Detecting Fake Amazon Reviews With Machine Learning

Springboard Data Science Career Track

Capstone 2 by Chantel Clark

The Problem

- The majority of consumers (63.6%, Reviewtrackers) base their online purchase decisions on customer reviews and ratings
- Amazon is flooded with incentivized or paid customer reviews
    - 5-star glowing reviews to promote
    - Malicious reviews with low ratings for competing businesses
- Fake reviews are not limited to Amazon
- Why it is a problem:
    - Makes it difficult for the average company to have their products show up in user searches
    - Consumers are deceived

https://www.reviewtrackers.com/reports/online-reviews-survey/

Aim of this Project

Use machine learning to determine what makes a suspicious reviewer:

- Text from review body
- Reviewer behavior

The resulting model could be used in the development of a web application to:

- Help customers to determine the authenticity of a review
- Help businesses to eliminate fake reviews from their e-commerce site

Dataset

Open Amazon dataset of customer reviews (from AWS S3 bucket)

- U.S. Apparel subset, over 5.8 million reviews
- Years 1995 - 2015
- Each row represents one review
- Columns
  - ‘marketplace’ (country code), **‘customer_id’**, **‘review_id’**, ‘product_id’, ‘product_parent’, ‘product_title’, ‘product_category’, **‘star_rating’**, **‘helpful_votes’**, ‘total_votes’, **‘vine’**, **‘verified_purchase’**, ‘review_headline’, **‘review_body’**, **‘review_date’**
- 785 reviews contained no text in the review body, removed

Exploratory Data Analysis

1. What is the mean and median number of reviews per customer?
2. What do the reviews of a highly active (>500 reviews) reviewer look like?
3. What do the reviews that a one review customer look like?
4. Do customers who write different amounts of reviews give the same distribution of star ratings?

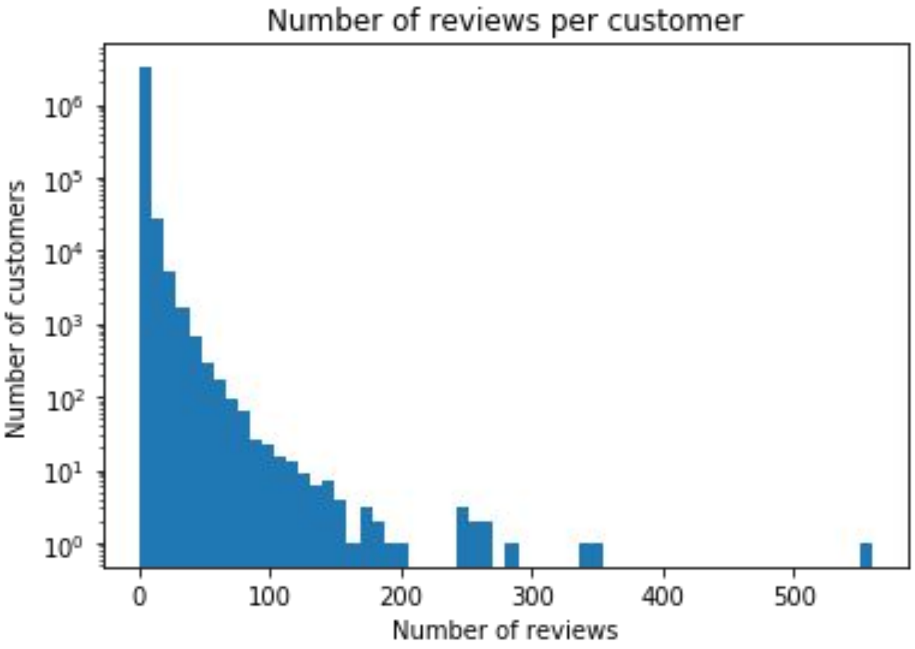
EDA: 1) What is the average and median number of reviews per customer?

Average:

1.83 reviews per customer

Median:

1 review per customer

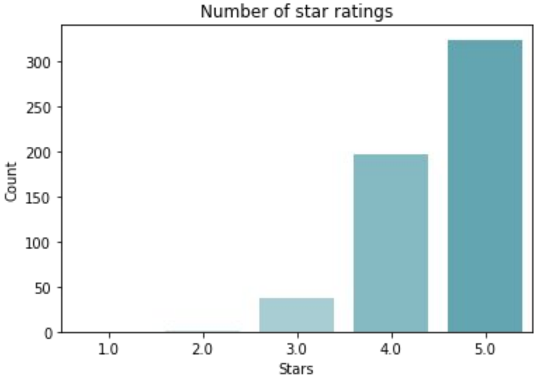


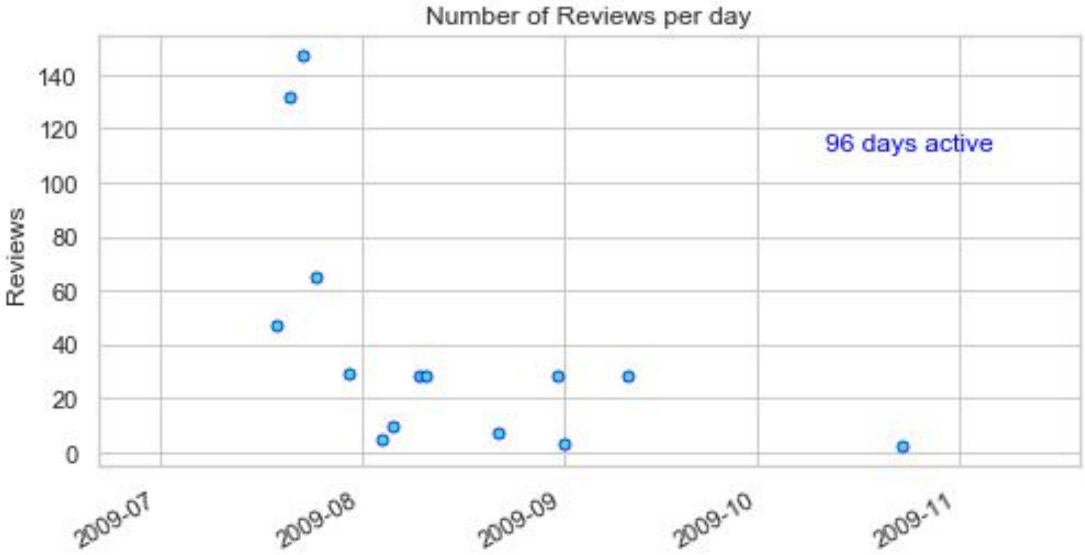
EDA: 2) What do the reviews of a highly active (>500 reviews) reviewer look like?

Total number of reviews:
559 (all unverified)

Max number of reviews in 1 day:
147

Average star rating: 4.5





EDA: 2) Sample reviews from most active reviewer

Stunning Cotton Kurti with gorgeous colored print. This ethnic kurti is master piece of Indian Ethnic Art.

This classic cotton boho tier skirt are perfect for all long and short kurtis. A fun & fashionable skirt by Mogul Interior! This printed skirt offers a bohemian style and tie design Skirt is stylish, easy to wear, and is one of this season's hottest looks! Cotton.

This classic cotton boho tier skirt are perfect for all long and short kurtis. A fun & fashionable skirt by Mogul Interior! This crinkle skirt offers a bohemian style and tie design Skirt is stylish, easy to wear, and is one of this season's hottest looks! Cotton.

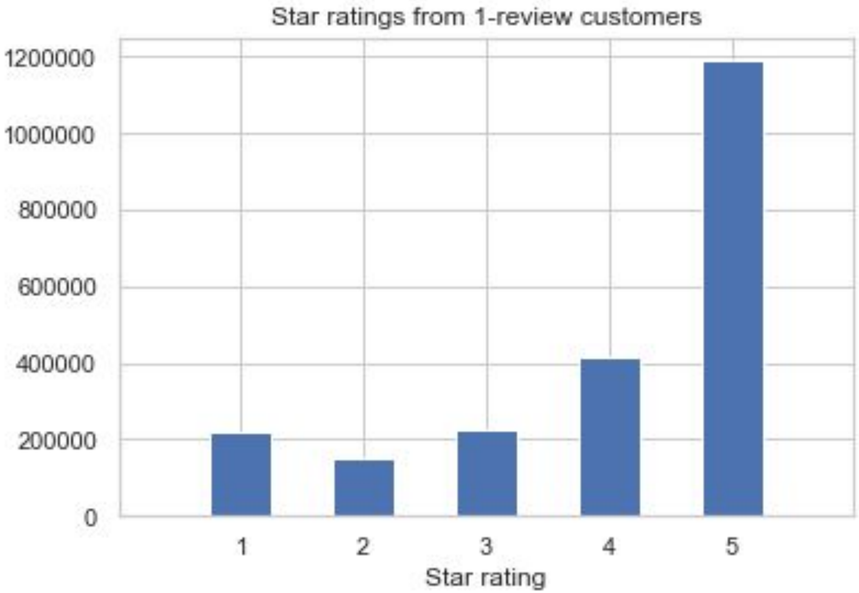
This classic cotton boho tier skirt are perfect for all long and short kurtis. A fun & fashionable skirt by Mogul Interior! This printed skirt offers a bohemian style and tie design Skirt is stylish, easy to wear, and is one of this season's hottest looks! Cotton.

EDA: 3) What do the reviews that a one review customer look like?

Random sample of 50 reviews inspected

Unverified reviews, with no ‘helpful’ votes **surprisingly well-written**

- *“Fits true to size BUT don't be alarmed when you put them on the hips are a little tight. Trust me that they will loosen to a comfortable fit.”*



EDA: 4) Do customers who write different amounts of reviews give the same distribution of star ratings?

Group	% of reviews	mean stars	median stars
A (1 review)	37.4	4.002	5
B (2-5 reviews)	40.1	4.130	5
C (6-9 reviews)	10.5	4.198	5
D (10-12 reviews)	3.6	4.228	5
E (13+ reviews)	8.4	4.280	5

Kruskal Wallis test:

Null hypothesis:

mean rank C = mean rank D = mean rank E

Alternative hypothesis:

at least 2 groups have different distributions

Result:

Test statistic =1298.9

P-value = 8.706e-283

At least two groups had significantly different star rating distributions

Feature and Label Selection

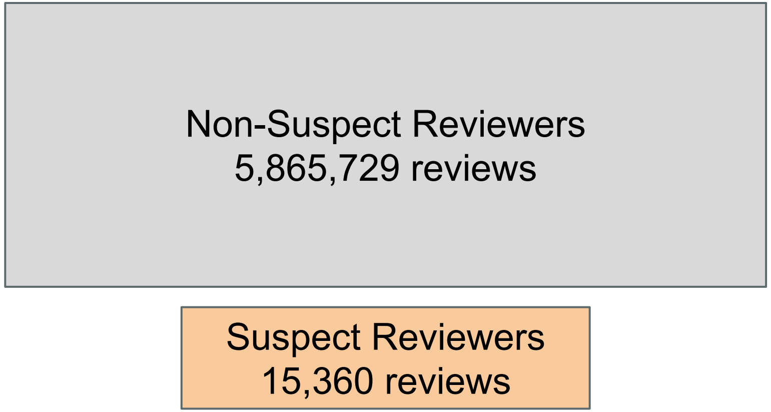
Suspect Label

0: Maximum number of reviews in a day < 30

1: Maximum number of reviews in a day ≥ 30

Features / Predictors

- Review body: text
- Star rating: ordinal category (1-5 stars)
- Helpful votes: integer
- Vine: category (‘y’ or ‘n’)
- Verified purchase: category (‘y’ or ‘n’)
- Cosine similarity of customer reviews

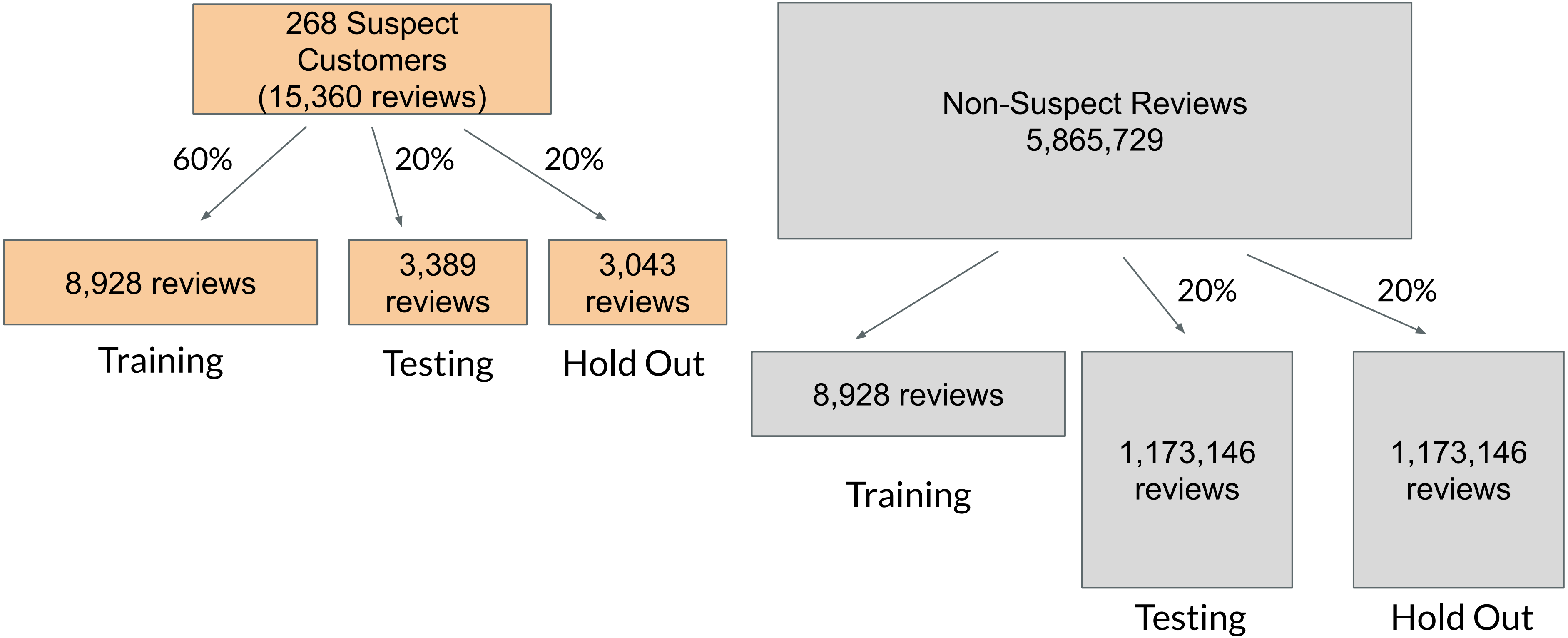


Non-Suspect Reviewers
5,865,729 reviews

Suspect Reviewers
15,360 reviews

0.26% suspect reviews

Modeling



268 Suspect Customers (15,360 reviews)

60%

20%

20%

8,928 reviews

Training

3,389 reviews

Testing

3,043 reviews

Hold Out

Non-Suspect Reviews 5,865,729

20%

20%

8,928 reviews

Training

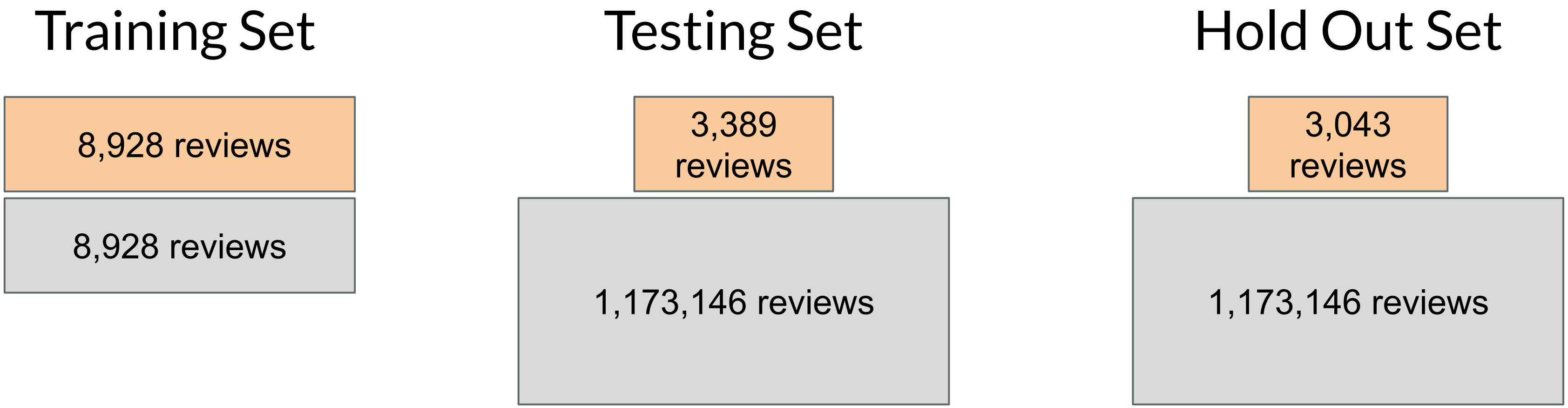
1,173,146 reviews

Testing

1,173,146 reviews

Hold Out

Modeling



Training Set

8,928 reviews

8,928 reviews

Testing Set

3,389 reviews

1,173,146 reviews

Hold Out Set

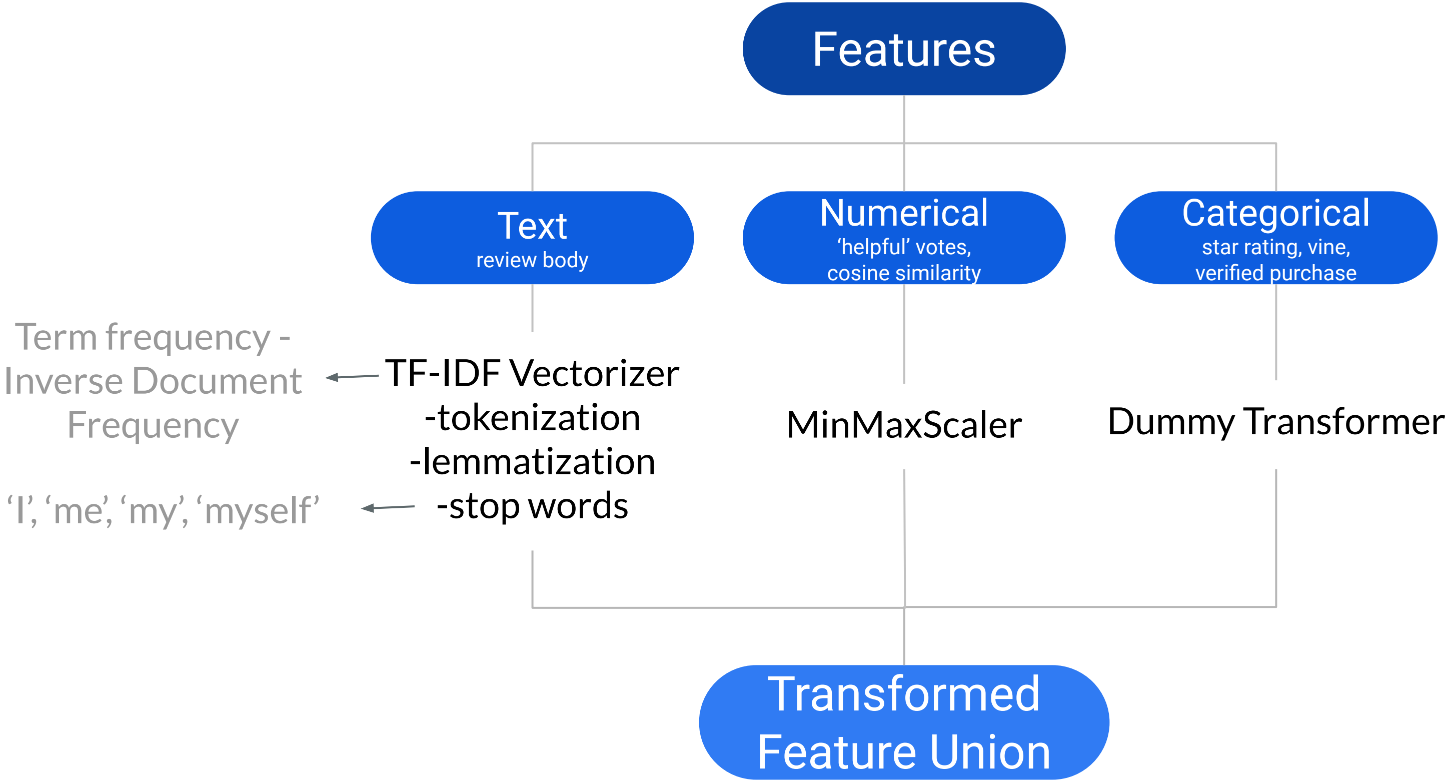
3,043 reviews

1,173,146 reviews

0.29% suspect labels

0.26% suspect labels

Feature Preprocessing Pipeline



Features

Text
review body

Numerical
'helpful' votes,
cosine similarity

Categorical
star rating, vine,
verified purchase

Term frequency -
Inverse Document
Frequency

TF-IDF Vectorizer
-tokenization
-lemmatization
-stop words

'I', 'me', 'my', 'myself'

MinMaxScaler

Dummy Transformer

Transformed
Feature Union

Model Selection

Binary classification, large number of observations and features (2819):

- Multinomial Naive Bayes: α = 1.0
- Random Forest: 200 estimators (trees), max depth = 10
- Linear SVM: L2 penalty, C=1.0

Model Evaluation

ROC AUC:

Summative measure to compare model true positive rate vs. false positive rate across different thresholds

PR AUC:

Summative measure to compare precision vs. recall across different thresholds

Recall:

What proportion of true suspect labels did the model catch?

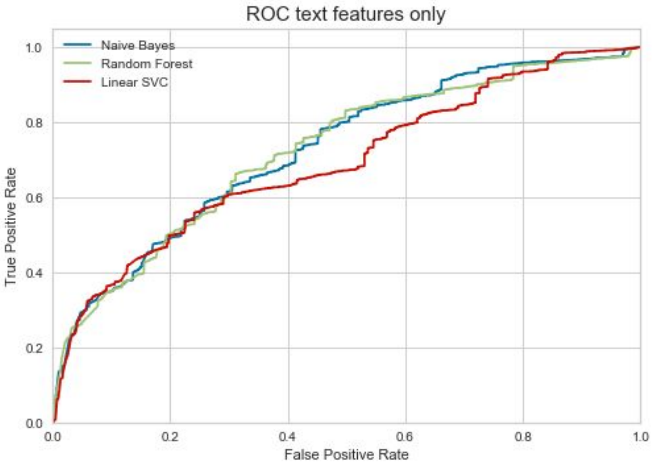
F2 measure:

A measure of precision and recall, giving more importance to recall

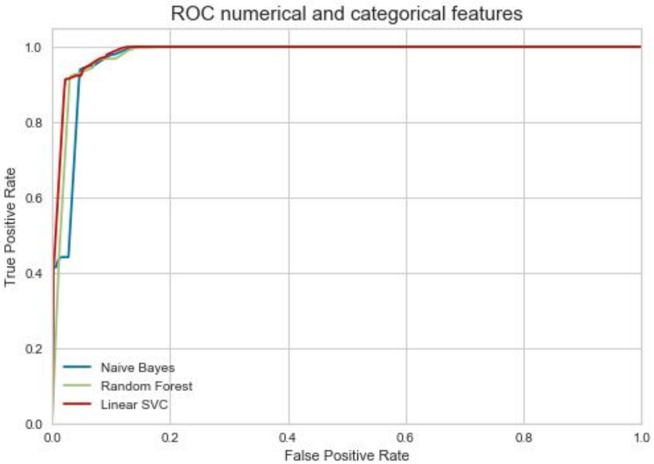
Findings (test set)

Features in Model	Number of Features	Naive Bayes	Random Forest	Linear SVC
Text only	2813	ROC AUC: 0.726 PR AUC: 0.012 Recall: 0.437 F2: 0.037	ROC AUC: 0.723 PR AUC: 0.012 Recall: 0.581 F2: 0.028	ROC AUC: 0.692 PR AUC: 0.010 Recall: 0.463 F2: 0.034
Numerical and categorical features	6	ROC AUC: 0.975 **PR AUC: 0.300** Recall: 1.00 F2: 0.043	ROC AUC: 0.980 PR AUC: 0.291 Recall: 0.998 F2: 0.087	**ROC AUC: 0.987** PR AUC: 0.176 Recall: 1.00 F2: 0.044
Text, numerical, and categorical features	2819	ROC AUC: 0.928 PR AUC: 0.048 Recall: 0.962 F2: 0.042	ROC AUC: 0.960 PR AUC: 0.085 Recall: 0.929 **F2: 0.142**	ROC AUC: 0.947 PR AUC: 0.049 Recall: 1.00 F2: 0.044

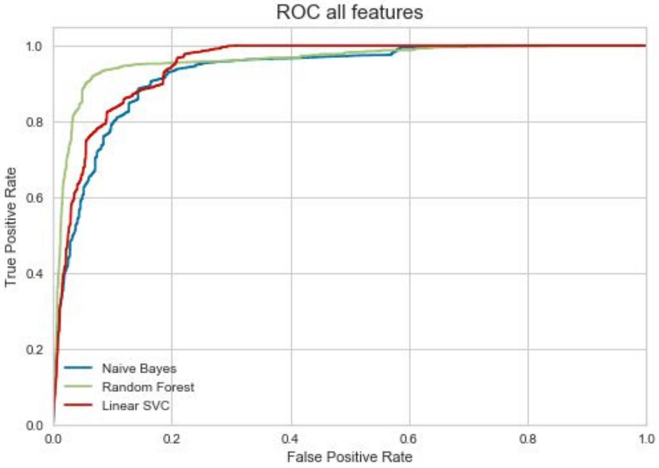
ROC curves



ROC text features only

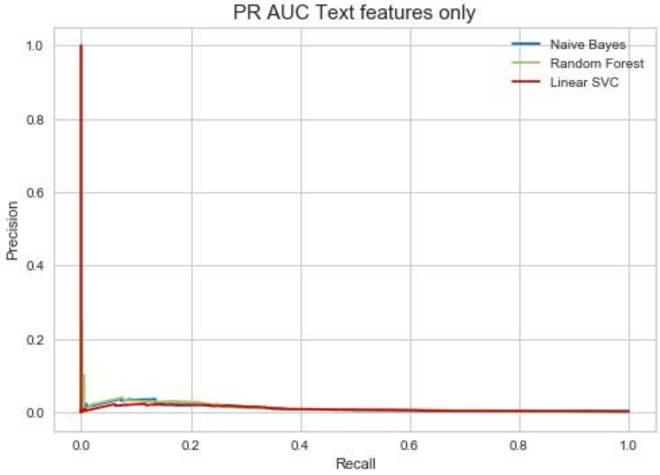


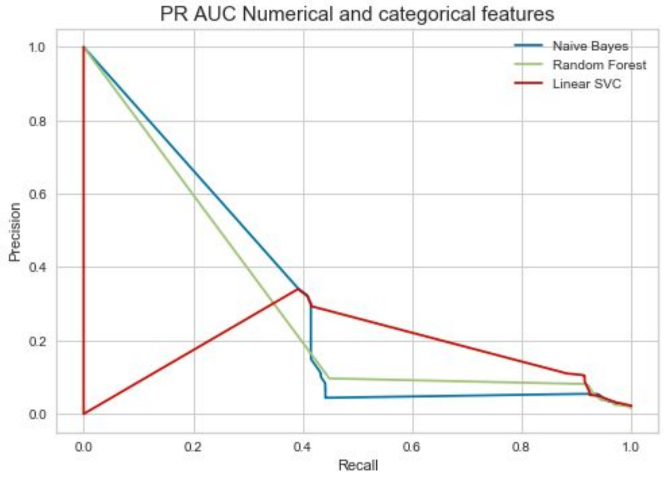
ROC numerical and categorical features

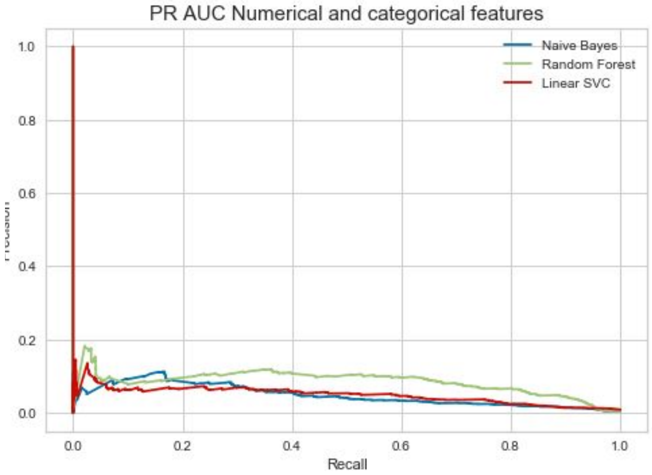


ROC all features

Precision Recall (PR) curves

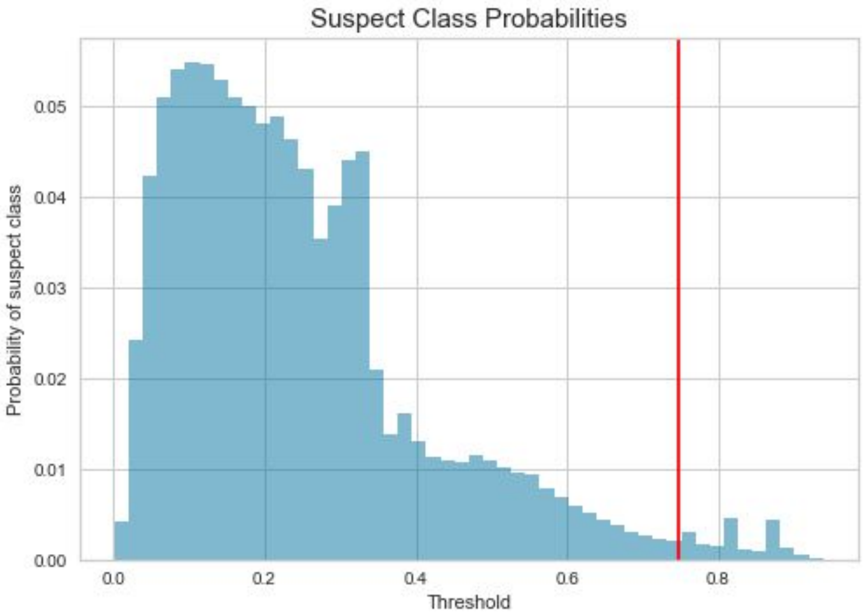

PR AUC Text features only

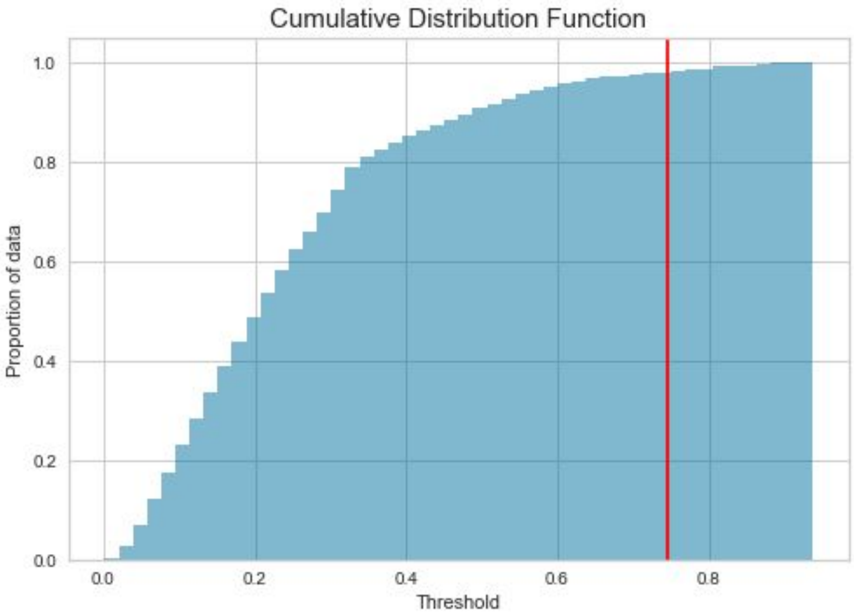

PR AUC Numerical and categorical features


PR AUC Numerical and categorical features

Threshold tuning: Random Forest all features

A maximum F2 measure was obtained on test set, when threshold = 0.746


Suspect Class Probabilities


Cumulative Distribution Function

Findings (hold out set)

Model	Number of Features	Test Set	Hold out set
Naive Bayes (numerical and categorical)	6	ROC AUC: 0.975 **PR AUC: 0.300** Recall: 1.00 F2: 0.043	ROC AUC: 0.964 PR AUC: 0.067 Recall: 0.156 F2: 0.116
Linear SVC (numerical and categorical)	6	**ROC AUC: 0.987** PR AUC: 0.176 Recall: 1.00 F2: 0.044	**ROC AUC: 0.982** PR AUC: 0.071 Recall: 0.073 F2: 0.069
Random Forest (numerical and categorical)	6	ROC AUC: 0.980 PR AUC: 0.291 Recall: 0.998 F2: 0.087	ROC AUC: 0.966 **PR AUC: 0.107** **Recall: 0.862** F2: 0.250
Random Forest (all features)	2819	ROC AUC: 0.960 PR AUC: 0.085 Recall: 0.929 **F2: 0.142**	ROC AUC: 0.974 PR AUC: 0.075 Recall: 0.658 **F2: 0.283**

Summary

- Models with text features only had lowest performance
- Best models (Random Forest) included all features (text, numerical and categorical data), attained ROC AUC 0.974, recall 65.8%, and F2 Measure 0.283

Next steps:

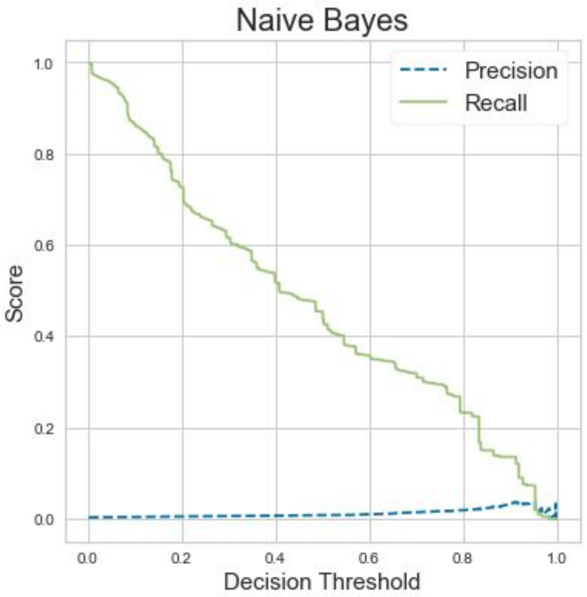
Increase target labels by adding criteria or using repeated

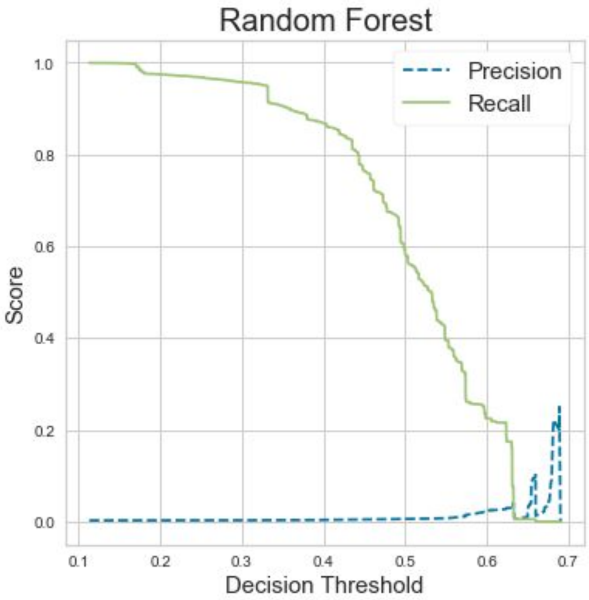
Feature engineering: emojis, sentiment, active time

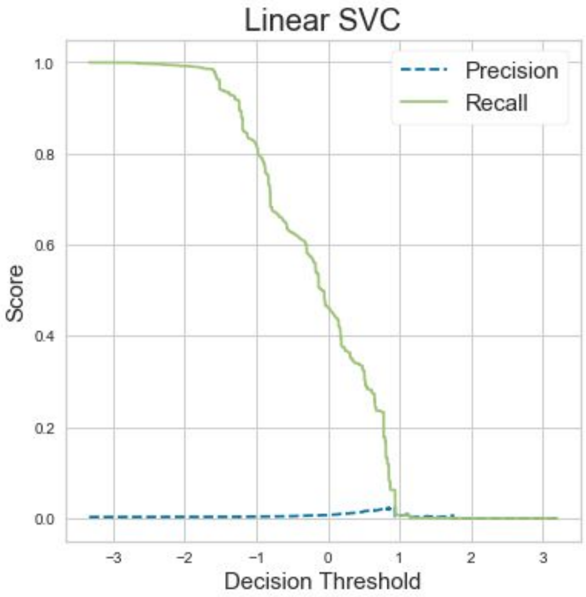
Suspect reviews

Extra images

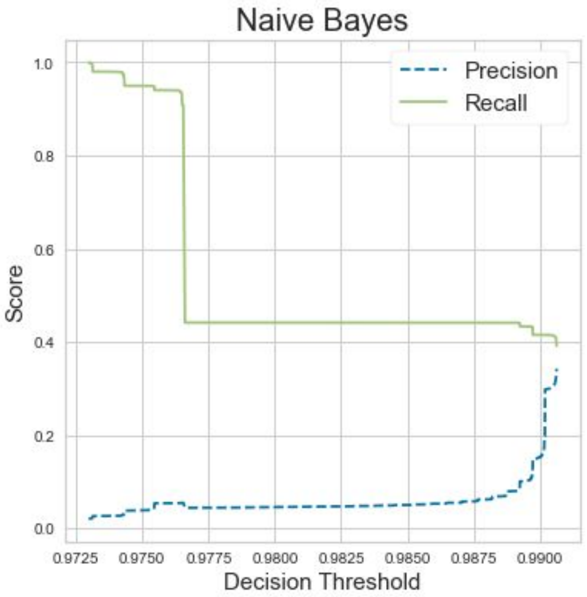
PR vs Threshold - text features only



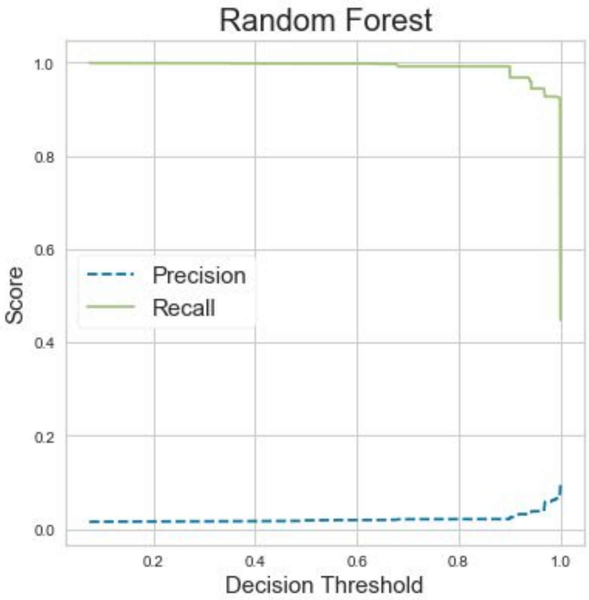




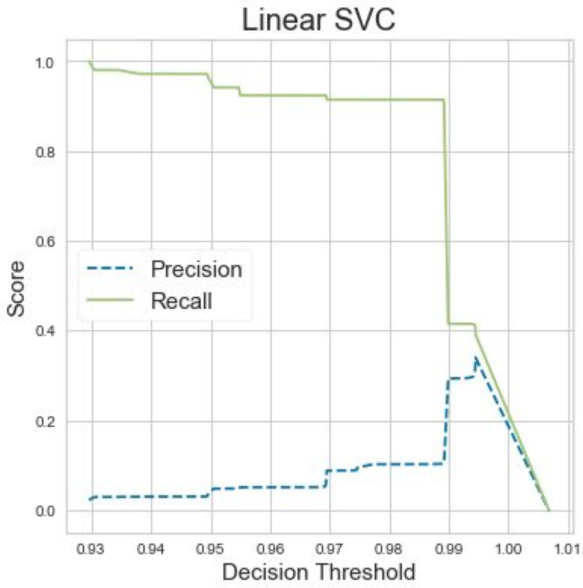
PR vs Threshold - numerical and categorical features



Naive Bayes

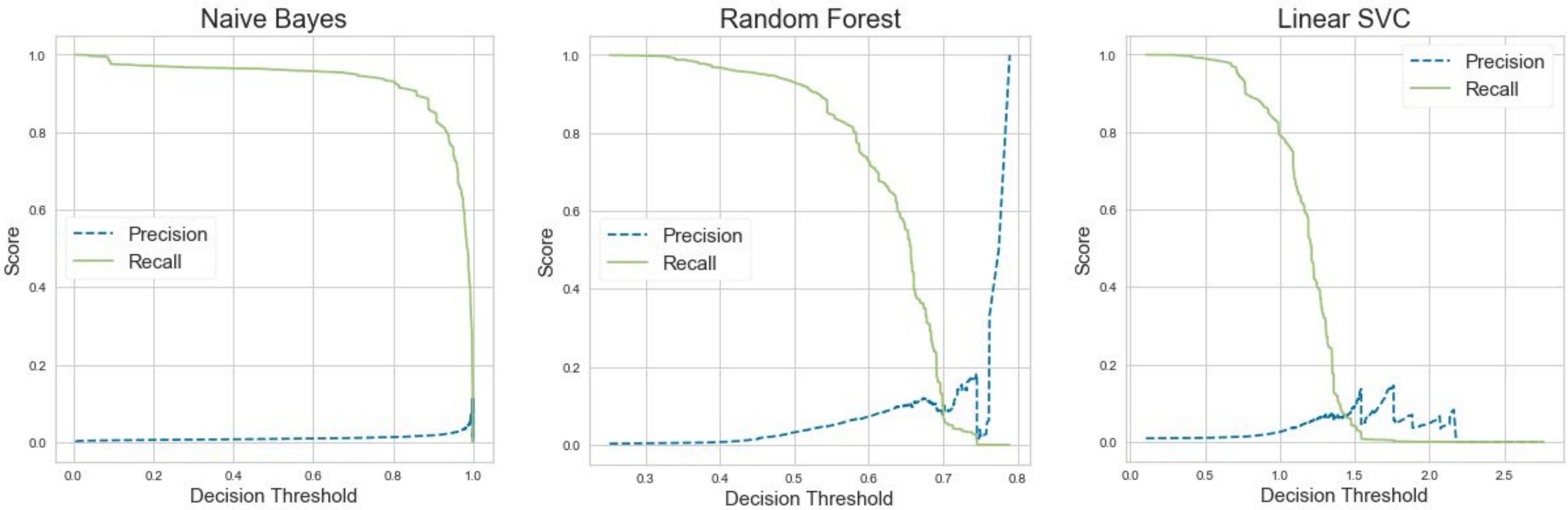


Random Forest

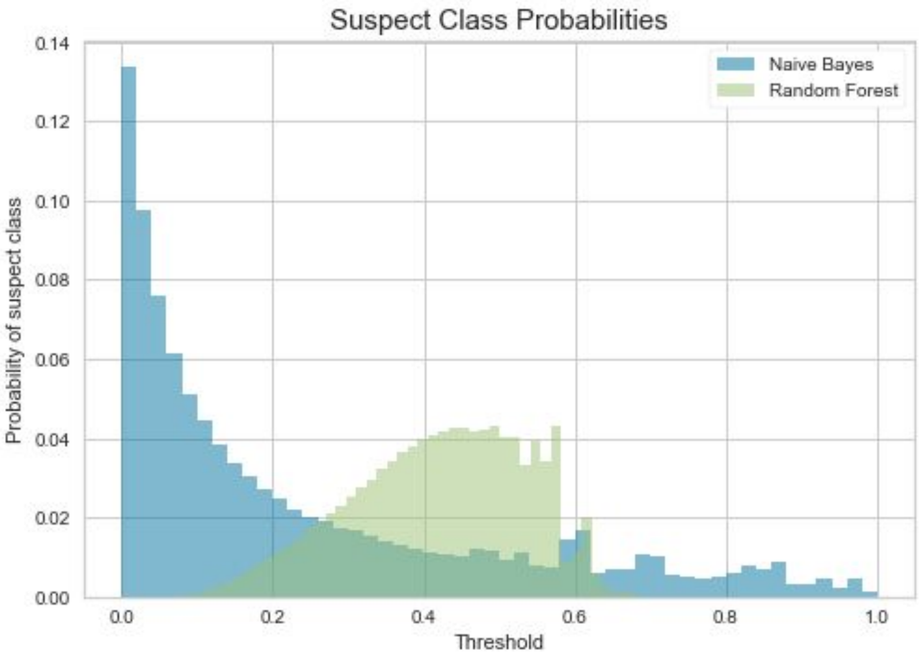


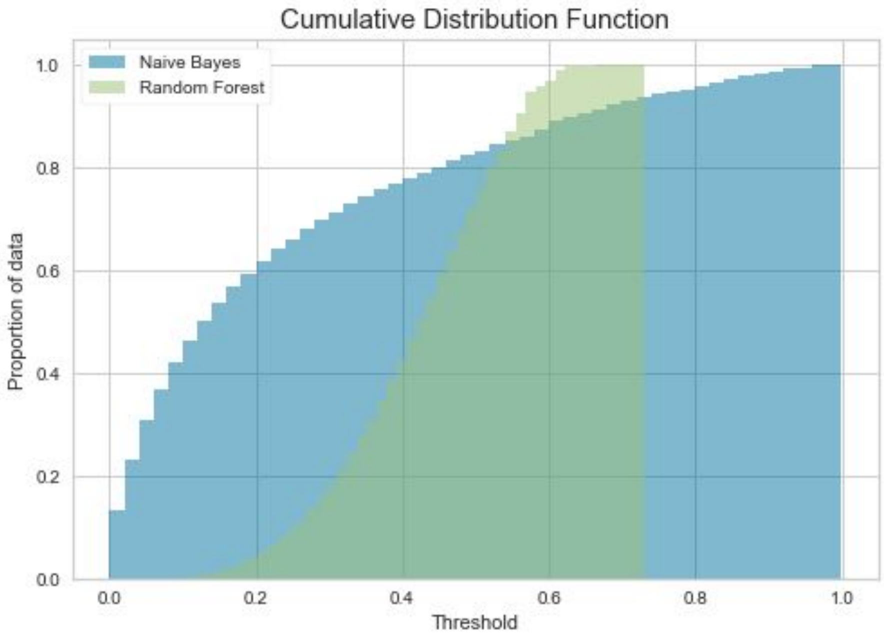
Linear SVC

PR vs Threshold - all features

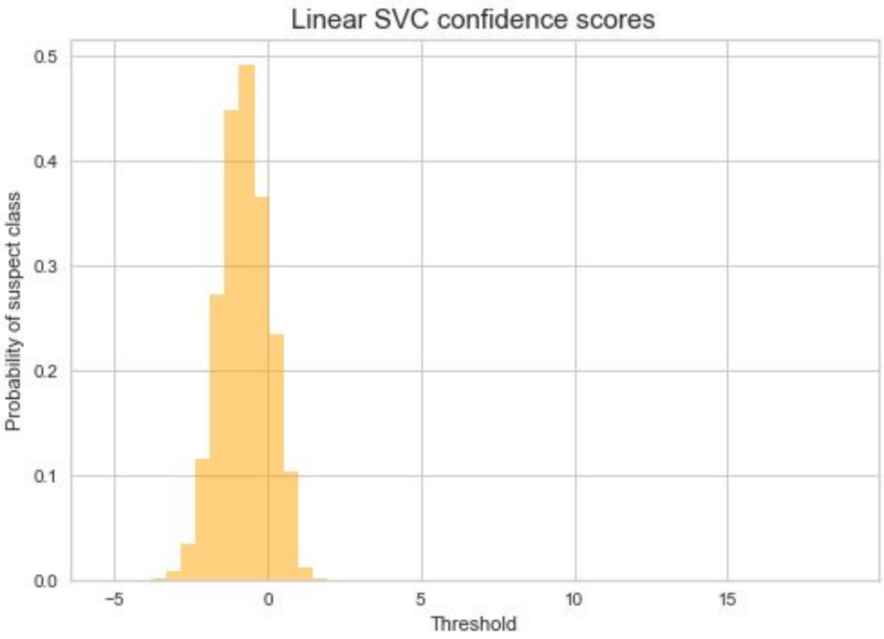


Class probabilities and CDF’s

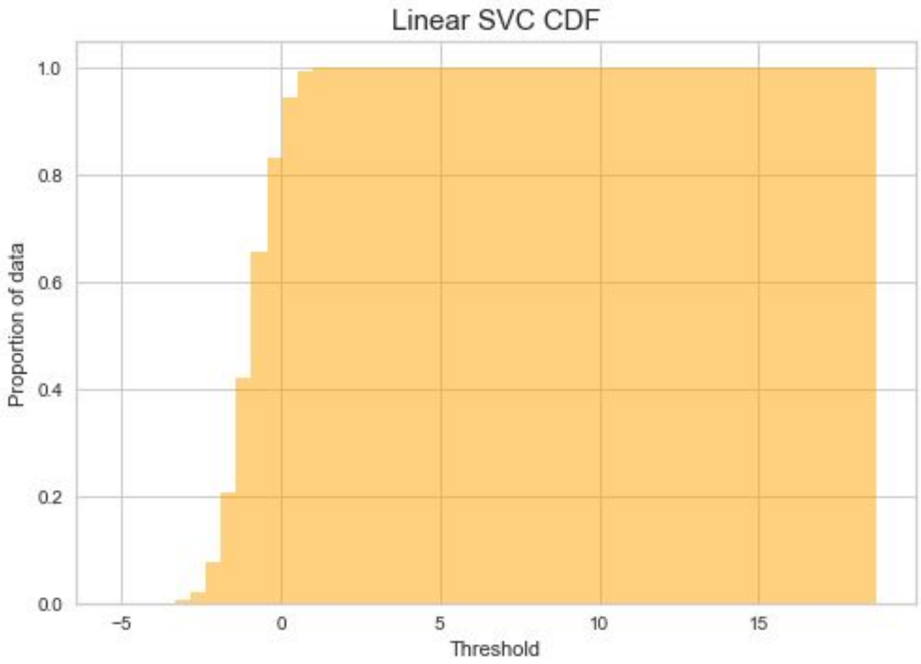




Class probabilities and CDF’s



Linear SVC confidence scores



Linear SVC CDF