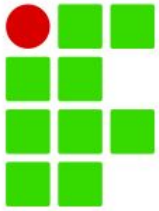
2nd KDD BR Competition 2018

Predicting palm oil production

Solução do time TSI

Rafael Alencar, João Paulo de Melo, Guilherme Domith
IF Sudeste MG, Campus Barbacena

Time TSI - Tecnologia em Sistemas para Internet



INSTITUTO FEDERAL
Sudeste de Minas Gerais | *Campus* **Barbacena**

Rafael Alencar <rafael.alencar@ifsudestemg.edu.br>
Professor/pesquisador
Instituto Federal do Sudeste de MG, Campus Barbacena

João Paulo <jpmdik@gmail.com>
Bolsista de IC no curso de Tecnologia em Sistemas para Internet
Instituto Federal do Sudeste de MG, Campus Barbacena

Guilherme Domith <guilhermedomith@gmail.com>
Bolsista de IC no curso de Tecnologia em Sistemas para Internet
Instituto Federal do Sudeste de MG, Campus Barbacena

Desafio

- Desenvolver um modelo preditivo para a **produção de óleo de palma**.

- Desafio de **regressão**.

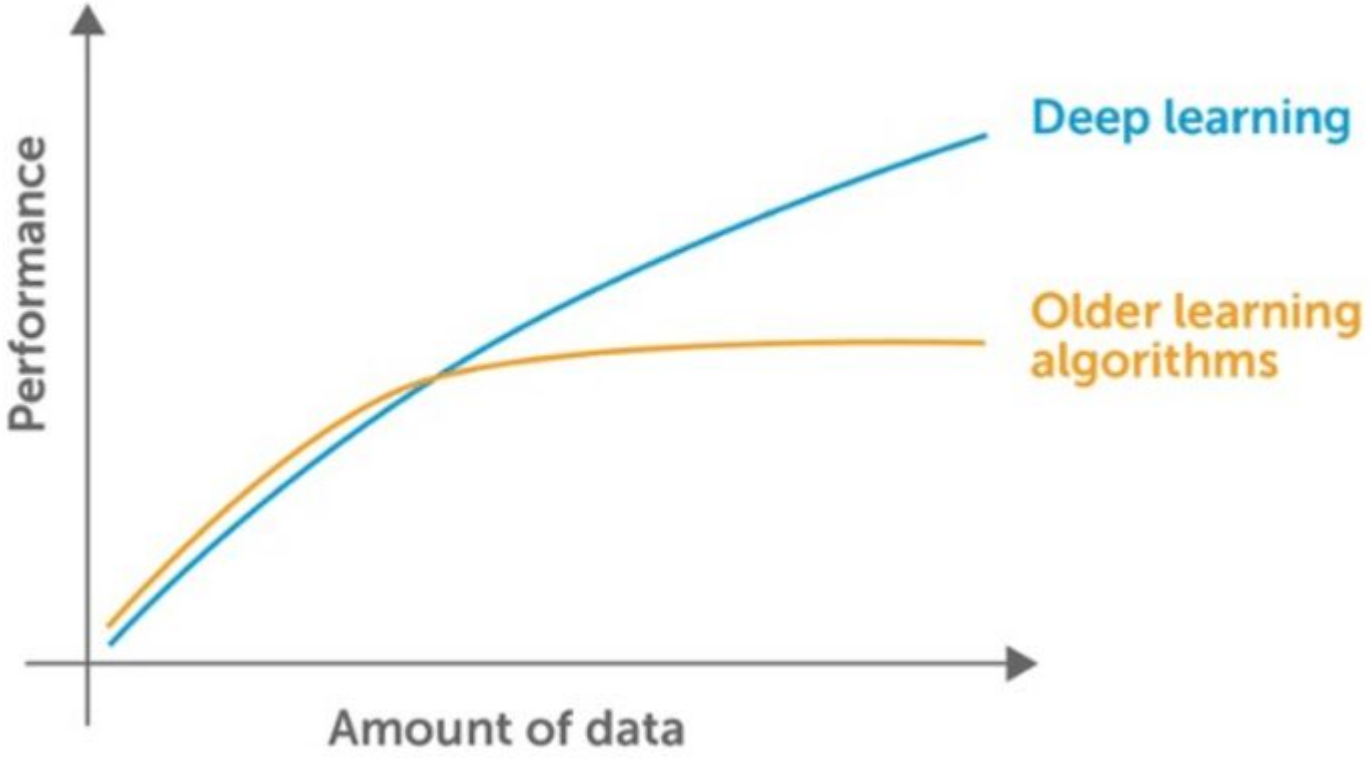
- Dados das plantações fornecidos pela AGROPALMA.

- Dados de solo (via SoilGrids).

- Dados atmosféricos (via ERA-Interim reanalysis dataset).

Escolha do modelo

Devido ao **tamanho reduzido do dataset**, optamos por utilizar **modelos tradicionais** de machine learning, não avaliando modelos baseados em deep learning.



Fonte: http://www.andrewng.org/

Escolha do modelo: XGBoost

- Modelo baseado em árvores, utilizando a técnica de **Gradient Boosting**.

- Amplamente utilizado em competições no Kaggle.

- Bom *trade-off* entre acurácia e tempo de treinamento.

- Hiperparâmetros para **controle de overfitting** como profundidade máxima das árvores e regularização *gamma*.

Dataset (treino)

	Id	field	age	type	harvest_year	harvest_month	production
0	0	0	19	5	2004	1	0.064071
1	1	0	19	5	2004	2	0.047658
2	2	0	19	5	2004	3	0.016866
3	3	0	19	5	2004	4	0.025525
4	4	0	19	5	2004	5	0.047690

Dataset: arquivos adicionais

Experimentos combinando os **dados de solo** soil_data.csv e **dados atmosféricos** field-*.csv com o dataset inicial:

- Correlação de *features*

- Importância de *features*

- Seleção sequencial de *features*

- Redução de dimensionalidade (PCA, NMF, etc.)

Estes novos dados não aprimoraram a acurácia do modelo, não sendo utilizadas na engenharia de features.

Transformação das features

Modelos baseados em árvores como o XGBoost não costumam se beneficiar de transformações nos dados, como:

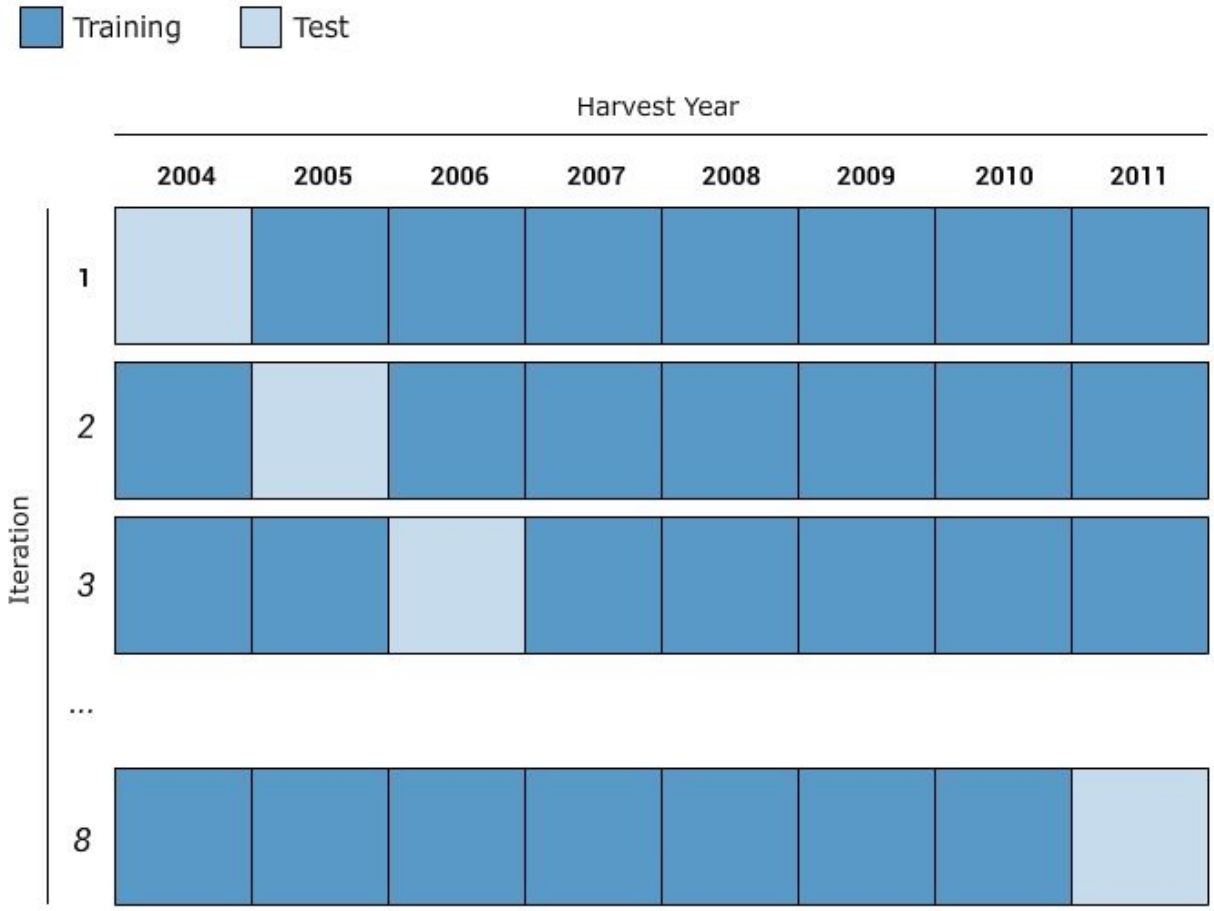
- Alteração na escala (StandardScaler, QuantileTransformer)

- Transformação de variáveis categóricas (OneHotEncoder)

Desta forma, não foram aplicadas transformação no dataset.

Validação local

Devido à **característica temporal** do dataset, utilizamos **validação k-fold** onde os *folds* foram **segmentados pelo ano da colheita**.



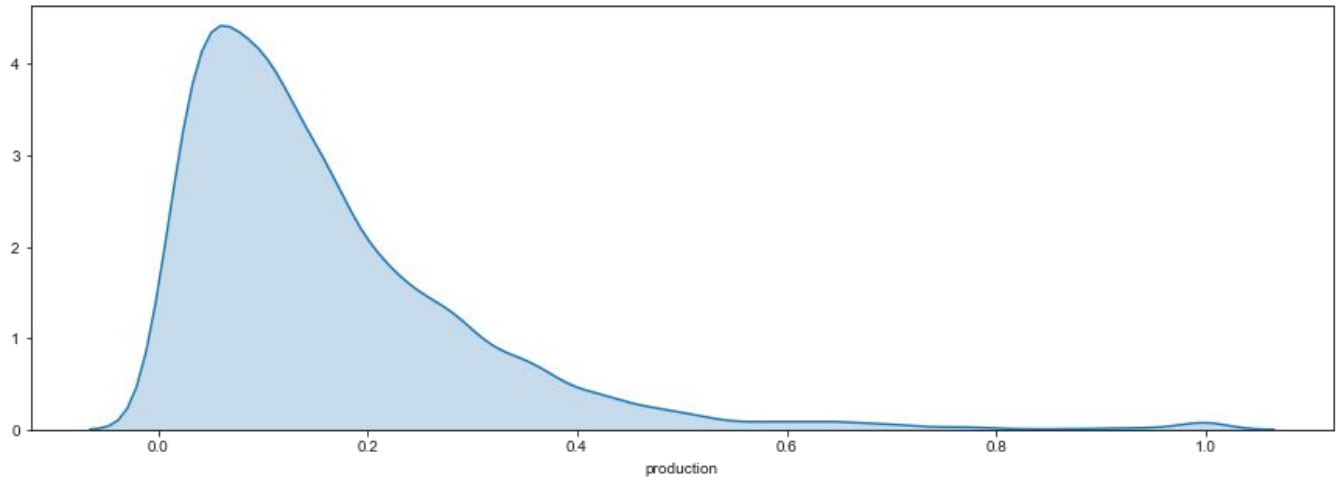
Validação local

```
validation(
    XGBRegressor(random_state=1),
    train.copy(),
    FEATURES
)
```

```
2004 - 0.08249160642854028
2005 - 0.07848157736198758
2006 - 0.07712762225183364
2007 - 0.07531411923053549
2008 - 0.0728327367800164
2009 - 0.08269558166787827
2010 - 0.06935449545319247
2011 - 0.08904272886528519

Mean score: 0.07841755850490867
```

Transformação do target



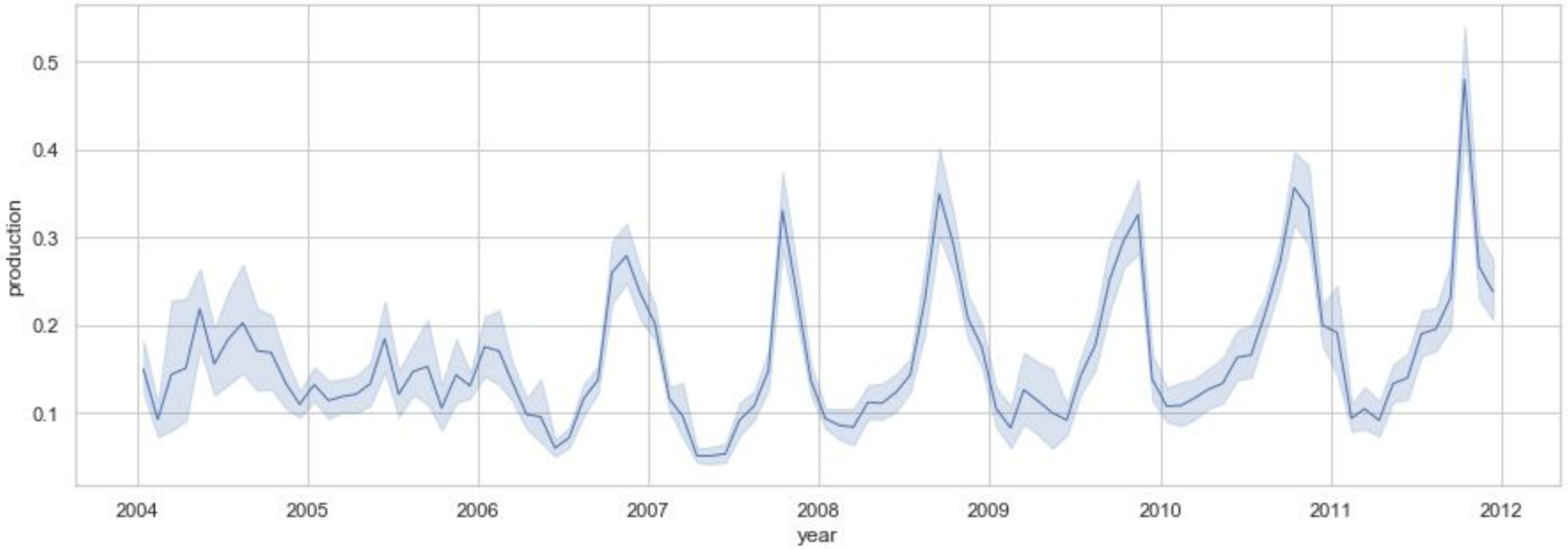
- Os **dados a serem previstos** possuem uma **distribuição estatística diferente da normal**, o que pode impactar na acurácia do modelo.

- Foram aplicadas transformações na coluna a ser prevista, para alterar sua distribuição.

- Durante a validação, encontramos como melhor transformação **elevar** os valores a serem previstos a **0.1**.

Seleção dos registros

Os **dois primeiros anos** do conjunto de treino (2004 e 2005) possuem valores de produção inconsistentes com os demais, não sendo utilizados no modelo.



Modelos

Foram utilizados **dois modelos XGBoost**, ajustando os parâmetros *max_depth* e *gamma* para evitar *overfitting*:

- max_depth: profundidade máxima das árvores

- gamma: torna o modelo mais conservador na divisão dos nodos das árvores

A previsão dos modelos foi **combinada** utilizando **média aritmética simples**, visando:

- Evitar *overfitting*

- Aumentar a generalização

Valores desconhecidos

- O conjunto de teste possui dois valores na coluna “type” que não aparecem no treino: -1 e 7.

- Como esses valores não foram vistos pelos modelos durante o treinamento, eles **não serão capazes de realizar uma previsão adequada**.

- Dessa forma, para uma previsão aproximada mais consistente, os registros de “type” -1 e 7 foram previstos com base na média dos registros semelhantes no treino.

	Id	best	mean
0	5243	0.764729	0.780892
1	5244	0.728357	0.753050
2	5245	0.720708	0.757387
3	5246	0.720301	0.739678
4	5247	0.721155	0.753678

Obrigado pela atenção!