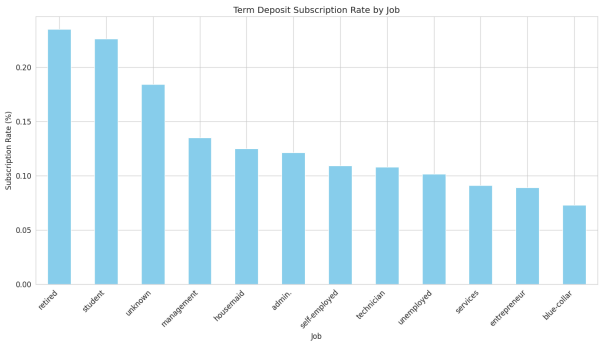
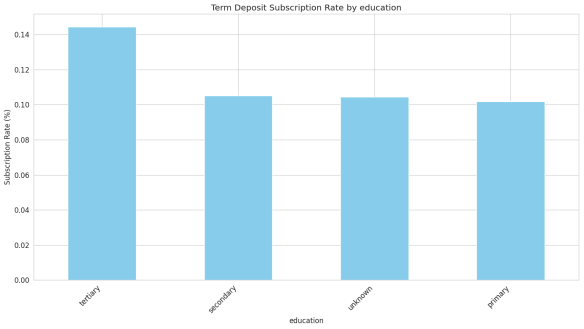
Cypherコンペ #02_最終レポート_チーム3

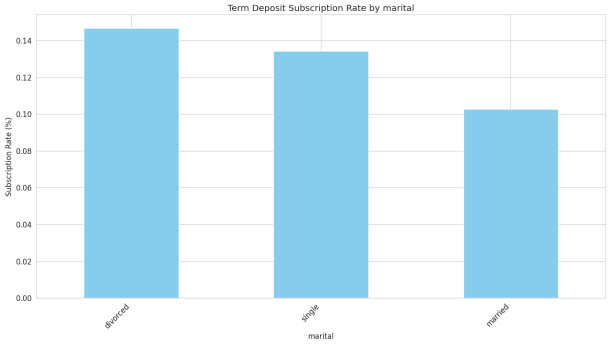
メンバー：濱元、手塚、山口

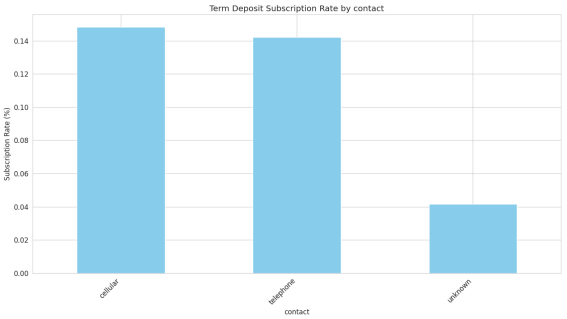
データに対する仮説

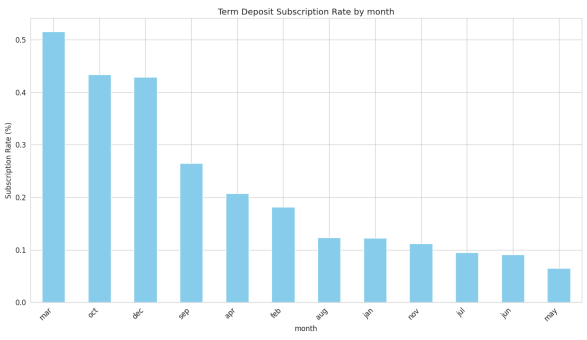
◯カテゴリ別の加入率から順番情報を付与したら良いのでは？
ロジスティック回帰は線形な情報に強いとのことだったので、幾つかのカテゴリ変数についてEDMのファイルでカテゴリ別の加入率を見て、順位をもとにカテゴリのラベルを数字に変換すれば線形になり、ロジスティック回帰の精度が上昇するのではないかという仮説を立てました。（以下水色の六枚の棒グラフ参照）





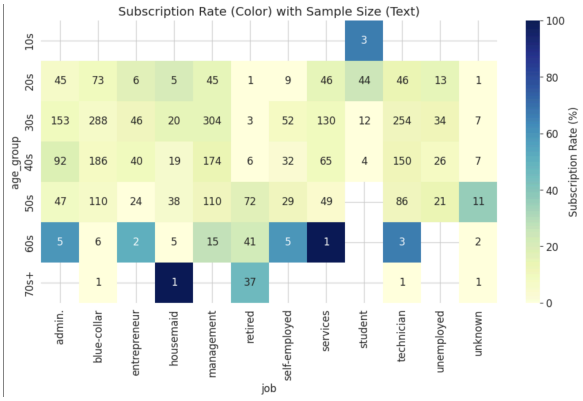


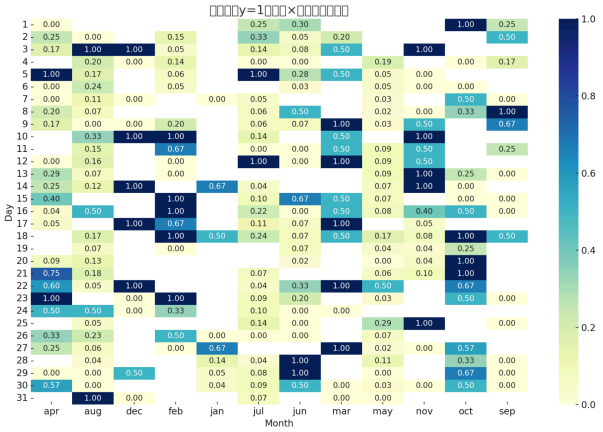




◯変数を組み合わせて新しい特徴値を作ることでより精度が上がるのではないか？

単一の値だけだとスコアを向上することが難しいと考えたので、いくつか変数を組み合わせて加入率の分布をヒートマップにして可視化しました。
以下に例として職業と年齢を組み合わせた特徴値、月と日にちを組み合わせた特徴値のヒートマップを添付しました。
job×age：全体的に高年齢層かつ特定の職業（例：60代自営業）に集中している傾向があるがまだばらつきがあるため他の変数との関わりがあるのではないかと仮説を立てました。
month×day：連日加入している割合が高い部分はキャンペーンと関係があるのではないかと考えました。

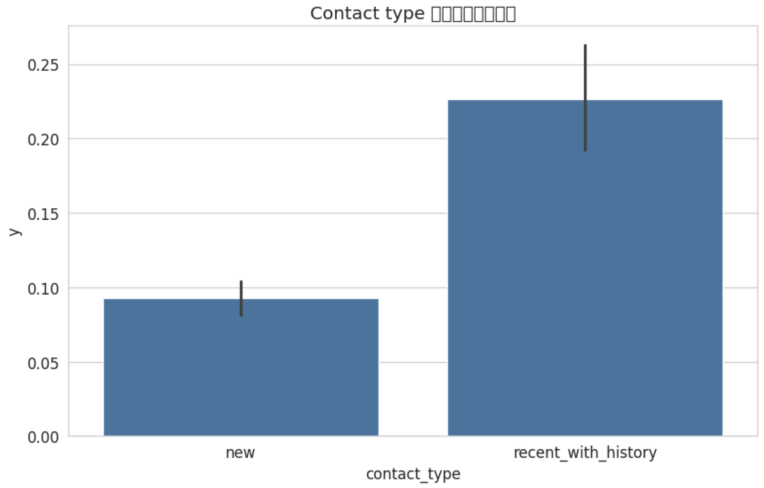




◯特殊な値を分離し意味づけすると、より有用性の高い特徴量になるのではないか？
顧客がこれまでにキャンペーンの接触歴があるかどうかを、pdaysとpreviousの2つの情報をもとに新たなカテゴリ変数contact_typeとして定義しました。

'new' : pdays = -1かつprevious = 0の場合 → 完全な新規顧客
'recent_with_history' : pdays >= 0かつprevious > 0の場合 → 過去にキャンペーン接触の履歴

<contact_type と y の関連性>



◯線形のデータに強いモデルには線形のデータを、非線形のデータに強いモデルには非線形のデータをそれぞれ学ばせて合わせれば精度が上がるのではないかという仮説を立てました。

特徴量エンジニアリング

それぞれで考えた仮説を特徴量エンジニアリングとして実装・集計し、最終的に大まかに以下の特徴量エンジニアリングを実装しました。

- 順序情報の付与（山口）(**`education_ordinal`**): 「高卒」「大卒」といった学歴データを、順番関係を持つ数値に変換することで、モデルが学歴レベルと契約率の相関関係を直接学習できるようにしました。
- 特殊な値の分離と意味付け（濱元）(**`pdays_transformed`**, **`was_contacted`**): 「未接触」を意味する `-1` という特殊な値を、接触済みのデータと明確に区別しました。
- 交互作用による文脈の追加 (手塚) (**`balance_vs_job_avg`**, **`pdays_x_poutcome`** など): 「個人の残高」を「職業グループの平均残高」と比較したり、「前回のキャンペーン結果」と「経過日数」を掛け合わせることで、単一の特徴量では表現できない顧客の相対的な状況や行動の文脈をモデルに与えました。

モデル概要

経緯：

線形のデータに強いモデルには線形のデータを、非線形のデータに強いモデルには非線形のデータをそれぞれ学ばせて合わせれば精度が上がるのではないかという仮説をもとにAIに相談していくうちに、それを4つのモデルで行い、最後に一つモデルで合わせるという形になりました。

選択したモデル

・ベースモデル:個別に4つのモデル（**LightGBM, XGBoost, CatBoost,** ロジスティック回帰）それぞれで学習する。

・メタモデル: ベースモデルの4つのモデルの予測値を新たな特徴量として受け取り、最終的な予測をLightGBMが行う。

理由:

1. 複数のモデルの良さを取り入れることができるため: 特徴の異なる4つのベースモデルを組み合わせて、異なる視点を取り入れ、全体の予測の安定性を高めました。
2. 合わせる時の精度を高めるため: ベースモデルたちの予測を単に平均するのではなく、そこにも学習モデルを噛ませることでより高い精度でベースモデルの予測結果を合わせました。

所感／学びの共有

山口:初めてデータサイエンスコンペティションに参加しましたが、仮説を考えて実装し、結果を集計する作業は科学の実験をしているようで楽しかったです。初めてなので、エンコーディングなど基本的な勉強をするところから始めました。ある日の夜に、モデルの特徴にデータの形を合わせ

たら良いのではと思い付き、それを実装して実際にF1スコアが上がった時は嬉しさのあまりすぐにスラックでチームメイトに報告しました。チームメイトは毎日「今日やったこと」「これからやること」をグラフやコードをまとめたドキュメントとセットで報告してくるので「自分もやらなければ…！」という気持ちが働き、それが強い原動力となり、改めてチームの大切さを学びました。
また、コード自体はほとんどAIに書いてもらったのですが、AIの威力を感じるとともに、AIに期待通りのアウトプットをしてもらったり、AIが書いたものを理解したりといった作業をするモチベーションや使いこなす能力の大切さを普段以上に感じることができました。

濱元：
データサイエンスコンペは初参加でしたが、仮説検証から特徴量設計、数値予測までの一連の流れを体験することで多くの学びを得ることができました。
一番難しいと感じた点は、与えられたデータをどう活用していくかという部分です。
カテゴリ変数や数値データは、そのまま利用することが難しく、データの組み合わせやデータ加工を繰り返していく中で、説明変数は絶対的な値として扱うのではなく、相対的な値として扱っていくことが大切だと感じました。
また、モデルに対する理解や特徴量エンジニアリングに関する知識、ChatGPTの活用方法への理解を高めていくことができたと実感しています。
加えて、今回はチームでの取り組みであったため、コミュニケーションや各メンバーの取り組みをうまく説明し、共有していくことが重要だと感じました。またお互いのチャットから、モチベーションも高められた点も良かったです。
結果的に、楽しみながら分析力とチームワークの両方を高めることができ、非常に有意義な経験となりました！

手塚：
データサイエンスコンペは初挑戦だったのですが、データ解析や機械学習に関する単語を一つひとつ調べながら徐々にデータの読み方や機械学習について理解を深めることができました。また、相互特徴値の組み合わせを工夫し、グラフ化することで膨大なデータからどの客層が定期預金に加入する傾向があるかを可視化し、それを元にF1スコアを改善していく過程が楽しかったです。
毎日チームで進捗を報告し合いコミュニケーションを取ることでプロジェクトによりコミットできたのが良かったです。チームメイトがうまくAIを活用しながら課題を取り組む姿を見て、どうAIに質問や提案を投げかけると効率的に作業ができるのか、勉強になりました。