AUTOMATING PRODUCT REVIEWS FOR THE CONSUMER PRICE INDEX

Bureau of Labor Statistics

Office of Prices and Living Conditions

Nicole Shepler — Transportation Commodities Section Chief

CIVIC DIGITAL FELLOWSHIP

U.S. BUREAU OF LABOR STATISTICS

AVENA CHENG
University of California, Berkeley
Data Science

OVERVIEW

Background

- The CPI requires products to be reviewed by Commodity Analysts every month to calculate price indices.

Problem

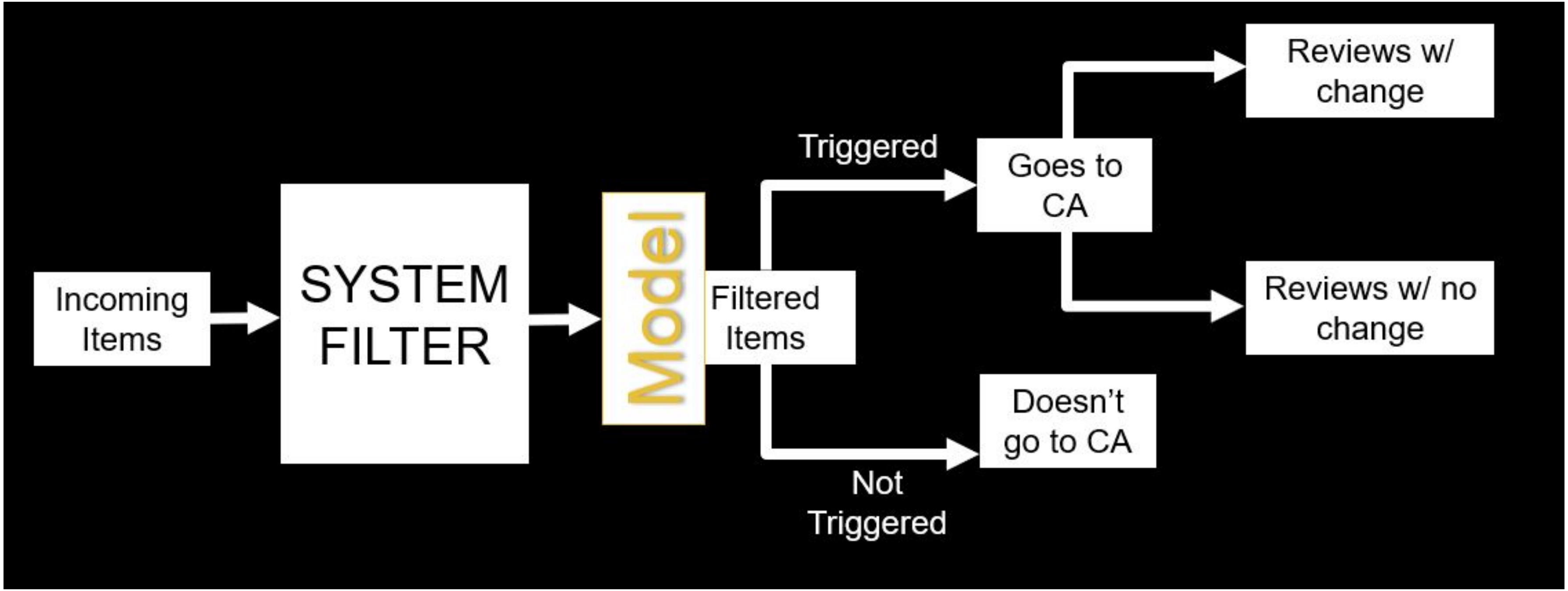
- Currently *< 25% of new items and < 12% of current items* require an analyst’s attention.
- The current system filter has an accuracy ~ 29%.

Goal/Impact

- Use machine learning to automate the review process and improve the current system’s filter to save time and money.

CIVIC DIGITAL FELLOWSHIP

INFRASTRUCTURE



CA = commodity analyst

CONTRIBUTIONS

Data

- Upsampled minority class to improve the model's learning
- Used TF-IDF and other feature engineering techniques to create more predictive variables
- Created a customized train-test function to handle duplicate entries* in both sets

Models

- Built 8 different Random Forest classifiers (one for each category in the CPI)
- Experimented with 6 other algorithms to determine the best predictive model
- Developed a post-prediction method to handle duplicate entries* while also improving recall

PERFORMANCE RESULTS

Section	F1 Score	Recall	Precision
Apparel	0.78	0.97	0.65
Housing	0.87	0.97	0.79
Food	0.79	0.90	0.70
Education/Communication	0.88	0.99	0.79
Transportation	0.84	0.97	0.75
Medical	0.93	1.0	0.88
Recreation	0.85	0.98	0.78
Goods & Services	0.87	0.96	0.81

Average Scores:

F1: 0.85
Recall: 0.97
Precision: 0.77

CIVIC DIGITAL FELLOWSHIP

OVERALL IMPACT

Total Avg Percentage of Reviews (Labor) Decreased: **85%**

Percentage decreased from each category:

- Apparel:	90%
- Housing:	87%
- Food:	94%
- Educ/Comm:	73%
- Transportation:	86%
- Medical:	78%
- Recreation:	83%
- Goods/Services:	88%

CIVIC DIGITAL FELLOWSHIP

Understanding the Metrics

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}}$$

$$\text{Recall} = \frac{\textit{True Positive}}{\textit{True Positive} + \boxed{\textit{False Negative}}}$$

```
Avg PR:     0.8327
F1-Score:   0.9082
Recall:     0.9883
Prec:       0.8402

CLASSIFICATION REPORT

             precision    recall  f1-score   support

          0       1.00      0.95      0.97      2647
          1       0.84      0.99      0.91       681

avg / total       0.96      0.96      0.96      3328
```

	No Review	Review
No Review	2519	128 (FP)
Review	8 (FN)	673

Predicted Result	0	1
Actual Result		
0	2519	128
1	8	673

CIVIC DIGITAL FELLOWSHIP

SOLUTION

1. Adjust Probability Threshold
2. Take the majority by rounding

id	P(No Review)	P(Review)	true	prediction	pred1, t = 0.5	pred2, t= 0.3
_13	0.696667	0.303333	0	1	0	1
_13	0.686667	0.313333	0	1	0	1
_13	0.696667	0.303333	0	1	0	1
_13	0.766667	0.233333	0	0	0	0
_13	0.766667	0.233333	0	0	0	0
_13	0.686667	0.313333	0	1	0	1
_13	0.696667	0.303333	0	1	0	1
_13	0.683333	0.316667	0	1	0	1
_13	0.696667	0.303333	0	1	0	1
_13	0.773333	0.226667	0	0	0	0
_13	0.773333	0.226667	0	0	0	0
_13	0.686667	0.313333	0	1	0	1

predicted "1": 8
Total # predicitons: 12

8/12 = 0.6667

round(0.6667) → 1

Final Prediction for id_13 = 1

Takeaway:

This will force the model to predict 1 more than 0, reducing the # of false negatives.

CIVIC DIGITAL FELLOWSHIP

DUPLICATE ENTRIES?

Some items have multiple rows
Each row has a different change

	id	predicted	actual
4239	8702729_001_201806_1	0	0
4240	8702729_001_201806_1	0	0
4241	8702729_001_201806_1	0	0
4242	8702729_001_201806_1	0	0
4243	8702729_001_201806_1	0	0
4244	8702729_001_201806_1	0	0
4245	8702729_001_201806_1	0	0
4246	8702729_001_201806_1	0	0
4247	8702729_001_201806_1	0	0
4248	8702729_001_201806_1	0	0
4249	8702729_001_201806_1	0	0
4250	8702729_001_201806_1	0	0
4251	8702729_001_201806_1	0	0
4252	8702729_001_201806_1	0	0

This can lead to different predictions.

	id	predicted	actual	error
1177	8200697_003_201902_4	0	1	True
1178	8200697_003_201902_4	1	1	False
1179	8200697_003_201902_4	1	1	False

CIVIC DIGITAL FELLOWSHIP