# Women Fashion Review

Khoon Ching Wong

10-August-2022

https://github.com/wongkhoon/Coursera/tree/main/Advanced%20Data%20Science%20with%20IBM/Advanced%20Data%20Science%20Capstone

# Use Case

## Fashion recommendation prediction according to textual review from consumers

# Data Source

## Name

Women's E-Commerce Clothing Reviews

## Source

Kaggle[1]

## Reason

- Publicly accessible and available
- Data contains both qualitative and quantitative

# EDA

# Data Set

## String

- Title
- Review
- Division Name
- Department Name
- Class Name

## Numeric

- Clothing ID
- Age
- Rating
  (1 Worst, to 5 Best)
- Positive Feedback Count
- Recommended IND
  (1= recommended, 0=not recommended)

# Data Set

```
---- ORIGINAL DATA ----

Data shape: Total of 23486 entries, 10 data columns

+----------------+-----+-------------------+
|Recommended IND|count|          proportion|
+----------------+-----+-------------------+
|               1|19314| 82.23622583666865|
|               0| 4172|17.763774163331348|
+----------------+-----+-------------------+

---- Post random sampling without replacement ----

Data shape: Total of 6000 entries, 10 data columns

+----------------+-----+-------------------+
|Recommended IND|count|          proportion|
+----------------+-----+-------------------+
|               1| 4929|              82.15|
|               0| 1071|17.849999999999998|
+----------------+-----+-------------------+
```

# Data Set

**---- POST-SAMPLING DATA ----**

**First five entries:**

| Clothing ID | Age | Title | Review Text | Rating | Recommended IND | Positive Feedback Count | Division Name | Department Name | Class Name |
|---|---|---|---|---|---|---|---|---|---|
| 858 | 39 | Shimmer, surprisi... | I ordered this in... | 4 | 1 | 4 | General Petite | Tops | Knits |
| 1065 | 34 | You need to be at... | Material and colo... | 3 | 1 | 2 | General | Bottoms | Pants |
| 1120 | 32 | Super cute and cozy | A flattering, sup... | 5 | 1 | 0 | General | Jackets | Outerwear |
| 949 | 33 | Huge disappointment | I have been waiti... | 2 | 0 | 0 | General | Tops | Sweaters |
| 1003 | 31 | Loved, but returned | The colors weren'... | 4 | 1 | 0 | General | Bottoms | Skirts |

only showing top 5 rows

# Data Quality Assessment

## Duplicates

Two identified

- Removed

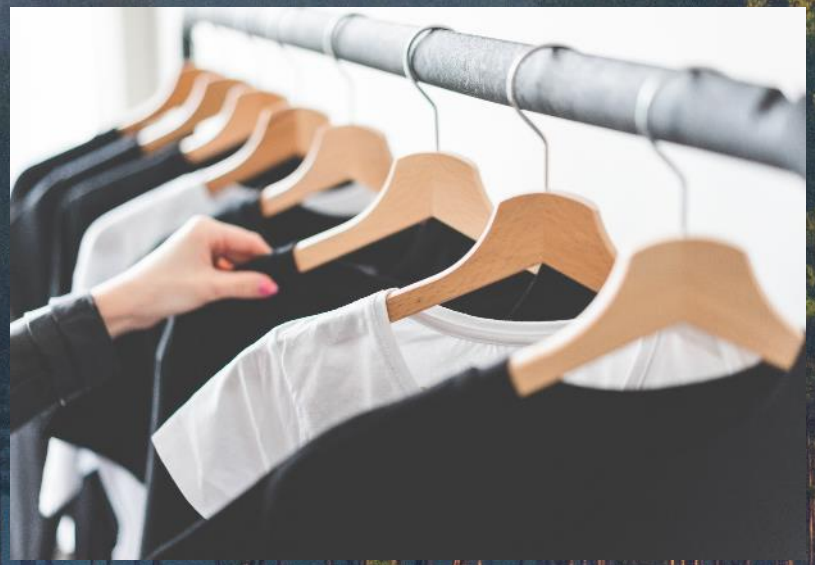
## Missing data

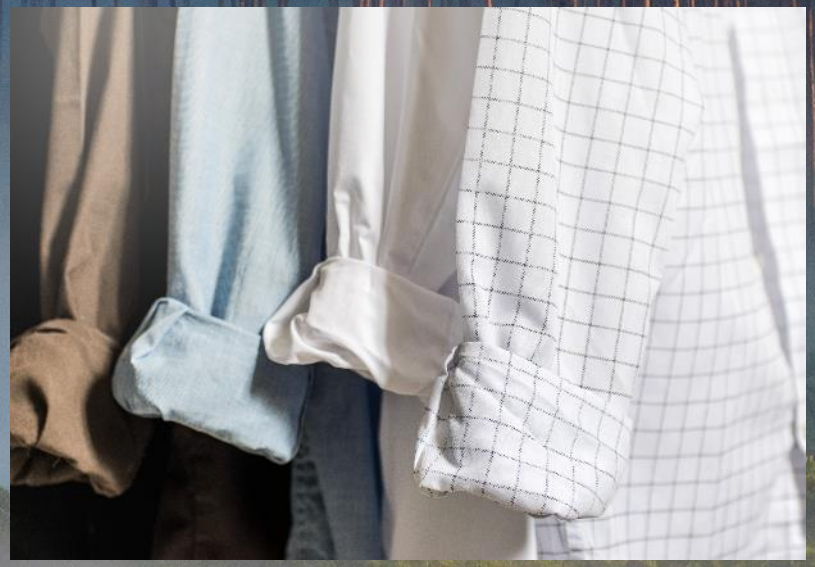
String columns, except for Review and Title

- Imputed by “unknown”

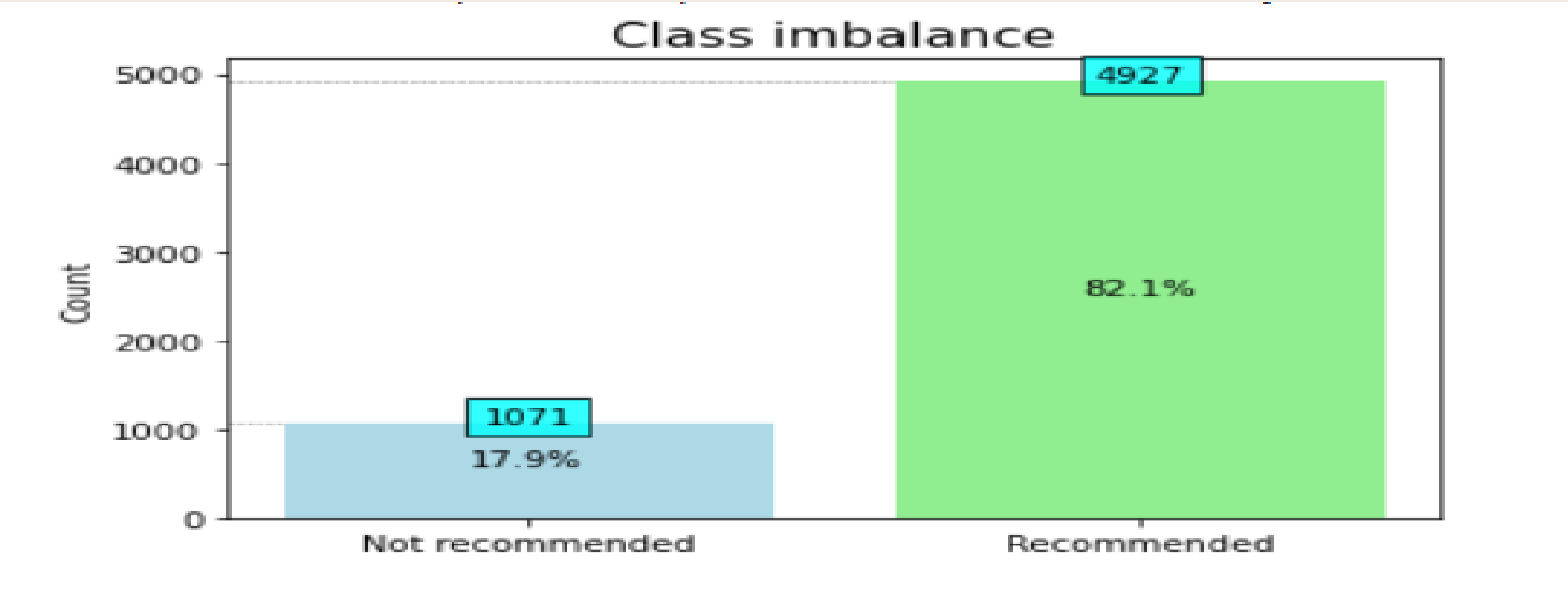
Numeric columns

- No missing





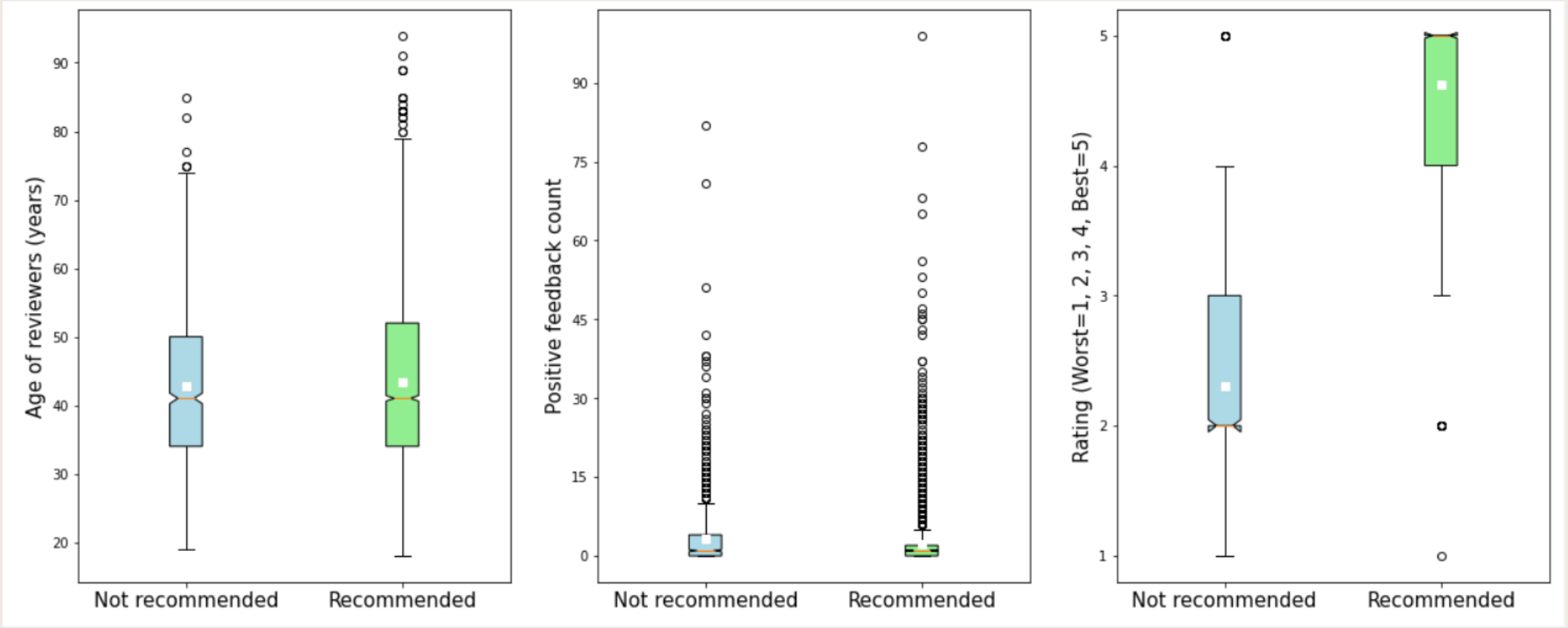
# Data Exploration/Visualization



# Class Weights

```
# https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html
# Estimate class weights for unbalanced datasets
weight0=train.count()/((train.select('Recommended IND').distinct().count())*(train.filter(train["Recommended IND"] == 0).count()))
weight1=train.count()/((train.select('Recommended IND').distinct().count())*(train.filter(train["Recommended IND"] == 1).count()))
```

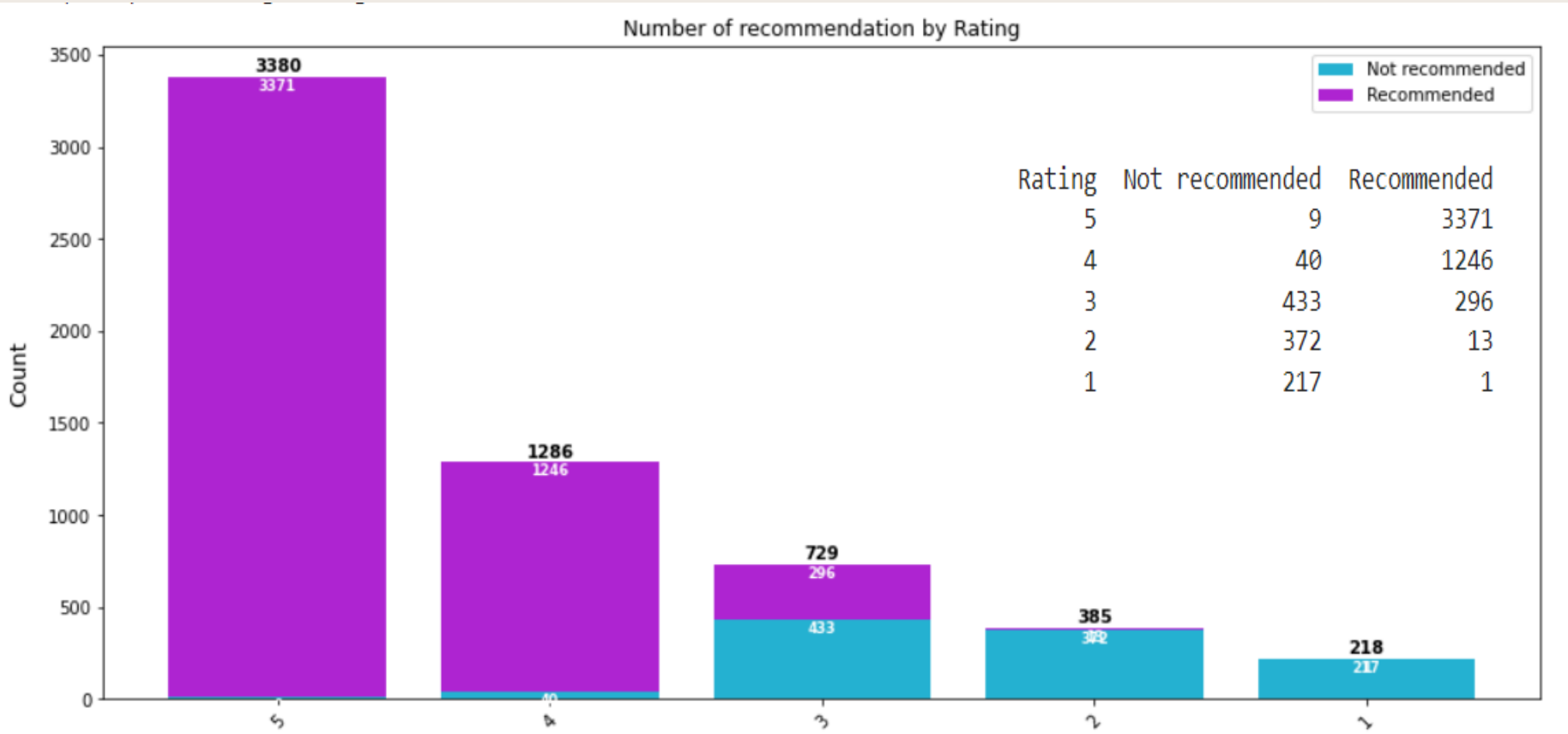
# Data Exploration/Visualization



# Data Exploration/Visualization

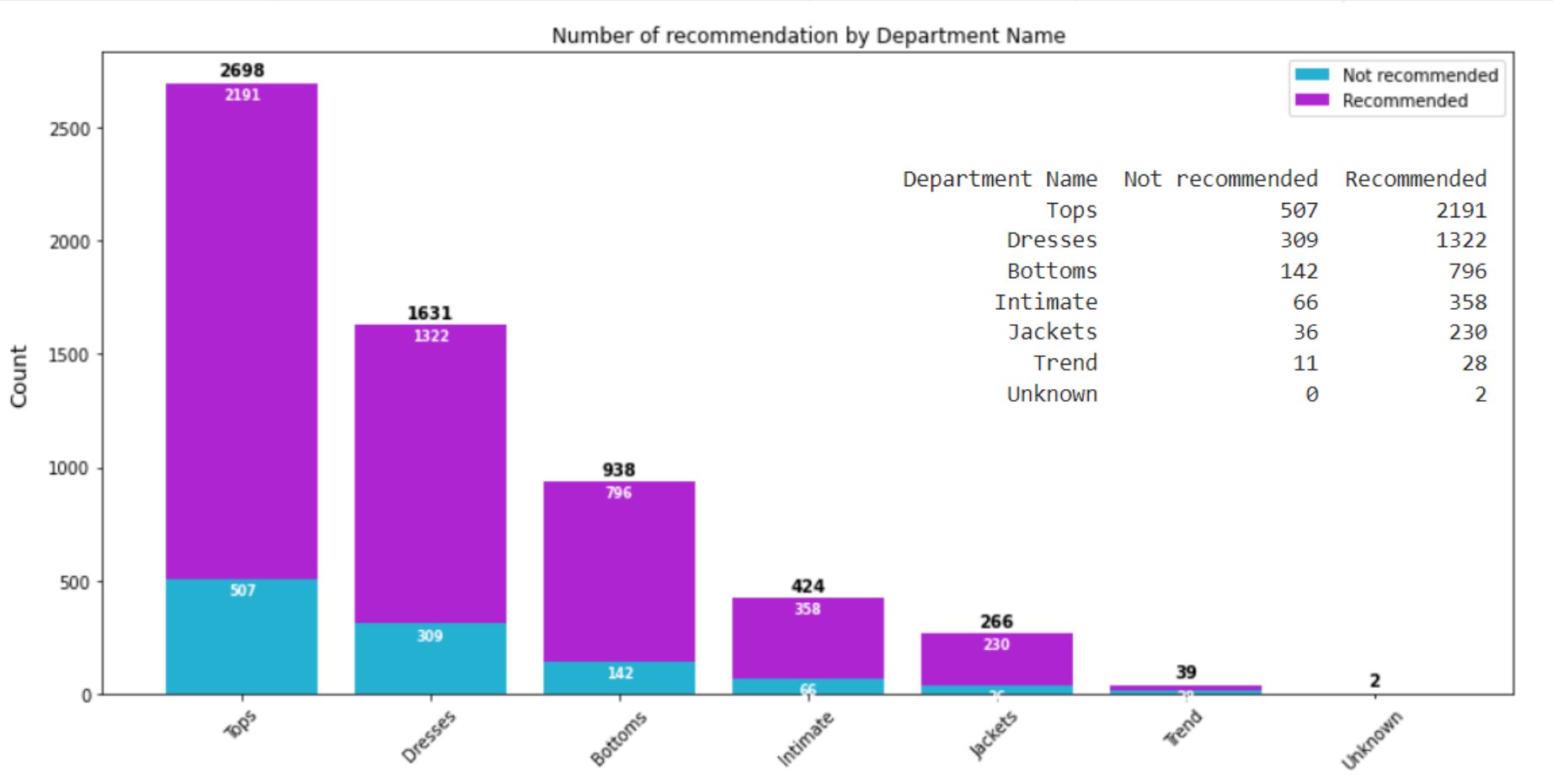
```
Pearson correlation:

Overall between numerical features
Age versus
     Rating                                          :   0.0285
     Positive Feedback Count                         :   0.0556
Rating versus Positive Feedback Count:  -0.0669

Not recommended group
Age versus
     Rating                                          :  -0.0232
     Positive Feedback Count                         :   0.0702
Rating versus Positive Feedback Count:  -0.0155

Recommended group
Age versus
     Rating                                          :   0.0324
     Positive Feedback Count                         :   0.0545
Rating versus Positive Feedback Count:  -0.0102
```

# Data Exploration/Visualization



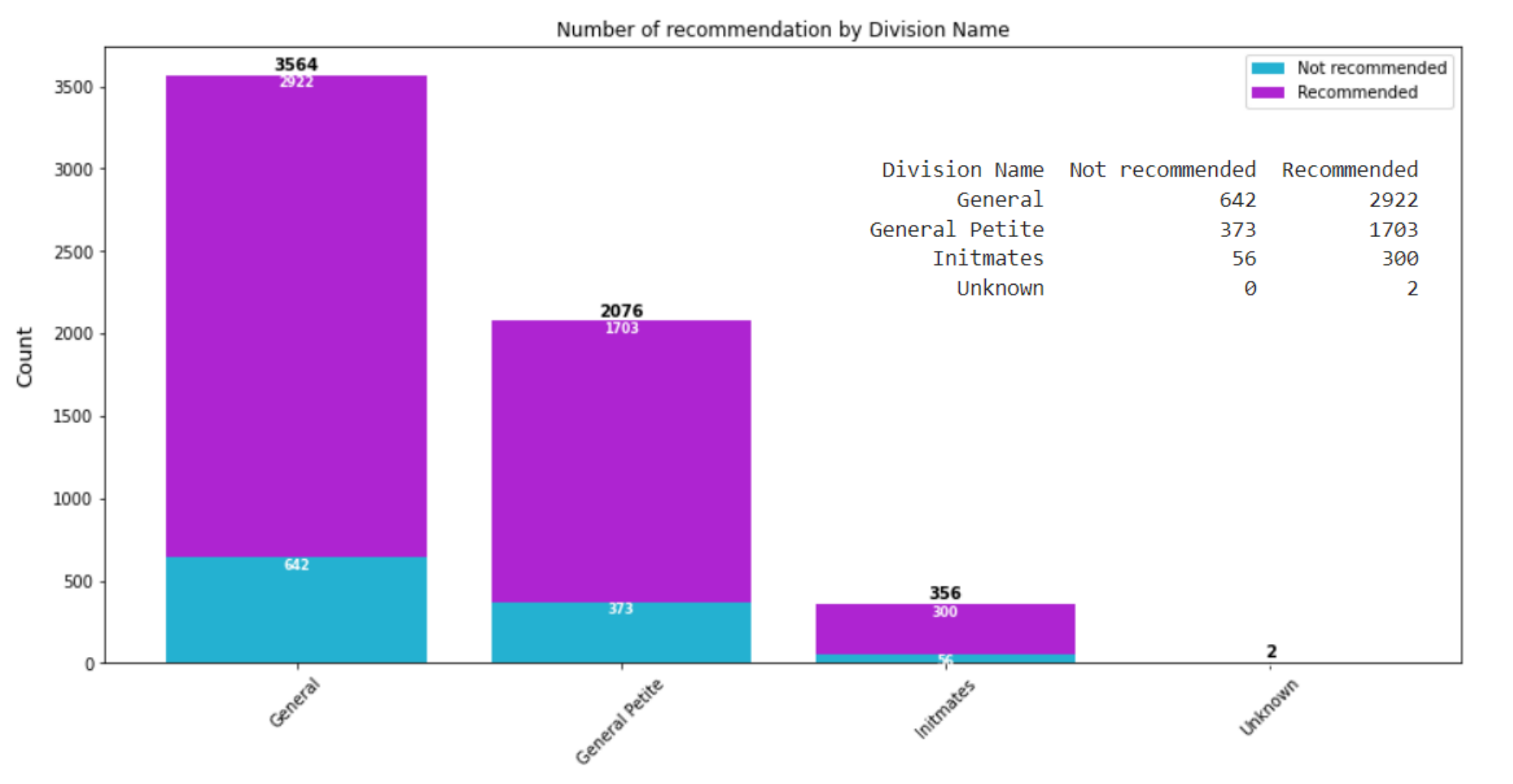

| Rating | Not recommended | Recommended |
|---|---|---|
| 5 | 9 | 3371 |
| 4 | 40 | 1246 |
| 3 | 433 | 296 |
| 2 | 372 | 13 |
| 1 | 217 | 1 |

# Data Exploration/Visualization



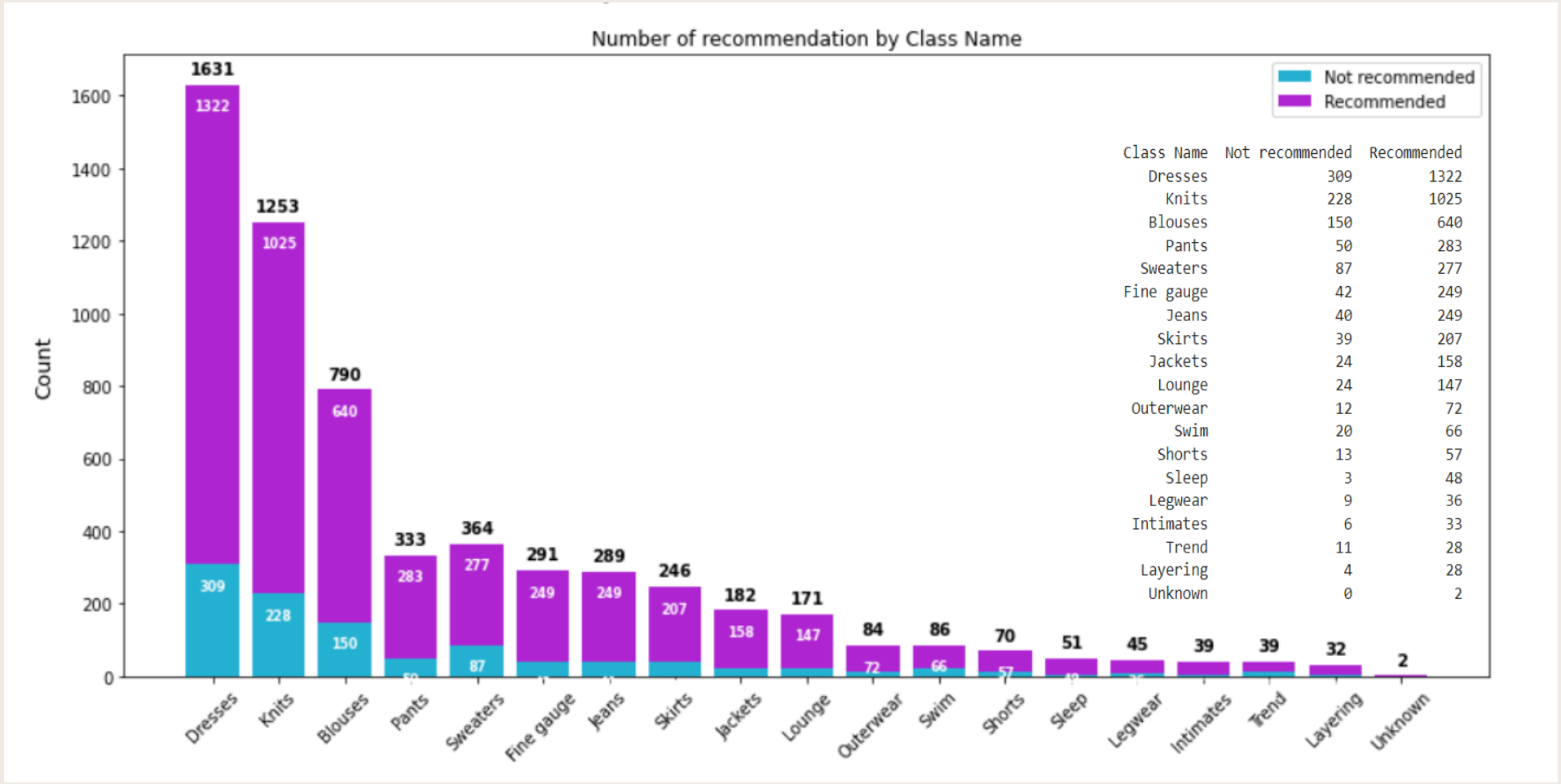

| Department Name | Not recommended | Recommended |
|---|---|---|
| Tops | 507 | 2191 |
| Dresses | 309 | 1322 |
| Bottoms | 142 | 796 |
| Intimate | 66 | 358 |
| Jackets | 36 | 230 |
| Trend | 11 | 28 |
| Unknown | 0 | 2 |

# Data Exploration/Visualization




| Division Name | Not recommended | Recommended |
|---|---|---|
| General | 642 | 2922 |
| General Petite | 373 | 1703 |
| Initmates | 56 | 300 |
| Unknown | 0 | 2 |

# Data Exploration/Visualization




| Class Name | Not recommended | Recommended |
|---|---|---|
| Dresses | 309 | 1322 |
| Knits | 228 | 1025 |
| Blouses | 150 | 640 |
| Pants | 50 | 283 |
| Sweaters | 87 | 277 |
| Fine gauge | 42 | 249 |
| Jeans | 40 | 249 |
| Skirts | 39 | 207 |
| Jackets | 24 | 158 |
| Lounge | 24 | 147 |
| Outerwear | 12 | 72 |
| Swim | 20 | 66 |
| Shorts | 13 | 57 |
| Sleep | 3 | 48 |
| Legwear | 9 | 36 |
| Intimates | 6 | 33 |
| Trend | 11 | 28 |
| Layering | 4 | 28 |
| Unknown | 0 | 2 |

# Feature Engineering

**Incorporating both text and non-text features**[2]

```
This item comes from {"Department Name} department
and division is {“Division Name”}, and is classifi
ed under {“Class Name”}. There are {"Positive Feed
back Count”} customers who found this review
positive. I am {“Age”} years old. I rate this item
{“Rating”} out of 5 stars.
```

# Classification algorithms

# Classifiers

## Simple logistic regression[3]

- Pyspark 3.3.0
- Baseline model
- Numerical features (Age, rating and Positive Feedback Count)
- Grid search for hyperparameter tuning
- 3-fold cross validation

## Logistic regression[3]

- Pyspark 3.3.0
- TF-IDF
- Grid search for hyperparameter tuning
- 3-fold cross validation

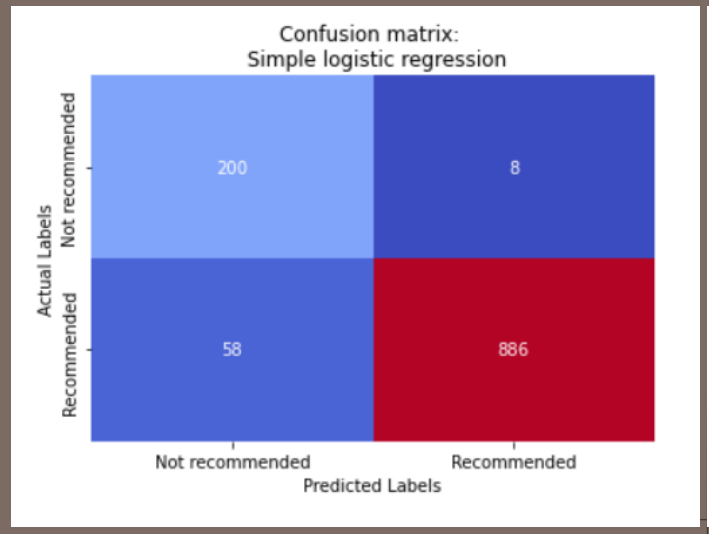
## Random forest[3]

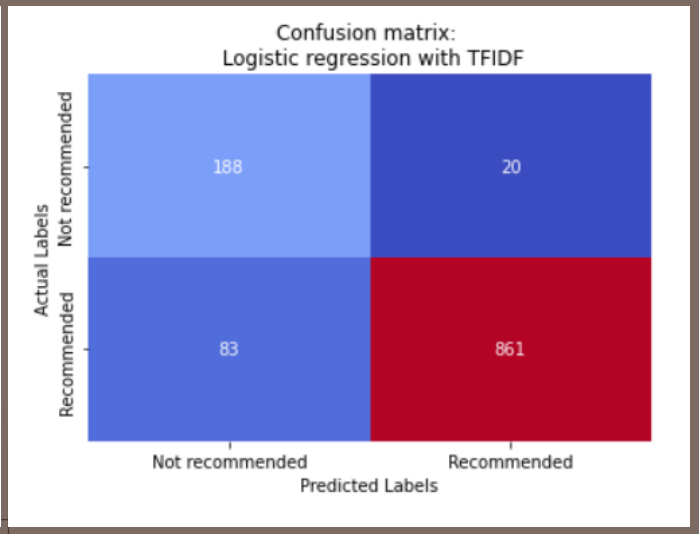
- Pyspark 3.3.0
- TF-IDF
- Grid search for hyperparameter tuning
- 3-fold cross validation

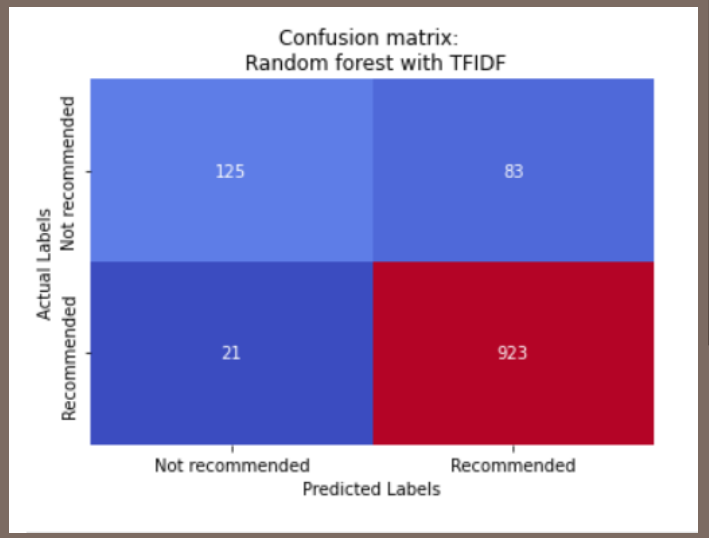
## Deep learning

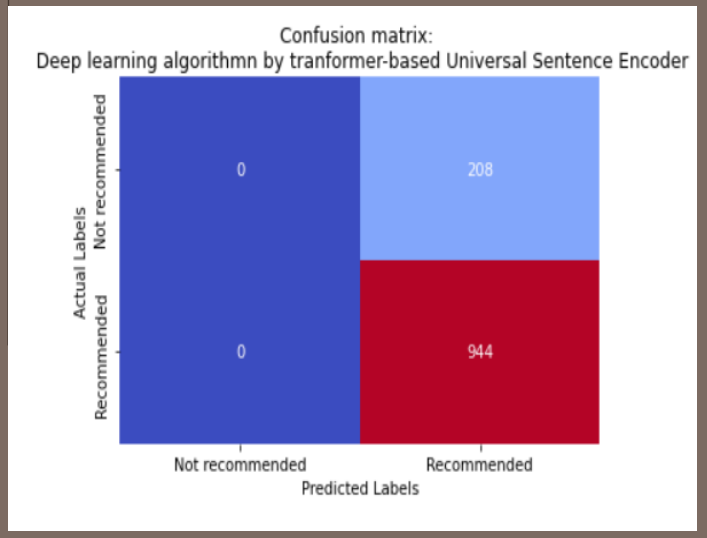
- Spark NLP 4.0.1 [4]
- Transformer-based Universal Sentence Encoder[5, 6, 7]
- Default hyperparameters[8]

# Confusion Matrix



# Performance metrics

| Metrics | Simple Logistic Regression | Logistic Regression (TF-IDF) | Random Forest (TF-IDF) | Deep Learning (Transformer-based Universal Sentence Encoder) |
|---|---|---|---|---|
| Accuracy | 0.94 | 0.91 | 0.91 | 0.82 |
| Precision | 0.99 | 0.98 | 0.92 | 0.82 |
| Recall | 0.94 | 0.91 | 0.98 | 1.00 |
| F1-score | 0.96 | 0.94 | 0.95 | 0.90 |
| | | | | |
| Area under the receiver operating characteristic (ROC) curve | 0.9794 | 0.9556 | 0.9856 | 0.5000 |
| Area under the precision-recall curve | 0.9957 | 0.9895 | 0.9906 | 0.8194 |
| | | | | |
| Training time taken (seconds) | 82.318608045578 | 397.9566535949707 | 9241.370339393616 | 420.804 |

# Conclusion

Studying online reviews with machine learning models allows businesses to develop their services with ideas and to meet consumer satisfaction through their buying trends and behavior.

# Resources/References


1. https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews
2. https://mccormickml.com/2021/06/29/combining-categorical-numerical-features-with-bert/
3. https://spark.apache.org/docs/latest/ml-tuning.html
4. https://towardsdatascience.com/text-classification-in-spark-nlp-with-bert-and-universal-sentence-encoders-e644d618ca32
5. https://nlp.johnsnowlabs.com/2020/04/17/tfhub_use.html
6. https://nlp.johnsnowlabs.com/docs/en/transformers
7. https://tfhub.dev/google/universal-sentence-encoder-large/5
8. https://nlp.johnsnowlabs.com/api/com/johnsnowlabs/nlp/annotators/classifier/dl/ClassifierDLApproach
9. https://www.coursera.org/learn/advanced-data-science-capstone/home/assignments


# Thank you