【材料開発における多変量解析】

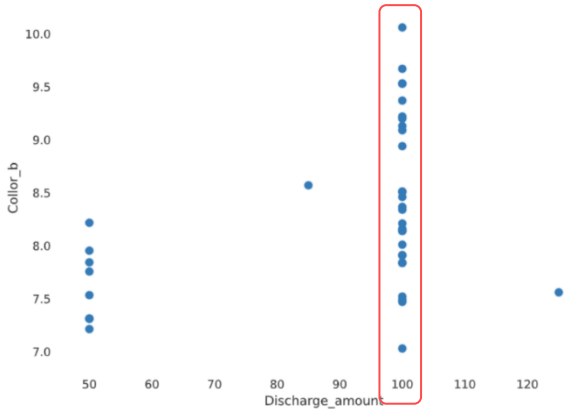
【多変量解析の目的】

材料系の研究開発では、一般的に蓄積できるデータ数が少ない傾向にある。また、実験計画法を用いずに蓄積したデータには偏りが生じている事も多く、データ解析をする際に精度が低くなる事例が多発している。そこで素材系の研究開発では「データ駆動型研究開発」と呼ばれる手法が取り入れられている。これはデータ解析が示す結果と、研究員の知見（理論科学に基づいた知見）をミックスさせて検討と行う手法である。今回は「データ駆動型研究開発」の一環として、素材系研究員のデータセットを用いて多変量解析を行い、先方の知見と解析結果の傾向が合致するかを事前に確認する。先方の知見と異なる結果が出た場合、その原因を考察し、不足分のデータを追加実験にて補う等の改善処置を行い、解析の精度を改善する事を目指す。

【多変量解析の方法】

まず先方のデータセットを用いて非線形の予測モデルを構築し、そのモデルから SHAP 値を算出する事で傾向分析を行う。SHAP 値は目的変数に対する各説明変数の貢献度（寄与度）を確認する事で行う。特筆すべき説明変数（研究開発においてキーとなる要因など）の寄与度が先方の知見と大きく外れていた場合、蓄積データから原因を考察し、対策を検討する。

【傾向分析の結果】



先方より、キーとなる説明変数として「PDR 温度」と「吐出量」の2つを紹介して頂いた。前提として「PDR 温度」は目的変数へ＋の影響を及ぼし、「吐出量」は目的変数へ-の影響を及ぼす事が理論科学の知見として分かっている。そして SHAP 値を確認してみたところ、「PDR 温度」は知見と合致したが「吐出量」は＋とも-とも言えない様な、特に大きな影響を及ぼさない結果となった。そこで、次に原因分析として、「吐出量」（横軸）と目的変数（縦軸）の関係をプロットで確認した。すると「吐出量」は4階級しかなく、データが100の階級に集中している事が分かった。100の階級のデータには、特に目的変数が高くなっているデータが多い事が分かった。

【考察と、今後の方針】

考察として、１００の階級の中の特に目的変数が高くなっているデータは、他の説明変数の影響によって値が高くなっている可能性がある。これについては先方の課題として取組んで頂く。この考察が真の場合、追加実験をしてデータを拡充して頂くが、その際は、「吐出量」と目的変数の関係が知見通りに現れる様に他の説明変数パラメータを調整して行って頂く。そうすると、その後の予測モデルによる解析の精度は向上すると期待される。