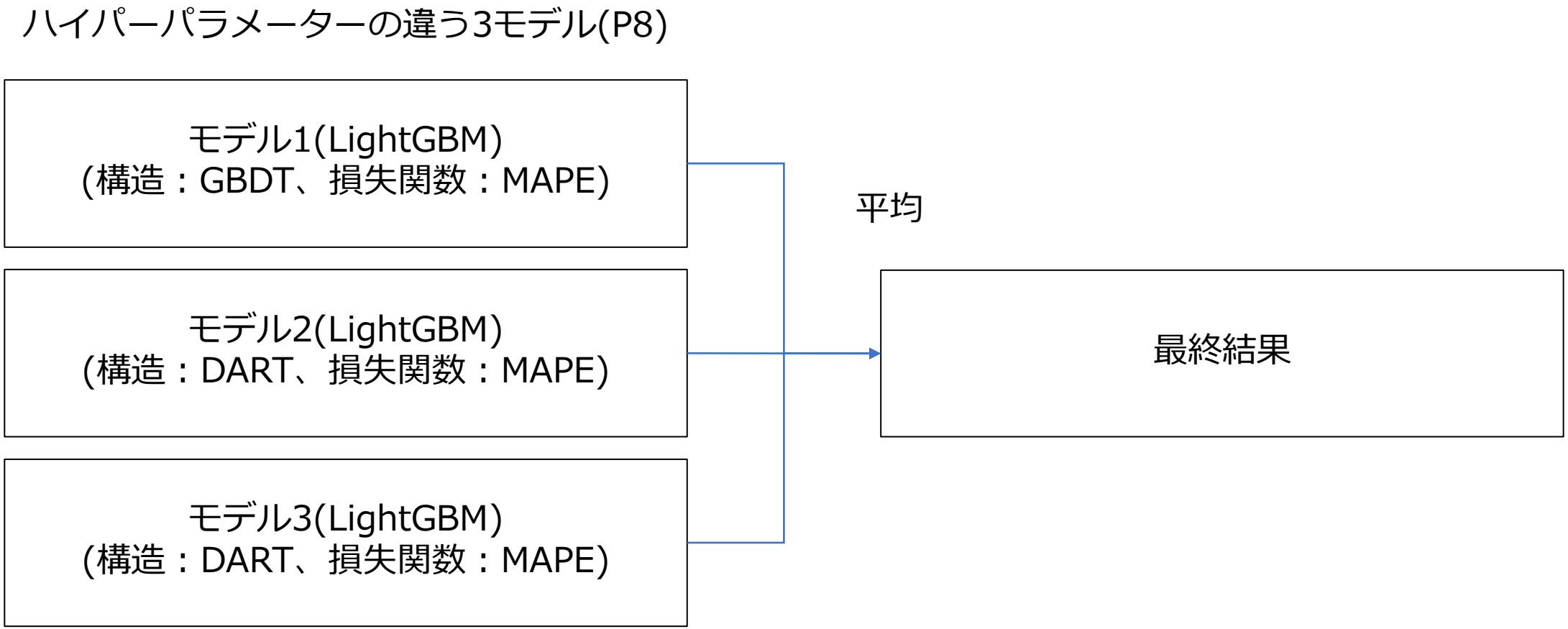
平均価格の小さい順に付与

目的

木構造モデルが学習しやすくなる効果を狙う

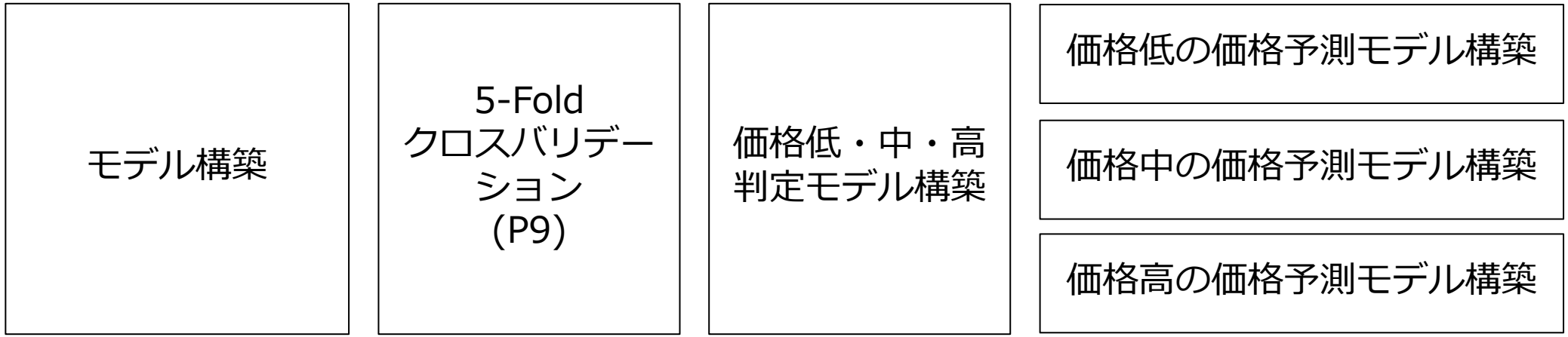
モデリング(全体像)

モデルはLightGBM、損失関数をMAPEとしてハイパーパラメーターの違う3つのモデルを作成し、最終的にその平均を取りました。各モデルの構造は次ページ以降です。



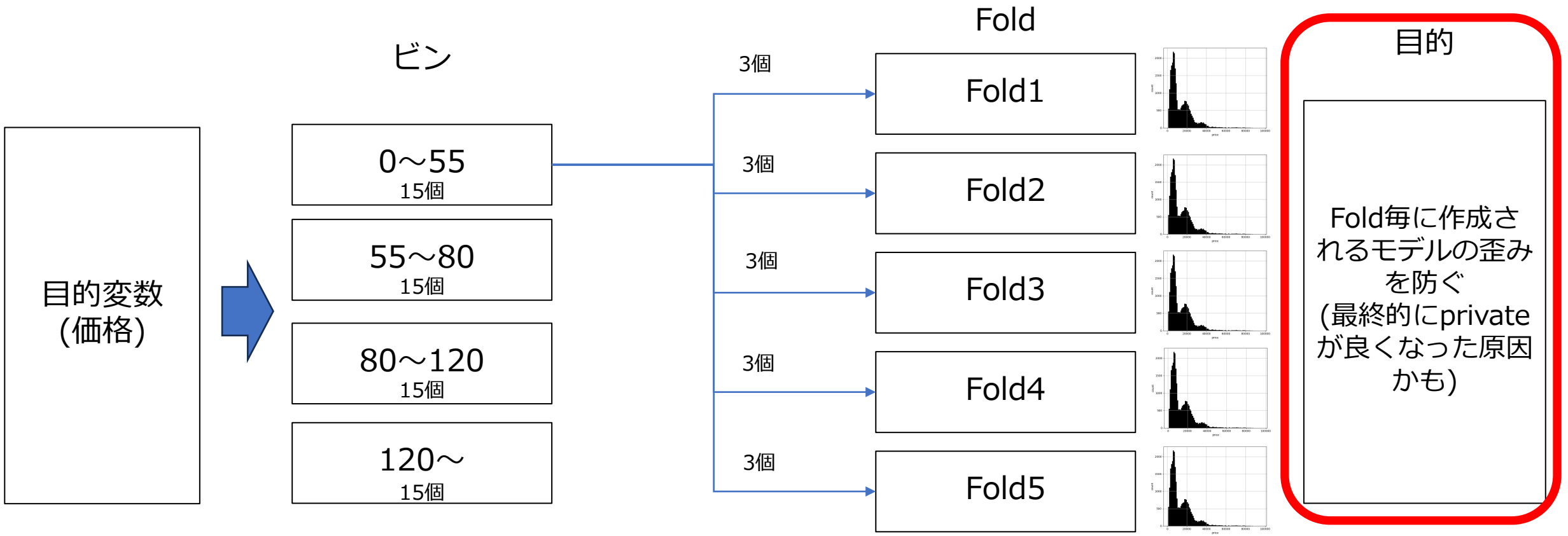
モデリング(各モデルの構造)

5-Foldのクロスバリデーション毎に価格を判定するモデルとそれぞれの判定した価格に対応する予測モデルを構築しました。



モデリング(バリデーション)

5-Foldのクロスバリデーションを作成しました。各Foldの目的変数の分布が等しくなるようにしました。目的は、Fold毎に作成されるモデルの歪みを防ぐためです。

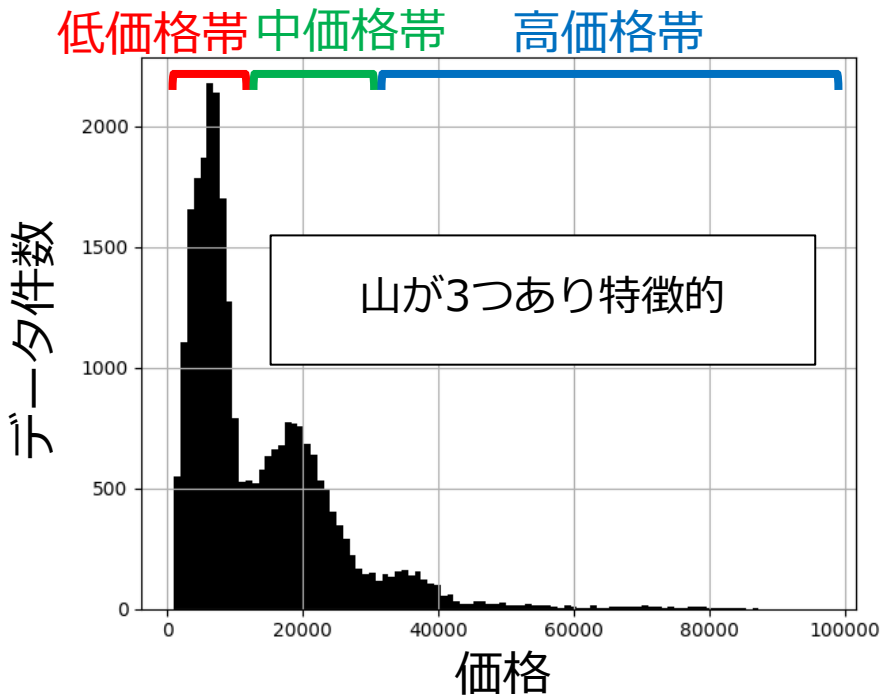


目的変数を大きさの順序でグループに分割する。
各グループ(ビン)のデータ数が等しくなる。

ビンのデータを各Foldに振り分ける
各Foldが含むビンのデータ数を等しくなるようにする

モデリング(判定モデルと価格予測モデル)

今回のデータは価格の分布は特徴的であり、単純に1つのモデルでは学習しきれないと想定しました。そのため、価格帯を判定後にそれぞれの価格帯に応じた価格予測モデルを適用する、というアプローチを取りました。



損失関数MAPEの1つの分布では学習しきれないかも

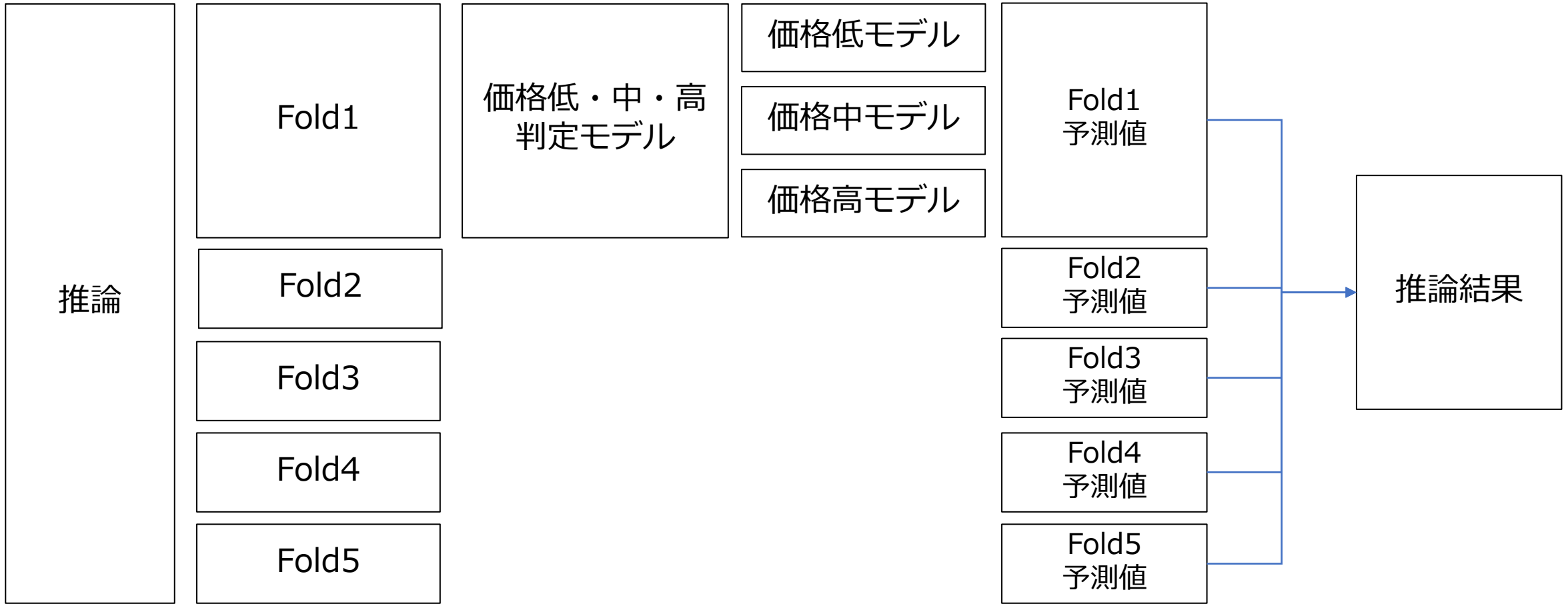
アプローチ

①まず価格帯を予測する
（判定モデル：低・中・高価格帯）

②価格帯に応じた価格予測モデルを
適用する

モデリング(各モデルの推論)

構築した各Fold毎の予測値を最終的に平均し、モデルの推論結果としています。



まとめ

今回のコンペは普段分析業務においても直面するような課題が多く含まれており（特に前処理の部分）、データサイエンスのスキルを磨くのに良いコンペという印象でした。

感想①	普段の分析業務においても直面するような課題が多く含まれており、データサイエンスのスキルを磨くのに良いコンペという印象でした
感想②	モデリングに入る前に目的変数の分布や説明変数の内容を丁寧に見ることなど、データの前処理の重要性を再確認できた学びの多いコンペでした。

関係者の皆様
このような素晴らしいコンペを開催・運営していただきありがとうございました。