# Text Classification of Ecommerce Clothing Line Reviews

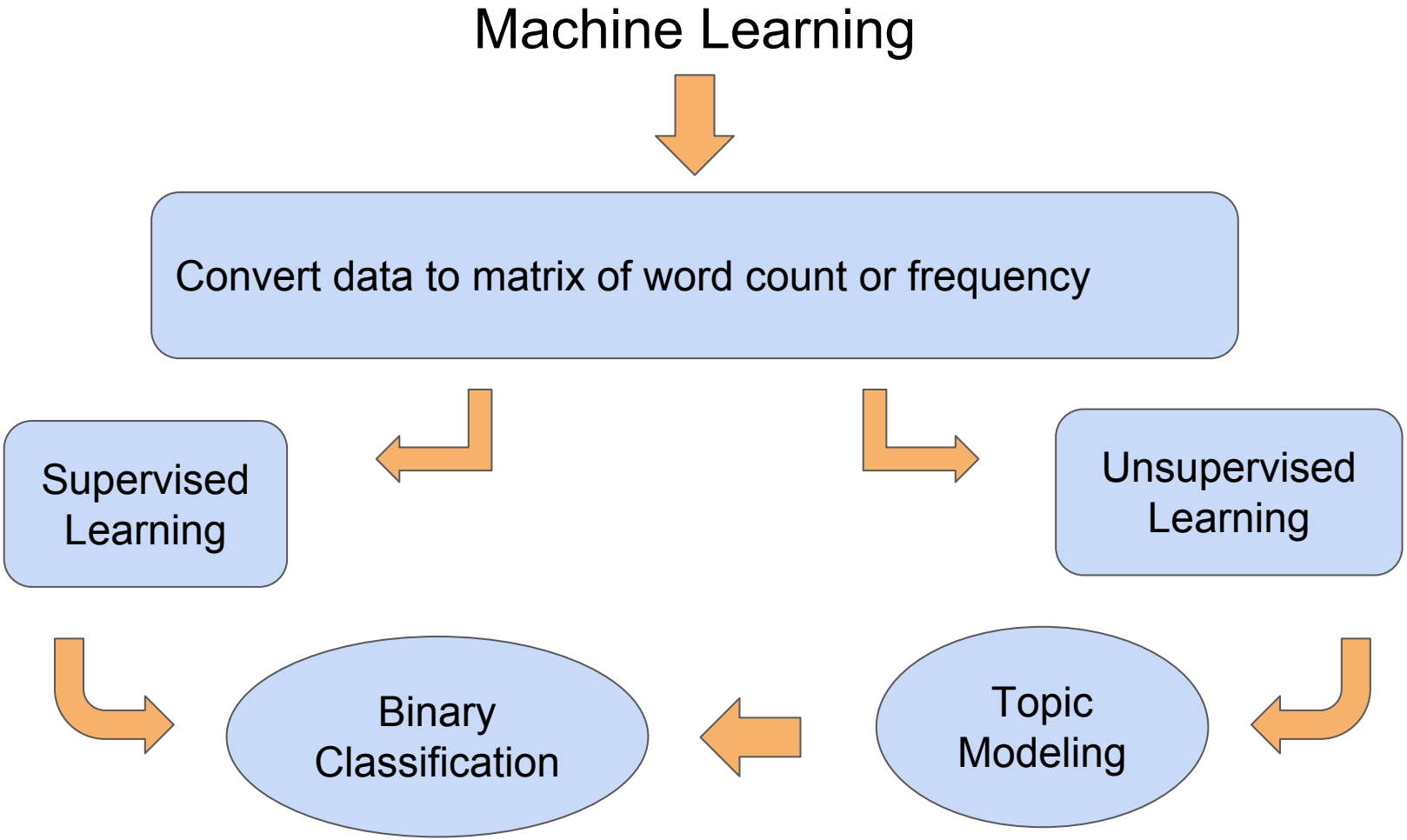
Dona Ray

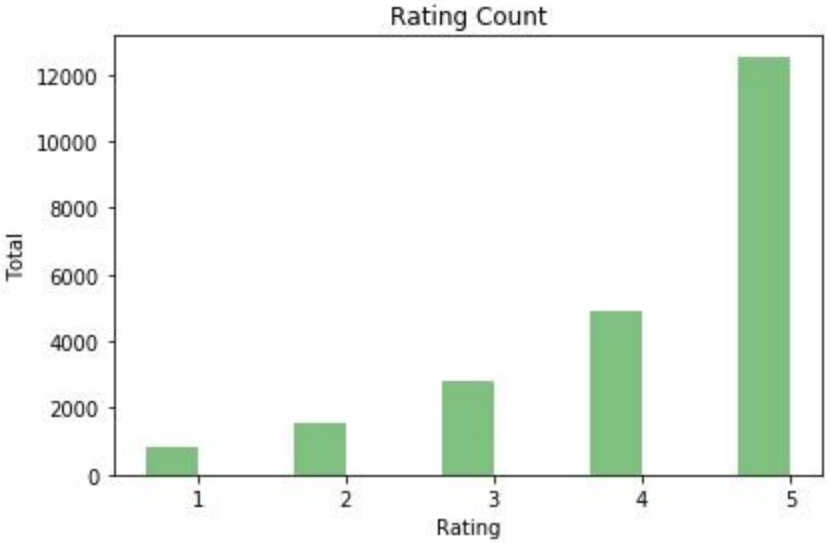
# The Problem

Predict rating based on text reviews

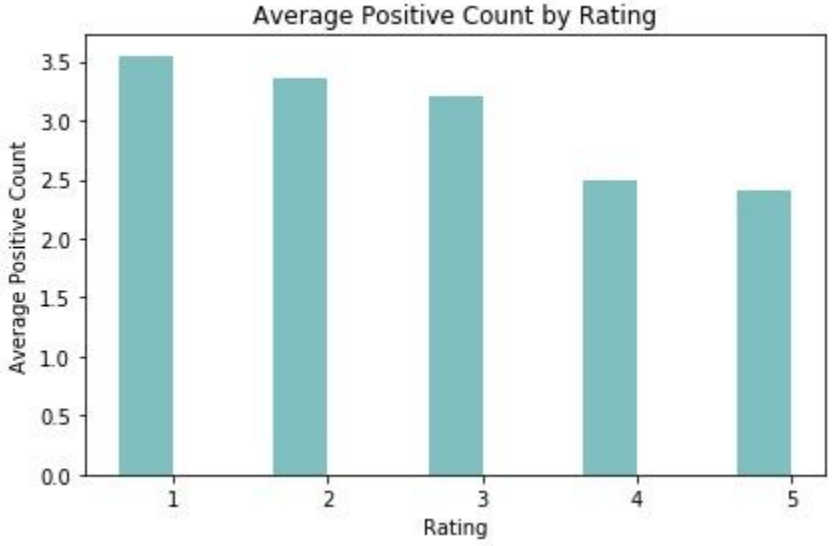
Text data of customer reviews for an online Women’s Clothing line

Ratings are on a 5-point scale

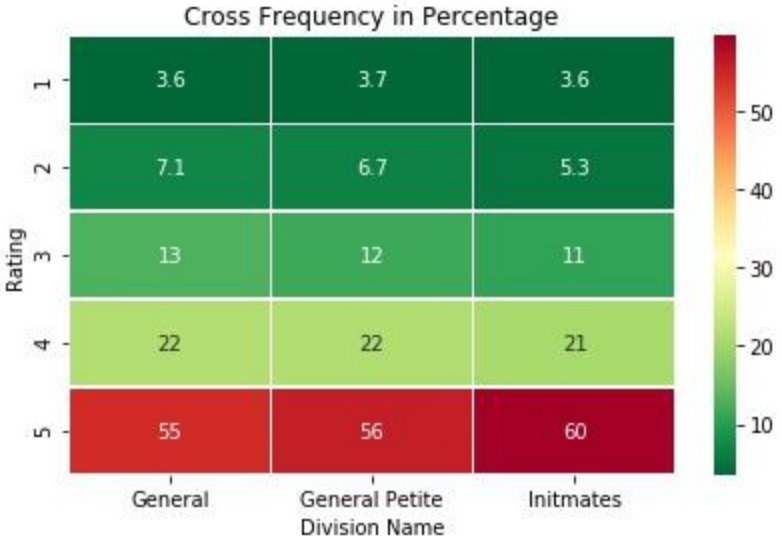

Machine Learning
Convert data to matrix of word count or frequency
Supervised Learning
Unsupervised Learning
Binary Classification
Topic Modeling

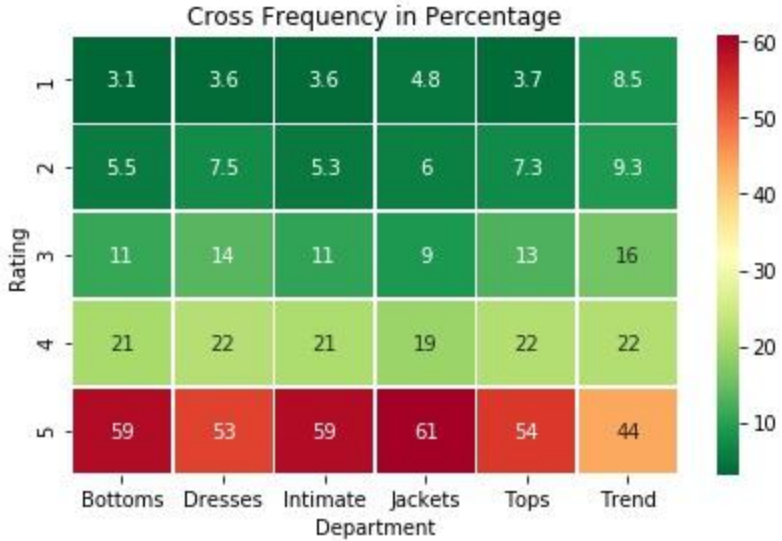



**77% ratings are 4 and 5 stars!**

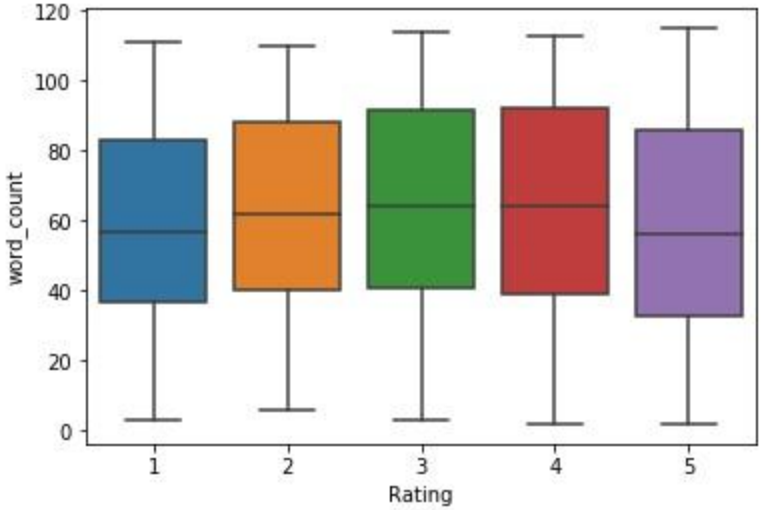


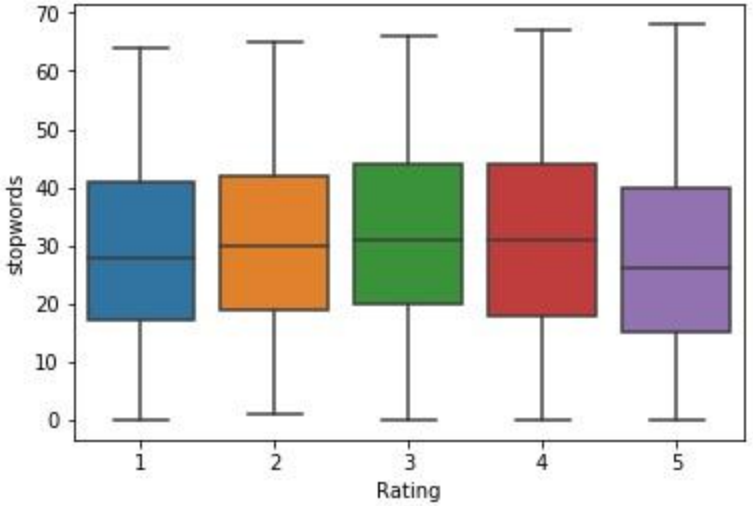

**1-star rating has the highest average positive feedback count**




**Distribution of percentage of ratings by Division and Department is uniform**




**Rating 3 has the highest word and stopword count!**

| | Total Words | Unique Words |
|---|---|---|
| **Overall Total** | 1,363,325 | 40,071 |
| **Convert to Lower Case** | 1,362,913 | 37,720 |
| **Remove Punctuation** | 1,362,913 | 19,386 |
| **Remove Stopwords** | 668,290 | 19,244 |
| **Remove Numerics** | 652,217 | 18,840 |
| **Stem words** | 652,217 | 17,483 |
| **Remove words that appear once** | 642,694 | 7,960 |

- Convert text data to numpy matrix
  - CountVectorizer
  - TF-IDF
- K-fold Cross-Validation
  - 80% Training and 20% Test data
  - GridSearch for Hyperparameter Tuning
- Supervised Learning
  - Multinomial Naive Bayes
  - Logistic Regression
  - Random Forest
  - SVM
- Unsupervised Learning
  - Topic Modeling
    - LDA
    - LSI

# Model Comparisons

| Model | Accuracy (Training) | Accuracy (Testing) | ROC |
|---|---|---|---|
| **Multinomial NB** | 0.99 | 0.88 | 0.80 |
| **Random Forest** | 0.82 | 0.82 | 0.82 |
| **Logistic Regression** | 1.00 | 0.88 | 0.82 |
| **SVM** | 0.97 | 0.87 | 0.85 |
| **TF-IDF (Multinomial)** | 0.93 | 0.87 | 0.76 |
| **LDA (Logistic)** | 0.78 | 0.77 | 0.77 |
| **LSA (Logistic)** | 0.87 | 0.85 | 0.85 |

# Key Findings

**Supervised learning methods do well classifying ‘good’ and ‘bad’ reviews**

**Achieve an accuracy score of 0.88 on testing data**

**SVM does the best with ROC score of 0.85**

**Word counts do better than frequencies since document size is small**

**Reducing dimension using topic modeling (LSI) gives a ROC score of 0.84**