**eコマース製品のカテゴリ予測**

**コンペの概要**

[**https://www.kaggle.com/competitions/tabular-playground-series-may-2021/overview**](https://www.kaggle.com/competitions/tabular-playground-series-may-2021/overview)

電子商取引の商品について、商品の様々な属性を記述している説明変数をもとに、商品にかんする目的変数のカテゴリを予測することを目指すコンペ。変数は全て匿名化されており、数値だけが分かる状態。

**EDAで分かったこと**

・訓練データ数:100000

・テストデータ数:50000

・５０種類の説明変数(全てカテゴリ変数)

・説明変数のクラス数は最大５８

・目的変数は、(0,1,2,3)の４つのクラスからなっている。

**誤差関数**

クロスエントロピー誤差

**分析手順と結果**

①特徴量作成

この工程はやらない。理由は、データが匿名化されているので、新たに特徴量作成の手がかりが得られない。分析精度の向上は、次の手順②で行う。

②特徴量選択、モデル選択

工程１．カイ二乗検定を用い、目的変数とは独立と考えられる説明変数を省く。P値と特徴　　量が対応するリストを作る。

結果：検定のP値を使って明らかに省ける変数は50個のうち１つのみだった。

工程２.　選ばれたすべての説明変数をつかい、LightGBMを使ってハイパーパラメーターチューニングなどを行いできるだけ精度の良い分析モデルを作成し、特徴量重要度の順位表を出力する。LightGBMを選んだ理由は工程３の計算コストを抑えるため。

結果：図１のようなグラフが得られた。上位６個～４５個までは特徴量重要度にあまり差はなく、図１からはどの特徴量を削除すれば良いか曖昧で、明確に予測精度を悪くしそうな特徴量がわかりにくい。

グラフ, ヒストグラム

自動的に生成された説明

図１ LightGBMによって得られた特徴量ごとの重要度

工程3. 工程２で得られた特徴量重要度順に、5~49個まで５個刻みに特徴量を選び、誤差を評価する。この作業を何通りかためし、最適な特徴量数を選び出す。特徴量数を入力したら、対応する予測誤差が返ってくる関数を作る。もし、工程２のモデルよりも精度の高いモデルが得られれば、それに対してハイパーパラメーター調整し、モデルを再学習し精度を上げる。

結果：図２のように、特徴量をすべて使用したモデルが最も精度が高かった。

考えられる理由としては

１：工程2において特徴量をすべて使用したモデルでハイパーパラメーター最適化をおこなっており、特徴量を減らしたモデルではハイパーパラメーター最適化していない。

２: 基本的に特徴量は多いほうが学習に有利、などがあげられる。

テーブル

自動的に生成された説明

図2 特徴量数と得られた誤差

③分析結果の解釈

・工程3で得られた最高の分析精度を持つモデルに、SHAPを使い、目的変数に対する説明変数の影響を詳しく見る。

　結果：詳細な結果は 下のリンクのパワポのSHAPの章で説明している。

<https://github.com/tomeneurodive/portfolio/blob/main/%E3%83%9D%E3%83%BC%E3%83%88%E3%83%95%E3%82%A9%E3%83%AA%E3%82%AA%E9%9B%86/e-commerce/e%E3%82%B3%E3%83%9E%E3%83%BC%E3%82%B9%E8%A3%BD%E5%93%81%E3%81%AE%E3%82%AB%E3%83%86%E3%82%B4%E3%83%AA%E4%BA%88%E6%B8%AC_tomeno.pptx>

**全体のまとめ**

・eコマース製品についての属性データ（説明変数)をもとに、その製品のもつカテゴリー(目的変数)を予測するプログラムを作った。

・モデルとしてLightGBMを使い、分析に使用する特徴量を削減することで、 予測精度を高めようとした。しかし特徴量の多いほうが予測精度は高かった。 KagglePrivate/Publicスコア : 1.09047 / 1.08752