Back-testing and machine learning models

bruno.abia@icomp.ufam.edu.br
twitter: @brunoabia

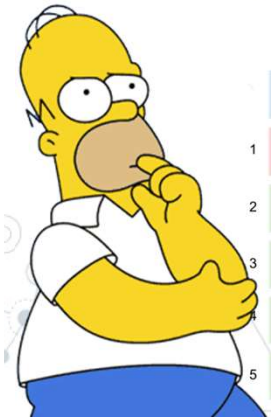
Salut!

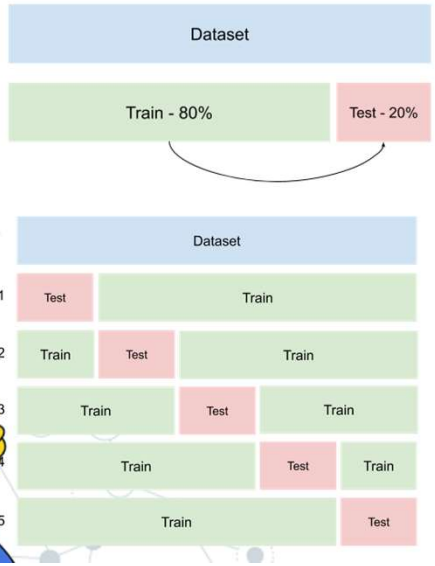
Je m’appelle Bruno Ábia

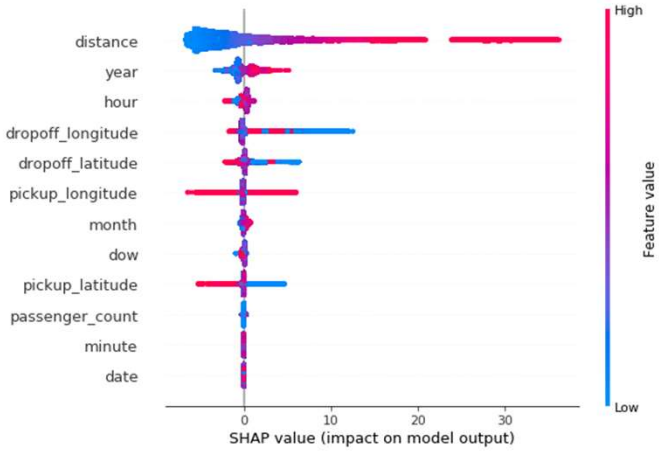
- Ingénieur informatique
- Maîtrise en informatique
- Conseiller chez CGI
- Youtuber - DevDojo Academy

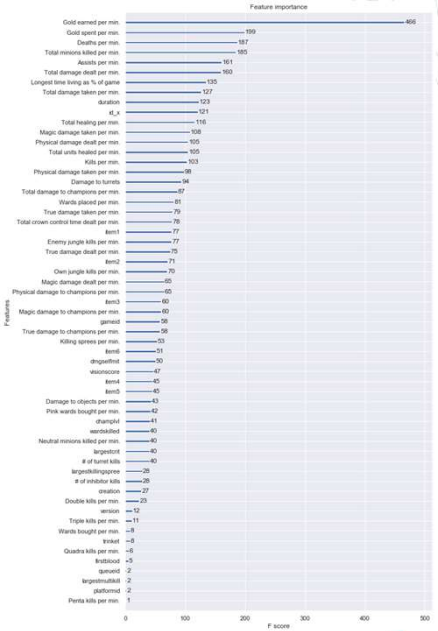
Why is important to study about it?

When we start to study data science is common to begin with traditional methods to understand and evaluate our machine learning models. For example:



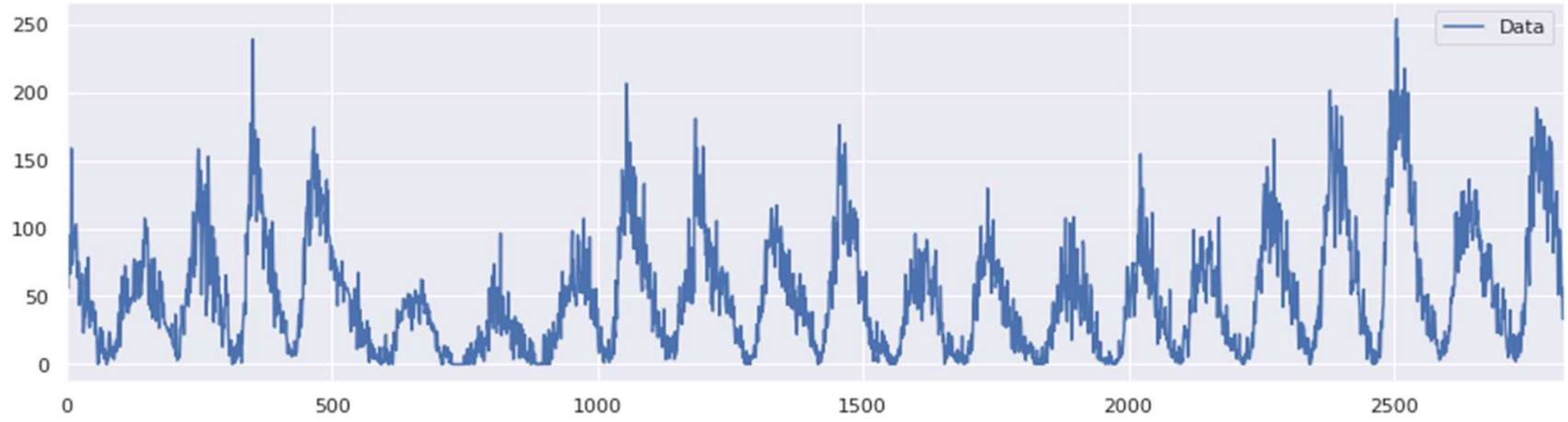






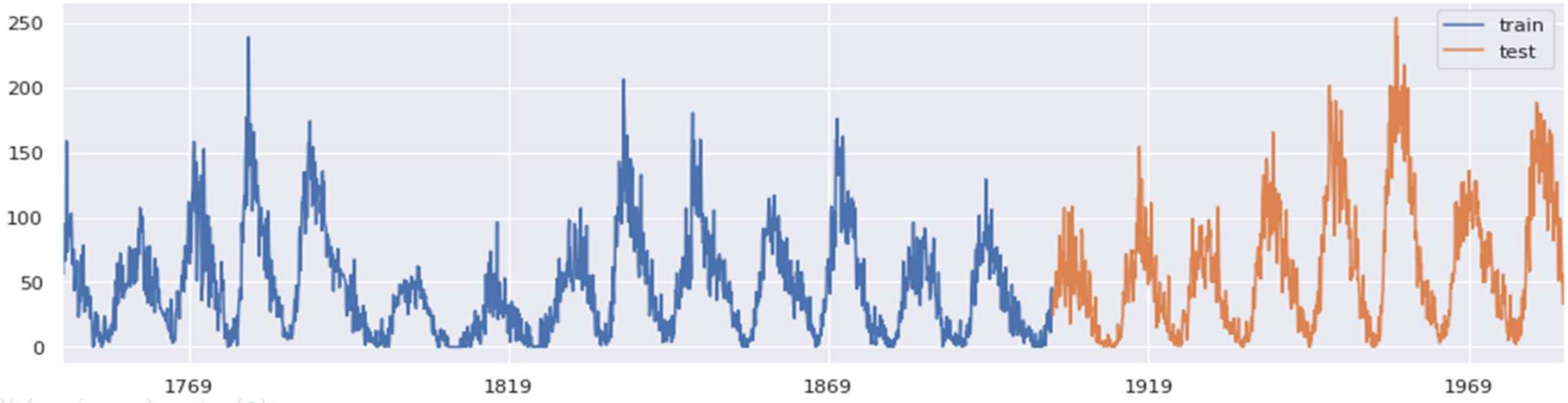
What is back-testing?

The back-testing is a technique to evaluate the machine learning models when you want to consider temporal components.



Where I can use back-testing?

You can use Back-testing in **ANY** machine learning problem that you have temporal components. For example, if you work with sentiment analysis with social network data, you can consider collected date as component.



How I can apply back-testing?

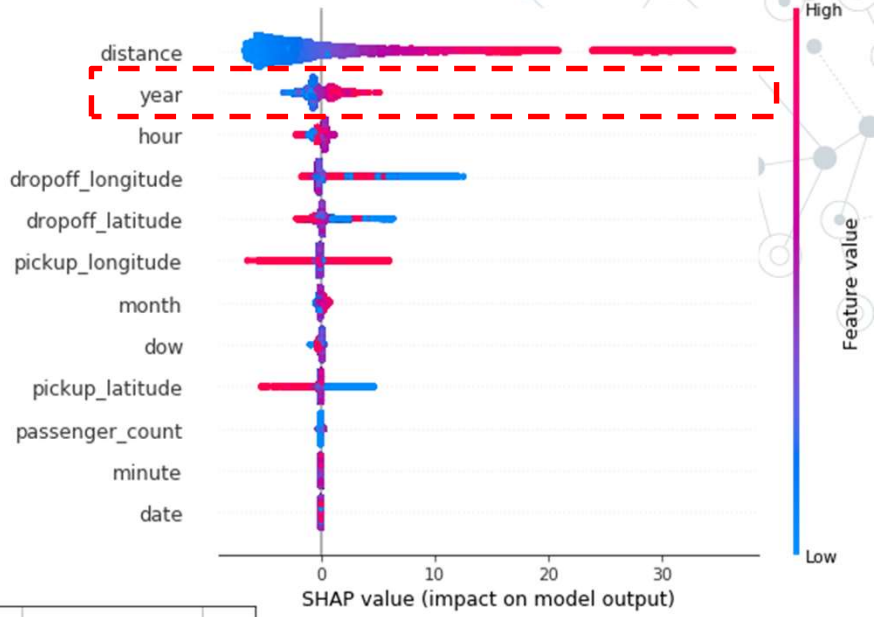
Example: **New York Taxi Fare Prediction Problem**

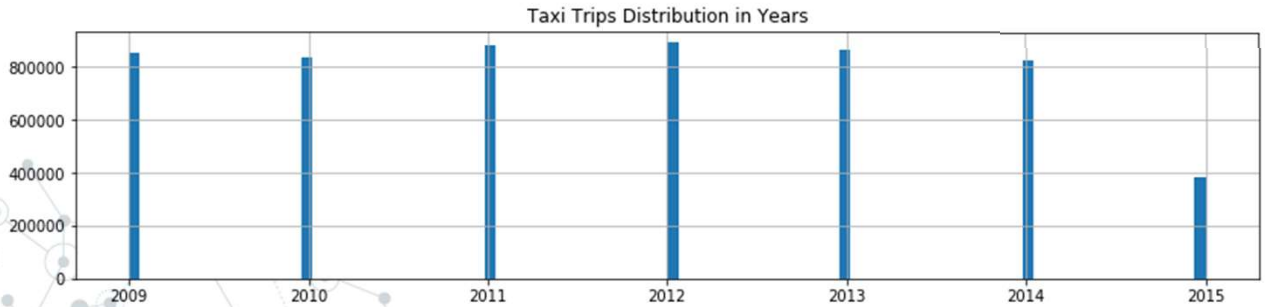
- The dataset has 55M rows;
- I started collecting a random sample of 10% of the dataset (about 5,5 million rows);
- Model used: XGB Regressor
- Linear Regression problem
- Evaluation metric: R^2 (regression score function)

- Total Features: 12
  - pickup_longitude
  - pickup_latitude
  - dropoff_longitude
  - dropoff_latitude
  - passenger_count
  - Year
  - Month
  - Date
  - Day
  - Hour
  - Minute
  - distance

Example: **New York Taxi Fare Prediction Problem**

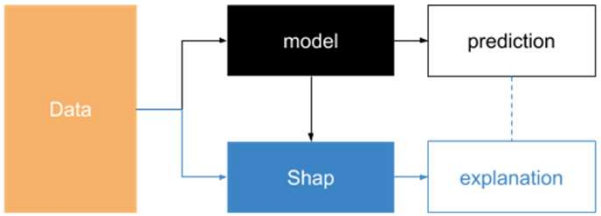
To create the tests, I used the years as a temporal component

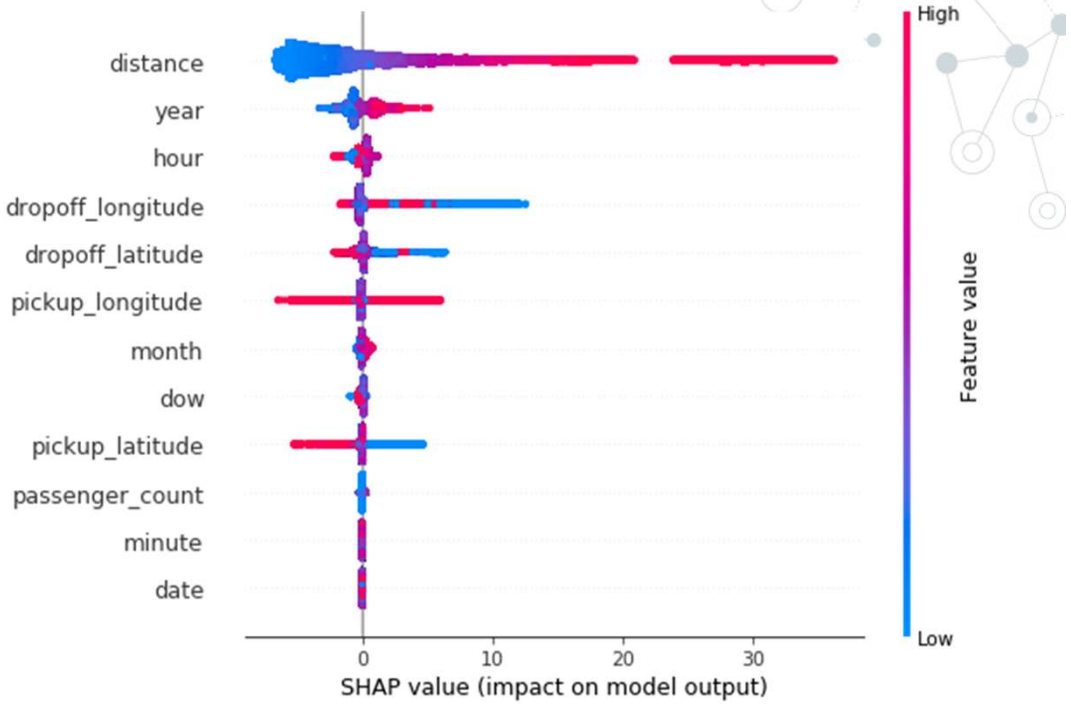




Explanation with Shap Valeus

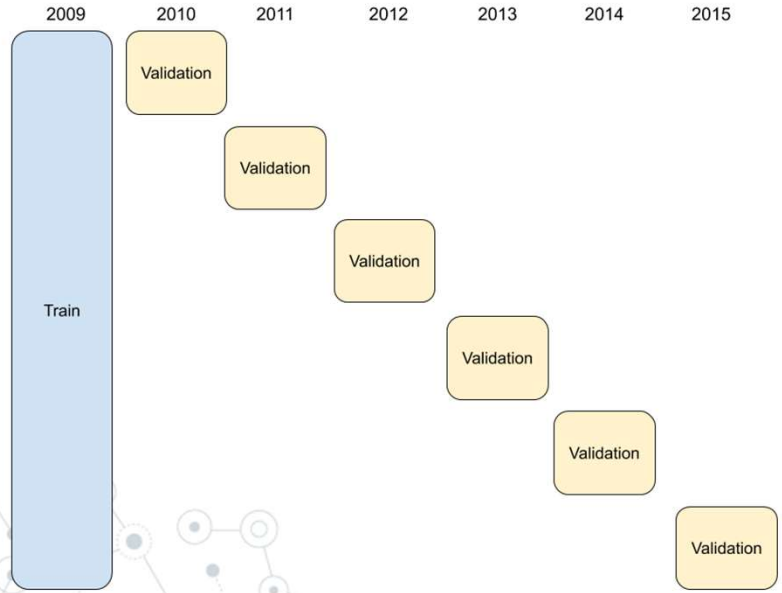
SHAP (SHapley Additive exPlanations) is a unified approach to explain the output of any machine learning model. SHAP connects game theory with local explanations, uniting several previous methods.

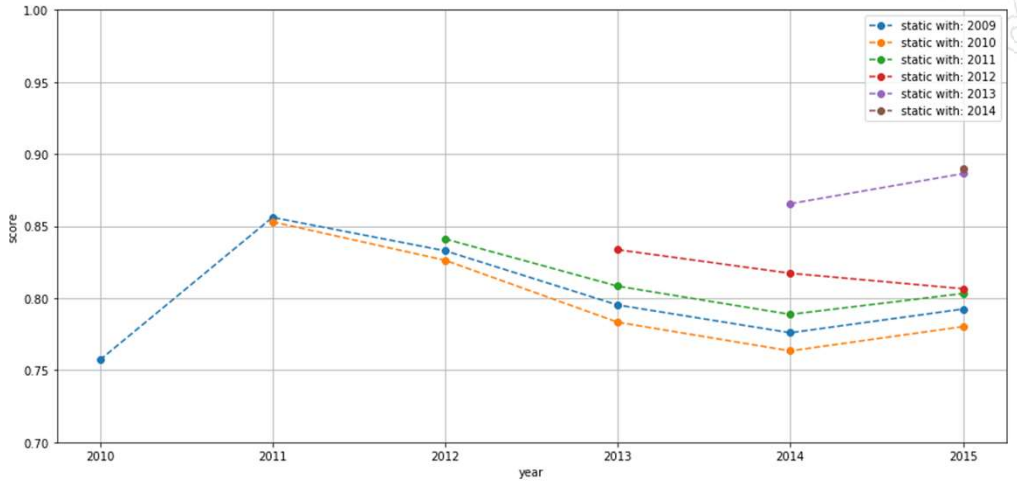




Walk-forward - Static

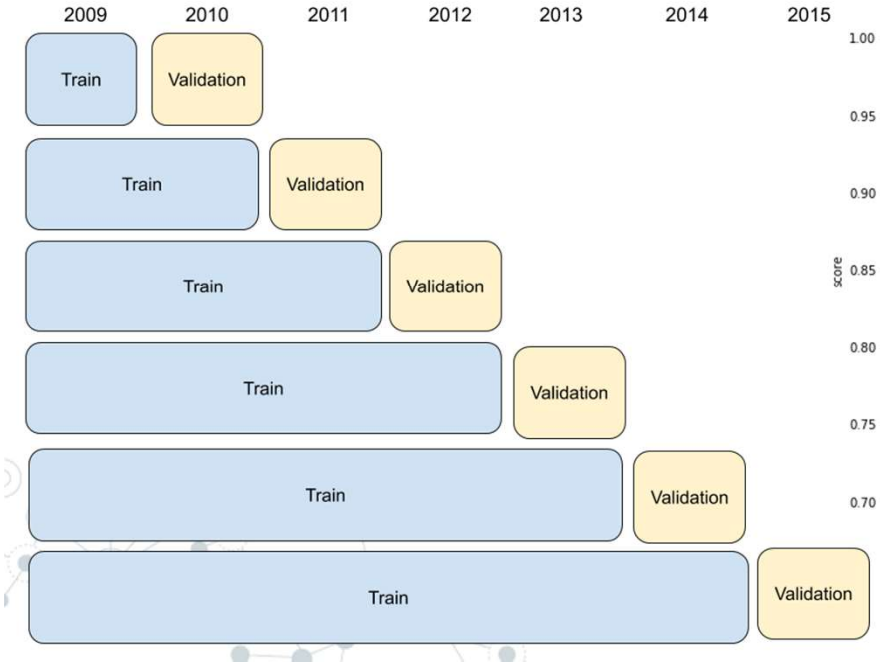
This approach suggests static training model of data, and making predictions for the following period.

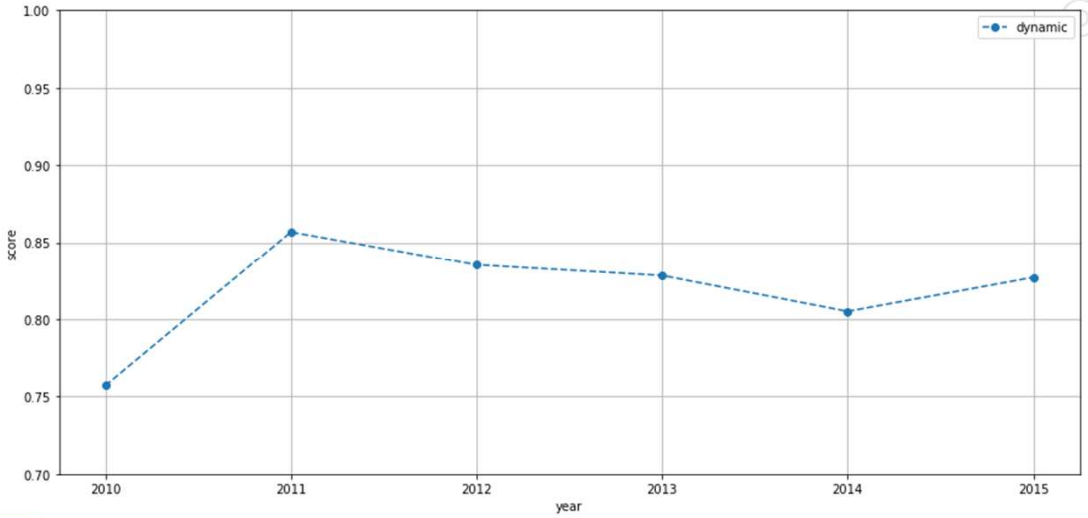




Walk-forward - Dynamic

This approach suggests training models on successive (potentially sliding) windows of data, and making predictions for the following period.



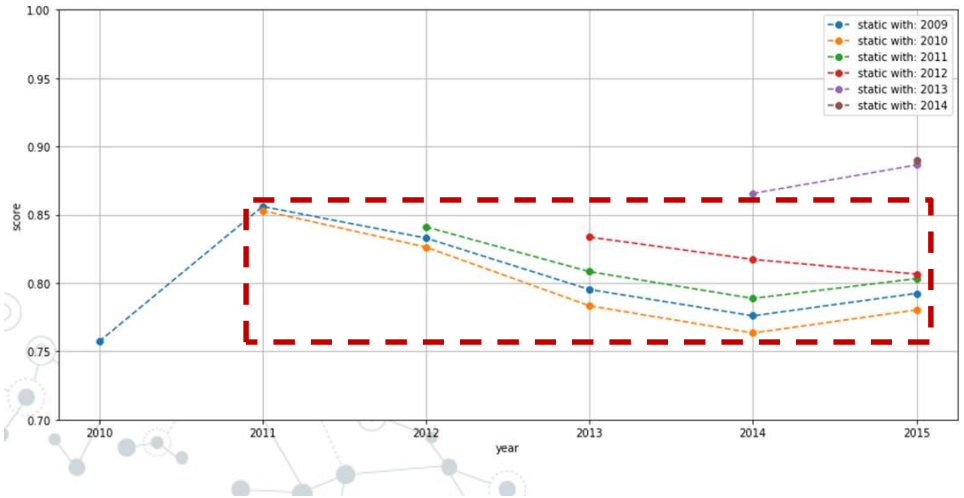


Example: **New York Taxi Fare Prediction Problem**

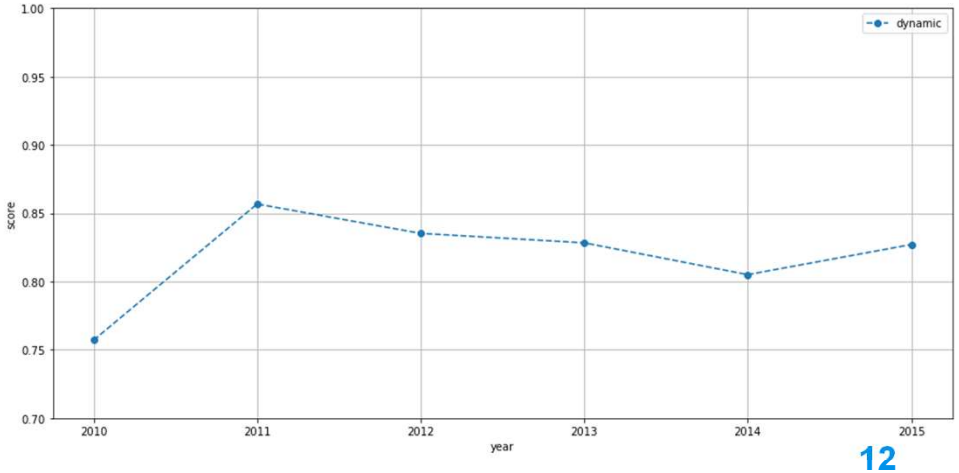
For each step, you can combine the traditional techniques to help you to understand the results:

- Estimating the Hyperparameters
- Calculating the important features
- Applying Shap Values

Static

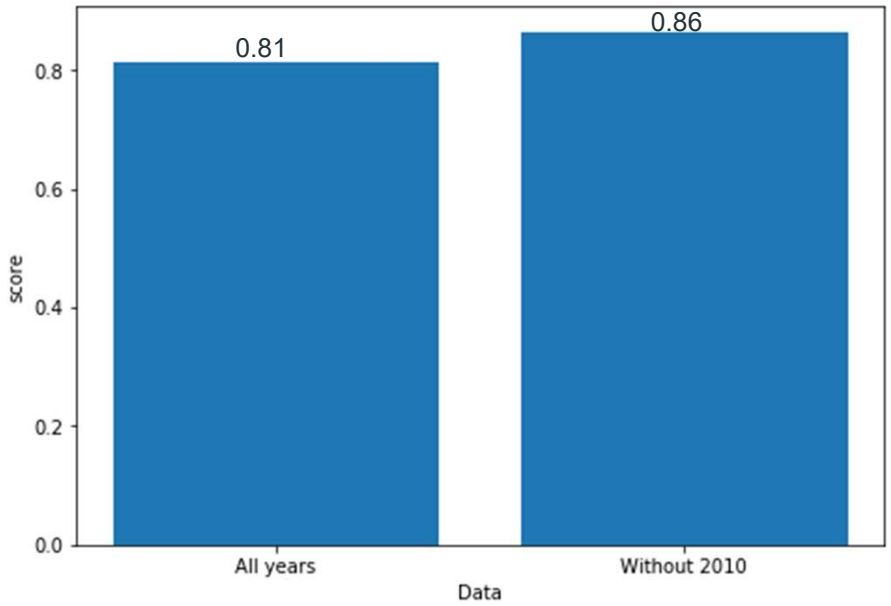


Dynamic



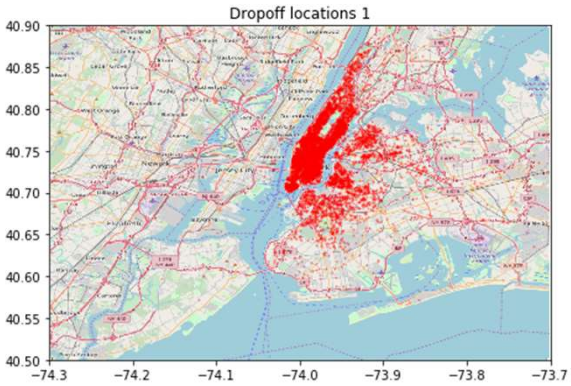
Example: **New York Taxi Fare Prediction Problem**

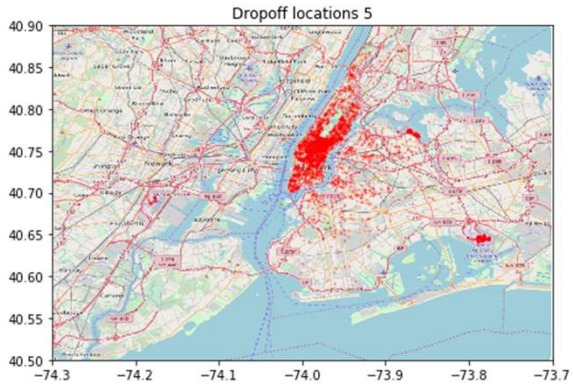
It observed with back-testing the samples of 2010 are not "important" for model prediction. If we train the model without 2010. The result is:

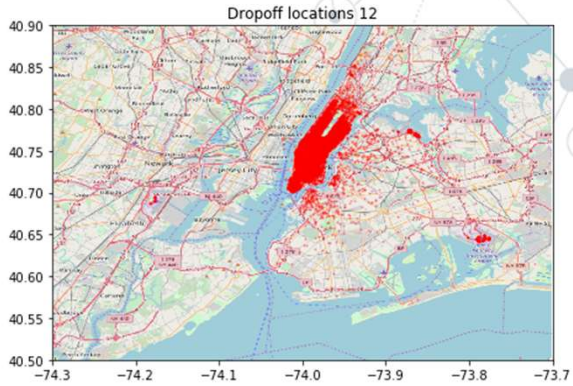


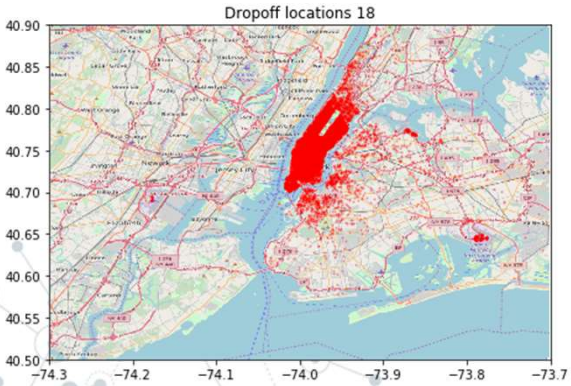
We can do too..

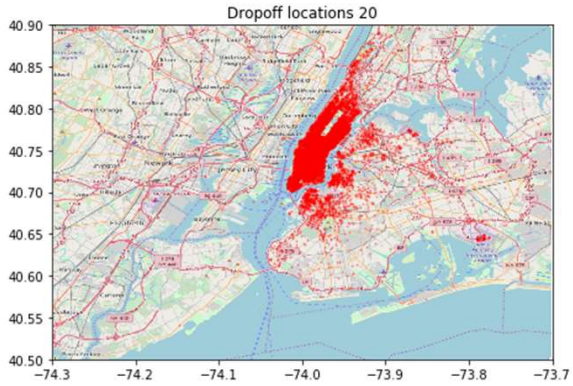
Applying back-testing with hours as a temporal component.

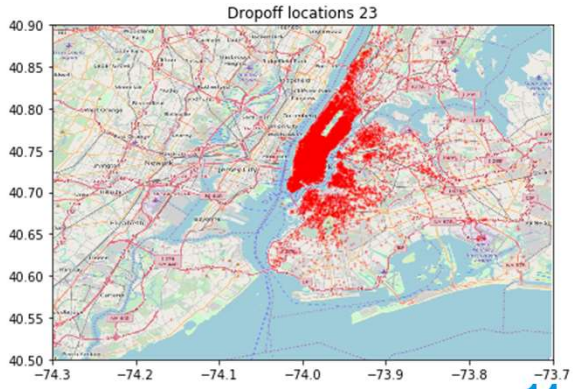






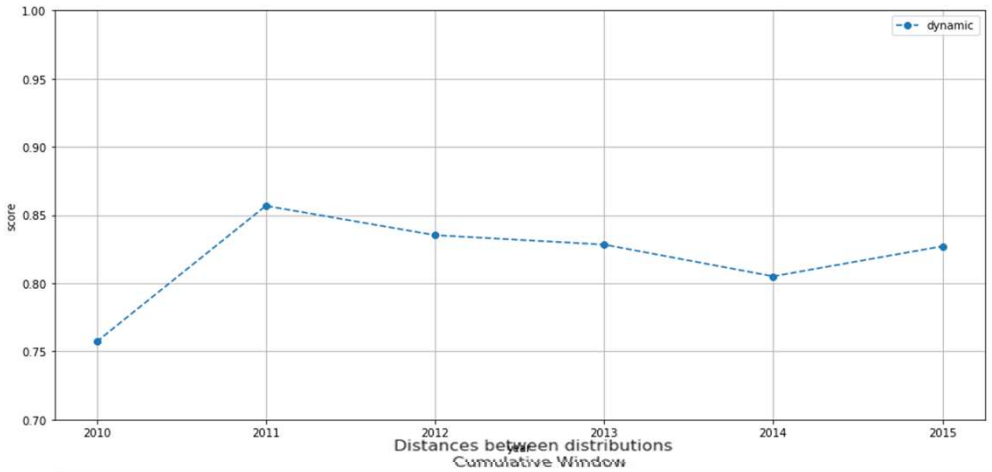


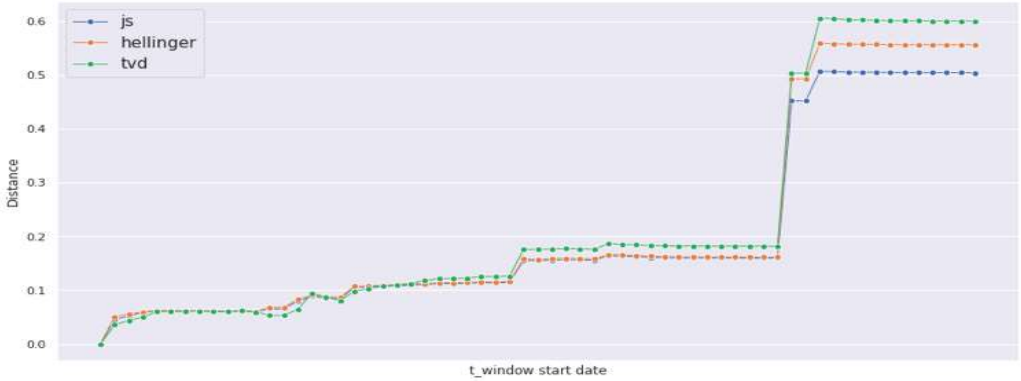




The Concept Drift

The **Concept Drift** means that the statistical properties of the target variable, which the model is trying to predict, change over time in unforeseen way.





I want to identify a big change behavior in our system. When I found, I come back to data study and back-testing.

Conclusion

- Back-testing allows us to understand our data;
- We can select the best model based on temporal components;
- The training process is long, but the results are cool **=)** ;
- Shap Values allows us to see the important features;

References

- Thais Almeida, Bruno Ábia, Eduardo Nakamura, Fabíola Nakamura (2017, October). Detecting hate, offensive, and regular speech in short comments. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web* (pp. 225-228).
- *Shapley sampling values:* Strumbelj, Erik, and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions." Knowledge and information systems 41.3 (2014): 647-665.
- *LIME:* Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.
- Shap Values: https://github.com/slundberg/shap
- Back-testing: https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/
- Dataset: https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/overview

Merci Beaucoup!

Any questions?

You can find me at:

Twiiter: @brunoabia &
bruno.abia@icomp.ufam.edu.br