ML Methods for Labeling CPI Data

Ostap Stefak

Harvard University

ostapstefak@college.harvard.edu

August 26, 2021

National Bank
of Ukraine

Overview

The Task: Overview

- Categorize products scraped from online stores into their correct CPI category (200+ categories).
- Difficulties/constraints: computational, time, number of categories.

The Task: Literature Review

- UK Office for National Statistics paper [Sands, 2020] $\rightarrow$ XGBoost is the best performing model, Logistic Regression and Random Forest not far behind.
- Netherlands Statistical Office [Harms, A. (2019)] $\rightarrow$ best model is Logistic Regression, Random Forest and XGBoost similar.
- US Census Bureau paper [Harms, Roberson 2021] $\rightarrow$ Logistic Regression is best.
- Simple methods for text encoding suggested by all surveyed papers (e.g. one-hot encoding, n-grams etc)

The Task: A Closer Look at Our Data

- Important columns listed here. Other columns are date/time related information and country of origin.

	Unnamed: 0	X	retailer	city	product_id	CPI	origin	name	category	category_full
0	(57.0, 2272109)	48603	tavriav	odesa	27515435	57.0	NaN	Йогурт Агуша дитяч. питний 27 % 200 г бут. малина	Молокосырыяйца:Йогуртыдесерты:Йогурты:323:15	323
1	(57.0, 15769809)	310674	furshet	kyiv	4820052666254	57.0	NaN	Бифидойогурт слива/лен 3% Активиа 115г	iogurt	dairy-eggs:iogurt
2	(57.0, 1809922)	44229	tavriav	odesa	13133	57.0	NaN	Йогурт Кілія Елітний 450 г бут. н/ж Злаки з медом	Молокосырыяйца:Йогуртыдесерты:Йогурты:323:9	323
3	(57.0, 2090786)	46885	tavriav	odesa	209071	57.0	NaN	Йогурт Локо Моко 115 г 15% з персиком	Молокосырыяйца:Йогуртыдесерты:Йогурты:323:9	323

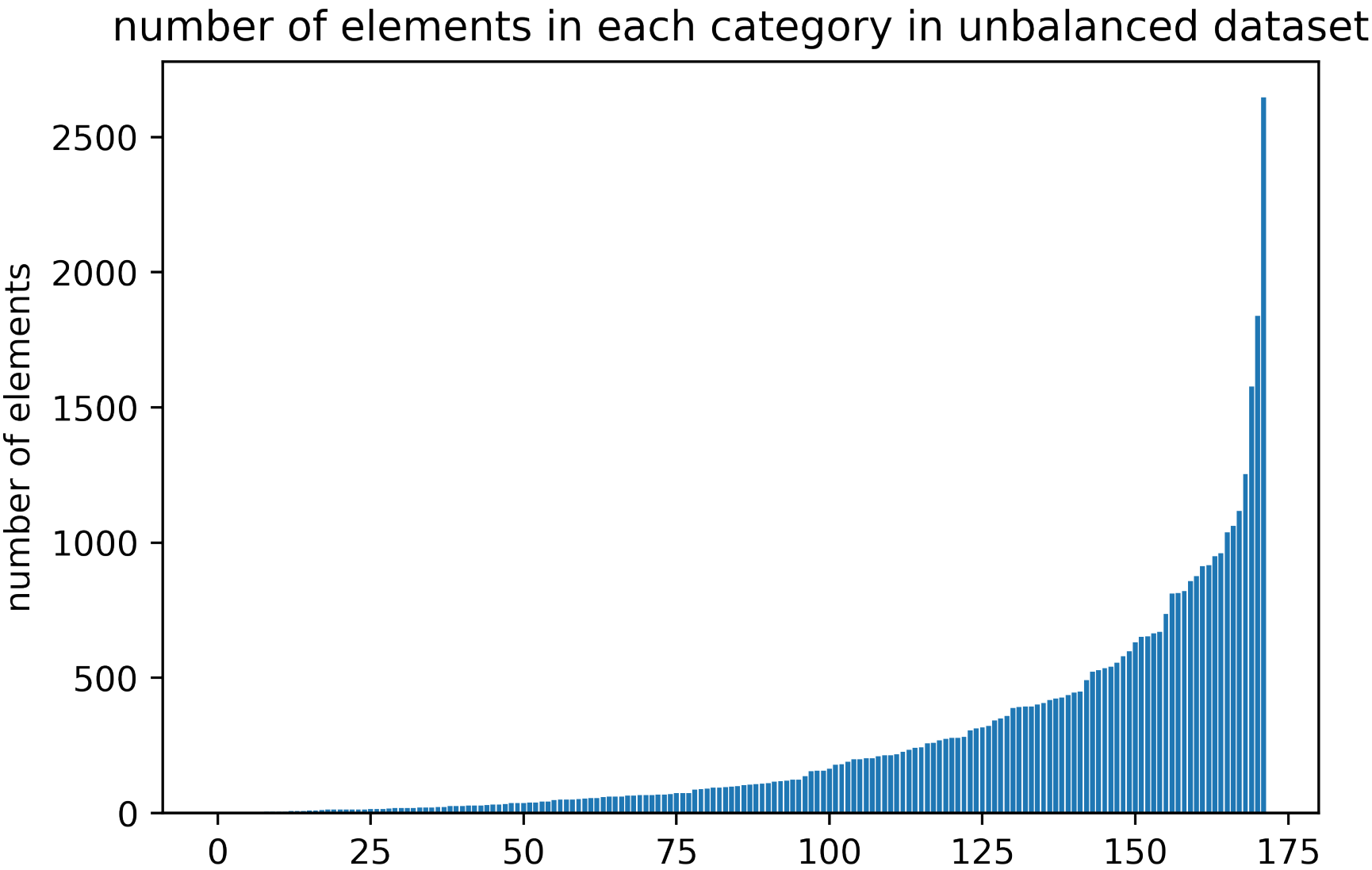
Data Processing: Data Cleaning

- Remove duplicates I: The initial dataset has 1.77 million observations. Many of these are duplicates. Removing $\rightarrow$ shrinks to 103k observations
- Removing duplicates II: There can also be duplicates with just the same name. It is important to remove duplicates of this form too $\rightarrow$ shrinks dataset to 101k observations.
- Remove observations with missing/unknown CPI tag $\rightarrow$ shrinks the dataset to 44k observations (of labeled, training data)

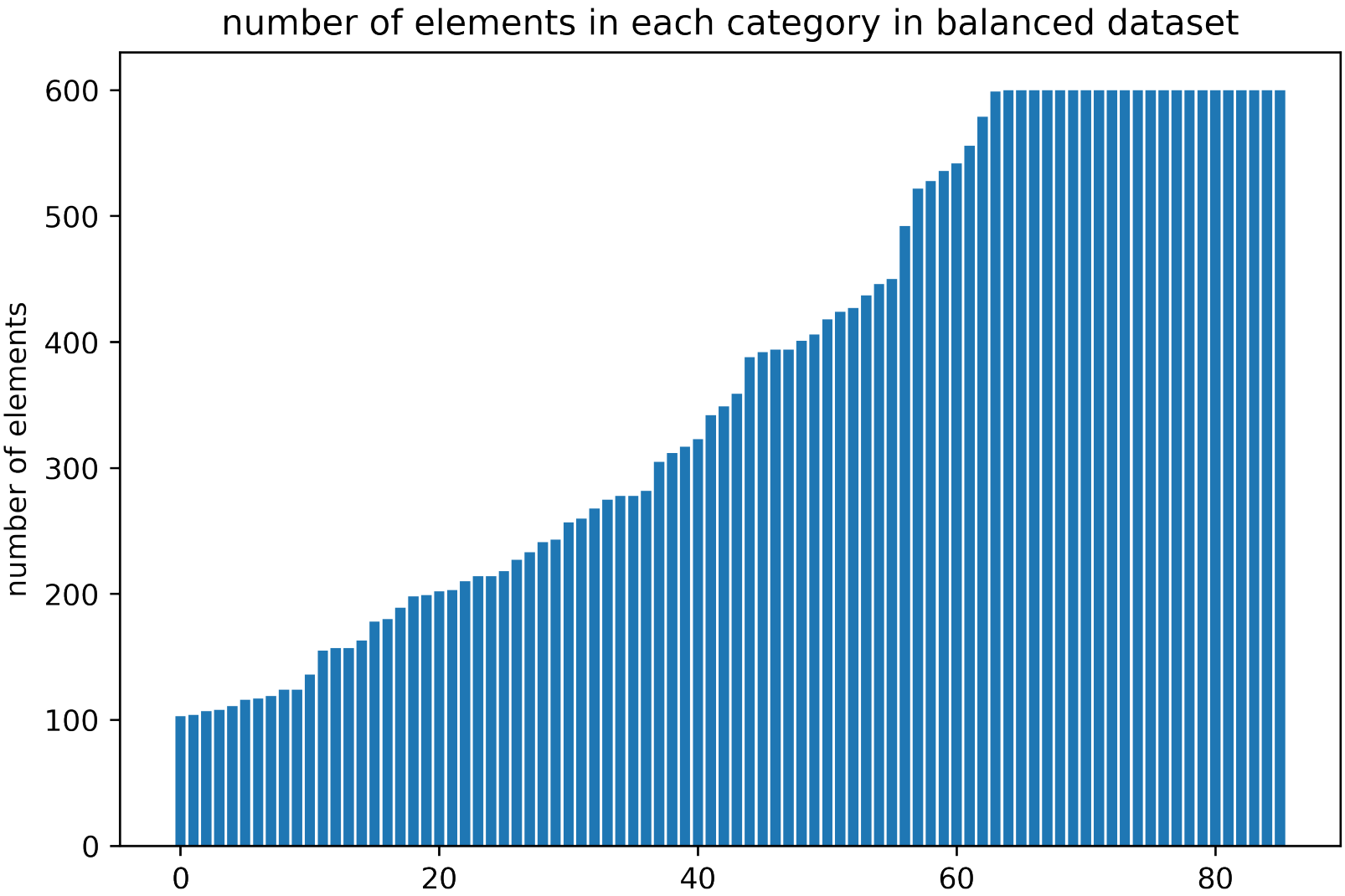
Modeling choice I

Remove items that belong to categories with fewer than 100 items assigned $\rightarrow$ reduces number of categories from 172 to 86; shrinks dataset to 41k observations.

Data Processing: Data Cleaning



Data Processing: Data Cleaning



Data Processing: Text Transformations

Modeling choice II

As a baseline, use only a text transformation of the 'product name' field.

- **Method Used:** CountVectorizer() from sklearn.

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

- **Reasoning:** simplest method that encapsulates a lot of information. Not context dependent.
- **Intuition:** Allows machine learning methods to find patterns that otherwise would need to be found manually using search and exclusion terms (what UK trialled in 2016 and what Poland trialled more recently).

Data Processing: Training Setup

- The dataset is split into train and test chunks (70% train, 30% test).
- **Features:** the outputs from CountVectorizer() on the product name column, which results in slightly more than 20k features.
- **Output**: The integer value of the inflation category class.

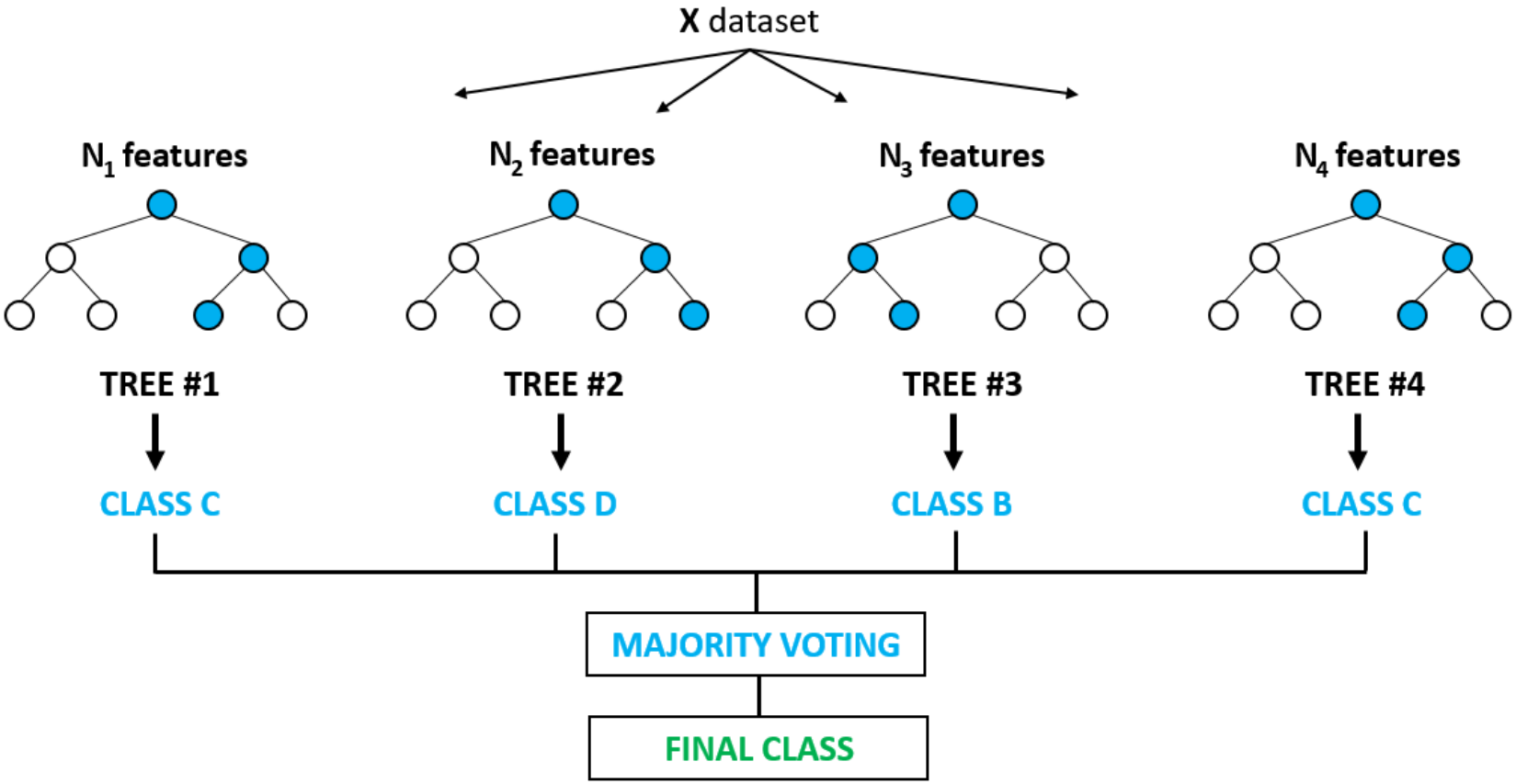
Methods w/o cross-validation: Reasoning

Modeling choice III

Methods applied: Random Forest (with various numbers of trees), Logistic Regression.

- **Reasoning:** Relatively short training time, high expressiveness, recommendation from literature reviews.

Methods w/o cross-validation: Random Forest



Methods w/o cross-validation: Performance

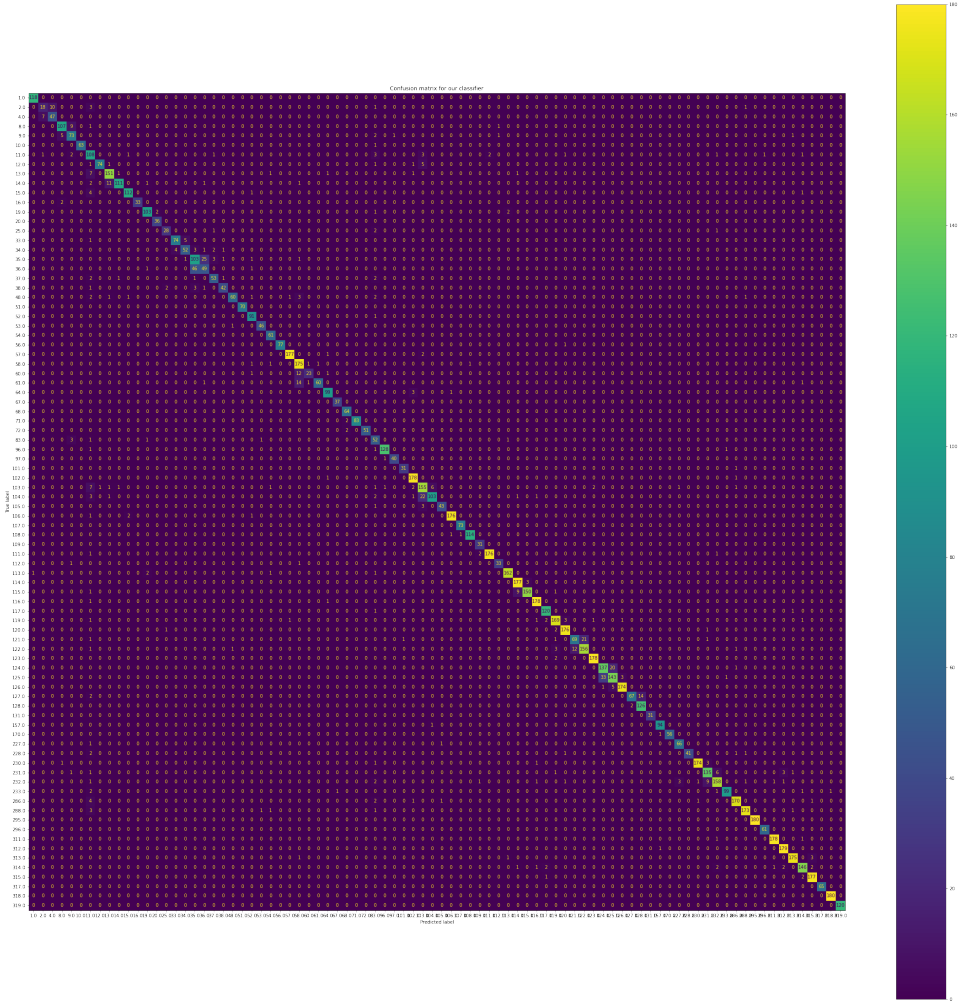
Insight I

Un-tuned Logistic Regression has better performance than un-tuned random forest methods.

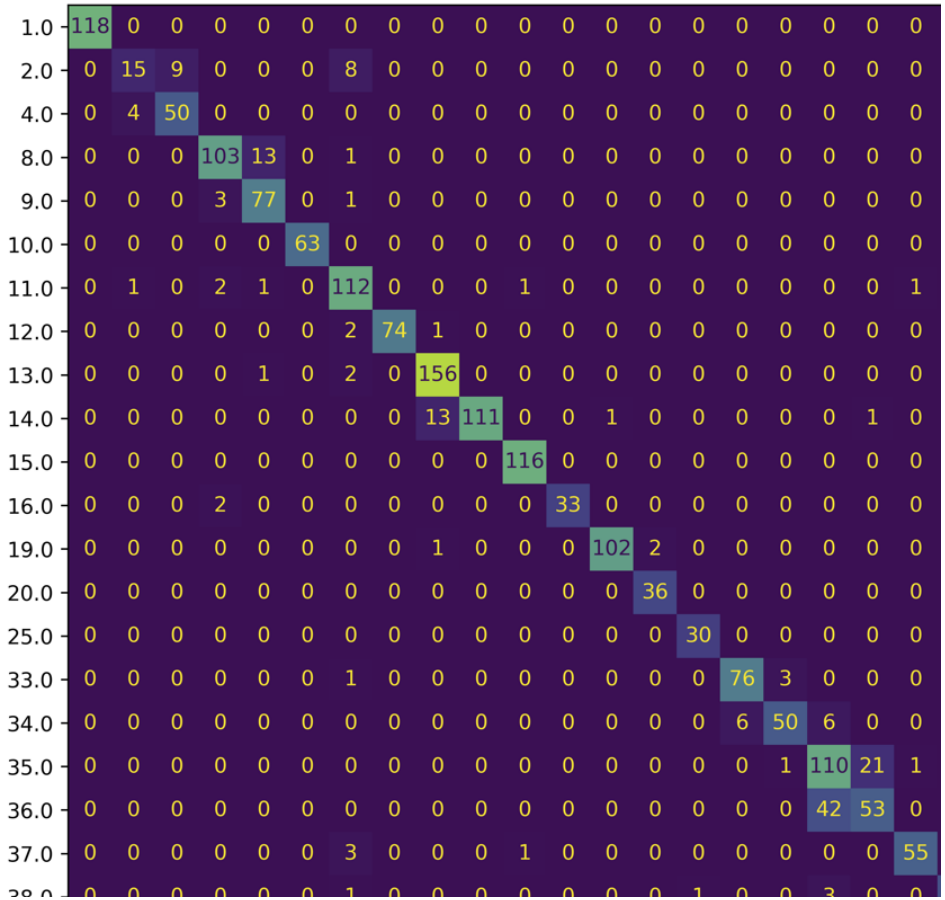
Type	**Accuracy**	**Precision**	**Recall**	**F1**
RF, 10	.933	.938	.923	.928
RF, 100	.946	.947	.935	.939
RF, 200	.947	.949	.939	.942
Logistic regression, $C = 1$	**.957**	**.959**	**.951**	**.954**

Table: Untuned random forest and logistic regression performance

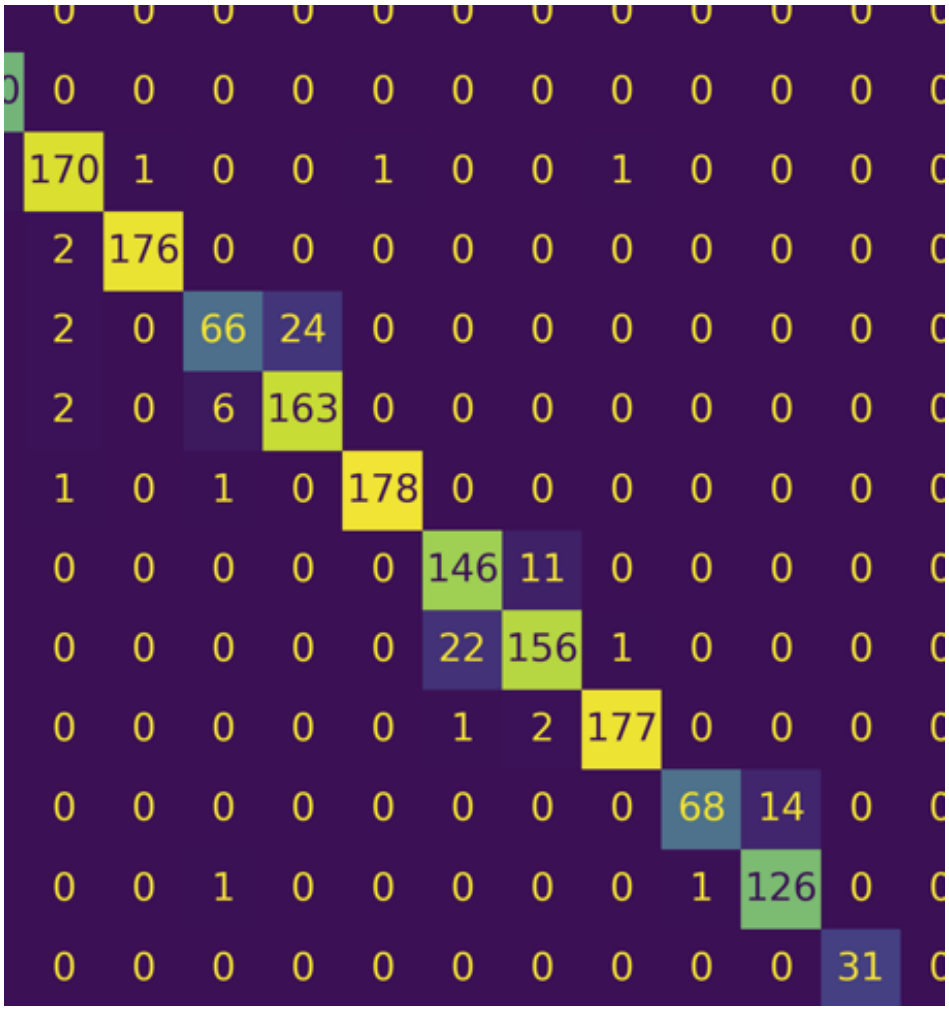
Methods w/o cross-validation: RF 200 Confusion Matrix



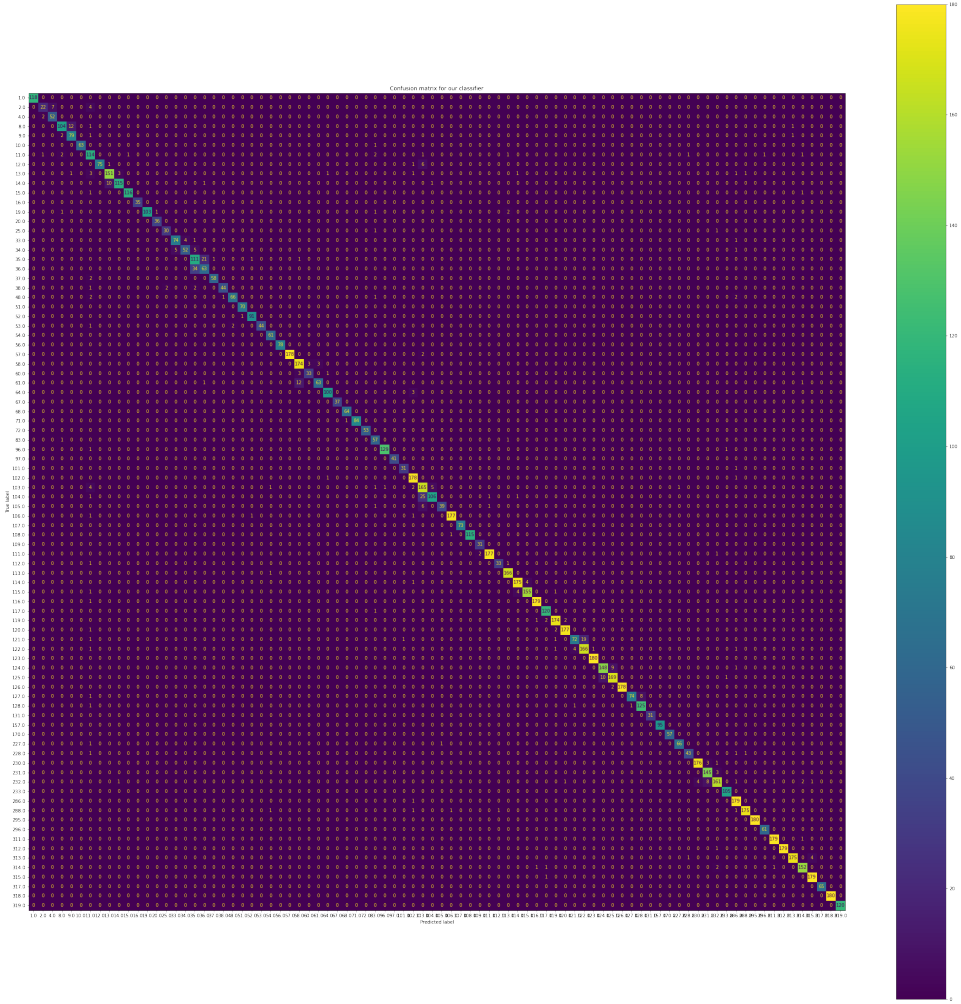
Methods w/o cross-validation: RF 200 Confusion Matrix Zoomed I



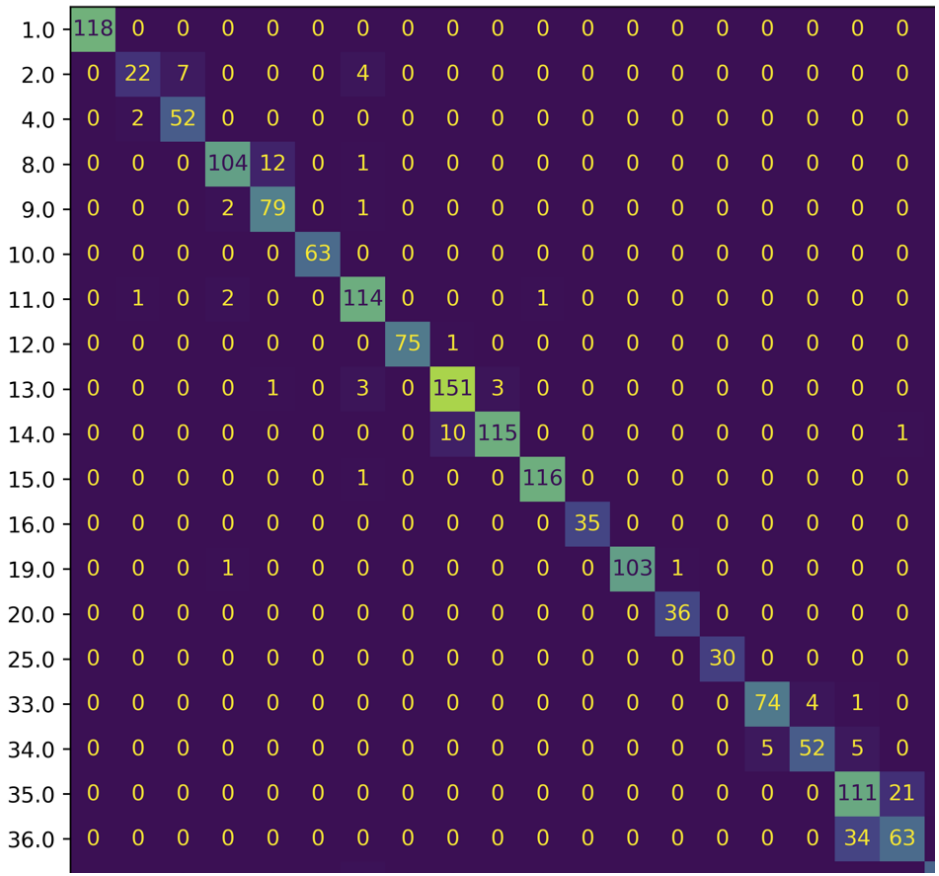
Methods w/o cross-validation: RF 200 Confusion Matrix Zoomed II



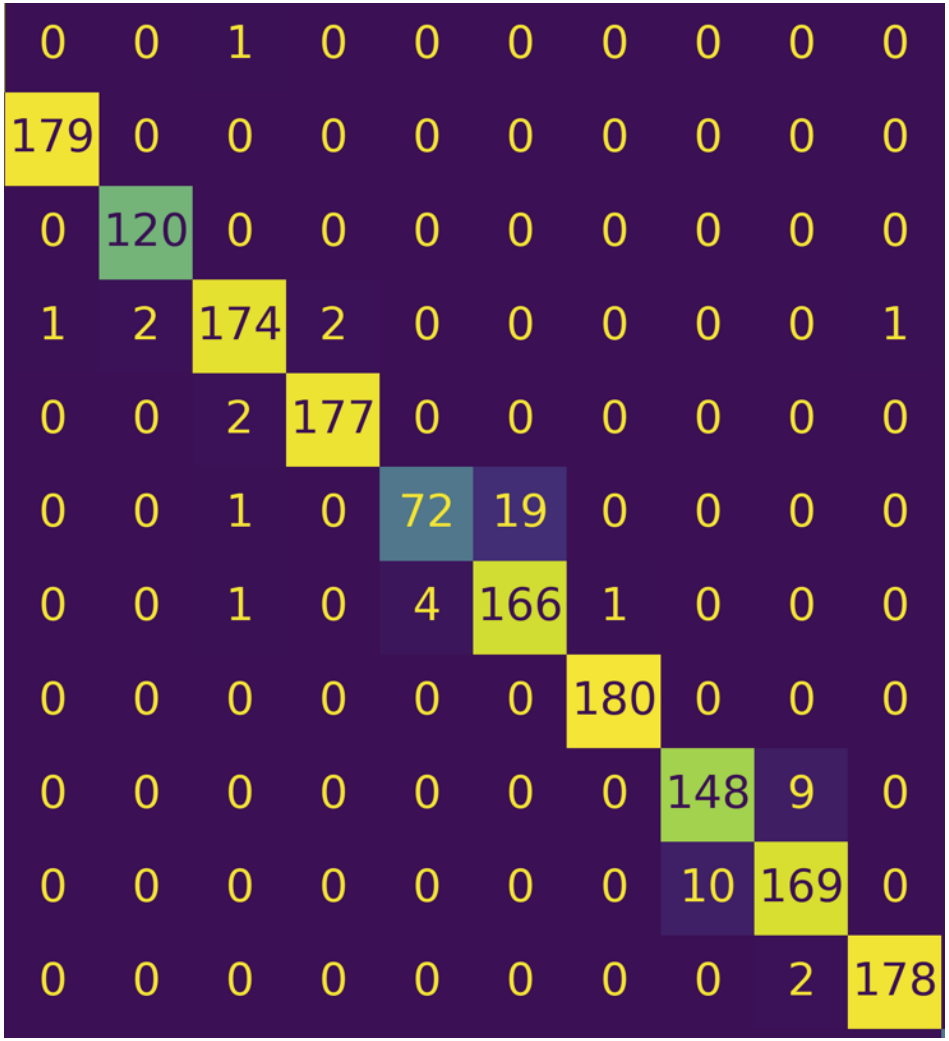
Methods w/o cross-validation: LR Confusion Matrix



Methods w/o cross-validation: LR Confusion Matrix Zoomed I



Methods w/o cross-validation: LR Confusion Matrix Zoomed II



Confusion Matrix Summary

Insight II

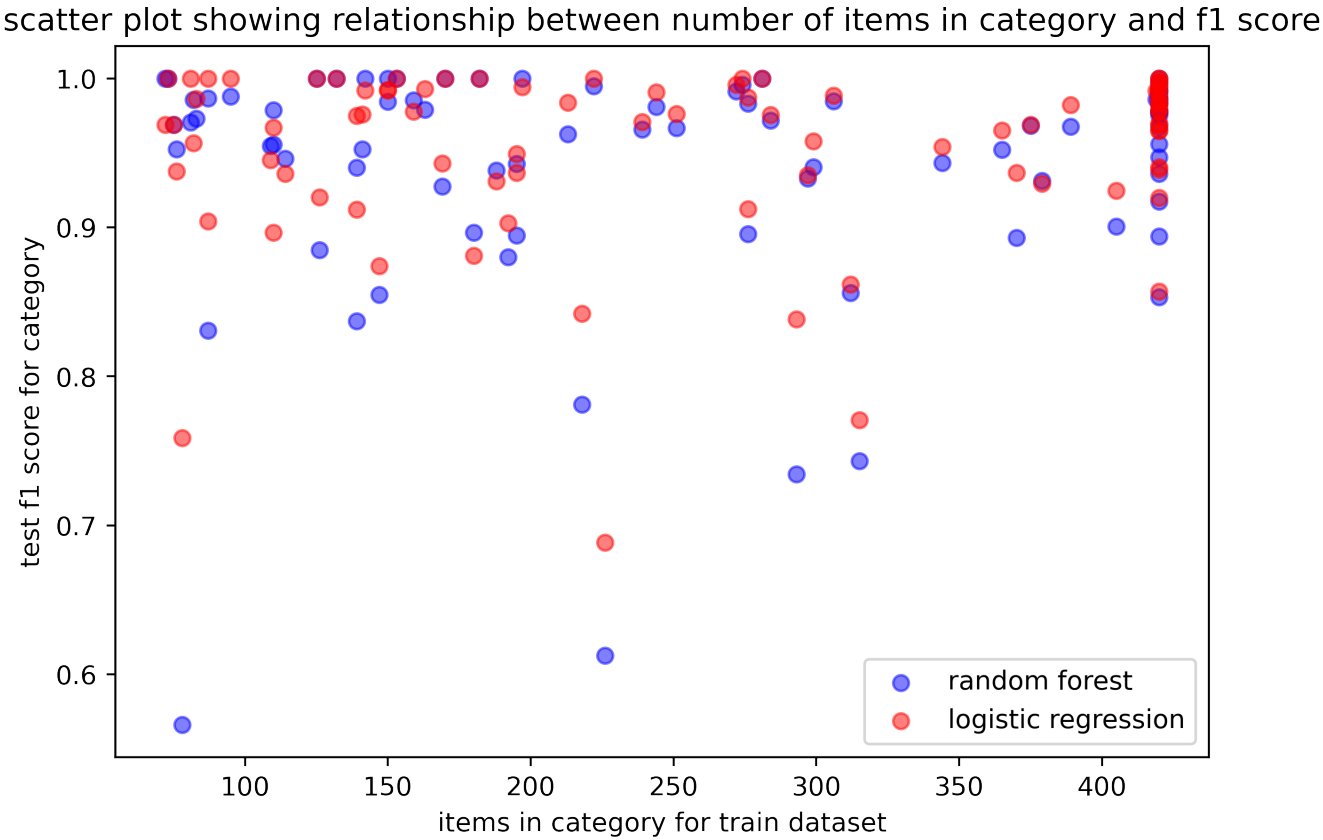
The vast majority of errors are assignments to adjacent or near-adjacent categories.

- This means that accuracy could be increased to 98-99% by grouping together these categories, if necessary.
- This also suggests that this problem could be approach in a hierarchical (2-step) way. First level classifying between broader categories Second level classifying finer distinctions, e.g. distinction between 35 (kovbasy vareno-kopcheni) and 36 (kovbasy syro-kopcheni).

Methods w/o cross-validation: More Performance Metrics

Insight III

No clear relationship between number of items in category in the train dataset, and f1 score on that category.



Methods w/o cross-validation: Comparing RF and LR

Insight IV

Lower-performing methods may still do better on some categories.

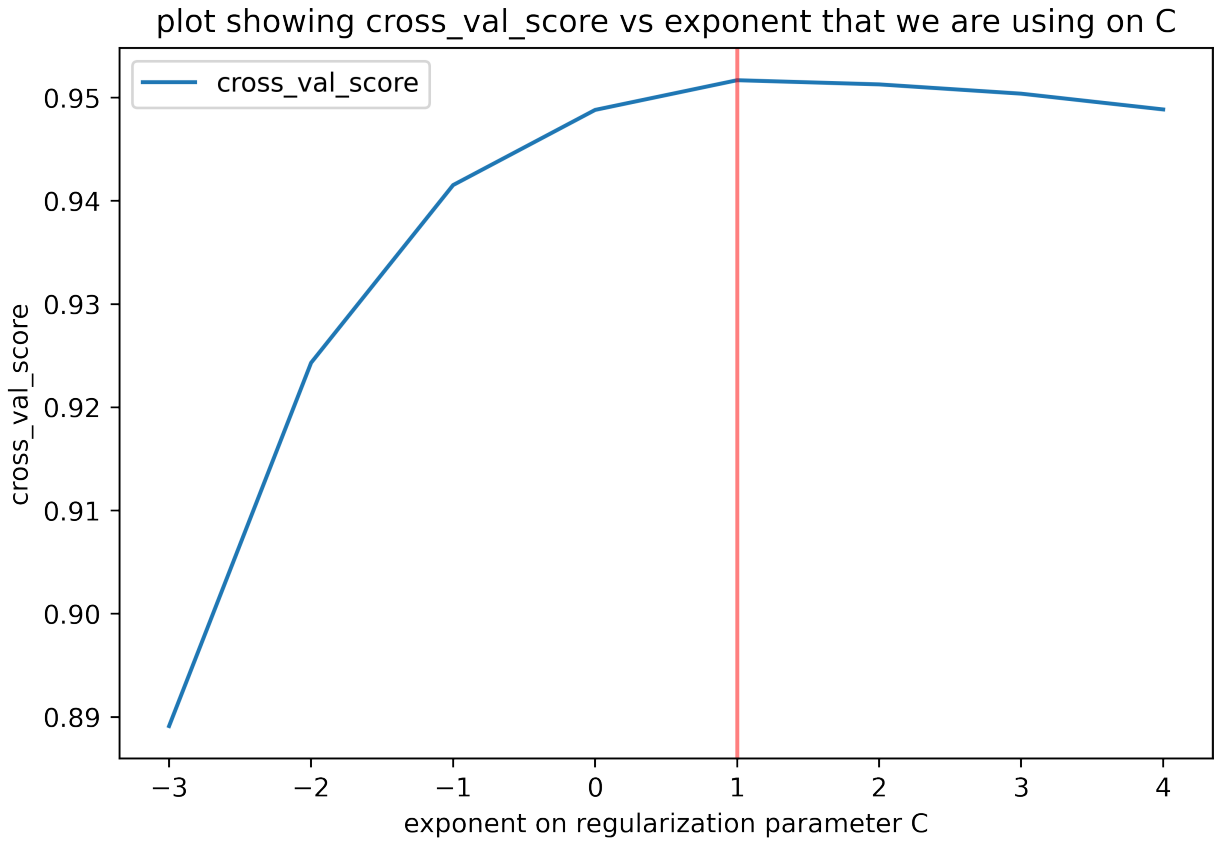
	f1_lr	f1_rf200
13.0	0.929231	0.931343
25.0	0.937500	0.952381
33.0	0.930818	0.938272
38.0	0.936170	0.946237
53.0	0.967033	0.978723
61.0	0.881119	0.896552
68.0	0.992248	1.000000
71.0	0.994083	1.000000
102.0	0.964770	0.967391
105.0	0.896552	0.955556
106.0	0.986072	0.986150
109.0	0.968750	1.000000
112.0	0.956522	0.985507
114.0	0.969529	0.975342
227.0	0.977778	0.985294
228.0	0.945055	0.954545
296.0	0.991870	1.000000
313.0	0.977654	0.977778

Hyperparameter Tuning

- So far all methods shown have been trained on default or close-to-default settings. But hyperparameter tuning would improve performance.
- Memory issues with 16gb RAM, requires dimensionality reduction (e.g. Principal Components Analysis). Will not describe here.

Hyperparameter Tuning: Logistic Regression

- There is only one parameter here (C), which makes this easy to tune over.



Next Steps

- **Modelling:** Consider fitting a 2-level model.
- **Data:** Add categories with < 100 products. This restriction was probably too strict.
- **Feature Engineering:** Consider incorporating new features (e.g. store categories) into model.
- **Computational:** Tune hyperparameters of random forest model using cross-validation + try XGBoost on a computer/server with more RAM and processing power.

References

Harms, A (2019)

A comprehensive view of machine learning techniques for CPI production

Statistics Netherlands Discussion Paper, November.

Sands, H (2020)

Automated classification of web-scraped clothing data in consumer price statistics

UK Office For National Statistics pp. 1–19. Available at: https://www.ons.gov.uk/economy/inflationandpriceindices/articles/automatedclassifica 09-01

Harms, Roberson (2021)

Applying Machine Learning for Automatic Product Categorization

US Census Bureau 37(2), pp. 395–410.

happy independence day! (for 24th), and...

Thank you for listening

Technical Appendix

- Precision:

$$\frac{TP}{TP + FP}$$

- Recall:

$$\frac{TP}{TP + FN}$$

- F1 Score:

$$\frac{2 * Precision * Recall}{Precision + Recall}$$