

# Our model

We used xgboost to minimize the error: a technique that allows to combine an ensemble of weak predictors (decision trees). It uses a more regularized model formalization to control over-fitting



## ***Model performance:***

20% test set hold out evaluation

**Prediction Error** 3.18%

**R<sup>2</sup>** 95.71%

**R<sup>2</sup> month** 99.23%

We also tried to make a Random forest per Store, since each Store may have its own characteristics, with results comparable to a more generalist approach

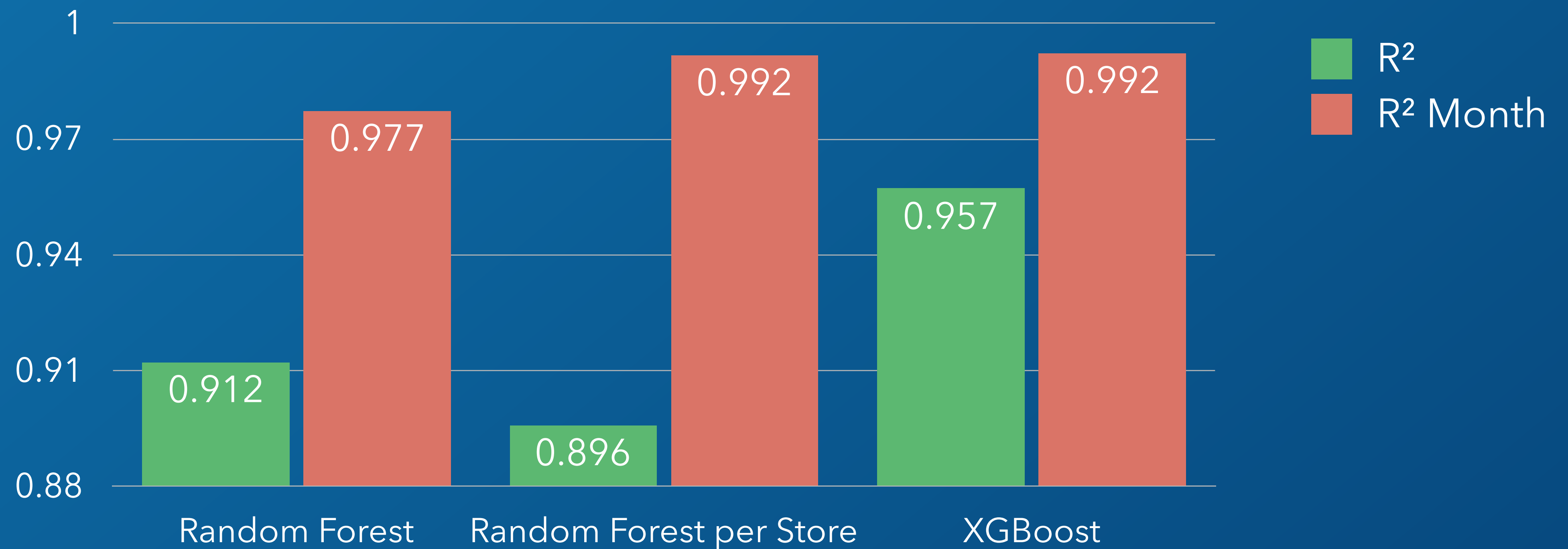
# Preprocessing

- 1 Dropped NumberOfCustomers  
*Would convey information that couldn't be used for the predictions*
- 2 Transformed categorical variables with One-hot Encoding  
*For the following features: Event, AssortmentType, Storetype*
- 3 Split Date into three different features  
*Allowed to compute the predictions per Month*
- 4 Scaled temperature to Kelvin  
*To reduce eventual errors with negative temperatures*



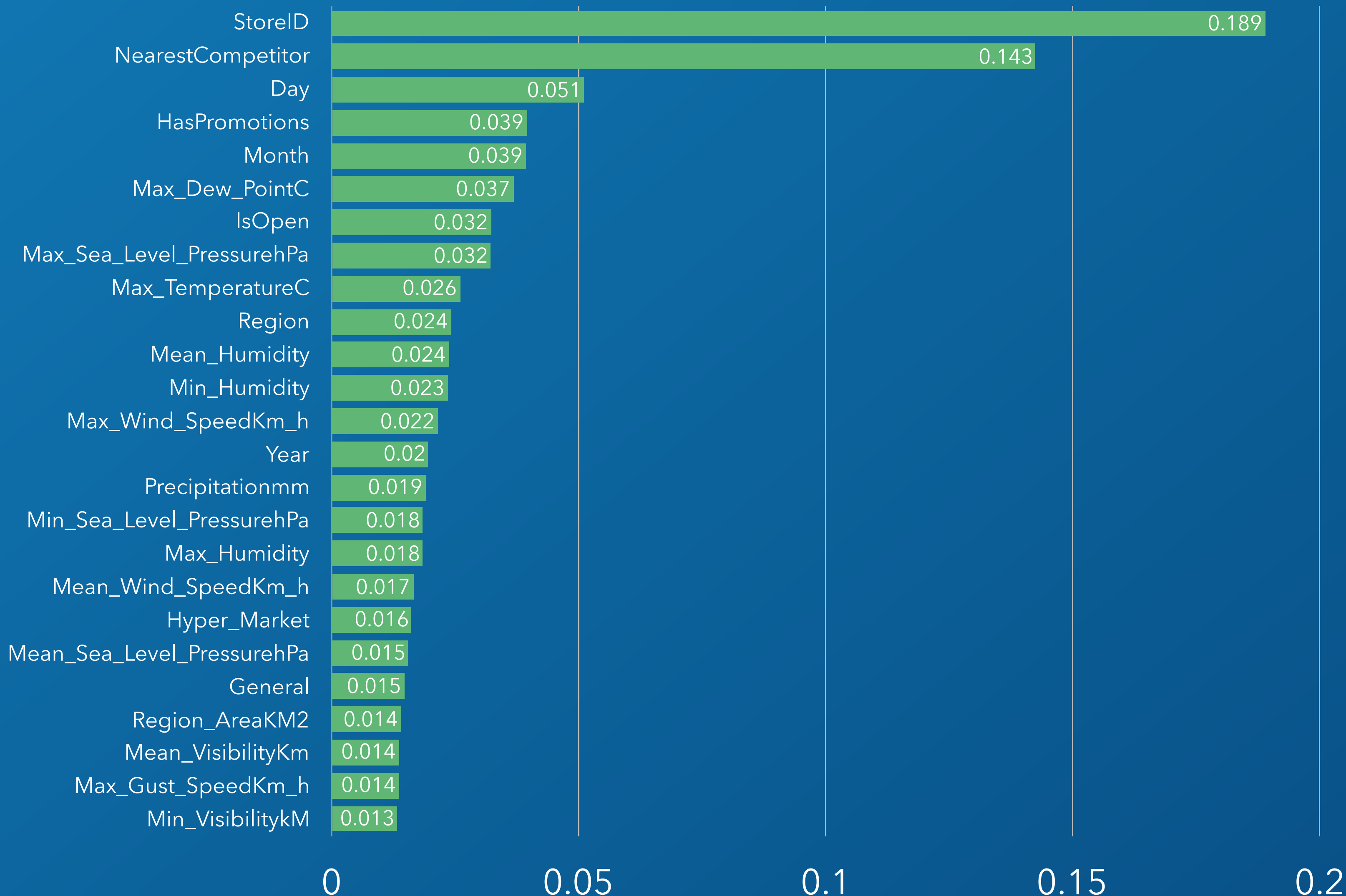
# *$R^2$ of models*

All regression models were made by predicting daily Number of Sells and summing the prediction per month. This explains why  $R^2$  Month is higher than  $R^2$ : sum of single prediction allows error reduction





# Features Ranking for xgboost



The most important feature seems to be the StoreID.

Weather information are not important as other features