コンペ XX 説明文

アカウント名

・欠損値の補完について

とにかくクリーンなコードを心がけていたので、欠損値を埋める際の材料は全て train.csv から持ってきています。つまり、補完におけるリークを極力避ける形にしました。

・Fare

Pclass=3, Parch=0, SibSp=0 となる人でグルーピングし、その中央値で補完しました。

・Embarked

EDA の結果から、最頻値の S で補完しました。(他の特徴量から、欠損した 2 名は一緒に乗船した可能性が高いことがわかりました。)

・Age

Pclass, Sex, Fare(補完済), Title(後に説明)を特徴量とし、ランダムフォレストを使って予測した値で補完しました。

・Cabin

欠損値が多すぎるので、欠損している所は別の文字で埋めることにしました。特徴量として採用するか非常に悩みました。

・特徴量エンジニアリングについて

結果的に特徴量は 41 個になりました。以下に、新たに作成した中で特に有効だった特徴量を挙げます。

・Family_Survival_Rate

ある人について、家族がいるならばその家族の生存率を計算し、特徴量に。(CV の枠組みで。)

・NameLength

名前の長さ。意外にも有効だった。

・Family_Size

Parch+SibSp+1 として家族数を計算。

・Is_Female_or_Child

女性または 12 歳以下ならば 1、そうでないなら 0 としたもの。全てのモデルで非常に高い重要度を記録しました。

・Title

カラム Name から Mr, Mrs などの敬称を抽出。同義な Miss, Mlle は Miss でまとめるなどしました。特に Title_Mr は大きな重要度を示しました。

・Fare/TicketGroup

同じ Ticket を持つ人でグルーピングし、Fare をその人数で割ったもの。

・使用したモデルについて

主に以下の 5 種類のモデルを使用しました。

・ランダムフォレスト

・k-近傍法

・LightGBM

・XGBoost

・ロジスティック回帰

各モデルについて Grid Search を行い、CV スコアが最も高いモデルを各モデルの base モデルとしました。そして、各モデルのハイパーパラメータを変更した派生モデル(max_depth を増やしたものなど)をいくつか作成し、合計で 14 個のモデルを用意しました。

また、ニューラルネットワークモデルも作成してみましたが、高いスコアが期待できなかったため、今回は除外しました。

・特徴量の選定について

各モデルについて特徴量重要度を表示し、重要度が 0 あるいは非常に少ない特徴量を逐次削除しました。K-近傍法だけは、手動で 6 つだけ選択しました。

・アンサンブルについて

結果的に「重み付きのフォワードセレクション」を採用しました。ひとまず沢山モデルを作って、いちばん OOF スコアが良かったものをベースモデル(B)とし、新たなモデル(N)と重み付きアンサンブルを行ったとき、少しでも OOF スコアが改善したならそのモデルを B に追加、改善しないならその N は落とす、という具合です。これによって RF_base, KNN_base, XGB_base, LR_best という 4 つのモデルが採用されました。この方法ではベストな重み付き和も算出されるのですが、それを提出しても PublicLB があまり高くない(0.81 くらい)ことから、ここからは手動で重みを調整して最も良さそうな重みを最終的に採用しました。これはおそらく過学習、すなわち OOF スコアと PublicLB に乖離がある(0.03 くらい)ことが原因で、私の今後の改善点でもあります。最終的に Public スコアは 0.822、Private スコアは 0.816 となりました。