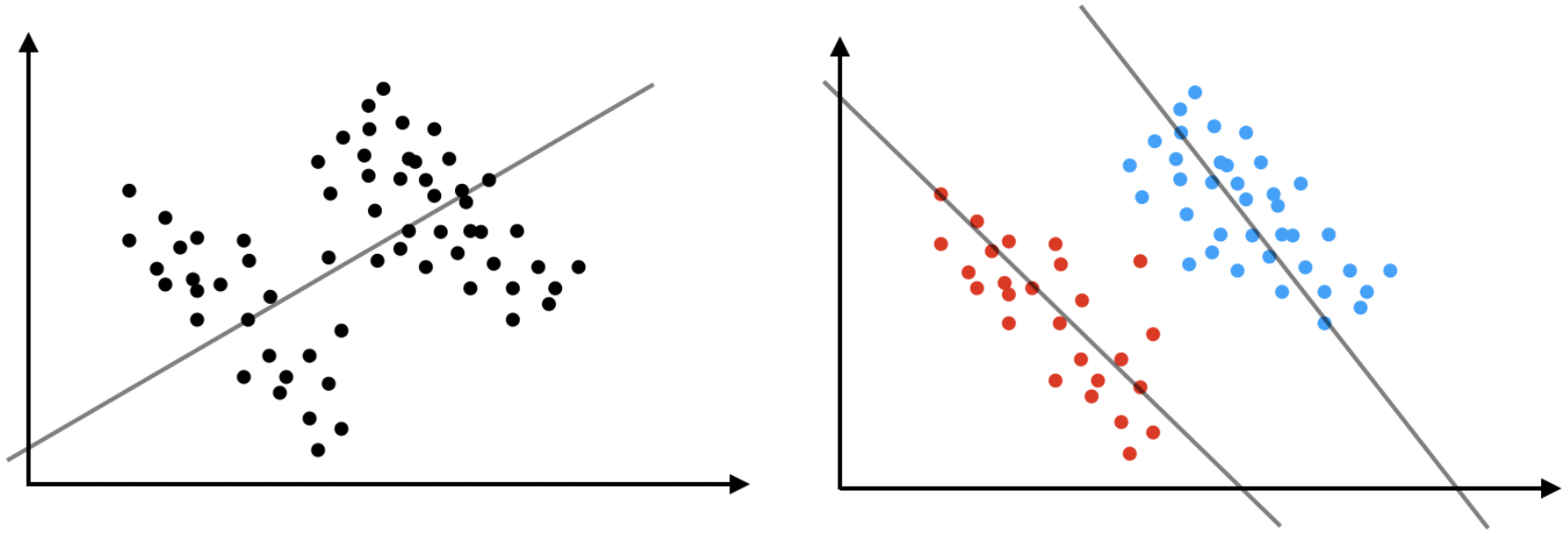
Hybrid of Regression Trees & Linear Regression (HRT)

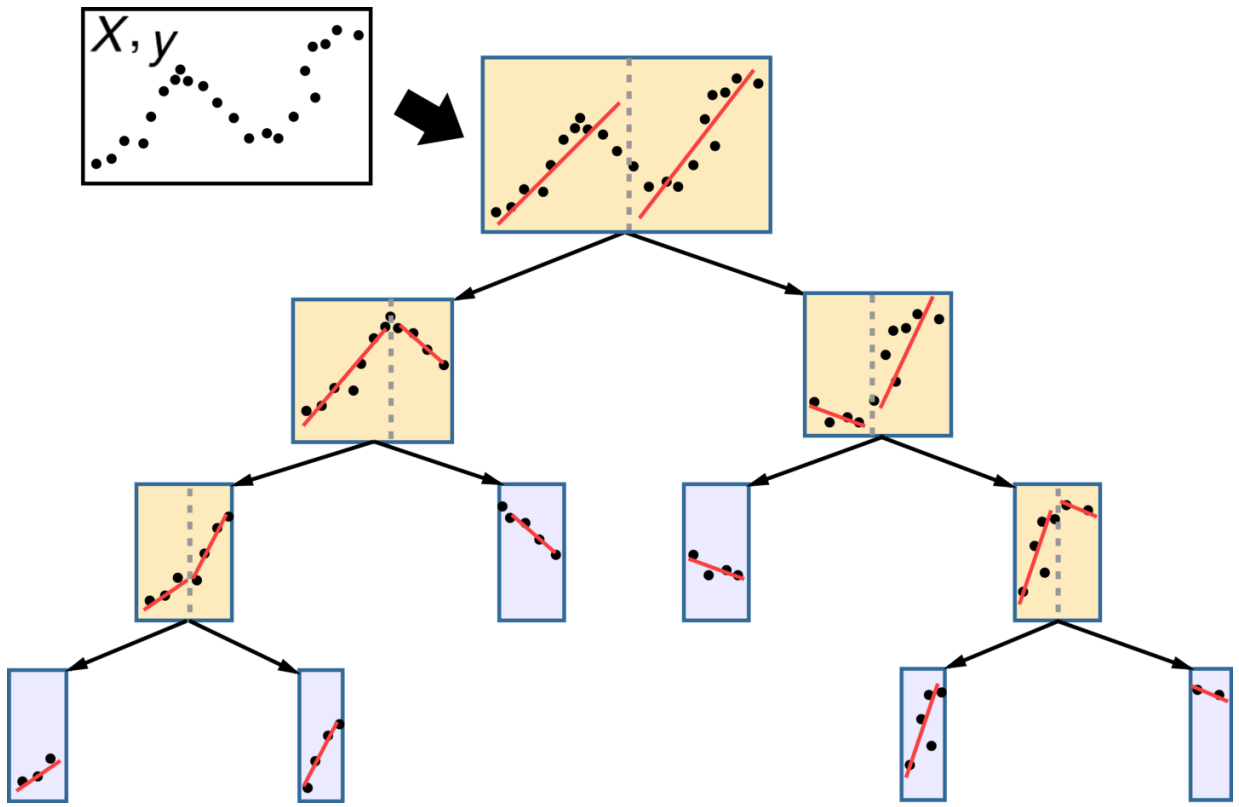
Problem Statement

Simpson's paradox



A result that is present when data is put into groups that reverses or disappears when the data is combined

HRT - General Concepts



How does it work, advantages

*Fits local linear regression in the leaves with alternative impurity measure**

+ Explainability
+ Non-linearity
+ Requires small tree depth
+ Can be run in parallel

- Is trained longer than LR, GBT
- Ensemble Hybrid is not explainable
- Poorly covers feature interaction

**Linear Regression in Regression Tree Leaves. A.Karalik (1992)*

Deep Dive in Theory

Hybrid Regression Tree

- Class value is estimated as **linear function of attributes**;
- Impurity measure of an example set E:

$$I(E) = \frac{1}{W(E)} \sum_{e_i \in E} (y_i - g(\vec{x}_i))^2$$

Function g represents the regression plane through the example set.

- Expected utility of the split - defined in similar way as for CART.

Basic CART

- Class value is estimated as **constant value**;
- Impurity measure of an example set E - estimate of variance of the class values:

$$\sigma^2(E) = \frac{1}{W(E)} \sum_{e_i \in E} w_i (y_i - \mu(E))^2$$

Here w_i is the weight of example. , i
$W(E)$ sum of example weights,
$\mu(E)$ mean class value.

- Expected utility of the split:

$$I_{exp} = p_l I_l + p_r I_r$$

Here p_l, p_r probabilities of transitions into left/right son of the node,
I_l, I_r corresponding impurities.

Datasets

PIK

Goal: Predict flats sales

Metric: RMSE

Shape: (8726, 47)

Data:

- Flats characteristics
- Macroeconomic data
- Geo data

StarSkill

Goal: Predict unique units made per timestamp

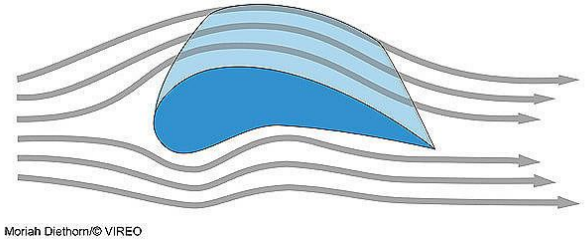
Metric: RMSE

Shape: (3395, 20)

Data:

- Game attributes
- Players actions

Airfoil



Goal: Sound pressure level prediction

Metric: RMSE

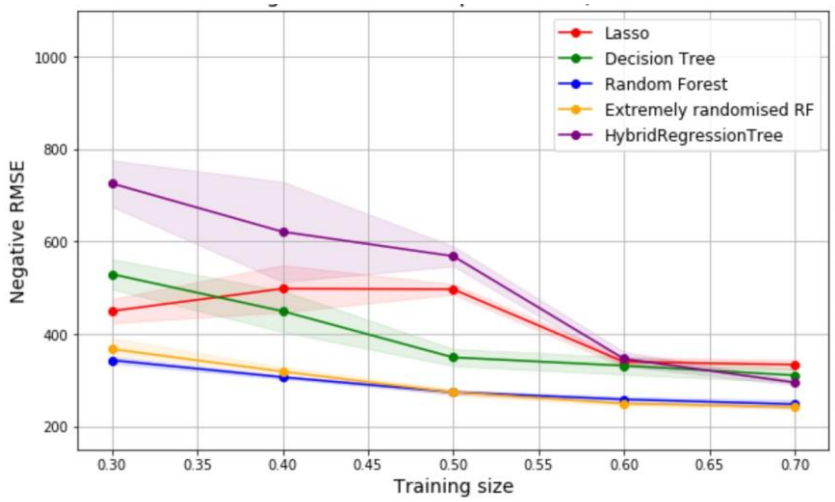
Shape: (1503, 6)

Data:

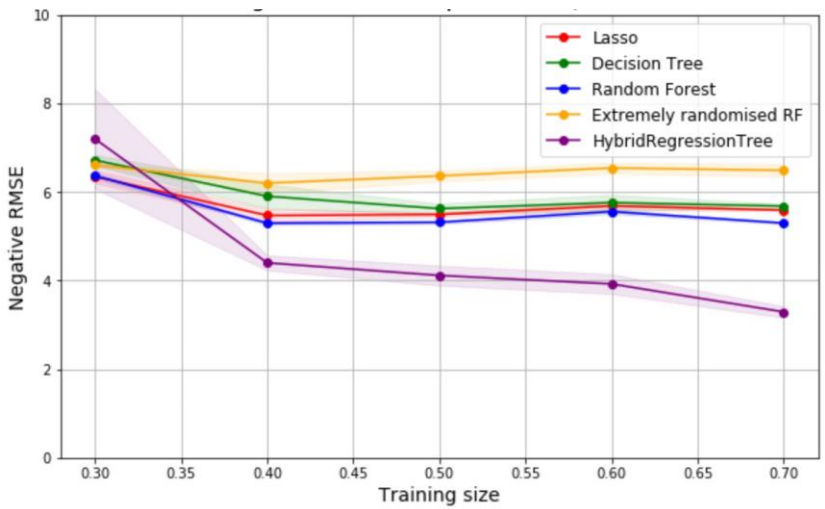
- NASA data, aerodynamics and acoustic tests of airfoil blade sections

Comparison – Forecasting Quality

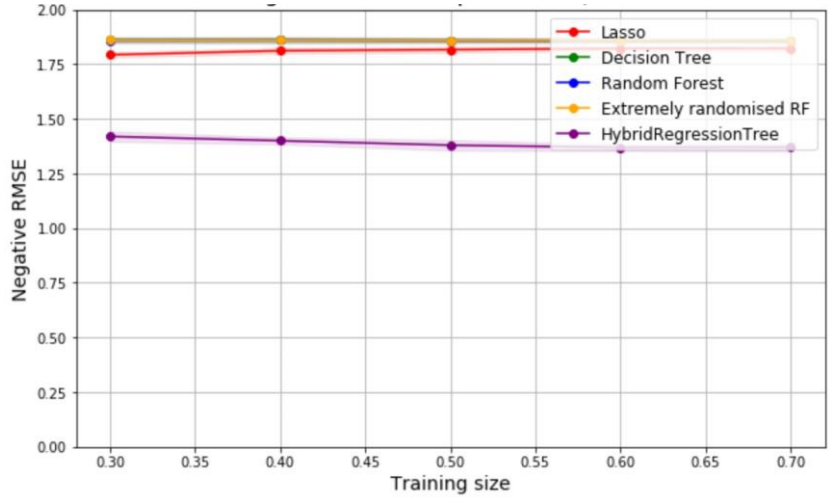
PIK



SkillCraft



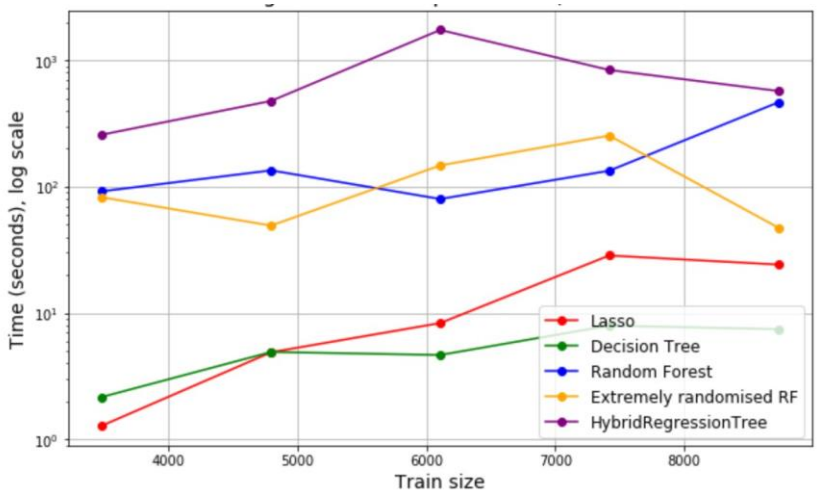
Airfoil



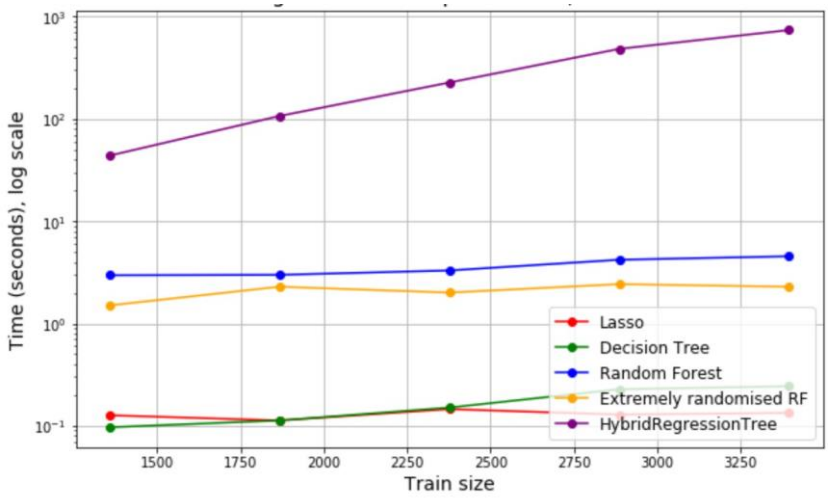
- HRT outperforms other methods for 2/3 datasets
- The better Lasso is, the better HRT is
- HRT is dependent on the number of data it is trained on

Comparison – Training Time

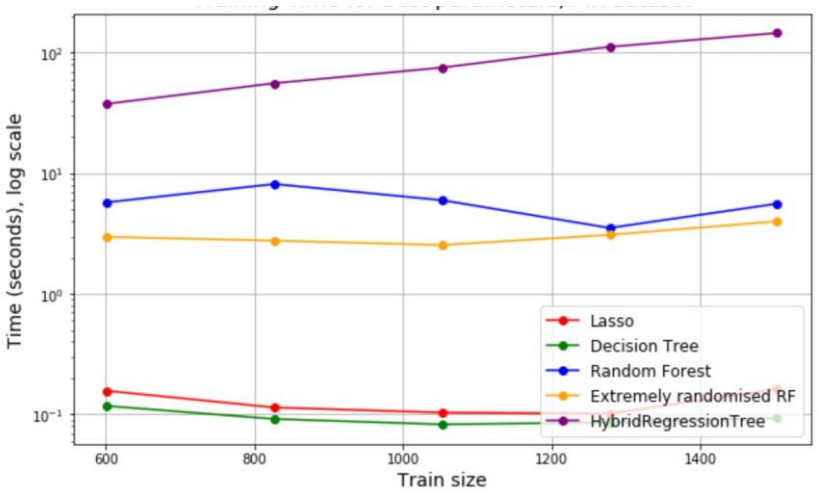
PIK



SkillCraft



Airfoil



- HRT is trained longer than all other methods (as expected)
- Its train time scale worse in train size than other methods
- For larger datasets it could perform better than RF, ERRF

Conclusion & Further steps

Conclusion

- For some cases HRT **outperforms** RF, Extremely Randomized RF and Lasso
- In general, HRT **improves Lasso** results
- HRT captures **non-linear** data structure
- HRT is **sensitive to the number of data** it is trained on
- HRT is **computationally expensive**
- HRT is suitable for middle size datasets

Further steps

- Compare HRT to GBT
- Consider other algorithm in leafs except linear regression
- Implement HRT with parallelization
- Apply HRT to big data sets

Our Team

Input of each team member was focused on, but not restricted to:

Gazdieva Milena:
- HRT theory
- HRT code
- Data cleansing

Maksimov Ivan:
- HRT ensembles
- Pipeline
- Benchmarks

Belikova Kristina:
- HRT code
- Benchmarks
- Results analysis

Nikita Alexeichyk:
- HRT debugging
- Pipeline
- Visualization

Mikhail Kuzin:
- HRT debugging
- Feature engineering
- Visualization