Playground Series – Season 3, Episode 11

Tabular Regression with Media Campaign Cost Dataset

**Problems Encountered**:

1. The size of the training dataset is quite large but the features themselves where not good predictors for target cost
2. A lot of categorical features which were disguised as numeric feature
3. **Just the mean of target gave us scores which were almost the same as most model predictions scores**
4. Hyperparameter tuning is slow because of large size
5. **Not all features are good for all models. Some feature increase score in one model and decrease score in another model**

**Solutions:**

|  |  |
| --- | --- |
| Solutions | Links |
| #1: A Zoo of Models | **https://www.kaggle.com/competitions/playground-series-s3e11/discussion/399401** |
| 17th place solution | Hand-Tuning and features | **https://www.kaggle.com/competitions/playground-series-s3e11/discussion/399393** |
| 17th Place Solution And Approach | **https://www.kaggle.com/competitions/playground-series-s3e11/discussion/399438** |
| #81: First Top 10% but It Could've Been Better... | **https://www.kaggle.com/competitions/playground-series-s3e11/discussion/399407** |

**What I Have Learned:**

* **Only a subset of the features is useful, although it's not completely clear which features belong to the subset. Being in doubt which subset is the right one, we can make models for different feature subsets and blend them. A diverse ensemble, averages out the prediction errors of the models.**
* **Use stratified kfold for regression too as it makes the model more robust see this notebook for reference ”https://www.kaggle.com/code/janmpia/feature-eng-xgb-cat-ensemble-0-29265/notebook”**
* **Pipelining and custom transformers**
* **Sometimes it is better to just hand tune model hyperparameters. So I need to learn more about each feature or to quickly learn about hyperparameters tuning for an particular model search online like Hyperparameter tuning for random forest**
* **See third solution for feature selection technique**