==========================================

○ベースライン作成

・データの概要確認（欠損、データ型、テーブル構成）

・最低限のテキトー前処理

・モデル構築（lightgbm）

※ベースラインのハイパーパラメータを固定（デフォルトor過去でよい値）

○バリデーション（データ分割）

・Hold分割要否の決定

　・学習とテストの分布確認（pairplot×hue）

・CV分割数の決定（３～５）

・CV分割シードの決定（カテゴリによるStratifyの要否検討）

・前提：lightgbmのランダム要素なしor固定で実験

○単体モデル構築

・手法選択（tree、linear、deep）

・EDA（変数重要度、pairplot、heatmap、個別レコード）

・前処理（テストデータへのリークに注意）

・特徴量エンジ

・特徴量選択（boruta、null\_importance）

・後処理（丸め等。特に線形モデル等の外れ値的予測）

○提出モデル構築（仕上げ）

・Hパラメータチューニング

　・CV分割シードは別物

　・時間が係るため様子みながら

・スタッキング

　・モデルの違い、特徴量の違い、ランダムシードの違い

　・各モデルの出力の相関重要

○事後分析

・重要度分析（データの理解と新しい特徴量作成への着想）

・エラー分析（大外れ、カテゴリによる外れ傾向）

==========================================

○前処理

・欠損、外れ（偏り）、スケール×数値、カテゴリ

※カテゴリも外れ値処理重要（ビンの丸め）

○特徴量作成

・線形用、ツリー用、ディープ用

・手動作成、自動作成

○特徴量

※単独作用と交互作用を熟慮（目的変数に直接影響のみ、間接影響あり）

※尺度（間隔、比例等）を認識して加工する

・名義：車種、色

・順序：品質カテゴリ、ランク

・間隔：年式（西暦）、温度

・比例：価格、走行距離

・カテゴリの階層化（カテゴリ⇒カテゴリ）

　・抽象化：車種⇒タイプ（セダン等）、色⇒白、黒、その他

　・細分化：カテゴリの組みあわせ

・数値のカテゴリ化（数値⇒カテゴリ）

　・グループ化：顧客層（高所得）、

　・ランク化：

　・フラギング：年式⇒10年落ち、走行距離⇒1万キロ超え

==========================================

○データセット作成（前処理＋特徴量エンジ）⇒piplineのための整理

・欠損補完（平均、中央値、…）->複数選択肢

・カテゴリ変数のエンコーディング（ラベル、頻度、TE）->複数選択肢

・外れ値処理（削除、丸め）->複数選択肢

・新規変数作成（ドメイン＋α）

・合成変数作成（集約特徴量）

・圧縮変数作成（PCA、LDA、…）->複数選択肢

・特徴量選択（Boruta、…）->複数選択肢

（・分布変換）

（・スケーリング）

==========================================

■実行時間

○データ作成（1200行）（colab）

・CPU：50sec

・GPU（A100）：55sec

・GPU（V100）：63sec

・GPU（T4）：59sec

・TPU：63sec

○モデル構築（1200行）（colab）

・CPU：59sec

・GPU（A100）：19sec

・GPU（V100）：20sec

・GPU（T4）：20sec

・TPU：23sec

==========================================

○特徴量エンジ

・ランク変換

・高度エンコーディング（頻度、平均、ターゲット）

・組み合わせ変数

・集約特徴量

・カテゴリ変数×topic\_model

・テキスト（BoW/TF-IDF/word2vec/topic\_model/model->Bert/FastText）

・圧縮（SVD/PCA/T-SNE）

○基礎調査

・pairplot、heatmap

・学習データ、テストデータの違いに注意（外挿。目的変数、説明変数）

・

○その他

・AutoML：手法毎の精度を簡易調査

○メモ

・例外：ビンテージもの

==========================================