SCIKIT-LEARN

A TOOL FOR BETTER MODEL RISK GOVERNANCE @ BNP PARIBAS CARDIF

Sébastien Conort
Chief Data Scientist

Tung Lam Dang
Data Scientist

ANNUAL WORKSHOP, SCIKIT-LEARN
Nanterre, 28th may 2019

BNP PARIBAS CARDIF

The insurer for a changing world

BNP Paribas Cardif – insurance branch of the Group

(2) Variation at constant scope and exchange rates

A DIVERSIFIED ACTIVITY

51% REVENUES* **PROTECTION**

49% REVENUES* **SAVINGS**

*Economic net banking income

World leader in credit insurance

GROSS WRITTEN PREMIUMS	ASSETS UNDER MANAGEMENT
31.8 €Bn	**239** €Bn
+9% / 2017(1)	+1% / 2017

Unique business model anchored in partnerships
500 partner distributors in a variety of sectors

- Banks and financial institutions
- Credit companies
- Automaker financing arms
- Financial advisors and brokers
- Retailers
- Telecommunications and energy companies

SERVING 100 MILLION POLICYHOLDERS AROUND THE WORLD

35 countries

10 000 EMPLOYEES

Models? What are talking about?

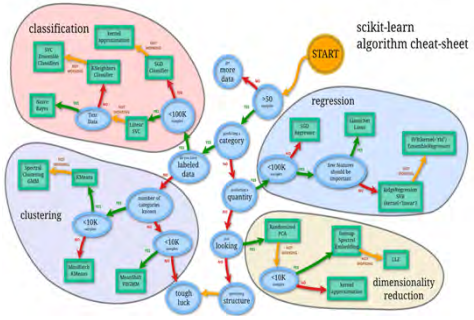
BNP Paribas RISK department : retained « model » definition

Quantitative method, system or approach that produces **quantitative estimates of uncertain values** and used to **make decisions** and/or to **make public communications**

In this talk, we will focus on models whose parameters are based on a statistical learning procedure

New challenges with Machine Learning models

NEW TYPES OF MODELS



NEW USES SPREAD ALL OVER THE COMPANY

HIGHER VOLUME OF MODELS IN PRODUCTION

Simple example: Error in data preprocessing

```
# Concat all texts into 1 column
X_train['all_texts'] = X_train.loc[:, TEXT_COLUMNS].apply(lambda row: ' '.join(row), axis=1)
X_test['all_texts'] = X_train.loc[:, TEXT_COLUMNS].apply(lambda row: ' '.join(row), axis=1)
```

```
# encapsulation helps
def concat_text_columns(df):
    return df.apply(lambda row: ' '.join(row), axis=1)
```

```
# this is much nicer
text_pipeline = Pipeline([
    ('concat_text', FunctionTransformer(concat_text_columns,
                                        validate=False)),
    ('vectorizer', TfidfVectorizer()),
    ('model', LogisticRegression()),
])

text_pipeline.fit(X_train.loc[:, TEXT_COLUMNS], y_train)

p_test = text_pipeline.predict(X_test.loc[:, TEXT_COLUMNS])
```

In a bigger context

```
text_pipeline = Pipeline([
    ('concat_text', FunctionTransformer(concat_text_columns,
                                        validate=False)),
    ('vectorizer', TfidfVectorizer())
])

global_pipeline = Pipeline(
    [
        (
            'transformer',
            ColumnTransformer(
                transformers=
                    [
                        ('text', text_pipeline, TEXT_COLUMNS),
                        ('cat', OneHotEncoder(), CAT_COLUMNS),
                        ('num', 'passthrough', NUM_COLUMNS),
                    ],
                remainder='drop'
            )
        ),
        (
            'model',
            LogisticRegression()
        ),
    ]
)
```

- **End-to-end pipeline**
  - pipeline.Pipeline
  - pipeline.FeatureUnion
  - compose.ColumnTransformer
- **Robust model selection and evaluation**
  - sklearn.model_selection
  - sklearn.metrics
- **Extensibility**
  - sklearn.base
  - sklearn.preprocessing

scikit-learn contribute to significantly lower risks

ROBUSTNESS of developments

- Strong community of developers
- Very large community of users
- Strong governance of developments

STANDARDISATION of classical steps

- Large scope of classical functions
- Key methodological steps packaged in functions or objects
- Simple, efficient and stable API design

EXTENSIBILITY of the API

Easy creation of custom :

- Metrics
- Transformers
- Regressors/classifiers
- Validation strategies

even based on other libraries (statsmodels, …)

Some of our best practices at BNP Paribas Cardif

Through our Analytics Governance, we strongly encourage our developers to:

- Not develop custom functions already available in scikit-learn
- Follow scikit-learn API for custom objects and put all steps in a Pipeline object for deployment.
  - Developers are required to justify in case this is not possible
- Always have a Baseline pipelines with scikit-learn when reporting results

Thank you!

Sébastien Conort
Chief Data Scientist
Contact
sebastien.conort@bnpparibas.com

Tung Lam Dang
Data Scientist
Contact
tunglam.dang@bnpparibas.com

THANK YOU!

BNP PARIBAS CARDIF
8, rue du Port
92728 Nanterre Cedex
Tel.: +33 (0)1 41 42 83 00
bnpparibascardif.com