Evaluation of product batches

ML approach

Agenda

1. Data preprocessing
2. Exploratory Data Analysis (EDA)
3. Modeling
4. Most important features
5. What could improve the hitherto analysis

Data preprocessing

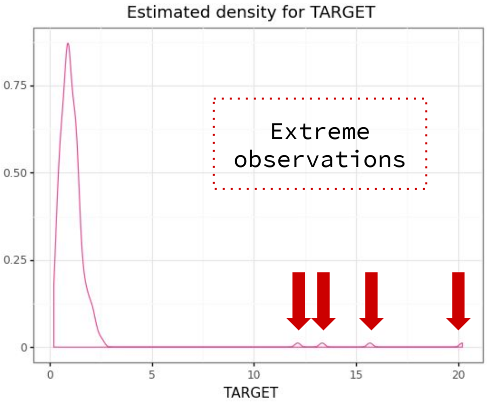
In total, at the modeling stage of the task **82** columns were preserved, among these **65** were used in the training. In the preprocessing of the data we needed to make some difficult decisions on which columns to remove.

Data preprocessing - removed columns

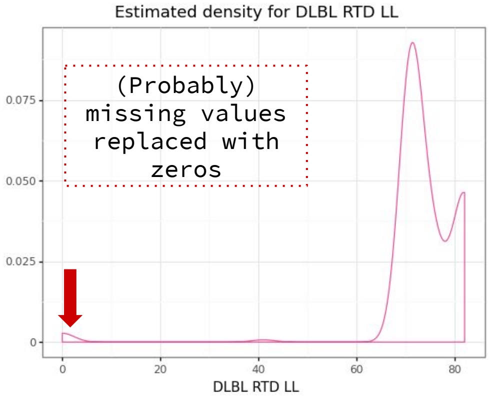
Reason	Num columns removed	Examples
Single value in the column	18	HLAT BASE DIAM TARGET, MLPF LENGTH TARGET
Not enough measurements (<50%)	69	DLBL TP BACK LENGTH ULL, DLBL RTD ULL
Almost no variability (one dominant value)	19	LL TLR LENGTH, LL PLLA LENGTH
Redundancy (correlation of > 0.999)	13	UL MLPF RTD, LL HLAT FRONT DIAM

In the preserved columns, **missing values** were **imputed.**

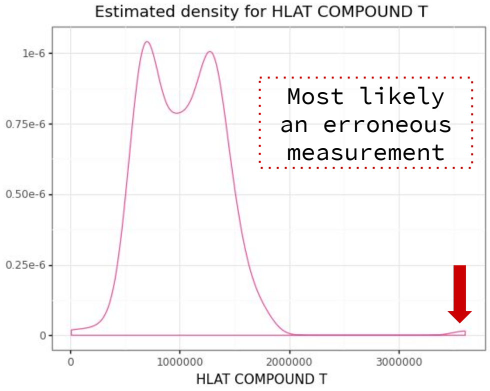
Data preprocessing - interesting cases (part of EDA)


Estimated density for TARGET

Extreme observations


Estimated density for DLBL RTD LL

(Probably) missing values replaced with zeros


Estimated density for HLAT COMPOUND T

Most likely an erroneous measurement

All of these situations were addressed.

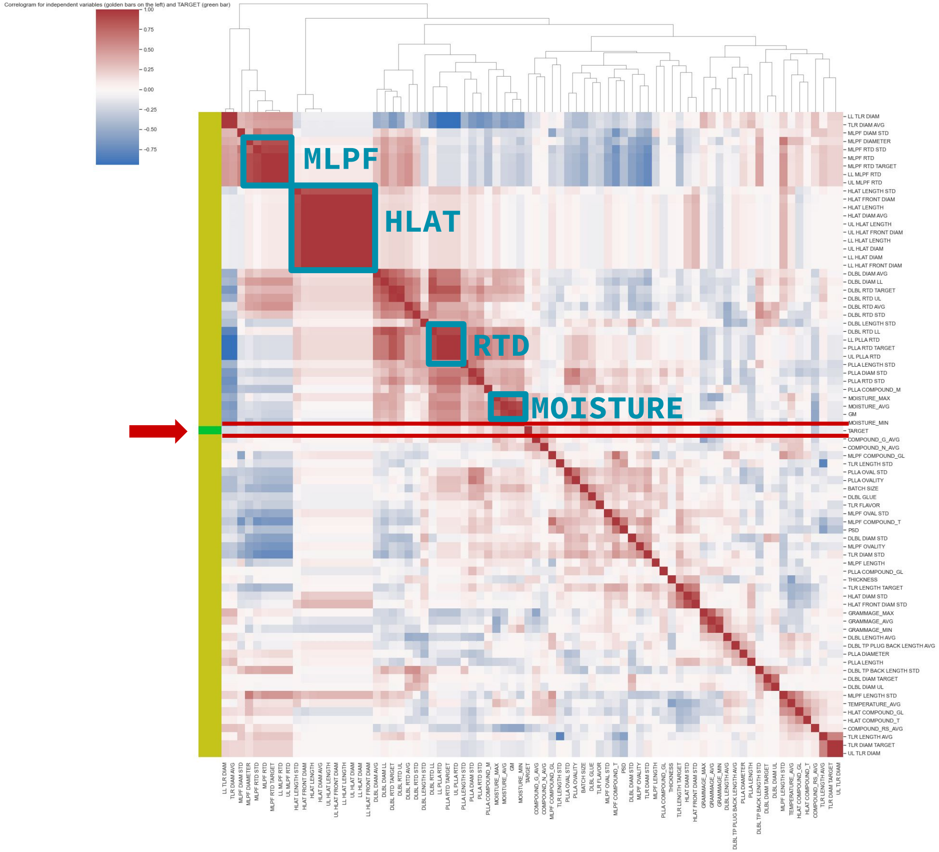
Data preprocessing - removed batches

There were also two types of observations (batches) that we had to discard from the analysis.

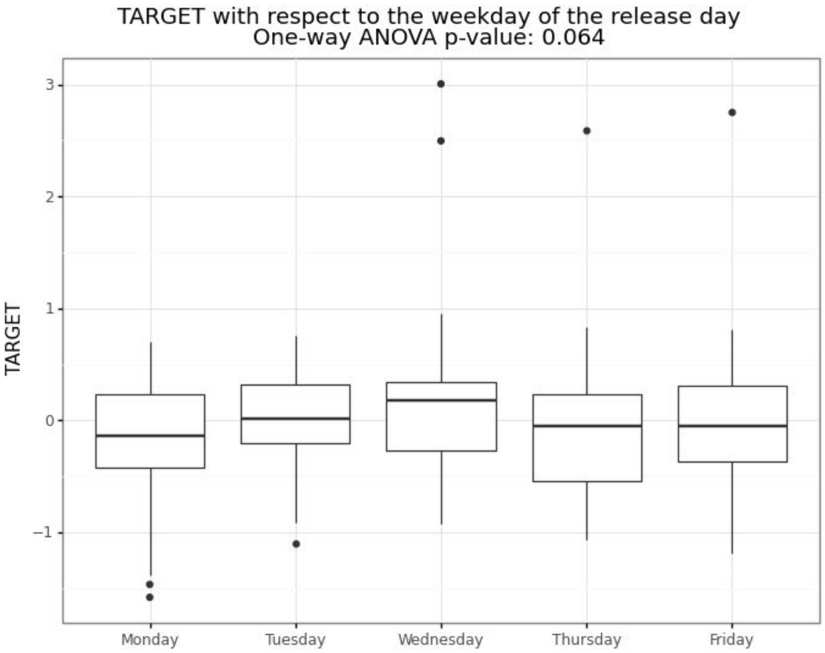
- 5 observations that had an atypical pattern of measurements for a number of columns (all the values were zeroes and these were the only zeros in these columns)
- 5 observations were deemed as outliers by an outlier-detecting algorithm (Isolation Forest)

Exploratory Data Analysis

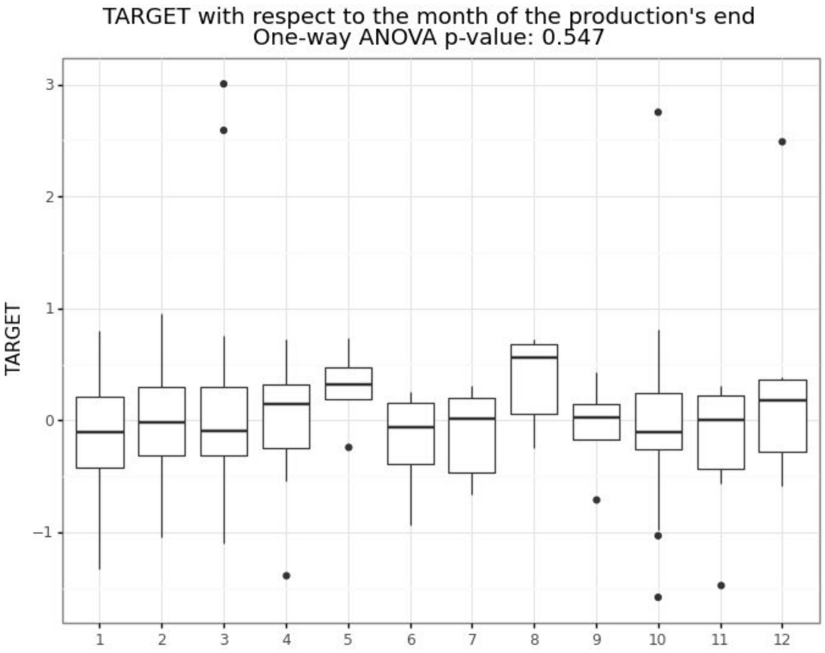
- Groups of highly correlated variables.
- Low correlation of TARGET with other variables.



Exploratory Data Analysis - temporal aspect



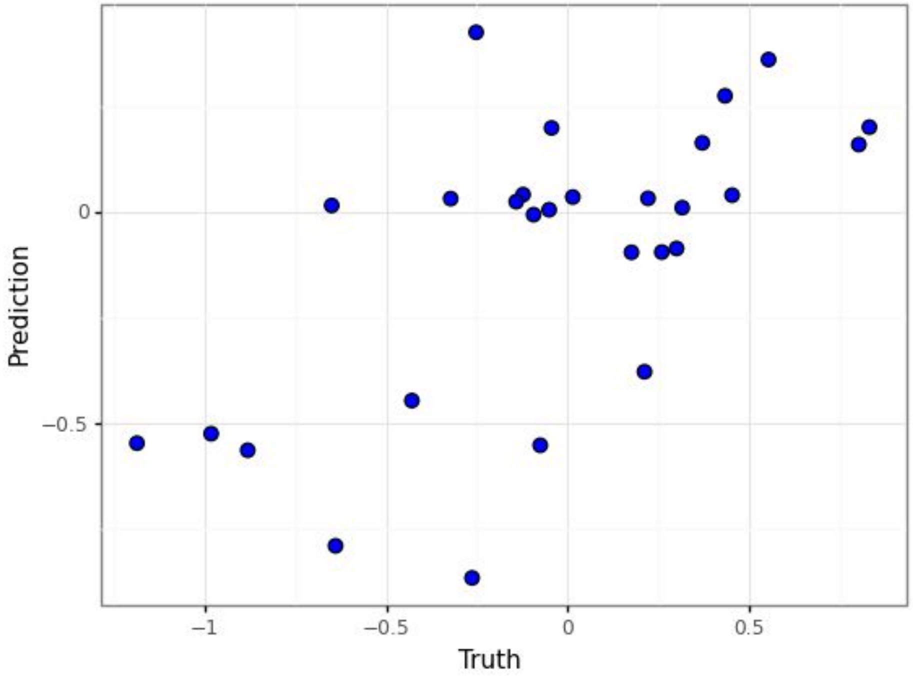
TARGET with respect to the weekday of the release day
One-way ANOVA p-value: 0.064

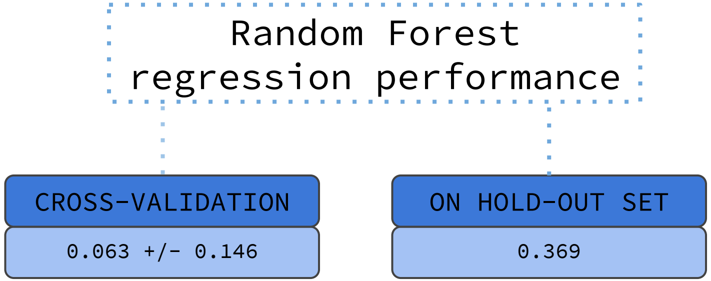


TARGET with respect to the month of the production's end
One-way ANOVA p-value: 0.547

Modeling (predicting the TARGET)

Predictions from random forest on holdout set vs. true values



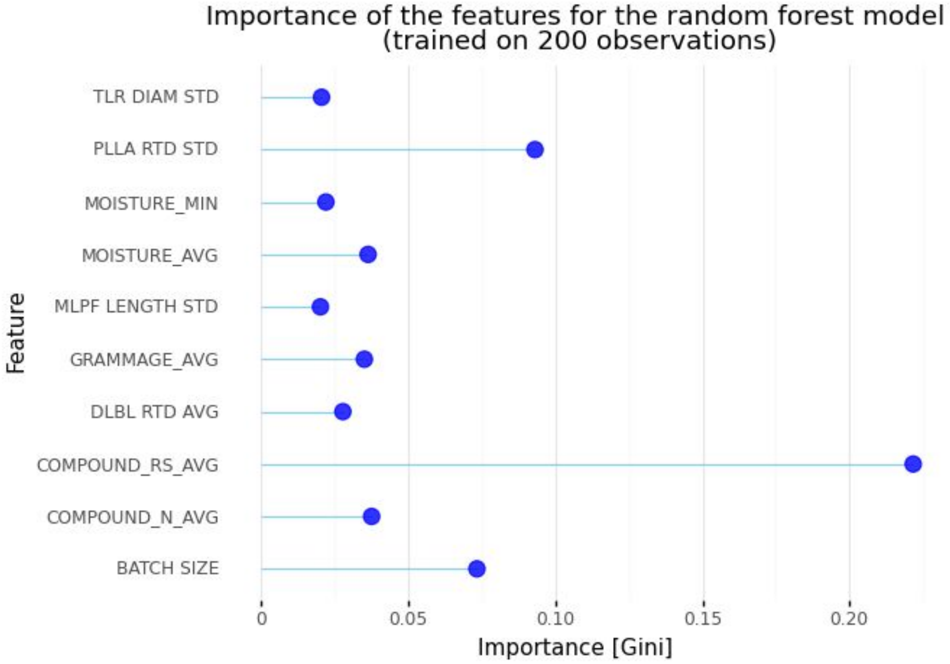


Random Forest regression performance

CROSS-VALIDATION	ON HOLD-OUT SET
0.063 +/- 0.146	0.369

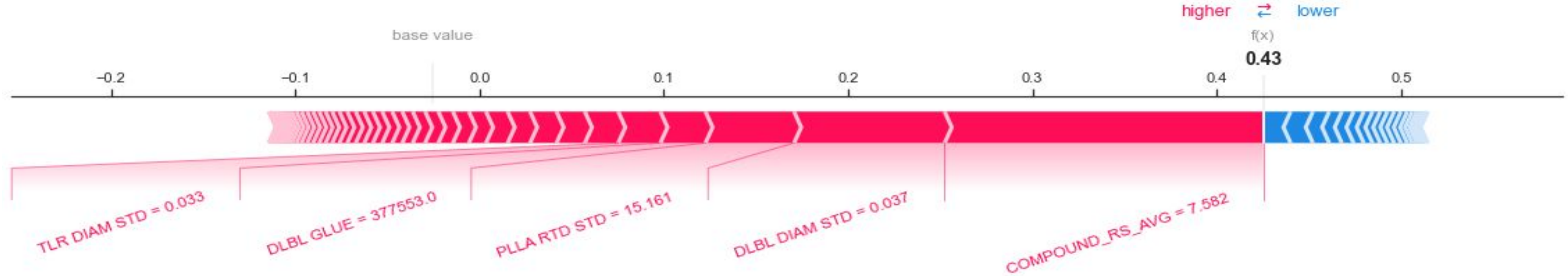
Such a discrepancy between the performance in cross-validation and on hold-out set may be indicative of an **insufficient testing procedure** - it needs to be conducted more thoroughly.

Feature importance at model's level

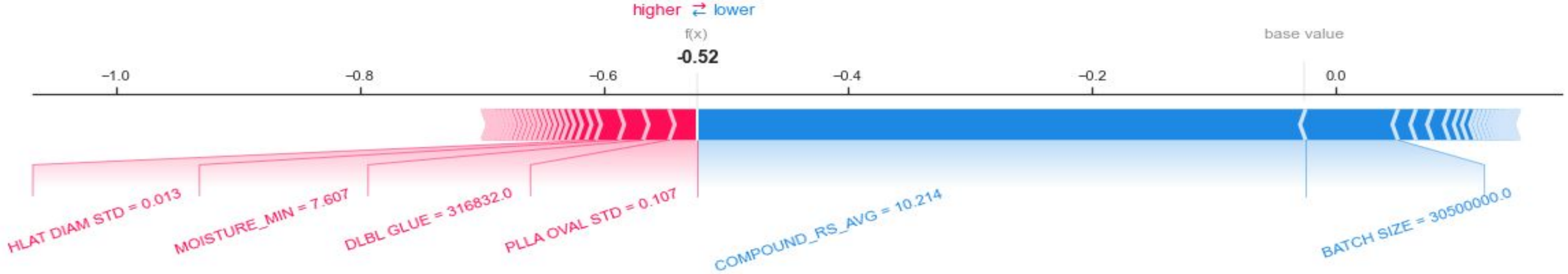


* 10 most important features are plotted

Feature importance at observation’s level



A batch with high TARGET value



A batch with low TARGET value

Future work

- **First and foremost - hitherto work requires consultation with domain experts**
- Other than that, we should have a more careful look at the preprocessing
- More strict model’s evaluation procedure
- Modeling of rejection/acceptance (with threshold for TARGET of 1.3) - it is an easier problem and result a fruitful approach
- Extended EDA