BBC News Article Sorting

By - Rushikesh Darge

Objective:

Development of a predictive model that takes a news as an input and in output it returns the category of that news. Model correctly tells us the given news from which category.

Benefits:

- Time reduce to sorting news
- Reduce human error
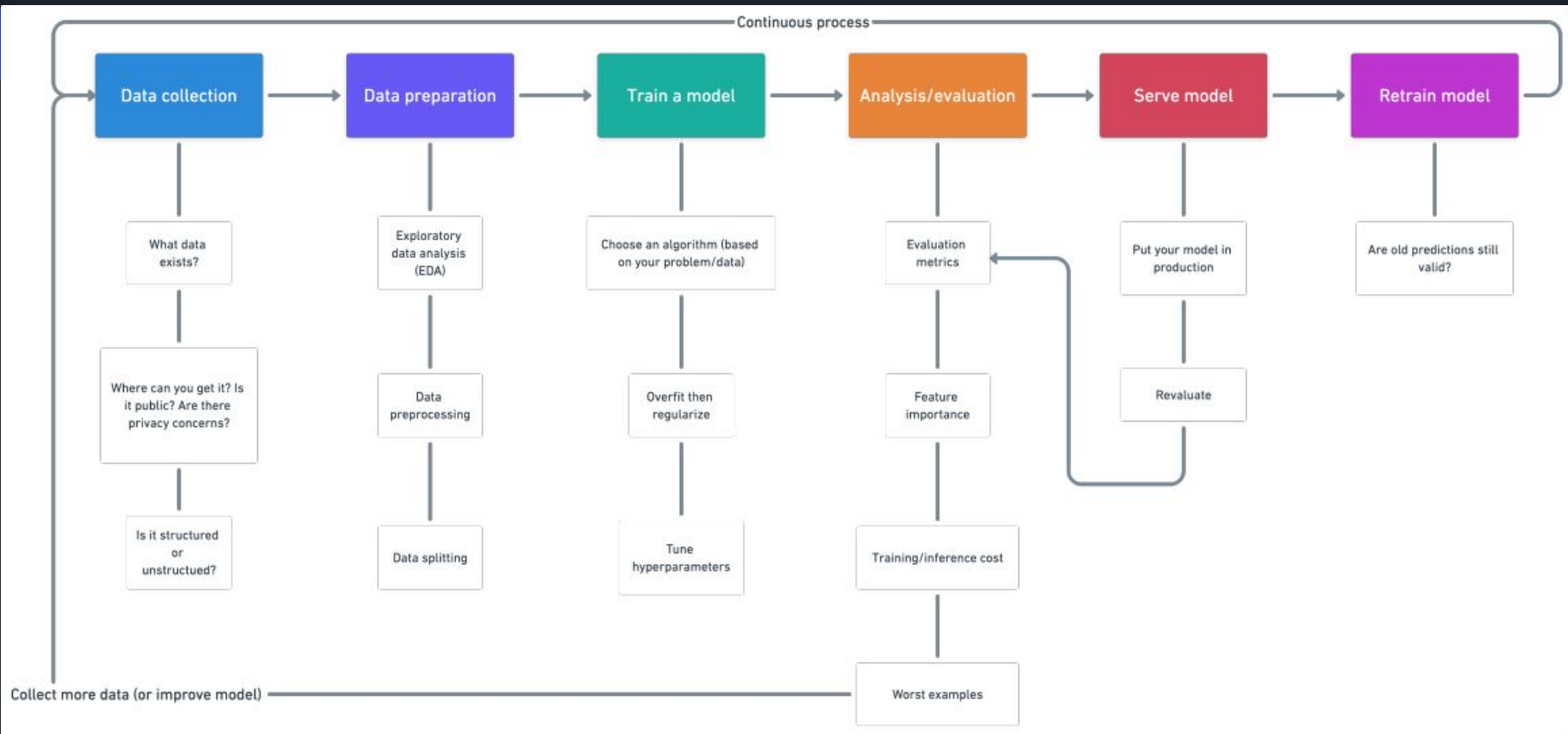- Task is automated

Data Sharing Agreement :

Dataset file name : BBC News Train.csv

Number of Columns : 3

Column names : ArticleId, Text, Category

Column data type : int, string, string

Continuous process

| Data collection | Data preparation | Train a model | Analysis/evaluation | Serve model | Retrain model |

Data collection

What data exists?

Where can you get it? Is it public? Are there privacy concerns?

Is it structured or unstructued?

Data preparation

Exploratory data analysis (EDA)

Data preprocessing

Data splitting

Train a model

Choose an algorithm (based on your problem/data)

Overfit then regularize

Tune hyperparameters

Analysis/evaluation

Evaluation metrics

Feature importance

Training/inference cost

Worst examples

Serve model

Put your model in production

Revaluate

Retrain model

Are old predictions still valid?

Collect more data (or improve model)

Data Validation and Data Transformation :

Data : We download the csv files form kaggle platform for training

Data Preprocessing : See samples of rows and get insight of data. Some contain different forms of the same words like good, better, best or do, doing like words. We prepreocess them and also we remove all characters except alphabets and numbers.

Null values in the columns : We don't have any null values in our dataset

Model Selection:

After the PCA visualization , we find the best model out of these 2 algorithms "Logistic Regression" and "Decision Tree". We calculate the Accuracy and F1 scores for both models and select the model with the best score.

Deployment:

BBC News Article Sorting

Enter the news for classifying

Input Field

This news belongs from Output Field

Thank You!!!