Two-Stage Direct Response Predictive Model

The Nature Conservancy
GWSB Group 8
Yimi Chen, Xiwen Gui, Kylie Loudermilk, Giovanna Maria

Overview:

Executive Summary

Problem Understanding

Data Preprocessing

Feature Selection

Model Process

Model Performance

Potential Next Steps

Risk Considerations

Client Introduction

The Nature Conservancy, a global environmental non-profit, which is advancing conservation in all 50 states and U.S. territories and in 70 countries around the world.

Founded in the U.S. through grassroots action in 1951, The Nature Conservancy (TNC) has grown to become one of the most effective and wide-reaching environmental organizations in the world. Thanks to more than a million members and the dedicated efforts of our diverse staff and over 400 scientists, we impact conservation in 79 countries and territories: 37 by direct conservation impact and 42 through partners.

- Mission: To conserve the lands and waters on which all life depends.
- Vision: A World where the diversity of life thrives, and people act to conserve nature for its own sake and its ability to fulfill our needs and enrich our lives.

Executive Summary and Problem Understanding

TNC relies on fundraising efforts to support its mission, so we need to develop a two-stage direct response model to maximize revenue efficiently.

- ❖ The Nature Conservancy faces the challenge of optimizing its direct-mail fundraising appeals program to maximize net revenue.
- ❖ Traditional approaches lack precision, leading to inefficiencies. Appeal to all donors risks financial losses, while targeting only the most responsive donors neglects high-value, less responsive ones.
- ❖ Balancing responsiveness and value are critical for success.

Project Objectives

Build a two-stage direct response model to predict (1) probability of donor response and (2) expected gift amount.

Maximize net revenue of direct-mail fundraising appeal campaign by identifying donors with highest expected gift potential.

The Nature Conservancy

Brief Steps:

Target1stPrtyfile

Demofile

Transfile

Pre-processing Data

Hyper-parameter tuning

Addressing Outliers

Feature Selection

Model-Phrase 1:
Random Forest Classification

Model-Phrase 2:
Random Forest, Gradient Boost

Generate Predictions

The Nature Conservancy

Methodology: Data Analyzed

Internal data:

Target1stPrtyfile
- Past donation history. - Includes target variables. "TGTresp" and "TGTgiftamt".

Third party data:

Demofile	Transfile
- Includes demographic information such as age, gender, geographic location. - Donation history and responses to marketing campaigns.	- Performance metrics from previous marketing campaigns. - Includes response rates and revenue generated.

Target variable:

- TGTresp: 0/1, response or not.
- TGTgiftamt: the donation amount.

Data shape:

- 3276648 rows * 62 columns

Max/min/mean of target gift amount:

- $9636/ $0/ $0.88

The Nature Conservancy

Data Preprocessing

Merge three files	Field type manipulation	Handling missing values	Deal with duplicate values	Deal with redundant values
Used 'masterprimaryid' as the key. 'Masterprimaryid' corresponds to a donor.	Converted currency values from string format to float data type to facilitate numerical calculations.	Dropped the 'append_enviroconquintile' column due to a high proportion of missing data. Removed two rows in the 'append_HomeValue' column where data was absent. Filled missing values with the mean for some variables.	Removed columns 'birth_year' and 'append_age_indicator' as these fields contain overlapping age information.	Removed 'run_date_y', 'run_date_x' columns for their lack of contribution to predictive modeling.

Feature Engineering

Created dummy variables

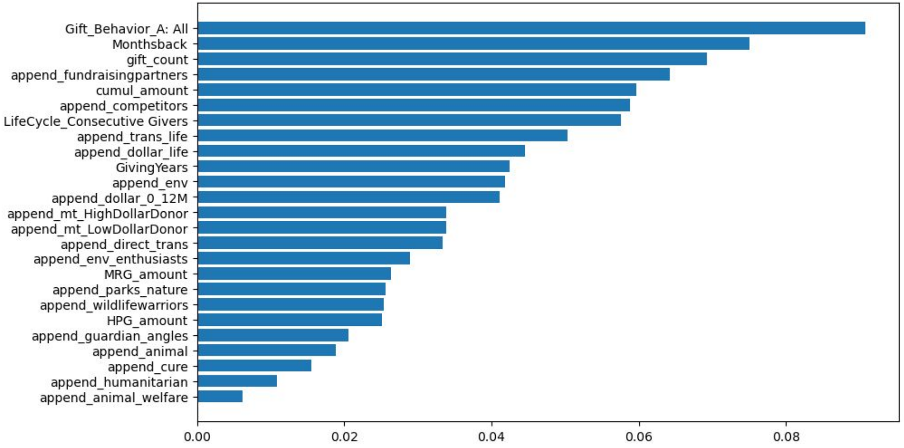
- 'Monthly_Donor', 'YE_Behavior', 'Gift_Behavior' reflect donation timing and patterns, identifying year-end donors, regular monthly contributors, and responses to varied fundraising campaigns.
- 'First_gift_channel', 'MRG_channel', 'HPG_channel', 'Prior_Channel_Behavior' captured the initial, most recent, highest, and preferred donation channels.
- 'LifeCycle', 'LifeCycleDetail' ,'Donor_status' for classifying the duration of continuous donations or the time since lapse.

Feature selection

- Utilize a grid search to tune hyper-parameters.
- Use selectKbest[1] for feature selection.
- Review all feature distributions to find the most important features.

The Nature Conservancy

Phase 1 Feature Selection

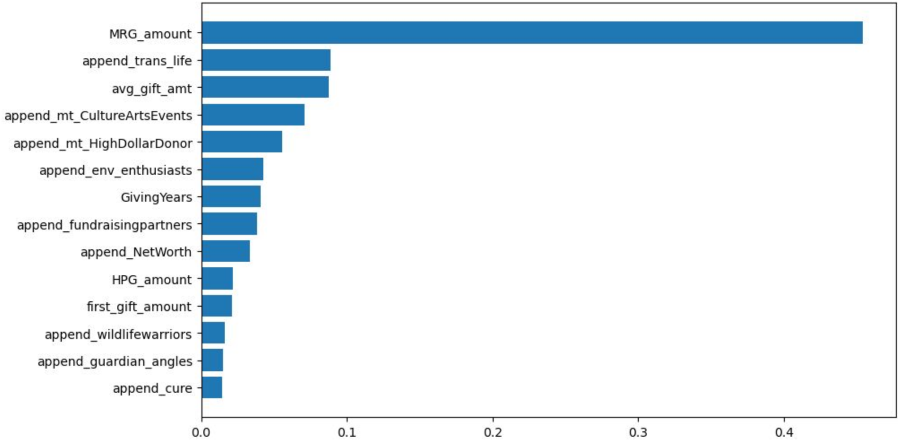


Column Name	Description
Gift_Behavior_A: All	Kinds of donor behavior: all types,including appeal, renewal and others.
Monthsback	Number of months since the donor's last contribution.
Gift_count	Count of donation transactions.
append_fundraisingpartners	Cumulative number of participants donating to non-profit merchandise categories in the past 12 months.
cumul_amount	cumulative donation amount
append_competitors	Transactions with competitor non-profits.
LifeCycle_Consecutive Givers	Donors with consecutive giving years.
append_trans_life	Overall transactions lifetime.
append_dollar_life	Lifetime total of donation dollars.
GivingYears	Number of donation years.
append_env	Number of transactions with environmental non-profits within lifetime.
append_dollar_0_12M	Overall dollars 0-12 month.
append_mt_HighDollarDonor	Likelihood to donate more than $500 to non-religious causes.
append_mt_LowDollarDonor	Likelihood to be Low Dollar Donors.
append_direct_trans	Past year number of direct transactions.

- The likelihood of a response is strongly linked to **donor's historical engagement**.
- Demographic factors appear to have a less pronounced correlation.

The Nature Conservancy

Phase 2 Feature Selection



Column Name	Description
MRG_amount	Most recent donation amount
append_trans_life	Overall transactions lifetime.
avg_gift_amt	Average donation amount.
append_mt_CultureArtsEvents	Epsilon's ranking for likelihood to attend cultural arts events.. Rank 1 is the best and 99 is the worst.
append_mt_HighDollarDonor	Likelihood to donate more than $500 to non-religious causes.
append_env_enthusiasts	Environment enthusiasts' participants in a lifetime.
GivingYears	Number of donation years.
append_fundraisingpartners	Cumulative number of participants donating to non-profit merchandise categories in the past 12 months.
append_NetWorth	Epsilon's net worth estimate. Value equals household's asset minus liabilities.
HPG_amount	Highest previous donation amount.
first_gift_amount	Amount of the first donation.
append_wildlifewarriors	Past year participants in non-profits for animal health and welfare.
append_guardian_angles	Cumulative number of participants in the past 12 months that are in search of a cure for disease or medical conditions that affected a loved one.
append_cure	Cumulative number of participants in the past 12 months who donate to medical causes

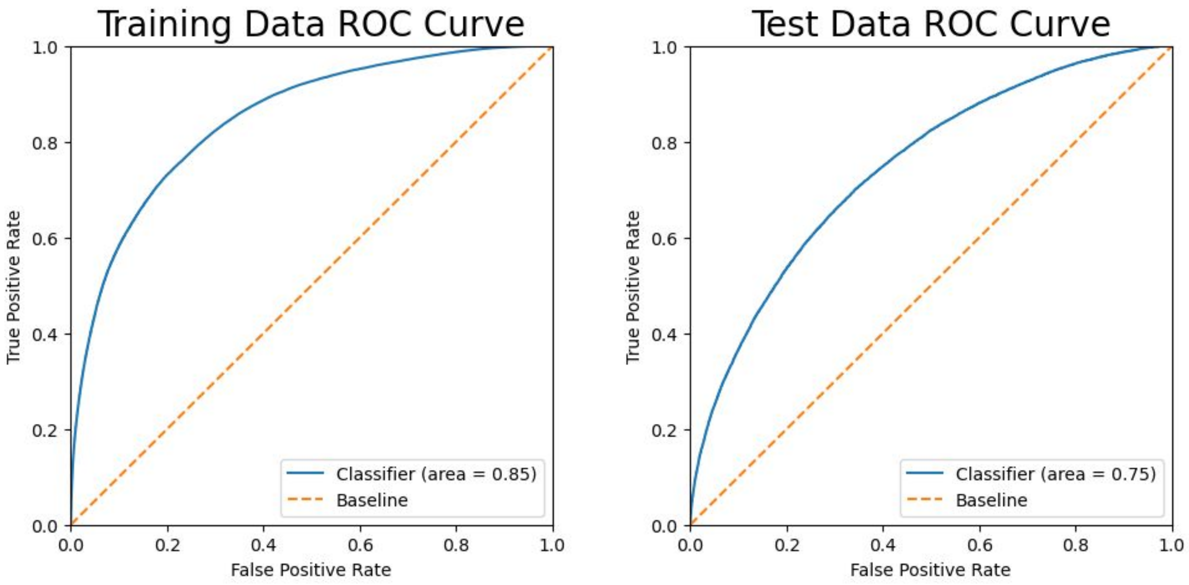
- There is a strong correlation between the target gift amount and the **last donation amount (MRG_amount).**
- Individuals who frequently participate in **cultural and arts events** tend to donate larger amounts.

The Nature Conservancy

Phase 1 Model Process

Feature Selection	Hyper-parameter tuning	Model Training	Generate Predictions
25 features are chosen for Phase 1.	Hyper-parameter tuning is done through a grid search function. Best model is chosen by AUC.	The following model is fitted with our undersampled training data: Random Forest Classification[2]	Remaining 30% of the data is used to test model predictions and results are recorded.

Model 1 Performance



Top 5 Predictor Variables

- Gift Behavior A: All
- Monthsback
- gift_count
- cumul_amount
- append_competitors

The Nature Conservancy

Phase 2 Model Process

Addressing Outliers

Boxplots are drawn to view the presence of outliers in the data.

A target variable threshold of 250 is set and data is changed accordingly.

Feature Selection

14 features are selected for Phase 2.

Model Training

2 models are fitted with our positive response training data:

Random Forest Regression[3]

Gradient Boosting Regression[4]

Remediation, Re-train

We identify and plot residuals.

Removed residual outliers from the data, and train the models an additional time.

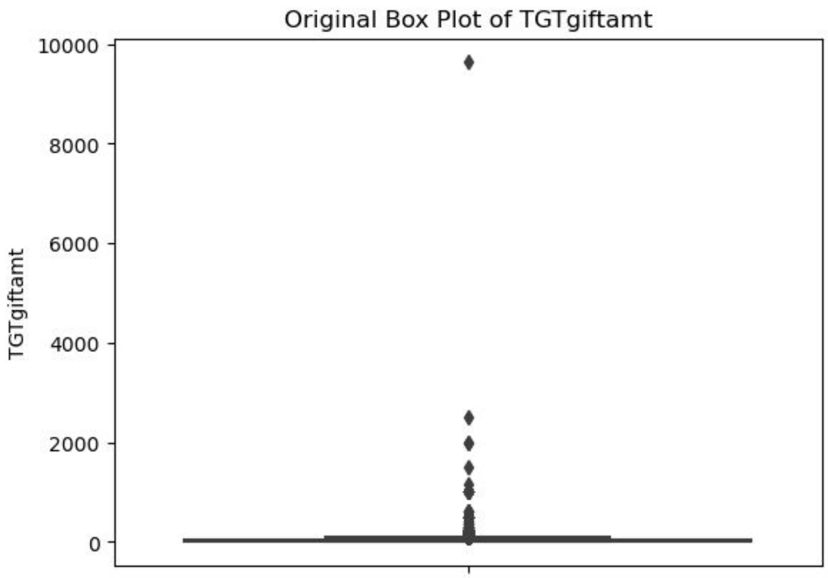
Generate Predictions

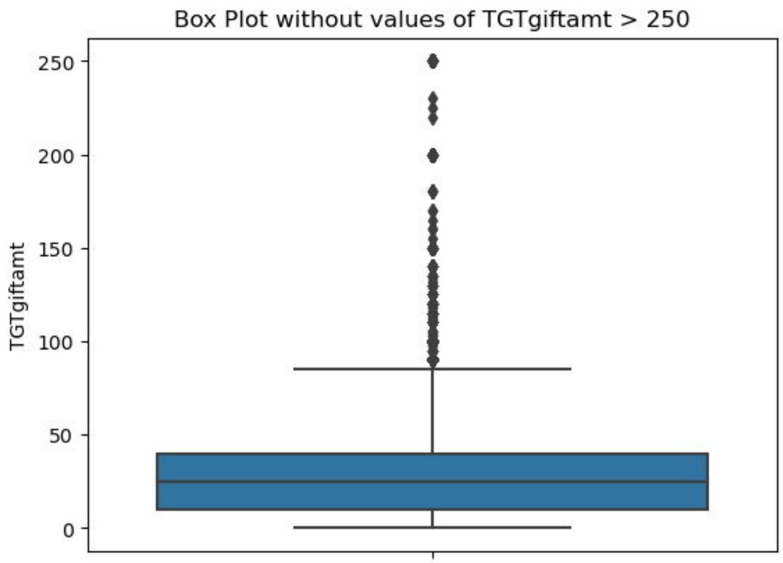
Remaining 30% of the data is used to test models predictions.

The models are assembled by prediction result averages to reach our phase 2 results.

The Nature Conservancy

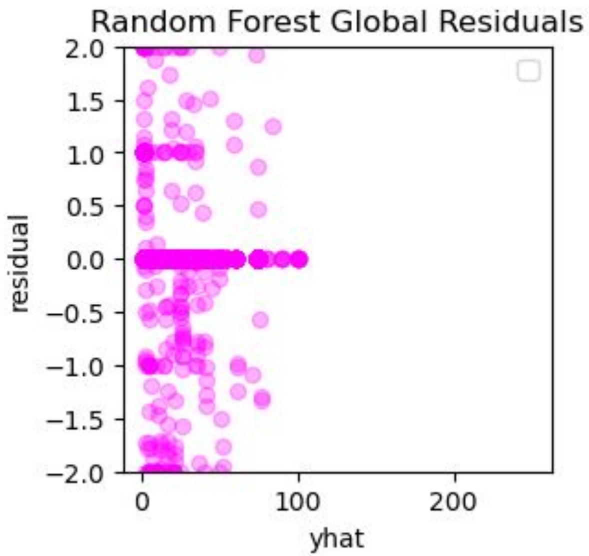
Model 2 Outlier Visualization

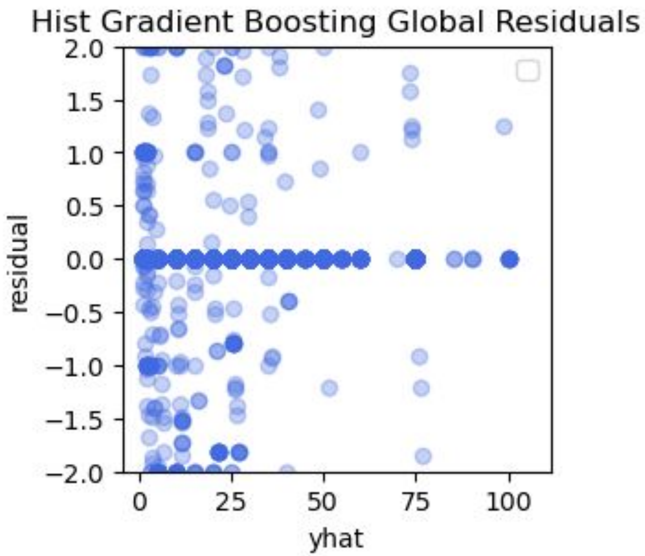




The Nature Conservancy

Model 2 Performance Pt. 1





Final MAE results*:

Random Forest: 29.45

Decision Tree: 29.32

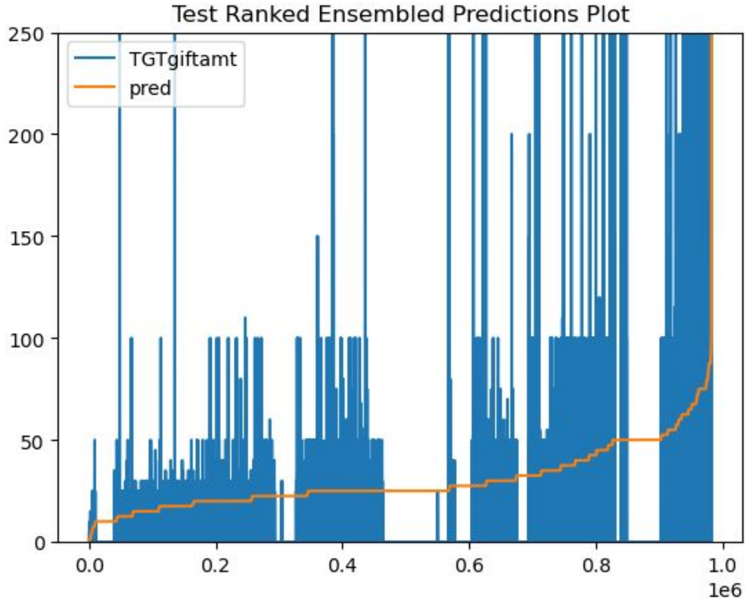
* MAEs above calculated with test data used for final result predictions.

Remediation steps performed:

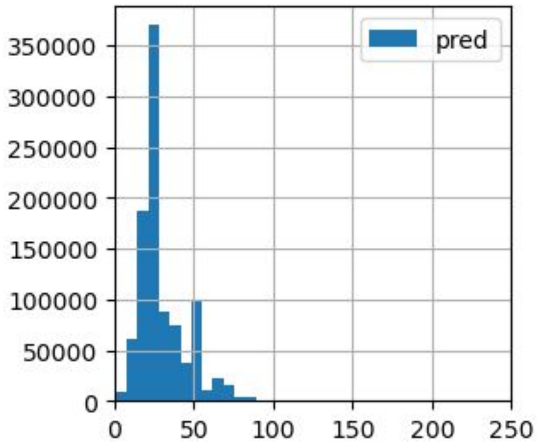
- Quantile feature transformation.[5]
- Data remediation by way of removing rows with some of the largest residual outliers.

The Nature Conservancy

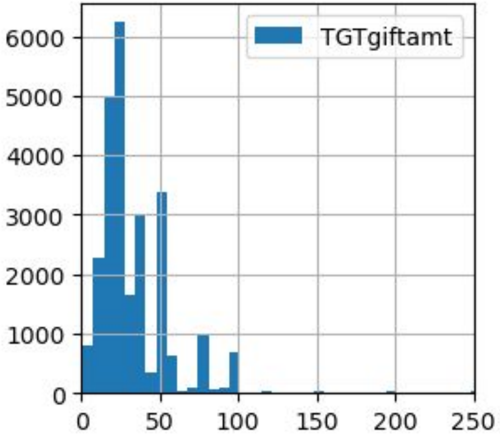
Model 2 Performance Pt. 2



MAE: 29.38



Distribution of gift amount predictions from final validation data

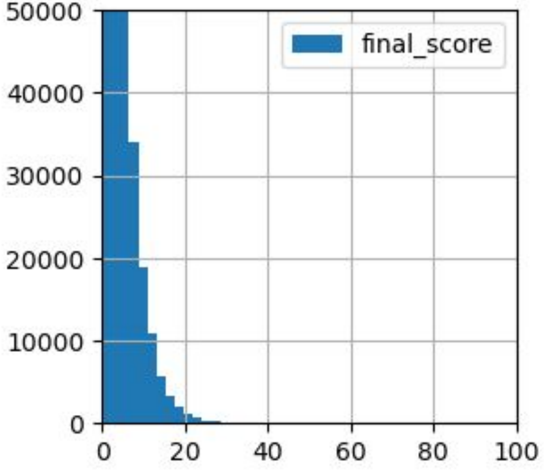


Distribution of TGTgiftamt from phase 2 validation data

The Nature Conservancy

Formula: *classifier['predict_proba'] * regressor['predict']*

Results



Final score distribution*

*Y axis cut off for visualization purposes

Decile	Responses	% of Total
1	4053	27.1%
2	2607	17.5%
3	2056	13.8%
4	1618	10.8%
5	1264	8.5%
6	1016	6.8%
7	842	5.6%
8	608	4.1%
9	512	3.1%
10	353	2.3%
Total	14929	100%

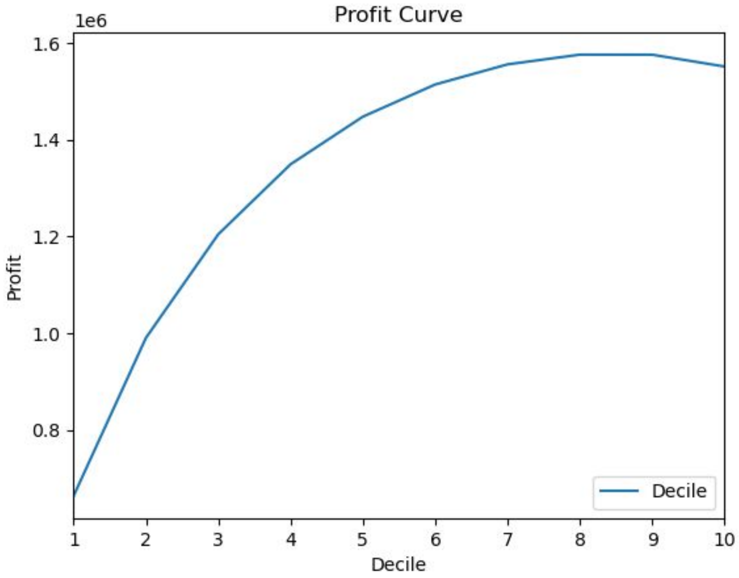
Final output decile analysis*

*Duplicates were removed for decile analysis (highest rank kept).

The Nature Conservancy

Profit Visualization

These results suggest that the maximum profit potential will be reached by sending appeals to the top 80% of the final score ranking.



* These results assume a cost of $0.80 per mailed appeal.

Potential Next Steps

- Future enhancements to the models should focus on handling outlier donors. This model's output is limited to predicting expected gift amounts within our min and max ($0.00-$100.02) after accounting for expected gift amount and probability of response.
  - In our original training data, analysis found that outliers were gift amounts over $72.00.

- Conduct a zero-inflated regression to effectively address the excess of zero values in the model due to non-respondents.

- Explore integration with other fundraising channels, such as online campaigns or events.
  - Conduct further analysis on lower deciles to determine if they are more responsive to other donation channels.

- Continuously monitor and update the model to adapt to evolving donor behaviors and campaign dynamics.

The Nature Conservancy

Risk Considerations

- Over-reliance on predictive models may overlook the human element and unique donor motivations.
- Ethical considerations regarding data privacy and transparency in model deployment.
- Potential for unintended consequences, such as alienating donors or reducing engagement if targeting strategies are too aggressive.

GW

The Nature Conservancy

Thank you!

Appendix

1. sklearn.feature_selection.SelectKBest()
2. RandomForestClassifier(class_weight={0: 1.0, 1: 3.0}, max_depth=16,
   min_samples_leaf=19, n_estimators=110, random_state=12345, n_jobs=10)
3. RandomForestRegressor(random_state=12345, n_jobs=N_CORES, max_features=m)
4. HistGradientBoostingRegressor(learning_rate=0.9, max_depth=91, max_iter=60,
   min_samples_leaf=15, random_state=12345)
5. sklearn.preprocessing.QuantileTransformer()

The Nature Conservancy