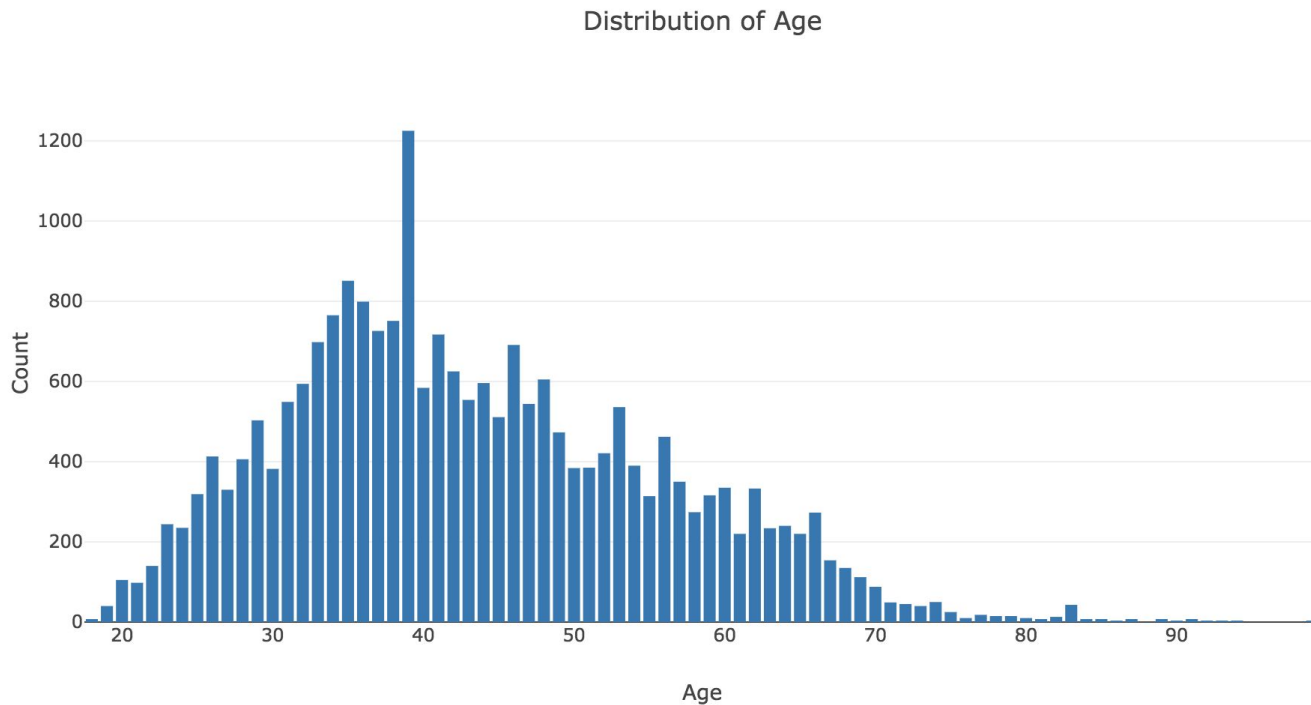


Predicting Age from E-commerce clothing Reviews

Deena Liz John

Data

- 22,628 reviews and product information available
- 845 null rows were excluded

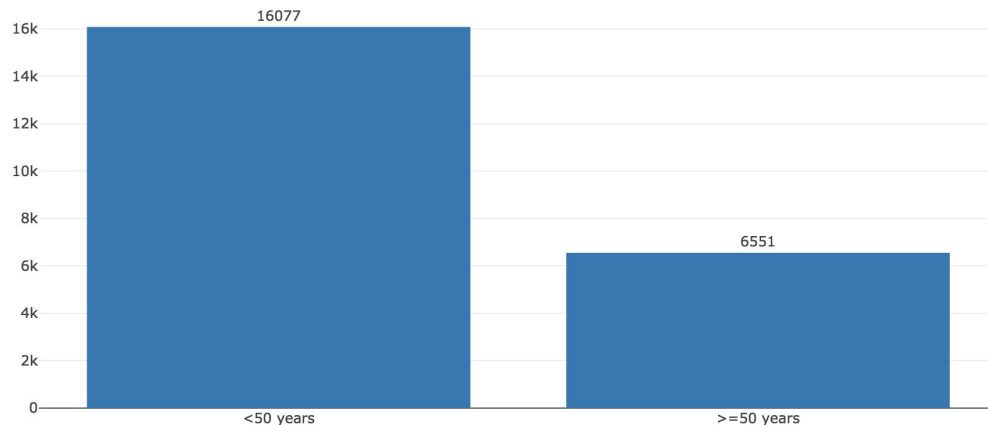


Age Bucketing

- Bucketed age into two groups: < 50 years and ≥ 50 years

	Age	Review Text
0	33	Absolutely wonderful - silky and sexy and comf...
1	34	Love this dress! it's sooo pretty. i happene...
2	60	I had such high hopes for this dress and reall...
3	50	I love, love, love this jumpsuit it's fun, fl...
4	47	This shirt is very flattering to all due to th...

Distribution of age group



[illegible]

fabric length
good long look well large
sweater small tried flattering
great model longer black little
one skirt gorgeous
fit right blouse first looks
soft waist comfortable size
medium lbs see enough light
tight short runs quality summer
online work design sleeves perfectly
jeans wearing true top blue shirt bit
jacket lovely fits weight cut
need going many super front way white
store big price still think fall really wanted
even usually petite color feel
back colors pants better purchased
nice style beautiful pretty
like retailer try ordered wear
much dress bought
loved cute made
perfect
love

Word Embeddings and Feature Engineering

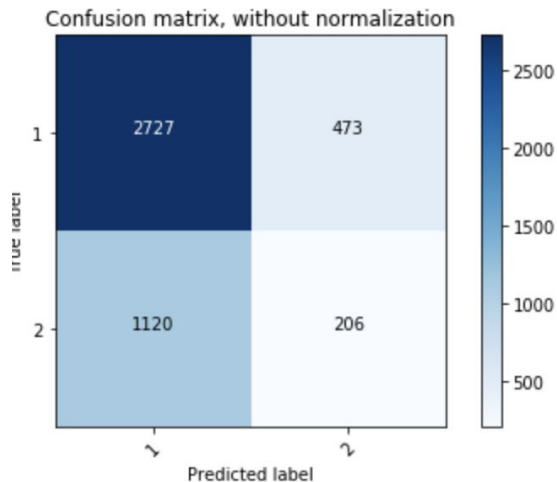
- Cleaned text:
 - Removed stop words and punctuations to remove randomness
 - Considered words of length > 2
- Used Glove embeddings of 200 dimensions to encode reviews
- Calculated mean embeddings of each review
- Feature Engineering: Added columns like Department Name, Division Name, Rating and Clothing ID

Classification Models

- Divided into 80-20 train and test set

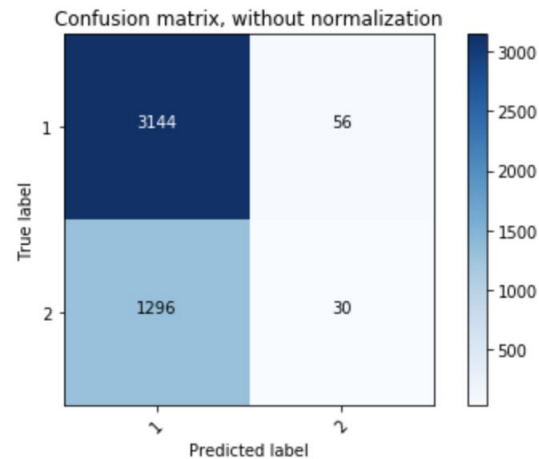
Random Forest

Accuracy: 64%



Logistic Regression

Accuracy: 70%



Classification Model 3: Neural Networks using Keras

```
model.add(Dense(input_dim=200, output_dim=100, activation='sigmoid'))
```

```
model.add(Dense(output_dim=6, activation='softmax'))
```

- Used only word embeddings of reviews
- Sequential model with two layers
- Batch size = 32, epochs = 5
- Metrics:
 - Accuracy: 0.80
 - Loss: 0.39