Capstone: What makes a Yelp review useful?

Problem Statement

Yelp is a platform that allows users to make reservations, leave reviews, and find businesses. In addition, users can mark which reviews they find useful. Reviews can be both very helpful and very harmful for businesses. Being able to determine what makes a useful review can help businesses create better, more targeted listings.

What is a “Yelp”?

- Find and connect with businesses
  - Appointments
  - Waitlists
  - Reservations
- Business info
  - Hours of operation
  - Location
- Reviews, photos
  - Funny/Cool/Useful

The Data!

- Available from https://www.yelp.com/dataset
- 6,990,280 reviews
- 150,346 businesses
- 200,100 pictures

Data Preparation

- Data was randomly sampled to approx 20,000 records
- No imputation was necessary
- New lines and punctuation were cleaned up
- Outliers removed
    - Useful column had extreme outliers

Vectorization

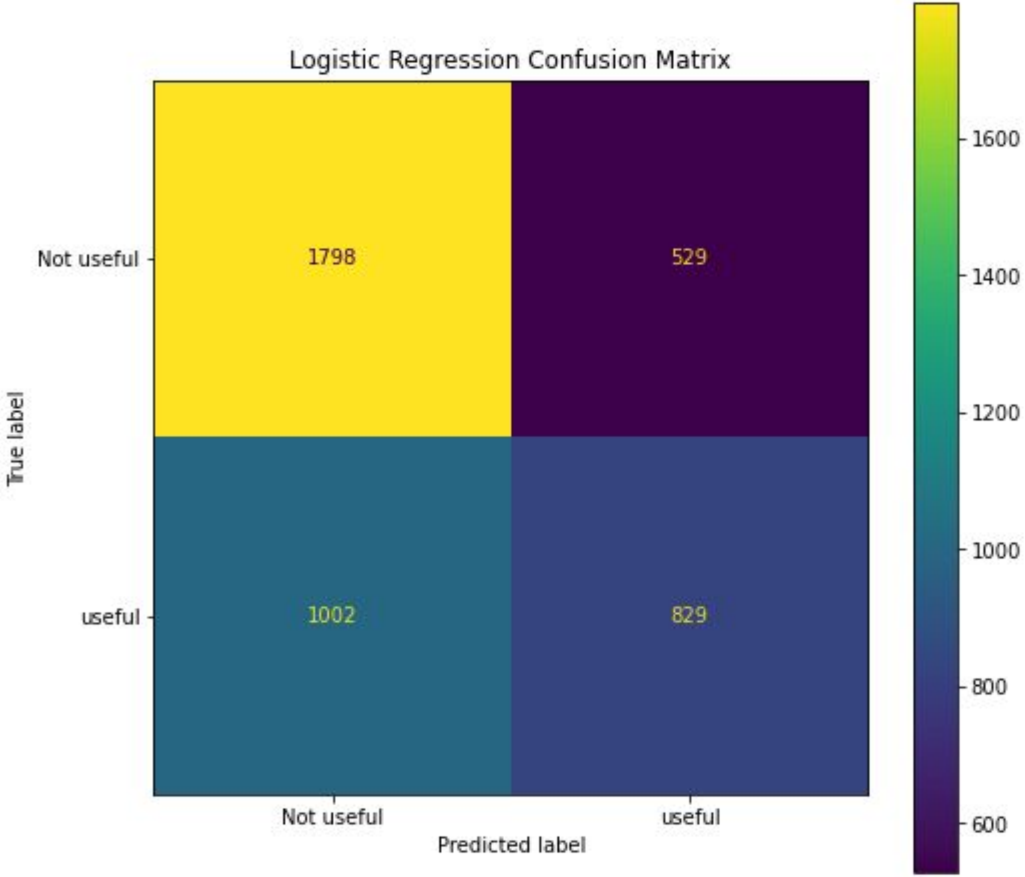
- TF-IDF
  - Term frequency-inverse document frequency
  - A measure of how relevant a word is in a document or corpus
  - Max depth: 200

LassoCV

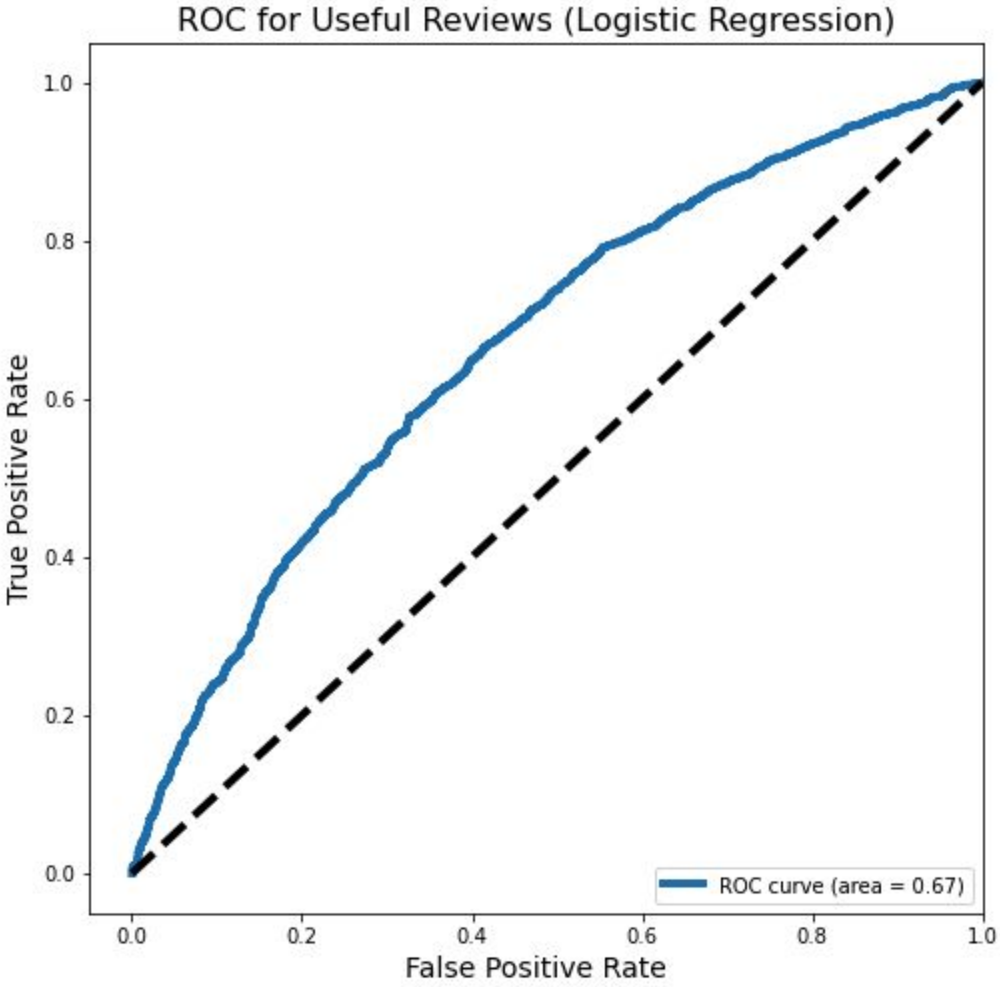
- Reduce columns / multicolinearity

Logistic Regression Model

- Parameters
  - C=100
  - Solver: liblinear
  - Penalty: l2
- Metrics
  - 0.62 +/- 0.036
  - Precision: 0.61
  - Recall: 0.45
  - F1-score: 0.52

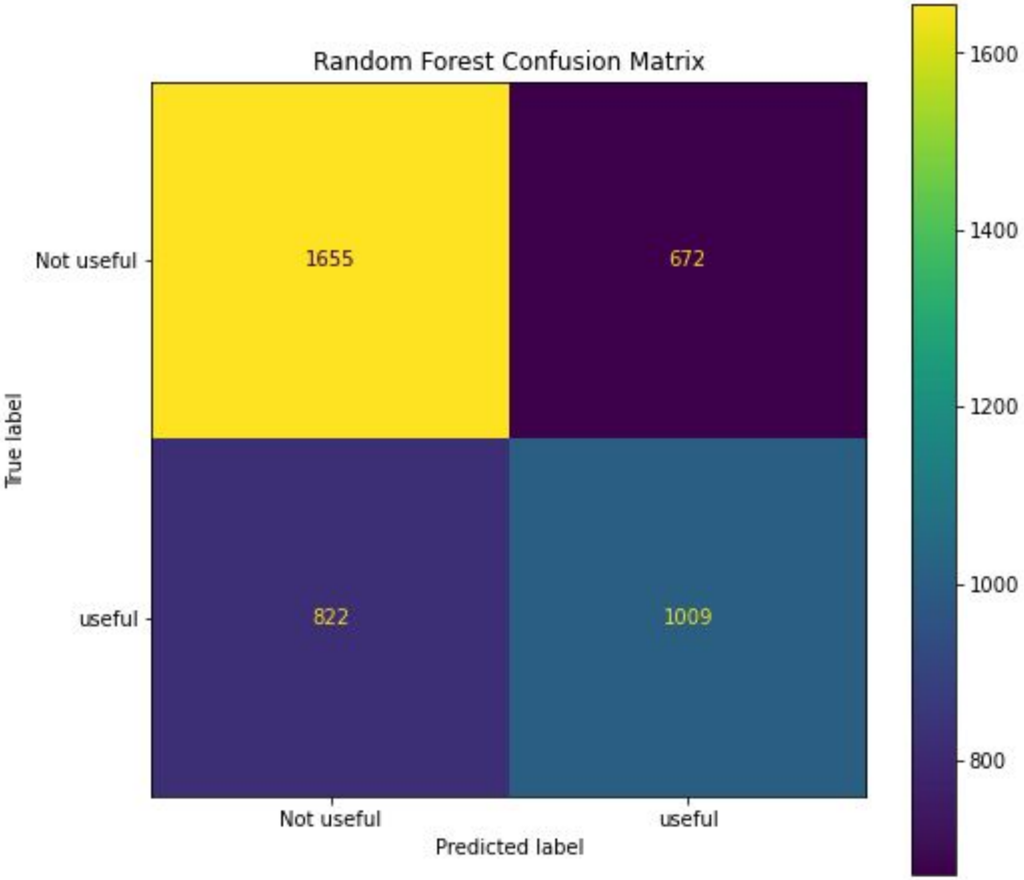


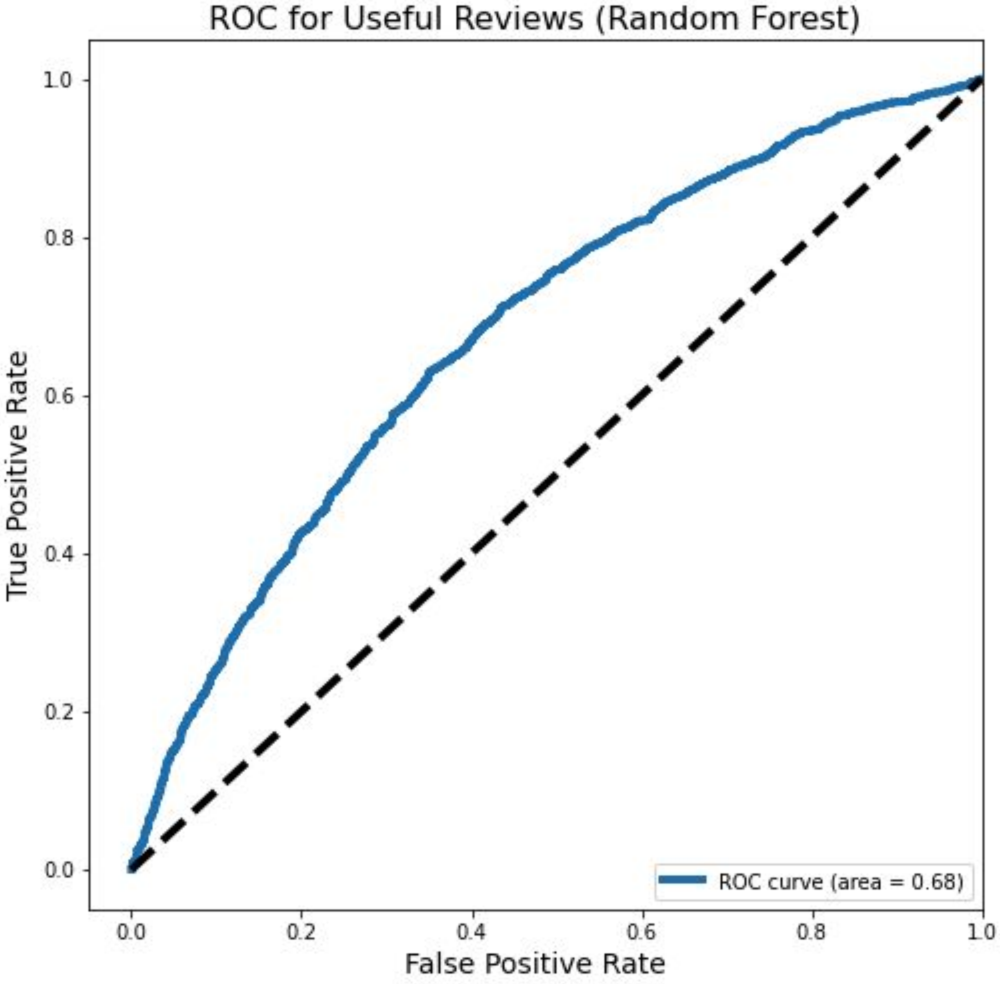
True label \ Predicted label	Not useful	useful
Not useful	1798	529
useful	1002	829


ROC for Useful Reviews (Logistic Regression)
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.67)

Random Forest

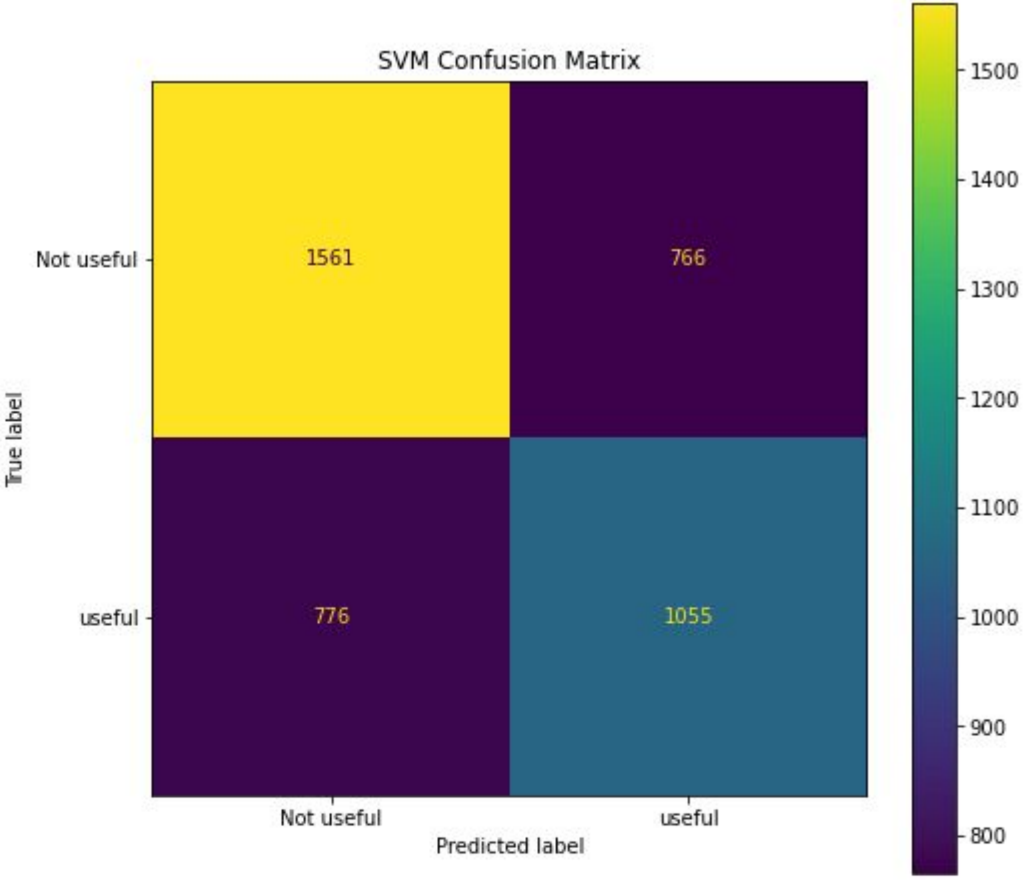
- Parameters
    - N_estimators: 200
    - Min_samples_leaf: 10
    - Warm start: true
- Metrics
    - 0.64 +/- 0.038
    - Precision: 0.60
    - Recall: 0.55
    - F1-score: 0.57

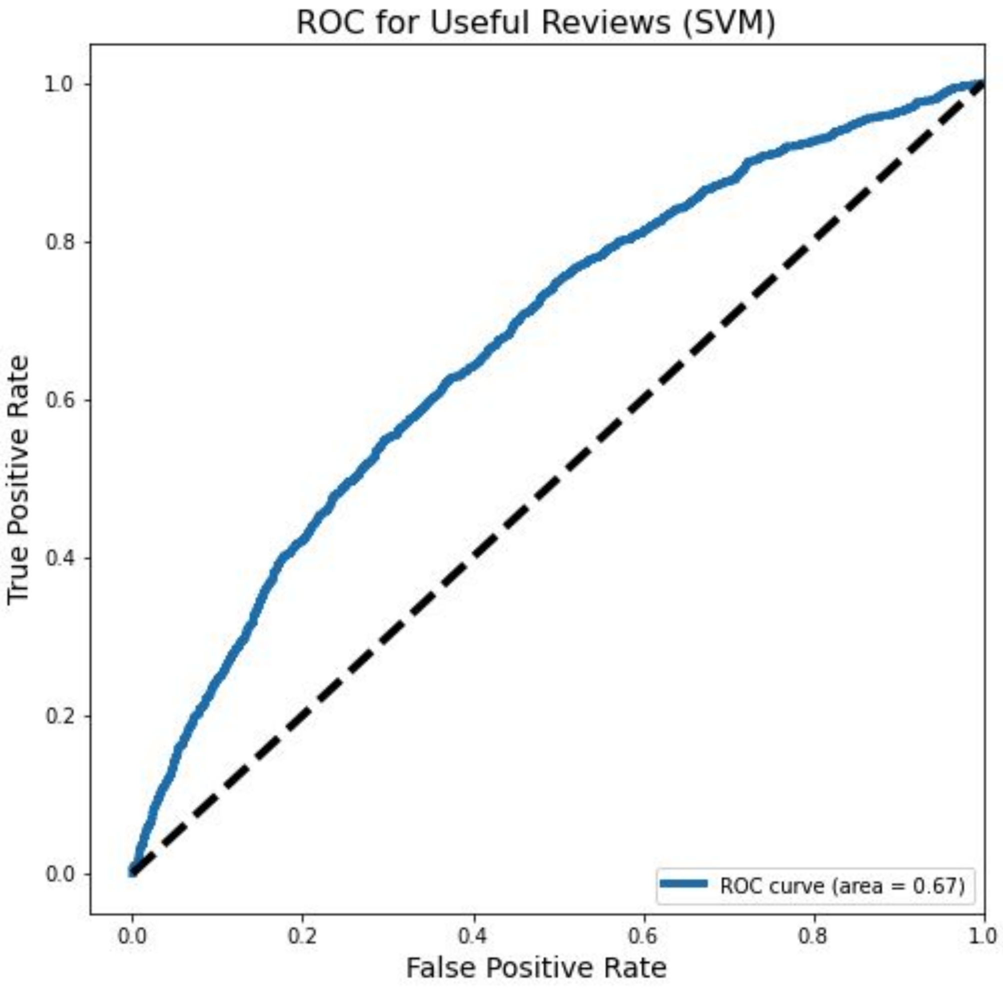

Random Forest Confusion Matrix
Not useful
useful
True label
1655
672
822
1009
Not useful
useful
Predicted label
1600
1400
1200
1000
800


ROC for Useful Reviews (Random Forest)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.68)

SVM

- Parameters
    - C=0.0001
    - Max_iter: 900
    - Class_weight: balanced
- Metrics
    - 0.62 +/- 0.043
    - Precision: 0.58
    - Recall: 0.58
    - F1-score: 0.58


SVM Confusion Matrix
True label
Not useful
useful
Predicted label
Not useful
useful
1561
766
776
1055
1500
1400
1300
1200
1100
1000
900
800


ROC for Useful Reviews (SVM)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.67)

Neural Network

- Parameters
  - 1 input, 3 hidden, 1 output layer
  - Kernel regularization = l2
  - Activation function = ReLU for non-output layers
  - Activation function = signmoid for output layer
- Metrics
  - 0.63

Themes

- Location!
  - Shop
  - Place
  - Store
- Praise
  - Great
  - Good
  - Like
- Temporal
  - New
  - Time

Best Model?

- All models beat the baseline (0.56)
- SVM?
  - Best true positive predictions
  - Lowest false negative predictions
  - Highest recall score
  - Highest f1-score
  - BUT score is less than Random Forest and NN
  - AND highest false positive rate

Conclusion

- Somewhat limited in how accurate we can get the models
    - Accuracy around 62-64%
- All models performed comparably
    - No clear winners based on multiple metrics
- Themes
- Future analysis might want to pull in business data
    - Location
    - Type