Challenge
ML Predictor Insulet

Yi Hernández

May 20th 2025

Objectives

- **Build ML Predictor to predict the target variable**
- **Reduce RMSE from AutoML**
- **Describe**
    - **Data Processing**
    - **Model Selection**
    - **Model Training**
    - **Model Evaluation**
    - **Feature Selection**

- **Deliver UI app to run the model:**

The running regressor is deployed in streamlit, the link for the app is:
https://interviewregreappr-mmgztdc72afbn9zsxc5gfy.streamlit.app/

Project github: https://github.com/cihernand/interview_regressor

Data Processing

Many Dates are repeated

Total dates repeated two or more times: 322

Date variable was transformed into:

Year - Numeric

Month name - Category

Day number - Numeric

Day name - Category

Categorical Variables were transformed with One-Hot Encoding

Numerical Variables were in different magnitudes, thus they were transformed to a normal distribution with

mean = 0

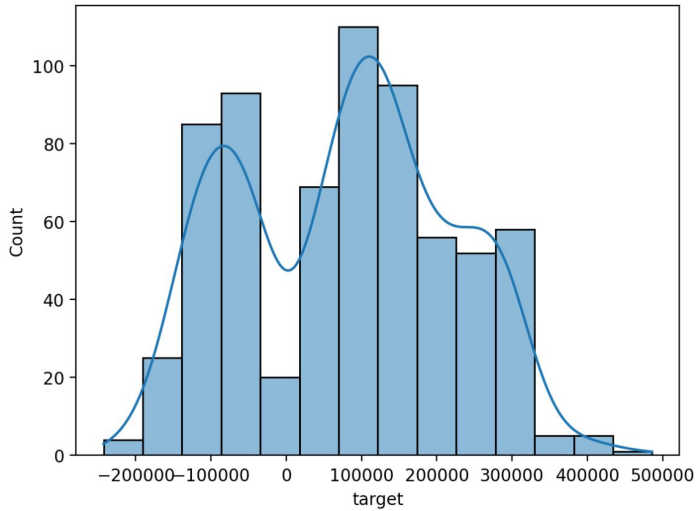
standard deviation = 1

Data leakage avoided by doing transformations separately for training and testing datasets.

Target Variable

Numerical Variable with a trimodal distribution

n = 678



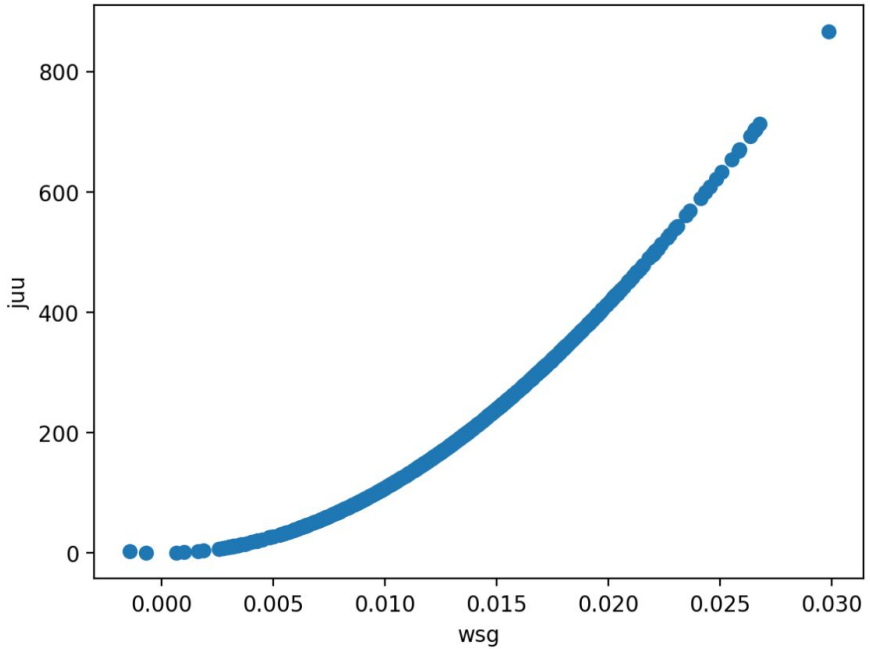
No correlation with Input Features

```
bar           0.058495
baz           0.008524
xgt           0.059559
qgg          -0.015258
lux           0.033065
wsg          -0.000386
yyz          -0.014862
drt           0.001962
gox          -0.011707
foo          -0.011086
boz           0.033299
fyt          -0.011562
lgh          -0.039907
hrt          -0.036169
juu          -0.005196
target        1.000000
date_year    -0.089509
date_day      0.108495
Name: target, dtype: float64
```

Search of multicollinearity

Variables juu and wsg are correlated 0.97

Variable wsg was excluded



Training and Testing models

Models selected included Linear regression and Ensemble models with decision trees.

Decision trees tend to do overfitting in the model training , thus the training dataset was splitted into X_training and X_testing to evaluate the models performance and monitor the presence of overfitting .

X Training data 80% (542 rows)

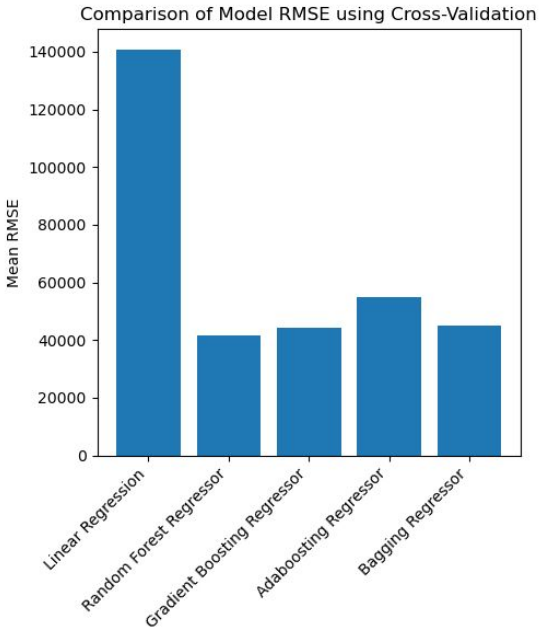
The model was trained with 5 fold cross validation scoring by the negative mean squared error and the r2 score

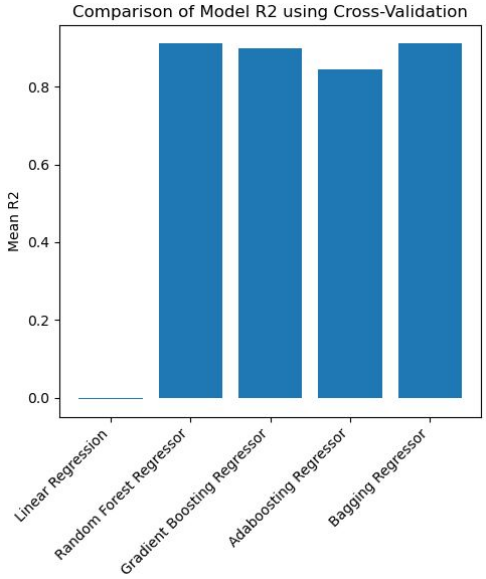
The random state was 42 to get reproducible results.

X Testing data 20% (136 rows)

Models performance using all features

Number of Decision Trees = 100





Best Model Random Forest

RMSE:

Cross Validation Training Mean 41580.54

Testing set

29315.81

R2:

Cross Validation Training Mean 0.91

Testing set 0.95

The model performs better in unseen data suggesting no overfitting

Feature Importance Random Forest regressor

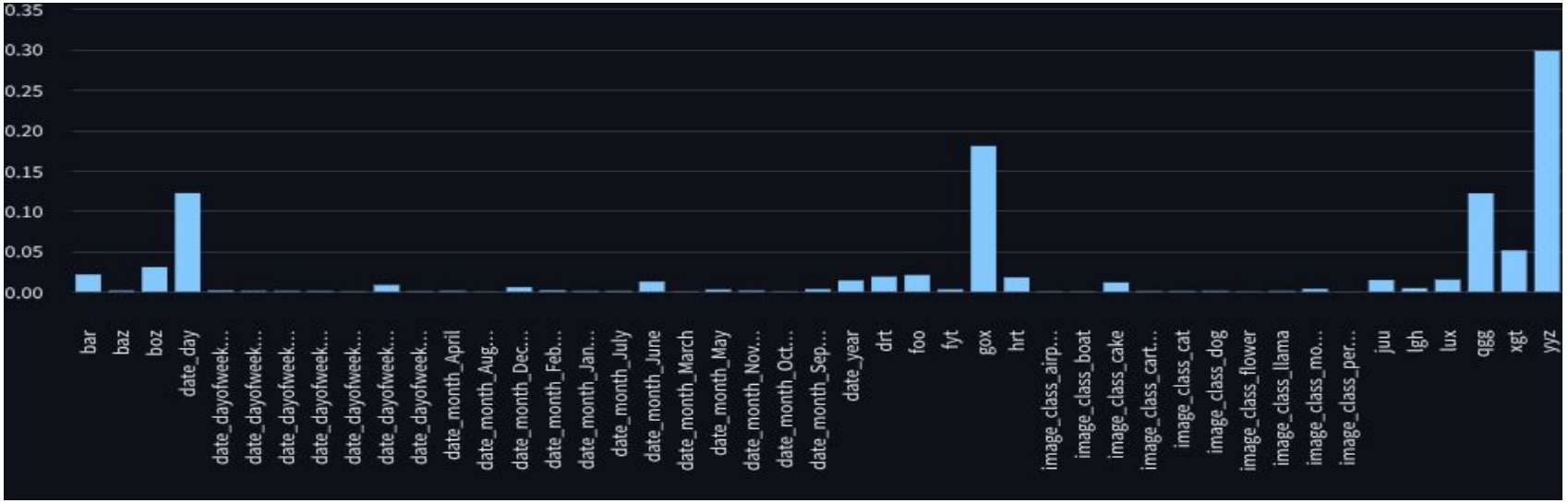


Image class is not relevant for the predictor.

The numeric day of the date is among the top features.

Features selection

A second model was build filtering by feature importance. Threshold >= 0.10

Features selected: qgg, yyz, gox, date_day

RMSE:

Cross Validation Training Mean 53685.81

Testing set 23575.80

R2:

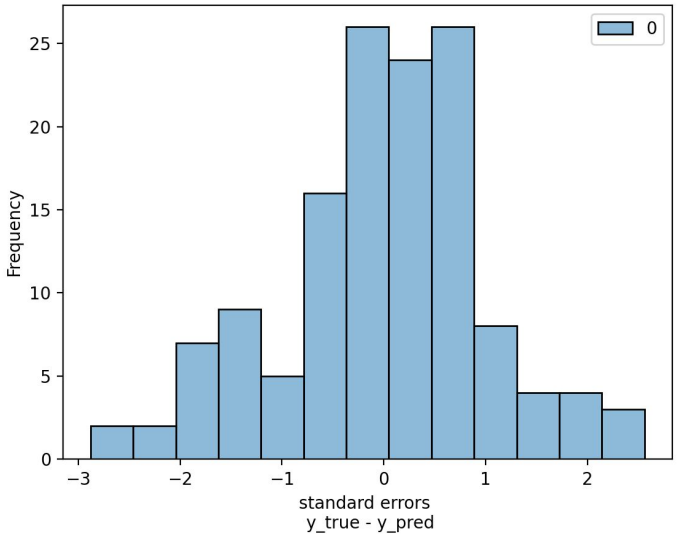
Cross Validation Training Mean 0.85

Testing set 0.97

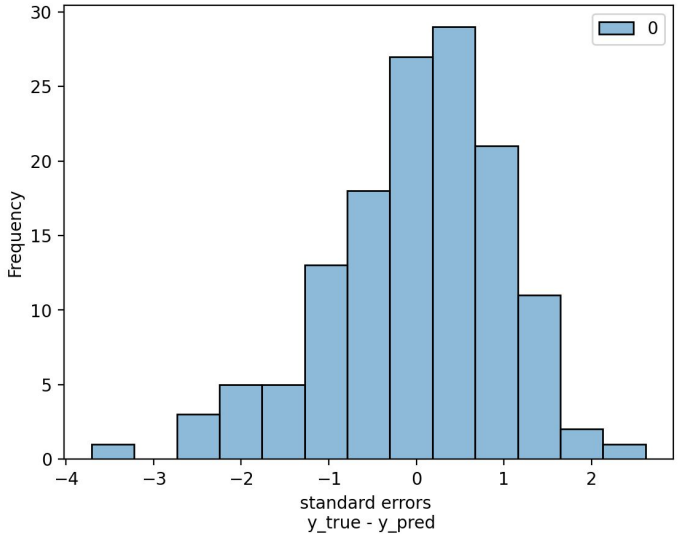
This second model performs better in the testing set

Standard Errors

The standard error distribution of the model with selected features has higher frequencies within +/- 1 standard deviation.



Model all features

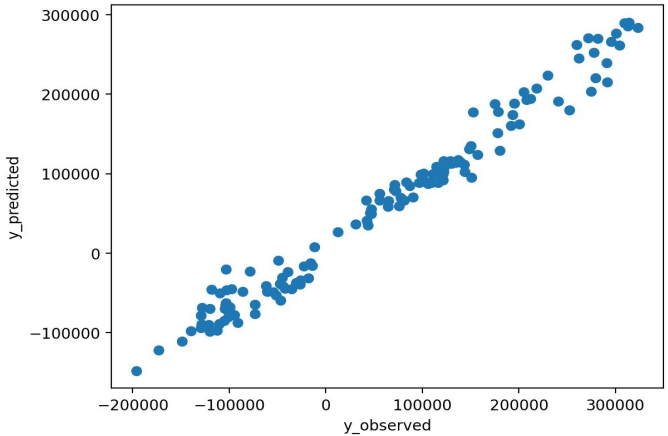


Model selected features

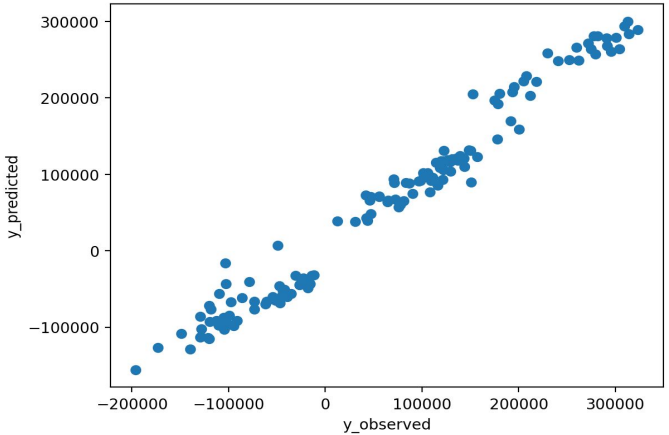
Observed vs Predicted

The predictions in the model with selected features splits into three subgroups:

values <0, values between 100,000 and 150,000 , and values > 200,000



Model all features



Model selected features

Conclusions

Model Comparison

RMSE Testing set:

All Features 29315.81

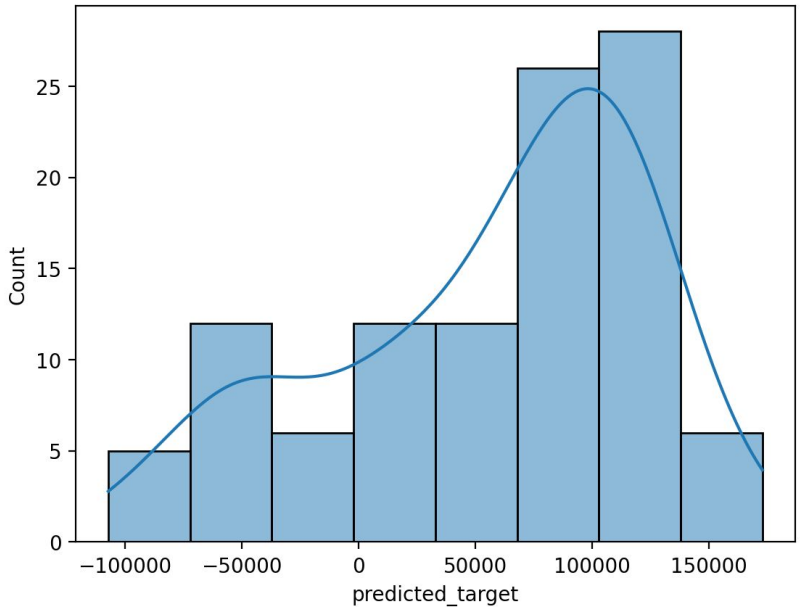
Selected Features 23575.80

R2 Testing Set:

All Features 0.95

Selected Features 0.97

Based on Model evaluation metrics on the Testing Set, the best model is the Random Forest with selected features: qgg, yyz, gox, date_day. The image class was not relevant.



Predicted Values

Next Steps

It is advised to do a grid search across the Random Forest parameters such as the number of trees and max_depth.

Define a drift value to evaluate the performance of the best model and re-train it with new data if RMSE increases and R2 decreases.

Thanks!