第三回コンペ解法

化学システム工学専攻修士 2 年見内伸之

1．解法の概要

今回のデータセットにはカテゴリカル変数が多かったので、その変数に対して Count Encoding と Target Encoding を行うという、オーソドックスな手法を採用しました。
Count Encoding と Target Encoding の簡単な説明を下に載せます。
Count Encoding: データセット中にそのカテゴリが登場した数にカテゴリを変換する手法。
Target Encoding: あるカテゴリを、train データ中のそのカテゴリの y(今回の場合は”is_arrested”)の平均値に変換する手法。

工夫した点は、カテゴリカル変数同士を文字列として組み合わせたものを計算し、新しい記述子として利用した点です。その様子を下図に示します。

	stop_date	county_name	stop_date_county_name
0	2014-09-23	Hartford County	2014-09-23_Hartford County
1	2014-08-10	Fairfield County	2014-08-10_Fairfield County
2	2013-12-17	Windham County	2013-12-17_Windham County
3	2014-02-07	Middlesex County	2014-02-07_Middlesex County
4	2014-11-05	Tolland County	2014-11-05_Tolland County
5	2014-09-16	Fairfield County	2014-09-16_Fairfield County

もともと記述子として存在した stop_date と county_name という記述子から中身を単純に結合して作成された stop_date_county_name という新しい特徴量を作りました。

組み合わせる記述子としては以下の 12 変数を選びました。
“stop_date”, “stop_time”, “location_raw”, “county_name”, “fine_grained_location”, “police_department”, “officer_id”, “driver_gender”, “driver_race_raw”, “violation_raw”, “search_type_raw”, “stop_duration”
選んだ基準は①カテゴリカル変数であること。②”driver_race”, “driver_race_raw”のように raw と raw でないもの両方が登録されていたら raw の方を残す。です。”state”は全てに“CT”

が入っていて分類には役に立たないので、カテゴリカル変数ではありましたが除外しました。この 12 種類の特徴量に対して今回の処理を施すと、新たに 66 記述子が作成されます。この 66 記述子に対して Count Encoding, Target Encoding を行い、132 記述子を用意しました。これらを用いて予測を行いました。

2. 学習モデル

統計手法としては Light GBM を用いました。また、今回のデータセットの特徴として正例と負例の数の差が大きかったので、under sampling も用いました。under sampling は正例と負例の数の差が大きいデータセットに対して、正例の数に合わせて負例を sampling したデータを学習データに用いて予測を行うものです。Under Sampling を何度か行った予測結果を ensemble したものを提出しました。

3. 感想

もともとは全く違う方針で記述子を作成していて、総当たり的に作った 1500 個ほどの記述子を用いたモデルを最終日まで使っていました。そのモデルの AUC は leaderboard 上で 0.874 であり、順位は 6 位というものでした。そこでもう一押ししようと、今回の記述子を作成しました。実装はあまり時間がかからずに終わり、元の記述子に追加して計算したみたところ、手元の評価で AUC が 0.9 を超えました。コンペ期間中初めて 0.9 を超えたので、驚いたことを覚えています。試しに新たに作成した記述子のみで予測を行ったところ、そちらの結果のほうが「元々の記述子」のみの結果はもちろん、「元々の記述子」+「新たな記述子」の結果よりも良い結果を示すことがわかりました。最終的には今まで作っていた 1500 個ほどの記述子を全て捨て、最終日に作成した 132 個の記述子で作った予測モデルの結果を提出しました。

結局総当たり的に作成された記述子よりも、対象をしっかり説明できる記述子の発見が最後の勝敗を分けるのだと身をもって感じたコンペでした。