CA Tech Lounge
課題1

河面雄樹

課題

covertypeデータセットを使って、気候や標高などの環境条件から、森林を占める木の種類を予測する多クラス分類問題を解いてください。

問1

適当な機械学習モデルで cross validation 及びそれによって生成されたテストデータに対する予測を実行して、結果と考察を報告してください。

問2

複数のアプローチで問 1を解いてください。具体的には特徴量の比較や機械学習モデルの比較を行い示唆をまとめてください。

問 1

木の種類を予測

→ 火事が起きたことがあるか、標高はどのくらいか、土中水分量はどのくらいか、どの地域かなどの相互作用で決まりそう

→ 相互作用(交互作用)を見られる学習モデルが直観的に良い

→ 今回は決定木系のモデルの一つ、LightGBM を用いて予測する

問 1

値の確認:

欠損値は無し。
異常な値も無くはないが、全体を占める割合は極端に低いので、影響は少ないと考える。

問 1

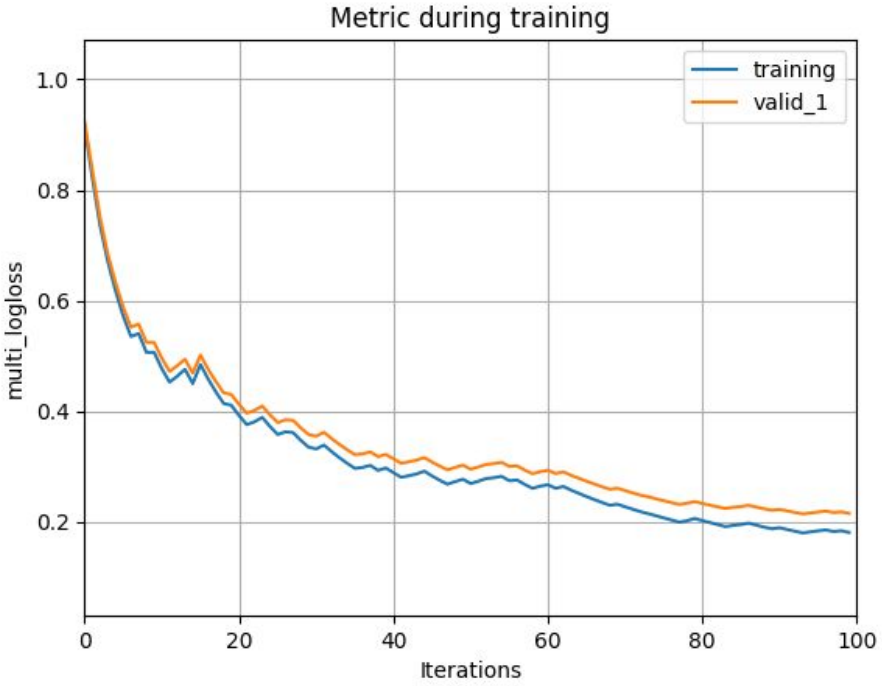
LightGBM モデルで多クラス分類を行った。

Optuna を用いて、交差検証を行い、ハイパーパラメータチューニングを行なった。

チューニングを行なったのは、{num_leaves', 'bagging_fraction', 'bagging_freq', 'feature_fraction', 'lambda_l1', 'lambda_l2', 'min_data_in_leaf'}　他はデフォルト。

ブースティングはdart で行なった。

訓練：検証：テスト　＝　0.6：0.2：0.2 でデータを分割した。



学習曲線

訓練データと検証データに分け、ブースティングの回数を増やすごとにLogloss(損失関数)がどの程度減るかを表している。

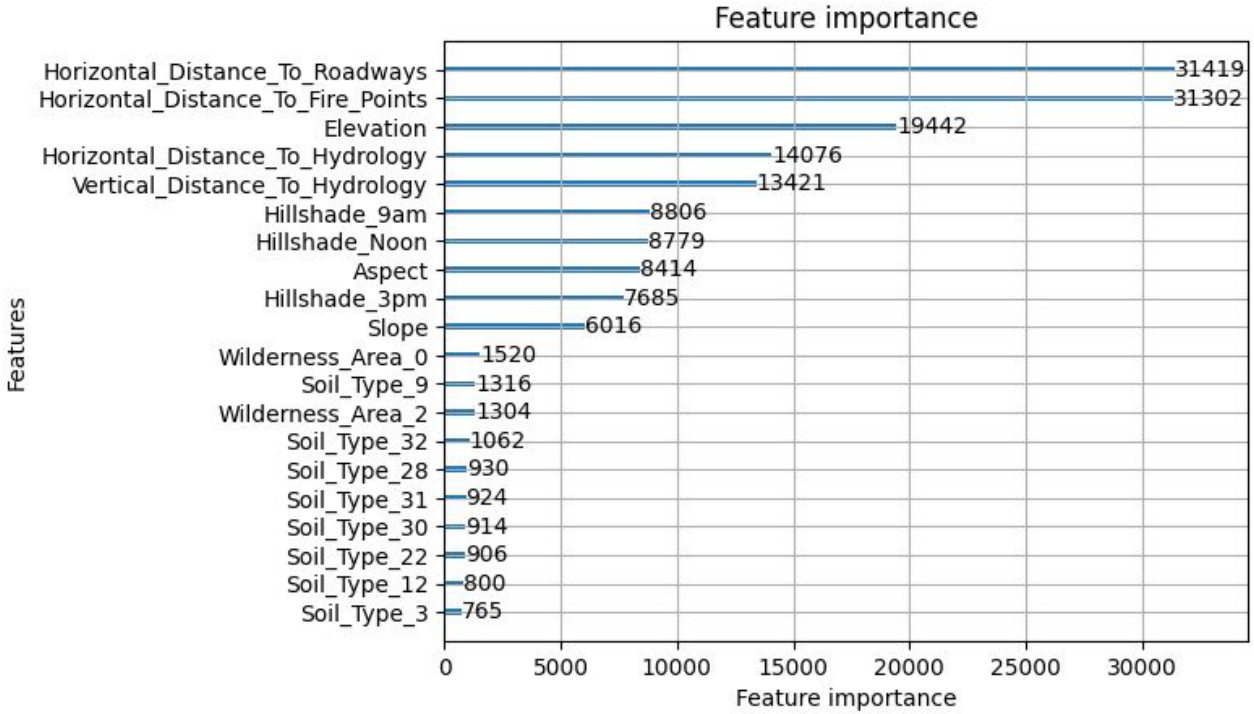
過学習は起こしていなさそう。

問 1

評価指標: Accuracy

= 0.9343820727519944 (テストデータ)

以下スライドにて、推論結果の考察をする



特徴量重要度

- 道路までの距離
- 火事発生地点までの距離
- 海抜何mか
- 一番近い水場までの距離

……

が決定木の分岐に大きな影響を与えている

Weight	Feature
0.4383 ± 0.0011	Elevation
0.1777 ± 0.0012	Horizontal_Distance_To_Roadways
0.1480 ± 0.0015	Horizontal_Distance_To_Fire_Points
0.0622 ± 0.0015	Horizontal_Distance_To_Hydrology
0.0526 ± 0.0009	Wilderness_Area_0
0.0334 ± 0.0008	Vertical_Distance_To_Hydrology
0.0294 ± 0.0003	Hillshade_Noon
0.0195 ± 0.0008	Hillshade_9am
0.0133 ± 0.0006	Aspect
0.0130 ± 0.0003	Wilderness_Area_2
0.0118 ± 0.0005	Soil_Type_28
0.0098 ± 0.0002	Soil_Type_31
0.0096 ± 0.0004	Soil_Type_21
0.0090 ± 0.0005	Soil_Type_32
0.0077 ± 0.0002	Hillshade_3pm
0.0070 ± 0.0003	Slope
0.0067 ± 0.0003	Soil_Type_30
0.0059 ± 0.0004	Soil_Type_22
0.0054 ± 0.0003	Soil_Type_3
0.0052 ± 0.0003	Wilderness_Area_1
… 34 more …	

Permutation Feature Importance

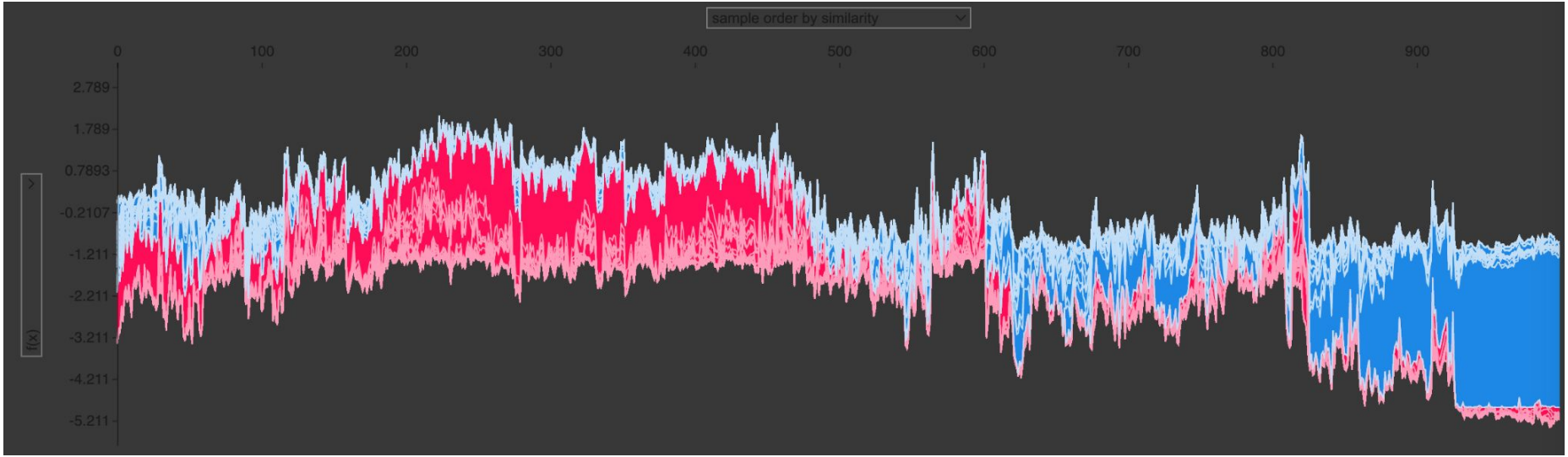
特徴量の値をデータ間で入れ替えてモデルに当てはめ直した時に、どの程度予測誤差が増えるかを測る指標

→ 誤差が大きいほど、その特徴量はモデルの予測に寄与していたと言える

→ 直接的な効果だけでなく、間接的な特徴量の効果(例:相互作用)も測れる

海抜と、道路、火事への距離は相変わらず予測に重要そう。

しかし、水場への距離は、予測値への影響は薄いか？
→ 他の変数との相互作用として予測に影響している可能性有り



•SHAP による個々の値の予測を 1000 例横に並べたもの (force_plot)
左の方の赤く太くなっている部分と、右の青く太くなっている部分は、 Elevation が持つ、個々の予測値に対する”力”（予測の値をどの程度正負方向に動かすか定量化した指標）
→全体を通して Elevation が予測に寄与している傾向がある

※七クラスある目的変数の内、一番目のクラスに対する SHAP の予測 (force_plot) を使用

問 1:考察まとめ

全体を通して、Elevation は木の分類分けに対して高い予測能力を持っている。
（他にも、道路や、火事発生地点への距離なども）

また、Elevation など、主たるpredictor との相互作用として、水辺への距離垂直的・並行的)も予測に対して影響力が大きいと考えられる。
→ 最初に建てた”相互作用が働いていそう”という推測は、合理的な推測であったと言えそう

テストデータでの精度も良く、過学習も起こしてなさそうなので、モデルは妥当そう

当該地域で森林を占める木の種類を知りたいときは、海抜高度と、道路、火事発生地点、水辺への距離、また、どの自然保護区域かなどが分かれば、約9 割の精度で推測が立てられる。

再掲：課題

covertypeデータセットを使って、気候や標高などの環境条件から、森林を占める木の種類を予測する多クラス分類問題を解いてください。

問1

適当な機械学習モデルで cross validation 及びそれによって生成されたテストデータに対する予測を実行して、結果と考察を報告してください。

問2

複数のアプローチで問 1を解いてください。具体的には特徴量の比較や機械学習モデルの比較を行い示唆をまとめてください。

問 2

まずベースラインモデルを立てる。
→ 多項ロジスティック回帰（正則化有り、全特徴量使用）R スクリプトにて実行）
→ テストデータ予測精度 :0.6981153

問 2

別アプローチ 1）

Tabnet を使った予測（全特徴量使用、各種ハイパーパラメータはチューニング済み）

評価指標 Accuracy：0.8731013829247093（テストデータ）

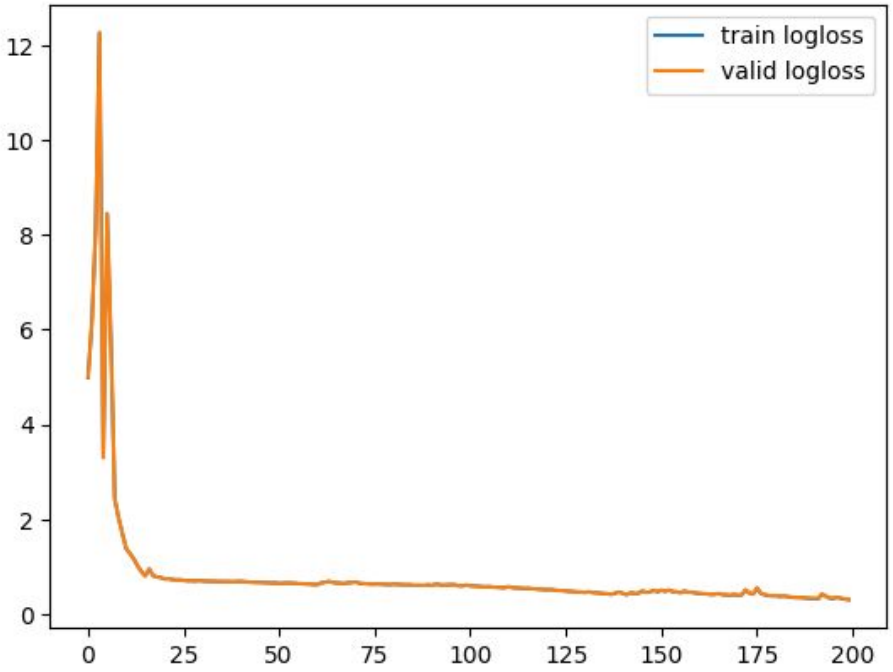
問 2

学習曲線:

過学習は起こしていなさそう。

気になる点:

Train と Valid があまりにも重なっている点



問 2

特徴量重要度:

やはりElevation が圧倒的に予測に寄与している(全体の 3 割)。
LightGBM と異なり、特定のSoil_Type もかなり影響する。
火事発生地点への距離は、Elevation の約 1/5 の影響力。
一方、水辺への距離や、道路への距離は重要度として低い。
→ WIlderness_Area_0 までで約 8 割の予測の説明が付く。

Elevation	2.902804e-01
Soil_Type_33	1.795125e-01
Soil_Type_27	1.341771e-01
Soil_Type_13	5.403456e-02
Horizontal_Distance_To_Fire_Points	4.572441e-02
Wilderness_Area_2	3.915821e-02
Soil_Type_28	3.535965e-02
Wilderness_Area_0	3.011381e-02
Soil_Type_37	2.725735e-02
Hillshade_Noon	2.327167e-02
Soil_Type_11	1.699547e-02
Wilderness_Area_3	1.690117e-02
Soil_Type_16	1.303577e-02
Soil_Type_3	1.220668e-02
Soil_Type_12	1.097772e-02
Horizontal_Distance_To_Hydrology	1.069449e-02
Soil_Type_30	7.379829e-03
Soil_Type_9	7.369230e-03
Soil_Type_32	6.727367e-03
Soil_Type_18	6.125742e-03

問 2

考察:

Elevation は Tabnet モデルでも共通して重要。
一方、水辺への距離、道路への距離は重要度は低く出ている。

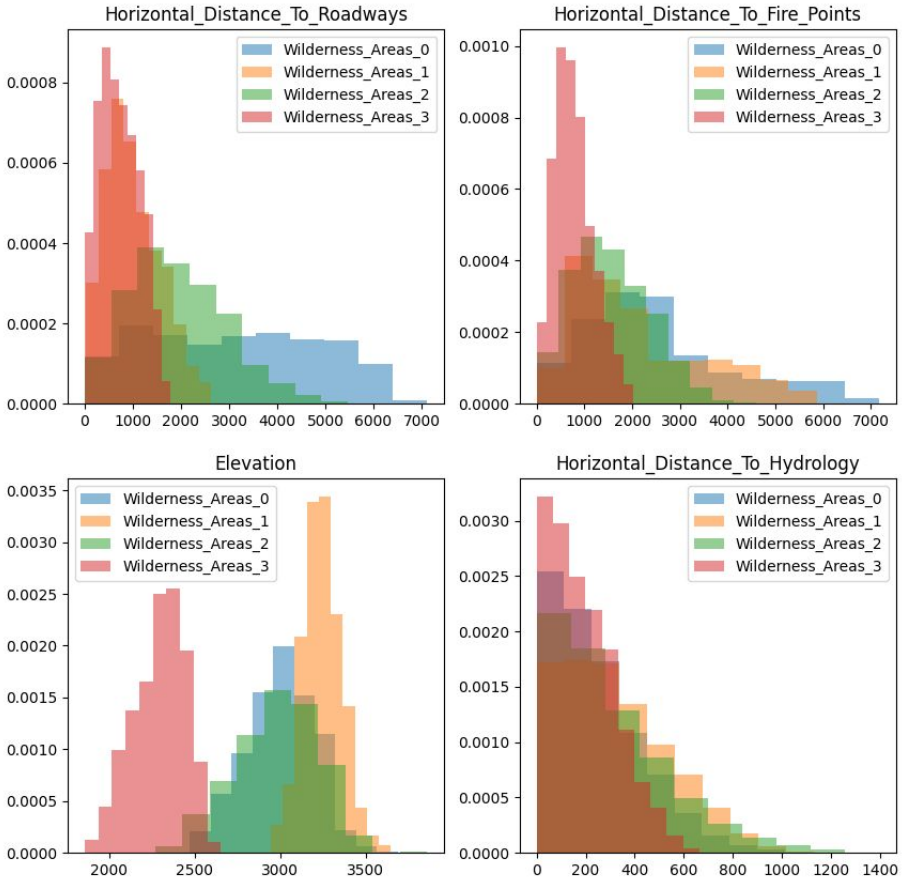
各種 Soil_Type が重要度が高く出ている。

→ 海抜高度、土壌の種類が木の種類の予測に強く影響する。
→ 直観的にもしっくりくる結果。
→ 水辺への距離や道路への距離の影響は、土壌の種類の影響に包含されている可能性。

問 2

別アプローチ2）

データを眺めると、自然保護区域の特徴量がある。
→ 自然保護区域ごとに標高や、道路への距離などが違って然るべき
→ 階層構造がある
→ multilevel model を使うのが自然か



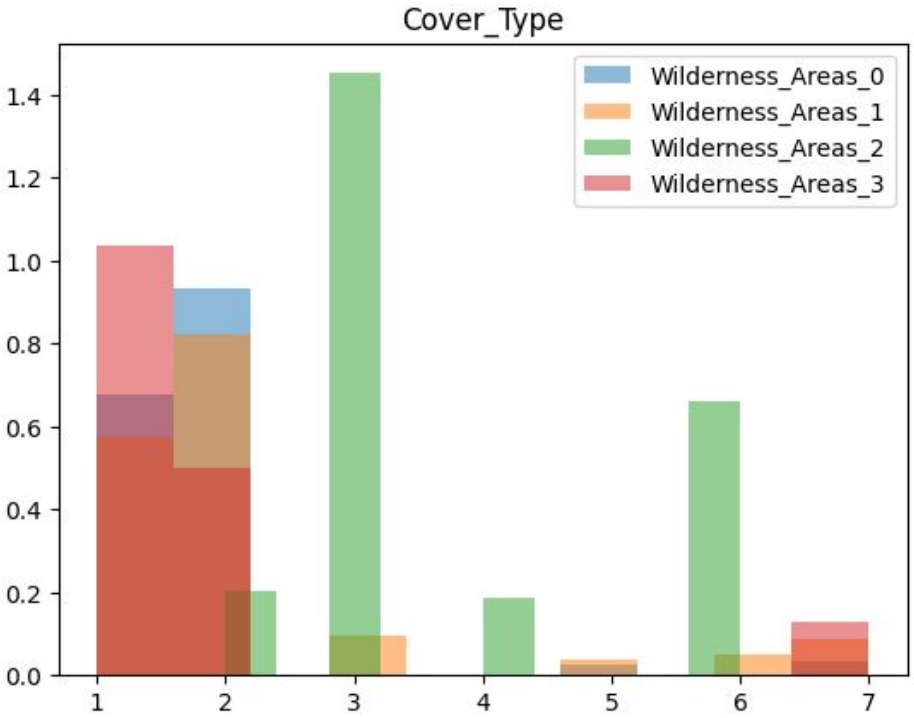
自然保護区域ごとに 4 つの特徴量をヒストグラムにした図

→ Elevation, 道路への距離に関しては、比較的分布の違いが出ている

→ 階層構造は確かに有りそう

自然保護区域ごとにデータを分割し、それぞれのデータの目的変数（木の種類）の値をヒストグラムにしたもの

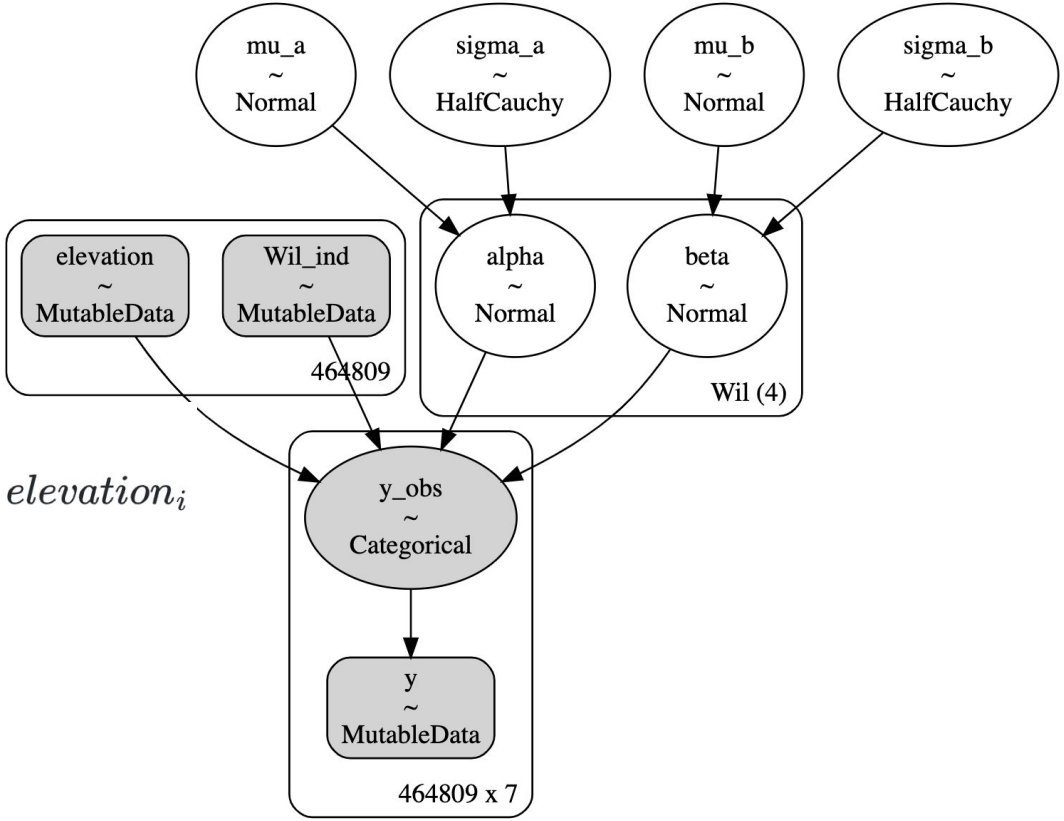
→ 実際に自然保護区域によって、目的変数の取る値が異なることが分かる



階層モデル

以下のような回帰モデルを立てた。

$$Y_i \sim Categorical(logit\_p = \mu_i)$$
$$\mu_i = \alpha_{\text{Wilderness\_Area[i]}} + \beta_{\text{Wilderness\_Area[i]}} \cdot elevation_i$$



階層モデル

評価指標:Accuracy

0.1403(テストデータ)

階層モデル

考察：
やはり、elevation 以外の特徴量もかなり重要。

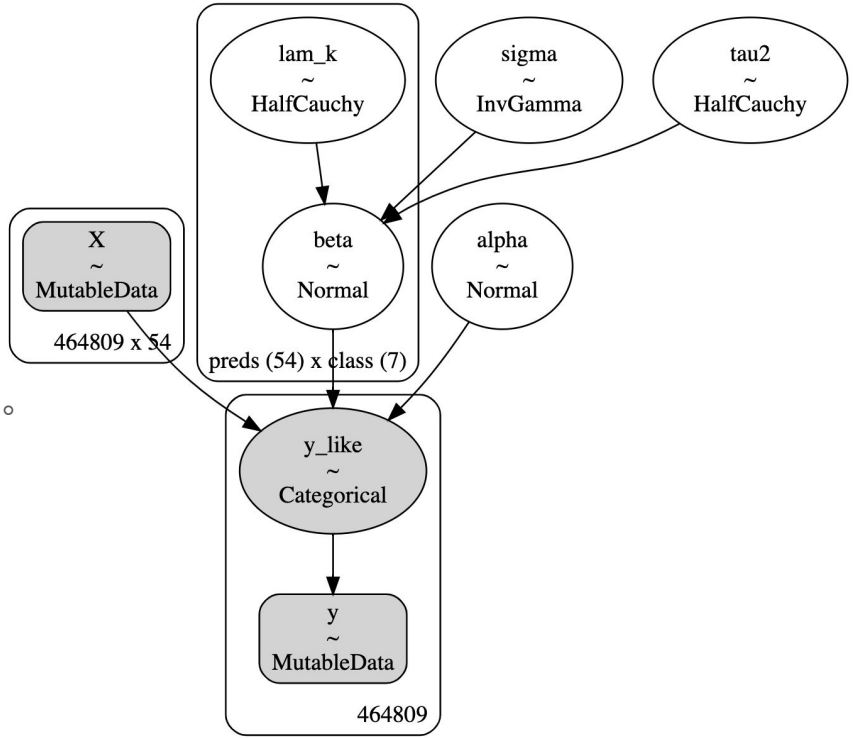
改善点：
技術的な話だが、複数の特徴量を階層モデルの中に組み込むコードが分からなかった。
特に、尤度関数（カテゴリカル分布）の挙動が上手くいかなかった。
コードが間違っている可能性も否定しきれない

問 2

別アプローチ3)

改めてベイズで事前分布を縮小事前分布として推論してみる。

評価指標 Accuracy: 0.3598271989535554 (テスト)



問 2

考察：
特徴量の単純な直接効果だけでなく、相互作用、もしくは非線形な効果が予測には活きてくる。

反省点：
コードが上手く書けている自信がない。
→ R の正則化付き多クラス分類(ベースライン)の精度が約0.7 あるのに、0.4 程度の精度しかないのは少し違和感

問 2 総括

- Elevation は各モデルを通して重要な特徴量。

- ただ、Elevation 単体ではなく、他の特徴量との相互作用が効いている可能性有り。
  → multilevel モデルより

- また、線形ではなく、非線形な関数で目的変数に影響を及ぼしている可能性が高い。
  → 縮小事前分布の線形回帰モデルより

- モデルとしてはLight GBM という決定木系のモデルが一番テストデータの予測精度が高かった。
  → 一番最初に建てた仮説(”相互作用が重要そう”)は正しそう。

- → 木の種類は、特に海抜高度などの様々な特徴量の相互作用が、非線形的に組み合わさって予測するのが一番。特に、決定木系の手法Light GBM は予測に有用。