Machine Learning Nanodegree Capstone project

Anton De Meester – 06-08-2019

Summary

My proposal for the ML Capstone project is an end-to-end automated Machine Learning model for Numerai. Numerai is a data science firm for Financial Services that provides an online competition with clean data to outsource part of its own predictions and to attract talent.

Background

Numerai is an investment company that outsources (a part of) its research to the community. They host a competition and provide rewards for the best predictions of a data set that they provide.

They provide a clean dataset of 1 time variable (era), 310 normalised features and 1 output variable. The training dataset is around 500.000 lines, and they also provide validation, testing and live data. The output variable is required as a fraction, as thus makes it a regression problem.

The goal is to predict the live and test set, which only Numerai will validate. If you join the competition your predictions will be compared to other community members.

You can either just compete or you can also stake some numer, a cryptocurrency related to Etherium. If you stake some numer, you can earn more numer if your predictions are better than average. The exact details for how and how much you earn can be found here.

Problem statement

The goal of Numerai is easy, provide the best predictions for the dataset they provide.

On top of that, there are some additional, personal goals:

- Do some EDA on the dataset to understand it a bit better
- Create the best model in a iterative way
    - Set a couple of initial models
    - Train them with limited dataset with hyperparameter tuning
    - Take the best ones
    - Train them on the total data set
    - Try to combine them (weighted averages)
    - Select the best model
- Automate this model creating
- Automate the entire workflow
    - Download data
    - Make predictions on the (already defined) best model with a batch
    - Upload the data to a central repository
    - (If possible) Upload the data to Numerai
    - Automate this so it runs every week automatically

One set of examples of solutions is provided in a Python packages, Numerox. Numerox is provides an interface to Numerai and also has a set of example models that you can use.

Dataset and inputs

The dataset has two parts, a training and a tournament data set.

Each data set has 310 features, divided in 6 types (intelligence, charisma, strength, dexterity, constitution, wisdom), 1 output variable (target_kazutsugi), an id, an time variable (era) and a data type (train, validation, test, live).

The training data set is 501808 rows with only training data, the tournament data is 385290 rows, divided on validation, test and live data.

Each feature is normalised and one of 0.00, 0.25, 0.50, 0.75 and 1.00. The target has the same values. The ear is between 1-120 for training data, 121-132 for validation data, 133-196 for test data and eraX for live data.

Solution statement & project design

As provided in the problem statement, the solution would be in 3 parts.

- EDA: A light EDA exercise to understand the data a bit better, and to provide some hint which models might be better
- Model training: The largest part of the project. Many models would be trained and there would be extensive part of hyperparameter training.
  As mentioned, the model training would be in 2 parts:
    - Individual model training: Creating several models (5+), and do extensive hyperparameter training on each of the models to obtain the best version. Each of trainings would be done with a selection of the data (proposed 10%)
    - Model comparison: after each of the individual models would be optimised, the optimised set would be trained with the full data set. Then each model would be compared, and possibly combined to obtain the best overall model
    - Models that I would use are:
        - Linear Regression
        - XGBoost
        - SGD Regressor
        - LightGBM
        - 2, 3, and 4 layer Neural Networks, with different kind of nodes, loss functions and optimizers
        - And any more that I find
    - Methods that I would use:
        - PCA to reduce the amount of features
        - Hyperparameter tuning
        - Ensemble methods to combine the best of all models
- Automation: The entire process should be automated.
    - Whenever a new data set is presented, the complete training can be redone “with the touch of a button (and 20€ for AWS)”
    - The complete competition should also be automated. The data can be downloaded, then batch processed with the current best model, saved to the correct S3 location, and (if possible) uploaded and staked on the Numerai site. This should be able to be scheduled with a cron job.

Benchmark models & evaluation metrics

Because Numerai has a competition, it's quite easy to compare the model to others. Additionally, the feedback of losing/gaining a stake is quite clear when and if your model does well.

Numerai uses the correlation coefficient to rate its submissions (`np.corrcoef`), and thus so will I. Additionally, during the development, I will also use accuracy the help train the model.

During training, I will split up the data in training, validation and test sets, as well as use the provided validation data with outputs provided by Numerai.

For the evaluation of the final model, I will provide my username and can then provide the feedback of Numerai which gives an indication how well it would score. (As far as I understand, they do this with the test or validation data). Final results from numerai are delivered a month after the round ends.

For the benchmark model, I will compare it to 2 things: a standard Linear Regression model and previous rounds of Numerai. Because they recently changed methodologies and also evaluation metrics, there is no round that has resolved the new metrics yet. However, it will be provided on August 13. They will be found here. My benchmark is to do better than 75% of the provided predictions, in the correlation coefficient.

Notes

Datasets can be found on the Numerai site: the big blue button on their site: https://numer.ai/homepage

The latest data set of this week is also on my S3 folder, which can be found here: https://sagemaker-eu-west-1-729071960169.s3-eu-west-1.amazonaws.com/captone-project/numerai_datasets.zip

The data set changes every week, but the structure remains constant. As far as I understand, the training set also doesn't change.